2018

Monitoring Infection of E. Coli By Bacteriophage T7 through Mass Spectrometry-Based Proteomics

Amy Swayne Schienschang
William & Mary - Arts & Sciences, amy.schienschang@gmail.com

Monitoring Infection of *E. coli* by Bacteriophage T7 through Mass Spectrometry-Based Proteomics

Amy Swayne Schienschang

Orange, Virginia

Bachelor of Science, Randolph College, 2009

A Thesis presented to the Graduate Faculty of The College of William & Mary in Candidacy for the Degree of
Master of Science

Department of Chemistry

College of William & Mary
January 2019

APPROVAL PAGE

This Thesis is submitted in partial fulfillment of
the requirements for the degree of

Master of Science

Amy Swayne Schienschang

Approved by the Committee November 2018

Committee Chair
Garrett-Robb-Guy Professor John Poutsma, Chemistry
College of William & Mary

Professor Lisa Landino, Chemistry
College of William & Mary

Assistant Professor Rachel O'Brien, Chemistry
College of William & Mary

ABSTRACT

Mass spectrometry-based proteomics has become an important and versatile tool in analytical chemistry, making sense of complex biological samples and shedding light on the intricate proteomes of living organisms. Bottom-up proteomics studies are used to elucidate the changes in gene expression of bacteriophage T7 over the course of infection of *Escherichia coli*. *E. coli* cultures were infected with T7, sampled over time, and proteins were isolated and enzymatically digested. Nanoflow liquid chromatography combined with tandem mass spectrometry was used to detect proteolytic peptides and identify host and phage proteins. Generally, phage proteins were detected on a time scale fitting the established lytic cycle for T7 phage, confirming the effectiveness of infection monitoring by mass spectrometry-based proteomics studies. Continued development of the experimental method sought to increase detection of proteolytic peptides and identify phage and host proteins to a higher level of confidence, and lead to the implementation of 1D SDS-PAGE as a fractionation method to reduce sample complexity and increase method sensitivity.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The writer would like to thank Professor Poutsma for his patience, understanding, and teaching her everything she knows about mass spectrometry. She would also like to thank Professors O'Brien and Landino for their time and constructive guidance on this manuscript. The contribution of *E. coli* samples for analysis from Professor Williamson and his lab were integral to this work. The writer especially thanks the stellar proteomics team of Christian Harris and Theresa Gozzo who assisted in this project and the Ion Lab at large.

This thesis is dedicated to my family, who listened to me ramble about proteins ad nauseam, to Tony, who always supported me and offered needed proofreading, and to Mac, who has always believed I am a good “sciencer.”

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction to Mass Spectrometry-Based Proteomics

1.1 Mass Spectrometry Based Proteomics

Twelve years after *Nature Biotechnology* declared the coming of a grand new technique for drug design, identification of biomarkers, and so much more, *Nature Methods* declared that mass spectrometry-based proteomics was finally ready for the "big time".[1-2] Advances in instrumentation, computing power, and sample preparation methods have made the once disappointing and unreliable field of mass spectrometry-based proteomics into the industry standard for high throughput protein analysis. Issues with reproducibility in early, high-profile studies and a failure to deliver on promises of revolutionary new work on biomarkers for disease diminished the early reputation of proteomics work,[2] but careful implementation of controls have rehabilitated the field. The rapid increase in genome sequencing technology since the beginnings of proteomics studies has consequently elevated the field. Though the term "proteome" refers to the entire complement of proteins that can be produced from a given genome, proteomics studies can be on the global protein complement of an organism or on more localized and targeted studies of a specific protein. From the growing list of species with fully sequenced genomes came new opportunities to study the protein products of these genomes. Attempting to characterize the entire protein complement of a genome is incredibly difficult, as all genes are not expressed equally at all times. Even single-cell organisms vary their gene expression with environmental conditions and cellular needs at any given time. The variety and amount of proteins will vary across even a single cell depending on the

intracellular location sampled. The entire array of possible proteins encoded in an organism's DNA is unlikely to be observed by any one analysis.

Despite being unable to practicably detect all possible products of a given genome, mass spectrometry-based proteomics studies offer a wealth of information about the system studied. Proteins present in the cell at a given time are a function of the cellular environment and can provide valuable insight into the state of the organism. One of the challenges of proteomics lies in the fact that proteins that are present in the cell will vary in concentration over time and low-abundance proteins can be particularly hard to detect in the amalgam of higher-abundance proteins. Differential proteomics experiments seek to focus analysis on the changes in protein expression as a result of some cell stressor rather than attempting to characterize all proteins in a sample. For example, by studying only the proteins that showed noticeable changes in concentration visible by 2-D gel electrophoresis, Ogada et al. were able to track the immune response of Western flower thrips, a common agricultural pest insect, when faced with viral infection.[3] Of the thousands of proteins present in an organism, especially one as complex as an insect, only 30 showed significant changes in concentration over the course of the infection. Zeroing in on the differences significantly cut down on analysis time and resources while elucidating valuable information about immune response. Differential proteomics experiments can also be used to identify potential early biomarkers of disease. In 2018, Aslebagh et al. identified several proteins that were differently expressed in human breast milk samples between the precancerous and healthy breasts of a woman that was diagnosed with

cancer in only one breast 24 months following sampling.[4] With further study, these early indicators of disease could point to new screening techniques for these biomarkers for heightened breast cancer risk.

Mass spectrometry-based proteomics has been bolstered as a field with the rapid improvement in DNA sequencing techniques as proteomics studies generally rely on sequenced genomes. Shortly after being declared “ready for the big time” by Nilsson et al. in 2010, the library of completely sequenced genomes included 3,969 prokaryotes, eukaryotes, and viruses[5] and has steadily grown in the intervening years. With sequenced genomes, potential protein products can be predicted. While mass spectrometry is a great technique to sequence proteins de novo, this is a time-consuming and labor intensive proposition even for an isolated protein. With the complex samples typical of cellular digests, sequencing proteins de novo from a veritable soup of peptides would be nearly impossible. Using sequenced genomes and the predicted protein products thereof, mass spectrometrists can instead identify observed peptides to reconstruct the proteome from the pieces. This technique of identifying proteolytic peptides to build up protein structures is referred to as “bottom-up” proteomics and can be employed for a purified protein or for a mixture of proteins, such as a result of cellular digestion or tissue extraction. Because of the similarity to shotgun genomic sequencing, where DNA is cleaved by various restriction enzymes and then sequenced using the assignment of overlapping fragments, analysis of a mixture of proteins is commonly called “shotgun” proteomics.

Alternately, proteomics experiments can be conducted in a “top-down” manner by analyzing intact proteins or a “middle-down” method that analyzes partially digested proteins. These two methods require mass spectrometers with high resolution and high mass ranges, such as a time of flight or Fourier-transform ion cyclotron resonance instruments. Due to instrumental availability, this work will be focused on bottom-up proteomics, which can be successfully implemented on a wider range of mass spectrometers.

1.2 Bottom-Up Proteomics Methodology

A bottom-up proteomics study begins with protein digestion, usually completed with a slate of proteases. Trypsin is far and away the most commonly used protease in proteomics assays. As trypsin is a serine protease that cleaves peptide bonds on the C-terminal side of arginine and lysine, peptides from tryptic digestion often acquire multiple positive charges when ionized using electrospray ionization. This can be advantageous as it enables use of a wide range of fragmentation methods in tandem mass spectrometry. Trypsin can be used in isolation or in conjunction with other common proteases. Enzymatic digestion selectively cleaves peptide bonds depending on the residues targeted by a specific enzyme, so the use of multiple enzymes can create complementary coverage of protein sequences. Protease selection is based on a variety of factors including target protein primary sequence and instrumentation mass range.

To ensure better peptide cleavage, steps are taken to prepare the sample for enzymatic digestion. Treatment of cell lysates with detergents and buffers

helps to disrupt hydrophobic and electrostatic interactions that contribute to protein folding. Reagents are added to disrupt disulfide bonds and guard against reformation of these bonds with protecting groups. Disulfide bridges can lead to cross-linking between protein strands that would create a nearly insolvable mess of peptides with little correlation to the predicted proteins from the organism's genome. Purposeful cross-linking in proteins can be used to elucidate interactions in a protein's quaternary structure, but this manner of study would certainly require isolation of the protein of interest. As proteins and their component peptides are identified by precursor mass and subsequent fragmentation patterns, cross-linking would lead to precursor and product ion masses that could not be assigned by automated means.

To analyze incredibly complex mixtures of peptides, separation methods are required before analysis with the mass spectrometer. While mass spectrometers are capable of analyzing simple mixtures as analytes are detected by their distinctive mass to charge ratio, even the highest resolution instrument could not identify the thousands of peptides present in a typical shotgun sample. The resulting mass spectrum would be essentially impossible to assign peptides to peaks. Separation using strong cation exchange chromatography (SCX),[6] 1 or 2-D gel electrophoresis,[4] and size-exclusion chromatography[7] has been used prior to mass spectrometry to great success to lower complexity of shotgun samples. For example, to obtain a quantitative profile of the human plasma proteome, Wang et al. used a combination of solution isoelectric focusing (IEF), liquid chromatography (LC), and 2-D difference gel electrophoresis (2DIGE) prior

to quantification with mass spectrometry (MS).[8] Fractionation methods are not required for all shotgun proteomics studies, especially where qualitative analysis is sufficient. However, fractionation does produce discrete samples that are required for use in matrix-assisted laser desorption/ionization (MALDI) analyses. MALDI provides efficient peptide ionization but requires additional sample preparation that can be difficult to integrate with LC-MS/MS usage. In contrast, coupling LC to electrospray ionization-MS (ESI-MS) allows for in-line analysis of shotgun samples and is therefore widely used in this field. On-line analysis of the proteolytic peptides allows for separation and detection to be completed on one instrument, in one step increasing the efficiency and throughput.

Nano-flow liquid chromatography and nanoelectrospray ionization (nESI) have greatly increased sensitivity over ESI-MS due to more efficient peptide ionization and separation. Both ionization methods use in-solution protonation to create ions in the sample solution prior to vaporization with the addition of a weak acid. As the name suggests, nanoelectrospray operates in the nanoliter per minute flow rate range, while ESI operates at a microliter per minute flow rate. Both techniques aspirate charged droplets of sample in solvent, which then decrease in size as solvent evaporates. This concentrates the positive charge until Coulombic repulsion explodes the charged sample ions free of the solvent, creating gas-phase ions for MS analysis. The difference in initial droplet size, in the μm range for ESI and roughly 180 nm for nESI, leads to different droplet fission pathways which result in higher rates of analyte ionization and more favorable signal to noise ratios for nESI.[9]

Whether ESI or nESI is used for ionization, a shotgun proteomics analysis will generate an enormous quantity of spectral data. As peptides elute from the LC, mass spectra are continuously taken and ions are fragmented in a data dependent manner. The mass range is scanned for peaks with intensities above a given signal strength and the four or five most abundant peaks are identified. These precursor ions are isolated in subsequent scans for fragmentation by a variety of methods. In a process that can span hours, ions are isolated and spectra are recorded on a millisecond timescale continuously throughout the chromatographic run. If the same precursor mass is identified more than once in short time period, it will be excluded temporarily from fragmentation to allow the mass spectrometer to examine lower-abundance peptides.

1.3 Peptide Fragmentation and Identification

Mass-selected fragmentation of selected precursor ions can be accomplished by various methods, most commonly collision induced dissociation (CID) and electron transfer dissociation (ETD). Both fragmentation methods will cleave the precursor ion along the peptide backbone, resulting in a characteristic pattern of fragment masses that can be used to identify the precursor ion. Product ions are designated by a letter and number identifier, indicating the length of the fragment and which end of the precursor from whence it was generated. The designation of b, c, y, and z ions depends on the location of the positive charge: remaining on the N-terminal side of the cleaved peptide for b and c ions and on the C-terminal end of the fragment for y and z ions. As seen in Figure 1, a and x ions can also be formed, but are produced from high-energy

fragmentation methods seen in a magnetic sector mass spectrometer, for example, or from secondary degradation of b or y ions.

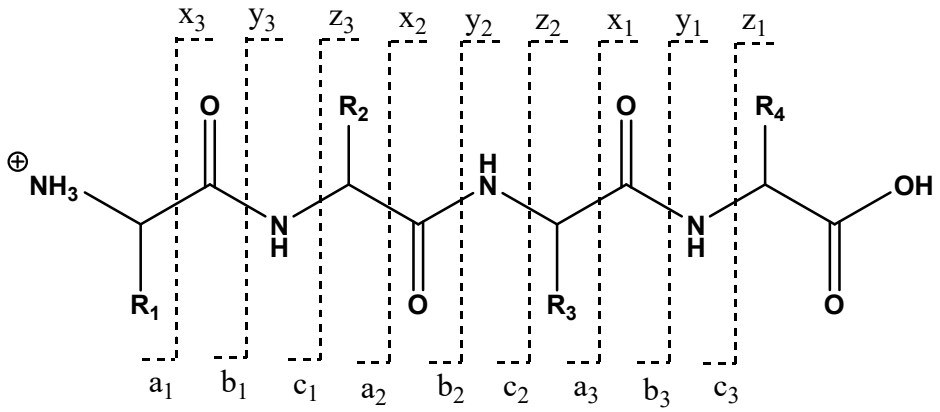

Figure 1: Peptide fragmentation locations along backbone.

CID uses the introduction of inert collision gas to collide with precursor ions to cause fragmentation and tends to result in product ions created from cleavage of peptide bonds, labelled b and y ions in Figure 1. CID is the most common and robust fragmentation method employed in bottom-up proteomics,[5] but cannot be used to study post-translational modifications (PTMs) of proteins, such as phosphorylation. Phosphorylation, the addition to and removal of phosphate groups from proteins, is often used in biological systems as important signaling mechanisms and thus these groups are of particular interest in many proteomics assays. Proteins can also be modified by adding carbohydrate or lipid groups, through glycosylation or lipidation, respectively, or a variety of other modifications to the side chains or the terminal groups. These modifications occur after the protein is translated from genetic material, and are therefore not encoded in the organism's DNA or RNA and can only be elucidated through protein analysis, as opposed to genetic analysis. Post-translational modifications

such as phosphorylation are lost in CID analyses as the sidechain modification bond is more labile than those of the backbone, as CID activates and cleaves the lowest energy bond. Therefore, if a phosphorylated peptide were to be activated with CID, the phosphate modification would be cleaved from the peptide and detectable only in the mass spectrometer as a neutral mass loss.

To preserve post-translational modifications, electron transfer dissociation can be used as an alternative or complementary fragmentation method to collision induced dissociation. ETD can only be used for multiply-charged peptides, as an electron is transferred from an electron-rich donor reagent to the positively charged peptides. Singly charged peptides would become neutral upon electron transfer and therefore undetectable in the mass spectrometer. The electron transfer initiates a radical process that results in cleavage of the peptide along the backbone, predominantly resulting in c and z ions as seen in Figure 2.

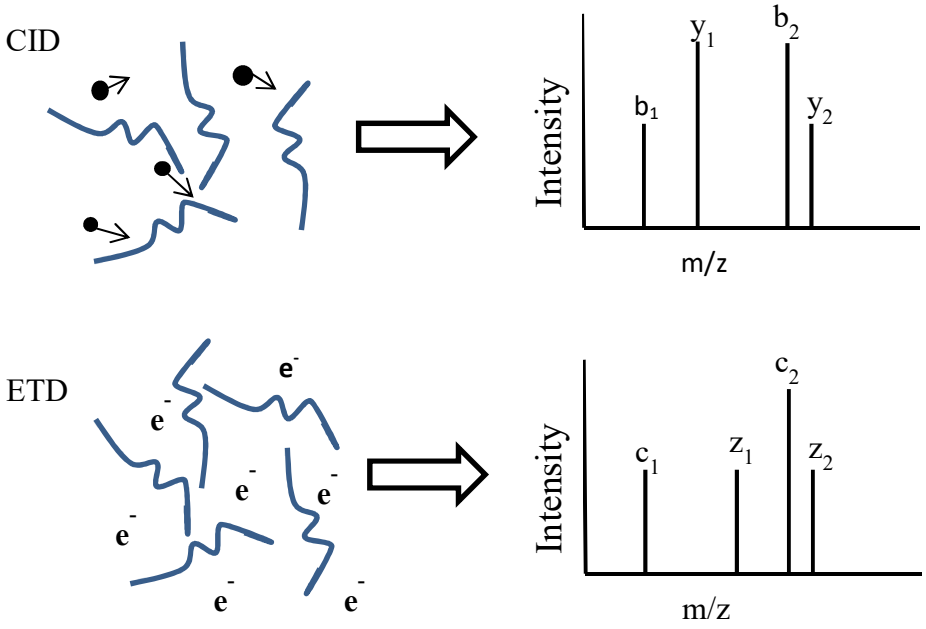

Figure 2: Expected product ions produced by common fragmentation methods.

As this dissociative process is localized to the peptide backbone, sidechain modifications remain intact and identifiable in the fragmentation spectra.

Regardless of the fragmentation method, precursor ion masses and product ion spectra are collected throughout the chromatographic run in proteomics studies, and the resulting thousands of mass spectra are aggregated. De novo sequencing and manual interpretation of all of the ion peaks is impractical for a sample size this large with this complex of a mixture. For organisms with sequenced genomes, theoretical proteins are predicted from the DNA and expected peptides are produced from *in silico* digestion using commercial bioinformatics software. Precursor masses detected in the proteomics analysis are compared to predicted peptide masses based on the sequenced proteome, and theoretical product ion spectra are generated for each peptide to compare to the experimental fragmentation spectra. Bioinformatic programs such as MASCOT and SEQUEST correlate the experimental and theoretical precursor and product ion spectra to identify proteins present in the shotgun sample. SEQUEST largely uses comparison of experimental product ion spectra to product ion spectra generated from the sequenced protein database and provides a statistical measure of the correlation, called XCorr.[5] The higher the XCorr value, the more confidently the identity of the peptide is assigned. MASCOT also incorporates mass fingerprinting when coupled with high-resolution mass spectrometry, using the exact mass to within 10ppm of the proteolytic peptides to identify proteins. Using either method, fragmentation spectra are used to identify peptides that are then pieced together to identify

proteins. The sequential construction of smaller to larger pieces of information is indicative of bottom-up proteomics. When controls are implemented carefully shotgun proteomics experiments can identify singular proteins, even those in low abundance, in the mess of a complex cellular digest eliminating or reducing the need for protein purification before analysis.

Chapter 2: Bacteriophage T7 infection of *Escherichia coli*

Diarrheal diseases, largely due to foodborne illness or contaminated drinking water, are one of the top ten causes of death worldwide, amounting to 1.4 million lives lost in 2016.[10] Outbreaks of *Escherichia coli* in food or water sources are frequently to blame for these diarrheal diseases. *E. coli* as a species encompasses an immense range of bacterial strains, many of which are harmless to humans and are used extensively in recombinant DNA research applications. This Gram-negative, facultative anaerobe occupies the intestines and feces of warm-blooded animals, and comprises 90% of the gut microbiota of humans.[11] Testing for coliform bacteria in public drinking water and food supplies is vital to reducing diarrheal diseases in a population and quickly identifying sources of contamination. Traditional methods of *E. coli* detection using microbial cultures can take days, slowing possible response time to outbreaks. DNA fingerprinting techniques can also be used to identify virulent strains, and the polymerase chain reaction (PCR) technique has helped to drastically speed up the process and provided lower detection limits, but cannot distinguish between live and dead bacteria. New methods using mass spectrometry based proteomics to analyze *E. coli* cultures over the course of bacteriophage infection are rapid, sensitive, and specific to live cultures.[12]

Bacteriophages are a class of viruses that infect bacteria and commandeer bacterial resources to replicate the phage prior to bacterial cell lysis and phage propagation. Phages were discovered nearly a century ago and have been used as an effective antibiotic treatment, particularly in former Soviet

satellite nations that lacked access to Western antibiotic pharmaceuticals. In combination with newer biological understanding of genetic engineering, bacteriophages are making a resurgence in medicinal interest with the rise of antibiotic resistant infections.[13] In addition to possible medicinal uses, bacteriophages can be effective in screening potentially contaminated food, drug, or water sources for specific bacterial cultures as phages will selectively infect host strains. Bacteriophage-based detection of bacteria is specific to the species of interest, and due to the rapid proliferation of phages, able to produce desirable signal to noise ratios in analytical detection. Of particular biological interest is enterobacteria phage T7, a heavily researched phage that infects most strains of *E. coli* and has a short lytic life cycle that leads to rapid proliferation of the phage and decline of the host. Each lytic cycle releases about one hundred new phages from the killed host cell, leading to exponential growth of the phage in a very short time, typically 25 to 30 minutes at human physiological temperature.[12] T7 phage is of particular usefulness due to its short lytic cycle, ability to survive in a variety of laboratory conditions, and its ability to infect a range of *E. coli* strains, including commonly used research strains. Additionally, T7 phage has a fully sequenced genome[14] that translates to 57 protein products, enabling the use of bioinformatic searching software in proteomics studies.

The genome of T7 phage has been fully mapped and the protein products thereof have been divided into three separate classes according to the order in which they are expressed during the lytic cycle. Class I proteins are essential for establishing favorable conditions for phage propagation and are expressed early

in the infection cycle, followed by Class II proteins and Class III proteins which are used predominately for DNA replication and packaging new phages, respectively.[15] Protein composition of the phage-host sample therefore, will change significantly across the period of infection. Identification of key proteins from each class could point to the state of phage replication and provide a window into the host-phage interaction.

While the number of phage proteins is dwarfed by the number of proteins expressed by the host, mass spectrometry-based proteomics is a sensitive technique that can zero in on the proteins of interest, without needing to segregate phage versus host proteins. *Escherichia coli* strain B/BL21(DE3), which is a common, nonpathogenic laboratory strain that was used in this study, has a fully sequenced genome that translates to 4,156 possible proteins,[16] which is more than 70 times the possible proteins produced by T7. Using shotgun proteomics to digest host and phage together, and identify their respective proteins in the same analysis gives a snapshot into the status of phage life cycle as well as the changes in protein expression in the host during viral infection. Exploring the protein composition of *E. coli* and T7 over time can offer insights into the progress and process of infection. As both organisms are well known and widely studied, this analysis can be used as foundational work to base the exploration of unknown phages and their effects on other bacterial systems.

Chapter 3: Experimental Procedure

3.1 In-solution Digestion of E. coli Cultures

Escherichia coli (strain B/BL21-DE3) samples were cultured by the Williamson Lab, Department of Biology, College of William and Mary, and infected with T7 phage before sampling at 0, 5, 20, 35, and 50 mins after infection. Each cell culture sample was suspended in tryptic soy broth and frozen at -80°C to arrest infection and culture growth. Later replicates of *E. coli* cultures were sampled at 0, 15, 30, 45, and 60 minutes by the same procedure. To begin mass spectrometry analysis of these samples, cell cultures were thawed and vortexed to suspend the cells prior to sampling. Sample preparation followed the provided instructions for the Pierce™ Mass Spec Sample Prep Kit for Cultured Cells (Thermo Scientific). Aliquots (1mL) of cell culture media were added to 1.5mL Eppendorf centrifuge tubes, centrifuged at low speed to avoid premature cell lysis, and the supernatant removed and discarded until the cell pellet is of sufficient size for analysis, roughly 20μL in volume. Generally, 6 to 8 milliliters of cell culture media were required to produce a satisfactory protein concentration for the sample preparation protocol and mass spectrometry analysis. The cell pellets were rinsed using pH 3.8 phosphate buffered saline (PBS), prepared by the Williamson Lab. Subsequent lysis, reduction, alkylation, and in-solution tryptic digestion of the cell pellets were performed in accordance with the Pierce™ sample prep kit[17] with reagents provided in the kit.

As prescribed by the Pierce™ mass spectrometry sample prep kit instructions, the protein concentration of the cell lysates was determined using a

bicinchoninic acid (BCA) colorimetric assay prior to the reduction and alkylation procedure. Reagents and bovine serum albumin (BSA) standards were obtained from Thermo Scientific, as part of the Pierce BCA Protein Assay Kit, and assay was completed following the provided instructions.[18] Absorbance measurements taken at 562nm using a Synergy™ HTX Multi-Mode Microplate Reader, per the provided microplate procedure for the BCA Assay Kit, were used to determine the protein concentration of the cell lysates. Most cell lysate samples had low protein concentrations, generally around 500µg/mL or roughly half of the recommended concentration for use of the Pierce Mass Spec Sample Kit. For these instances, volumes of reagents were adjusted proportional to the sample concentration. Vortexing the cell culture media prior to sampling was found to greatly increase the concentration of protein in the cell lysates, and this step was added to the sample preparation protocol. On average, protein concentrations of cell lysates that were vortexed prior to sampling increased nearly ninefold over previous samples.

After determination of the cell lysate protein concentration via the BCA assay, 100µg of cell lysate protein was transferred to a new Eppendorf tube in preparation for reduction, alkylation, and acetone precipitate to isolate protein from the remaining cellular debris in the samples. As mentioned above, cell lysate protein concentrations were often low, and 100µg of protein was not always available for transfer and further processing. In these instances, volumes of the following reagents were adjusted proportional to the actual amount of cell lysate protein present to maintain the same protein to reagent ratios as

prescribed in the Pierce Mass Spec Sample Prep Kit instructions. Lysate samples were incubated with freshly prepared 500mM dithiothreitol (DTT, No-Weigh™ tube, Thermo Scientific) at 50°C for 45 minutes, then cooled to room temperature before incubating at room temperature with freshly prepared 500mM iodoacetamide (IAA, Single-Use tube, Thermo Scientific while protected from light. Pre-chilled (-20°C) acetone was then added to quench the reaction and precipitate protein. The precipitated, dried protein pellet was then re-suspended in Digestion Buffer (provided in Pierce Mass Spec Sample Prep Kit, Thermo Scientific) to prepare for enzymatic digestion by Lys-C and trypsin. Samples were incubated with Lys-C at an enzyme to substrate ratio of 1:100 for 2 hours at 37°C before incubating with trypsin at a 1:50 enzyme to substrate ratio overnight at 37°C. After overnight digestion, samples were frozen at -80°C to stop enzymatic digestion.

Following tryptic digestion of the *E. coli* cell culture lysates, samples were dried in a speed vac to remove the digestion buffer as prescribed in the Pierce™ Mass Spec Sample Prep Kit, then re-suspended in sample buffer for a clean-up step with Pierce® C18 Spin Columns (Thermo Scientific). Due to the complex nature of the cell culture samples and the reagents, buffers, and detergents necessary to enzymatically digest the protein samples, spin columns were used to isolate the proteolytic peptides from other entities that could suppress signal in the mass spectrometer. Peptides were bound to, washed on, and eluted from the C18 resin per the manufacturer's guidance[19] before drying via speed vac. Final proteolytic peptide samples were re-suspended in Solvent A (98% deionized

water, 2% acetonitrile with 0.2% formic acid) for mass spectrometry analysis and stored at -20°C until ready for analysis.

3.2 Gel Electrophoresis

To reduce complexity of *E. coli* cell culture digests and increase sensitivity of the analysis, additional separation of protein samples was implemented via sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). The peptide mixture reaching the mass spectrometer was not sufficiently separating after liquid chromatography, resulting in fewer proteins being identified in cell culture samples than expected and with lower than desired confidence. Implementing an additional separation step should lower complexity of the *E. coli* lysate samples allowing for identification of a wider range of proteins, at the cost of significantly increasing the volume of samples and therefore sample preparation time. To prepare *E. coli* cell culture samples for SDS-PAGE separation, the procedure for the Pierce™ Mass Spec Sample Prep Kit for Cultured Cells[17] was followed as written, until the completion of the BCA protein concentration assay. After determination of the protein concentration, sufficient sample volumes were transferred into new tubes to provide 100µg of protein in 20µL deionized water. For samples with lower concentrations that required more than 20µL of sample solution to provide 100µg of protein, the sample solution was dried by speed vac then re-suspended in deionized water. For samples with higher concentrations that yielded volumes less than 20µL sample, sufficient deionized water was added to bring the sample volume up to 20µL. Protein samples for SDS-PAGE were prepared per the usage guidance for the 2X

Laemmli Sample Buffer (Bio-Rad) with the addition of 25µL 2X Laemmli Sample Buffer and 5µL 500mM dithiothreitol (DTT) prepared with No-Weigh™ DTT (Thermo Scientific, from Pierce kit) for a final volume of 50µL. Samples were heated per sample buffer guidance at 70°C for 10 minutes to denature the proteins prior to gel electrophoresis.

Precast polyacrylamide gels were purchased from Bio-Rad (Mini-Protean TGX Precast Gels, 10%, 10 well, 30µL wells) and used per package guidance with a Mini-Protean II gel electrophoresis tank. Running buffer was prepared fresh prior to run using 10X Tris/Glycine/SDS buffer (Bio-Rad) by diluting the buffer concentrate to ten times volume with deionized water. The precast gels were removed from their packaging and prepared as indicated in the instruction manual.[20] Prepared protein samples and a purchased protein standard (Precision Plus Protein Standard, Unstained from Bio-Rad) were loaded into the gel wells while submerging in running buffer. Voltage was then applied in constant voltage mode, first at 90V until the dye front compressed into a thin line, then at 120V for the remainder of the separation. Power was shut off when the dye front reached a black line on the precast gel cassette near the base of the gel, roughly an hour after initial application of voltage. The gel was carefully removed from the precast gel cassette per manufacturer's instructions and submerged in sufficient Coomassie Brilliant Blue R-250 Staining Solution (as purchased from Bio-Rad) to cover the gel. The gel remained in the staining solution overnight while shaking to visualize protein bands. Destaining solution was prepared with 600mL deionized water, 300mL methanol, and 100mL glacial acetic acid. The staining

solution was carefully poured off the gel, and destaining solution was added to cover the gel. The gel was moved to a shaker and allowed to shake for an hour before the destaining solution was removed and fresh destaining solution was added. This was repeated as necessary until the background of the gel was nearly clear and protein bands were clearly demarcated.

3.3 In-gel Digestion of E. coli Protein Samples

Following gel electrophoresis separation of proteins from the *E. coli* lysate samples, in-gel tryptic digestion of proteins was performed prior to extraction and mass spectrometry analysis. Each lane of the polyacrylamide gel corresponds to a particular *E. coli* sample, which was then subdivided into ten individual samples, numbered sequentially down the gel lane, with sample 1 beginning just below the well and sample 10 ending at the dye front or base of the gel. In-gel digestion was performed per the Arizona Proteomics Consortium Protocol for tryptic digestion of protein in gel bands[21] for all steps, except for peptide extraction which was carried out in accordance with Basic Protocol 1 from Gundry et al.[22] While both protocols are based on the same foundational publication[23] and have nearly identical procedures, the extraction procedures differ most significantly in the acid used to protonate the proteolytic peptides. The Arizona Proteomics Consortium Protocol chooses to extract peptides from the polyacrylamide gel using trifluoroacetic acid (TFA), which has been shown to suppress signal in electrospray ionization mass spectrometry[24] while the Gundry protocol suggests extraction with 5% formic acid and 100% acetonitrile for a more mass spectrometry-friendly sample preparation. Due to availability of lab

equipment, sonication was used in place of shaking for the extraction steps. Ice was added to the sonication bath as necessary to prevent unintended degradation of peptides or polyacrylamide gels as sonication increased the bath temperature.

Briefly, the in-gel digestion began by dividing each gel lane, corresponding to one *E. coli* cell culture lysate sample, into 10 sections, each of which were chopped into roughly $1mm^3$ pieces and loaded into an Eppendorf tube. The gel bands were then washed with solutions of acetonitrile (EMD Millipore) and 100mM ammonium bicarbonate (Fisher) to remove the Coomassie blue stain before treating the gel bands with dithiothreitol (DTT) and iodoacetamide (IAA) to sever and prevent reformation of disulfide bridges. DTT and IAA solutions were prepared using No-Weigh™ tubes, obtained from Thermo Scientific as part of the Pierce Mass Spec Sample Prep Kit. A solution of mass spectrometry-grade trypsin (Thermo Scientific) in 50mM ammonium bicarbonate was then added and the samples were incubated overnight at 37°C. Cleaved peptides were then extracted per the Gundry et al. protocol[22] with 5% formic acid and acetonitrile. Peptides samples were then dried via speed vac and stored dry at room temperature until ready for analysis.

Because of the potential for a wide range of peptide concentrations following gel electrophoresis and sample fractionation, the concentration of each completed sample was determined prior to LC/MS analysis. Working in batches, samples were re-suspended in 100µL Solvent A, and then analyzed for peptide concentration using a Nanodrop™ 2000 microvolume UV-Vis spectrophotometer

at a wavelength of 280nm. A blank measurement was taken using deionized water prior to sample analysis. Samples were loaded onto the Nanodrop stage by 2μL aliquots and the absorbance was measured. Per guidance from the Arizona Proteomics Consortium, ideal absorbance of peptide samples was assumed to be 1.0. In preparation for LC/MS analysis, injection volumes for each sample were calculated by dividing 1.0 by the measured absorbance. For example, a sample with observed absorbance of 0.2 would require an injection volume of 5μL when analyzed by LC/MS.

3.4 High Performance Liquid Chromatography and Tandem Mass Spectrometry

Proteolytic peptides samples, whether prepared via the in-solution or in-gel digestion, were analyzed by LC/MS-MS using an Eksigent NanoLC-2D and Finnigan LTQ ion trap mass spectrometer. Separation of peptides was achieved via reverse-phase liquid chromatography using a capillary column packed with Zorbax Eclipse XDB-C18 (Agilent, 5μm particle size) and an integrated, laser-pulled nanospray emitter. Peptides were first bound to a trap column (C18 PepMap^{TM} 100, Thermo Scientific, 5μm particle size) after sample injection at a flow rate of 1μL/min and washed for 5 minutes to remove remaining buffers, detergents, and other cellular debris that survived the protein isolation and digestion process. Flow was then reversed and peptides were eluted from the trap column by applying a mobile phase gradient at 350nL/min, moving from more polar to less polar solvents. Solvent A (98:2 water:acetonitrile) and Solvent B (98:2 acetonitrile:water) both contained formic acid (0.2% v/v) to protonate the peptides in solution in preparation for mass spectrometry analysis. The mobile

phase gradient ran from 5% to 95% Solvent B followed by a wash period at 95% Solvent A to prepare for the next injection. The gradient is provided in detail in Table 1, below.

Time (minutes)	% Solvent A	% Solvent B
0	95	5
5	95	5
35	65	35
40	55	45
41	5	95
46	5	95
47	95	5
85	95	5

Table 1: Mobile phase gradient for 85 minute run.

Peptides eluting from the column were detected by mass spectrometry in a data-dependent manner. Nanospray ionization was used as the ionization source for its high sensitivity and ability to be used in-line with HPLC assays. Data-dependent scans were taken throughout the 85 minute run, as the five highest intensity peaks in MS^1 were identified as precursor ions then sequentially isolated and fragmented. Precursor and product ion spectra were recorded to be later analyzed with SEQUEST, a data analytics software program used to identify proteins from tandem mass spectrometry.

3.5 Capillary Column and Nanospray Emitter Preparation

Nanospray emitters and capillary HPLC columns are integral to this mass spectrometry based proteomics experiment, but both items have limited lifespans, are prone to failure, and can carry large price tags. For this reason, a source of new, reliable, and cost-effective nanospray emitters was investigated. Commercially produced fused silica-based nanospray emitters were purchased from New Objective (uncoated, 360μm OD x 75μm ID, 15μm tip diameter) and

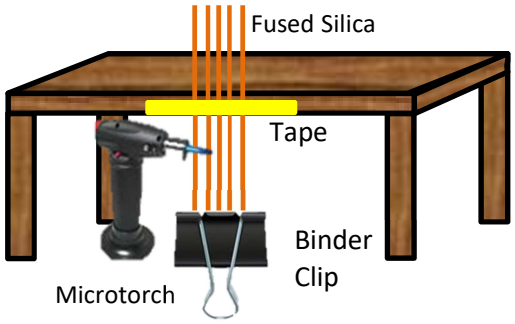

Figure 3: Binder clip method for pulled nanospray emitters

used successfully, but proved unsustainably expensive for repeated use. Nanospray emitters were pulled in-lab using 360µm OD x 250µm ID fused silica tubing per guidance from the University of Washington Proteomics Resource[25] using a microtorch and large binder clip as seen in Figure 3. Lengths of fused silica capillaries were cut and secured to a lab bench with tape before carefully burning off the coating near the middle of the capillaries. Capillaries were gently wiped with methanol to remove the charred coating and a large binder clip was clipped to the capillaries. The coating-free portion of the capillaries was then heated with the torch until melting began and the weight of the binder clip stretched the capillaries to a breaking point. Pulled tips were then cooled and observed under a microscope before carefully trimming the pulled end to the desired tip diameter and length. The 250µm inner diameter fused silica was extremely fragile after removal of the protective coating, particularly after pulling. Future studies will be conducted with fused silica tubing with a much smaller inner diameter and therefore thicker walls that are less likely to break after stretching. While usable nanospray emitters were produced in this manner, the inherent variations in this manual pulling technique led to unstable spraying that were not long-lived enough for use in proteomics studies. In future studies, pulled

nanospray emitters will be produced using a commercial laser puller specially designed for fused silica tubing. The P-2000 Laser Micropipette Puller (Sutter Instrument Company) can consistently and accurately pull fused silica into nanospray emitters that can be packed as capillary columns.

Capillary columns were packed using guidance from the University of Washington Proteomics Resource[25] with Zorbax Eclipse XDB-C18 (5µm particle size, obtained in bulk from Agilent). Lengths of 360µm x 75µm fused silica capillaries were cut with a rotary capillary cutter to roughly 25cm before preparing potassium silicate (KASIL) polymer frits based on the method described by Meiring et al.[26] Briefly, 50µL formamide was added to 200µL potassium silicate (29.1%, PQ Corporation) in an Eppendorf tube which was quickly vortexed to mix then centrifuged at 10,000 rpm for 2 minutes. Cut capillaries were then dipped in the KASIL solution for about 5 seconds to draw the solution in by capillary action. Excess KASIL was wiped from the outside of the capillaries before curing in a 90°C oven overnight. Cured frits were observed under a microscope before trimming the completed frits down to roughly 2mm to create a satisfactory amount of backpressure when applied to the HPLC after packing.

Fritted capillaries were then packed using a pressure cell as seen in Figure 4. A few micrograms of packing material were suspended in an appropriate solvent (80% acetonitrile) by vortexing then the lid was cut off the Eppendorf tube to fit in the pressure cell. The Eppendorf was then lowered into a small brass plug that was machined to hold the tube securely upright during the packing procedure. The lid of the pressure cell was tightly screwed on before

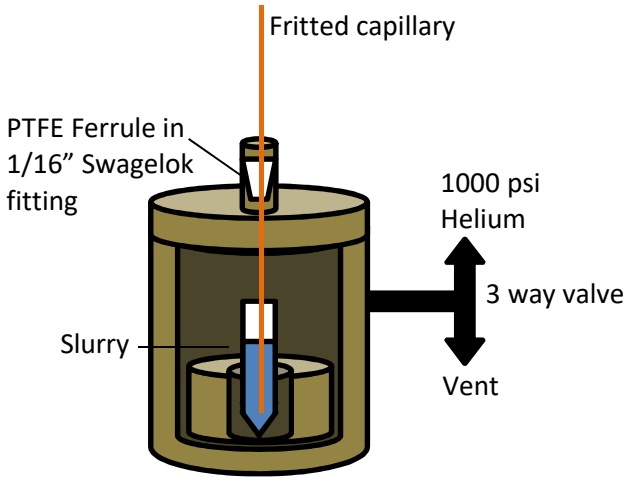

Figure 4: Pressure cell for packing capillary columns.

capillary into the PTFE (polytetrafluoroethylene) ferrule and positioning the open end of the capillary just above the bottom of the slurry tube before tightening the Swagelok fitting to finger-tightness. Helium was introduced into the cell via the three-way valve and the cell was pressurized to roughly 1000 psi. Pressure was kept on the system as long as drips formed at the fritted end of the capillary and packing material was visibly aggregating in the column. When no movement of material was observed, the system was depressurized, the slurry was re-suspended by vortexing, and the process was repeated. This continued until the column was of a suitable length (6-8cm). Once the column packing reached the desired length, the column was allowed to equilibrate under pressure by replacing the slurry with solvent and pressurizing the cell. The columns were then trimmed to eliminate excess dead volume.

Chapter 4: Methods Development

4.1 Early Work and Electrospray Ionization Studies

Because of the complex nature of cell culture digests, the heart of any shotgun proteomics study is effective separation technique. To this end, much effort was made to continually improve separation and therefore sensitivity in peptide detection throughout this work. Initial work was completed with a Shimadzu Prominence UFLC XR HPLC system with a 1:1000 flow splitter to reduce the flow rate to enable the use of nanospray ionization. When using the Shimadzu HPLC, a commercially packed column, EASY-Column^{TM} (C18-A2, 10cm, 75µm ID, 3µm particle size, Thermo Scientific) was used in conjunction with commercially prepared fused silica nanospray emitters (PicoTips, New Objective, uncoated, 360µm OD x 75µm ID, 15µm tip diameter), as opposed to the column with integrated emitter as described in Section 3.4. While use of the standard flow HPLC with a flow splitter created occasionally inconsistent spraying, the EASY-Column^{TM} provided adequate separation of proteolytic peptides and provided proof of concept results for *E. coli* cell culture samples. A 35 minute post-infection sample run on 23 May 2017, using the Shimadzu HPLC and Finnigan LTQ ion trap mass spectrometer, identified 3594 *E. coli* proteins and 43 T7 phage proteins, or 86% and 75% of possible proteins produced by the respective organism's genome. While many proteins were identified, the vast majority of the identified peptides were assigned sequences with low confidence by SEQUEST. Only 46 peptides identified to high confidence out of 15,629 total peptides assigned, or 0.3%. Therefore, improvements to system sensitivity were

sought out to increase the confidence of peptide and therefore protein identification. Unfortunately, the EASY-Column™ failed shortly after this run, and due to a lack of available capillary columns as well as continuing reliability issues with the nanospray ionization source, the flow splitter was removed and an electrospray ionization source was installed.

To accommodate the higher flow rate, the switching valve on the mass spectrometer was bypassed to avoid the precolumn, and the analytical column was changed to an ACE Excel 3 SuperC18 column (30mm x 2.1mm ID, 3µm particle size). The flow rate was set to 0.4 mL/minute and various mobile phase gradients were tested, ranging from 60 minutes to 120 minutes. To assess the LC/MS system before analyzing *E. coli* samples, a standard solution of MS Qual/Quant QC Mix (Sigma Aldrich) was prepared as recommended by the manufacturer in 20% acetonitrile with 0.1% formic acid. This pre-digested mix contained six proteins of variable abundance to assess the sensitivity and accuracy of experimental setups used for proteomics studies. Using a 60 minute gradient method as an initial test of the ESI source setup on 25 July 2017, all six proteins were identified, with a range of coverages from 46% to 74% as seen in Table 2. For each protein, a numerical score, percentage of sequence coverage,

Protein Description	Score	Coverage (%)	# Unique Peptides	# Peptides	# PSMs
Carbonic anhydrase 1	13.21	50.77	7	21	77
Carbonic anhydrase 2	8.04	69.50	8	26	116
NAD(P)H dehydrogenase [quinone] 1	0.00	58.24	1	35	119
C-reactive protein	0.00	45.67	2	17	66
Peptidyl-prolyl cis-trans isomerase A	0.00	73.51	0	35	158
Catalase	0.00	67.30	0	63	214

Table 2: Proteins identified in MS Qual/Quant QC Mix on 07/25/17, using ESI and 60 minute gradient

number of high confidence (denoted as unique by SEQUEST) peptides, total number of peptides identified per protein and the number of peptide spectrum matches (PSMs) is given. Proteins are listed in Table 2 in order of most abundant to least abundant, with the first pair of proteins being five times as abundant as the next pair, and the second pair of proteins being five times more abundant than the third. With the range of protein concentrations in the prepared mix, the most abundant proteins, the carbonic anhydrases 1 and 2, are expected to be the highest scored proteins, and this pattern of decreasing protein scores and peptide confidence as the concentrations of standard protein decreased was observed for all runs of the MS Qual/Quant QC Mix. These protein score values were calculated in a proprietary manner by SEQUEST, but derived from the cross correlation (XCorr) values and number of spectral matches for the peptides that make up a given protein and were therefore indicators of how confidently the identity of the protein was assigned.

Solvent gradient composition, injection volume, and length of run were varied to improve separation and identification of peptides in the predigested standard. The subsequent analysis of the MS Qual/Quant QC Mix on 1 September 2017 showed marked improvement in protein coverages and scores,

Protein Description	Score	Coverage (%)	# Unique Peptides	# Peptides	# PSMs
Carbonic anhydrase 1	16.41	84.23	8	33	176
Carbonic anhydrase 2	12.56	88.03	5	42	206
C-reactive protein	1.61	45.67	2	20	129
NAD(P)H dehydrogenase [quinone] 1	0.00	78.39	0	49	258
Peptidyl-prolyl cis-trans isomerase A	0.00	100.00	2	44	309
Catalase	0.00	91.44	1	95	468

Table 3: Proteins identified in MS Qual/Quant QC Mix on 09/01/17, using ESI and 120 minute gradient.

summarized in Table 3. While not scored, both proteins of lowest abundance were identified with coverage of over 90% of the protein sequence, indicating an improvement in sensitivity of the system. Coverage of all but one protein, C-reactive protein, increased appreciably, and protein scores increased for three of the higher abundance proteins. With these promising results in hand, an *E. coli* cell culture digest was prepared per the in-solution digestion procedure, followed by the C18 spin column clean-up procedure to further isolate peptides and wash away potential sources of ion suppression leftover from the digestion process. This digest was analyzed with the aforementioned ESI setup with a 120 minute gradient and 10 µL sample injection volume on 29 November 2017. From this 0 minute post-infection sample, 3119 total proteins were identified, of which 38 were T7 phage produced proteins. Of the 3119 proteins identified, only 111 had a nonzero score as assigned by SEQUEST, and even the highest scored protein had a score of only 28.13. As seen in the total ion chromatogram in Figure 5, the

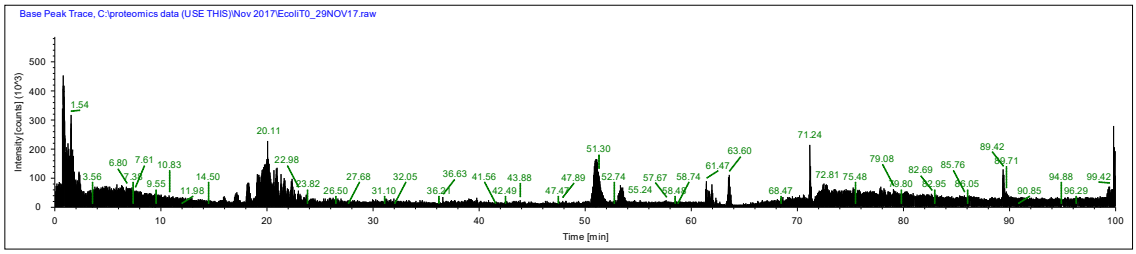

Figure 5: Total ion count over time for *E. coli* digest, 0 minutes post-infection, analyzed 11/29/17 using ESI and 120 minute gradient.

total ion count was low throughout the run, indicating poor detection of peptides eluting from the column. While the number of proteins identified and peptides identified with high confidence were roughly equivalent to the *E. coli* sample run on 23 May 2017 via nanospray ionization, the highest protein score for the

sample run with electrospray ionization was about 100 times lower than via nanospray ionization, 28.13 versus 2157.29, respectively. Additionally, neither analysis had a satisfactory percentage of confidently assigned peptides and therefore scored proteins. Both analyses were able to qualitatively identify a large number of proteins from both T7 phage and its host, but it appeared the electrospray ionization and direct injection method lacked the sensitivity to confidently identify the complex cell culture digests.

4.2 Nanospray Ionization Studies

With the acquisition of a new nanoflow HPLC, the ionization source on the LTQ was switched back to nanospray ionization. Lab-pulled nanospray emitters, produced via the binder clip method described in Section 3.5, were tested with direct injection from a syringe pump and stable spraying was observed. While this was promising, these emitters typically only sprayed consistently for roughly 15 minutes before clogging or otherwise failing. For this reason, commercial fused-silica nanospray emitters were utilized following a lab-packed capillary column containing roughly 12 centimeters of Pronto-SIL-120-5-C18AQ (Bischoff Chromatography). The precolumn was replaced and the LC/MS method was updated to include a five minute binding and wash period before reversing the mobile phase flow via the switching valve on the mass spectrometer and eluting peptides from the trap column onto the analytical column. For initial testing of the new HPLC setup, the MS Qual/Quant QC Mix was analyzed for an apples-to-apples comparison to the previous experimental setup with ESI. Using a 60 minute gradient at a flow rate of 250 nL/minute, all six proteins were identified

with coverages comparable to the 1 September 2017 analysis, as seen in Table 4. While the coverage percentages were comparable, none of the proteins were

Protein Description	Score	Coverage (%)	# Unique Peptides	# Peptides	# PSMs
Carbonic anhydrase 1	0.00	84.23	1	34	205
Carbonic anhydrase 2	0.00	97.30	1	46	278
NAD(P)H dehydrogenase [quinone] 1	0.00	78.02	0	51	346
C-reactive protein	0.00	42.79	0	19	115
Peptidyl-prolyl cis-trans isomerase A	0.00	100.00	0	34	290
Catalase	0.00	92.02	1	101	583

Table 4: Proteins identified in MS Qual/Quant QC Mix on 03/26/18, using NSI and 60 minute gradient.

scored, and only 3 out of 285 peptides were identified with high confidence. Nanospray ionization typically has better ionization efficiency and therefore sensitivity compared to electrospray ionization, but efficient separation is crucial to proteomics studies. While this system was clearly ionizing and identifying peptides at the same rate as the 1 September 2017 assay, the separation was highly suspect, as evidenced in the total ion count chromatogram provided in Figure 6. The slight increase in peptide spectral matches from the ESI to NSI

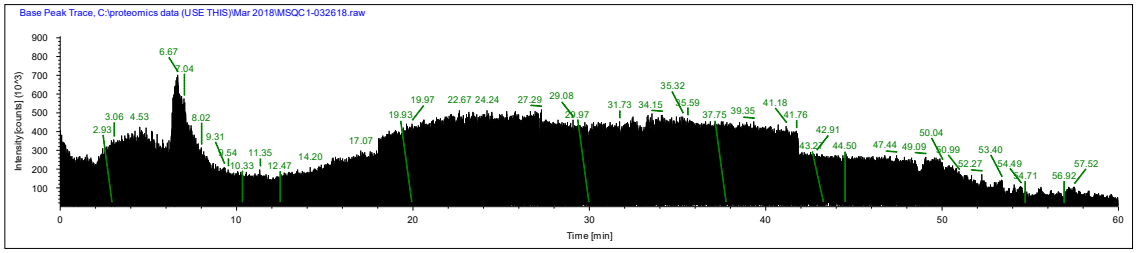

Figure 6: Total ion count over time of MS Qual/Quant QC Mix analysis on 03/26/18 using 60 minute gradient and NSI.

assays despite the clear lack of separation could likely be attributed to the difference in ionization source.

As was done with the ESI setup, length and composition of the mobile phase gradient was varied to improve separation of peptides in the QC mix as

well as *E. coli* lysate digests. Analysis time was increased to between 120 and 180 minutes and flow rate was decreased to allow for better separation. These changes improved the look of the total ion count chromatograms, but proved ineffective in meaningfully improving peptide separation. A step-wise 180 minute gradient run at a 250 nL/minute flow rate produced such a chromatogram, seen in Figure 7, for an *E.coli* sample taken 0 minutes post-infection with T7 phage.

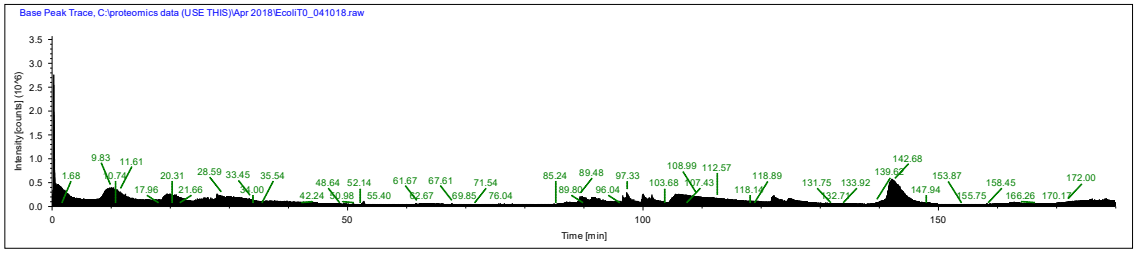

Figure 7: Total ion count over time of *E. coli* digest, 0 minutes post-infection, analyzed 04/10/18.

Much of the chromatogram was unremarkable with only poorly defined peaks. The small peaks seen, while not well resolved, indicated some level of separation that was likely a result of peptides sequentially eluting off of the trap column as the mobile phase composition changed. Despite the unremarkable chromatogram, 3150 proteins were identified of which 651 had a non-zero score assigned by SEQUEST. The highest scoring protein, a histone family DNA-binding protein encoded by *E. coli*, had a score of 1551.98 and 71% sequence coverage. Out of 8604 peptides identified, 61 were assigned sequences with high confidence. By these metrics, this analysis was as good as or better than previous analyses at identifying proteins in cell culture digests. However, as the total ion count included any noise detected by the mass spectrometer as well as peptides, this chromatogram presented a rosier picture of the analysis than when

looking at the elution profiles of individual peptides.

For example, the peptide ALEEAGAEVEVK was identified with high confidence, with an Xcorr value of 4.25, in the *E. coli* digest sample described above, and had 244 spectral matches throughout the length of the analysis. As seen in Figure 8 where each spectral match is indicated by a red line, this

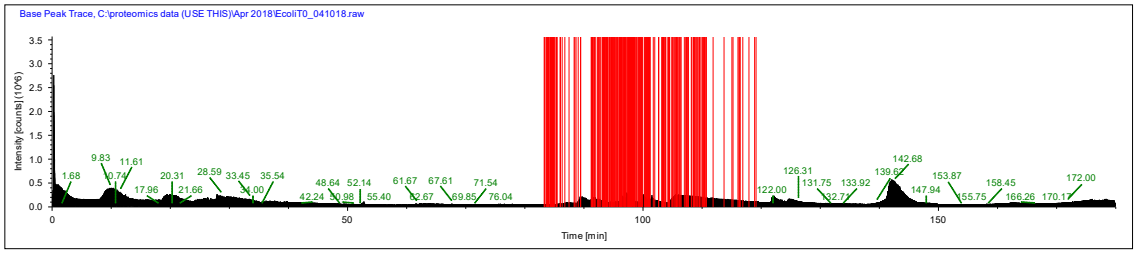

Figure 8: Spectral matches for peptide ALEEAGAEVEVK from *E. coli* digest, 0 minutes post-infection, analyzed 04/10/18

peptide was detected across a wide range of retention times. ALEEAGAEVEVK was first detected at around 83 minutes into the 180 minute analysis and was last detected at roughly 119 minutes for a peak width of 36 minutes or 20% of the entire analysis. During these 36 minutes, the percentage of Solvent B varied little as this peptide was first detected at 80% Solvent B and last detected at 90% Solvent B. As seen in the extracted ion chromatogram in Figure 9, the precursor mass for ALEEAGAEVEVK as indicated in SEQUEST, m/z = 623, was detected throughout the chromatographic run. This peptide was identified starting at 83 minutes, indicating the early signal in Figure 9 was likely due to another ion of similar m/z or the signal upon fragmentation was insufficient to make an assignment to even low confidence. SEQUEST logged spectral matches for ALEEAGAEVEVK starting at the left shoulder of the large peak and nearly continuously until around 119 minutes, when the precursor ion peak has tailed to

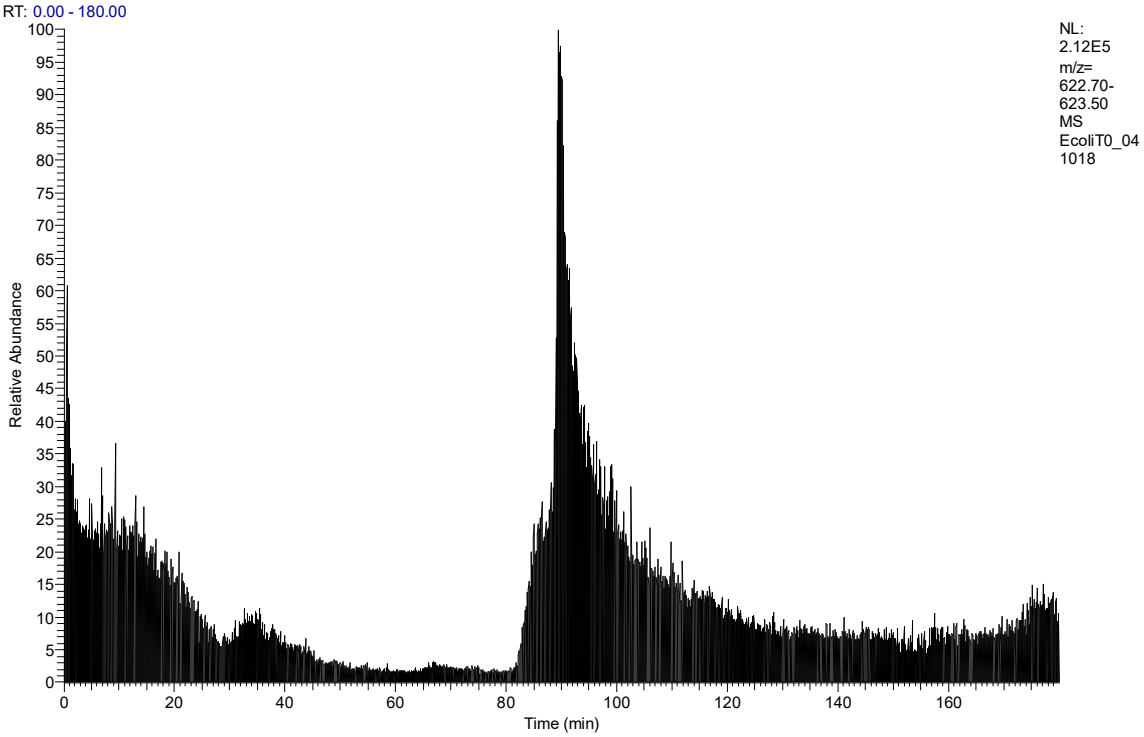

Figure 9: Extracted ion chromatogram for ALEEAGAEVEVK from *E. coli* digest, 0 minutes post-infection, analyzed 04/10/18

less than 15% of the maximum signal intensity (at 89 minutes).While a defined peak was observed for this peptide, the considerable tailing, presence of a significant shoulder, and failure to reestablish a steady baseline were evidence of poor peptide separation by the chromatographic system.

While the analysis was able to identify a comparable number of peptides to the previous NSI and ESI assays, there were clear separation issues that were hampering the detection of peptides. As the data dependent acquisition was only isolating and fragmenting the highest intensity peaks in a given scan, without proper separation of peptides, the lower abundance peptides would not be observed. Dynamic exclusion settings in the data-dependent acquisition would have allowed for other peptides to be isolated and fragmented if the same precursor mass was detected too frequently within a given time period, but this

cannot compensate entirely for poor chromatography. For the 36 minutes that ALEEAGAEVEVK was detected, this peptide would have produced one of the five most intense peaks in the full MS^1 scan leading to its isolation and fragmentation with each isolation event resulting in a spectral match. While the many spectral matches observed undoubtedly contributed to the high cross-correlation value calculated by SEQUEST for this peptide, the poor resolution in the extracted ion chromatogram caused lower abundance peptides to not be isolated and therefore detected. Had this peptide eluted over a shorter period of time and therefore produced a resolved peak in the chromatogram, other peptides likely could have been detected.

Despite the poor separation of peptides, this analysis proved to be the best case scenario for *E. coli* digest samples tested using the 180 minute method and the lab-packed Pronto-SIL column, as subsequent *E. coli* digest samples varied widely in the number of peptides and proteins identified despite various optimization attempts. To decrease complexity of samples separated by the questionable lab-packed column, the MS Qual/Quant QC Mix was tested again after maintenance of the Eksigent nanoLC. This analysis, on 13 June 2018, could not identify any of the six standard proteins with a non-zero score, and the highest sequence coverage was only 31%. As the system failed this test, an injection of bradykinin (Sigma-Aldrich, acetate salt in 100% acetonitrile with 1% formic acid) was made to test the resolving power of the column. In theory, a peptide alone in solution should have created a single peak in the chromatogram, but while bradykinin was correctly identified and scored highly by SEQUEST, it

was ubiquitous across the entire length of the run. As the column failed this easy test, a new column was packed as described in Section 3.5 with Pronto-SIL-120-5-C18AQ. A newly packed column of the same material failed again to provide sufficient separation of peptides and while separation was lacking in the nanospray ionization experimental setup, the increased ionization efficiency contributed to the identification of comparable number of proteins as the ESI assay. Specifically, the NSI analysis of *E. coli* on 10 April 2018 was able to identify roughly the same number of proteins while meeting or exceeding the same metrics for assignment confidence using an analytical column that failed to meaningfully separate peptides as the ESI analysis on 29 November 2017 using a compatible analytical column which indicates the importance of ionization source for mass spectrometry-based proteomics. Both analyses are summarized in Table 5 for direct comparison of results. As seen in Table 5, both ionization

		# Proteins	# Scored Proteins	Highest Score	# Peptides	# High Confidence Peptides
ESI (11/29/17)	Total	3119	111	28.13	8383	48
	T7 Phage	38	1	1.78	128	0
NSI (04/10/18)	Total	3150	651	1551.98	8604	61
	T7 Phage	36	6	26.84	124	0

Table 5: Summary of identified proteins in *E. coli* culture digests, 0 minutes post-infection.

methods enabled the identification of roughly the same absolute number of peptides and proteins, but the later analysis using NSI was able to do so to much higher confidence as evidenced by the nearly sixfold increase in scored proteins and fiftyfold increase in highest score obtained by a single protein. The increased ionization efficiency and sensitivity of nanospray ionization overcame some of the deleterious effects of poor separation. While the electrospray analysis likely

represented the best case scenario for the electrospray ionization experimental setup, the nanospray ionization setup allowed for further optimization and can be expected to improve with further modification.

4.3 Incorporation of Analytical Column with Integrated Nanospray Emitter

When this newly packed and equilibrated column still failed to meaningfully separate peptides, the experimental setup was changed to accommodate a column with an integrated nanospray emitter to use columns packed by the Arizona Proteomics Consortium, with Zorbax Eclipse XDB-C18 packing material. This column packing material provided better separation of proteolytic peptides and the integrated nanospray emitter reduced the incidence of needle clogging as the packing material acted as a filter for the emitter. With new columns more appropriate for the mobile phase gradient and protein analysis and an updated gradient as recommended by the Arizona Proteomics Consortium, *E. coli* cell cultures were lysed and digested via the in-solution enzymatic digestion procedure then analyzed utilizing nanospray ionization and consistent, if preliminary, results were obtained. Consistent identification of T7 phage proteins was possible across all time points available for infection of *E. coli* as will be discussed in depth in Chapter 5. Separation of proteolytic peptides saw marked improvement using the Zorbax Eclipse packing material compared to the Pronto-SIL packing material, as evidenced by the tight grouping of spectral matches in Figure 10 The total ion chromatogram showed a characteristic pattern of peaks common to proteomics studies. While lacking baseline resolution between peaks, peptides elute to form defined peaks throughout the

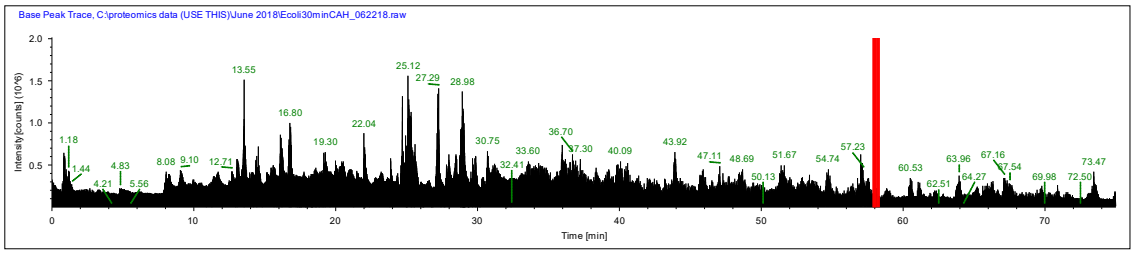

Figure 10: Total ion count over time for *E. coli* digest, 30 minutes post-infection, using NSI and 85 minute gradient. Red lines indicate spectral matches for peptide SGETEDATIADLAVGTAAGQIK.

chromatographic run. The highlighted peptide, SGETEDATIADLAVGTAAGQIK, was identified to high confidence with a XCorr value of 6.29 and 25 spectral matches. Compared to the distribution of peptide spectral matches in Figure 8 that were dispersed across 36 minutes of a 180 minute gradient, all 25 spectral matches fall within a window of roughly one minute. The extracted ion chromatogram for the selected peptide was also radically improved, with clear demarcation of precursor mass peaks. As SEQUEST identified precursor masses within a 1.5 Da window, the peptide SGETEDATIADLAVGTAAGQIK had a range of precursor masses with slightly different retention times, each accounting for a peptide spectral match. When the peptide ion chromatogram was extracted, the precursor ion with the best cross correlation value was marked with a red line, but all peaks in Figure 11 were the result of the same peptide. It is important to note that the extracted ion chromatogram does not

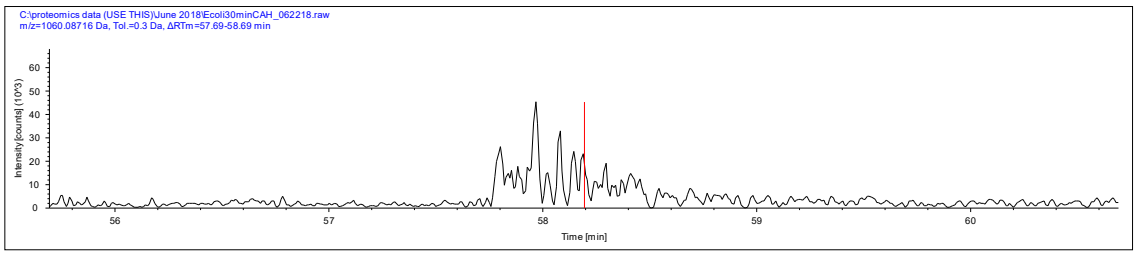

Figure 11: Extracted ion chromatogram of SGETEDATIADLAVGTAAGQIK from *E. coli* digest on 06/22/18, using NSI and 85 minute gradient.

have the same scale as the total ion chromatogram, but is zoomed in to the retention time of the peptide to show more detail. In the time before the first peaks appeared for the peptide precursor mass in Figure 11, a steady baseline was established and then reestablished shortly following the minute over which peptide spectral matches were recorded. As the Zorbax Eclipse column packing material provided superior separation over the Pronto-SIL packing material, future studies will be conducted with lab-packed Zorbax Eclipse columns with integrated nanospray emitters pulled with a laser pipette puller. While these results were promising, sensitivity could be improved by initiation of additional, off-line separation of proteins to reduce complexity of the proteolytic peptides.

4.4 Reducing Sample Complexity and Improving Sensitivity

With an appropriate gradient and appropriate analytical column, separation of peptides was improved which led to more confident assignments of the protein composition of these cell culture samples. However, utilizing an ion trap mass spectrometer instead of a high-resolution mass spectrometer limited the detection of peptides in the highly complex samples. To decrease the complexity of the *E. coli* samples and potentially identify more proteins to a higher confidence, one dimensional denaturing gel electrophoresis of *E. coli* cell culture lysates prior to enzymatic digestion was initiated. While implementing SDS-PAGE before protein digestion and fractionating cell culture lysate samples greatly increased the time of sample preparation and volume of samples, the increased sensitivity should compensate for relatively low resolution of the mass spectrometer. Even under optimized conditions, the thousands of proteins in cell

culture lysates overwhelmed the system and lower abundance proteins were overshadowed by higher abundance proteins. Decreasing the number of proteins in each sample prepared for LC/MS analysis should allow for more confident identification of proteins across the course of the T7 phage infection. Generally, analysis of simpler protein samples, either the MS Qual/Quant QC Mix of six proteins or a single purified protein, produced higher coverage of protein sequence, more high confidence peptides, and higher protein scores. Fractionating each *E. coli* cell lysate sample via SDS-PAGE into ten samples as described in Section 3.2 will radically reduce sample complexity and provide a deeper understanding of host-phage interactions during viral infection. Reducing sample complexity and therefore increasing analysis sensitivity could be potentially most helpful in monitoring phage proteins in early stages of infection, when phage protein concentrations are low. The rapid reproductive cycle of T7 phage will quickly increase these concentrations as the phage propagates, but valuable information about the initial expression of Class I proteins would be inaccessible without a highly sensitive technique.

Additionally, by separating the T7 and *E. coli* proteins by molecular weight prior to digestion and mass spectrometry analysis, future studies could reduce analysis time by identifying which gel bands carry proteins of most interest and preferentially processing only these bands. For example, the major capsid protein 10A coded for by T7 phage was detected in all time points sampled from 0 minutes to 60 minutes post-infection, as can be expected for a major structural protein that comprises the majority of the viral capsid by the arrangement of 415

copies of gp10A in an icosahedral shell.[27] As this protein will be present whenever the phage is present, regardless of infection progression, it has less value to a differential proteomics study. With a molecular weight of 36.5kDa, this protein will migrate in SDS-PAGE nearly alongside the 37kDa protein standard making identification simple. If no other proteins of interest have similar molecular weights, the band can be excluded from study. Conversely, proteins that are detected in only one time point can be selectively excised for further in-depth analysis.

In addition to reducing sample complexity by implementing preliminary separation by gel electrophoresis before LC/MS analysis, determination of the concentration of proteolytic peptides present in the sample prepared for LC/MS helped ensure the appropriate amount of analyte was deposited on the analytical column. The in-solution and in-gel digestion methods as described in Chapter 3 both required the determination of protein concentration prior to digestion and electrophoresis, respectively. However, due to inevitable sample losses during the sample preparation and clean-up procedures, this protein concentration could not be assumed to be the same as the peptide concentration following digestion. Analytical HPLC columns are designed to separate only a limited range of analyte masses dependent on column size, and overloading the column will result in poor resolution and increased analyte carryover in subsequent injections. Alternatively, injecting too little analyte will result in a poor signal to noise ratio and hamper detection of analyte. Implementation of the determination of peptide concentration following in-gel digestion and extraction sought to

ensure an appropriate amount of peptide loaded onto the analytical column. While equally concentrated protein solutions were applied to the gel for each *E. coli* sample, the manual excision of protein bands based only on stain intensity virtually guaranteed a range of peptide concentrations in the completed samples for LC/MS. For example, peptide concentrations determined using the NanoDrop 2000 as described in Section 3.3 varied widely within a single *E. coli* digest as seen in Table 6 for a 15 minute post-infection *E. coli* digest prepared by in-gel digestion. Concentration of proteolytic peptides following in-gel digestion varied

Sample ID	Peptide Concentration	Unit	A280	Injection Vol (µL)
t15 sx1	0.201	mg/ml	0.201	5
t15 sx2	0.225	mg/ml	0.225	4
t15 sx3	0.205	mg/ml	0.205	5
t15 sx4	0.229	mg/ml	0.229	4
t15 sx5	0.154	mg/ml	0.154	6
t15 sx6	0.196	mg/ml	0.196	5
t15 sx7	0.038	mg/ml	0.038	26
t15 sx7*	0.046	mg/ml	0.046	22
t15 sx8	0.076	mg/ml	0.076	13
t15 sx9	0.119	mg/ml	0.119	8
t15 sx10	0.268	mg/ml	0.268	4

Table 6: Peptide concentrations and ideal injection volumes of *E. coli* digest samples, 15 minutes post-infection, determined by NanoDrop2000 on 09/06/18. Note: Sample 7 was vortexed and absorbance measured again as first calculated concentration was an outlier.

from 0.046 mg/mL to 0.268 mg/mL, likely resulting from unequal excision of stained protein bands from the polyacrylamide gel. Injection volumes calculated in the Table 6 will be used when the LC/MS analysis of these samples occurs. The variable injection volume provides an opportunity to control the amount of peptide deposited on the analytical column without concentrating or diluting each individual sample, saving preparation time in an already lengthy sample preparation process.

Reducing sample complexity and optimizing chromatographic conditions should improve the confidence with which proteins are identified in the *E. coli* cell culture digests, enabling more reliable and in-depth insight into changes that occur throughout the course of viral infection. Continued method development and analysis of viral-host interactions could identify markers of phage infection in the host organism, which could then be used to screen bacterial cultures for signs of infection by uncharacterized phages. While identification of phage proteins via bottom-up proteomics studies require sequenced genomes to predict protein sequences, identifying key protein markers of infection in host cell digests could guide targeted investigation of such cell cultures.

Chapter 5: Preliminary Results

E. coli cultures were lysed and digested as described for the in-solution tryptic digestion in Section 3.1 from the following time points of T7 phage infection: 0, 15, 30, 45, and 60 minutes following infection. These samples were analyzed via LC/MS-MS using nanospray ionization utilizing a capillary column packed with Zorbax Eclipse XDB-C18 by the Arizona Proteomics Consortium with an integrated, pulled nanospray emitter and the 85 minute mobile phase gradient as detailed in Table 1. As the peptide concentration determination using the NanoDrop 2000 had not yet been implemented prior to analysis of these samples, 10 µL injection volumes were used for all samples. All samples were analyzed between 22 June 2018 and 28 June 2018 in this manner.

Of the 57 predicted protein products of the T7 phage genome, 49 phage-produced proteins were identified across the five *E. coli* lysate samples tested or 86% of the T7 phage proteome. The 45 minute post-infection sample contained the most phage proteins, 37, while the 15 minute post-infection sample contained the fewest at 21 phage proteins identified. On average each time point sample contained 29 different phage proteins, as each sampling point captured a different snapshot of gene expression during the course of the infection. When taken together, these snapshots form a full picture of viral protein production during host infection. A summary of the number of phage and total proteins identified in each *E. coli* cell culture lysate is provided in Table 7. Peptides in phage proteins were not assigned with high confidence and only 6 of 49 phage proteins identified across all sampling time points were assigned a non-zero

score, with all but one scored phage protein observed in later sampling points. The 0 and 15 minute post-infection samples identified the fewest phage proteins despite identifying roughly the same number of total proteins as the 30 minute sample. An increase of phage protein concentration due to proliferation of new phages released at the completion of the first lytic cycle would allow for detection

		# Proteins	# Scored Proteins	Highest Score	# Peptides	# High Confidence Peptides
0 minute	Total	1876	110	355.82	3031	44
(06/25/18)	T7 Phage	27	1	7.01	45	0
15 minute	Total	1682	39	124.77	2540	21
(06/28/15)	T7 Phage	21	0	0.00	38	0
30 minute	Total	1934	146	436.82	3115	31
(06/22/18)	T7 Phage	29	2	5.14	41	0
45 minute	Total	2850	210	289.75	6954	71
(06/27/18)	T7 Phage	37	4	3.40	101	0
60 minute	Total	3095	151	95.04	7996	11
(06/26/18)	T7 Phage	33	2	1.75	103	0

Table 7: Summary of total and phage proteins identified in *E. coli* cell culture sampled at given time points.

of more proteins in samples taken later in the infection cycle. As more *E. coli* cells had their cellular machinery taken over by T7 phage, fewer *E. coli* proteins would be produced in favor of phage proteins. Indeed, the *E. coli* culture samples taken after the initial lytic cycle of T7 phage identified more phage proteins with more non-zero scores with higher sequence coverage. As the number of scored phage proteins also increased in the later samples and scored proteins generally result from an increase in precursor ion signal intensity or number of spectral matches for their composite peptides, it can be inferred that the concentration of phage proteins relative to *E. coli* proteins had increased.

Only 15 phage proteins appeared in all time points sampled, which are listed in Table 8 with the name, class of protein, and gene number that encoded each protein. Gene numbers were assigned to the sequenced and mapped genetic code for bacteriophage T7, with integer numbers representing gene products that were essential to phage survival and non-integer numbers representing gene products that were non-essential.[14] The proteins identified in

Protein Description	0 min	15 min	30 min	45 min	60 min	Class	Gene Number
Protein 2.8	x	x	x	x	x	II	2.8
Endonuclease I	x	x	x	x	x	II	3
DNA primase/helicase	x	x	x	x	x	II	4
Inhibitor of toxin/antitoxin system	x	x	x	x	x	II	4.5
Protein 4.7	x	x	x	x	x	II	4.7
DNA-directed DNA polymerase	x	x	x	x	x	II	5
Exonuclease	x	x	x	x	x	II	6
Portal protein	x	x	x	x	x	III	8
Major capsid protein 10A	x	x	x	x	x	III	10
Minor capsid protein 10B	x	x	x	x	x	III	10
Internal virion protein gp14	x	x	x	x	x	III	14
Internal virion protein gp15	x	x	x	x	x	III	15
Peptidoglycan hydrolase gp16	x	x	x	x	x	III	16
Tail fiber protein	x	x	x	x	x	III	17
Spanin, inner membrane subunit	x	x	x	x	x	III	18.5

Table 8: Phage proteins identified in all *E. coli* digest time points.

all time points tested belonged to Class II and III, composed of gene products expressed after initial infection is established and replication of DNA and capsid proteins predominate. Class III proteins are largely structural, and can be expected to be present in a shotgun sample at any part of the lytic cycle, as phages were digested alongside host bacterial cells. Regardless if the generation of new structural proteins is occurring at a given sampling time, these proteins comprise the phage capsid and will be detected whenever the phage is present.

Because these proteins were detected across all time points of infection, a high abundance protein could be selected from this list to screen bacterial cultures for the presence of T7 phage. Major capsid protein, gene product 10A, was detected in all time points, which was expected as it was found to be the most abundant protein in T7 phage in previous work by Dunn and Studier.[14] Of the eight Class III proteins observed in Table 8, six are structural proteins while the other two proteins, peptidoglycan hydrolase and the inner membrane spanin subunit (i-spanin), participate in the injection of viral DNA into a host cell[28] and initiate host cell lysis and viral release.[29] Class II proteins, produced in the intermediary stage between early establishment of viral infection and packaging and release of progeny, predominantly function in roles of DNA reproduction. Of the seven Class II proteins detected in *E. coli* at all time points sampled, four were directly involved in DNA reproduction and editing: endonuclease I, exonuclease, DNA primase/helicase, and DNA-directed DNA polymerase. Of the remaining three proteins, two have functions that are as of yet undetermined and are thus identified only by their gene product (gp) numbers: protein 2.8 and protein 4.7. As indicated by the non-integer gene product numbers, these proteins are non-essential to phage survival. The final identified protein was an inhibitor of the toxin/antitoxin system (GP 4.5) that was found to interrupt bacterial defense mechanisms against phage infection by preventing conversion of the inert antitoxin to an active toxin.[30] As bacterial resistance to phage infection is a dynamic process and not limited to the onset of cellular invasion, expression of a

protein product that counteracts cellular defenses can be expected to be present throughout the infection cycle.

Of particular interest are the unique proteins that appeared in only one time point across the infection cycle, as these are likely low abundance proteins or those with time-sensitive expression. Table 9 summarizes the unique proteins identified with the *E. coli* culture sample they appear in, class of protein, and gene number. These unique proteins can point to changes in gene expression

Protein Description	0 min	15 min	30 min	45 min	60 min	Class	Gene Number
Protein 19.3	x					III	19.3
Gene 0.4 protein			x			I	0.4
Spanin, outer lipoprotein subunit			x			III	18.7
Protein 19.5			x			III	19.5
Overcome classical restriction gp0.3				x		I	0.3
Protein 1.8				x		II	1.8
Bacterial RNA polymerase inhibitor				x		II	2
Protein 7.7				x		III	7.7
Protein 4.1					x	II	4.1
Tail tubular protein gp12					x	III	12

Table 9: Proteins uniquely identified in T7 phage infected *E. coli* samples.

over time, as well as potential deficiencies in analytical method if an ubiquitous protein fails to be identified consistently. Unfortunately, four of these unique proteins have functions that are as of yet undetermined: protein 1.8, protein 4.1, protein 7.7, and protein 19.3, and are therefore likely low abundance, non-essential proteins that provide little information to the status of gene expression at the given time. Protein 19.5 does not have a well-defined function as well, but was found to play an important, if non-essential role in degrading host cell DNA.[31] Gene product 0.4, identified only in the 30 minute post-infection sample, is a Class I protein typically expressed only in the first 2 minutes of viral infection that

inhibits *E. coli* cellular division.[32] As this protein was detected only after 30 minutes following the initial infection, it was an indicator that the 30 minute sample captured T7 after completion of the first lytic cycle, and early into infection of *E. coli* cells by progeny phages. The concentration of this protein was likely too low during the sampling points prior to phage replication. As the phage concentration radically increases with each lytic cycle, the phage protein product concentration should increase accordingly.

Of the total 49 phage proteins identified across all *E. coli* samples, 34 proteins were detected in at least one, but not all time points indicating measurable variation in gene expression over the T7 phage infection. Table 10 summarizes the proteins that varied in expression, their class, and gene numbers. The disappearance of a protein from one time point to the next may be an indication of downregulation, just as the appearance of a new protein may be an indication of upregulation. As bacteriophage T7 typically reproduces on a roughly 30 minute timescale at physiological temperature,[12] the experiment encompassed at least two lytic cycles of the phage. This replication greatly increased the concentration of phage proteins in later *E. coli* samples, as the infection spread and more cells had their DNA transcription machinery hijacked into producing viral proteins. As T7 phage reproduced, new copies of each protein expressed would be produced in the progeny phages, which upon enzymatic digestion would generate new peptides above the detection limit of the mass spectrometer.

Protein Description	0 min	15 min	30 min	45 min	60 min	Class	Gene Number
Terminase, small subunit gp18	X	X	X	X		III	18
Protein 7	X	X	X			III	7
Protein 0.6B	X	X		X	X	I	0.6
Protein 1.6	X	X		X		I	1.6
Protein 6.7	X	X		X		II	6.7
Uncharacterized protein 1.1	X	X				I	1.1
Protein 4.3	X		X	X		II	4.3
Single-stranded DNA-binding protein gp2.5	X			X	X	II	2.5
Protein 3.8	X			X	X	II	3.8
Probable RecBCD inhibitor gp5.9	X			X	X	II	5.9
Capsid assembly scaffolding protein	X			X	X	III	9
Protein 19.3	X					III	19.3
Protein kinase 0.7			X	X	X	I	0.7
DNA ligase			X	X	X	I	1.3
Endolysin			X	X	X	II	3.5
Terminase, large subunit gp19			X	X	X	III	19
Tail tubular protein gp11			X	X		III	11
Protein 5.3			X		X	II	5.3
Fusion protein 5.5/5.7			X		X	II	5.5
Protein suppressor of silencing			X		X	II	5.5
Gene 0.4 protein			X			I	0.4
Spanin, outer lipoprotein subunit			X			III	18.7
Protein 19.5			X			III	19.5
Inhibitor of dGTPase				X	X	I	1.2
Nucleotide kinase gp1.7				X	X	I	1.7
Protein 6.5				X	X	II	6.5
Protein 7.3				X	X	III	7.3
Overcome classical restriction gp0.3				X		I	0.3
Protein 1.8				X		II	1.8
Bacterial RNA polymerase inhibitor				X		II	2
Protein 7.7				X		III	7.7
Protein 4.1					X	II	4.1
Tail tubular protein gp12					X	III	12

Table 10: Phage proteins identified in *E. coli* digests, arranged by time point.

From the pattern of proteins identified in Table 10, an example of differential gene expression over the period of infection potentially emerged from the interaction of the small subunit of terminase (gp18, denoted terminase-S) and

the tail tubular protein (gp12). Terminase-S was detected in the 0, 15, 30, and 45 minute post-infection samples while the tail tubular protein was only identified in the 60 minute post-infection cell culture. Terminase-S acts as essential chaperone in the DNA packaging process in preparation for release of new phages, but is generally not detected in completed virions.[33] After packaging DNA into the protocapsid, terminase is replaced with the tail tubular protein (gp 12) prior to release of completed new phages.[34] Downregulation of gp18 was evident prior to the 60 minute post-infection sampling, when gp12 was being upregulated, as evidenced by the detection of the tail tubular protein for the first time in the 60 minute sample. While this cycle of production of gp12 and downregulation of terminase-S presumably occurred earlier during the infection period, it is likely that the 100x increase in phage concentration after the first lytic cycle pushed the concentration of gp12 above the threshold for detection. Interestingly, while tail tubular protein (gp12) is only identified in the 60 minute post-infection sample, gp12 was identified to a higher confidence than most of the other phage proteins as it was one of the six phage proteins assigned a non-zero score. Of the 49 different phage proteins detected in the infected *E. coli* culture, gp12 was the only protein to be only identified at one time point and have a non-zero protein score assigned by SEQUEST. Four of the six scored phage proteins were identified in all time points sampled even if the protein was not scored in all sample digests. The remaining scored protein was scored in one time point but identified in three time points. As the tail tubular protein appeared in only one time point, but was identified confidently enough to be assigned a

score by SEQUEST, it was an indicator of changing gene expression and not a deficiency in experimental setup. Identification of a predictable pattern of gene expression across the infection cycle helps to confirm the ability of this mass spectrometry-based proteomics experiment to track and monitor viral infection. Method improvements implemented following the analysis of the *E. coli* digests discussed here can provide a path forward to investigating other viral-bacterial interactions with greater sensitivity.

While the identification of phage proteins in a predictable manner consistent with available literature confirmed the adequacy of the experimental design, one of the unique proteins identified in Table 9 demonstrated the systemic limitations of the method. Identified in only the 30 minute post-infection sample, the outer lipoprotein subunit of spanin (gp18.7, o-spanin) was unlike the gp0.4 protein identified in a single time point due to gene expression only over a limited timeframe. The outer and inner subunits of spanin form a complex that spans the periplasmic space of a Gram-negative host cell, like *E. coli*, and are linked by interactions between the C-terminal ends of each subunit.[29] Therefore, o-spanin and i-spanin should not be identified without the presence of both subunits. The inner membrane subunit was identified in all time points with a range of sequence coverages, from 5.59% in the 30 minute post-infection sample to 37.76% in the 60 minute post-infection sample. The increase of T7 phage concentration improved the sequence coverage percentages after the second lytic cycle, evident in the 45 and 60 minute *E. coli* samples which nearly doubled the coverage of the 0 and 15 minute samples.

A potential contribution to the discrepancy in identification of the two spanin subunits was their relative sizes, with the inner membrane portion comprised of 143 amino acids and the outer lipoprotein portion comprised of 83 amino acids. As the length of i-spanin was nearly double that of o-spanin, vastly more possibilities for producing peptides with m/z in the mass range of the ion trap existed for the inner spanin subunit. The most significant contribution to the poor detection of o-spanin despite its implied presence wherever i-spanin appeared was the primary sequence of o-spanin. The primary sequence lacked basic residues at the C-terminal end that upon tryptic digestion would create peptides within the m/z ratio range detectable by the ion trap mass spectrometer used. Past residue 32, only two peptides would be generated upon digestion and both with m/z outside of the detectable range, and indeed only peptides produced from the N-terminal end of o-spanin were observed as seen in Figure 12. Additionally, the numerous basic residues present at the N-terminal end of

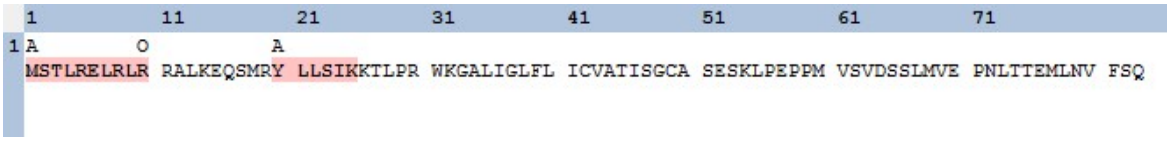

Figure 12: Primary sequence of spanin, outer lipoprotein subunit, from 30 minute post-infection *E. coli* sample. Portions highlighted in red were observed peptides.

the protein, would render peptides too short with m/z ratios too low to be selected for fragmentation by the data-dependent acquisition algorithm upon complete tryptic digestion. As SEQUEST was set to search for peptides of at least five residues, full cleavage by trypsin at each lysine and arginine would result in at least six peptides too short to be assigned even if the precursor ion was isolated and fragmented. The region between the two identified peptides in Figure 12,

highlighted in red due to the low confidence of peptide assignment, contained three basic residues in a nine residue sequence which would generate small, multiply charged peptides assuming full cleavage. Missed cleavages are common and in this instance, it may be possible to improve coverage of the N-terminal end of o-spanin, but the C-terminal end past residue 33 will likely never be detected within the mass range of the LTQ mass spectrometer with a trypsin digestion. Incorporating other enzymes could cleave the protein differently to produce more peptides within the available mass range, but would also potentially create new peptides too short to be detected or identified by SEQUEST. Though data analytics software was an integral component of bottom-up proteomics studies, particularly for a complex sample such as a cell culture digest, automation of mass spectra interpretation requires setting parameters that will always exclude some percentage of possible information. Peptides of length below the cutoff were assuredly present and even presumably detected by the mass spectrometer but would not be identified by SEQUEST. However, assigning peptides of only a few residues within a proteome for an organism as complex as *E. coli* would be effectively meaningless, as many proteins have similar sequences. That the outer lipoprotein spanin subunit should have been present in all samples that the inner spanin subunit appeared, but did not was an indication of the inherent limitations of bottom-up proteomics.

Despite these limitations, these early results obtained for *E. coli* infected with T7 phage provided proof of concept data to validate the experimental setup. The method, even when used with the in-solution digestion, provided valuable

infection monitoring information and was able to identify 86% of the phage proteome over the course of the infection. Taken separately, each time point provided a snapshot into phage and host gene expression and taken together, much of the phage proteome was categorized. Implementation of SDS-PAGE separation prior to enzymatic digestion should decrease sample complexity and allow for higher confidence in protein assignment moving forward, enabling deeper insight into phage-host interactions.

Bibliography

(1) Persidis, A. An ambitious drug development platform attempts to link gene sequence to expressed phenotype under various physiological states. *Nature Biotech.* **1998**. *16*, 393-394.

(2) Nilsson, T.; Mann, M.; Aebersold, R.; Yates III, J. R.; Bairoch, A.; Bergeron, J. J. M. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nature Methods*. **2010**. *7*, 681-685.

(3) Ogada, P. A.; Kiirika, L. M.; Lorenz, C.; Senkler, J.; Braun, H.; Poehling, H. Differential proteomics analysis of *Franklinella occidentalis* immune response after infection with *Tomato spotted wilt virus* (Tospovirus). *Dev. Comp. Immunol.* **2017**. *57*, 1-7.

(4) Aslebagh, R.; Channaveerappa, D.; Arcaro, K. F.; Darie, C. C. Comparative two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) of human milk to identify dysregulated proteins in breast cancer. **2018**. *Electrophoresis*. *39*, 1723-1734.

(5) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.; Yates III, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **2013**. *113,* (4), 2343-2394.

(6) Kwiatkowski, M.; Krösser, D.; Wurlitzer, M.; Steffen, P.; Barcaru, A.; Krisp, C.; Horvatovich, P.; Bischoff, R.; Schlüter, H. Application of displacement chromatography to online two-dimensional liquid chromatography coupled to tandem mass spectrometry improves peptide separation efficiency and detectability for the analysis of complex proteomes. **2018**. *Anal. Chem. 90*, (16), 9951-9958.

(7) Benedikter, B.J.; Bouwman, F. G.; Vajen. T.; Heinzmann, A. C. A.; Grauls, G.; Mariman, E. C.; Wouters, E. F. M.; Savelkoul, P. H.; Lopez-Iglesias, C.; Koenen, R. R.; Rohde, G. G. U.; Stassen, F. R. M. Ultrafiltration combined with size exclusion chromatography efficiently isolates extracellular vesicles from cell culture media for compositional and functional studies. **2017**. *Sci. Rep. 7*, (1), 15297.

(8) Wang, H.; Clouthier, S. G.; Galchev, V.; Misek, D. E.; Duffner, U.; Min, C.; Zhao, R.; Tra, J.; Omenn, G. S.; Ferrara, J. L. M.; Hanash, S. M. Intact-protein based high-resolution three-dimensional quantitative analysis system for proteome profiling of biological fluids. *Mol. Cell. Proteomics*. **2005**. *4*, (5), 618-625.

(9) Juraschek, R.; Dulcks, T.; Karas, M. Nanoelectrospray - more than just a minimized-flow electrospray ionization source. *J. Am. Soc. Mass Spectrom.* **1999**. *10,* (4), 300-308.

(10) WHO Fact Sheet. The top 10 causes of death. **2018**. http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death. (Accessed 29 September 2018).

(11) Tenaillon, O.; Skumik, D.; Picard, B.; Denamur, E. The population genetics of commensal *Escherichia coli*. *Nat. Rev. Microbiol.* **2010**. *8,* 207-217.

(12) Banu, M.; Ng, D.; Zheng, L.; Goh, L.; Bi, X.; Ow, D. S. Rapid quantification of *Escherichia coli* in food and media using bacteriophage T7 amplification and liquid chromatography-multiple reaction monitoring tandem mass spectrometry. *J. Biotechnol.* **2014**. *192*, 50-58.

(13) Lu, T. M.; Koeris, M. S. The next generation of bacteriophage therapy. **2011**. *Curr. Opin. Microbiol. 14*, 524-531.

(14) Dunn, J.; Studier, F.; Gottesman, M. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol*. **1983**. *166,* (4), 477-535.

(15) Hernandez, A. J.; Richardson, C. C. Gp2.5, the multifunctional bacteriophage T7 single-stranded DNA binding protein. **2018**. *Semin. Cell Dev. Biol.* https://doi.org/10.1016/j.sernCdb.2018.03.018.

(16) Jeong, H.; Barbe, V.; Lee, C. H.; Vallenet, D.; Yu, D. S.; Choi, S.; Couloux, A.; Lee, S.; Yoon, S. H.; Cattolico, L.; Hur, C.; Park, H.; Ségurens, B.; Kim, S. C.; Oh, T. K.; Lenski, R. E.; Studier, F. W.; Daegelen, P.; Kim, J. F. Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). **2009**. *J. Mol. Biol. 394*, 644-652.

(17) Pierce™ Mass Spec Sample Prep Kit for Cultured Cells Instruction Manual. **2013**. https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0011864_Pierce_MassSpecSamplePrep_CulturedCells_UG.pdf. (Accessed 11 October 2018).

(18) Pierce BCA Protein Assay Kit Instruction Manual. **2015**. https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0011430_Pierce_BCA_Protein_Asy_UG.pdf. (Accessed 11 October 2018).

(19) Pierce® C18 Spin Column Instruction Manual. **2011**. https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0011495_ Pierce_C18_SpinCol_UG.pdf. (Accessed 12 October 2018).

(20) Mini-Protean Precast Gels Instruction Manual and Application Guide. **2011**. http://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_1658100.pdf. (Accessed 11 October 2018).

(21) Arizona Proteomics Consortium Protocol: Tryptic digestion of protein in gel bands. **2012**. http://proteomics.arizona.edu/sites/proteomics.arizona.edu/files/In_gel_trypsin_digestion_May2012.pdf. (Accessed 12 October 2018).

(22) Gundry, R. L.; White, M. Y.; Murray, C. I.; Kane, L. A.; Fu, Q.; Stanley, B. A.; Van Eyk, J. E. Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. *Curr. Protoc. Mol. Biol.* **2010**. 9*0*, (1), 10.25.1-10.25.23.

(23) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal. Chem.* **1996**, *68*, (5), 850-858.

(24) Eshraghi, J.; Chowdhury, S. K. Factors affecting electrospray ionization of effluents containing trifluoroacetic acid for high-performance liquid chromatography/mass spectrometry. *Anal. Chem.* **1993**. *65*, 3528-3533.

(25) University of Washington Proteomics Resource. Packing capillary columns and pre-columns (traps). **2013**. https://proteomicsresource.washington.edu/docs/protocols05/Packing_Capillary_Columns.pdf. (Accessed 17 October 2018).

(26) Meiring, H. D.; van der Heeft, E.; ten Hove, G. J.; de Jong, A. P. J. M. Nanoscale LC-$MS^{(n)}$: technical design and applications to peptide and protein analysis. **2002**. *J. Sep. Sci. 25*, 557-568.

(27) Ionel, A.; Velázquez-Muriel, J. A.; Luque, D.; Cuervo, A.; Castón, J. R.; Valpuesta, J. M.; Martín-Benito, J.; Carrascosa, J. L. Molecular rearrangements involved in the capsid shell maturation of bacteriophage T7. **2011**. *J. Biol. Chem. 286*, (1), 234-242.

(28) Moak, M.; Molineux, I. J. Role of the Gp16 lytic transglycosylase motif in bacteriophage T7 virions at the initiation of infection. **2000**. *Mol. Microbiol. 37*, (2), 345-355.

(29) Summer, E. J.; Berry, J.; Tran, T. A.; Niu, L.; Struck, D. K.; Young, R. Rz/Rz1 lysis gene equivalents in phages of Gram-negative hosts. **2007**. *J. Mol. Biol. 373*, (5), 1098-1112.

(30) Sberro, H.; Leavitt, A.; Kiro, R.; Koh, E.; Peleg, Y.; Qimron, U.; Sorek, R. Discovery of functional toxin/antitoxin systems in bacteria by shotgun cloning. **2013**. *Mol. Cell. 50*, (1), 136-148.

(31) Kim, S-H.; Chung, Y-B. Isolation of a mutant bacteriophage T7 deleted in nonessential genetic elements, gene *19.5* and *m*. **1996**. *Virology*. *216*, 20-25.

(32) Kiro, R.; Molshanski-Mor, S.; Yosef, I.; Milam, S. L.; Erickson, H. P.; Qimron, U. Gene product 0.4 increases bacteriophage T7 competitiveness by inhibiting host cell division. **2013**. *Proc. Natl. Acad. Sci. U. S .A.*, *110*, (48), 19549-19554.

(33) White, J. H.; Richardson, C. C. Processing of concatemers of bacteriophage T7 DNA *in vitro*. **1987**. *J. Biol. Chem. 262*, (18), 8851-8860.

(34) Cuervo, A.; Pulido-Cid, M.; Chagoyen, M.; Arranz, R.; González-García, V. A.; Garcia-Doval, C.; Castón, J. R.; Valpuesta, J. M.; van Raaij, M. J.; Martín-Benito, J.; Carrascosa, J. L. Structural characterization of the bacteriophage T7 tail machinery. **2013**. *J. Biol. Chem. 288*, 26290-26299.