Temporal gene expression and regulation in T4 phage

by
Mandara Alexis Levine

B.S. Biology, Haverford College, 2018

SUBMITTED TO THE MICROBIOLOGY GRADUATE PROGRAM IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2024

Signature of Author:

Mandara A. Levine
Microbiology Graduate Program
November 17, 2023

Certified by:

Michael T. Laub
Professor of Biology
Thesis Supervisor

Certified by:

Gene-Wei Li
Associate Professor of Biology
Thesis Supervisor

Accepted by:

Jacquin C. Niles
Professor of Biological Engineering
Co-Director of the Microbiology Graduate Program

Temporal gene expression and regulation in T4 phage

by
Mandara Alexis Levine

Submitted for the Microbiology Graduate Program on November 17, 2023 in partial fulfillment of the requirement for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology

ABSTRACT

As the most abundant biological entity in the biosphere, bacteriophages play a critical role in shaping the microbial diversity, and thus overall ecosystem health. They are also essential tools in molecular biology, shedding light on fundamental biological concepts. T4 phage, with its complex lifecycle and genetic content, has been instrumental in many such discoveries. However, many questions regarding gene regulation in T4 phage remain unanswered. In this study, we employ end-enriched RNA-seq (Rend-seq) and ribosome profiling to examine T4 RNA and protein synthesis throughout the course of infection, gaining new insights at the transcriptional, translational, and genomic level. At the transcriptional level, we identified transcript boundaries, novel putative promoters, and new potential cleavage sites for the T4 endoribonuclease RegB. At the translational level, we identified many instances of previously unreported changes in translational efficiency over the course of infection, indicating the presence of intricate and uncharacterized mechanisms of regulation. Collectively, transcriptional and translational controls lead to precisely tuned protein synthesis rates during infection, as exemplified by the phenomenon that components of T4 protein complexes are synthesized according to their stoichiometry— a principle that has been observed in organisms during steady-state growth. Finally, we identified and experimentally validated T4's 290th gene, *61.-1*. Though non-essential to T4 in laboratory conditions, this gene has homologs present in a number of other phage and drastically impacts *E. coli* growth when ectopically expressed. This study provides insights into T4 phage biology, paving the way for further exploration into molecular biology, virology, and biotechnology; our rich data set can be utilized by future studies to answer a diverse array of inquiries.

Thesis Supervisor: Michael T. Laub, Title: Professor of Biology
Thesis Supervisor: Gene-Wei Li, Title: Associate Professor of Biology

ACKNOWLEDGMENTS

To my microbiology graduate program cohort, thank you: Veda, Annika, Jade, and Irene; you were the first friendly faces of MIT for me. With only the five of us, we all became very close very fast, and truly pulled each other through grad school. We studied together, offered moral support, and got each other in trouble in class for sending the worst and funniest photos we could find (I will never forget cholera girl). I am so grateful for you all. To Veda in particular, thank you for becoming my closest friend at MIT. Our walks (taking turns listening and giving much needed free therapy), our Newbury Street adventures (spending entirely too much of our stipends), and our cooking experiences (sushi and summer rolls in particular) are some of my most treasured memories.

To my lab mates, thank you: Mirae Parker and Manraj Gill, I feel so lucky to have joined the lab at the same time as you and to have progressed through this adventure together. You are both incredibly kind and selfless people who are ready to help anyone at any time, and I am so glad that we were on this journey together. Lydia Herzel, Matthew Tien, Julia Dierkshiede, Julian Stanley, Julia Ramseyer, Daniel Saxton, and Sriram Srikant; you have helped me grow so much as a scientist. From brainstorming sessions to help with specific challenging protocols, you all were always so patient with me and engaged with my work, and that has been invaluable to me during my time as a graduate student.

To my previous mentors, thank you. Mr. Shurtz, you were the first real non-family mentor I ever had. The way you believed in me in high school was so impactful that I still think of you often, all these years later. Iruka Okeke, Paul Brodfuehrer, and Mark Goulian, you allowed me into your labs during my summers in undergrad, and truly cultured my love for scientific research. Rachel Hoang and Phil Meneeley, you helped me throughout my time at Haverford to become a better scientist. Doing my undergrad thesis with you, Rachel, showed me that I could become a true scientist by letting me plan, execute, analyze, and share my experiments. Phil, I will never forget your enthusiasm at the idea of me going to graduate school, and that really gave me the confidence to apply to schools at the end of my senior year, knowing I had your full support.

To my friends, thank you. Zoe McDoom, you were my very first friend, and I will forever be grateful to you for being an amazing friend to me, and helping me learn how to become a good friend as well. Marisa Egan and Tera Egan, thank you for being such wonderful friends to me throughout our childhood. Marisa, we really have lived in such parallels and supported each other steadfastly throughout every experience. I feel so lucky to be a part of your life. Maggie Sawyer, Amy Zamora, and Amanda Soled, thank you for pulling me through undergrad. I didn’t have many friends, but I had you three and I never doubted the strength of our friendships, which is worth so much to me. Maggie, I am so grateful that we were placed together as roommates freshman year; we share so many wonderful memories and continue to support each other throughout every new stage of life that we approach. I love our regular chats and all the pictures you send me of Egg.

To my family, thank you. Grampa and Grampa, you are my inspiration for everything in life. Even when I feel completely lost and directionless, I know that I want to be like you two, and that always pulls me through. You are my forever role models, and your support throughout my entire life has meant the world to me. Mom, no one has the kind of bond we have, and I am so incredibly thankful for our relationship. I don’t think I spent even a fraction as many hours on the phone with anyone during grad school as I did with you; I always love talking to you (or singing songs, repeating silly jokes, or making bizarre noises) and just knowing you are there for me on the other side of the line. Kamila, you are the most wonderful little sister. You always want me to be happy and you will do anything you can to make that happen; every time you tell me I am your best friend, I am so grateful that we are able to be friends in addition to sisters. Neither of us remember a life without the other, and it is so special that we still are here supporting each other to this day. To Dad, thank you for spending so much time and effort helping me be able to pursue the things that I loved. Doing things like karate and reading, which you always supported, really helped me to grow as a person. Uncle Mark, thank you for being my go-to scientific mentor from the start. You have always done everything in your power to help me achieve my goals, and to make sure I enjoy myself along the way as well.

To my husband, thank you. Max, though I entered MIT not knowing you, I now cannot imagine my life without you. The years I have spent in graduate school have been full of the most difficult times, to the point where I truly wasn’t sure I could make it through. But every single time I have struggled, you have pulled me up from even the deepest, darkest places that I’ve fallen into. You are the kindest, strongest, and most selfless person I have ever met. You know every version of me and love me for all of them, which is something I never believed could happen. I love you so much and I am so proud of you. I really believe that together, we will build the most amazing life and family together. There is so much more to come, and I can’t wait to do it all with you.

TABLE OF CONTENTS

Chapter 1: Introduction

INTRODUCTION

A fascinating topic within molecular biology is the study of viruses. Viruses are infectious agents composed of genetic material, either DNA or RNA, enclosed in a protein coat. Viruses have the ability to replicate within a host, but are unable to do so independently. They are extremely diverse, and responsible for numerous diseases in humans, as well as many plant and animal infections. Their unique mode of infection involves attaching to host cells, injecting their genetic material, and hijacking the host's cellular machinery to reproduce. Understanding viruses is essential for combating diseases and has led to the development of vaccines and antiviral treatments, demonstrating their profound impact on biology and medicine. Because viruses are so complex and so crucial to understand, there are still many unanswered questions about them, some of which have never been posed. Because of this, viruses are a perfect candidate for exploratory science.

PHAGE OVERVIEW

Phage biology

Phages, also known as bacteriophages, are viruses that infect bacterial cells; because they are a type of virus, they cannot replicate outside of their hosts. Phages are the most abundant biological agent on earth and are extremely diverse in size, morphology, and genomic organization (Kasman et al., 2022). All phages consist of a nucleic acid genome encased in a shell of phage-encoded capsid proteins, which protect the genetic material and mediate its delivery into the next host cell (Kasman et al., 2022).

Once a phage has infected its host, there are two replication strategies which it may enact: lysogenic or lytic. In temperate phage, once the phage’s genetic material is injected into the host cell, phage DNA integrates itself into the host's chromosome (thereby becoming what is known as a "prophage"). This prophage replicates along with the host's DNA and is passed on to the next generation of cells during cell division. The prophage can remain dormant for many generations, but can then be activated to enter the lytic cycle under certain conditions. In the lytic cycle, which is the only cycle lytic phage undergo, the phage takes over the host's cellular machinery to produce new viral genomes and capsid proteins, which will assemble into multiple copies of the original phage. The host cell is rapidly destroyed, releasing new progeny phage to infect other cells (Clokie et al., 2011).

Phages are also in a constant arms-race with the bacteria they infect; bacteria have developed phage defense systems, leading to the evolution of systems within phage that help them to overcome the phage defense systems carried by the bacteria. There are many known examples of this, particularly on the side of phage defense systems. Some well-characterized systems found in bacteria are restriction-modification, abortive infection, and CRISPR (clustered regularly interspaced short palindromic repeats). Restriction-modification systems are able to recognize DNA that is not host cell DNA by detecting unmethylated DNA, and therefore degrading any other DNA that is not methylated, such as that from an infecting phage (Rodic et al., 2017). CRISPR aids a bacterial cell to create a "memory" of different viruses so it is able to attack the phage DNA and resist infection (Gostimskaya et al., 2022). Abortive infection systems act by killing the host cell during phage infection, before the phage can form fully-developed progeny, and in this way the bacterium sacrifices itself to prevent multiplication of the phage— a process that could lead to infection of many other bacterial cells in the surrounding population (Lopatina et al., 2020).

Genome annotation

Genome annotation is an essential step in genome analysis that helps to identify the functional elements of DNA, provides an overview of gene content, and helps to understand the function of a gene and how variation might affect its function. It is an important process to carry out in regards to viruses, in order to identify functional elements within a viral genome, such as genes, regulatory regions, and non-coding sequences. This annotation is crucial for understanding the biology, evolution, and pathogenicity of viruses. There are numerous methods that are used for genome annotation, including both experimental and computational approaches.

To predict the location of genes within a genome computationally, two main methods have been harnessed: ab initio methods and homology-based methods. It is also possible to predict genes using a combination of the two methods. Ab initio methods are based on statistical models of gene features, and predict gene structure without relying on any additional information, which is useful for predicting genes in newly sequenced genomes or genomes that have relatively few resources for their annotation (Baker et al., 2023). Homology-based methods use external evidence, such as similarity to annotated sequences. Part of this process can involve the use of expressed sequence tags (ESTs), in which cDNA representing certain expressed genes is isolated and used to find the matching portion of chromosomal DNA (Ejigu et al., 2020).

Often, computational methods are used as a starting point to identify the presence of genes, and then experimental techniques are employed to validate gene function (Shoemaker et al., 2001). Methods such as microarrays or RNA-sequencing are used to ensure that predicted genes are indeed expressed, and then more specific approaches must be taken to investigate the precise function of genes (Shoemaker et al., 2001). These approaches could involve the creation of genetic knockouts, the employment of gene overexpression, and protein-protein interaction analysis. Knockout experiments demonstrate if a gene is essential to an organism, and can elucidate which biological process a gene is involved in. Similarly, overexpression experiments can reveal the role the gene plays by phenotypic observation of what occurs when there is an elevated level of gene expression, and whether too much of a certain protein can cause damage or even render an organism non-viable. Protein-protein interaction analysis is also useful, as identifying the binding partner of an unknown protein can give insight into the biological process it is involved in, based on the known function of the binding partner.

Even with the combination of computational and experimental techniques, it can be difficult to identify all genes within an organism. For example, some ORFs contain multiple start sites and it can be unclear what the correct start site is, or if multiple proteins are encoded by the ORF using internal start sites, as is the case with T4 gene *17*. Another key challenge is the presence of many start and stop codons in frame with each other throughout the genome, though they do not all code for proteins. Because of this, gene prediction tools often set minimum ORF length thresholds to prevent inaccurate identification of all small ORFs as true genes, when many are not (Fremin et al., 2022; Hyatt et al., 2010). However, there are still a number of validated genes that are quite small, such as the T4 gene *stp*, which encodes a protein that is only 26 amino acids long. While the length cutoff does prevent an overwhelming number of false positive gene identifications, that does not truly rule out all the smaller ORFs as being possible true genes, which is why high throughput methods that detect RNA and protein production in a cell can be helpful in assessing which ORFs are truly genes that produce transcripts and proteins.

T4 PHAGE

T4 lifecycle

The discovery of phage occurred in 1915, in the hands of William Twort (Clokie et al., 2011). Subsequently, in 1917 Felix d'Herelle discovered that phage were able to kill bacteria (Clokie et al., 2011). Though antibiotics were easier to administer, and therefore were used over phage in cases of infectious disease, phage research continued. Early phage work focused mainly on phages that infected types of *E. coli*, and ended up uncovering many fundamental tenants of molecular biology as a whole. T4 phage in particular has been instrumental in the understanding of many fundamental biological concepts. One such concept being the discovery of DNA as the genetic material across the tree of life; this was elucidated through an experiment where the sulfur in proteins and phosphorus in DNA were radiolabeled, and it was observed that it was DNA that entered the cell infected by the phage and allowed for the production of new phage (Hershey et al., 1952).

Not only did T4 allow for the discovery that DNA was the genetic material, it also helped to show that the genetic code was read in groups of three bases, also known as codons (Crick et al., 1961). It was also presumed that the genetic code was also degenerate in the sense that different codons could encode the same amino acid (Crick et al., 1961). Another discovery made through the use of T4 phage was that bacterial mutations for phage resistance arise in the absence of selection, rather than being a response to selection, and that bacteria do indeed have genes, as it was previously believed that bacteria did not have such features (Luria et al., 1943). T4 was also key in the discovery that introns were present in organisms other thank eukaryotes (Miller et al., 2003).

Studying T4 has also given insight into phage defense systems, most prominently restriction-modification systems (Luria et al., 1952). Restriction-modification systems (R-M systems) protect the host bacterial cell from phage infection through recognition and cleavage of foreign DNA. This is achieved through the host cell modifying its own DNA and producing an endonuclease that will cleave any DNA that is unmodified (Enikeeva et al., 2010). Many other phage defense systems and counter defense systems have been discovered since the discovery of R-M systems, and each play an important role in understanding the interplay between bacteria and phage. Ultimately, this knowledge has the potential to be exploited in the search for new therapies targeting bacterial infections. It is clear that T4 phage has provided scientists with an abundance of knowledge about phage, and also about fundamental molecular biology as a whole.

T4 is a double-stranded DNA virus that infects *E. coli*. It is also a lytic phage, whose life cycle begins after binding and injecting its DNA into the host cell, and completes itself over the course of approximately 30 minutes at 37˚C. T4's infection cycle begins with its adsorption to the surface of the host *E. coli* cell. T4 then injects its genetic material into the host cell wall, directly through the cell wall and into the cytoplasm of the host bacterium (Maghsoodi et al., 2019). This injection is facilitated by the contraction of the T4 tail sheath, which acts like a hypodermic needle to deliver the phage DNA. Following injection, the T4 phage DNA enters the host bacterium's cytoplasm. During this stage, new phage DNA and proteins are synthesized using the host cell's machinery (Kutter et al., 2018). Following this, the newly synthesized T4 phage components, including the DNA, proteins, and other structural elements, come together to form new phage particles. As the newly assembled phage particles mature, they undergo final structural modifications and maturation processes. This involves the acquisition of a tail, tail fibers, and the assembly of a head containing the phage DNA (Yap et al., 2014). Once the new T4 phage particles are fully mature, they trigger the lysis of the host bacterium. Enzymes produced by the phage weaken the bacterial cell wall, and the cell lyses, releasing the newly formed T4 phages into the surrounding environment (Maghsoodi et al., 2019). The mature T4 phage particles then become free to infect new host bacteria. They attach to the receptors on the surface of other *E. coli* cells, inject their DNA, and the cycle begins again.

The structure of T4 phage is exquisitely complex; over 40% of T4's genetic information is dedicated to creating and assembling the head, tails, and tail fibers (Miller et al., 2003). The T4 phage head in particular is a prolate icosahedral structure that is composed of hexameric capsomers made from Gp23, Gp24, and Gp20. (Gamkrelidze et al., 2014). The head also contains a portal protein complex that serves as the entry and exit point for the viral genome during infection; this complex is composed of 12 copies of Gp20 and one copy of Gp17 (Rao et al., 2023). The head also contains decoration proteins (Hoc and Soc) that are involved in stabilizing the capsid structure; though these two proteins are non-essential for infection, they have been utilized by scientists to display pathogen epitopes or antigens on the surface of the head for the purpose of designing vaccines (Rao et al., 2023). The head is also the component of the phage that packages the genetic material. T4's DNA is packaged by the headful, which yields circularly permuted and terminally redundant ends, ensuring the encapsulation of the entire genome. T4 packaging is terminated at a random sequence following packaging of approximately 102% (one headful) of the viral genome, to yield the terminal redundancy of 3.3 kb (Rao et al., 2023). The T4 phage tail is a very large macromolecular complex, comprised of about 430 polypeptide chains. It consists of a sheath, an internal tail tube, and a baseplate, situated at the distal end of the tail (Leiman et al., 2010). The baseplate is responsible for attaching to the host cell, while the tail tube is responsible for injecting the viral genome into the host cell (Leiman et al., 2010). The long tail fibers are built from Gp34, Gp35, Gp36, and Gp37; they compose the part of the tail structure that is responsible for recognizing and binding to specific receptors on the host cell surface (Hyman et al., 2017). The assembly of the T4 phage tail is a complex process that involves the coordinated action of multiple proteins. Gp18 subunits compose the tail sheath, and once the sheath is built to reach the length of the tail tube, Gp15, the tail terminator protein binds to Gp3, which is the head-proximal tip of tail tube, and the final row of Gp18 proteins. (Leiman et al., 2010). Following this, the tail is ready to be attached to the head in the assembly of progeny phage.

In order to release the newly assembled phage, lysis must occur. There are two main proteins involved in this process: Gpe and Gpt. Gpt is a holin, and serves to create holes the inner membrane of *E. coli* to grant access to the peptidoglycan layer to the lysozyme, Gpe, for the purpose of degradation, so that the new phage can escape the cell (Dressman et al., 1999). In the case that there are many phage in the surrounding environment, T4 is able to sense this and delay lysis to wait for additional bacteria to appear in order for the progeny phage to have hosts to infect (Dressman et al., 1999).

T4 Phage transcription

T4 does not encode its own RNA polymerase (RNAP), making it completely dependent on the host core RNAP for transcription of its RNAs. However, T4 does encode many RNAP-associated factors that act to control the timing of phage gene expression (Hinton, 2010). Genes expressed early in infection utilize the host RNAP containing the sigma factor σ^{70}, which is the same sigma factor used in *E. coli* during exponential growth (Hinton, 2010). T4 early promoters contain sequences that match the host σ^{70} - RNAP recognition sequences; it is predicted that T4 early promoters may even have more regions of matching recognition sequence than the *E. coli* promoters have, which, in conjunction with active host RNA degradation, offers some explanation as to how the phage so quickly commandeers the RNAP for its own RNAs (Hinton,

2010) (Wolfram-Schauerte et al., 2022). In addition, it is thought that the T4 protein Alt is able to increase the functioning of certain early promoters through ADP-ribosylation of RNAP, and that the T4 protein Alc can specifically terminate transcription from host DNA (Hinton, 2010).

About one minute after infection, T4 middle transcription begins. Aiding in this transition is the T4 protein RegB, which cleaves certain early mRNAs at a GGAG motif, rendering them inert (Piešiniene et al., 2004). However, RegB does not cleave all GGAG motifs; it is believed to recognize a particular RNA structure which is stabilized by the 30S ribosomal protein S1 (Piešiniene et al., 2004). When RegB cleaves at these motifs, it leaves behind a 5′-OH and a 2′, 3′-cyclic phosphate at the resulting termini (Durand et al., 2012). T4 polynucleotide kinase/phosphatase (PNK) can phosphorylate the 5′-OH end, allowing for host RNase G and RNase E to make cuts that lead to the degradation of these mRNA fragments (Durand et al., 2012).

Middle transcription is also highly mediated by two T4 proteins: AsiA and MotA. AsiA binds to σ^{70} in order to inhibit transcription of early T4 promoters and any host promoters (Adelman et al., 1998). It also serves to enhance transcription of middle promoters (Adelman et al., 1998). MotA binds to a DNA recognition element stretching from the -32 and -27 positions of the T4 middle promoters known as the “MotA box” in order to set in motion transcription with the AsiA-bound RNAP (Hinton, 2010). Both proteins are necessary for the activation of middle promoters, making these two proteins responsible for the precise timing of middle transcription during the infection cycle (Hinton, 2010). It is also believed that the ADP-ribosylating enzymes ModA and ModB act to turn off many early promoters in the transition to middle transcription (Hinton, 2010).

Late transcription begins when the sliding clamp of the T4 replisome, Gp45, recruits the RNAP to late promoters through interactions with two phage-encoded RNAP subunits: Gp33 and Gp55 (Geiduschek et al., 2010). Gp33 is the co-activator of late transcription that is bound to the DNA, while Gp55 serves as the late promoter recognition protein that is bound to RNAP (Geiduschek et al., 2010). Also essential to the start of late transcription are Gp44 and Gp62, which form the clamp-loading complex that loads Gp45 onto DNA; this loading occurs at single-strand breaks or primer-template junctions, which are also sites at which DNA replication begins, explaining why DNA replication is necessary for late transcription to occur (Nechaev et al., 2008).

T4 Phage translation

Much like how T4 transcription relies on host RNAP, T4 translation relies on host ribosomes and tRNAs, though it does also encode some tRNAs that are more commonly used to make T4 proteins than *E. coli* proteins (Miller et al., 2003). Upon infection, T4 is able to quickly commandeer host translational machinery, redirecting translational efforts to create almost exclusively phage proteins (Kutter et al., 2018). While there is no concrete answer as to how this is accomplished, it is thought that T4 may modify host ribosomes in some way. This theory is formed in part due to how transcription of host mRNA can be induced after phage infection, but no host mRNA is found to be associated with ribosomes (Kennell, 1970). Additionally, when ribosomes are isolated from T4-infected cells, they appear to be primed to translate mRNA from that phage, rather than host mRNA or mRNA from a different phage (Hsu et al., 1969). Research in this area performed on jumbo phage ϕKZ uncovered a phage factor that binds the 5S ribosomal RNA, but it is unconfirmed what the exact effect the factor has on protein synthesis (Gerovac et al., 2023).

Despite a lack of understanding in the switch from host to T4 protein production, there are examples of T4 translational phenomena that have been characterized, including a process known as “ribosome hopping” that has been identified in gene 60, where ribosomes skip over a gap of 50 nucleotides while translating its mRNA (Herbst et al., 1994). There are also better understood examples of translational regulation in T4. Gp32 and Gp43 are known to repress their own translation, as the proteins bind to the RBSs of their own mRNAs, occluding ribosomes (Andrake et al., 1988). RegA performs a similar self-repression, but also represses translation of a number of other early genes (Winter et al., 1987).

Importance of studying T4 phage

Over a century after their discovery, phages still are a topic of great interest among scientists. A growing field of research regarding phage is phage display, a technique in which phage capsids are modified in such a way that they display foreign proteins of interest. This allows for the discovery of binding partners to a specific protein using a large diversity of variant proteins as the potential partner. In T4, this is accomplished with the use of the non-essential capsid decoration proteins Hoc and Soc. A T4 strain containing a deletion of either *hoc* or *soc* is used to infect bacteria expressing fusion proteins on a plasmid. In these plasmids, either *hoc* or *soc* is translationally fused to the test protein of interest in order to localize the test protein to the phage capsid during the production of progeny phage (Gamkrelidze et al., 2014). A large diversity of proteins can be represented using a library of different fusions to the capsid protein. Phage display holds great promise in the area of vaccine development, as T4 particles displaying different antigens can cause substantial antibody responses. In fact, it has been shown that T4 can display multiple antigens fused to Hoc are able to be displayed on the same capsid at the same time (Shivachandra et al., 2007). This is critical in terms of vaccines, as most pathogens encode numerous virulence factors, and in order to effectively defend against the pathogen, vaccines must result in antibodies against multiple of these factors. In fact, in one study, it was demonstrated that T4 particles simultaneously displaying multiple antigens against anthrax was able to elicit an immune response in mice (Shivachandra et al., 2007). The use of T4 phage display technology in vaccine development has the potential to revolutionize the field of vaccinology by providing a platform for the rapid and efficient development of vaccines against a wide range of infectious diseases.

Phages have also been studied as an alternative to antibiotics in clinical care. The use of phages as an alternative to antibiotics has been considered for over a century, and there have been many anecdotal examples of it being used effectively in the treatment of bacterial infections. In recent years, there have been a few documented cases of drug-resistant bacterial infections being treated with phage therapy. The most well-known of these if the treatment of a multi-drug-resistant *Acinetobacter baumannii* infection; though the efficacy and safety of treating someone with phage was not deeply understood, the FDA authorized the administration of phage in this case in a desperate attempt to save the patient (Schooley et al., 2017). Samples from the infected individual were taken and tested in the laboratory against environmentally-isolated phage. Though the bacteria initially adapted to be resistant to the phage, the repetition of the phage selection process was finally able to cure the patient (Schooley et al., 2017). This process, though effective, was very time-consuming as it needed to be tailored to the specific bacterial strain of interest, and thus far there have not been protocols developed for widespread use of phage therapy (Hatfull, 2022). It is of great importance to ensure that any phage used in clinical settings be free of any toxic genes, so as to not cause further decline in the patient, underlining the importance of gaining a more in-depth understanding of a wide range of phages (Hatfull, 2022).

There is also value in simply understanding T4 at a deeper level. The genes found in T4 may be conserved in other phage, and the processes carried out during infection may parallel processes carried out in organisms across the tree of life. There are over 120 predicted ORFs in T4's genome that are still of unknown function, and investigating their functions in the T4 lifecycle has the potential to reveal insights reaching well beyond T4 itself (Nolan et al., 2006). T4 has proven itself to be an excellent model system to be explored in the laboratory, and therefore is a perfect starting place to answer such questions.

RNA AND PROTEIN SYNTHESIS

Rend-seq

There have been numerous methods used in the past to measure the expression of genes across the genome in a high-throughput manner. Microarrays are one such method, which function by having oligos representing each gene on an array; RNA is isolated from the system of interest, reverse transcribed into cDNA, and allowed to hybridize to the oligos on the array (Govindarajan et al., 2012). The DNA fragments have fluorescent markers attached to them, so a stronger fluorescent signal indicates a higher degree of gene expression for a particular fragment (Govindarajan et al., 2012). However, this method can only detect signal for locations that correspond to the oligos used on the array rather than capturing all gene expression across the genome. RNA-seq is a more advanced and less biased technique, as it captures RNA expression levels more precisely than microarrays and uses sequencing in order to directly determine any given fragment's sequence (Wang et al., 2009). Rend-seq (end-enriched RNA-seq) gives the same kind of quantification of gene expression levels, but is designed to also give information on transcript ends. End-enrichment is obtained by randomly and sparsely fragmenting RNAs, then selecting for short RNA fragments that are between 15-45 nucleotides long (Lalanne et al., 2018). As a result of this, many of the selected fragments will have one of their ends corresponding to the 5′ or 3′ end of a transcript, since it is unlikely to have two cuts close enough together to make a short fragment when the fragmentation is sparse (Lalanne et al., 2018). This results in a higher degree of signal at the ends of the transcript, with a smaller amount of signal throughout the body of the transcript.

Pinpointing the location of these peaks is valuable in answering many inquiries; this technique gives information on whether or not transcripts contain multiple isoforms, if genes appear to be monocistronic or polycistronic, what the content of a 5′ or 3′ UTR is for a transcript, and where sites of RNA processing may occur.

Ribosome profiling

One way to measure the amount of different proteins within a cell at any given time is through mass spectrometry, but understanding the production rate of a protein requires a different method. Ribosome profiling is a technique used to measure translation comprehensively and quantitatively by deeply sequencing ribosome-protected mRNA fragments. In order to accomplish this, the RNA harvested from cells is extracted and subjected to an RNase that degrades any RNA not protected by a ribosome (Ingolia, 2016). From the resulting monosomes, the protected RNA is extracted and sequenced, revealing the exact location that was being translated at the time of cell harvest.

Understanding translation across the genome is important when inquiring into subjects such as the translational efficiency (TE) for any given gene. TE can be defined as the rate of protein production per RNA for a gene at a particular moment in time, and can be calculated by dividing the normalized signal from ribosome profiling by the normalized signal from Rend-seq for a specific gene. This measurement can be useful in defining and understanding different modes of translational regulation throughout the genome.

Other techniques

In order to measure the ends of RNAs, some additional techniques that can be used include Northern blots, capped analysis of gene expression (CAGE), and RNA ligase-mediated rapid amplification of cDNA ends (RLM-RACE). Northern blots can be used to assess the presence as well as the size of RNA molecules, which can indirectly lead to identification of the ends of the RNA molecules. Higher resolution may be achieved by CAGE or RLM-RACE. In CAGE, the 5′ ends of capped RNAs are identified in order to gain insight into transcriptional start sites (TSSs). This is achieved through a modified cap structure, which captures the 5’ end of RNA molecules and allows for reverse transcription and sequencing of the captured RNA (Morioka et al., 2020). RLM-RACE can be used to identify either the 5′ or 3′ end of an RNA transcript. For 5′ RACE, an adaptor is ligated to the 5′ end of a transcript, followed by reverse transcription, PCR amplification of the cDNA using gene-specific primers, and sequencing (Adkar-Purushothama et al., 2017). 3′ RACE is similar, with the exception that the RNA needs to be reverse transcribed first, then a poly(A) tail is added to the cDNA molecules before PCR amplification (Adkar-Purushothama et al., 2017).

When it comes to quantifying protein synthesis, assessments of bulk protein production can be made using techniques such as pulse-chase experiments or metabolic labeling with fluorescent amino acids. Pulse-chase experiments utilize a short “pulse” of a labeled amino acid followed by a “chase” with excess unlabeled amino acids to reveal newly synthesized proteins (Hou et al., 2013). Metabolic labeling with fluorescent amino acids adds non-radioactive amino acids that are labeled with fluorescent dyes, which allows for protein synthesis to be visualized and quantified in live cells via microscopy (Tom Dieck et al., 2012). To get a more detailed look at protein synthesis for particular proteins, one techniques that can be used is stable isotope labeling by amino acids in cell culture (SILAC), which is then paired with mass spectrometry to get protein identifications. In SILAC, some population of cells are grown in media containing amino acids with the natural isotope, and another population is grown in media containing stable isotope labeled amino acids (Lu et al., 2022). The cells grown in the stable isotope labeled amino acids will then create proteins containing these labeled amino acids, which can be differentiated from proteins containing natural isotopes using liquid chromatography mass spectrometry (LC-MS). This technique is especially useful in comparing protein production under two different populations of cells in different conditions.

Proportional synthesis

A fundamental question regarding protein synthesis in different organisms is whether they follow the principles of proportional synthesis. If an organism exhibits proportional synthesis, that indicates that in the case of multi-protein complexes, they synthesize proteins in the correct ratio needed for the complex (Li et al., 2014). Under this regime, excess protein production for any particular protein in the complex is prevented, which is beneficial as it avoids the presence of excess subunits that may misfold or aggregate (Taggart et al., 2018). This concept has been shown to hold true across a number of organisms, including both prokaryotes and eukaryotes (Taggart et al., 2018). However, proportional synthesis has only been shown to exist in organisms that are in steady state growth (Taggart et al., 2018).

SUMMARY

Viruses have always been a topic of great interest in biology, offering insight into basic biological tenants. The bacteriophage T4 in particular has served as an excellent model system for understanding molecular biology as a whole. In this project, we utilized recently developed experimental techniques to delve into the biology of T4 phage in a way that had not yet been done. At the start of this project, most of the research into T4 had been conducted using low-throughput experiments, though there was a microarray study conducted in 2002 to observe T4 gene expression for all genes during the course of infection (Luke et al., 2002). In our study, we aimed to carry out a more extensive look at both gene transcription and translation throughout infection using the newer techniques of end-enriched RNA-seq (Rend-seq) and ribosome profiling. During the course of our experiments, another study was released studying T4 gene expression over time using RNA-seq. While a valuable contribution to the study of phage, our dataset offers a more comprehensive look at the lifecycle of T4, since Rend-seq gives all the information of RNA-seq with additional information on transcript ends, and ribosome profiling shows not only which mRNAs exist but also their levels of translation.

In Chapter 2 of this thesis, I describe the results gathered and resulting conclusions made from our experiments. The majority of these conclusions were derived from the Rend-seq and ribosome profiling data. Using Rend-seq, we were able to locate the boundaries of different transcripts created during T4 infection, identify putative novel RegB cleavage sites, and detect new potential promoters. Using ribosome profiling, we had the ability to both identify new instances of translational regulation in genes that had not been previously thought to be regulated at this level, as well as highlight the presence of proportional synthesis for the first time in any virus. Using a combination of the two methods we were not only able to quantify and track changes in transcription and translation levels at many time points across infection, but were also able to detect the presence of new potential genes. One of these genes, referred to as *61.-1*, showed great promise as being T4's 290th gene, as it showed a strong signal in Rend-seq and ribosome profiling as well as contained a clear late promoter, ribosome binding site (RBS), and open reading frame (ORF). Using this information, we were able to conduct additional experiments on *61.-1*, learning that it is toxic to *E. coli* when ectopically expressed, but is not essential for T4 growth under laboratory conditions. In Chapter 3, I discuss the impact of our data on the field of phage research and biology as a whole. I also review potential future directions for following up on this work in order to answer some of the remaining questions around T4's complex lifecycle.

References

Adelman, Karen, Edward N. Brody, and Malcolm Buckle. “Stimulation of Bacteriophage T4 Middle Transcription by the T4 Proteins MotA and AsiA Occurs at Two Distinct Steps in the Transcription Cycle.” *Proceedings of the National Academy of Sciences* 95, no. 26 (December 22, 1998): 15247–52. https://doi.org/10.1073/pnas.95.26.15247.

Adkar-Purushothama, Charith Raj, Pierrick Bru, and Jean-Pierre Perreault. “3′ RNA Ligase Mediated Rapid Amplification of cDNA Ends for Validating Viroid Induced Cleavage at the 3′ Extremity of the Host mRNA.” *Journal of Virological Methods* 250 (December 1, 2017): 29–33. https://doi.org/10.1016/j.jviromet.2017.09.023.

Andrake, M, N Guild, T Hsu, L Gold, C Tuerk, and J Karam. “DNA Polymerase of Bacteriophage T4 Is an Autogenous Translational Repressor.” *Proceedings of the National Academy of Sciences* 85, no. 21 (November 1988): 7942–46. https://doi.org/10.1073/pnas.85.21.7942.

Baker, Lonnie, Charles David, and Donald J Jacobs. “Ab Initio Gene Prediction for Protein-Coding Regions.” *Bioinformatics Advances* 3, no. 1 (January 1, 2023): vbad105. https://doi.org/10.1093/bioadv/vbad105.

Clokie, Martha RJ, Andrew D Millard, Andrey V Letarov, and Shaun Heaphy. “Phages in Nature.” *Bacteriophage* 1, no. 1 (2011): 31–45. https://doi.org/10.4161/bact.1.1.14942.

Crick, F. H. C., Leslie Barnett, S. Brenner, and R. J. Watts-Tobin. “General Nature of the Genetic Code for Proteins.” *Nature* 192, no. 4809 (December 1961): 1227–32. https://doi.org/10.1038/1921227a0.

Dressman, Holly Kloos, and John W. Drake. “Lysis and Lysis Inhibition in Bacteriophage T4: rV Mutations Reside in the Holin t Gene.” *Journal of Bacteriology* 181, no. 14 (July 1999): 4391–96.

Durand, Sylvain, Graziella Richard, François Bontems, and Marc Uzan. “Bacteriophage T4 Polynucleotide Kinase Triggers Degradation of mRNAs.” *Proceedings of the National Academy of Sciences* 109, no. 18 (May 2012): 7073–78. https://doi.org/10.1073/pnas.1119802109.

Ejigu, Girum Fitihamlak, and Jaehee Jung. “Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing.” *Biology* 9, no. 9 (September 18, 2020): 295. https://doi.org/10.3390/biology9090295.

Fremin, Brayon J., Ami S. Bhatt, Nikos C. Kyrpides, Aditi Sengupta, Alexander Sczyrba, Aline Maria da Silva, Alison Buchan, et al. “Thousands of Small, Novel Genes Predicted in Global Phage Genomes.” *Cell Reports* 39, no. 12 (June 21, 2022): 110984. https://doi.org/10.1016/j.celrep.2022.110984.

Gamkrelidze, Mariam, and Krystyna Dąbrowska. “T4 Bacteriophage as a Phage Display Platform.” *Archives of Microbiology* 196, no. 7 (July 1, 2014): 473–79. https://doi.org/10.1007/s00203-014-0989-8.

Geiduschek, E Peter, and George A Kassavetis. “Transcription of the T4 Late Genes.” *Virology Journal* 7 (October 28, 2010): 288. https://doi.org/10.1186/1743-422X-7-288.

Gerovac, Milan, Kotaro Chihara, Laura Wicke, Bettina Böttcher, Rob Lavigne, and Jörg Vogel. “Immediate Targeting of Host Ribosomes by Jumbo Phage Encoded Proteins.” bioRxiv, February 26, 2023. https://doi.org/10.1101/2023.02.26.530069.

Gostimskaya, Irina. “CRISPR–Cas9: A History of Its Discovery and Ethical Considerations of Its Use in Genome Editing.” *Biochemistry. Biokhimiia* 87, no. 8 (2022): 777–88. https://doi.org/10.1134/S0006297922080090.

Govindarajan, Rajeshwar, Jeyapradha Duraiyan, Karunakaran Kaliyappan, and Murugesan Palanisamy. “Microarray and Its Applications.” *Journal of Pharmacy & Bioallied Sciences* 4, no. Suppl 2 (August 2012): S310–12. https://doi.org/10.4103/0975-7406.100283.

Hatfull, Graham F. “Phage Therapy for Nontuberculous Mycobacteria: Challenges and Opportunities.” *Pulmonary Therapy* 9, no. 1 (December 30, 2022): 91–107. https://doi.org/10.1007/s41030-022-00210-y.

Herbst, K L, L M Nichols, R F Gesteland, and R B Weiss. “A Mutation in Ribosomal Protein L9 Affects Ribosomal Hopping during Translation of Gene 60 from Bacteriophage T4.” *Proceedings of the National Academy of Sciences of the United States of America* 91, no. 26 (December 20, 1994): 12525–29.

Hershey, A. D., and Martha Chase. “INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE.” *The Journal of General Physiology* 36, no. 1 (September 20, 1952): 39–56.

Hinton, Deborah M. “Transcriptional Control in the Prereplicative Phase of T4 Development.” *Virology Journal* 7, no. 1 (October 28, 2010): 289. https://doi.org/10.1186/1743-422X-7-289.

Hou, Tieying, Cornelia H Rinderknecht, Andreas V Hadjinicolaou, Robert Busch, and Elizabeth Mellins. “Pulse-Chase Analysis for Studies of MHC Class II Biosynthesis, Maturation, and Peptide Loading.” *Methods in Molecular Biology (Clifton, N.J.)* 960 (2013): 411–32. https://doi.org/10.1007/978-1-62703-218-6_31.

Hsu, Wen-Tah, and Samuel B. Weiss. “SELECTIVE TRANSLATION OF T4 TEMPLATE RNA BY RIBOSOMES FROM T4-INFECTED Escherichia Coli.” *Proceedings of the National Academy of Sciences* 64, no. 1 (September 1969): 345–51. https://doi.org/10.1073/pnas.64.1.345.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11, no. 1 (March 8, 2010): 119. https://doi.org/10.1186/1471-2105-11-119.

Hyman, Paul, and Mark van Raaij. “Bacteriophage T4 Long Tail Fiber Domains.” *Biophysical Reviews* 10, no. 2 (December 4, 2017): 463–71. https://doi.org/10.1007/s12551-017-0348-5.

Ingolia, Nicholas T. “Ribosome Footprint Profiling of Translation throughout the Genome.” *Cell* 165, no. 1 (March 24, 2016): 22–33. https://doi.org/10.1016/j.cell.2016.02.066.

Kasman, Laura M., and La Donna Porter. “Bacteriophages.” In *StatPearls [Internet]*. StatPearls Publishing, 2022. https://www.ncbi.nlm.nih.gov/books/NBK493185/.

Kennell, David. “Inhibition of Host Protein Synthesis During Infection of Escherichia Coli by Bacteriophage T4.” *Journal of Virology* 6, no. 2 (August 1970): 208–17.

Kutter, Elizabeth, Daniel Bryan, Georgia Ray, Erin Brewster, Bob Blasdel, and Burton Guttman. “From Host to Phage Metabolism: Hot Tales of Phage T4’s Takeover of E. Coli.” *Viruses* 10, no. 7 (July 21, 2018): 387. https://doi.org/10.3390/v10070387.

Leiman, Petr G., Fumio Arisaka, Mark J. van Raaij, Victor A. Kostyuchenko, Anastasia A. Aksyuk, Shuji Kanamaru, and Michael G. Rossmann. “Morphogenesis of the T4 Tail and Tail Fibers.” *Virology Journal* 7, no. 1 (December 3, 2010): 355. https://doi.org/10.1186/1743-422X-7-355.

Li, Gene-Wei, David Burkhardt, Carol Gross, and Jonathan S. Weissman. “Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources.” *Cell* 157, no. 3 (April 24, 2014): 624–35. https://doi.org/10.1016/j.cell.2014.02.033.

Lopatina, Anna, Nitzan Tal, and Rotem Sorek. “Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy.” *Annual Review of Virology* 7, no. 1 (September 29, 2020): 371–84. https://doi.org/10.1146/annurev-virology-011620-040628.

Lu, Kun, Yun-Chung Hsiao, Chih-Wei Liu, Rita Schoeny, Robinan Gentry, and Thomas B. Starr. "A Review of Stable Isotope Labeling and Mass Spectrometry Methods to Distinguish Exogenous from Endogenous DNA Adducts and Improve Dose–Response Assessments." *Chemical Research in Toxicology* 35, no. 1 (January 17, 2022): 7–29. https://doi.org/10.1021/acs.chemrestox.1c00212.

Luke, Kimberly, Agnes Radek, XiuPing Liu, John Campbell, Marc Uzan, Robert Haselkorn, and Yakov Kogan. "Microarray Analysis of Gene Expression during Bacteriophage T4 Infection." *Virology* 299, no. 2 (August 1, 2002): 182–91. https://doi.org/10.1006/viro.2002.1409.

Luria, S. E., and Mary L. Human. "A NONHEREDITARY, HOST-INDUCED VARIATION OF BACTERIAL VIRUSES1." *Journal of Bacteriology* 64, no. 4 (October 1952): 557–69.

Luria, S E, and M Delbrück. "MUTATIONS OF BACTERIA FROM VIRUS SENSITIVITY TO VIRUS RESISTANCE." *Genetics* 28, no. 6 (November 20, 1943): 491–511. https://doi.org/10.1093/genetics/28.6.491.

Maghsoodi, Ameneh, Anupam Chatterjee, Ioan Andricioaei, and Noel C. Perkins. "How the Phage T4 Injection Machinery Works Including Energetics, Forces, and Dynamic Pathway." *Proceedings of the National Academy of Sciences* 116, no. 50 (December 10, 2019): 25097–105. https://doi.org/10.1073/pnas.1909298116.

Miller, Eric S., Elizabeth Kutter, Gisela Mosig, Fumio Arisaka, Takashi Kunisawa, and Wolfgang Rüger. "Bacteriophage T4 Genome." *Microbiology and Molecular Biology Reviews* 67, no. 1 (March 2003): 86–156. https://doi.org/10.1128/mmbr.67.1.86-156.2003.

Morioka, Masaki Suimye, Hideya Kawaji, Hiromi Nishiyori-Sueki, Mitsuyoshi Murata, Miki Kojima-Ishiyama, Piero Carninci, and Masayoshi Itoh. "Cap Analysis of Gene Expression (CAGE): A Quantitative and Genome-Wide Assay of Transcription Start Sites." *Methods in Molecular Biology (Clifton, N.J.)* 2120 (2020): 277–301. https://doi.org/10.1007/978-1-0716-0327-7_20.

Nechaev, Sergei, and E. Peter Geiduschek. "Dissection of the Bacteriophage T4 Late Promoter Complex." *Journal of Molecular Biology* 379, no. 3 (June 6, 2008): 402–13. https://doi.org/10.1016/j.jmb.2008.03.071.

Nolan, James M, Vasiliy Petrov, Claire Bertrand, Henry M Krisch, and Jim D Karam. "Genetic Diversity among Five T4-like Bacteriophages." *Virology Journal* 3 (May 23, 2006): 30. https://doi.org/10.1186/1743-422X-3-30.

Piešiniene, Lina, Lidija Truncaite, Aurelija Zajančkauskaite, and Rimas Nivinskas. "The Sequences and Activities of RegB Endoribonucleases of T4-Related Bacteriophages." *Nucleic Acids Research* 32, no. 18 (2004): 5582–95. https://doi.org/10.1093/nar/gkh892.

Rao, Venigalla B., Andrei Fokine, Qianglin Fang, and Qianqian Shao. "Bacteriophage T4 Head: Structure, Assembly, and Genome Packaging." *Viruses* 15, no. 2 (February 2023): 527. https://doi.org/10.3390/v15020527.

Rodic, Andjela, Bojana Blagojevic, Evgeny Zdobnov, Magdalena Djordjevic, and Marko Djordjevic. "Understanding Key Features of Bacterial Restriction-Modification Systems through Quantitative Modeling." *BMC Systems Biology* 11, no. Suppl 1 (February 24, 2017): 1–15. https://doi.org/10.1186/s12918-016-0377-x.

Schooley, Robert T., Biswajit Biswas, Jason J. Gill, Adriana Hernandez-Morales, Jacob Lancaster, Lauren Lessor, Jeremy J. Barr, et al. "Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter Baumannii Infection." *Antimicrobial Agents and Chemotherapy* 61, no. 10 (September 22, 2017): 10.1128/aac.00954-17. https://doi.org/10.1128/aac.00954-17.

Shivachandra, Sathish B., Qin Li, Kristina K. Peachman, Gary R. Matyas, Stephen H. Leppla, Carl R. Alving, Mangala Rao, and Venigalla B. Rao. "Multicomponent Anthrax Toxin Display and Delivery Using Bacteriophage T4." *Vaccine* 25, no. 7 (January 26, 2007): 1225–35. https://doi.org/10.1016/j.vaccine.2006.10.010.

Shoemaker, D. D., E. E. Schadt, C. D. Armour, Y. D. He, P. Garrett-Engele, P. D. McDonagh, P. M. Loerch, et al. "Experimental Annotation of the Human Genome Using Microarray Technology." *Nature* 409, no. 6822 (February 2001): 922–27. https://doi.org/10.1038/35057141.

Taggart, James C., and Gene-Wei Li. "Production of Protein-Complex Components Is Stoichiometric and Lacks General Feedback Regulation in Eukaryotes." *Cell Systems* 7, no. 6 (December 26, 2018): 580-589.e4. https://doi.org/10.1016/j.cels.2018.11.003.

Tom Dieck, Susanne, Anke Müller, Anne Nehring, Flora I. Hinz, Ina Bartnik, Erin M. Schuman, and Daniela C. Dieterich. "Metabolic Labeling with Noncanonical Amino Acids and Visualization by Chemoselective Fluorescent Tagging." *Current Protocols in Cell Biology* Chapter 7 (September 2012): 7.11.1-7.11.29. https://doi.org/10.1002/0471143030.cb0711s56.

Wang, Zhong, Mark Gerstein, and Michael Snyder. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews Genetics* 10, no. 1 (January 2009): 57–63. https://doi.org/10.1038/nrg2484.

Winter, R B, L Morrissey, P Gauss, L Gold, T Hsu, and J Karam. "Bacteriophage T4 regA Protein Binds to mRNAs and Prevents Translation Initiation." *Proceedings of the National Academy of Sciences* 84, no. 22 (November 1987): 7822–26. https://doi.org/10.1073/pnas.84.22.7822.

Yap, Moh Lan, and Michael G Rossmann. "Structure and Function of Bacteriophage T4." *Future Microbiology* 9 (October 2014): 1319–27. https://doi.org/10.2217/fmb.14.91.

Chapter 2: Temporal gene expression and regulation in T4 phage

Abstract

As the most abundant biological entity in the biosphere, bacteriophages play a critical role in shaping microbial diversity, and thus overall ecosystem health. They are also essential tools in molecular biology, shedding light on fundamental biological concepts. T4 phage, with its complex lifecycle and genetic content, has been instrumental in many such discoveries. However, many questions regarding gene regulation in T4 phage remain unanswered. In this study, we employ end-enriched RNA-seq (Rend-seq) and ribosome profiling to examine T4 RNA and protein synthesis throughout the course of infection, gaining new insights at the transcriptional, translational, and genomic level. At the transcriptional level, we identified transcript boundaries, novel putative promoters, and new potential cleavage sites for the T4 endoribonuclease RegB. At the translational level, we identified many instances of previously unreported changes in translational efficiency over the course of infection, indicating the presence of intricate and uncharacterized mechanisms of regulation. Collectively, transcriptional and translational controls lead to precisely tuned protein synthesis rates during infection, as exemplified by the phenomenon that components of T4 protein complexes are synthesized according to their stoichiometry— a principle that has been observed in organisms during steady-state growth. Finally, we identified and experimentally validated T4's 290th gene, *61.-1*. Though non-essential to T4 in laboratory conditions, this gene has homologs present in a number of other phage and drastically impacts *E. coli* growth when ectopically expressed. This study provides insights into T4 phage biology, paving the way for further exploration into molecular biology, virology, and biotechnology; our rich data set can be utilized by future studies to answer a diverse array of inquiries.

Introduction

Bacteriophages are viruses that infect bacteria and have been instrumental tools in uncovering many mechanisms underlying molecular biology. T4 phage, in particular, has significantly contributed to our understanding of biological concepts, such as DNA as the genetic material (Kutter et al., 2018), the function of mRNAs (Cobb, 2015), and the triplet code (Crick et al., 1961). T4 phage encodes many proteins that are structurally and functionally similar to proteins within other organisms across the tree of life, making it a prime model organism for laboratory studies (Kutter et al., 2018; Mosig et al., 2006).

T4 exhibits a remarkably complex lifecycle, beginning with irreversible adsorption to a host *E. coli* cell using the phage's tail fibers (Maghsoodi et al., 2019). Upon attachment, T4 punctures the host with a needle-like apparatus to deliver the phage DNA to the inner membrane, where it is pulled into the cytoplasm using the membrane's electrochemical potential (Yap et al., 2014). T4 then initiates the process of replication. Similar to many other phages, T4 exhibits time-dependent gene expression, and its genes are categorized into four groups based on their expression timing: early, delayed early, middle, and late (Liu et al., 2013; Luke et al. 2002). T4 relies on the host RNA polymerase throughout infection to transcribe its genes; in early infection the polymerase is not altered by the phage, but as time progresses there are different modifications to the polymerase to modulate its processivity and allow sequential recognition of different classes of promoters (Mosig et al., 2006). This progression through different stages of gene expression is also aided by other proteins: The T4-encoded endoribonuclease RegB will cleave many early genes at their ribosome binding sites to inactivate the transcripts, and AsiA and MotA will stimulate the transcription of middle genes as the phage moves through its lifecycle (Mosig et al., 2006). Once transcribed, genes are translated into proteins by host ribosomes. Some T4 proteins are believed to modify these ribosomes to prevent the translation of any host genes, since T4 ribosome binding sites are very similar to those of *E. coli*, although these mechanisms have yet to be fully elucidated (Kutter et al., 2018). Translational repressors, such as Gp32, Gp43, and RegA, also modulate the translation of certain T4 genes by occluding ribosomes from the target RNA (Russel et al., 1976; Andrake et al. 1988). Upon completing its lifecycle and accumulating new phages within the host cell, T4 phage causes cell lysis using two proteins, Gpt and Gpe. Gpt creates a hole in the inner membrane to allow Gpe, the endolysin, to access and degrade the peptidoglycan layer. Subsequently, the new phages are released into the environment, and the cycle repeats.

Though there is a wealth of information available in the literature about T4 phage, many questions still remain unanswered, a number of these relating to gene expression in T4. In order to address these questions, there have been efforts to better characterize patterns of gene expression over time during infection; microarrays were the first technique used as a way to follow each gene's expression pattern in a high-throughput manner (Luke et al., 2002), followed more recently by RNA-seq (Wolfram-Schauerte et al., 2022). To gain a deeper understanding of the T4 proteome, 2D-PAGE gels were initially used (Cowan et al., 1994), while liquid chromatography-mass spectrometry (LC-MS) has been utilized in recent years (Wolfram-Schauerte et al., 2022). While these improved techniques have yielded deeper understandings of the inner workings of T4, they remain limited for several reasons. RNA-seq provides more a comprehensive view of gene expression versus microarrays, but the technique is unable to elucidate many aspects of gene transcription such as transcript boundaries. The information gained through LC-MS is often biased due to a priori assumptions of previously known proteins (similar to how microarray probes are designed for previously-identified specific target genes).

There remains a need for studies leveraging techniques that result in less biased and more comprehensive examination of all gene transcription, gene translation, and protein production throughout the time course of T4 infection.

Here we provide new insights into T4 transcription and translation over the course of infection by using the recently developed techniques of end-enriched RNA-seq (Rend-seq) and ribosome profiling (Lalanne et al., 2018; Ingolia et al., 2016). Rend-seq allows for the identification of distinct molecules of RNA by highlighting 5′ and 3′ transcript ends while also showing expression levels across the gene body (Lalanne et al., 2018). The first step in achieving this is the random and sparse cleavage of isolated RNAs. These fragments have a higher probability of a 5′ and 3′ end on one side because this scenario requires only one cut, whereas having a fragment with two ends within the gene body requires two separate cuts. Ribosome profiling, on the other hand, allows for identification of the exact regions being translated in the cells at any particular moment, as well as quantification of the rate of protein synthesis in those regions (Ingolia et al., 2019; Johnson et al., 2018). After rapid harvesting of the cells, they are flash frozen to preserve the location of the ribosomes on their transcripts, and then thawed in chloramphenicol for the same reason. The mRNA that is not protected by the ribosomes is then degraded with the use of an RNase, and the mRNA that is within the ribosomes is extracted, ligated to a DNA adaptor, reverse transcribed, and amplified via PCR before being sequenced. These reads can then be mapped to the organism of interest’s genome to observe which genes were being actively transcribed immediately preceding cell harvest.

Our data offers a host of new information about T4 transcription, translation, and its genome. In terms of transcription, we were able to provide the most comprehensive database to date summarizing aspects of T4 transcription. Specifically, by analyzing the 5′ and 3′ Rend-seq read count peaks as well as the ribosome profiling density, we were able to identify start and end sites of many transcripts, and in doing so have identified sets of polycistronic genes. Similarly, our analysis enabled us to expand upon prior work by identifying probable new promoters as well as RegB cleavage sites, which were previously identified incrementally across multiple different experiments (Uzan, 2001).

With regards to translation, we were able to examine T4's genes and their respective changes in translational efficiency over time, illustrating that the number of genes of exhibiting time-dependent translational regulation in T4 is far more extensive than what has been previously shown. Beyond the presence of additional translation regulation, we showed the existence, for the first time in any phage and more broadly viruses, of proportional synthesis – a specific form of regulation in which proteins in a complex are produced in the precise ratio needed for that complex. This has only previously been documented in cells undergoing steady-state growth, something that does not exist in phage infection (Lalanne et al., 2018).

Finally, while the current understanding of T4 genome assumes the presence of 289 protein coding genes (identified over decades of experimental and computational analysis), we have identified, investigated, and experimentally validated the presence of the 290th gene of T4, referred to as "*61.-1*," which we have shown is present within the genomes of many other phages. Taken together, this work thus provides a more comprehensive understanding of several fundamental aspects of T4, spanning new mechanisms to new genes. Updated T4 annotations, combined with new experimental datasets, set the foundation for continued exploration of T4 biology.

Results

T4 exhibits complex, time-dependent gene expression over the course of infection

To evaluate the time-dependent expression of genes in T4, we carried out both end-enriched RNA-seq (Rend-seq) and ribosome profiling. We isolated RNA for these analyses by setting up in parallel five flasks of *E. coli* growing in M9 medium (containing the carbon sources of glycerol and glucose) at 30˚C and waited for the cultures to enter exponential growth with an OD600 of ~0.3. We then infected each of these flasks simultaneously with T4 at a multiplicity of infection (MOI) of 5 and waited 2, 5, 10, 15, or 20 minutes post-infection to capture expression throughout the T4 lifecycle; previous work showed that expression of the latest genes occurs by 20 minutes post-infection (Luke et al., 2002). We harvested the cells from each time point using rapid filtration and then preserved them in their current state by flash freezing them in liquid nitrogen (Johnson et al., 2018). This ensures that we capture a snapshot of the exact transcription and translation occurring at the moment of interest (Fig 1A). We then separated each sample, using one portion for Rend-seq and another for ribosome profiling to capture the progression of both transcription and translation throughout the course of T4 infection for genes in the T4 genome.

It has already been demonstrated that T4 gene expression is time-dependent, with different genes being maximally expressed at different stages of infection. Given our Rend-seq data, we were similarly able to calculate the time of peak expression for each gene, and then compared and contrasted our results with prior studies. To determine the time of peak expression for a gene, we first converted the raw read counts into RPKM (reads per kilobase per million mapped reads) values. We then determined the relative expression level for each time point by calculating the percent of total gene expression occurring within a given time point (e.g., the calculated RPKM for the gene within that time point, described above) given the total summation of that gene’s expression across all time points (e.g., the total of all RPKMs across all time points for that gene). This normalized the results for each gene to that gene’s total expression level, allowing different genes with different total expression levels to be compared. Lastly, we assigned genes to the categories of early, delayed early, middle, and late (categories that have been previously established in the field (Luke et al., 2002)) based upon their time of peak expression (e.g., the time point with the highest percentage of gene expression, described above); previous studies have categorized genes by their time of first appearance, but the resolution of our data is such that many genes show levels of expression at all time points, making it more practical for us to categorize genes by time of peak expression. An identical normalization and analysis was performed for the ribosome profiling results. Although the time point of peak expression for transcription was often the same time point as peak translation, there were several exceptions where this was not the case, illustrating complex translational regulation, something that will be discussed in a later section of this paper (Fig 1B).

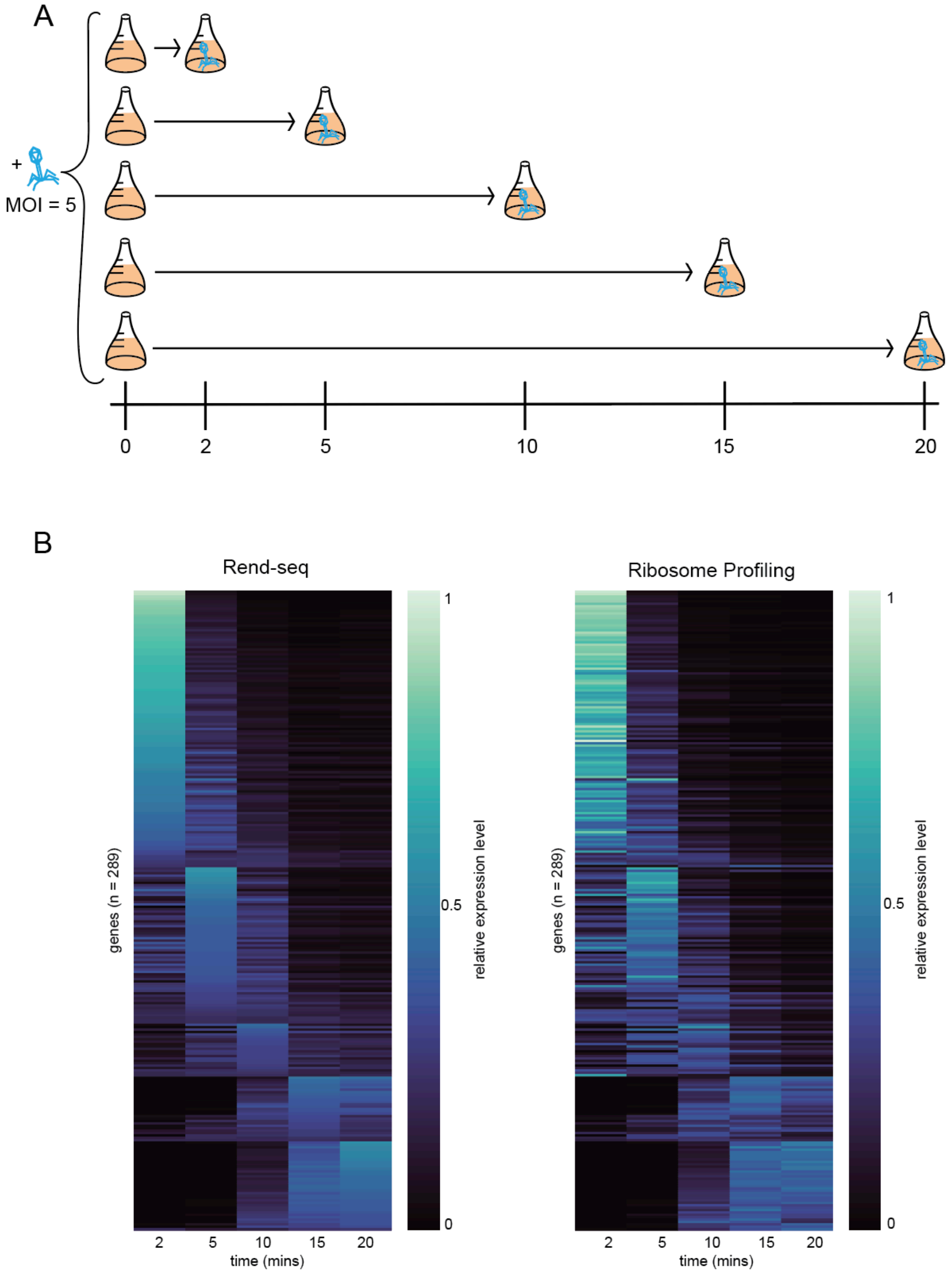


Figure 1

(A) Schematic overview of experimental layout. T4 phage was added to flasks of E. coli in mid-log phage (OD600 0.3) and then the infected cells were harvested at different time points post-infection (2, 5, 10, 15, or 20 minutes). The cells were then processed; part were used for Rend-seq and part were used for ribosome profiling.

(B) Heatmaps illustrating the time-dependency of gene transcription (left) and translation (right) throughout infection. Relative expression levels are determined by taking the sum of all RPKMs for a gene across all time points, and then dividing the RPKMs at a given timepoint by that sum. Both heatmaps are laid out in the same order (that of Rend-seq expression levels for each gene) grouping genes by time of peak expression, and within that by highest expression levels at that time.

We compared our assigned gene time categorization of peak expression with the categorization in prior work, specifically the RNA-seq data from Wolfram-Schauerte et al. (Wolfram-Schauerte et al., 2022). We grouped all the genes in the T4 genome based on their time of peak expression in our dataset and then carried out a side-by-side comparison with the Wolfram-Schauerte data (Wolfram-Schauerte et al., 2022). Although the time points used in each study were not the same (2, 5, 10, 15, and 20 minutes in this study vs. 1, 4, 7, and 20 minutes in the other study), we were able to compare similarities and differences in the datasets (Fig 2). Many genes peaked in expression at very similar times post-infection in the two data sets (Fig 2A), but for a number of genes the Wolfram-Schauerte data provided a time of maximal expression that was very different from the calculated time of maximal expression from our Rend-seq data.

To further explore discrepancies between these works, we examined the genes that exhibited the most substantial discrepancies between these two works (defined as any work categorizing that gene's time of peak expression as equal to or greater than 10 minutes different from the time of peak expression from any of the other work (Fig 2B). For all 20 of such genes, the Wolfram-Schauerte study characterized them as peaking in expression at 20 minutes post-infection, whereas we categorized them as peaking at either 2 or 5 minutes (Wolfram-Schauerte et al., 2022). One reason for these discrepancies could be that the study from 2022 did not analyze any time points between 7 and 20 minutes post-infection, causing any genes with higher expression at 20 minutes post-infection than at 7 minutes post-infection to be placed in the category of peaking at 20 minutes, although it is entirely possible that these genes could have peaked at any time after 7 minutes (Wolfram-Schauerte et al., 2022).

We examined each of the 20 genes that exhibited large discrepancies in time of peak expression based on the data available. 11 of these genes are not well characterized, and therefore could not be further investigated as to which assignment would be more appropriate based on gene function. For two genes, *55* and *31*, it can be understood why they would be characterized as peaking at 20 minutes. *55* encodes a sigma factor recognizing late promoters, and *31* encodes a co-chaperone for GroEL, which is necessary for the structural assembly of progeny phage (Snyder et al., 2005). *ipIII* also logically should be expressed late in infection, as it encodes a protein that is packaged within the heads of the newly manufactured phage capsids (Kuhn et al., 2022), and while our Rend-seq data characterized it as peaking in expression at 5 minutes post-infection, in reality we see that expression levels in our data remained relatively stable over time for *ipIII* as well as *ipII*, with only a small variation leading to *ipIII* being called as peaking at 5 minutes rather than the 15 minute time point that *ipII* is characterized as peaking at. *rnlA* encodes an RNA ligase, which is used throughout the cycle of infection, and again although it peaked in expression at 5 minutes in our Rend-seq data, it remained relatively stable in level of expression over the course of infection (Snopek et al., 1977). *denA* encodes an endonuclease that restricts dC-containing DNA, which is found in the host *E. coli* genome, and would be present near the beginning of infection. Finally, genes *43*, *46*, *nrdA*, and *nrdG* all play critical parts in the replication of the T4 genome. *nrdA* and *nrdG* are ribonucleotide nucleotide reductase subunits, converting the building blocks for RNA to what is needed for DNA replication, which would need to begin early in infection in order to provide materials for genome replication, as this begins not long after infection begins as well (Hendricks et al., 1997). *43* and *46* are needed for genome replication as well, as they encode DNA polymerase and a recombination protein respectively. Origin-dependent DNA replication begins early in the infection process, and as time goes on recombination-dependent replication takes over, so these genes would need to be expressed earlier in infection to be ready for replication (Belanger et al., 1998). As discussed above, the vast majority of the discrepancies for these genes were associated with the Wolfram-Schauerte study strongly biasing their assigned time of peak expression to the "late" 20 minute time category, although there are some discrepancies that cannot be explained. Future work is needed to resolve the remaining ambiguities documented here.

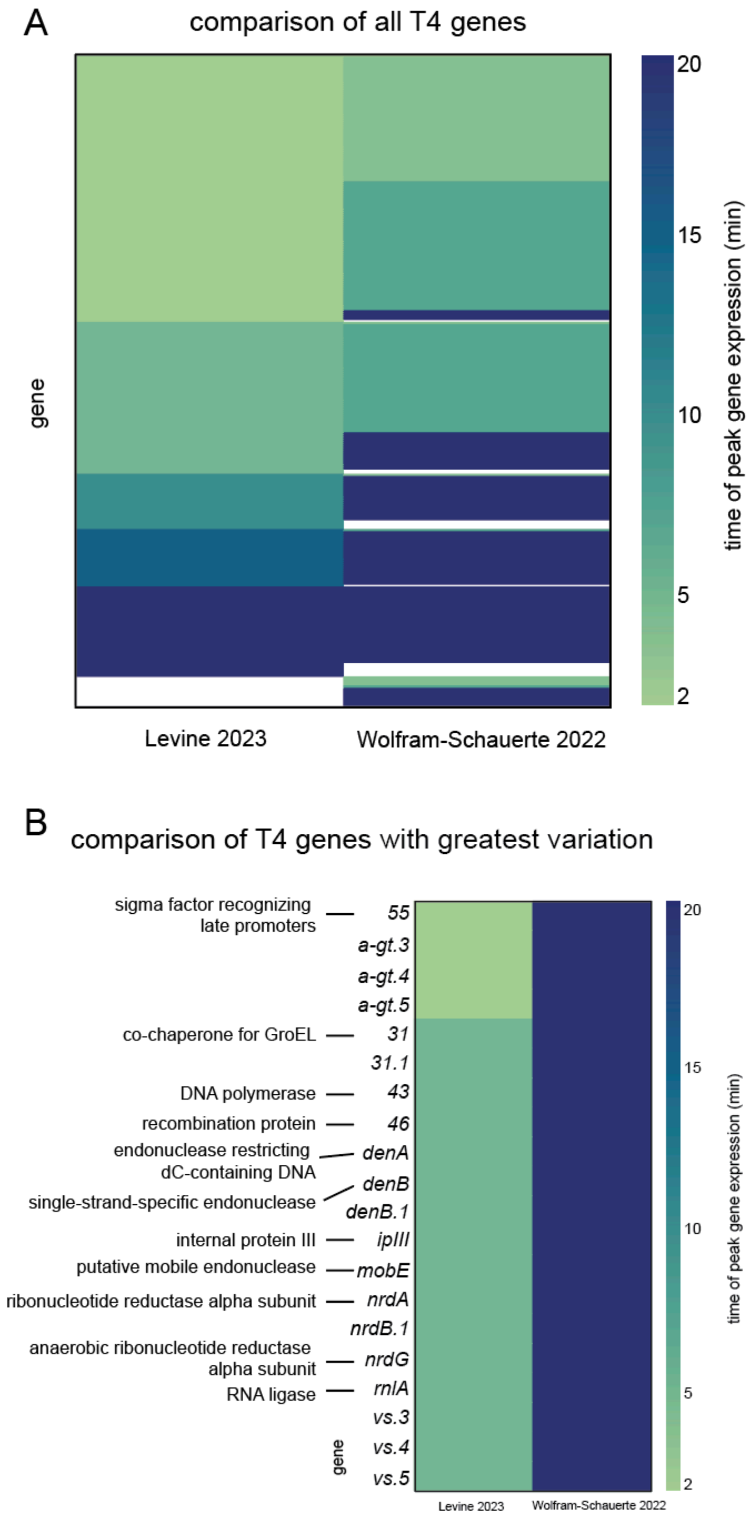

A
comparison of all T4 genes
gene
Levine 2023
Wolfram-Schauerte 2022
20
15
10
5
2
time of peak gene expression (min)
B
comparison of T4 genes with greatest variation
sigma factor recognizing late promoters
co-chaperone for GroEL
DNA polymerase
recombination protein
endonuclease restricting dC-containing DNA
single-strand-specific endonuclease
internal protein III
putative mobile endonuclease
ribonucleotide reductase alpha subunit
anaerobic ribonucleotide reductase alpha subunit
RNA ligase
55
a-gt.3
a-gt.4
a-gt.5
31
31.1
43
46
denA
denB
denB.1
ipIII
mobE
nrdA
nrdB.1
nrdG
rnlA
vs.3
vs.4
vs.5
gene
Levine 2023
Wolfram-Schauerte 2022
20
15
10
5
2
time of peak gene expression (min)

Figure 2

(A) Heatmaps illustrating the classification of peak expression time in this study vs. the recent Rend-seq study by Wolfram-Schauerte (2022). Both columns are ordered based on expression timing of genes in this study.

(B) Heatmap highlighting the genes whose classifications differ the most between the two studies.

Transcript boundaries can be identified using Rend-seq

A major advantage of using Rend-seq over RNA-seq is that we can determine transcript boundaries. The readout from our Rend-seq experiment is a list of read counts at different positions in the genome, corresponding to either the forward or reverse strand, and either a 5′ or a 3′ line. In our visualizations of these data, orange lines correspond to 5′ ends of RNA fragments from our sequencing, and blue lines correspond to 3′ ends. Because the RNA was sparsely fragmented and then size selected between 15-45 base pairs, we see an accumulation of reads at the ends of the transcripts (as fragments this small resulting from sparse fragmentation will statistically be biased towards including a 3′ or 5′ transcript end at one side of the fragment (Lalanne et al., 2018). This creates the signature orange (5′) and blue (3′) peaks that can be seen in our visualizations (Fig 3). In addition to the transcript end peaks, there is some amount of read density throughout the gene body, as well as what we refer to as "shadows" (Lalanne et al., 2018). These shadows are the increased number of 5′ reads inside the gene body directly adjacent to the 3′ peak or vice versa. This occurs because while many of the fragments that do indeed map to the 5′ or 3′ end of the transcript, they will not all be the exact same size, and therefore this will lead to increased, but slightly spread out, density next to the true transcript ends.

To formally determine 5′ transcript start sites and the 3′ transcript end sites, we followed a process in which we first selected the Rend-seq data corresponding to that gene's time point of peak expression (described above). We then defined the search range for the Rend-seq peak, which we describe as any peak identified within 500 base pairs of the end of the gene's coding region (or by the end of the adjacent gene, as a transcription start site located beyond an adjacent gene would likely represent the transcription start site for that adjacent gene). We then used a Rend-seq peak detection software (Parker, in process) which leverages the statistical Z-score to determine statistically significant peaks within the Rend-seq data to search for peaks that fall within the search range. Finally, every identified peak was then manually inspected for acceptance or rejection (rejection criteria is based on whether the peak is located at a potential RegB site, discussed below, whether the peak is within a shadow from a 3′ peak, and whether there is uncertainty between multiple potential peaks within the search region) (Fig 7). Through this method, we were able to identify 106 5′ peaks and 93 3′ peaks. It is important to note that transcription start and end sites identified through Rend-seq, while robust, are still susceptible to errors, such as post-transcriptional editing of the transcripts that results in modification of the primary transcript.

Leveraging these newly-identified transcriptional start and end sites, we identified T4 genes that are monocistronic (Fig 3A). An example of one such gene is *motB*, which had an evident 5′ peak at position 7165, and 3′ peak at position 6580 (Fig 3A). The ORF for *motB* lies inside the identified peaks, with the start at position 7,141 and the end at position 6653. We were also able to identify likely polycistronic messages, indicated by the presence of a single 5' and a single 3' peak flanking multiple ORFs, which prior work indicated in common in T4 (Samuel, 1989) (Fig 3B). One example is the transcript encompassing the genes *24.3*, *24.2*, and *rnlB*, which could be clearly identified as polycistronic because of the clear 5′ peak at position 110,107 and 3′ peak at 108,607 bp in the Rend-seq data (Fig 3B). The ORFs of the three genes fall within this transcript, with the beginning of the first ORF in this transcript, *24.3*, at 109,915 bp, and the end of the last ORF, *rnlB*, lying at 108,636 bp.

Previous work also demonstrated that some genes can be expressed in different mRNA isoforms, sometimes as monocistronic messages and sometimes as polycistronic messages (Young et al., 1981). One example in our data was the gene uvsY.-2, which was polycistronic early in infection but became monocistronic later on, starting at 10 minutes post-infection and rising in expression of the monocistronic RNA up to 20 minutes post-infection (Fig 3C). Upstream of the polycistronic RNA (926 bp) there was a motif resembling a middle promoter at position 115370 in the T4 genome, and upstream of the 5′ peak of the monocistronic form of the RNA (236 bp) at position 114,680 there was the signature of a late promoter. Our full dataset can be further used to investigate the expression of particular genes of interest in the future.

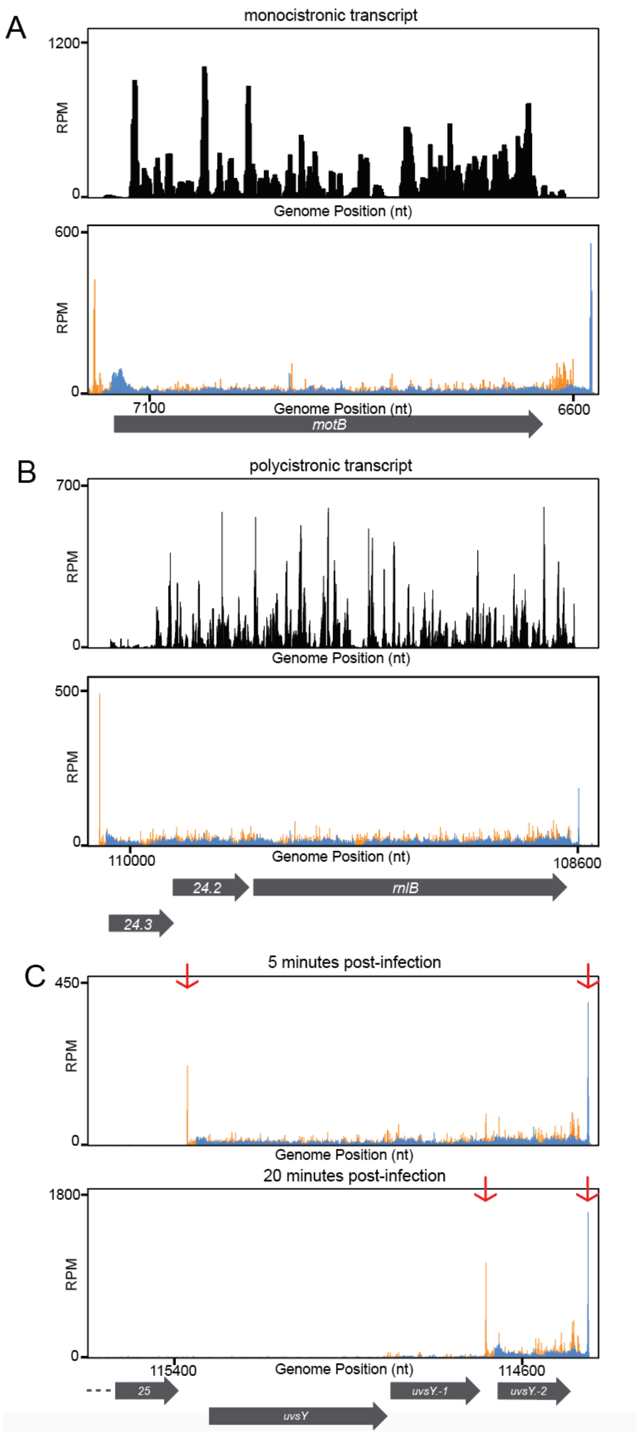


Figure 3

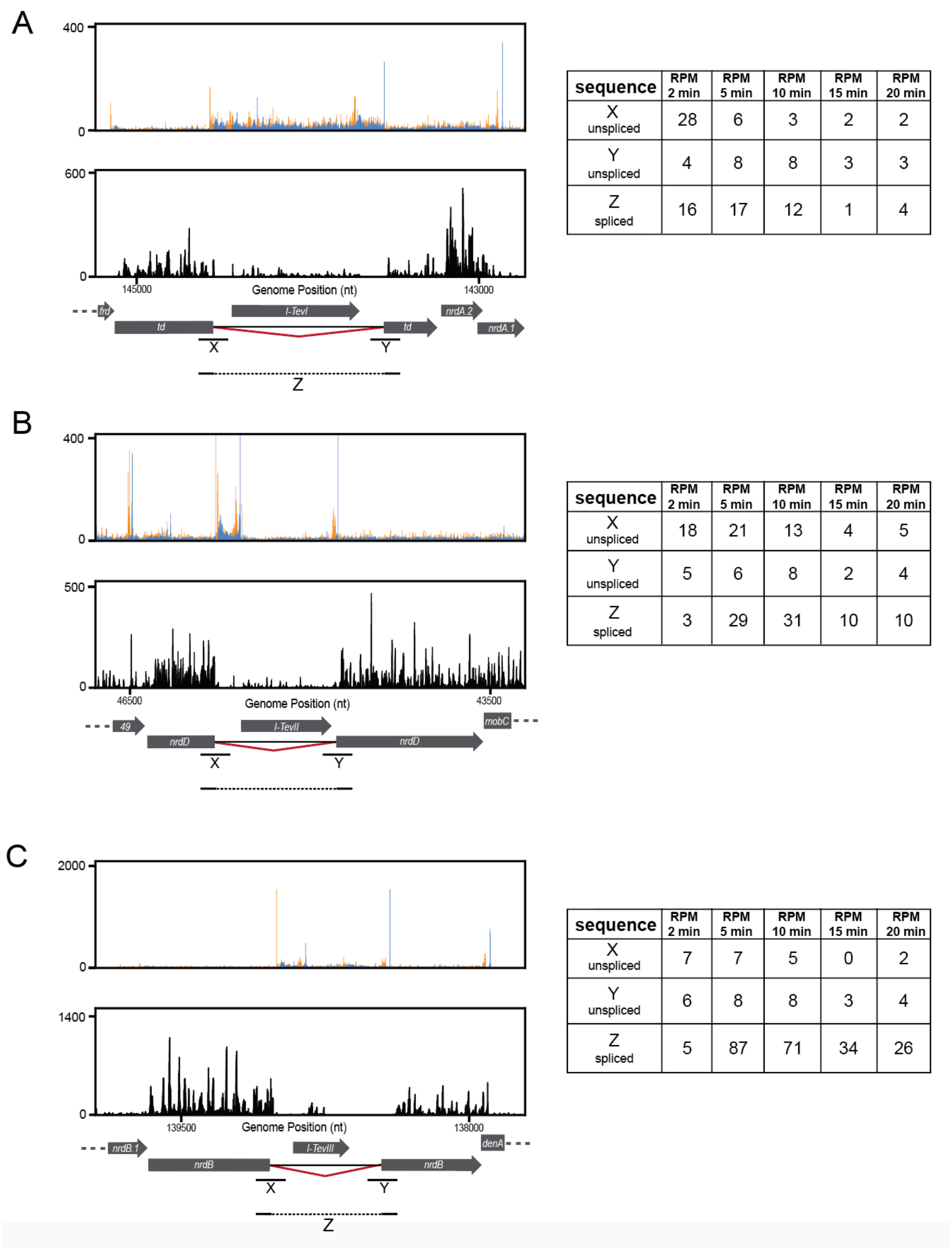
(A) Ribosome profiling (top) and Rend-seq (bottom) illustrations of reads per million (RPM) over a segment of the genome containing a monocistronic transcript, as illustrated by the 5′ peak (orange) and 3′ peak (blue) containing only this one gene.

(B) Similar illustration, but showing an example of a polycistronic transcript.

(C) Similar illustration, but showing how, over time, transcripts can shift from being polycistronic to monocistronic. Red arrows highlight the peaks in order to bring attention to the movement of the 5′ peak.

Transcriptional and translational regulation events are more prevalent in T4 than previously recognized

It has been well documented that there are three genes in T4 (*td*, *nrdD*, and *nrdB*) that are each spliced, joining two exons and leaving an intron that contains a homing endonuclease that can be independently translated (Sandegren et al., 2007). We searched our dataset for evidence of this process in these three known spliced transcripts, looking for sequences that included the 15 nucleotides at the 3′ end of the first exon and the 15 nucleotides at the 5′ end of the second exon joined together, removing the intervening sequence of the intron. Though there were few reads at the start of infection, we did detect an increase in the presence of these reads reporting on spliced transcripts after 2 minutes post-infection, with peak transcript levels occurring 5 minutes post-infection (Fig 4).



sequence	RPM 2 min	RPM 5 min	RPM 10 min	RPM 15 min	RPM 20 min
X unspliced	28	6	3	2	2
Y unspliced	4	8	8	3	3
Z spliced	16	17	12	1	4

sequence	RPM 2 min	RPM 5 min	RPM 10 min	RPM 15 min	RPM 20 min
X unspliced	18	21	13	4	5
Y unspliced	5	6	8	2	4
Z spliced	3	29	31	10	10

sequence	RPM 2 min	RPM 5 min	RPM 10 min	RPM 15 min	RPM 20 min
X unspliced	7	7	5	0	2
Y unspliced	6	8	8	3	4
Z spliced	5	87	71	34	26

Figure 4

(A) Rend-seq (top) and ribosome profiling (bottom) illustrations of the splicing of the T4 gene *td*. X and Y correspond to RPMs going straight through the splicing junction (i.e. no splicing has occurred). Z corresponds to RPMs bridging the outside of the two splicing junctions (i.e. splicing has occurred). The sequences of numbers represent the changing RPMs over time for each section.

(B) Similar illustration, but for the T4 gene *nrdD*.

(C) Similar illustration, but for the T4 gene *nrdB*.

We also used our dataset to identify the locations of existing and new potential promoters. Consensus sequences exist for each class of promoter in T4 (early, middle, and late) based on previously identified promoters, which were mapped by primer extension of mRNA extracted from T4-infected cells, promoter-cloning vectors, or using information content software (Miller et al., 2003). These approaches identified 121 promoters total: 40 early promoter, 32 middle promoters, and 49 late promoters (Miller et al., 2003). Our approach differed from previous methods as we had access to 5′ transcript start sites through Rend-seq, allowing us to search the genome in the correct locations upstream from the individual transcripts to identify new, putative promoters. Specifically, we searched the genome for the T4 promoter consensus sequences located upstream from the 5′ transcript starts sites identified through Rend-Seq. In our analysis, we only searched for the early and middle promoter sequences, as these promoters have two regions of consensus sequences that are a specific number of base pairs apart; to predict a new promoter we required that both regions match. Late promoters have only a single consensus region which is largely composed of As and Ts, which the T4 genome is already biased towards containing, making it difficult to predict late promoters (Miller et al., 2003). Although we did not identify any new suspected early promoters, we identified nine putative middle promoters that had not yet been annotated. The published consensus sequence for middle promoters contains a GCTT motif, with the G at position -30, and a motif containing TATAAT, with the last T being at position -12 (Fig 5A) (Miller et al., 2003). While each putative promoter did not capture these motifs perfectly, as would be expected when looking for a motif from a sequence logo, every one of the sequences highlighted shared many bases with the known motifs.

Of the potential promoters identified, we further split them into three classes. (Fig 5B). The first class were those in areas where there were no existing promoters identified, suggesting an entirely new site of regulation. The second category was promoters in areas where there were already promoters identified, and the new promoter and known promoter were of the same time class (e.g., a single transcript containing two promoters of the same timing at different locations upstream from its transcription start site). The third category was promoters in areas with a known promoters, but of a different time class as the new suspected promoter (e.g. one was an early promoter and the other was a middle promoter). These cases could indicate that there is separate regulation occurring for these transcripts at different time points during infection.

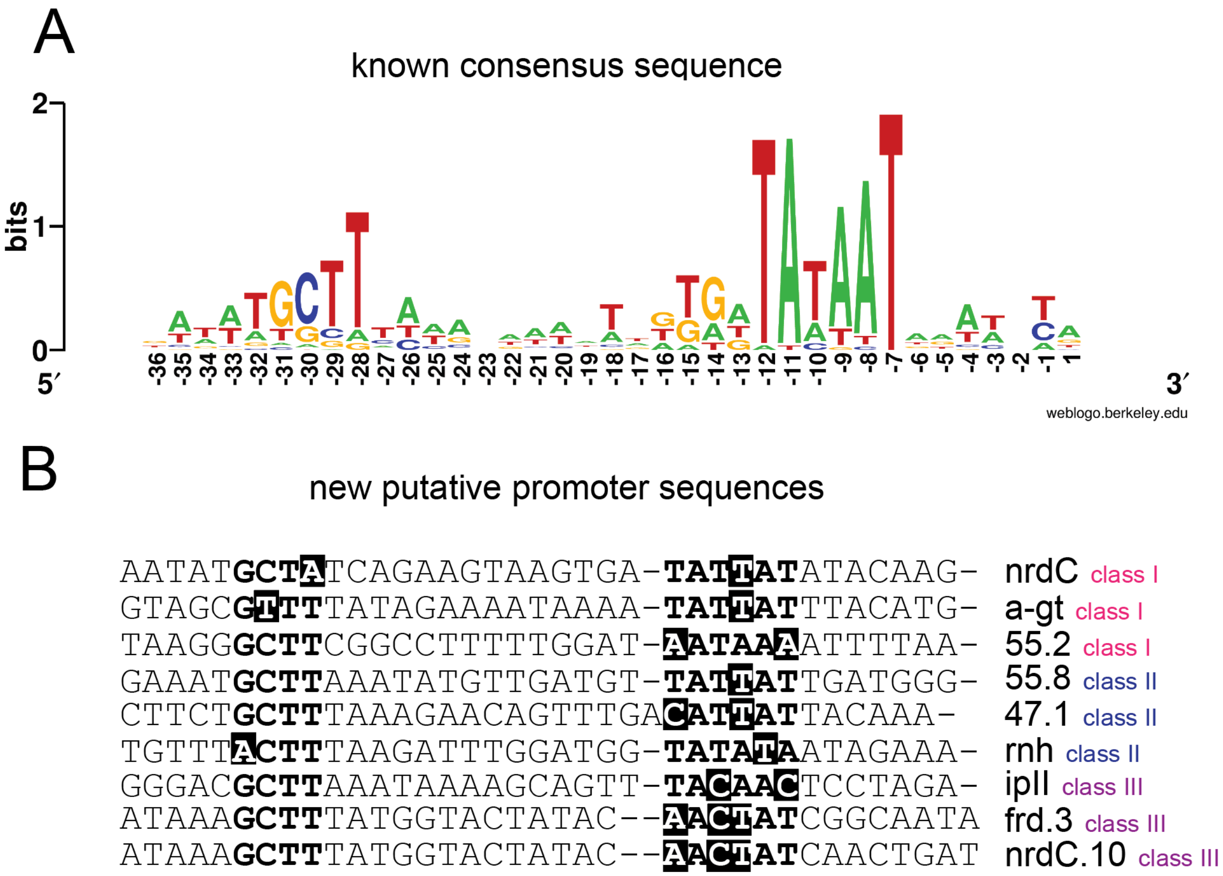

A
known consensus sequence
bits
2
1
0
5′
-36 -35 -34 -33 -32 -31 -30 -29 -28 -27 -26 -25 -24 -23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1
3′
weblogo.berkeley.edu
B
new putative promoter sequences
AATATGCTATCAGAAGTAAGTGA-TATTATATACAAG- nrdC class I
GTAGCGTTTTATAGAAAATAAAA-TATTATTTACATG- a-gt class I
TAAGGGCTTCGGCCTTTTTGGAT-AATAAAATTTTAA- 55.2 class I
GAAATGCTTAAATATGTTGATGT-TATTATTGATGGG- 55.8 class II
CTTCTGCTTTAAAGAACAGTTTGACATTATTACAAA- 47.1 class II
TGTTTACTTTAAGATTTGGATGG-TATATAAATAGAAA- rnh class II
GGGACGCTTAAATAAAAGCAGTT-TACAACTCCTAGA- ipII class III
ATAAAGCTTTATGGTACTATAC--AACTATCGGCAATA frd.3 class III
ATAAAGCTTTATGGTACTATAC--AACTATCAACTGAT nrdC.10 class III

Figure 5

(A) Consensus sequence compiled from all known middle promoters listed in Miller 2003, showing several distinct motifs.

(B) Individual sequences from all newly discovered putative middle promoters in this study, with the distinct motifs from above bolded and any variations in these motifs highlighted. Class numbers are assigned to denote whether the putative promoter is the only promoter assigned to the gene (Class I), an additional putative promoter associated with a gene that has an additional middle promoter (Class II), or an additional putative promoter associated with a gene that has a different type of promoter (Class III).

We also searched our dataset for cleavages driven by the endoribonuclease RegB, which cleaves mRNAs during infection to inactivate production of early gene products while sparing middle and late mRNAs (Piešiniene et al., 2004). To identify potential RegB cleavage sites, we first identified all cases of the RegB recognition sequence, GGAG, in the genome, finding 603 sites: 312 on the reverse strand and 291 on the forward strand. For each recognition sequence identified, we then determined if any 5′ Rend-seq peaks coincided with that region (Parker, in process). Lastly, we confirmed that the corresponding gene containing the RegB motif exhibits early peak expression. All three requirements needed to be met to select that site as a potential RegB cleavage site (Fig 7).

Through this analysis, we identified 29 new suspected RegB sites in addition to the 22 RegB sites that had previously been reported (Sanson et al., 1995). Both the known RegB cleavage sites and the new putative, new RegB sites demonstrated a clear 5′ peaks within the GGAG sequence, indicating that these peaks likely arise from transcript cleavage (Fig 6A). We did not see a corresponding 3′ peak next to the 5′ peaks in these regions, as RegB cleavage likely allows for accelerated mRNA decay by creating an entry site for 3′-5′ exoribonucleases (Durand et al., 2012). RegB sites are found in different regions of RNAs: intergenic regions, within the ribosome binding site (RBS), or within the coding regions (CDs). Of the new putative RegB sites that we uncovered, 3 were in intergenic regions, 14 in RBSs, and 18 in CDSs (Fig 6B).

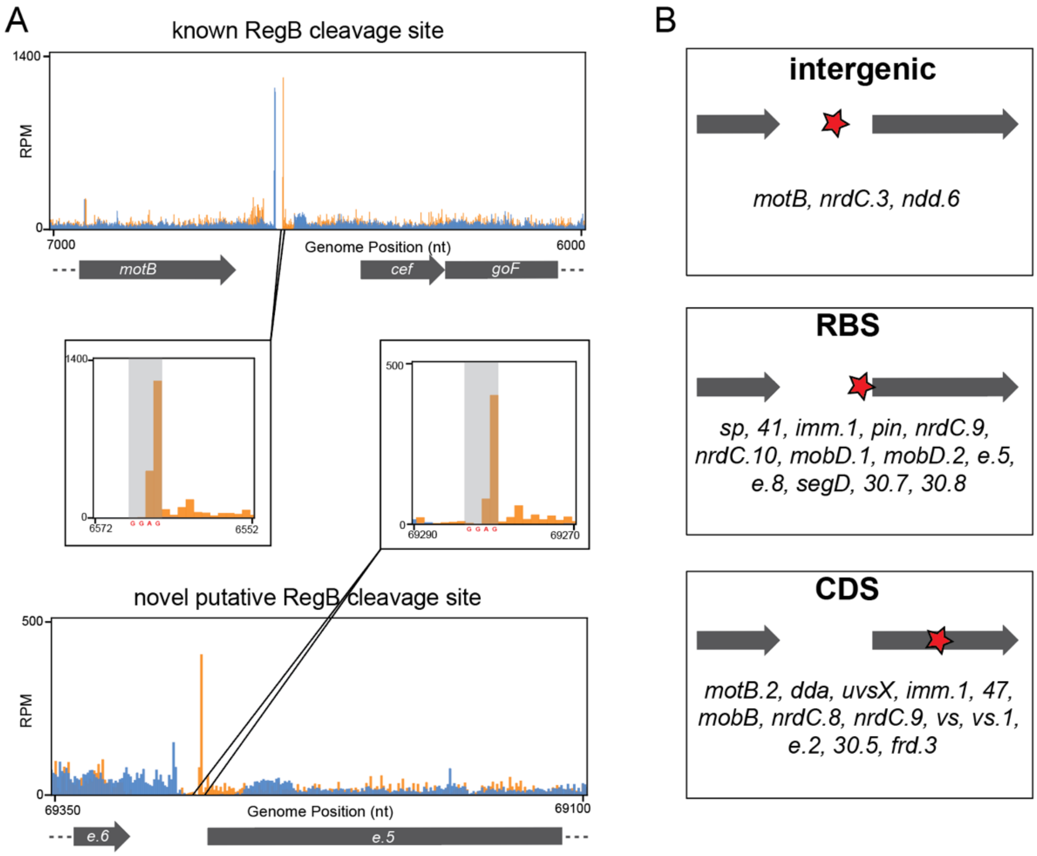


Figure 6

(A) Rend-seq data comparing the appearance of a known RegB cleavage site and a novel putative RegB cleavage site, with zoomed in views in the middle to emphasize the similarity in appearance.

(B) Classification of all novel putative RegB cleavage sites based on whether they are found in an intergenic region of the genome (top), in the RBS of a gene (middle), or within the CDS of a gene (bottom).

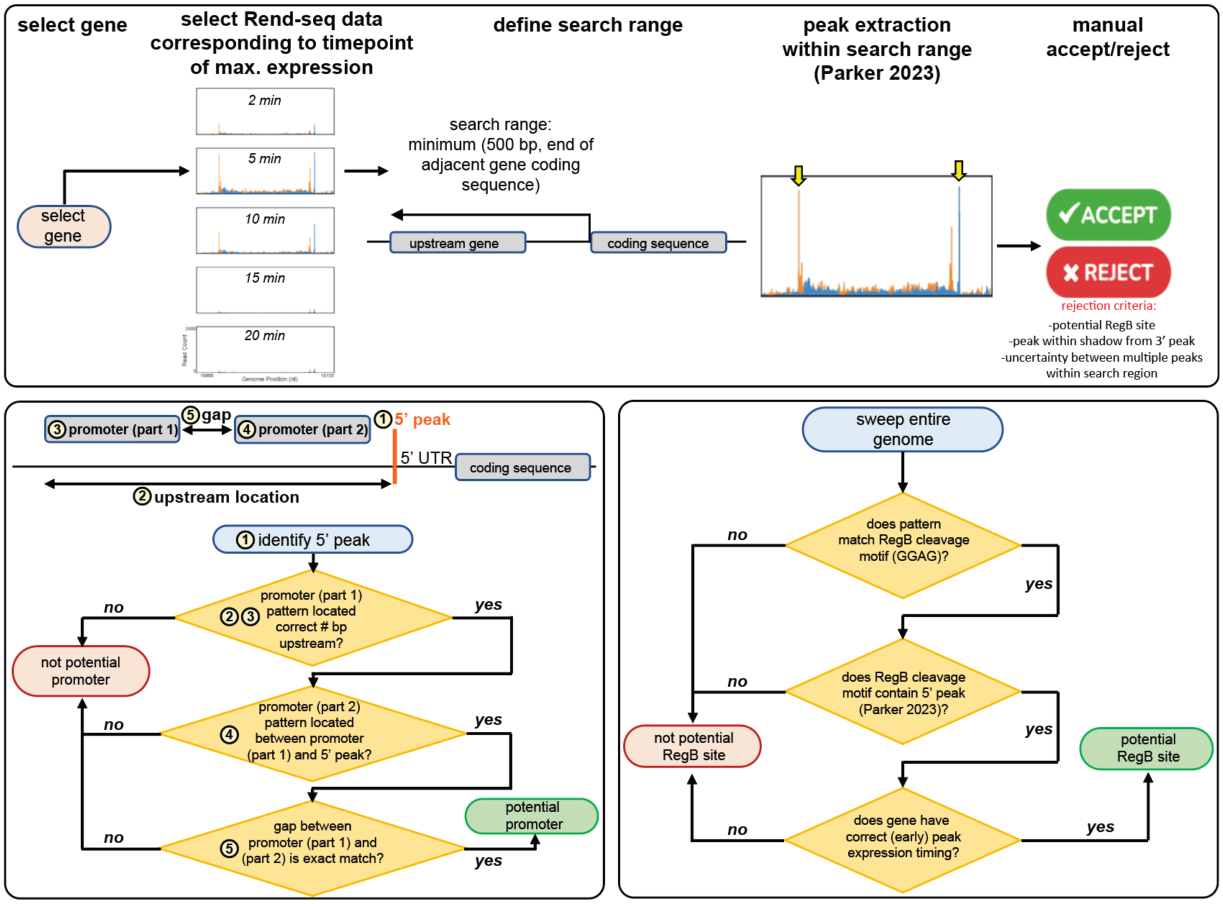

select gene
select Rend-seq data corresponding to timepoint of max. expression
define search range
peak extraction within search range (Parker 2023)
manual accept/reject
select gene
2 min
5 min
10 min
15 min
20 min
Read Count
Genome Position (nt)
search range: minimum (500 bp, end of adjacent gene coding sequence)
upstream gene
coding sequence
✓ACCEPT
✖ REJECT
rejection criteria:
-potential RegB site
-peak within shadow from 3' peak
-uncertainty between multiple peaks within search region
③ promoter (part 1)
⑤ gap
④ promoter (part 2)
① 5' peak
5' UTR
coding sequence
② upstream location
① identify 5' peak
②③ promoter (part 1) pattern located correct # bp upstream?
no
yes
not potential promoter
④ promoter (part 2) pattern located between promoter (part 1) and 5' peak?
no
yes
⑤ gap between promoter (part 1) and (part 2) is exact match?
no
yes
potential promoter
sweep entire genome
does pattern match RegB cleavage motif (GGAG)?
no
yes
does RegB cleavage motif contain 5' peak (Parker 2023)?
no
yes
not potential RegB site
potential RegB site
does gene have correct (early) peak expression timing?
no
yes

Figure 7

Flowchart describing criteria used to determine Rend-seq peaks for each gene (top), putative new promoters (bottom left), and rotative new RegB sites (bottom right).

T4 regulates translational events using complex mechanisms, some of which are uncharacterized

T4 is known to employ a variety of translational regulatory mechanisms to control the levels of protein synthesized from its mRNAs. By comparing our Rend-seq and ribosome profiling datasets, we sought to document and catalog translational regulation during T4 infection.

One known translational regulatory process in T4 is "ribosome hopping," a phenomenon that occurs during translation where the ribosome bypasses a sequence of nucleotides on the mRNA and resumes translation downstream. It can result in the production of truncated proteins or proteins with altered amino acid sequences. A well-known example of ribosome hopping occurs during the translation of Gp60 in T4. In this case, ribosomes hop over a 50 nucleotide non-coding segment of mRNA between the two open reading frames of gene *60* to synthesize a full-length protein (Herr et al., 2001). Our ribosome profiling data allowed us to observe the ribosome occupancy of different sections of mRNAs, thus showing us which regions were being actively translated. Black lines anchored at each nucleotide with heights corresponding to RPMs at each position demonstrated relative levels of protein synthesis across a section of the genome. When looking at the ribosome profile for gene *60*, we saw almost no ribosome density on the section of the gene known to be "hopped" over by the ribosome (Fig 8). Although this visualization helped to support the idea of ribosomes skipping over that particular portion of *60*, we were not able to use our ribosome profiling data to reveal new possible regions of ribosome hopping. This is because there is naturally a great deal of variation in the signal from ribosome profiling over any given mRNA such that an internal region of a transcript with almost no read counts would not necessarily indicate that a ribosome was consistently skipping over that region. The source of this variation is not well understood, although it may reflect different speeds of the ribosome along a transcript, with areas of slower elongation having higher signal (Ingolia et al., 2019).

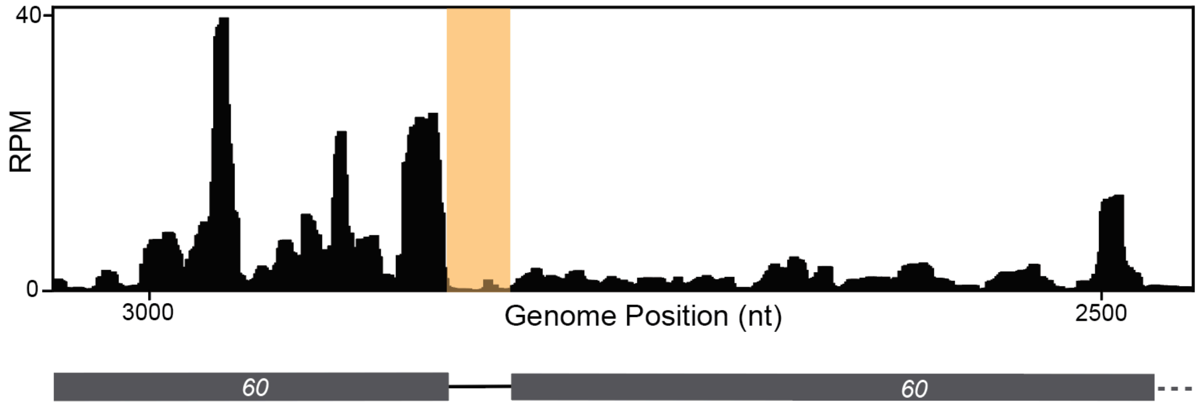


Figure 8

Visualization of the phenomenon known as “ribosome hopping” on T4’s gene *60*. There is almost no ribosome density on the part of the gene that has been shown to be skipped over by the ribosome, which is highlighted in orange.

T4 also uses translational control mechanisms to regulate the timing of protein synthesis. For instance, the endolysin gene, *e*, which is needed for the final step of infection, host cell lysis, is transcribed immediately after infection begins, but is not efficiently translated until the endolysin is needed at later time points. Our Rend-seq data revealed that transcript levels of *e* were nearly as high at 2 minutes post-infection as they were at 20-minutes post-infection, when the endolysin protein is required for cell lysis to release progeny phage. Our ribosome profiling data revealed that there was almost no protein synthesis from these *e* transcripts at 2 minutes post-infection, so even though the transcripts for *e* exist, they were not being used to make proteins (Fig 9D). However, we saw significantly higher levels of protein synthesis at 20 minutes post-infection. We used our data to calculate translational efficiency (TE), the rate of protein production per RNA for a gene at a specific time point. We calculated TE values by dividing ribosome profiling RPKMs by Rend-seq RPKMs. The higher the TE, the more efficiently a given transcripts is being translated. In the case of *e*, its TE increased dramatically over the course of infection, peaking at 20 minutes post-infection (Fig 9D).

There are other examples of translational regulation in T4, including Gp32, Gp43, and RegA. Gp32 and Gp43 bind their own mRNAs to prevent translation of existing transcripts without actively degrading the transcripts themselves (Fig 9A), and RegA binds a number of early mRNAs in addition to its own to achieve the same result (Fig 9B) (Winter et al., 1987). Our data revealed that there are other genes in T4 that have changes in TE at the same magnitude, if not higher, than the genes known to exhibit this regulation (Fig 9E). Although many genes have TEs that remained fairly stable over time, there were also many for which the TE changed significantly over the course of infection (Fig 9E), suggestive of translational regulation. We conclude that time-dependent translational regulation is more extensive that previously appreciated, with dynamic changes in translational efficiency throughout the infection process.

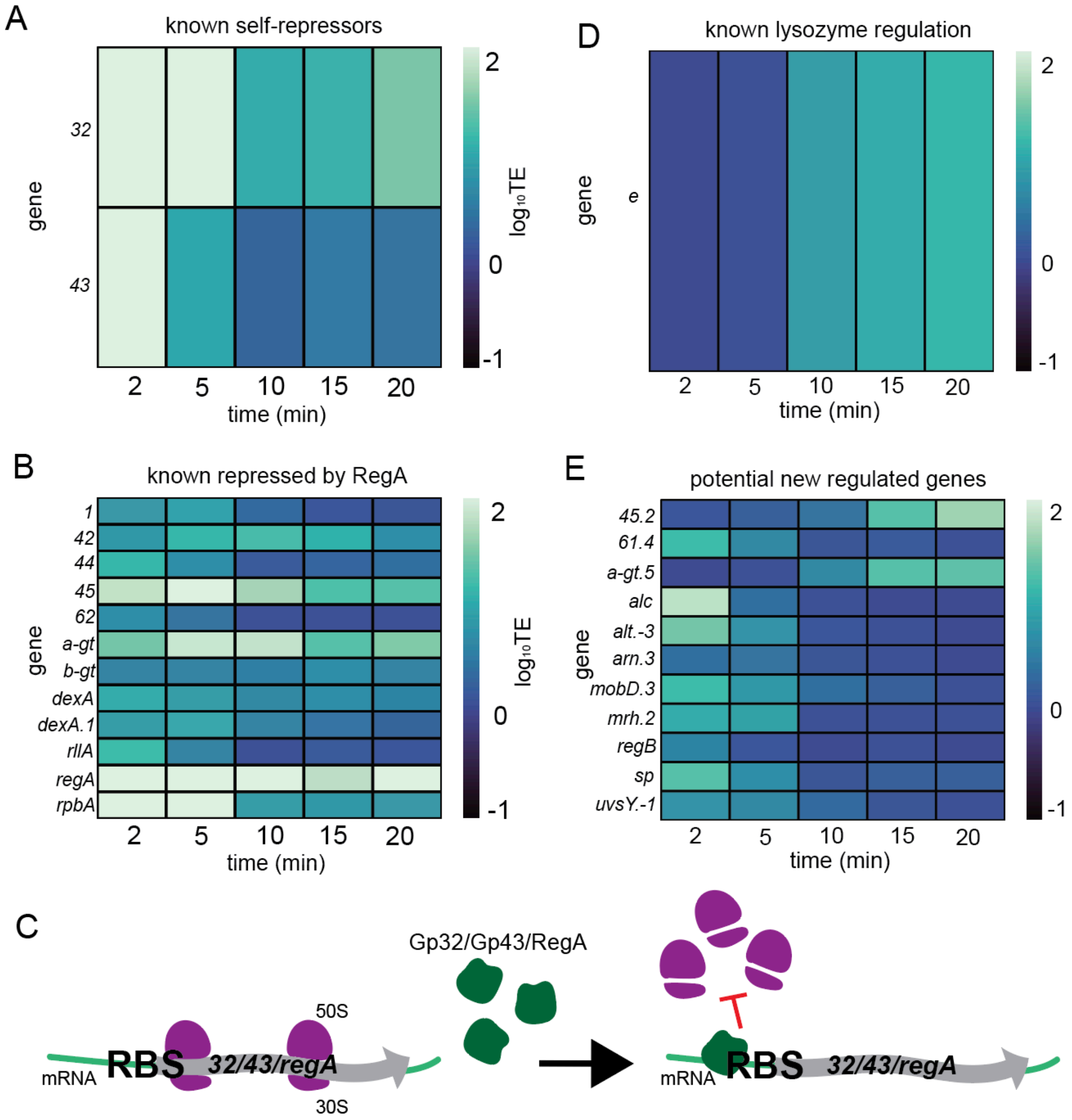


Figure 9

(A) Heatmap illustrating changes in translational efficiency (TE) over time for genes *32* and *43*, which are known to exhibit self-imposed translational regulation.

(B) Heatmap illustrating known translational regulation, in which all of the listed genes are repressed by *regA*.

(C) Representative illustration of the mechanism of action of translational regulation carried out by Gp32, Gp43, and RegA. All three proteins bind to the ribosome binding site (RBS) of their own mRNA, preventing translation.

(D) Heatmap illustrating an increase in TE over time for gene *e*, which is known to produce transcripts early in infection but only translate them as the infection progresses.

(E) Heatmap illustrating some of the genes with the largest changes in TE over the course of infection. None of these genes have been previously characterized as being regulated translationally.

Proportional synthesis is a conserved process found in T4 in addition to higher organisms

Our data also enabled us to document proportional synthesis in T4 phage, the first such example for any viral system. Proportional synthesis is a sophisticated form of translational control that has been predominantly studied in eukaryotic and prokaryotic cells, in which proteins that are parts of complexes are produced in the ratio that they are needed in the complex (Lalanne et al., 2018; Li et al., 2014). Our assessment of proportional synthesis in T4 focused on nine protein complexes: long tail fibers, ATP-dependent clamp loader, primosome, head scaffold, neck, MR complex, tail wedge, head, and terminase (Fig 10A). These complexes each have strict stoichiometric requirements for the involved proteins in order to function correctly. The structural proteins of T4 cannot be properly assembled without enough of each protein component, but excess protein production during an infection process in which time and resources are of the essence could be detrimental to the productivity of the infection as a whole.

An especially clear visualization of proportional synthesis is laid out for the protein complex that serves as the T4 head (Fig 10B), which needs 960 copies of Gp23 but only 55 copies of Gp24 and 12 copies of Gp20. Thus, gene *23* needs to be translated at a much higher level than gene *24* or *20*, and indeed we see that the measured ribosome profiling RPKMs for Gp23 are substantially higher than Gp24 and Gp20 with an approximately linear and proportional relationship between the experimentally measured ribosome profiling RPKMs for each protein in the complex (y-axis) and the proportion of each protein in the complex (x-axis). A similar pattern was observed for each of the nine protein complexes considered (Fig 10B). In other organisms that have been shown to exhibit proportional synthesis (Lalanne et al., 2018; Li et al., 2014), this form of regulation is important for maintaining a balance between synthesis and degradation of proteins to achieve cellular homeostasis. However, T4, and lytic phage in general, do not exhibit homeostasis—their conditions are never steady since they undergo a time-limited lifecycle throughout which they must be constantly changing in order to create new phage. And yet, they seem to achieve the same type of proportional synthesis, likely to promote proper formation of each complex and to avoid wasteful synthesis of proteins that cannot be incorporated into complexes.

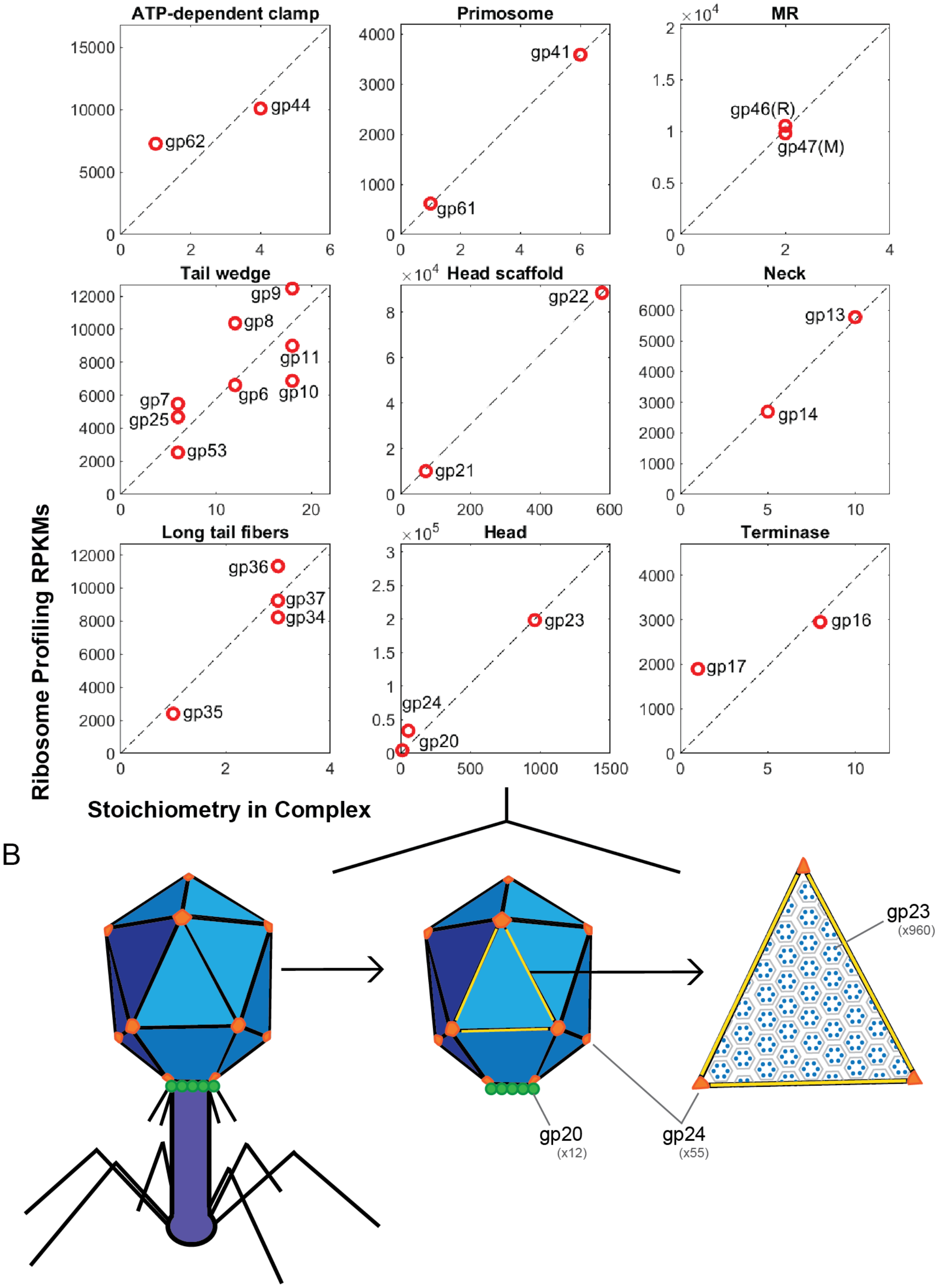


Figure 10

(A) Scatter plots representing various protein complexes in T4, depicting amounts of each protein per complex known in the literature (stoichiometry in complex) vs. synthesis levels of each protein in this study (ribosome profiling RPKMs). The dotted line is the line of best fit that passes through the origin.

(B) Visual representation of proportional synthesis for T4's head, illustrating why the different proteins in the head would need to be made in different proportions.

The T4 genome contains a previously undescribed gene, which produces a functional protein

Genome annotation is the process of identifying the location and function of genes and other functional elements in a genome. It is an essential step in understanding an organism, as it gives meaning to its genome by identifying the functional elements of DNA, such as genes, regulatory sequences, and non-coding RNAs. There are two main computational approaches to genome annotation: ab initio and homology-based. Ab initio methods identify genes and their structures using mathematical models and the DNA sequence as the only input, while homology-based methods predict genes by aligning a protein or RNA sequence with the genome sequence that needs to be annotated (Ekigu et al., 2020). In our study, we used our Rend-seq and ribosome profiling data to annotate the genome and identify regions of active transcription and translation across the T4 genome. This effort led to the identification of a previously unknown gene.

As discussed above, Rend-seq can be used to identify transcripts in the genome by finding a clear 5′ (orange) peak upstream of a 3′ (blue) peak with some amount of read density between the two. When examining our dataset, we were able to call distinct peaks in most regions where the presence of peaks would be expected, *i.e.* in regions with annotated genes. However, we also were able to detect evidence for three transcripts in regions with no prior annotations. The three transcripts were spread throughout the genome; the first occurring between nucleotides 11,897 and 11,975 (Fig 11A), the second between nucleotides 17,721 and 18,025 (Fig 11B), and the third between nucleotides 70,798 and 70,956 (Fig 11C). While all three appeared to be transcripts, only two of them (the second and third) had substantial and clearly detectable ribosome density shown within the bounds of these putative transcripts, indicating that the first could potentially be a noncoding RNA. The third potential transcript showed ribosome density, but without a clear open reading frame (ORF). We found no AUG start codons within the ORF, but did find a GUG sequence that could possibly serve as an alternative start codon, though the corresponding ORF would only contain 10 amino acids. The second transcript has an ORF of 66 amino acids contained between the Rend-seq peaks, as well has having ribosome density in region of the ORF. We therefore decided to further investigate this second transcript, which will be referred to as *61.-1*, as T4 gene naming convention states that new hypothetical or uncharacterized ORFs should be named in reference to the preceding known gene, which in this case in *61*, and as this new gene is located in the opposite direction of *61*, it has the designation of being *61.-1* (Miller et al., 2003).

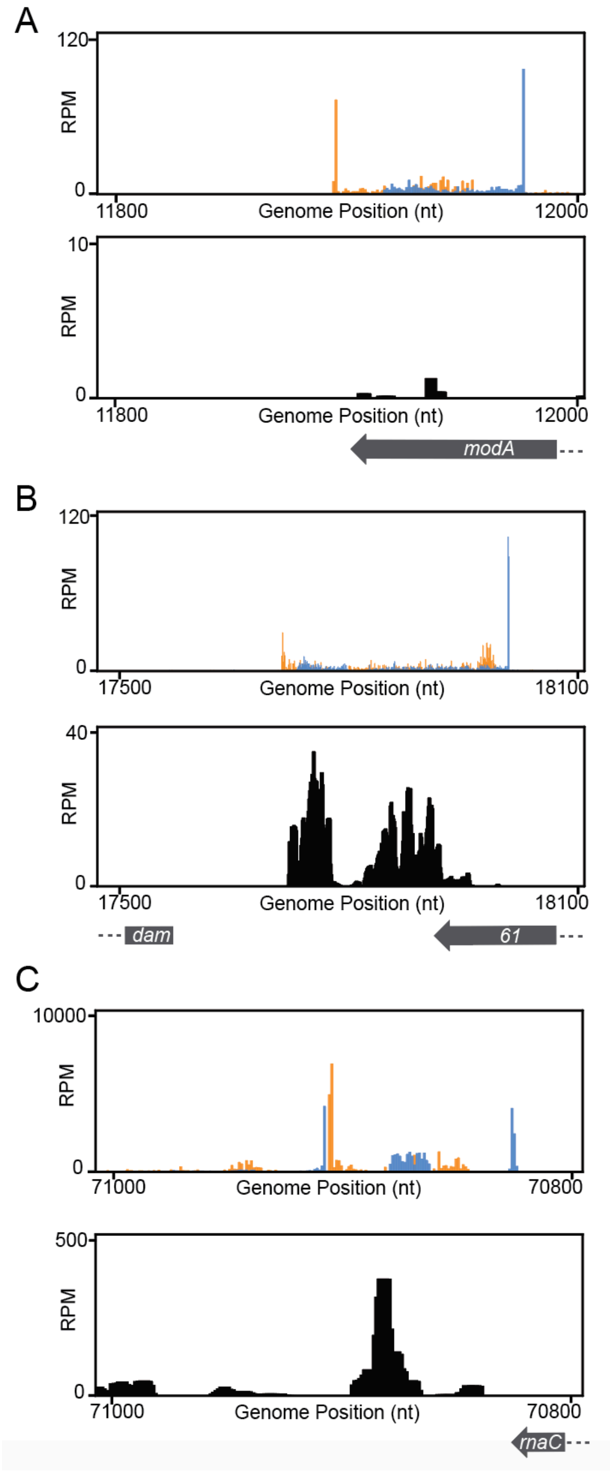


Fig 11

(A) Rend-seq (top) and ribosome profiling (bottom) data illustrating the presence of the first new putative transcript in an unannotated area of the genome.

(B) As described above, but for the second putative transcript.

(C) As described above, but for the third putative transcript.

This new gene, "*61.-1*," was expressed in a time-dependent manner and peaked at the 20 minute time point, suggesting it is a late gene (Fig 12A). As discussed above, there were no annotated genes that were also on the forward strand in the region where the *61.-1* transcript was found, but the location of this gene in the genome was particularly interesting because it was found right between two genes on the reverse strand: *dam* (DNA-adenine methyltransferase (Hattman et al., 1985; Zinoviev et al., 1998)) and *61* (DNA primase which requires interaction with Gp41 helicase for priming at unique sequences (Miller et al., 2003; Hinton et al., 1987)). This positioning was atypical because forward genes tend to be clustered in certain areas of T4's genome (Miller et al., 2003). In fact, the 5′ end of *61.-1* is 25,671 nucleotides away from the 3′ end of the nearest forward gene, *t*.

Further genomic evidence supporting our discovery of *61.-1* was that upstream of the transcript there were features of a traditional T4 late promoter (ATAAATA), and within the transcript, a T4 Shine-Dalgarno sequence (AGGA) precedes the 66-amino acid open reading frame (ORF). The start and stop codons of the ORF lined up precisely with the beginning and end of the protein synthesis signal from the ribosome profiling data (Fig 12A). Moreover, the amino acid sequence of Gp61.-1 was well conserved among a number of other phage species, including both lytic phages and prophages found within the genomes of some bacteria (Fig 12B), including *E. coli*, *Bacillus cereus*, *Shigella*, and *Citrobacter*. Out of the 66 amino acids in Gp61.-1, the homolog in the phage with the most differences still differs by only four amino acids, with three of those conservative substitutions. In addition, Gp61.-1 and its homologs are always positioned between two annotated genes: one of the genes always being a DNA primase, and the other often being a dCTP pyrophosphatase (in four of the five homologs). Much like *61.-1*, the homologs are also always on the opposite strand of the annotated genes they lie between (Fig 12C).

To begin assessing the possible function of Gp61.-1, we generated a predicted protein structure using AlphaFold (Jumper et al., 2021) (Fig 12D). The Gp61.-1 protein appeared to be composed of three alpha-helices, connected by unstructured regions. When we searched for homologs and conserved domains using HHPred, the closest matches were a putative copper-exporting P-type ATPase A and a membrane protein (Söding et al., 2005). We were also able to search existing mass spectrometry data to confirm that Gp61 is indeed produced during the T4 infection cycle, and is likely membrane-bound or part of polysomes due to where it was located in the gradient used for the experiment (Vogel, unpublished).

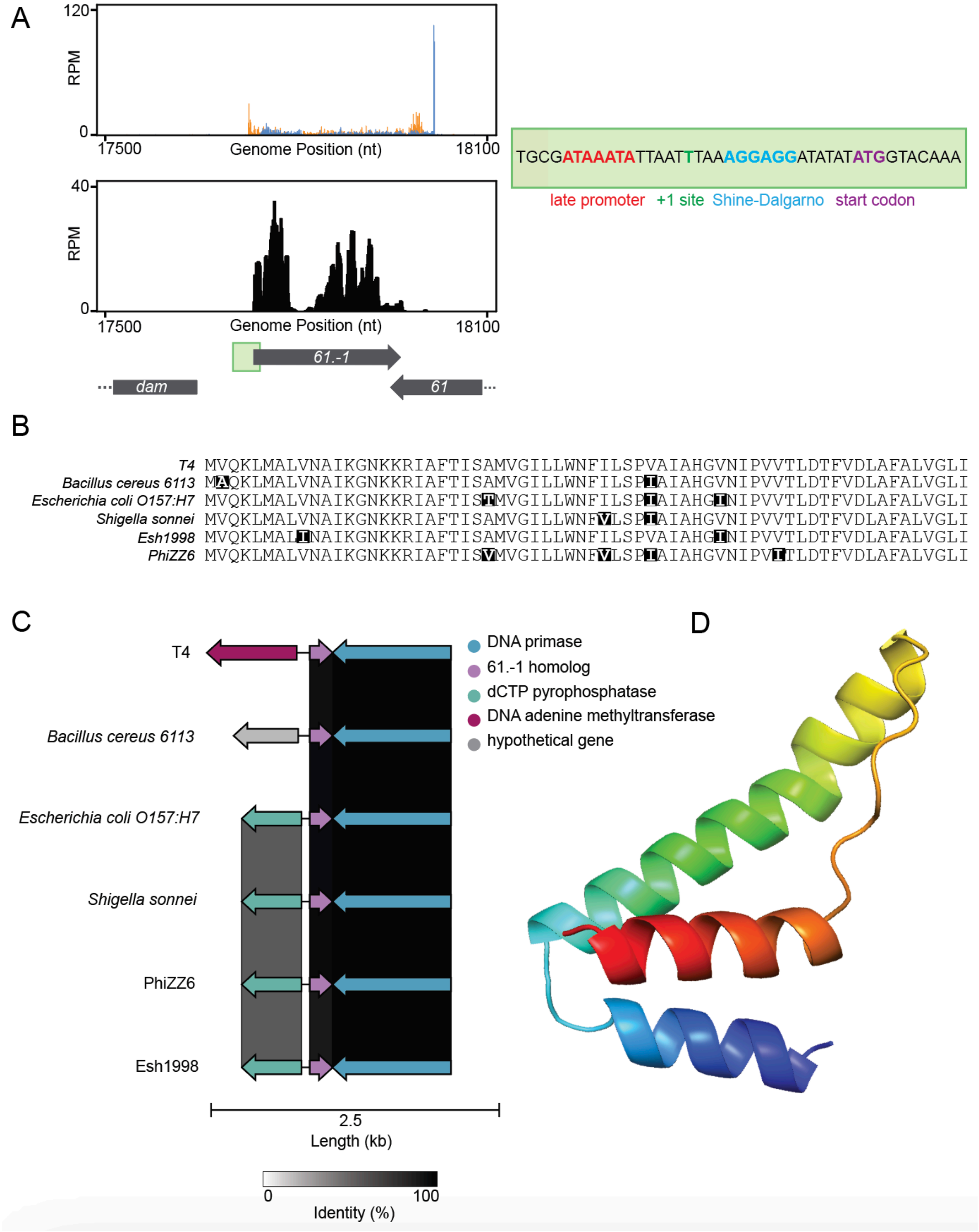

A
120
RPM
0
17500
Genome Position (nt)
18100
40
RPM
0
17500
Genome Position (nt)
18100
61.-1
dam
61
TGCGATAAATATTAATTTAAAGGAGGATATATATGGTACAAA
late promoter
+1 site
Shine-Dalgarno
start codon
B
T4 MVQKLMALVNAIKGNKKRIAFTISAMVGILLWNFILSPVAIAHGVNIPVVTLDTFVDLAFALVGLI
Bacillus cereus 6113 MAQKLMALVNAIKGNKKRIAFTISAMVGILLWNFILSPIAIAHGVNIPVVTLDTFVDLAFALVGLI
Escherichia coli O157:H7 MVQKLMALVNAIKGNKKRIAFTISTMVGILLWNFILSPIAIAHGINIPVVTLDTFVDLAFALVGLI
Shigella sonnei MVQKLMALVNAIKGNKKRIAFTISAMVGILLWNFVLSPIAIAHGVNIPVVTLDTFVDLAFALVGLI
Esh1998 MVQKLMALINAIKGNKKRIAFTISAMVGILLWNFILSPVAIAHGINIPVVTLDTFVDLAFALVGLI
PhiZZ6 MVQKLMALVNAIKGNKKRIAFTISVMVGILLWNFVLSPIAIAHGVNIPVITLDTFVDLAFALVGLI
C
T4
Bacillus cereus 6113
Escherichia coli O157:H7
Shigella sonnei
PhiZZ6
Esh1998
● DNA primase
● 61.-1 homolog
● dCTP pyrophosphatase
● DNA adenine methyltransferase
● hypothetical gene
2.5
Length (kb)
0
100
Identity (%)
D

Figure 12

(A) Rend-seq (top) and ribosome profiling (bottom) traces on the forward strand, depicting the signal produced which indicates the presence of *61.-1*.

(B) Depiction of the amino acid sequence alignment for *61.-1* and its homologs, with amino acids that differ from T4's *61.-1* highlighted.

(C) Genetic neighborhood analysis of different *61.-1* homologs in various lytic phage and prophage contained in bacterial species.

(D) Visual representation of a putative structure for *61.-1*, produced by AlphaFold. The structure is colored red to blue (N terminus to C terminus).

To better understand the function of *61.-1* we explored the effect it had on wild-type *E. coli* when expressed ectopically in the absence of phage infection. We created a strain of *E. coli* that contained a plasmid (pBAD18) carrying *61.-1* where expression could be repressed by glucose or induced by the addition of arabinose. We then performed a growth curve experiment in which we grew six different flasks of *E. coli*: three flasks with bacteria carrying an empty vector (as a negative control) and then three flasks of bacteria carrying the vector with *61.-1*. We started the experiments with all flasks at an OD600 of 0.001, growing in M9 minimal media with glucose at 37˚C. At three different optical densities (0.01, 0.1, and 0.3) we used rapid filtration in the two relevant flasks for the time point to isolate the cells from the media (Johnson et al., 2018), then took the nitrocellulose filters and put them into 50 mL conical tubes containing prewarmed new M9, this time with arabinose added. The media with the cells was then transferred to a flask, which was returned to the shaker.

This experiment revealed that ectopic expression of *61.-1* affected *E. coli* growth by causing a drop in optical density. Interestingly, we found that there was a lag in the timing between induction and the beginning of cell death, which became shorter the higher the OD is when induction began, suggesting that *61.-1* may affect *E. coli* in a growth phase-dependent manner (Fig 13A). When the same experiment was performed with a strain expressing a *61.-1* mutant in which all start codons had been mutated to prevent protein production, we did not see this effect, which suggests that it is the protein *61.-1* rather than the transcript itself that caused this effect. The induction of *61.-1* likely causes cell death rather than simply stasis; CFU plating performed after roughly 7 hours of growth with arabinose added to the media showed that wild type and empty vector strains were able to grow on an LB plate overnight, while the strain expressing *61.-1* was not able to do so (Fig 13B).

Microscopy analysis was performed on cells harboring the empty vector or the plasmid containing inducible *61.-1*. We started cultures that were back-diluted to OD_{600} 0.001 in M9 media without glucose or arabinose, and once they reached an OD of 0.3, we pipetted 0.5 µL of each strain onto an agar pad containing arabinose (to induce production of *61.-1*) and propidium iodide (a stain used to differentiate between healthy cells and cells with compromised membranes, as it is not membrane-permeable and will thus only enter and stain cells with damaged membranes). The agar pads were photographed every 10 minutes, and by the time the cells had been on the agar pads for three hours, we were able to see that there were overall more bacterial cells for the empty vector cells than the cells expressing *61.-1*, though there were cells for both strains. However, of the cells present, nearly all the cells expressing *61.-1* were fluorescent, while most of the empty vector cells had only a few cells that fluoresced. This result indicated compromised cell membranes in cells expressing *61.-1*, as nearly all cells expressing *61.-1* fluoresced when stained with propidium iodide (Fig 13C).

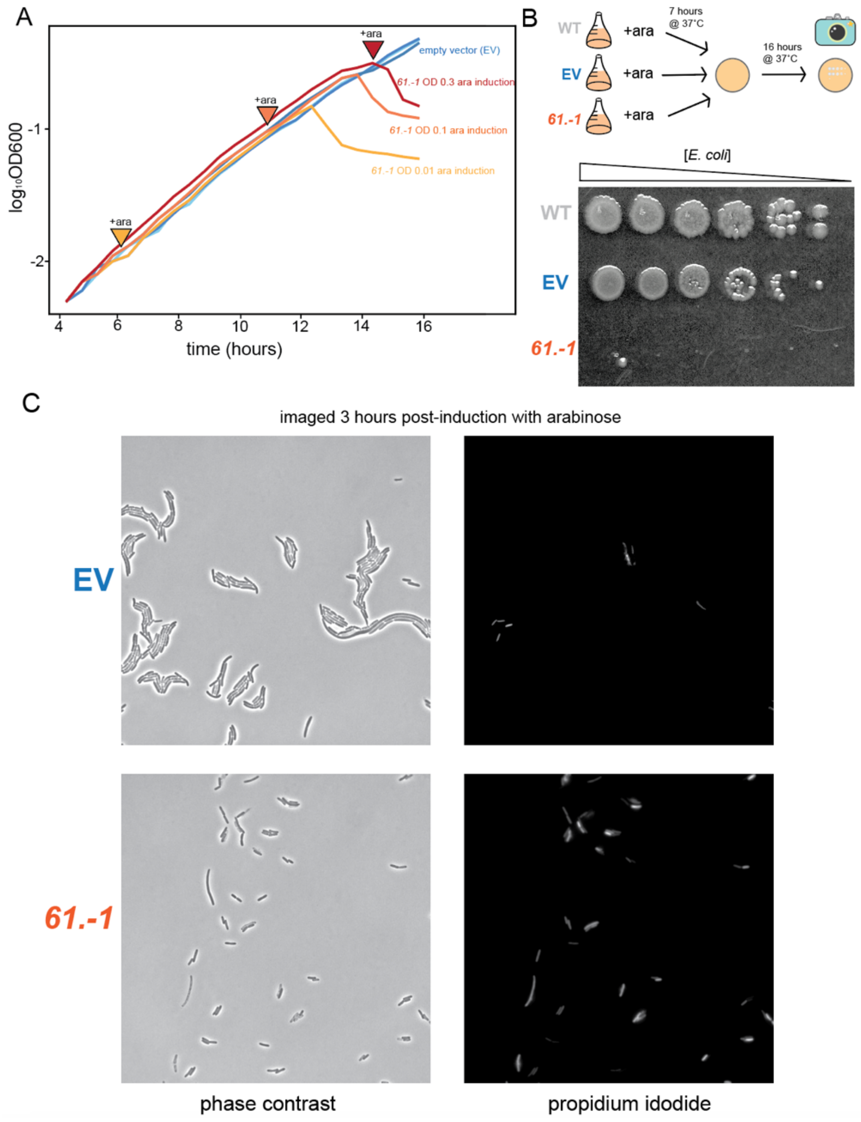


Fig 13

(A) Growth curves of MG1655 with either the empty vector (EV) or the vector containing *61.-1*. All strains started growing in 0.4% glucose, and at the listed OD the cells were filtered from the glucose media and resuspended in media containing 0.1% arabinose.

(B) CFU plating of wild-type MG1655 (WT), MG1655 with the empty vector (EV), and MG1655 with the *61.-1*-containing vector (*61.-1*). Bacterial samples were taken after 7 hours of growth in arabinose and spotted in serial dilutions to be visualized the next day.

(C) Microscopy images of MG1655 with empty vector (EV) or MG1655 with *61.-1*-containing vector (*61.-1*) when added to and incubated on an agar pad including arabinose (for vector induction) and propidium iodide (for staining cells with compromised cell membranes).

Given our findings that *61.-1* is expressed by T4 and affected *E. coli* growth when ectopically expressed, we wanted to gain a better understanding of what role *61.-1* might hold in T4 during infection. We used CRISPR to generate a deletion mutant in which *61.-1* was disrupted in T4's genome. We designed guide RNAs to target *61.-1* in T4; fortunately wild type T4 was susceptible to these, allowing us to avoid the use of genetically altered T4 that does not glucosylate its DNA, which is often needed for genetic modification. We also constructed plasmids containing a region of DNA homologous to the region in T4 where *61.-1* is found, but with the portion of *61.-1* that does not overlap with gene 61 completely removed, and the overlapping part recoded to ensure that *61.-1* would not be expressed. We infected E. coli containing both plasmids with wild type T4, and this created Δ*61.-1*, a modified strain of T4 which did not express *61.-1*.

We compared infection of wild-type and Δ*61.-1* T4 in *E. coli* through a one-step growth curve (Kropinski, 2018). This experiment revealed that the deletion of *61.-1* did not significantly affect the burst size (Fig 14A) or latent period (Fig 14B) of T4, indicating that *61.-1* may play a role in pathways not directly related to the timing of lysis or amount of phage production per infection cycle. However, this is only one measure of how *61.-1* may be affecting the life cycle of T4. There remains evidence that *61.-1* is a well-conserved gene that produces a protein that adversely affects E. coli following ectopic expression. Thus, *61.-1* represents a novel candidate for further exploration. Its potential functions and interactions with other T4 proteins warrant investigation, as it may play crucial roles in the phage's lifecycle and its interactions with the host, with implications that expand beyond T4 due to the highly conserved nature of *61.-1*.

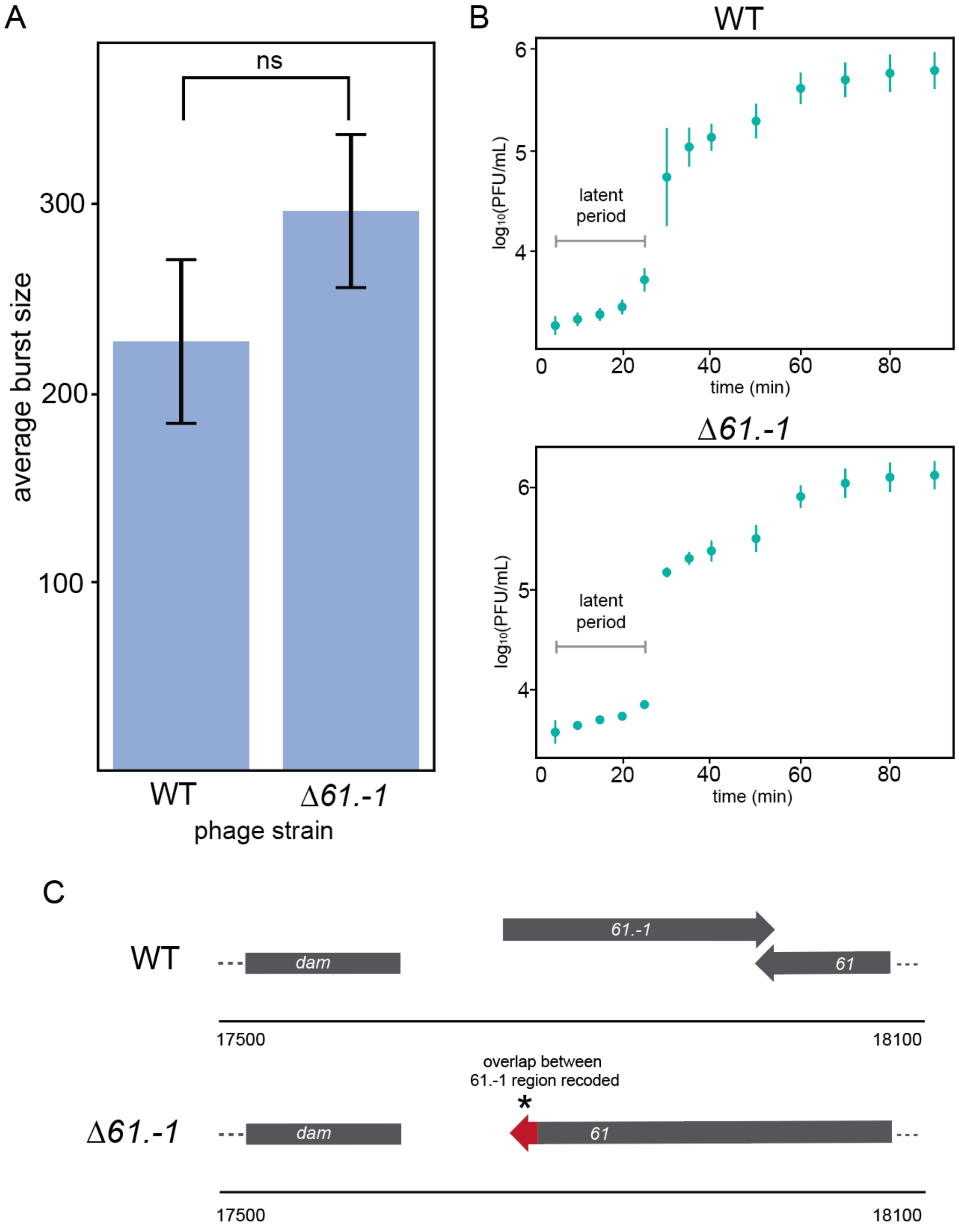


Fig 14

(A) Quantification of average burst size from infection of MG1655 with either wild-type T4 (WT) or T4 with *61.-1* knocked out (Δ*61.-1*), based on four replicates of a one-step growth curve. Bars indicate error bars from four replicate experiments, significance measured by a two-sample t-test.

(B) Plots illustrating the outcome of four one-step growth curves with either wild-type T4 (WT) or T4 with *61.-1* knocked out (Δ*61.-1*), displaying the PFU/mL over time and illustrating the length of the latent period.

(C) Diagram illustrating the region of the T4 genome mutated by CRISPR, highlighting the differences between wild-type T4 (WT) or T4 with *61.-1* knocked out (Δ*61.-1*), as the *61.-1* region was deleted and the overlap between *61.-1* and *61* was recoded.

Discussion

Studying T4 phage has proven to be of immense significance, providing valuable insights into various fundamental aspects of molecular biology and virology. This paper has highlighted the importance of investigating T4 phage through the application of advanced techniques, such as Rend-seq and ribosome profiling, and has expanded our understanding of its transcriptional and translational regulation. Rend-seq has enabled us to understand not only transcript levels at any given time during the course of infection, but to also identify transcript ends. This information can be valuable in many ways, including the discovery of new promoters and regulatory cleavage sites. More broadly, it allows a deeper look into T4 transcription at throughout infection, revealing transcript ends that may be within a gene, indicating secondary start sites or post-transcriptional processing. There may also be transcript ends that do not appear to be associated with a previously annotated gene, indicating the presence of uncharacterized mRNAs. As we pair Rend-seq with ribosome profiling, we are also able to access information regarding which transcripts are being translated at various times, and to what extent they are being translated at those times. In conjunction with our Rend-seq data, this allows us to understand more nuanced aspects of gene expression and regulation in T4; not only are we able to view levels of transcripts present as well as levels of translation of those transcripts, we are able to combine the two to understand the translational efficiency (TE) of any transcript. By using both sets of information, we can calculate the rate of protein production per mRNA, and since we have data from multiple time points throughout infection, we can observe and quantify changes in TE for particular transcripts over time. Changes in TE over time indicate some form of translational regulation, so these measurements open the door to answer many other questions regarding gene regulation in T4. We are also able to view where translation is occurring across the genome, and this information, combined with the transcript information from Rend-seq, can alert us to the presence of previously undescribed genes that lie in the T4 genome.

Our data provides us with a deeper understanding of temporal gene expression and regulation in T4 phage. One of the findings in this study is the confirmation of T4's time-dependent gene expression, a phenomenon that has been previously documented but remains crucial to comprehend more fully. By capturing snapshots of transcription and translation at specific time points during infection, we are able to visualize the progression of gene expression levels. We additionally characterize transcript ends, allowing us to designate some transcripts as monocistronic or polycistronic, as well as identify new putative promoters and RegB cleavage sites within the T4 genome, giving us further insight into the complexity of transcriptional regulation in T4. This provides valuable information for deciphering similar processes in other organisms, highlighting the broader implications of studying phages for the purpose of better understanding gene expression control. Furthermore, our analysis of translational regulation in T4 phage reveals a remarkable level of complexity, extending beyond what has been previously reported. The discovery of genes exhibiting substantial changes in translational efficiency points to the presence of additional translation regulatory mechanisms yet to be fully elucidated. The first observation of proportional synthesis in any phage or viral system also underscores the similarity between phages and other biological entities in fine-tuning gene expression based on the needs of protein complexes.

The identification and experimental validation of T4's 290th gene, *61.-1*, represents a significant contribution to the T4 genome annotation. This novel gene, conserved among other phages, offers a promising avenue for future investigations into its functions and interactions with other T4 genes. Conducting Rend-seq and ribosome profiling with our Δ*61.-1* strain of T4 has the potential to grant us a deeper understanding of the role of *61.-1* during infection. We know that *61.-1* does not affect the latent period or bust size resulting from infection, but we have no information on its genetic regulation potential at the transcriptional or translational level. With a dataset equivalent to the one described in this paper, but using the knockout strain of T4 instead, we would be able to identify changes in the presence of different transcripts, as well as changes in the levels of translation occurring over these transcripts. By identifying specific genes whose expression levels are altered by the deletion of *61.-1*, we could gain insight into the role of *61.-1* in T4's life cycle in a time-resolved manner. In this way, we may be able to shed light on its impact on the host and even reveal new targets for therapeutic interventions against bacterial infections, as *61.-1* appears to be well conserved in other species of phage.

The study of T4 phage continues to be of paramount importance in expanding our knowledge of molecular biology, virology, and gene regulation. The findings presented in this paper provide a stepping stone for further exploration of T4 phage biology and its implications in broader scientific fields. By delving into the complexities of T4 phage biology, researchers can continue to unravel the mysteries of phages and their potential applications in biotechnology and medicine, ultimately shaping the future of scientific inquiry and advancing our understanding of life's fundamental processes.

References

Andrake, M, N Guild, T Hsu, L Gold, C Tuerk, and J Karam. "DNA Polymerase of Bacteriophage T4 Is an Autogenous Translational Repressor." *Proceedings of the National Academy of Sciences* 85, no. 21 (November 1988): 7942–46. https://doi.org/10.1073/pnas.85.21.7942.

Belanger, Karyn Goudie, and Kenneth N Kreuzer. "Bacteriophage T4 Initiates Bidirectional DNA Replication through a Two-Step Process." *Molecular Cell* 2, no. 5 (November 1998): 693–701. https://doi.org/10.1016/S1097-2765(00)80167-7.

Cobb, Matthew. "Who Discovered Messenger RNA?" *Current Biology* 25, no. 13 (June 2015): R526–32. https://doi.org/10.1016/j.cub.2015.05.032.

Crick, F. H. C., Leslie Barnett, S. Brenner, and R. J. Watts-Tobin. "General Nature of the Genetic Code for Proteins." *Nature* 192, no. 4809 (December 1961): 1227–32. https://doi.org/10.1038/1921227a0.

Durand, Sylvain, Graziella Richard, François Bontems, and Marc Uzan. "Bacteriophage T4 Polynucleotide Kinase Triggers Degradation of mRNAs." *Proceedings of the National Academy of Sciences* 109, no. 18 (May 2012): 7073–78. https://doi.org/10.1073/pnas.1119802109.

Ejigu, Girum Fitihamlak, and Jaehee Jung. "Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing." *Biology* 9, no. 9 (September 18, 2020): 295. https://doi.org/10.3390/biology9090295.

Hattman, S, J Wilkinson, D Swinton, S Schlagman, P M Macdonald, and G Mosig. "Common Evolutionary Origin of the Phage T4 Dam and Host Escherichia Coli Dam DNA-Adenine Methyltransferase Genes." *Journal of Bacteriology* 164, no. 2 (November 1985): 932–37. https://doi.org/10.1128/jb.164.2.932-937.1985.

Hendricks, Stephen P., and Christopher K. Mathews. "Regulation of T4 Phage Aerobic Ribonucleotide Reductase: SIMULTANEOUS ASSAY OF THE FOUR ACTIVITIES *." *Journal of Biological Chemistry* 272, no. 5 (January 31, 1997): 2861–65. https://doi.org/10.1074/jbc.272.5.2861.

Hinton, D. M., and N. G. Nossal. "Bacteriophage T4 DNA Primase-Helicase. Characterization of Oligomer Synthesis by T4 61 Protein Alone and in Conjunction with T4 41 Protein." *Journal of Biological Chemistry* 262, no. 22 (August 5, 1987): 10873–78. https://doi.org/10.1016/S0021-9258(18)61045-2.

Ingolia, Nicholas T. "Ribosome Footprint Profiling of Translation throughout the Genome." *Cell* 165, no. 1 (March 24, 2016): 22–33. https://doi.org/10.1016/j.cell.2016.02.066.

Ingolia, Nicholas T., Jeffrey A. Hussmann, and Jonathan S. Weissman. "Ribosome Profiling: Global Views of Translation." *Cold Spring Harbor Perspectives in Biology* 11, no. 5 (May 1, 2019): a032698. https://doi.org/10.1101/cshperspect.a032698.

Johnson, Grace E., and Gene-Wei Li. "Chapter Ten - Genome-Wide Quantitation of Protein Synthesis Rates in Bacteria." In *Methods in Enzymology*, edited by Agamemnon J. Carpousis, 612:225–49. High-Density Sequencing Applications in Microbial Molecular Genetics. Academic Press, 2018. https://doi.org/10.1016/bs.mie.2018.08.031.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596, no. 7873 (August 2021): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Kropinski, Andrew M. "Practical Advice on the One-Step Growth Curve." *Methods in Molecular Biology (Clifton, N.J.)* 1681 (2018): 41–47. https://doi.org/10.1007/978-1-4939-7343-9_3.

Kuhn, Andreas, and Julie A. Thomas. "The Beauty of Bacteriophage T4 Research: Lindsay W. Black and the T4 Head Assembly." *Viruses* 14, no. 4 (April 2022): 700. https://doi.org/10.3390/v14040700.

Kutter, Elizabeth, Daniel Bryan, Georgia Ray, Erin Brewster, Bob Blasdel, and Burton Guttman. "From Host to Phage Metabolism: Hot Tales of Phage T4's Takeover of E. Coli." *Viruses* 10, no. 7 (July 2018): 387. https://doi.org/10.3390/v10070387.

Lalanne, Jean-Benoît, James C. Taggart, Monica S. Guo, Lydia Herzel, Ariel Schieler, and Gene-Wei Li. "Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry." *Cell* 173, no. 3 (April 2018): 749-761.e38. https://doi.org/10.1016/j.cell.2018.03.007.

Li, Gene-Wei, David Burkhardt, Carol Gross, and Jonathan S. Weissman. "Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources." *Cell* 157, no. 3 (April 24, 2014): 624–35. https://doi.org/10.1016/j.cell.2014.02.033.

Liu, Xiaoqiu, Huifeng Jiang, Zhenglong Gu, and Jeffrey W. Roberts. "High-Resolution View of Bacteriophage Lambda Gene Expression by Ribosome Profiling." *Proceedings of the National Academy of Sciences* 110, no. 29 (July 16, 2013): 11928–33. https://doi.org/10.1073/pnas.1309739110.

Luke, Kimberly, Agnes Radek, XiuPing Liu, John Campbell, Marc Uzan, Robert Haselkorn, and Yakov Kogan. "Microarray Analysis of Gene Expression during Bacteriophage T4 Infection." *Virology* 299, no. 2 (August 1, 2002): 182–91. https://doi.org/10.1006/viro.2002.1409.

Maghsoodi, Ameneh, Anupam Chatterjee, Ioan Andricioaei, and Noel C. Perkins. "How the Phage T4 Injection Machinery Works Including Energetics, Forces, and Dynamic Pathway." *Proceedings of the National Academy of Sciences* 116, no. 50 (December 10, 2019): 25097–105. https://doi.org/10.1073/pnas.1909298116.

Miller, Eric S., Elizabeth Kutter, Gisela Mosig, Fumio Arisaka, Takashi Kunisawa, and Wolfgang Rüger. "Bacteriophage T4 Genome." *Microbiology and Molecular Biology Reviews* 67, no. 1 (March 2003): 86–156. https://doi.org/10.1128/mmbr.67.1.86-156.2003.

Piešiniene, Lina, Lidija Truncaite, Aurelija Zajančkauskaite, and Rimas Nivinskas. "The Sequences and Activities of RegB Endoribonucleases of T4-Related Bacteriophages." *Nucleic Acids Research* 32, no. 18 (2004): 5582–95. https://doi.org/10.1093/nar/gkh892.

Russel, M, L Gold, H Morrissett, and P Z O'Farrell. "Translational, Autogenous Regulation of Gene 32 Expression during Bacteriophage T4 Infection." *Journal of Biological Chemistry* 251, no. 22 (November 25, 1976): 7263–70. https://doi.org/10.1016/S0021-9258(17)32967-8.

Samuel, Charles E. "Polycistronic Animal Virus mRNAs." In *Progress in Nucleic Acid Research and Molecular Biology*, edited by Waldo E. Cohn and Klvle Moldave, 37:127–53. Academic Press, 1989. https://doi.org/10.1016/S0079-6603(08)60697-2.

Sandegren, Linus, and Britt-Marie Sjöberg. "Self-Splicing of the Bacteriophage T4 Group I Introns Requires Efficient Translation of the Pre-mRNA In Vivo and Correlates with the Growth State of the Infected Bacterium." *Journal of Bacteriology* 189, no. 3 (February 2007): 980–90. https://doi.org/10.1128/jb.01287-06.

Sanson, Bénédicte, and Marc Uzan. "Post-Transcriptional Controls in Bacteriophage T4: Roles of the Sequence-Specific Endoribonuclease RegB." *FEMS Microbiology Reviews* 17, no. 1–2 (August 1, 1995): 141–50. https://doi.org/10.1111/j.1574-6976.1995.tb00196.x.

Snopek, T J, W B Wood, M P Conley, P Chen, and N R Cozzarelli. "Bacteriophage T4 RNA Ligase Is Gene 63 Product, the Protein That Promotes Tail Fiber Attachment to the Baseplate." *Proceedings of the National Academy of Sciences* 74, no. 8 (August 1977): 3355–59. https://doi.org/10.1073/pnas.74.8.3355.

Snyder, Larry, and Hye-Jeong Tarkowski. “The N Terminus of the Head Protein of T4 Bacteriophage Directs Proteins to the GroEL Chaperonin.” *Journal of Molecular Biology* 345, no. 2 (January 14, 2005): 375–86. https://doi.org/10.1016/j.jmb.2004.10.052.

Söding, Johannes, Andreas Biegert, and Andrei N. Lupas. “The HHpred Interactive Server for Protein Homology Detection and Structure Prediction.” *Nucleic Acids Research* 33, no. Web Server issue (July 1, 2005): W244–48. https://doi.org/10.1093/nar/gki408.

Uzan, Marc. “[38] - Bacteriophage T4 RegB Endoribonuclease.” In *Methods in Enzymology*, edited by Allen W. Nicholson, 342:467–80. Ribonucleases - Part B. Academic Press, 2001. https://doi.org/10.1016/S0076-6879(01)42567-5.

Winter, R B, L Morrissey, P Gauss, L Gold, T Hsu, and J Karam. “Bacteriophage T4 regA Protein Binds to mRNAs and Prevents Translation Initiation.” *Proceedings of the National Academy of Sciences* 84, no. 22 (November 1987): 7822–26. https://doi.org/10.1073/pnas.84.22.7822.

Wolfram-Schauerte, Maik, Nadiia Pozhydaieva, Madita Viering, Timo Glatter, and Katharina Höfer. “Integrated Omics Reveal Time-Resolved Insights into T4 Phage Infection of E. Coli on Proteome and Transcriptome Levels.” *Viruses* 14, no. 11 (November 12, 2022): 2502. https://doi.org/10.3390/v14112502.

Yap, Moh Lan, and Michael G Rossmann. “Structure and Function of Bacteriophage T4.” *Future Microbiology* 9, no. 12 (December 2014): 1319–27. https://doi.org/10.2217/fmb.14.91.

Young, E T, R C Menard, and J Harada. “Monocistronic and Polycistronic Bacteriophage T4 Gene 23 Messages.” *Journal of Virology* 40, no. 3 (December 1981): 790–99. https://doi.org/10.1128/jvi.40.3.790-799.1981.

Zinoviev, V. V., A. A. Evdokimov, Y. A. Gorbunov, E. G. Malygin, V. G. Kossykh, and S. Hattman. “Phage T4 DNA [N6-Adenine] Methyltransferase: Kinetic Studies Using Oligonucleotides Containing Native or Modified Recognition Sites.” *Biological Chemistry* 379, no. 4–5 (1998): 481–88. https://doi.org/10.1515/bchm.1998.379.4-5.481.

Experimental Model and Subject Details

The strain of *E. coli* used for all experiments was MG1655. For experiments in liquid media, either Luria broth (LB) or M9 media were used; the recipe for 1 L of M9 media is as follows: 500 mL 2x M9 salts, 2 mL 1 M $MgSO_4$, 200 µL 0.5 M $CaCl_2$, 10 mL 10% Casamino acids, 10 mL 40% glycerol, and water up to 1L). Glucose was added to a final concentration of 0.4% where stated. Antibiotics were used at the following concentrations: carbenicillin at 100 µg/mL, chloramphenicol at 20 µg/mL, kanamycin at 50 µg/mL.

Method Details

Phage propagation - Top agar was melted and then cooled enough to be handled while not killing the bacteria (but not so much that it hardened). This was mixed with 1 µL of phage stock, 99 µL of LB, 100 µL of *E. coli* overnight culture. This mixture was incubated for 6 to 8 minutes, then added to 5 mL of melted and cooled top agar in a glass culture tube, mixed well, and poured onto LB plates. These plates were incubated overnight at 37˚C. The next day, an overnight culture of *E. coli* was back diluted and the new culture was grown up to OD_{600} 0.3. A pipette tip was touched to a single plaque and ejected into the culture, which was then incubated overnight. The next day, 10% volume chloroform was added to the infected culture, incubated for 10 minutes, then pipetted off the culture from the chloroform into a fresh falcon tube. This was spun down at 8000 rpm for 6 minutes, pipetted off the supernatant into a fresh falcon tube, and filtered through 0.22 µm filters. This stock was then stored at 4˚C for further use.

Cell harvesting - Five different flasks containing 300 mL of M9 minimal + 0.4% glucose media were prewarmed to 30˚C, and then each flask was inoculated with an overnight culture of *E. coli* so that it was at an OD_{600} of 0.001. The cultures then grew on a shaker at 30˚C, during which time the filter apparatus, filters, and reagent diggers needed for cell harvest were brought into the 30˚C room. As the OD_{600} neared 0.3, the filter apparatus was assembled and liquid nitrogen was dispensed into 50 mL labeled conical tubes (one per flask) with about six holes poked through their caps with a syringe, which were then stored in an insulated box that was also filled with liquid nitrogen. As the cells reached OD 0.3, they were inoculated with T4 at a multiplicity of infection (MOI) of 5. The cells were collected by pouring the culture through the filter apparatus attached to a vacuum line, scraping the cells off of the filter with a reagent digger, immediately submerging the digger in one of the conical tubes full of liquid nitrogen, and scraping cells off the digger and into the tube using a pre-chilled spatula. The cells were collected from the flasks one at a time, at 2, 5, 10, 15, and 20 minutes. Once the cells were collected from all five flasks at the five different time points, all 50 mL conical tubes were capped and tilted to pour out the liquid nitrogen, then transferred immediately to a -80˚C freezer for storage until ready to proceed with Rend-seq or ribosome profiling.

Mixermilling - A 1.5 mL lysis buffer was prepared for each sample (30µL of 1M Tris pH 8.0, 150 µL of 1M NH_4Cl, 15 µL of 1M $MgCl_2$, 24 µL of 25% Triton X-100, 7.5 µL of 20% NP-40, 10 µL of 155 mM chloramphenicol, 15 µL of 10 U/µL RNase-free DNase, and 1248.5 µL of DEPC H_2O). liquid nitrogen was into a 50 mL conical tube with holes poked through its cap, and 650 µL of lysis buffer was pipetted into the conical tube in order to form lysis buffer pellets, taking care to not position the pipette tip too close to the liquid nitrogen in order to avoid the buffer freezing inside the tip. The 10 mL mixermill cannister (Retsch) was cleaned with EtOH and then Millipore water, dried using paper towels and compressed air (any remaining water can freeze the two halves of the cannister to freeze together, preventing them from opening after mixermilling), and then added to liquid nitrogen to pre-chill them. The cannister was removed from the liquid nitrogen and filled with the combined lysis pellets and sample of flash frozen cells in the cannister, along with the metal ball required to break up the sample. This was mixermilled inside the cannister (using a Qiagen Tissuelyzer 2) five times at 15 Hz for 3 minutes each time, re-cooling the cannister in liquid nitrogen between each round. After completing this, the cannister was opened and each half was placed, open side up, into a shallow water bath to thaw. As soon as the sample inside thawed, the lysate was immediately transferred to a 1.5 mL tube and placed on ice. The lysate was centrifuged at 20,000 RPM at 4˚C for 10 minutes, after which it was transferred to a new tube very carefully as the pellet is very soft. To measure the concentration of RNA in the sample, the supernatant was diluted 1:100 in 10 mM Tris 7.0 and measured A_{260} with Nanodrop (after blanking with 1:100 dilution of lysis buffer) then the concentration of RNA was calculated in the undiluted sample (1 A_{260} = 40 μg/mL). There should be roughly 1 mg of RNA. Each sample was split in two, with 0.5 mg to be used for ribosome profiling, and the rest set aside for Rend-seq and stored at 4˚C.

Ribosome Profiling – For footprinting, the part of the lysate taken for ribosome profiling (0.5 mg of RNA) was centrifuged and combined with 4 μL of MNase (750 U), 5 μL of

Superasin (100 U), 1 µL CaCl2 (1M), and enough lysis buffer to being the solution volume to 200 µL. This was incubated at 25˚C for 1 hour, then quenched by adding 2.4 µL EGTA (0.5 M) to the 200 µL, and left on ice while preparing the sucrose solutions. There were two sucrose solutions made: one 10% sucrose solution and one 55% sucrose solution. For these solutions there was either 1.5 g of sucrose (for the 10% solution) or 8.25 g (for the 55% solution) of sucrose added to 300 µL Tris 8.0 (1 M), 1.5 mL of NH4Cl (1M), 150 µL MgCl2 (1M), 97.5 µL chloramphenicol (155 mM), and enough DEPC H2O to bring the solution to 15 mL. The ultracentrifuge was pre-chilled to 4˚C, and as well as the rotor and buckets. Roughly 6 mL of the 55% sucrose solution was layered below roughly 7 mL of 10 µL sucrose solution in Seton tubes for the SW41 rotor. The sucrose gradient was created using the 7-47% setting on the gradient maker, then each sample was loaded onto the top of one of the sucrose gradient tubes, as tubes that would be across from each other in the rotor with lysis buffer were balanced. The outside of each tube was wiped with an ethanol-soaked wipe (to remove any sucrose that may have gotten onto the side of the tube, as this sucrose could cause the tube to become stuck in the buckets). The tubes were lowered into the buckets of the rotor and the buckets were attached to the rotor before slowly lowering the rotor into place in the centrifuge. The ultracentrifuge was run at 35,000 rpm for 2.5 hours. After this time, one bucket at a time was removed for fractionation, where the tube was fractionated at 0.2 mm/s, measuring A_{260} every second. The monosomes were collected (roughly 1.5 mL) in 2 mL cryovials, then added to liquid nitrogen to flash freeze them before storing at -80˚C.

For RNA isolation, two volumes of 0.75 mL of acid phenol chloroform per sample were prewarmed to 65˚C, taking just the bottom phase of the acid phenol chloroform as the top is buffer. 80 µL of 20% SDS was added to 1.4 mL of fractionated monosomes, then this two tubes was split in two with 0.75 mL each and 0.75 mL of pre-warmed acid phenol chloroform was added to each sample. These were incubated these at 65˚C for 5 minutes at 1400 rpm in a thermomixer and then chilled on ice for 5 minutes. These samples were spun at 20,000 g for 2 minutes and the top aqueous layer was transferred to a fresh tube. To this fresh tube, 0.7 mL acid phenol chloroform was added and the tube was incubated at room temperature for 5 minutes with the occasional vortex, then spun at 20,000 g for 2 minutes and the top aqueous layer was transferred to a fresh tube. To this fresh tube, 600 µL chloroform was added and the tube was vortexed for 30 seconds at room temperature, then spun at 20,000 g for 1 minute and the top aqueous layer was transferred to a fresh tube. This solution was precipitated by adding 78 µL of 3 M NaOAc (pH 5.5) and 0.75 mL isopropanol. This was vortexed and then chilled at -80˚C for 30 minutes. The samples were spun at 20,000 g for 30 minutes at 4˚C, supernatant was removed, pellet was washed in 750 µL 80% EtOH at 4˚C, and resuspended the samples in 11 µL 10mM Tris 7.0, pooling the two tubes for each sample together, resulting in one tube of RNA per sample.

For size selection, a 1:10 dilution of total RNA in 10 mM Tris 7.0 was made and each sample was quantified by Nanodrop, then diluted to 20 µg in 5 µL 10 mM Tris 7.0. A 15% TBE-Urea PAGE gel (Invitrogen, 10 well) was pre-run in 1x TBE at 200 V for 1 hour. 5 µL of 2x TBE-Urea sample loading buffer was added to each sample, o199-P oligo (1 µL of 20 µM O199-P, 3 µL of 10 mM Tris 7.0, and 5 µL of 2x loading buffer) was prepared as well as ladder (0.5 µL of 10 bp ladder, 4 µL of 10 mM Tris 8.0, and 5 µ: of 2x loading buffer). All samples were denatured at 80˚C for 2 minutes before returning to ice. The samples were run on the gel at 200 V for 65 minutes, stained with SYBR gold for 2 minutes, and the gel was photographed. The gel was transferred onto a UV box and a razor blade was used to excise between 15-45 bases from the gel, then the gel slice was placed into a 0.5 mL tube with the bottom poked through. The 0.5 mL tube was put in a 2 mL screw cap tube, spun at 20,000 g for 3 minutes, and the remaining gel pieces were collected, added 0.5 mL 10 mM Tris 7.0, and shaken at 70˚C for 10 minutes on the thermomixer. The samples were transferred to Spin-X cellulose acetate columns, spun at 20,000 g for 3 minutes, and the elute was transferred to a new tube. To precipitate this elute, 55 µL 3 M NaOAc (ph 5.5), 2 µL glycoblue, and 550 µL isopropanol were added and the solution was vortexed. This was incubated at -80˚C for 30 minutes, spun at 20,000 g for 30 minutes at 4˚C, and the supernatant was removed, then it was washed with 750 mL of 80˚C EtOH at 4˚C and resuspended in 15 µL 10 mM Tris 7.0.

For dephosphorylation, 15 µL of each sample was mixed with 2 µL of T4 PNK buffer w/o ATP, 1 µL of Superasein, and 2 µL of T4 PNK. This mixture was incubated at 37˚C for 1 hour, heat inactivated at 75˚C for 10 minutes, then precipitated by adding 500 µL of 10 mM Tris 7.0, 55 µL of 3 M NaOAc, 2 µL glycoblue, and 550 µL isopropanol, and vortexed. This was chilled at -80˚C for 30 minutes, pelleted at 20,000 g for 30 minutes at 4˚C, washed with 750 µL of 80% EtOH, resuspended in 7 µL of 10 mM Tris 7.0, and transferred to a new tube.

For ligation and size selection, 1 pmol of RNA was diluted to 6 µL in 10 mM Tris 7.0, using the qubit to determine concentration. This was denatured at 80˚C for 2 minutes before being returned to ice. Being sure to stay at room temperature or above, a solution was made containing 6 µL of RNA, 10 µL PEG 8000, 2 µL of 10x T4 RNA ligase 2 buffer, 1 µL of linker-1 (0.1 mM), and 1 µL of T4 ligase 2 truncated. This was incubated at 25˚C for 2.5 hours, then precipitated by adding 0.5 mL of mM Tris 7.0, 55 µL of 3 M NaOAc (pH 5.5), 2 µL of glycoblue, and 550 µL of isopropanol. This was vortexed and then stored at -80˚C for 30 minutes. This was pelleted at 20,000 g for 30 minutes at 4˚C, washed with 750 µL of 80% EtOH, and resuspended in 6 µL of 10 mM Tris 7.0. 6 µL of 2x TBE-urea loading buffer was added and the solution was denatured at 80˚C for 2 minutes, run on a 10% TBE-urea gel for 50 minutes at 200 V, and excised from 35-65 bases. This was precipitated, resuspended in 10 µL of 10 mM Tris 7.0, and transferred to a new tube.

For reverse transcription and size selection, a mixture was created containing 10 µL of RNA, 1 µL of oCJ485 (25 µM), and 1.5 µL of DEPC H2O. This was denatured at 65˚C for 5 minutes and placed back on ice. A mixture of 4 µL 5x FSB buffer, 1 µL of Superasein, 1 µL of DTT (0.1 M), and 1 µL dNTP (10 mM) was added to sample, followed by 1 µL of Superscript III. This was incubated at 50˚C for 45 minutes, then quenched by adding 2.3 µL of 1 M NaOH to hydrolyze RNA, and incubated at 95˚C for 15 minutes. 23 µL of TBE-urea loading buffer was added then the sample was denatured at 70˚C for 2 minutes, and run on a 10% TBE-urea gel at 200V at 80 minutes, using 2 wells per sample. Products of expected size were excised, being careful to avoid carrying over the RT primer, then the gel was crushed as before but using 500 μL of Tris 8.0 instead of Tris 7.0. This was passed through Spin-X cellulose acetate columns and then precipitated with 32 μL 5 M NaCl, 1 μL 0.5 M EDTA, 2 μL glycoblue, and 550 μL isopropanol. This was chilled for 30 minutes at -80˚C, spun at 20,000 g for 30 minutes at 4˚C, washed with 750 μL EtOH, resuspended in 15 μL 10 mM Tris 8.0, and transferred to a new tube.

For circularization, the following were combined: 15 μL DNA, 1 μL 1mM ATP, 2μL 10x CircLigase buffer, 1 μL 50 mM $MnCl_2$, 1 μL CircLigase. This mixture was incubated at 60˚C for 1 hour, after which another 1 μL of CircLigase was added, and then was incubated for 1 more hour at 60˚C. Following this, the reaction was inactivated at 80˚C for 10 minutes before returning the mixture to ice.

For PCR and size selection, a mixture was made from 17 μL 5x HF buffer, 1.7 μL 10 mM dNTP, 4 μL 10 μM o231, 4 μL 10μM indexing primer, 52.3 μL μL DEPC water, 1 μL HF Phusion, 5 μL DNA. 17 μL of the PCR mix was aliquoted into 4 different PCR strip tubes and run with the initial step being 98˚C for 30s, denaturation 98˚C for 10s, annealing 60˚C for 10s, and extension 72˚C for 5s. Each strip was removed after 6, 8, 10, 12 cycles, respectively. Then 3.5 μL 6X DNA loading dye was added to each sample and the ladder was set up as well with 1 μL 10 bp ladder, 16 μL 10 mM Tris 8, 3.5 μL 6x DNA loading dye. The samples and ladder were run on an 8% TB PAGE gel at 180 V for

45 minutes, and extracted from the gel, and the products were resuspended in 11 µL of 10 mM Tris 8.

Oligo	Nucleotide sequence
o199-P	5’_AUGUACACGGAGUCGACCCGCAACGCGA3phos_3’
Linker-1	5'_App/CTGTAGGCACCATCAAT/3ddC_3'
ocj485	/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT/iSp18/CAAGCAGAAGACGGCATACGAGATATTGATGGTGCCTACAG
Indexing primer	5'_aatgatacggcgaccaccgagatctacacgatcggaagagcacacgtctgaactccagtcacNNNNNNacactctttccctacac_3'
o231	5'_CAAGCAGAAGACGGCATACGA_3'

Rend-seq – For tRNA removal, phenol chloroform extracted RNA was diluted it to 100 µL. Then 350 µL of RLT/BME and 250 µL of 100% EtOH was added to the sample. The sample was added to an RNeasy column and the standard RNeasy Mini Kit instructions were used from there, eluting the sample in 85 µL of DEPC-H_2O. To this, 10 µL of TurboDNase buffer and 5 µL of TurboDNase were added and then the mixture was incubated at 37˚C for 20 minutes. The RNA was cleaned up using RNA Clean & Concentrator-5 kit from Zymo, and eluted in 30 µL of DEPC-H_2O.

For rRNA removal, the MICROBExpress Kit was used, resuspending the pellet in 40 µL Tris 7.0.

For RNA fragmentation, the 40 µL of RNA from the rRNA removal step was incubated at 95˚C for 2 minutes in a PCR block, and then placed on ice for >1 minute. Following this, 4.4 µL of 10X fragmentation buffer was added on ice, and then fragmented for 25 seconds in the PCR block that was preheated to 95˚C. After this, 5 µL stop buffer was added, then the sample was mixed well and quickly placed back on ice. This was done

one sample at a time due to the very short fragmentation time. The sample was diluted with 500 µL 10 mM Tris 7, 55 µL 3M NaOAc (pH 5.5), 2 µL glycoblue, 550 µL isopropanol, and then vortexed to mix. The samples were chilled at -80˚C for >30 minutes, and then the RNA was pelleted at 20,000 g for 1 hour at 4˚C. The supernatant was carefully removed and the pellet was gently washed with 0.8 mL ice cold 80% ethanol, then resuspended the pellet in 5 µL 10 mM Tris 7 and transferred to a new tube. For size selection, a 15% TBE-Urea PAGE gel (Invitrogen, 10 well) was pre-run in 1x TBE at 200 V for 1 hour. 5 µL 2x TBE-Urea sample loading buffer was added to each sample, as well as to oligo o199-P and a 10 bp ladder. All samples were denatured at 80˚C for 2 minutes and returned to ice. Then, the samples were run on the gel at 200 V for 65 minutes, stained with SYBR gold for 2 minutes, and photographed. The gel was excised between 15-45 bases and the gel slice was transferred into a 0.5 mL tube (with the bottom poked through with a needle) which as placed into a 2 mL screw cap tube. The samples were spun at 20,000 g for 3 minutes, then the remaining gel pieces were collected and added to 0.5 mL 10 mM Tris 7, then shook at 70˚C for 10 minutes. After this, the samples were transferred to Spin-X cellulose acetate columns and spun at 20,000 g for 3 minutes, then the elution was transferred to a new tube and precipitated by adding 55 µL 3 M NaOAc (pH 5.5), 2 µL glycoblue, and 550 µL isopropanol. This was vortexed to mix, then incubated at -80˚C for >30 minutes. Then, the samples were spun at 20,000 g for 30 minutes at 4˚C, the supernatant was removed, the pellet was washed with 750 µL 80 ethanol at 4˚C, and resuspended in 15 µL 10 mM Tris 7.

Dephosphorylation was performed as described above for ribosome profiling.

For ligation and size selection, 3 pmol RNA was diluted to 6 µL in 10 mM Tris 7 and denatured at 80˚C for 2 minutes before the samples were returned to ice. The following were mixed by pipetting: 6 µL RNA, 10 µL PEG 8000 50%, 2 µL 10x T4 RNA ligase 2 buffer, 1 µL Linker-1 0.1 mM, and 1 µL T4 ligase 2 truncated. This was incubated at 25˚C for 2.5 hours and precipitated by adding 0.5 10 mL Tris 7, 55 µL NaOAc pH 5.5, 2 µL glycoblue, and 550 µL isopropanol, then vortexed to mix. This was chilled at -80˚C for >30 minutes. Then, the samples were spun at 20,000 g for 30 minutes at 4˚C, the supernatant was removed, the pellet was washed with 750 mL 80 ethanol at 4˚C, and resuspended in 6 µL 10 mM Tris 7. 6 µL 2x TBE-urea loading buffer was added and the solution was denatured at 80˚C for 2 minutes, then run on a pre-run 10% TBE-Urea PAGE gel at 200 V for 50 minutes. This was gel-purified from 35-65 bases, then was precipitated and resuspended in 10 µL 10 mM Tris 7 and transferred to a new tube.

For reverse transcription and size selection, the following were combined: 10 µL RNA, 1 µL oCJ485 25 µM, and 1.5 µL DEPC water, then this was denatured at 65˚C for 5 minutes and placed back on ice. The following were then added to each sample: 4 µL 5x FSB buffer, 1 µL Superasein, 1 µL DTT 0.1 M, and 1 µL dNTP 10 mM. Following this, 1 µL of Superscript III was added and incubated at 50˚C for 45 minutes. The reaction quenched by adding 2.3 µL 1 M NaOH to hydrolyze RNA and incubated at 95˚C for 15 minutes. 23 µL TBE-urea loading buffer was added, then the solution was denatured at 70˚C for 2 minutes, and run on a 10% TBE-urea gel at 200 V for 80 minutes (2 wells per sample). Products of expected size were excised, being careful to avoid carrying over RT primer. The gel was crushed as before, but using 500 µL Tris 8 instead of Tris 7. The solution was then passed through Spin-X cellulose acetate column, precipitated with 32 µL 5 M NaCl, 1 µL 0.5 EDTA, 2 µL glycoblue, and 550 µL isopropanol. The samples were spun at 20,000 g for 30 minutes at 4˚C, the supernatant was removed, the pellet was washed with 750 µL 80 ethanol at 4˚C, and resuspended in 15 µL 10 mM Tris 8.

Circularization was performed as described above for ribosome profiling.

PCR and size selection was performed as described above for ribosome profiling.

***61.-1* Cloning**

The portion of the T4 genome containing *61.-1* was amplified using primers that encompassed the entire *61.-1* transcript (based on our Rend-seq data), with ~10 ng of T4 genomic DNA as the template DNA, using Q5 polymerase. In parallel, restriction enzyme double digest was used to linearize vector pBAD18, using enzymes EcoRI-HF and HindIII-HF. Both the amplified section of T4 genomic DNA as well as the digested plasmid were run on a 1% agarose gel at 120 V for 30 minutes. A gel extraction was performed for each of the bands (*61.-1* region of the genome and digested plasmid). For this, the DNA fragment from the agarose gel was extracted using a razor blade and then was transferred to a 1.5 mL tube. 3 volumes of ADB were added to each volume of agarose excised from the gel (for each 100 mg of gel slice, 300 µL of ADB was added) and the mixture was incubated at 50˚C for 5-10 minutes until the gel slice dissolved completely. The melted agarose solution was transferred to a Zymo-Spin I Column in a

Collection Tube, centrifuged at 10,000 x g for 60 seconds and the flow-through was discarded. 200 µL of Wash Buffer was added to the column and it centrifuged at 10,000 x g for 30 seconds. The flow-through was discarded, then the wash step was repeated. Then ≥ 6 µL of water was added directly to the column matrix, the column was placed in a 1.5 mL tube and centrifuged at 10,000 x g for 60 seconds to elute the DNA. A Gibson assembly was performed with the vector and insert to join them together. The mixture was transformed into Zymo Mix & Go Competent Cells (strain DH5α), ~1 mL of the mixture was outgrown in SOB at 37˚C for ~1 hour, then plated 50 µL on LB + carb plates and grown at 37˚C overnight. The following day, overnight liquid cultures were made from single colonies on the plate. The following day, glycerol stocks were created from each overnight culture, then the plasmids from each overnight culture were miniprepped as well. The Zymo DNA Clean and Concentrator kit was used to purify the plasmids and then they were submitted for sequencing. Once the sequencing verified that the plasmid was correctly assembled, the plasmid was transferred from DH5α cells to lab strain MG1655 *E. coli* using TSS transformation. To create the competent MG1655 cells, a 1x TSS solution was created, composed of 5 g PEG 3350, 1.5 mL 1M MgCl2 and LB to 50 mL, which was filer sterilized using a 0.22 µm filter then added 2.5 mL of DMSO. The MG1655 culture was then grown to an OD_{600} of 0.3-0.5 in 50 mL of LB. Once the correct OD was reached, the culture was placed on ice for 15 minutes, then the cells were spun down at 5000 x g at 4˚C, placed on ice and the supernatant was removed. The cells were resuspended in 2 mL of 1 x TSS solution, incubated on ice for 20 minutes, and made into 50 µL aliquots of cells in 1.5 mL tubes, which were stored at -80˚C for future use. To perform the actual transformation, competent MG1655 cells were thawed on ice, 10-20 ng of plasmid was added, the tube was gently mixed, and incubated on ice for 20 minutes. After this, 1 mL of LB was added and the cells were recovered at 37˚C for 1 hour, then 50 µL was plated on warmed selective plates. The following day individual colonies were picked and grown them overnight in LB + carb. The next day glycerol stocks were made from these overnight cultures.

Growth Curves

An overnight culture of the MG1655 + pBAD18/*61.-1* cells as well as control MG1655 + pBAD18 cells were back-diluted to an OD_{600} of 0.001 in 50 mL of M9 minimal media + 0.4% glucose. These cultures (three cultures with control cells and three cultures with *61.-1* cells) were incubated at 30˚C, shaking at 225 rpm. At three different OD600s (0.01, 0.1, and 0.3), one control culture and one *61.-1* culture were taken, filtered using rapid filtration (described in Rend-seq section), and the cells were resuspended in M9 media + 0.1% arabinose (in order to induce expression of *61.-1*). On a separate occasion a growth curve was performed using MG1655 + pBAD18/*61.-1*trx, with *61.-1*trx referring to the fact that *61.-1*'s coding sequence was mutated with all ATGs changed to GTAs to ensure that the transcript of *61.-1* would exist, but no proteins were created from it.

CRISPR

The Genious CRISPR guide tool was used to pick guides that would target and knock out *61.-1* in T4 once the CRISPR was completed. The primers for each guide were inserted into the pCas9 plasmid by phosphorylating the primer pairs with T4 PNK, digesting pCas9 with BsaI restriction enzyme and heat inactivating at 80˚C for 20 minutes, annealing primers into pCas9 using T4 DNA ligase, transforming each reaction (1 per guide) into a tube of Mix & Go cells, and plating on LB + cam plates. The next day overnight cultures were grown from each guide RNA plate. The following day the plasmids from the overnight cultures were miniprepped and sent out for sequencing. The three plasmids with the three different guides that had successfully transformed into Mix & Go cells were then transformed into MG1655 TSS competent cells. The next day overnight cultures of MG1655 + guide in LB + cam were made. The following day a phage spotting assay was performed with the MG1655 + guide strains. For this, three LB plates (one for each strain that harbored a separate guide) were warmed at 37˚C. Top agar was melted and 8 mL was added to each of three glass vials, cooled enough to be held comfortably while still molten, and 1/50th of the overnight culture was added. Each vial was mixed and poured on top of their respective warmed LB plates. Then 1x FM buffer was added (9 µL per plate, ex: if plating on three separate plates, use 27 µL) to columns 1-6 of a qPCR plate, with however many rows as types of phage being used. Phage stock (1 µL per plate) was added to the first well of the column, and repeated for all phage, only keeping one stock at a time open and cleaning the pipette in between to avoid stock contamination. Serial dilutions were performed out to the 6th column of the plate, then all phage were spotted on the plates, going from most dilute (column 6) to least dilute (column 1). The plates dried and were put them (with the agar on the bottom) at 37˚C overnight. In order to design rescue plasmids, a gBlock (Integrated DNA Technologies) was ordered containing the transcript of *61.-1* but with the coding sequence on *61.-1* deleted up to the overlap with *61* and then the overlap with *61* recoded with different nucleotides to produce the same amino acids in the translation of *61*. Then pRescue was amplified using primers to create an overlap with the gBlock, then a Gibson assembly was performed to introduce our gBlock into pRescue. The product was on a 1% agarose gel at 120 V for 30 minutes, then the band of interest was gel extracted, transformed into Mix & Go cells, and ~1 mL of the mixture was outgrown in SOB at 37˚C for ~1 hour, then plated 50 µL on LB + kan plates and grown at 37˚C overnight. The plasmid was then miniprepped and transferred into a competent strain of *E. coli* containing pCas9. This was outgrown for ~1 hour and plated on LB + cam + kan plates and grew at 37˚C overnight. A phage spotting assay was performed with the strain of *E. coli* containing pCas9 + pRescue, using wild-type *E. coli*. To amplify the phage created by the CRISPR mutant that was generated in the phage spotting assay, a culture of *E. coli* with pCas9 was grown to an OD of 0.3, a pipette tip was touched to a single plaque on the plate from the spotting assay, the tip was ejected into the culture, and the flask was incubated overnight. The next day the infected cultures were spun down at 3000 x g and filtered through a 0.22 µm filter to get a phage stock. The mutation in this new phage (Δ*61.-1*) was confirmed through PCR and a gel as well as through sequencing.

Microscopy

Two different flasks containing 50 mL of LB media to 37˚C were prewarmed, then one flask with was inoculated with an overnight culture of *E. coli* containing the empty vector pBAD18 and one flask with *E. coli* containing *61.-1* in the pBAD18 vector so that each flask was at an optical density OD_{600} of 0.001. The cultures grew on a shaker at 37˚C until they reached an OD of 0.3. At this point, 0.5 µL of each culture was pipetted onto an agar pad (M9L agar with arabinose and propidium iodide). The agar pad was placed into a microscope kept at 37˚C and was imaged at different locations every 10 minutes using both phase and the propidium iodide filter.

One-Step Growth Curves

An overnight culture of wild-type *E. coli* was diluted in 50 mL of LB with 2 mM of $CaCl_2$. Then 8 flasks (2 Adsorption Flasks, 2 Flask A's, 2 Flask B's, and 2 Flask C's) were laid out. Once the culture hit an OD of 0.3, 9.9 mL of LB was pipetted into 2 Flask A's, 9 mL into 2 Flask B's, and 9 mL into 2 Flask C's (one set of flasks for wild-type T4 and one set for Δ*61.-1*). These flasks were kept at 37˚C. 40 15 mL glass tubes were prepared with 5 mL of molten top agar each (kept in bead bath at 42˚C to keep molten), 40 LB agar plates were labeled for specific samples, and phage dilutions were prepared in glass tubes (need 100 μL of phage that is 10^7 PFU/mL). 9.9 mL of the *E. coli* culture (at OD 0.3) was added to each of the two Adsorption Flasks, then 100 μL of each phage was added to its respective Adsorption Flask. This mixture was incubated for 5 minutes at 37˚C while shaking at 200 rpm. Then 100 μL was transferred from the Adsorption Flask to Flask A for each phage. 1 mL from Flask A (after mixing thoroughly) was then added to a glass tube containing 30 μL of chloroform, which was then vortexed and placed on ice to serve as out adsorption control. 1 mL from Flask A was added to Flask B, mixed well, then 1 mL from Flask B was added to Flask C and mixed well. At each time point, 100 μL was removed from the designated flask, pipetted into a prepared top agar flask, and 100 μL of overnight *E. coli* culture were added before plating on an LB plate. The time points were laid out as such: Flask A was sampled at 5, 10, 15, 20, 25, 30, and 35 minutes; Flask B was sampled at 25, 30, 35, 40, and 50 minutes; Flask C was sampled at

35, 40, 50, 60, 70, 80, and 90 minutes. Once the samples from all time points were collected and plated, each adsorption control was plated onto plates using the same method as above. These plates were incubated overnight and the next day the plaques on each plate were counted.

Quantification and Analysis

Gene Expression- In order to calculate expression levels throughout the genome, RPKMs were calculated for each gene. In order to assign specific peaks to specific genes, the Rend-seq data was examined for a peak corresponding to the time point of that gene's maximal expression. The 500 base pairs upstream of a gene's coding sequence or up to the end of the upstream gene's coding sequence if that occurred at less than 500 base pairs upstream, because at that point the peaks may be related to the UTR of the upstream gene rather than the gene of interest, was searched. Using the methods described in Parker 2023, prominent peaks in that area were identified, which were then manually screened. Peaks were rejected if they were potential RegB cleavage sites or within a shadow from a 3′ peak, or if there were multiple peaks within the search region that caused uncertainty on the peak corresponding to the transcript end.

Splicing - For X and Y (the unspliced reads), 15 nucleotides on either side of the splice junction (not removing the section that would be spliced out) were taken and that 30 nucleotide sequence was searched for in the fasta files for each time point, converting read counts to Reads Per Million (RPMs). For Z (spliced reads), 15 nucleotides upstream of the first splice junction and 15 nucleotides downstream of the second splice junction (removing the section that is spliced out) were taken and that 30 nucleotide sequence was searched for in the same way as described above. Comparing the RPMs for X and Y vs. Z gives an idea of the splicing efficiency for each spliced gene.

Promoters - All identified 5′ peaks were taken and then the first part of the early or middle promoter consensus sequence was searched for as well as the second part of the consensus sequence. If these two sequences appeared in the correct spacing from each other and from the 5′ peak, it was labeled it as a potential promoter.

RegB Cleavage Sites - The entire genome was swept for GGAG sequences (the known RegB cleavage motif. If a potential RegB cleavage site contained an identified 5′ peak that was had early peak expression timing, it was labeled as a potential RegB cleavage site.

Proportional Synthesis - Using the available literature, a list was compiled of nine known protein complexes found in T4. For each, the stoichiometry in the complex from the literature was compared to the RPKMs of each gene from our dataset.

***61.-1* Homolog Comparisons** - NCBI protein BLAST tool was used to search the amino acid sequence of *61.-1*. From the results, five entries were picked that contained proteins similar in sequence to *61.-1*, capturing a variety of different species to demonstrate the diversity of organisms this gene can be found in. The clinker python package was then used to create a comparison diagram of the homologs and their surrounding genes.

One-Step Growth Curves - Data was normalized by multiplying the number of plaques on the Adsorption Control plates as well as on the Flask A plates by 10, the plaques on Flask B plates by 100, and plaques on Flask C by 1000 to get PFU/mL. All data was plotted in a scatter plot, which revealed an S-shaped curve. The average of the bottom line of points (Average 1) minus the number of plaques in the Adsorption Control gives the average number of infected cells. This number, divided by the average of the top line of points in the curve (Average 2), gives the average burst size for the phage. The intersection between the Average 1 line and the slope reveals the latent period for the phage. Four replicates of this experiment were performed and the mean of the four experiments was plotted, with the standard deviation shown in a bar passing through the dot corresponding to the mean.

Chapter 3: Conclusions and Future Directions

CONCLUSIONS

Exploring phage in the laboratory has had significant impacts on biology as a whole, and there are certainly many more discoveries to be made using phage. In my graduate work, I focused on T4 phage, and exploring both its transcriptional and translational profiles throughout the course of its infection in the hopes of uncovering new insights that could possibly be applicable beyond just T4. I believe that understanding gene expression and protein synthesis at the level of phage can provide information that help us explore evolution and better understand diseases that are in dire need of treatments.

From this study, I was able to uncover several pieces of information about T4, some of which I believe are more broadly applicable to biology as a whole. We were able to harvest RNA from T4-infected E. coli at several time points throughout phage infection in order to get a global picture of transcription and translation over time. Our Rend-seq enabled us to view transcript boundaries, while our ribosome profiling revealed levels of protein synthesis of the transcripts. We used this information in order to gain insight into sites of transcriptional regulation, including new potential cleavage sites by the T4-encoded endoribonuclease RegB, as well as new putative promoter sites. From the list of putative promoters, most were assigned to a gene with an already annotated promoter, exhibiting the possibility of complex layers of transcriptional control.

We also gained a view into translational regulation in T4. Using our ribosome profiling signal divided by our Rend-seq signal for each gene, we were able to calculate their translational efficiencies (TEs). Using the changes in TE over time for each gene, we were able to confirm known instances of translational regulation in T4, as well as identify new genes that exhibit changes in TE as great as, or even larger than, the genes already categorized as being regulated. This indicates the presence of additional translational regulation within the T4 genome.

Our ribosome profiling data allowed us to uncover the presence of proportional synthesis in T4. The presence of proportional synthesis indicates that proteins that are part of multi-protein complexes are being produced in the same proportions that they are needed for the complex, avoiding expression of any protein in excess of what is needed. This is notable as proportional synthesis has been noted before in both prokaryotes and eukaryotes, but never in viruses. Even more remarkable, all previous observations have been observed when the organism in question is in steady-state growth. T4 does not exhibit steady-state, as it is constantly changing as it progresses through a cycle of infection.

Finally, we were able to combine both our Rend-seq and ribosome data to identify the presence of a new gene in T4, known as *61.-1*. This gene is located on the forward strand of the genome, in a region containing only reverse strand genes. It is well conserved across many species of phage, not only in its protein sequence, but also in its positioning between neighboring genes that are always on the opposite strand as it. *61.-1* is toxic when ectopically expressed in *E. coli*, though the time of cell death relative to *61.-1* induction is longer when induced at a lower optical density (OD) and shorter when the induction occurs as OD approaches 0.3, which is mid-log phase. This OD-dependent toxicity has not yet been fully comprehended, but is interesting nonetheless. What is known is that *61.-1* is not essential for T4 infection under laboratory conditions, as a CRISPR-generated Δ*61.-1* strain of T4 had no difference in latent period or burst size during an infection, but the exact role of *61.-1* needs further investigation.

Taken together, this thesis has provided insights into many forms of gene expression and regulation in T4 phage over the course of infection. This work has also set the stage for follow-up work on some of the concepts investigating, particularly translational regulation in T4 and the role of *61.-1*. I also believe that this data will be valuable to other members of the phage community studying T4, and will hopefully be used to answer many other questions.

FUTURE DIRECTIONS

Translational Regulation in T4

In our study, we were able to use a combination of our Rend-seq data and our ribosome profiling data to calculate translational efficiency (TE) for each gene throughout the course of T4 infection. TE is calculated by dividing ribosome profiling RPKMs by Rend-seq RPKMs, giving an idea of the rate of protein production per RNA for a gene at a specific time point. This measurement shines light on whether translational regulation is taking place for a certain gene over time; if there is no regulation, then the translational efficiency should remain constant across time points as levels of transcription and translation would stay in roughly the same proportion to each other. However, if there is regulation occurring, then the ratio will change, as there will be a different amount of protein production per RNA at some time points relative to other time points.

For the genes that seem to exhibit large changes in TE, there are numerous ways to investigate the potential mechanisms by which they are translationally regulated. It is possible that some genes exhibit similar forms of translational regulation as those already known in Gp32, Gp43, and RegA. In these cases, the protein of interest bind to its own mRNA or to the mRNA of another gene in order to occlude ribosomes and prevent translation from occurring (Shamoo et al., 1993; Tuerk et al., 1990; Gordon et al., 1999). One possible way to investigate this mechanism of translational regulation in the newly identified genes would be to use SELEX, where a given protein is provided to random oligonucleotides and the binding sequence(s) become enriched and able to be sequenced (Tuerk et al., 1990). It is also possible to use RNA immunoprecipitation (RIP) or cross-linking immunoprecipitation (CLIP) to pull down on RNAs that are bound to specific proteins (Hafner et al., 2021).

Furthermore, it is possible that translational regulation could be occurring through mRNA folding, which can be mediated by many different stimuli, such as temperature, pH, metabolites, or macromolecules (Tollerson et al., 2020). In order to identify RNA structure, one approach that could be used is chemical probing with a probe such as dimethyl sulfate DMS. In this instance, RNAs could be exposed to DMS, which methylates unpaired adenine and cytosines; when the RNAs are then reverse transcribed, an incorrect base will be added in the place of the methylated base (Lempereur et al., 1985; Matthews et al., 2010). These mutations can be sequenced and the sequences can be compared to the true RNA sequence, revealing that any bases with high mutation rates may be unpaired. Performing this assay on the RNAs of interest at different time points during infection could reveal changing structures, which could play a role in translational regulation.

The role of 61.-1

Using the data described in this thesis, we were able to uncover the presence of a novel gene within T4's genome. Visualizations of our Rend-seq data showed the presence of a transcript in an unannotated area, and our ribosome profiling data indicated the presence of translation over the body of that transcript. The gene, named *61.-1*, also had the signs of a classic T4 late promoter, a ribosome binding site (RBS), and an open reading frame (ORF). Furthermore, ectopic expression of this gene in *E. coli* led to cell death, but in an interesting growth-dependent manner. Essentially, inducing *61.-1* expression at a lower optical density (OD) led to a longer delay time between induction and toxicity. Once the OD reached mid-log phase (OD of ~ 0.3), induction led to toxicity without delay. We found this to be extremely interesting, and the question still remains as to why this delay exists, and whether it reflects the protein's function within the T4 lifecycle.

In order to address this, I propose to first carry out additional growth curve experiments, testing induction at an even greater number of ODs. Since we know that toxicity occurs immediately when *61.-1* is induced in cells at an OD of 0.3, it would be interesting to see if the immediacy continues when cells are induced at a higher OD, or if perhaps the time between induction and response begin to separate again. One hypothesis might be that the cells need to be in a particular phase of growth to have a strong response to the expression of *61.-1*. It is clear from our data that the closer the culture gets to log phase growth, the more immediate the toxicity response occurs. However, what is unclear is whether log phase is the optimal time for *61.-1* activity, or if a higher OD simply correlates to a faster response time. If the cells need to be in log phase for *61.-1* to have its effect, this could point to the gene playing a role in sensing growth rate or population density to help T4 lyse at an optimal time for further infection. For example, if T4 were to lyse cells in a culture too early, there may not be enough other hosts for the progeny phage to infect. However, if T4 were to lyse cells in a culture too late, the cells might be in stationary phase and not provide an ideal cellular environment for phage production.

Ectopic expression of *61.-1* in *E. coli* can give hints as to what the role of the gene may be, but another way to approach the question of function is to perform experiments with T4 phage itself. In our study, we used CRISPR to knock out *61.-1* in T4 by removing the section of *61.-1* that does not overlap with any other genes and recoding the section of *61.-1* that overlaps with *61*. We performed one-step growth curves with both wild type T4 and Δ*61.-1* T4, and found that there were no differences in terms of latent period or burst size from the infection. While this does give us some information on what *61.-1* is not doing, there are still many avenues to pursue in terms of understanding how it does function.

To gain insight into the role of *61.-1* in phage infection, it could be interesting to tag *61.-1* with an affinity tag and then, during late infection, perform a pull-down assay in order to see if Gp61.-1 interacts with any other T4 proteins. If there are interactions with other proteins, then examining the roles of these proteins could give a sense as to what the role of Gp61.-1 is in the cell. For example, Gp61.-1 could interact with Gpt, the holin that creates pores in the inner membrane to allow the phage's endolysin to degrade the bacterial cell wall and cause lysis (Ramanculov et al., 2001). In that case, it could be postulated that *61.-1* plays a role in mediating the process of host cell lysis. Another possible experiment would be to tag *61.-1* with a fluorescent tag, and then follow the location of Gp61.-1 using microscopy. It could be interesting to see where in the host cell Gp61.-1 is localized; it could be along the membrane or be spread our throughout the host cell.

Finally, combining the ectopic expression of *61.-1* with T4 infection, it would be fascinating to try infecting *E. coli* that is expressing *61.-1* with wild type T4 phage. As of now, it is unknown whether that infection could be productive or not. It is possible that the level at which *61.-1* is expressed in *E. coli* is much too high and causes cell death because of its quantity, thereby preventing successful T4 infection. However, it could be that the amount of 61.-1 expressed in *E. coli* is able to assist the phage by bolstering the effect that 61.-1 normally has during infection. Although our CRISPR experiment revealed that *61.-1* is not essential under laboratory conditions, it is possible that extra *61.-1* during infection could have a beneficial effect for T4.

CONCLUDING REMARKS

Phage are extremely abundant and fascinatingly complex organisms that toe the line between living and non-living entities. Yet, studying them has revealed many fundamental biological concepts that extend to organisms across the tree of life. And not only do phage serve as excellent model organisms in the laboratory, they also play a major role in microbial ecosystems and have the potential to contribute greatly to advances in medicine and biotechnology.

This study into T4 phage gene expression and regulation has provided valuable insights into the intricate processes that control its lifecycle. We emerge from this work with a deeper understanding of transcriptional and translational regulation, well-conserved processes, and new genetic insights. Not only this, but our data can be utilized by many others to answer remaining question from our study, or to aid in answering completely new questions about phage.

References

Gordon, J., T. K. Sengupta, C. A. Phillips, S. M. O'Malley, K. R. Williams, and E. K. Spicer. "Identification of the RNA Binding Domain of T4 RegA Protein by Structure-Based Mutagenesis." *The Journal of Biological Chemistry* 274, no. 45 (November 5, 1999): 32265–73. https://doi.org/10.1074/jbc.274.45.32265.

Hafner, Markus, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. "CLIP and Complementary Methods." *Nature Reviews Methods Primers* 1, no. 1 (March 4, 2021): 1–23. https://doi.org/10.1038/s43586-021-00018-1.

Lempereur, L, M Nicoloso, N Riehl, C Ehresmann, B Ehresmann, and J P Bachellerie. "Conformation of Yeast 18S rRNA. Direct Chemical Probing of the 5' Domain in Ribosomal Subunits and in Deproteinized RNA by Reverse Transcriptase Mapping of Dimethyl Sulfate-Accessible." *Nucleic Acids Research* 13, no. 23 (December 9, 1985): 8339–57.

Mathews, David H., Walter N. Moss, and Douglas H. Turner. "Folding and Finding RNA Secondary Structure." *Cold Spring Harbor Perspectives in Biology* 2, no. 12 (December 2010): a003665. https://doi.org/10.1101/cshperspect.a003665.

Ramanculov, Erlan, and Ry Young. "Genetic Analysis of the T4 Holin: Timing and Topology." *Gene* 265, no. 1 (March 7, 2001): 25–36. https://doi.org/10.1016/S0378-1119(01)00365-1.

Shamoo, Y., A. Tam, W. H. Konigsberg, and K. R. Williams. "Translational Repression by the Bacteriophage T4 Gene 32 Protein Involves Specific Recognition of an RNA Pseudoknot Structure." *Journal of Molecular Biology* 232, no. 1 (July 5, 1993): 89–104. https://doi.org/10.1006/jmbi.1993.1372.

Tollerson, Rodney, and Michael Ibba. "Translational Regulation of Environmental Adaptation in Bacteria." *Journal of Biological Chemistry* 295, no. 30 (July 24, 2020): 10434–45. https://doi.org/10.1074/jbc.REV120.012742.

Tuerk, C., and L. Gold. "Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase." *Science (New York, N.Y.)* 249, no. 4968 (August 3, 1990): 505–10. https://doi.org/10.1126/science.2200121.