Finalizing the Genome Annotation of Model Cluster E Mycobacteriophage Ukulele via RNA-Seq Analysis

Jackson R. Foley
University of Maine

Follow this and additional works at: https://digitalcommons.library.umaine.edu/honors

Part of the Biochemistry Commons

Recommended Citation
https://digitalcommons.library.umaine.edu/honors/327

This Honors Thesis is brought to you for free and open access by DigitalCommons@UMaine. It has been accepted for inclusion in Honors College by an authorized administrator of DigitalCommons@UMaine. For more information, please contact um.library.technical.services@maine.edu.
FINALIZING THE GENOME ANNOTATION OF MODEL CLUSTER E
MYCOBACTERIOPHAGE UKULELE VIA RNA-SEQ ANALYSIS

by

Jackson R. Foley

A Thesis Submitted in Partial Fulfillment
of the Requirements for a Degree with Honors
(Biochemistry)

The Honors College
University of Maine
May 2018

Advisory Committee:
Sally D. Molloy, Ph.D., Assistant Professor of Genomics, Advisor
Hamish Greig, Ph.D., Assistant Professor of Stream Ecology
David Gross, Ph.D., Adjunct Associate Professor of Honors (English)
Keith W. Hutchison, Ph.D., Professor Emeritus of Microbiology
Benjamin L. King, Ph.D., Assistant Professor of Bioinformatics
ABSTRACT

Mycobacteriophage (phage) are viruses that infect bacteria of the genus *Mycobacterium*, including pathogenic *M. tuberculosis* and non-pathogenic *M. smegmatis*. Temperate phages are capable of undergoing both lytic and lysogenic infection. In lytic infections, phage lyse the host cell after replication. In lysogenic infection, the phage integrates its genome into the host genome (prophage) and replicates with the host. All pathogenic strains of *Mycobacterium* carry prophage that potentially contribute to host virulence and fitness. Formation and maintenance of these prophage is not well understood, particularly for cluster E phage. This project characterizes gene product (gp) 53, a potential Cro-like protein, in cluster E mycobacteriophage Ukulele. Cro-like proteins promote lytic gene expression while repressing lysogenic gene expression. To confirm gp53’s function as a Cro-like protein, strains of *M. smegmatis* overexpressing Ukulele gp53 are being constructed. The impact of gp53 expression on Ukulele infection will be tested on these strains and compared to wildtype.
# TABLE OF CONTENTS

List of Tables v
List of Figures vi

1.0 Introduction 1

2.0 Literature Review 4

2.1 Bacteriophage: the ubiquitous entity 4
2.2 Mycobacteriophage: our bacteriophage of interest 5
2.3 Cluster E mycobacteriophage 7
2.4 Phage lifecycles: a delicate decision 11
2.5 Transcription regulation in bacteriophage via intrinsic termination 16
2.6 Benefits of transcriptome analysis by RNA-seq 21

3.0 Materials and Methods 25

3.1 Mycobacteriophage and bacterial strains 25
3.2 Media and growth conditions 25
3.3 Lysogen isolation 26
3.4 Polymerase Chain Reaction (PCR) amplification of Ukulele gp53 26
3.5 Plasmid preparation 27
3.6 Restriction endonuclease digestion (RD) 27
3.7 Agarose gel electrophoresis 27
3.8 Recombinant plasmid preparation 28
3.9 Bacterial Transformations 28
3.10 DNA sequencing 29
3.11 Total RNA isolation and RNA-seq 29
3.12 Transcriptomic analysis of Ukulele and *M. smegmatis* 30
3.13 Reannotation of the Ukulele genome 31

4.0 Results 32

4.1 Translational start sites were changed for six predicted Ukulele genes based on RNA-seq data 32
4.2 RNA-seq uncovered two unpredicted genes in the Ukulele genome: gp1 and gp59 33
4.3 Sixteen predicted intrinsic terminators align with termination of RNA-seq reads
4.4 RNA-seq revealed two unpredicted sites of transcription termination
4.5 Recombinant plasmids containing Ukulele gp53 may be lethal to bacterial transformants due to cryptic promoters

5.0 Discussion

5.1 Cluster E genomes contain false predictions that can be improved by RNA-seq analysis
5.2 Transcription termination occurs primarily via I-type intrinsic terminators in cluster E
5.3 Cryptic expression of Ukulele gp53 is lethal to host cells
5.4 Conclusions and future directions

References

Author’s Biography
LIST OF TABLES

Table 1. Translational start sites changed based on RNA-seq data 23
Table 2. Putative intrinsic terminators in the Ukulele genome supported by RNA-seq analysis 24
Table 3. Primers used for PCR amplification of Ukulele gp53 26
LIST OF FIGURES

Figure 1. Putative genes and intrinsic terminators in the Ukulele genome  11
Figure 2. The integration cassette of Ukulele resembles the Lambda model  16
Figure 3. Gene calls were re-annotated based on RNA-seq data  33
Figure 4. An unpredicted gene upstream of gp1 was identified by RNA-seq  36
Figure 5. An unpredicted gene between gp57 and gp58  37
   was identified by RNA seq
Figure 6. RNA-seq analysis supports the presence of functional  39
   I-type terminators in the Ukulele genome
Figure 7. RNA-seq analysis supports the presence of functional  40
   L-type terminators in the Ukulele genome
Figure 8. Ukulele gp53 is likely not the immunity repressor  43
Figure 9. Ukulele gp53 may encode two distinct proteins  44
1.0 INTRODUCTION

Viruses that infect bacteria (bacteriophage or simply phage) are the most abundant biological entity on Earth, clocking in at an estimated $10^{31}$ particles worldwide and outnumbering their bacterial hosts roughly 10-fold\textsuperscript{1–5}. They are remarkably diverse\textsuperscript{1,2,4–9} and exist in all environments where bacteria are found, including within humans and other animals\textsuperscript{3–5}. The two are so intimately linked that approximately 60% of all sequenced bacteria contain elements of viral genomes integrated into their chromosomes (prophage) that often help confer virulence and pathogenicity to the bacterial host\textsuperscript{10}. Though the pathogenicity contributions of some prophage, such as the toxin-encoding CTX prophage in pathogenic strains of \textit{Vibrio cholera}, have been characterized, the impact of most prophage on the pathogenicity of their hosts remains a mystery\textsuperscript{18}. As obligate symbionts, phage provide a vehicle for better understanding not only the evolutionary mechanics of genomic diversity, but also the impact that phage have on the physiology and fitness of their hosts\textsuperscript{3–5,10–13}.

Mycobacteriophages are bacteriophages that infect the bacteria of the genus \textit{Mycobacterium}, which includes human pathogens like \textit{M. tuberculosis} and \textit{M. leprosy} and nonpathogenic \textit{M. smegmatis}\textsuperscript{1,2,6,7}. All characterized mycobacterial pathogens contain prophage elements within their genomes that likely contribute to aspects of their fitness and pathogenicity. Though they infect bacteria of the same genus, mycobacteriophage are extremely diverse in genetic organization and content\textsuperscript{1,2,7,20} and are grouped by genomic content and nucleotide similarity into clusters with some clusters being further divided into subclusters for patterns of extreme nucleotide similarity\textsuperscript{2,6,20}. Phages that do not
exhibit significant similarity to any known clusters are deemed ‘singletons’. To date, there are 115 different clusters (A–Z, AA, AB, AC, etc…), 32 of which have multiple subclusters, and 60 singletons\(^2\). Although some clusters have been well characterized, cluster E remains largely unexplored in the literature.

The genome of model cluster E mycobacteriophage Ukulele has been sequenced, annotated, and submitted to GenBank\(^2\). Translational start sites, gene functions, and transcriptional regulatory sequences have been predicted via bioinformatics approaches, however these have not yet been experimentally confirmed\(^2\,^2\). In order to experimentally verify translational start sites of putative genes, lifecycle-dependent gene functions, and transcriptional regulators in cluster E mycobacteriophage, RNA-seq analysis was performed on the cluster E phage Ukulele. Transcriptomic analysis of Ukulele revealed two previously unpredicted genes, six translational start sites that required reannotation, and 18 putative intrinsic transcriptional terminators—14 of which do not follow the canonical L-type structure described in the literature\(^3\), but instead follow the I-type structure lacking a poly-U trail. Though there is some debate in the literature about whether I-type intrinsic terminators are sufficient for terminating transcription\(^3\,^4\), this data suggests not only that I-type terminators are functional in mycobacteriophage, but also that they are the more abundant in mycobacteriophage genomes than are L-type intrinsic terminators.

RNA-seq analysis of the Ukulele transcriptome was also employed to investigate the integration machinery of cluster E mycobacteriophage. As temperate phage, members of cluster E including Ukulele form stable mycobacterial lysogens, however it is not yet understood how lysogeny is achieved and maintained by cluster E mycobacteriophage.
Cluster E lysogens exhibit superinfection immunity to other cluster E phage, but are susceptible to superinfection by phage of other clusters \(^{21}\). And though this indicates the presence of a Lambda-like repressor-mediated system for lysogen formation and maintenance, the only gene product involved in lysogeny regulation that has been identified in cluster E thus far is a tyrosine integrase (Figure 2)\(^{29}\). Though an immunity repressor has been predicted in many cluster E genomes (Ukulele gp53), work in our lab suggests this gene product does not function as a lysogeny-inducing repressor, but that it may instead be a lytic-inducing Cro-like or excise protein. Structural and transcriptomic analysis of Ukulele gp53 also support its function as a regulatory protein contributing to the lytic cycle rather than lysogeny. Secondary and tertiary folding predictions by Phyre\(^{2}\) indicate that gp53 is structurally most similar to the Xis protein of mycobacteriophage Pukovnik based on the presence of a winged helix-turn-helix DNA-binding domain\(^{34}\). Further, RNA-seq analysis suggests that gp53 is most highly expressed during early lytic growth, rather than during lysogeny\(^{29}\). This expression pattern is more typical of a lytic-inducing Cro-like protein. Part of this study aims to address the function of Ukulele gp53 by performing plaque assays of Ukulele plated on a strain of \textit{M. smegmatis} overexpressing Ukulele gp53. If gp53 does indeed contribute to the regulation of lytic growth like Cro or Xis, incoming Ukulele particles should only undergo lytic growth and exhibit clear plaques as a result.
2.0 LITERATURE REVIEW

2.1 Bacteriophage: the ubiquitous entity

Viruses that infect bacteria (bacteriophage or simply phage) are the most abundant biological entity on Earth, clocking in at an estimated $10^{31}$ particles worldwide and outnumbering their bacterial hosts roughly 10-fold$^{1-5}$. They are remarkably diverse$^{1,2,4-9}$ and exist in all environments where bacteria are found, including within humans and other animals$^{3-5}$. The two are so intimately linked that approximately 60% of all sequenced bacteria contain elements of viral genomes integrated into their chromosomes (prophage)$^{10}$. As obligate symbionts, phage provide a vehicle for better understanding not only the evolutionary mechanics of genomic diversity, but also the impact that phage have on the physiology and fitness of their hosts$^{3-5,10-13}$.

Phage that form prophage are temperate, meaning they can carry out both lytic and lysogenic infection$^{1,2,5,11}$. In the well-studied and ubiquitous lytic lifecycle, infecting phage synthesize and assemble progeny phage particles from within the host which are then released to the environment via cell lysis by viral proteins$^{1,2,6,7,14,15}$. This occurs when the host cell is healthy upon infection$^{5,14,15}$. When environmental conditions are not optimal to produce more viral particles however, temperate phage are able insert their genome into that of the host and replicate quiescently along with the cell without killing it$^{1,5,10,11,15,16}$. Often, the integrated prophage helps ensure the survival of its host (lysogen) by expressing genes whose products contribute to the fitness of the host$^{5,10,11}$.

Subsequently, if conditions improve, the phage can excise its genome from the host’s and re-enter lytic growth$^{1,7,10,15}$. Each lifecycle is controlled by a distinct subset of
viral genes that both promotes the selected lifecycle and inhibits the other\(^7,15,16\). Though lysogeny maintenance has been characterized for some temperate phages\(^7,13,15,16\), it remains a mystery how most prophage initiate and maintain integration.

During lysogenic infection, phage have a widespread impact on the physiology of their bacterial hosts\(^5,10,12,17,18\). Often, prophage carry with them genes that help enhance host fitness, which in turn benefits the virus hiding out within it\(^5,10,12,16,17,18\). In fact, a vast majority of pathogenic bacterial strains contain one or more of these prophage elements that help contribute to varying aspects of virulence and pathogenicity\(^5,10,12,18\). For example, the bacterial pathogen responsible for cholera in humans, \textit{Vibrio cholerae}, is on its own nonpathogenic\(^18\). However, pathogenic strains of \textit{V. cholerae} contain the CTX prophage which encodes the cholera toxin that causes symptoms in humans\(^18\). However, not all prophages are this direct in their contributions to host virulence.

Many prophage help enhance host virulence more indirectly by increasing the host’s fitness in stressful and otherwise uninhabitable environments like the human body\(^5,10\). Overall, the variety and nature of these fitness-conferring prophage-host interactions remain substantially unexplored and enigmatic. Unraveling this relationship between phage infection and host physiology is therefore greatly important for understanding the function and evolution of bacterial pathogens.

\subsection*{2.2 Mycobacteriophage: our bacteriophage of interest}

Mycobacteriophages are bacteriophages that infect the bacteria of the genus \textit{Mycobacterium}, which includes human pathogens like \textit{M. tuberculosis} and \textit{M. leprosy} and nonpathogenic \textit{M. smegmatis} \(^1,2,6,7\). Recent advances in the field of genomics and bioinformatics and the SEA PHAGES program\(^19,20\) have significantly accelerated the
ability to characterize phage on a large scale. As of March 27, 2018, 9,679 mycobacteriophage have been isolated and 1,576 of these have had their genome sequenced and submitted to the PhagesDB database—a 162% increase over the past three years\textsuperscript{21, 22}. Broad analyses of these phage genomes have revealed that even though they infect bacteria of the same genus, mycobacteriophage are extremely diverse in genetic organization and content\textsuperscript{1, 2, 7, 20}. This is primarily due to horizontal gene transfer\textsuperscript{1, 2, 6, 7}.

Illegitimate recombination events over the course of evolution cause phage to gain and lose genetic material, and thus phage can range from having nearly identical nucleotide sequences to being ostensibly unrelated in any genomic aspect\textsuperscript{2, 6, 11} Although this can happen anywhere throughout the genome, natural selection typically presses for recombination events occurring at the boundaries of genes or groups of related genes (operons or cassettes)\textsuperscript{2, 6, 11}. This results in phage containing interchangeable modules of genetic content that can be fine-tuned throughout evolution, driving a continuum of phage diversity\textsuperscript{1, 2, 6}.

In order to make sense of this continuum, phages are grouped by genomic content and nucleotide similarity into clusters with some clusters being further divided into subclusters for patterns of extreme nucleotide similarity\textsuperscript{2, 6, 20}. Phages that do not exhibit significant similarity to any known clusters are deemed ‘singletons’. To date, there are 115 different clusters (A–Z, AA, AB, AC, etc…), 32 of which have multiple subclusters, and 60 singletons\textsuperscript{21}. Cluster-specific studies have been instrumental in identifying various mechanisms of phage biology that demonstrate the remarkable diversity among mycobacteriophage. While some clusters such as A\textsuperscript{23}, K\textsuperscript{24}, M\textsuperscript{25}, N\textsuperscript{16}, O\textsuperscript{26}, Q\textsuperscript{13} and even some singletons\textsuperscript{27} have been well-studied, many remain uncharacterized. For example,
studies in cluster A, Q, and N have revealed three novel systems for establishing and maintaining lysogeny and the genetic elements required for each\textsuperscript{13,16,23}. Though many clusters have been studied in great detail, cluster E remains uncharacterized in the literature.

### 2.3 Cluster E mycobacteriophage

Cluster E contains 89 members whose genomes average 75,505 bp in length with 143 genes, two tRNA’s, and a GC content of 63\%\textsuperscript{21}. Phage in this cluster have a \textit{Siphoviridae} particle morphology, possessing a non-enveloped icosahedral capsid and a long non-contractile tail\textsuperscript{6}. Cluster E phage are temperate and form stable lysogens in \textit{M. smegmatis}\textsuperscript{21,22}. Across cluster E phage, genetic content and function is highly conserved and there are no gene products warranting novel Phamly classification\textsuperscript{22}. Though cluster E phage share many characteristics common to most mycobacteriophage, they also exhibit a wealth of unique features that are not yet fully understood\textsuperscript{22}. Overall, cluster E remains poorly characterized and has been overlooked in the literature.

The genome of the model cluster E mycobacteriophage Ukulele has been sequenced and annotated\textsuperscript{22}. Translational start sites, gene functions, and transcriptional regulatory sequences have been predicted, however these have not yet been experimentally confirmed\textsuperscript{22,28}. The presence of genes can be predicted via bioinformatics analysis with a decent level of accuracy\textsuperscript{2,13}, however the boundaries of these genes do not always line up with the most likely prediction. For example, a typical practice in genome annotations is to prioritize the longest possible ORF when predicting translational start sites for putative genes. Though utilizing the longest ORF allows the phage to avoid carrying extra unused genomic content in many cases, intergenic regions are often useful
for containing other regulatory sequence elements including transcriptional regulators like promoters, terminators, and operators\textsuperscript{1,2}. Thus, every gene call is merely a prediction based on computational evidence that must be confirmed or improved by experimental data.

Preliminary genome annotations also require experimental evidence to support gene function calls. Although the presence of genes can be predicted with a decent level of accuracy, functions can only be predicted by bioinformatics approaches for 25\% of the genes in the Ukulele genome, leaving a clear majority without any putative function. Preliminary annotations of cluster E genomes have revealed several features that are shared among mycobacteriophage in general and others that are cluster E specific\textsuperscript{22}. The overall genome of cluster E phage follows the typical structure of phage genomes\textsuperscript{1, 6, 22}: structural and lysis genes are located on the left arm of the genome, genes involved in nucleic acid metabolism are located on the right arm of the genome, and the integration cassette is located centrally in the genome (Figure 1)\textsuperscript{22}. Beyond these structural proteins and enzymes which are conserved to some extent in all mycobacteriophages, there are several genes that are predicted to be specifically conserved across cluster E\textsuperscript{22}. Among these putative gene products are an Lsr2 DNA-bridging regulatory protein, a WhiB transcriptional regulator, a tyrosine integrase, and five proteins with putative DNA-binding domains\textsuperscript{22, 29}. Though putative functions have been called for approximately 25\% of cluster E genes, none have yet been confirmed in the lab\textsuperscript{22, 29}. Therefore, experimental analysis is required to further improve and confirm genome annotations of cluster E phage.
As temperate phage, members of cluster E form stable mycobacterial lysogens, however it is not yet understood how lysogeny is achieved and maintained by cluster E mycobacteriophage. Cluster E lysogens exhibit superinfection immunity to other cluster E phage, but are susceptible to superinfection by phage of other clusters \(^{21}\). And though this indicates the presence of a Lambda-like repressor-mediated system for lysogen formation and maintenance, the only gene product involved in lysogeny regulation that has been identified in cluster E thus far is a tyrosine integrase (Figure 2)\(^{29}\). Though an immunity repressor has been predicted in cluster E phage, work in our lab suggests this gene product does not function as a lysogeny-inducing repressor, but that it may instead be a lytic-inducing Cro-like or excise protein. The huge variability in lysogeny maintenance systems across viruses highlights the importance of characterizing lifecycle regulation in different groups of phage, especially since these prophages are contributing to the fitness and virulence of their potentially pathogenic hosts. In order to fully characterize the mechanism by which cluster E mycobacteriophage form stable lysogens however, it is vital to supplement predictive genome annotations with experimental evidence.

Transcriptional regulatory sequences responsible for transcription initiation (promoters) and termination (terminators) have also been predicted in the Ukulele genome using a combination of computational methods\(^{28}\). These predictions rely on algorithms designed based around consensus sequences promoters and terminators characterized in the literature however, and are viable to produce false predictions or fail to predict functional transcriptional regulators. Though all putative promoters and terminators called in the Ukulele genome lie within intergenic gaps, a majority of
putative terminators do not entirely comply with the typical structure of bacterial terminators characterized in the literature\textsuperscript{30-32}. Further, these putative terminators have not yet been experimentally confirmed to function in effectively terminating transcription.

Though preliminary genome annotations have been carried out in multiple cluster E phage, a vast majority of putative translational start sites, gene functions, and transcriptional regulators have not yet been experimentally verified. In order to better understand cluster E mycobacteriophage, their genomes, and the mechanisms by which they impact mycobacterial hosts, it is essential that experimental data is utilized to improve cluster E genome annotations.
Lysogeny provides temperate phage with an advantageous method for surviving in nutrient-poor conditions by allowing them to hide out within their host until proper environmental conditions for producing progeny arise \(^1,5,10,11,15,16\). If conditions are suitable and the host is growing rapidly upon infection, phage enter the lytic cycle and use host metabolic machinery to synthesize progeny, culminating in lysis of the cell to
release viral particles into the environment\textsuperscript{1,2,6,7,14,15}. Should conditions not be optimal for lytic growth however, temperate phages utilize lysogeny to endure silently until the health of the host cell improves\textsuperscript{5,10,11,15}. Often, this relies heavily on the ability of the phage genome to integrate itself into the host chromosome and replicate quiescently from within it\textsuperscript{1,2,5,7}, however the phage genome may alternatively be maintained as an extrachromosomal plasmid in some cases\textsuperscript{33}.

Though the mechanisms for establishing and maintaining lysogeny exhibit great variability among bacteriophage, many typically follow some variation of the integration-dependent system described for the \textit{E. coli} phage Lambda. This mechanism for regulating lysogeny utilizes a multi-component system to gauge the health of the host cell and flip a genetic switch controlling the exclusive expression of lytic- or lysogeny-inducing genes\textsuperscript{15}. This switch is controlled by an operator containing three binding-sites (OR\textsubscript{1}, OR\textsubscript{2}, and OR\textsubscript{3}) located between two divergent promoters responsible for expression of lytic- and lysogenic-specific operons (P\textsubscript{R} and P\textsubscript{RE}/P\textsubscript{RM})\textsuperscript{15}. These operator binding sites are bound by the DNA-binding transcriptional regulators for lytic and lysogenic growth, Cro and CI (repressor) respectively\textsuperscript{15}. Slight overlap between the two outer binding sites (OR\textsubscript{1} and OR\textsubscript{3}) and the genes adjacent to them combined with preferential binding affinities of each site for Cro and CI allows for transcription initiation from the appropriate promoter in the given cellular conditions while also blocking RNA polymerase from interacting with the incorrect promoter\textsuperscript{15}. This is primarily mediated by another transcriptional regulator known as CII that determines which of these two will be produced\textsuperscript{15}.

Upon infection of healthy cells, host proteases are abundant and degradation of CII induces transcription from the promoter P\textsubscript{R} and the lytic-inducing Cro protein is
expressed\textsuperscript{15}. Cro is a DNA-binding protein that interacts first with $O_{R3}$ followed by a lower affinity binding to $O_{R2}$\textsuperscript{15}. When this occurs, the promoters controlling genes involved in establishing and maintaining lysogeny ($P_{RE}$ and $P_{RM}$, respectively) are blocked and lysogeny is suppressed\textsuperscript{15}. Additionally, Cro binding at $O_{R3}$ and $O_{R2}$ helps recruit RNA polymerase for the transcription of genes involved in the lytic process such as structural and lysis genes\textsuperscript{15}.

In unhealthy cells, however, CII is protected from degradation by the protein CIII and expression of integrase and CI occurs from the promoters $P_I$ and $P_{RE}$, respectively\textsuperscript{15}. Expression of integrase catalyzes site-specific recombination of the viral genome into the host chromosome to establish the lysogen\textsuperscript{15}. CI is a repressor that binds to $O_{R1}$ then $O_{R2}$ to block transcription initiation of lytic genes from $P_R$ and induce expression from $P_{RM}$ and $P_{RE}$\textsuperscript{15}. As a result, lytic growth is inhibited and lysogeny is maintained by the continued expression of repressor\textsuperscript{15}. Typically, enough repressor is synthesized to suppress lytic growth not only by itself, but also by any superinfecting phage that share similar operator-binding sequences\textsuperscript{15}. This process is not necessarily permanent however, and prophage have a way of re-entering lytic growth should host nutrient conditions improve\textsuperscript{15}.

As lysogeny is a mechanism for surviving hard times and not the primary reproduction strategy, it is crucial that phage be able to switch back to lytic growth when feasible\textsuperscript{15}. This process is mediated by another DNA-binding protein called excise\textsuperscript{15}. Excise is responsible for initiating site-specific recombination of the viral genome by integrase in the opposite direction as occurs during integration, effectively removing the
phage from the host chromosome. All phage excise proteins contain some variation of a winged helix-turn-helix DNA-binding domain.

This simplistic genetic switch provides the phage with an energetically and temporally efficient system for adapting its lifecycle to the environmental conditions. Though this schematic was elucidated from an *E. coli* phage, it has proven useful as a model in studying repressor-dependent integration systems in a wide variety of temperate phage, including some clusters of mycobacteriophage. However, many phage are known not to rely on this repression-dependent system for prophage formation and maintenance. Superinfection immunity assays between cluster E phage suggest the presence of repressor-mediated immunity, and gp50 has been implicated as the integrase in the model cluster E phage Ukulele, but other components of the cluster E lysogeny regulation system remain a mystery.

Preliminary annotations of all cluster E genomes identify a putative immunity repressor located three genes downstream from the integrase (Ukulele equivalent gp53, Pham 5753) because it is close to the integrase, located at a region of divergent transcription, and contains a putative DNA-binding domain (Figure 2). However, experimental data suggests that this is not a valid prediction. When deleted from the genome, loss of gp53 is lethal and pure deletion mutants are not able to be recovered. If gp53 were indeed the repressor, its deletion should result in lytic growth, indicated by distinctively clear plaques rather than cell death. If gp53 encodes a Cro-like protein on the other hand, deletion would result in the inability for Ukulele particles to undergo lytic growth and would therefore be fatal— as observed. Though it is likely not the repressor, it is possible that Ukulele gp53 encodes a different component of the lifecycle regulatory
machinery utilized by Ukulele, functioning as a Cro-like protein, Xis, or even a novel regulatory protein not yet characterized in other phages.

Structural and transcriptomic analysis of Ukulele gp53 also support its function as a regulatory protein contributing to the lytic cycle rather than lysogeny. Secondary and tertiary folding predictions by Phyre² indicate that gp53 is structurally most similar to the Xis protein of mycobacteriophage Pukovnik based on the presence of a winged helix-turn-helix DNA-binding domain³⁴. Further, RNA-seq analysis suggests that gp53 is most highly expressed during early lytic growth, rather than during lysogeny²⁹. This expression pattern is more typical of a lytic-inducing Cro-like protein. As these are simply predictions though, experimental evidence is required to confirm whether gp53 truly functions to support lytic growth in Ukulele and if so, to elucidate the mechanism by which it contributes. Part of this study aims to address the function of Ukulele gp53 by performing plaque assays of Ukulele plated on a strain of M. smegmatis overexpressing Ukulele gp53. If gp53 does indeed contribute to the regulation of lytic growth like Cro or Xis, incoming Ukulele particles should only undergo lytic growth and exhibit clear plaques as a result.
2.5 Transcription regulation in bacteriophage via intrinsic termination

Bacteriophage rely on transcription-level regulation of genes needed for successful lytic and lysogenic gene expression\textsuperscript{15,30,35}. Identifying sequences involved in transcriptional gene regulation improves our understanding of the biology of the phage and can also improve the quality of genome annotations. Sequences that signal transcription start points (promoters) and transcription termination (terminators) can be identified through sequence analysis. These transcriptional regulators are genomic
sequence elements that interact with the host RNA polymerase in various ways to alter its ability to bind to and express open reading frames (ORFs)\textsuperscript{30-32,35}. Although their general structure and function has been briefly characterized in some bacterial hosts, there is still some debate about the nature of these transcriptional regulatory elements in \textit{Mycobacterium} and in mycobacteriophage\textsuperscript{30-33}.

There are two general mechanisms for transcription termination in prokaryotes: Rho-dependent and Rho-independent, or intrinsic, termination. Rho-dependent termination, as its name suggests, relies on the RNA-binding protein Rho to disrupt transcription elongation and displace transcribed RNA from the elongation complex\textsuperscript{36-38}. To do so, Rho binds to a C-rich sequence in the transcript called the Rho utilization site and exhibits ATPase-dependent helicase activity to separate the DNA-RNA hybrid and release the transcript\textsuperscript{36-38}. Because this mechanism for terminating transcription does not solely rely on genomic sequence elements however, Rho-dependent terminators cannot be identified or predicted through sequence analysis\textsuperscript{36-38}.

Intrinsic termination on the other hand, is mediated entirely through genomic sequence elements without relying on a protein like Rho to assist in disrupting the elongation complex, and therefore can be studied via sequence analysis. Rho-independent termination utilizes the stable secondary folding of transcribed sequences called intrinsic terminators to stall and disrupt RNA-polymerase, releasing it from the elongation complex\textsuperscript{39}. Intrinsic terminators are typically located within 50 nt of a stop codon, but can be found up to 250 nt downstream of a stop codon in some cases\textsuperscript{30,31}. These rho-independent terminators typically have a two-component structure: a GC-rich stem-loop hairpin that forms from secondary folding of a transcribed inverted palindromic sequence.
and a U-rich trail sequence following it (≥ 5 U’s). Though this is the canonical structure of intrinsic terminators primarily described in the literature, there is also some evidence of functional intrinsic terminators that lack a poly-U trail in certain actinobacterial genera, including *mycobacterium*. Because of their structure, intrinsic terminators containing a poly-U trail are deemed L-type terminators and ones lacking a poly-U trail are known as I-type terminators. However, there is an ongoing disconnect in the literature about the importance of the poly-U trail and whether these I-type terminators are indeed functional.

Various studies investigating the ability for I-type intrinsic terminators to disrupt transcription elongation by mycobacterial RNA polymerase *in vitro* and *in vivo* have reported conflicting results concerning the necessity of a partially intact poly-U trail. Noncanonical I-type terminators have been studied *in vivo* using the mycobacteria-specific termination selection vector pTER5 with mycobacterial intrinsic terminators cloned upstream of a CAT reporter gene. A comparable reduction in transcription (~85%) of the CAT reporter gene is observed whether the cloned terminator is L-type, I-type with a partial poly-U trail (≥ 3 U’s), or I-type completely lacking a poly-U trail. Transcription termination is also observed at a similar efficiency for I-type terminators tested *in vitro* using RNA polymerase purified from *M. smegmatis*, suggesting that these terminators are indeed functioning in a factor-independent manner. Termination efficiency is greatest when intrinsic terminators are located within 37–50 nt of the stop codon but also showed effective termination up to 250 nt downstream of the stop codon. This study also notes that *M. tuberculosis* and *M. smegmatis* almost exclusively rely in I-type intrinsic terminators, but do contain some functional L-type terminators as...
well. Importantly, termination is not effective when intrinsic terminators were placed within coding regions, likely due to the close coupling of bacterial transcription and translation preventing the formation and detachment of the hairpin from the RNA in the coding region.

Another study investigating the termination efficiency of mycobacterial I-type terminators in vitro using M. bovis RNA polymerase however, indicates that I-type terminators are only effective if the poly-U trail is still partially intact (≤3 U’s). Even in the presence of the conserved elongation regulators NusA and NusG, which function to assist factor-dependent termination at sites where Rho is not alone sufficient, I-type terminators were ineffective for terminating transcription. I-type terminators with a partial poly-U tract on the other hand, did effectively terminate transcription, suggesting that this is a necessary feature for effective intrinsic terminators but that it may not be as tightly regulated as previously thought. This also begs the question of what counts as a partially intact poly-U trail and whether previous studies claiming to investigate I-type terminators completely lacking a poly-U actually contained an overlooked partial poly-U trail. This debate highlights not only the diversity of systems by which prokaryotes regulate gene expression at the transcriptional level, but also the importance of further studying the nature of intrinsic termination in mycobacteriophage and their hosts.

To date, only L-type terminators have been predicted in mycobacteriophage genomes. However, recent advances in bioinformatics approaches for predicting terminators within genomes provide a way of gaining insight into the potential termination system used by mycobacteriophage. The web-based program WebGester (Genome Scanner for Terminators), for example, uses an algorithm to search genomic
sequences for potential intrinsic terminators by identifying GC-rich palindromic
sequences with secondary structure that forms a stable hairpin with a highly negative $\Delta G$
(-15 – 25 kcal/mol) capable of terminating transcription$^{30}$. It also uses the context of
ORFs within the genome to only provide potential terminator sequences that lie outside
of coding regions$^{30}$. Additionally, WebGEster can be used to identify U-shaped and X-
shaped terminators which are simply multiple terminators (I- or L-type) that occur in
tandem in unidirectional or bidirectional orientation, respectively$^{30}$. Another web-based
program, ARNold, identifies intrinsic terminators via a similar algorithm, yet it only
identifies L-type terminators containing a poly-U trail$^{32}$. Also in contrast to WebGEster,
ARNold recognizes potential terminators originating from intragenic sequences in
addition to intergenic regions$^{32}$. This is problematic not only because intragenic
terminators have been shown to be ineffective in terminating transcription elongation$^{31}$,
but also because an effective terminator in the middle of a coding region would result in a
truncated gene product and likely interfere with its subsequent function.

These programs have been previously utilized in our lab to identify 19 potential
strong terminators in the Ukulele genome, with a majority (82%) being I-type$^{28}$. Belisle
identified convergent X-type terminators and U-shaped tandem terminators in cluster E
mycobacteriophage$^{28}$. These structures were highly conserved in all cluster E
mycobacteriophage further suggesting that I-type terminators may play an important role
in transcription termination in cluster E and mycobacteriophage in general. Although
these analyses provide evidence of transcription regulation in mycobacteriophage by
various types of intrinsic terminators, experimental methods are necessary for confirming
their function in-vivo.
2.6 Benefits of transcriptome analysis by RNA-seq

Preliminary genome annotations carried out using bioinformatics approaches, as was done for Ukulele, are purely predictive and gene calls require experimental evidence to confirm their presence and function. RNA-seq analysis is one method for improving genome annotations by giving transcriptomic evidence for the placement of genes and the timing of their expression\textsuperscript{13, 16, 41, 42}. On top of providing quality control for the relative locations of translational start sites and unpredicted genes, this allows genes with lifecycle-dependent functions to be identified by the timing of their expression and can also elucidate regions where transcriptional regulation occurs\textsuperscript{13, 16, 41}. Though transcriptome analysis by RNA-seq has been utilized to improve the genome annotations of some mycobacteriophage including the cluster Q phage Giles\textsuperscript{13}, this approach has not yet been applied to cluster E mycobacteriophage.

RNA-seq provides precise locations of transcription boundaries with single-base resolution and thus can be used to confirm predicted genes, improve predicted translational start sites, and identify unpredicted genes\textsuperscript{41}. Alignment of transcript reads to predicted genes provides evidence for the validity of the gene call, and the single-base resolution of these reads allows for a predicted translational start site to be adjusted if it lies upstream of the experimentally determined transcriptional start site\textsuperscript{41}. Further, predicted genes may be extended if a longer ORF is available for which the translational start site is still downstream of the origin of transcript reads. Transcripts arising from genomic regions that do not contain a predicted gene can also improve genome annotations by indicating potential genes that were initially overlooked\textsuperscript{41, 42}. Though the presence of mapped reads does not necessarily prove the presence of a gene, it can
provide evidence for the necessity of further investigation by both bioinformatics and experimental approaches.

Transcriptomic evidence can also be used to improve predictions of transcriptional regulators like intrinsic terminators\textsuperscript{16,41,42}. Regions where transcript reads show a sharp decline indicate locations where the elongation complex was disrupted and transcription was terminated\textsuperscript{41}. By comparing these locations with predictive and manual approaches described above, intrinsic terminators can be more accurately predicted and selected for further experimental confirmation\textsuperscript{41}.

Additionally, predictions of gene function can be made and improved based on their expression levels during different phases of growth\textsuperscript{13,16,41,42}. Previous studies using this method have identified genes essential for lytic and lysogenic growth in the cluster Q mycobacteriophage Giles\textsuperscript{13}. This included the identification of a gene encoding a repressor that lacks a DNA-binding domain—a feature common to most repressor proteins, without which its identification would have been exceedingly difficult and unlikely\textsuperscript{13}. Combining transcriptomic evidence of heightened expression during lysogeny with functional evidence from deletion mutant and superinfection immunity tests successfully elucidated the identity of the repressor in Giles, validating the usefulness of this approach in determining viral gene products with lifecycle-specific functions\textsuperscript{13}. This approach was also utilized to identify a set of genes involved in non-repressor-mediated superinfection immunity in cluster N phage\textsuperscript{16}. RNA-seq data revealed lysogenic expression of a five-gene operon between the lysis and integration cassettes, resulting in the hypothesis that they serve some function in viral-mediated immunity\textsuperscript{16}. After being identified, efficiency of plating tests between cluster N lysogens and 80 lytic and
temperate phages demonstrated that these genes play a role in heterotypic viral defense to superinfection by reducing the efficiency of plating in over 70 phages that are not similarly affected by presence of the cluster N integrase by itself\textsuperscript{16}. In a similar approach as was used in Giles, gene expression profiles were used to determine the function of novel proteins based on the timing of their expression— further validating the value of this approach\textsuperscript{16}. This is important not only for characterizing lysogeny maintenance systems in temperate phage, but also for identifying candidates for prophage-encoded genes that contribute to host fitness, especially in pathogens.

This study aims to utilize RNA-seq transcriptomic analysis of the model cluster E mycobacteriophage Ukulele to improve the genome annotations and therefore the general understanding of cluster E phage. In doing this, translational start sites have been adjusted for eight genes, with seven being shortened and one being extended, and two unpredicted genes have been identified in the Ukulele genome (Table 1). Additionally, transcriptome analysis by RNA revealed that Ukulele gp53 is likely not the immunity repressor, as is predicted across many cluster E genomes, since it is most highly

<table>
<thead>
<tr>
<th>Gene</th>
<th>Orientation</th>
<th>Function</th>
<th>Original Start (bp)</th>
<th>Updated Start (bp)</th>
<th>Stop Codon (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp1</td>
<td>FWD</td>
<td>NKF</td>
<td>—</td>
<td>10</td>
<td>270</td>
</tr>
<tr>
<td>gp8</td>
<td>REV</td>
<td>NKF</td>
<td>2,106</td>
<td>2,223</td>
<td>1,924</td>
</tr>
<tr>
<td>gp21</td>
<td>FWD</td>
<td>Tail Assembly Chaperone</td>
<td>12,159</td>
<td>12,231</td>
<td>13,102</td>
</tr>
<tr>
<td>gp53</td>
<td>REV</td>
<td>NKF (Potential Cro-like or Xis)</td>
<td>39,804</td>
<td>39,711</td>
<td>39,364</td>
</tr>
<tr>
<td>gp59</td>
<td>FWD</td>
<td>NKF</td>
<td>—</td>
<td>41,827</td>
<td>42,015</td>
</tr>
<tr>
<td>gp131</td>
<td>REV</td>
<td>NKF</td>
<td>70,632</td>
<td>70,542</td>
<td>70,285</td>
</tr>
<tr>
<td>gp141</td>
<td>REV</td>
<td>NKF</td>
<td>73,370</td>
<td>73,385</td>
<td>73,173</td>
</tr>
<tr>
<td>gp143</td>
<td>REV</td>
<td>NKF</td>
<td>73,835</td>
<td>73,799</td>
<td>73,644</td>
</tr>
</tbody>
</table>

Putative genes with translational start sites that were changed and unpredicted genes that were identified via RNA-seq and their relative location within the Ukulele genome. Genes are numbered per the new annotation. A “—” in the Original Start column denotes a previously unidentified gene.
expressed during early lytic growth (Figure 7). Lastly, RNA-seq analysis of the Ukulele transcriptome for evidence of transcription regulation has supported the presence and function of putative noncanonical I-type intrinsic terminators in the Ukulele genome, with I-type terminators representing a majority (87.5%) of the identified intrinsic terminators (Table 2).

### Table 2. Putative intrinsic terminators in the Ukulele genome supported by RNA-seq.

<table>
<thead>
<tr>
<th>Forward Terminators</th>
<th>Gene product</th>
<th>Gene Function</th>
<th>Distance From Stop Codon (nt)</th>
<th>Terminator Start Position (nt)</th>
<th>Terminator 5'–3' Sequence [Stem – Loop – Stem (Poly-U)]</th>
<th>Type</th>
<th>ΔG [kcal/mol]</th>
</tr>
</thead>
<tbody>
<tr>
<td>UT1</td>
<td>NKF</td>
<td>NKF</td>
<td>6</td>
<td>1,679</td>
<td>GGACTGACGCTGCGGGGCGGACGATTTTCTGTCA</td>
<td>I</td>
<td>-29.94</td>
</tr>
<tr>
<td>UT2</td>
<td>NKF</td>
<td>NKF</td>
<td>16</td>
<td>180</td>
<td>CCGGGCGCGGCGGCAATCTGCCCTTACCCGCTGTCA</td>
<td>I</td>
<td>-18.13</td>
</tr>
<tr>
<td>UT3</td>
<td>NKF</td>
<td>NKF</td>
<td>52</td>
<td>200</td>
<td>GCCTGACGCGGCGGCAATCTGCCCTTACCCGCTGTCA</td>
<td>I</td>
<td>-16.2</td>
</tr>
<tr>
<td>UT4</td>
<td>NKF</td>
<td>NKF</td>
<td>52</td>
<td>154</td>
<td>CCGGGCGCGGCGGCAATCTGCCCTTACCCGCTGTCA</td>
<td>I</td>
<td>-16.2</td>
</tr>
<tr>
<td>UT5</td>
<td>NKF</td>
<td>NKF</td>
<td>52</td>
<td>121</td>
<td>CCGGGCGCGGCGGCAATCTGCCCTTACCCGCTGTCA</td>
<td>I</td>
<td>-16.2</td>
</tr>
<tr>
<td>UT6</td>
<td>NKF</td>
<td>NKF</td>
<td>52</td>
<td>10</td>
<td>CCGGGCGCGGCGGCAATCTGCCCTTACCCGCTGTCA</td>
<td>I</td>
<td>-16.2</td>
</tr>
<tr>
<td>UT7</td>
<td>NKF</td>
<td>NKF</td>
<td>124</td>
<td>130</td>
<td>CCGGGCGCGGCGGCAATCTGCCCTTACCCGCTGTCA</td>
<td>I</td>
<td>-16.2</td>
</tr>
<tr>
<td>UT8</td>
<td>NKF</td>
<td>NKF</td>
<td>124</td>
<td>7</td>
<td>CCGGGCGCGGCGGCAATCTGCCCTTACCCGCTGTCA</td>
<td>I</td>
<td>-16.2</td>
</tr>
</tbody>
</table>

Sequences and relative locations of putative intrinsic terminators in the Ukulele genome confirmed or identified by RNA-seq analysis. A "*" in the terminator name indicates that it was not previously predicted.
3.0 MATERIALS AND METHODS

3.1 Mycobacteriophage and bacterial strains

The temperate cluster E mycobacteriophage Ukulele was isolated from flower garden soil located in Old Orchard Beach, Maine in 2011 as described by Beacham et al. 2015. Ukulele was isolated using the host *M. smegmatis* mc2155 (ATCC: 700084; NC_008596.1) and the enrichment method. Transformations were performed in NEB 5-alpha competent *E. coli* according to the manufacturer’s recommendations (New England Biolabs (NEB), Ipswitch, MA).

3.2 Media and growth conditions

*M. smegmatis* mc2155 cells were grown in complete media (7H9 broth; 1 mM calcium chloride; 10% AD supplement; 50 µg/mL carbenicillin (Sigma, St. Louis, MO); 10 µg/mL carboheximide). When necessary, cells were supplemented with 25 µg/mL kanamycin. Cells were incubated at 37ºC with shaking at 220 rpm until growth reached late-log stage. *M. smegmatis* was additionally cultured on 7H10 agar plates. Ukulele *M. smegmatis* lysogens were initially incubated at room temperature in 0.005% Tween-80 (Fisher Scientific, Fair Lawn, NJ) for one week. Lysogens were then sub-cultured into complete media without Tween-80 and grown for one week until cells reached the late-log growth stage. *E. coli* XL1-Blue cells were cultured in L-broth supplemented with 50 µg/mL kanamycin.
3.3 Lysogen isolation

Lysogens of Ukulele were isolated from Ukulele-infected *M. smegmatis* using a modified version of the procedure described by W. Pope, G. Sarkis, & G. Hatfull and Greg Broussard. Late-log stage *M. smegmatis* was diluted, plated on 7H10 agar seeded with high-titer Ukulele lysate. Resultant colonies were tested for lysogens by spot plating Ukulele lysate on bacterial lawns of potential lysogens. Absence of lysis, or superinfection immunity, was used as indication of true lysogens.

3.4 Polymerase Chain Reaction (PCR) amplification of Ukulele gp53

PCR amplification of Ukulele gp53 was carried out using 1 ng of Ukulele genomic DNA, 0.5 µM of pST-KT_gp53 forward and reverse primers (Table 3), and Q5 High-Fidelity Master Mix (NEB, Ipswitch, MA) in 25-µL reaction volumes per the manufacturer’s recommendations. Ukulele gp53 amplification reactions were incubated at 98°C for 30 sec followed by 35 cycles of 98°C for 10 s and 72°C for 45 s, and a final extension at 72°C for 2 min using a Techne TC-plus thermal cycler (Bio-techne, Minneapolis, MN).

<table>
<thead>
<tr>
<th>Primer ID</th>
<th>Sequence 5’ – 3’</th>
</tr>
</thead>
<tbody>
<tr>
<td>pST-KT_gp53_1</td>
<td>CGCAATGCACCACCCACCACCCACCACATGCTTGGACAGACGCAAGCGCGTGGAGCTGTCGGAGTGCA</td>
</tr>
<tr>
<td>pST-KT_gp53_2</td>
<td>CTTGTCGTGCTGCTCTTTGTACTACAGCGAGAGAGCGTCTCGCGGTAC</td>
</tr>
<tr>
<td>gp53ORF_F</td>
<td>GTGTGTGCAATGGGGGTCA</td>
</tr>
<tr>
<td>gp53ORF_R</td>
<td>TCTGAACCTTTAAGATTGGAA</td>
</tr>
</tbody>
</table>

Nucleotide sequences specific to the gp53 sequence are indicated in blue for Gibson cloning reactions. Primers were designed with specificity to both gp53 (blue) and the pST-KT sequence (red).

PCR amplification of recombinant pST-KT plasmid containing Ukulele gp53 recovered from *E. coli* transformants was carried out using 100 ng recombinant plasmid DNA, 0.5 µM of pST-KT_gp53 forward and reverse primers (Table 3), and Q5 High-
Fidelity Master Mix (NEB) in 25-μL reaction volumes per the manufacturer’s recommendations. Recombinant plasmid PCR reactions were incubated at 98°C for 30 sec followed by 35 cycles of 98°C for 10 s and 72°C for 45 s, and a final extension at 72°C for 2 min using a Techne TC-plus thermal cycler (Bio-techne).

3.5 Plasmid preparation

Plasmid preparations were performed on 5-mL cultures of *E. coli* using a Wizard Plus SV Minipreps DNA Purification kit (Promega, Inc. USA, Madison, WI) according to the manufacturer’s recommendations. Plasmid pST-KT is an inducible bacterial expression plasmid containing a kanamycin resistance gene, an N-terminal histidine tag, and a C-terminal FLAG tag. Plasmid DNA was stored in TE buffer (10 mM Tris-HCl; 1 mM EDTA, pH 7.5) and was quantified via spectrophotometry using a Nanodrop ONE spectrophotometer (Nanodrop Technologies, Rockland, DE).

3.6 Restriction endonuclease digestion (RD)

Restriction endonuclease double digests were performed on pST-KT plasmid DNA in preparation for Gibson Cloning Assembly reactions. Double digests were carried out using CutSmart Buffer, 19.47 μg plasmid DNA, 100 units of BamHI-HF, and 160 units of HindIII-HF per the NEB protocol (NEB). Digests were incubated at 37°C for 3 h and HindIII-HF was heat-inactivated at 67°C for 20 min.

3.7 Agarose gel electrophoresis

Successful PCR amplification of Ukulele gp53 was confirmed by separation on a 1.2% Seakem LE (Lonza, Rockland, ME) agarose gel in 1X TAE Buffer (40 mM Tris; 20 mM acetic acid; 2 mM EDTA). Gels were stained using Sybr Safe DNA Gel Stain.
gp53 PCR amplicons were purified on 3% NuSieve GTG Low-melt (Lonza) agarose gels in 1X TAE Buffer and extracted using a QIAquick DNA Gel Extraction Kit (QIAGEN, Hilden, Germany) according to the manufacturer’s directions. Band sizes were determined by comparison to NEB 1 kb Plus DNA Ladder (NEB).

Digested pST-KT was analyzed for removal of the 63 bp fragment between the BamHI and HindIII cut sites by separation on a 1.5% SeaPlaque Low-melt (Lonza) agarose gel in 1X TAE Buffer. Band sizes were determined by comparison to NEB 1 kb Plus DNA Ladder (NEB). Successfully digested plasmid was extracted using a QIAquick DNA Gel Extraction Kit (QIAGEN) per protocol.

3.8 Recombinant plasmid preparation

Recombinant pST-KT plasmids containing Ukulele gp53 were constructed using the Gibson Assembly Cloning protocol described by Gibson et al. 2009\(^{45}\). Cloning reactions were performed in Gibson Master Mix with 100 ng plasmid DNA in both 2:1 and 3:1 Ukulele gp53:pST-KT molar ratios in 20-µl total reaction volumes per the manufacturer’s recommendation (NEB). Reactions were incubated at 50ºC for 15 min and stored at -20ºC.

3.9 Bacterial transformations

Transformations of recombinant pST_gp53 were performed on NEB 5–alpha Competent E. coli cells using 2 µl of each (2:1 & 3:1) Gibson Assembly cloning reaction according to the manufacturer’s recommendation (NEB). Successful kanamycin-resistant
transformants were selected for via plating on L-agar containing 50 µg/ml of kanamycin. Transformant colonies were inoculated in L-broth with 50 µg/ml kanamycin. Plasmid DNA was recovered from 5-ml cultures using Wizard Plus SV Minipreps DNA Purification kit (Promega) according to the manufacturer’s specifications.

3.10 DNA sequencing

Recombinant pST-KT_gp53 plasmids were checked for successful insertion of Ukulele gp53 by sequencing at the University of Maine DNA Sequencing Facility (Orono, ME).

3.11 Total RNA isolation and RNA-seq

Ukulele M. smegmatis lysogens and uninfected M. smegmatis were grown at 37°C to an OD$_{600}$ of 1.0. Cells were pelleted and resuspended in phage lysate at a multiplicity of infection (MOI) of 3.0. Control M. smegmatis cells were resuspended in equal volumes phage buffer (10 mM Tris, pH 7.5; 10 mM MgSO$_4$; 68 mM NaCl; 1 mM CaCl$_2$). Cells were harvested in triplicate from the control flask at 0 min and from the phage-treated flask at 30 min and 2.5 h. Harvested cells were treated with 6 mL of RNAProtect Bacteria Reagent (QIAGEN) for 5 min. Cells were centrifuged 1 min at 5,000 x g and the supernatant was removed. Pellets were resuspended in 100 µL of RNase-free TE containing lysozyme. Resuspended pellets were incubated at room temperature for 10 min, following which 700 µL of RLT buffer (QIAGEN) was added. Samples were then transferred to 2.0-mL Lysing Matrix B tubes (MP Bio, Santa Ana, CA) and subjected to bead beating in ice-cold blocks of the TissueLyser (QIAGEN) for ten 20 s pulses of 30 Hz. Samples were then centrifuged for 1 min at 12,000 x g. Total RNA was isolated from
the samples using the RNAeasy Mini Kit (QIAGEN) per the manufacturer’s recommendations. Samples were treated with DNase twice. During isolation, RNA was treated with DNaseI on the column (QIAGEN). RNA was eluted in 50 L of water and samples were treated with Turbo DNaseI (Ambion, Foster City, CA). RNA was analyzed for quality by agarose gel electrophoresis and quantified by spectrophotometry on a Nanodrop ND-1000 (Nanodrop). Samples were sent to the Delaware Biotechnology Institute (Newark, Delaware) for ribo-depletion, quality control analysis, and paired-end RNA-seq library preparation. Libraries were sequenced by 50-bp read HiSeq Illumina sequencing (Illumina, Inc., San Diego, CA).

3.12 Transcriptomic analysis of Ukulele and *M. smegmatis*

Diagnostic analysis of FastQ data from each sample was performed on Galaxy (Penn State University). Reads were trimmed to remove specific adapters and low-quality reads using Trimmomatic\(^{46}\). Short reads were aligned to *M. smegmatis* and Ukulele reference genomes with Bowtie to generate SAM files\(^{47}\). SAM files were then converted to BAM files via samtools. The GenBank file for the Ukulele genome was converted into a GTF file using a custom Perl script\(^{29}\). Differential analysis of counts per read mapping to each feature (gene) was generated using htseq-count\(^{48}\). EdgeR was used to normalize reads by utilizing the trimmed mean of M values to estimate the dispersion based on a trended mean\(^{48}\). This generated tab-delimited text files complete with annotated lists of differentially expressed genes and their corresponding statistical significance. BAM files containing aligned reads were visualized using Integrative Genomics Viewer (IGV) to perform quality control of the Ukulele genome annotation and evaluate expression levels across samples\(^{49}\).
3.13 Reannotation of Ukulele genome

BAM files containing aligned transcript reads were visualized using IGV as a means of improving predictive annotations of open reading frames (ORFs) and transcriptional regulatory elements in the Ukulele genome. ORFs with predicted translational start sites upstream of corresponding transcript reads were analyzed using DNA Master for potential start sites downstream of the aligning reads. Coding potential, amino acid alignments and predicted protein structure of the ORF was evaluated using Glimmer, GeneMark, NCBI BLAST, HHpred, SCOPe, Pfam, NCBI_Conserved_Domains(CD), Phamerator to predict the most appropriate new translational start site. Regions of the Ukulele genome without predicted ORFs that aligned to transcript reads were analyzed using DNA Master, Glimmer, GeneMark, NCBI BLAST, HHpred, Phamerator to identify the correct ORF. IGV was used to assess transcriptional regulatory element predictions by identifying intergenic regions in the Ukulele genome where transcript reads initiated and terminated.
4.0 RESULTS

4.1 Translational start sites were changed for six predicted Ukulele genes based on RNA-seq data.

In order to improve the genome annotation of model cluster E mycobacteriophage Ukulele, RNA-seq analysis was performed and gene calls were realigned to better correspond with the transcriptomic data. By using the program Integrative Genome Viewer (IGV) to visualize transcript reads, predicted open reading frames (ORFs) were checked to ensure that translational start sites were downstream of the experimentally determined transcriptional start sites (figure 3). Across the Ukulele genome, six translational start sites were changed (Table 1). Of these six genes, five are reverse transcribed and one is forward transcribed, and only two have predicted functions (gp21, tail assembly chaperone; gp53, Cro-like or Xis) (Table 1). In five instances, the ORF had to be shortened to align with transcriptomic evidence (Table 1). Shortened ORFs ranged from 72 – 155 bp shorter than the original ORFs, with an average difference of 105 bp. In one case, gp141, the ORF was extended by 15 bp (Table 1). In cases of functional operons, carry-through of reads between multiple concurrently expressed genes occasionally made it difficult to discern the start of transcripts.
RNA-seq uncovered two new genes in the Ukulele genome: gp1 and gp59. In addition to adjusting translational start sites of predicted ORF’s, transcriptomic data was used to identify potential unpredicted ORF’s in Ukulele. A high number of reads mapped to three intergenic locations in the Ukulele genome indicating a potential gene that was not included in the original annotation. In relation to the numbering of the original annotation, these regions lie upstream of gp1, between gp57 and gp58, and...
between gp82 and gp83. Of these three regions, only two (pre-gp1 and gp57–gp58) showed significant evidence in support of a gene present at that location such as coding potential and structural homology to other known proteins (Table 2).

A gene was not inserted into the gap between gp82 and gp83 due to a lack of coding potential and significant similarity to any known proteins in the database. Though the region between gp82 and gp83 is highly conserved among all cluster E phage at the nucleotide sequence level, GeneMark detects no coding potential in this region (data not shown). The only possible ORF in this region is 111 nucleotides long and the putative gene product shows no significant amino acid sequence alignment or structural homology to any proteins in the database (data not shown). The start site of this potential ORF is also slightly upstream of the beginning of mapped transcript reads in this region (data not shown). Therefore, a new gene was not called in this region.

A previously unidentified gene upstream of Ukulele gp1 was identified by RNA-seq. The region upstream of gp1 is highly conserved across all cluster E phage at the nucleotide level and contains considerable coding potential per GeneMarkS and GeneMark M. smegmatis (Figure 4). There is a potential ORF from 10–271 bp, corresponding decently with the start and stop of the aligning transcripts (Figure 4B). This gene is called in 57 other cluster E mycobacteriophage, including Kostya, Goku, and 244, though a function is not predicted in any of these genomes (Figure 4C). The putative gene product of this ORF also has amino acid similarity to genes from various bacterial species including a LysR transcriptional regulator from Lactobacillus (data not shown). No conserved domains are detected and HHPred and Phyre² do not predict significant structural homology to any protein templates in the databases (data not shown).
A second unpredicted gene was added to the Ukulele annotation between gp57 and 58 (Table 1). This region is highly conserved across all cluster E phage at the nucleotide level, and contains a called gene in 30 cluster E phage (Figure 5C). Coding potential for the ORF (41,827–42,015) was detected by self-trained GeneMarkS, but not by M. smegmatis-trained GeneMark (Figure 5D). There is an ORF that aligns well with transcript reads mapping to this region (Figure 5B). The putative gene product of this ORF does not contain any conserved domains according to NCBI BlastP, and HHPred and Phyre² do not indicate significant structural homology to any proteins in the databases (data not shown). It is likely that an unpredicted gene lies between gp57 and gp58, however a function cannot yet be predicted for its gene product and further experimental confirmation of the gene product is required to determine its function.
Figure 4. An unpredicted gene upstream of gp1 was identified by RNA-seq. (A) RNA-seq transcript reads (grey) mapping to the region upstream of gp1 (blue bar) during early lytic, late lytic, and lysogenic growth. The height of RNA-seq indicates relative expression level with the scale listed in parentheses on the left. (B) Potential ORFs upstream of gp1 identified by DNA Master. Small and large tick marks indicate start and stop codons, respectively. Green bars indicate forward transcribed genes with the gene number listed above. One ORF (green line) aligns with RNA-seq transcript reads. (C) A Phamerator map comparing the left arm of the Ukulele genome to that of cluster E mycobacteriophage 244. Boxes indicate putative genes and are colored by Phamily. The background is colored by nucleotide sequence similarity, with purple indicating >99% similarity. 244 and other closely related cluster E phage have a gene called in this region. (D) Typical (black) and atypical (red) coding potential on the plus strand of the region upstream of Ukulele gp1 generated by GeneMarkS. ORFs are represented as black lines with upticks and downticks indicating start and stop codons, respectively. Black bars represent genes predicted by GeneMarkS. GeneMarkS predicts a gene in this region, and all coding potential is captured by the ORF identified by
Figure 5. An unpredicted gene between gp57 and gp58 was identified by RNA-seq. (A) RNA-seq transcript reads (grey) mapping to the region between forward genes 57 and 58 (blue bars) during early lytic, late lytic, and lysogenic growth. The height of RNA-seq reads represents relative number of transcripts mapping to the region (0–10,000). (B) Potential ORFs between gp58 and gp59 identified by DNA Master. Small and large tick marks indicate start and stop codons, respectively. Green bars indicate forward transcribed genes with the gene number listed above. One ORF (green line) aligns well to RNA-seq transcript reads. (C) A Phamerator map comparing this region in the Ukulele genome to the corresponding region in cluster E mycobacteriophage Goku. Boxes indicate putative genes and are colored by Phamily. The background is colored by nucleotide sequence similarity, with purple indicating >99% similarity. Goku and other closely related cluster E phage have a gene called in this region. (D) Typical (black) and atypical (red) coding potential on the plus strand of the region between Ukulele gp57 and gp58 generated by GeneMarkS. ORFs are represented as black lines with upticks and downticks indicating start and stop codons, respectively. Black bars represent genes predicted by GeneMarkS. All typical coding potential is captured by the ORF identified by DNA Master.
4.3 Sixteen predicted intrinsic terminators align with termination of RNA-seq reads

Predictions of transcriptional regulators in cluster E genomes were improved by analyzing transcriptomic RNA-seq data for evidence of regions in which transcription terminates abruptly. These regions were compared to previously predicted intrinsic terminators identified in the Ukulele genome for tentative confirmation of their function in vivo. Sixteen of the nineteen predicted intrinsic terminators in the Ukulele genome aligned well with genome regions to which RNA-seq reads diminished abruptly (Figure 1) (Table 2). As described by Belisle previously, terminators transcribed on the forward strand of the Ukulele genome are denoted by UT and terminators transcribed on the reverse strand are denoted by URT28.

Of these 16 tentatively confirmed terminators, 14 (87.5%) are I-type lacking a poly-U trail (Figure 6) and two are L-type (Figure 7) (Table 2). Additionally, there is one X-type pair of convergent terminators (UT1 and URT1) and one U-shaped pair located in tandem (≤ 50 bp apart) downstream of gp52 (URT5 and URT6). Seven of these terminators are on the forward strand (UT) of the Ukulele genome and the other nine are located on the reverse strand (URT). Forward terminators lie downstream of genes 6, 13, 19, 20, 29, 30, 32, 38, and 51 (Table 2). Reverse terminators are located downstream of genes 7, 39, 49, 52, and 124 (Table 2). They are all located in intergenic regions and lie between 4 (UT3) and 240 (URT4) bp downstream of the gene they follow with an average distance of 84.56 bp (Table 2). The ΔG values of these terminators range from -27.21 (UT2) to -16.2 (URT5) kcal mol⁻¹ with an average of 20.87 kcal mol⁻¹ (Table 2). They have an average stem length of 15.12 bp with fewer than three mismatches (Table 2). Both L-type terminators (UT3 and URT9) have 4 U’s in the poly-U trail (Table 2).
Figure 6. RNA-seq analysis supports the presence of functional I-type terminators in the Ukulele genome. (A) RNA-seq reads (grey) were analyzed for regions of abrupt termination within 250 nt downstream of genes. Blue bars represent putative genes with gene numbers listed below and white arrows indicating orientation. The height of transcript reads represents the number of mapped transcripts with scales listed in parentheses on the left. (B) The intergenic region between gp13 and gp14 with bases colored by nucleotide identity generated by Geneious. Yellow and red bars denote putative genes and terminators, respectively. For L-type terminators, the poly-U trail is indicated as a labeled red bar separate from the stem-loop. (C) Predicted secondary structure of the putative terminator downstream of gp13 (UT2) generated by mFold. Base pairing probabilities are color coded high–low according to the rainbow.
RNA-seq analysis of the Ukulele transcriptome revealed two regions in the genome where transcription termination occurred abruptly but a terminator had not been predicted (Table 2). These occurred downstream of gp19 and gp37 of the original annotation (now gp20 and gp38) (Table 2). The nucleotide sequences of these regions were analyzed manually for GC-rich inverted palindromic sequences that may form stem-loop structures upon secondary folding. One strong terminator was identified in each.

4.4 RNA-seq revealed two unpredicted sites of transcription termination

RNA-seq analysis supports the presence of functional L-type terminators in the Ukulele genome. (A) RNA-seq reads (grey) were analyzed for regions of abrupt termination within 250 nt downstream of genes. Blue bars represent putative genes with gene numbers listed below and white arrows indicating orientation. The height of transcript reads represents the number of mapped transcripts with scales listed in parentheses on the left. (B) The intergenic region between gp38 and gp39 with bases colored by nucleotide identity generated by Geneious. Yellow and red bars denote putative genes and terminators, respectively. For L-type terminators, the poly-U trail is indicated as a labeled red bar separate from the stem-loop. (C) Predicted secondary structure of the putative forward terminator downstream of gp38 (UT8) generated by mFold. Base pairing probabilities are color coded high–low according to the rainbow.
region (UT4 and UT8), both being L-type with strong poly-U tails (≥ 5 U’s) (Table 2). UT4 and UT8 lie 20 and 72 bp downstream of gp20 and gp38 with ΔG values of -35.00 and -37.80 kcal mol⁻¹, respectively (Table 2). Both have a stem length of 24 bp with fewer than three mismatches (Table 2).

4.5 Recombinant plasmids containing Ukulele gp53 may be lethal to bacterial transformants due to cryptic promoters

Because previous data from our lab indicates that gp53 is required for lytic growth and not lysogeny, its function in regulating phage lifecycles was investigated through attempts to generate a strain of *M. smegmatis* containing Ukulele gp53. Construction of *M. smegmatis* strains containing gp53 were attempted using the anhydrotetracycline-inducible mycobacterial plasmid pST-KT and the Gibson Assembly method for molecular cloning. Attempts using the new shortened ORF were unsuccessful, as recombinant plasmids were not able to be recovered from transformants. In one case, 13 transformant colonies were recovered, however these only contained the 30 bp of the pST-KT_gp53 F primer corresponding to the 5’ end of gp53, not the entire ORF. Primer sequences were correct and diagnostic gel electrophoresis indicates that PCR amplification of Ukulele gp53 was successful (Table 3). Restriction digests by *Bam*HI and *Hind*III were successful, demonstrated by the presence of a band corresponding to the excised plasmid fragment in gel electrophoresis.

To determine whether the inability to successfully recover recombinant plasmid was the result of difficulties in cloning or transformation reactions, construction of an *M. smegmatis*_gp53 strain was attempted by ligation using the general TA cloning vector pGEM-T Easy. Transformations were unsuccessful and no colonies were observed.
Primer sequences were correct (Table 3), and diagnostic gel electrophoresis indicated successful PCR amplification of Ukulele gp53.

In order to investigate potential reasons plasmids containing Ukulele gp53 may be lethal to transformants, the 5’ region of gp53 was examined for putative weak promoters that may drive cryptic expression \textit{in-vivo}. Two weak promoters were identified by the program PhiSITE as having motifs captured by the pST-KT\_gp53\_F primer (Figure 9) that may initiate unwanted expression in the absence of the inducer, anhydrotetracycline.
Figure 8. Ukulele gp53 is likely not the cluster E immunity repressor. (A) A phamerator map showing the gp53 equivalent (purple) across the cluster E phages ABCat, Asriel, Barbarian, and Ukulele. Boxes above and below the genome ruler indicate forward- and reverse-transcribed genes, respectively, with putative gene product functions listed above or below the corresponding gene. Genes are colored by Phamily with the Pham number listed above or below the gene and the number of Phamily members listed in parentheses. Many cluster E phage have predicted the gp53 equivalent to encode the immunity repressor. (B) The predicted secondary and tertiary folding structure of gp53 generated by Phyre². The ribbon diagram is colored according to the rainbow from N–C terminus. Ukulele gp53 contains a winged helix-turn-helix DNA-binding domain and most closely resembles Xis from mycobacteriophage pukovnik. (C) RNA-seq transcript reads aligning to gp53 (blue bar) during early lytic, late lytic, and lysogenic growth. Reads are scaled 0–750 for all three growth phases. Ukulele gp53 is most highly expressed during early lytic growth.
Figure 9. Ukulele gp53 may encode two distinct proteins. (A) The sequence of Ukulele gp53 corresponding to the updated translational start site at 39,711 (yellow) and the first start site downstream of the cryptic promoters at 39,663 (dark blue) generated by Geneious. Two putative cryptic promoters (green) are present in the 5’ region of gp53 and overlap the pST-KT_gp53_1 primer (light blue) used in Gibson cloning reactions. (B) The first start codon downstream of the cryptic promoters (blue arrow) captures all typical coding potential (black peak) detected by GeneMarkS. (C) An increase in RNA-seq read counts (grey) aligning to gp53 (blue bar) is observed just upstream of the first stop codon at 39,663 (blue arrow). (D) The gene product produced from translation of the start codon at 39,663 shows low confidence (64.1%) structural homology to the crystal structure of YabA from Bacillus subtilis according to Phyre2. The 3D ribbon diagram is colored N-terminus to C-terminus according to the rainbow.
5.0 DISCUSSION

The evolutionary relationship between pathogenic bacteria and prophage has been documented for quite a while\textsuperscript{51}, however very few phage-host relationships have yet been well characterized\textsuperscript{5,11}. Though some prophages confer pathogenicity directly by encoding a toxin\textsuperscript{18}, many do not encode an obvious toxin but instead contribute incrementally to host fitness\textsuperscript{12}. Among the poorly understood phage-host relationships are those involving mycobacterial pathogens such as \textit{M. tuberculosis} and \textit{M. leprae}, the causative agents of Tuberculosis and Leprosy, respectively\textsuperscript{12}. In order to better characterize these pathogens and their evolution, it is crucial to understand the viral genomes that impact their physiology and fitness\textsuperscript{3,5,11}. This study was aimed to enhance the understanding of cluster E mycobacteriophage through improving genome annotations via transcriptomic analysis by RNA-seq using Ukulele as a model. Analysis of the Ukulele transcriptome has helped better the understanding of cluster E mycobacteriophage genomes by providing experimental evidence for improving predictions of translational start sites, previously unidentified genes, transcriptional regulators, and gene functions.

5.1 Cluster E genomes contain false predictions that can be improved by RNA-seq analysis

A significant problem with predictive computational genome annotations is the ability for incorrect translational start calls to become reiterated in other related genomes if the genome containing the incorrect call is used as a model for the annotation of subsequent genomes. Typically, genome annotations are performed by using a similar genome in the database (if one is available) as a model, and thus an error in the first model genome can become prolific as the number of similar genomes sequenced
increases. For example, the program Starterator predicts the most likely translational start site for an ORF based on what is most often called in other similar genomes\textsuperscript{55}. This also applies to the practice of calling the longest ORF whenever possible in order to eliminate ostensibly unnecessary gaps from the genome\textsuperscript{21}. However, this is problematic if the calls in other genomes have not been experimentally determined and are incorrect predictions. Transcriptomic analysis by RNA-seq provides one piece of experimental evidence that can be used to validate or improve gene calls, which can then be tentatively applied to other phage with similar genomes.

Analysis of the Ukulele transcriptome via RNA-seq revealed eight instances in which the original genome annotation did not align with experimental data (Table 3). Most of these discrepancies (five) were cases where a gene had been called, but the predicted translational start site was upstream of the beginning of transcript reads (Table 3). Notably, three of these five genes are located in regions where a need for transcriptional promoters may require additional intergenic space. Two shortened genes (gp8 and gp53) lie at regions of divergent transcription and one (gp21) is located directly downstream of a putative terminator (Figure 1). These intergenic regions are much larger than is typically thought to be required for mycobacterial promoters, though.

The canonical structure of mycobacterial promoters typically only requires a $\sim$50 bp gap to contain the $-10$ and $-35$ motifs\textsuperscript{35,56}, with regions of divergent transcription requiring a $\sim$100 bp gap to account for a promoter in each direction. However, the intergenic gaps observed between genes 8 and 9 (183 bp) and 53 and 54 (310 bp) are much larger than 100 bp and the gap between genes 20 and 21 (121 bp) is much larger than 50 bp. In each of these cases, it is possible that initial annotations overlooked the
need for a promoter at these locations or that a larger intergenic gap is required than is obvious from the size of mycobacterial promoter sequences alone.

In the instance of gp20, it is likely that this large gap accounts for both the putative intrinsic terminator and the promoter sequence. The sequence of the terminator in this location constitutes 77 nt of the gap in this region (Table 1). Together, the ~50 nt needed for a promoter sequence and the 77 nt containing the putative terminator account for the entirety of the 121 bp gap observed between the 3′ end of gp20 and the updated start site for gp21. Though the transcriptional regulatory elements in this region have yet to be experimentally confirmed, this layout complies with the general trend of mycobacteriophage genomes being extremely compact and spatially efficient\textsuperscript{1,2,6}. As these genes are both transcribed in the same direction and belong to the same functional operon, this region is not an obvious candidate for transcriptional regulatory elements and thus the need for this gap was likely overlooked in the original annotation. Therefore, it is important to use RNA-seq analysis to improve genome annotations within functional operons of genes transcribed in the same direction.

Annotations for regions of divergent transcription can also benefit from RNA-seq analysis. Regions of divergent transcription require a promoter in either direction and therefore need a ~100 bp intergenic gap\textsuperscript{21,57}. However, the gap between gp53 and gp54 is more than 3x larger than this (310 bp). The larger gap between divergently transcribed gp53 and gp54, in contrast to gp20, may be required to accommodate a different kind of regulatory sequence: an operator. Being that gp53 is located within the integration cassette at a site of divergent transcription and is predicted to encode some component of the transcriptional regulatory mechanism responsible for regulating phage lifecycles, it is
possible that there are operator-binding sites located upstream of gp53 similar to Lambda OR\textsubscript{1,2,3}. In Lambda, these operator sites are 17 bp long and are spaced only a few nucleotides apart\textsuperscript{15}. A series of 16-bp conserved repeat sequences has previously been reported in the Ukulele genome that share multiple characteristics with known operator sequences, two of which (CR1-2 and CR1-3) lie just upstream of gp53\textsuperscript{22}. In support of their potential function as operator sequences, these conserved repeats are in the same orientation as the downstream genes and overlap their respective putative promoters\textsuperscript{58}. A third similar conserved repeat (CR1-4) is also located downstream of gp58\textsuperscript{22} which could indicate that at one point gp53 and gp58 were adjacent to one another with a functional operator dividing them, and that a recombination event inserting gp54–gp57 disrupted this structure and introduced the large gap between gp53 and gp54. However, the integration cassette is a highly important operon for temperate phage like Ukulele, and disruption of the transcriptional regulatory machinery for these genes would likely have deleterious effects on the ability of the phage to properly regulate its lifecycles\textsuperscript{6}. Since Ukulele forms stable lysogens, this explanation for the large gap between gp53 and gp54 revealed by RNA-seq data is not likely unless: gp53 does not encode a repressor or Cro-like protein necessary for regulating phage lifecycles; this operator does not function by the same mechanism as the one characterized in lambda and thus does not rely on the binding sites being adjacent; or the insertion of gp54–gp57 introduced a more effective mechanism for regulating lysogeny than the one that previously existed. Though it is unclear if the gap between Ukulele gp53 and gp54 plays a role beyond housing a pair of divergent promoters, RNA-seq analysis has revealed that it is larger than originally predicted, denoting it as a site of interest for further study and highlighting the
importance of complementing predictive genome annotations with experimental transcriptomic evidence. Similarly, the gap between genes 8 and 9 is nearly twice as large as is required for a set of divergent promoters but does not contain a terminator, conserved repeats, or any other obvious feature that would require an extended gap. However, the transcriptomic data supporting this large gap suggests the need for further investigation that would have otherwise been overlooked and again demonstrates the importance of using RNA-seq to improve genome annotations so that a lack of understanding doesn’t lead to incorrect assumptions and faulty predictions.

The two unpredicted genes identified in the Ukulele genome by RNA-seq also directly demonstrate the ability for transcriptomic analysis to circumvent some of the limitations presented by predictive annotations. These genes were likely missed in preliminary annotations because they are smaller than the lower limit described by the SEA-PHAGES Bioinformatics Annotation Guide. Typically, ORF’s encoding an amino acid sequence less than 100 residues long are considered too small to call, and the putative gene products encoded by gp1 and gp59 have amino acid sequences that are only 88 and 63 residues, respectively. However, RNA-seq data, coding potential output, and amino acid sequence similarity to conserved domains in the NCBI database suggest that these are in fact genes. Thus, RNA-seq provides a valuable tool not only for identifying genes that would otherwise not be predicted, but also for updating the parameters used for predictive genome annotations in the future. RNA-seq also has its limitations since not every set of reads necessarily corresponds to a gene, demonstrated by the reads observed between gp82 and gp83. However, as long as RNA-seq is combined with computational analyses, problems arising from this limitation can be minimalized.
5.2 Transcription termination occurs primarily via I-type intrinsic terminators in cluster E 
mycobacteriophage

Though there is ongoing debate about the ability for I-type intrinsic terminators to effectively terminate transcription in prokaryotic systems\textsuperscript{31,40}, the majority (78\%) of intrinsic terminators identified in Ukulele are I-type (Table 2). Some studies have reported the need for a partially intact poly-U trail\textsuperscript{40}, however only five of the 14 identified I-type terminators contain a partially intact poly-U trail (figure 6) while the remaining nine completely lack any distinguishable poly-U trail. This supports previous findings that I-type intrinsic terminators lacking a poly-U are not only functional for disrupting transcription elongation by mycobacterial RNA polymerases\textsuperscript{30,31}, but also that they represent a majority of mycobacterial intrinsic terminators\textsuperscript{30}. Though these previous studies have focused on mycobacterial hosts rather than mycobacteriophage, the reliance on host RNA polymerase for transcription of viral genes likely subjects the phage genome to the same transcriptional regulatory elements as its hosts. Therefore, it is likely that both mycobacteriophage and their mycobacterial hosts do in fact rely in I-type terminators for rho-independent termination, regardless of the presence of a partial poly-U trail.

The study previously reporting the need for a poly-U trail that is at least partially intact utilized \textit{M. bovis} RNA polymerase that was purified after being overexpressed in \textit{E. coli}\textsuperscript{40}. Though measures were taken to reduce the chance for interaction between the purified \textit{M. bovis} RNA polymerase and subunits of the \textit{E. coli} RNA polymerase\textsuperscript{40}, it is possible that there were some residual interactions driving this specific mycobacterial RNA polymerase to function more like an \textit{E. coli} RNA polymerase to some extent. This
could contribute to the discrepancy, being that transcription elongation by *E. coli* RNA polymerase occurs at a faster rate than does elongation by mycobacterial RNA polymerases\(^40\). A slower rate of elongation may allow the stem-loop to have a greater impact on the stalling of the elongation complex, facilitating its disruption without the need for a U-rich sequence to weaken the binding of the DNA-RNA hybrid\(^40\). This is reflected by the fact that even though Czyz et al. reported the necessity of a partial poly-U for both mycobacterial and *E. coli* intrinsic terminators, this requirement was less pronounced for disruption of elongation by the mycobacterial RNA polymerase\(^40\).

These findings suggest that I-type terminators are likely the dominant system for intrinsic transcription termination in mycobacteriophage and their hosts, but also highlight the idea that transcription regulation mechanisms are markedly diverse across microbial systems and that transcriptome analysis is a useful approach for investigating this diversity. Though it is possible that RNA-seq evidence of transcription termination at some of these regions was due to factors beyond intrinsic termination such as factor-dependent termination or an error in read-mapping, the strong alignment of these regions to previously predicted or manually identified intrinsic terminators suggests that this is not the case and that intrinsic termination by I-type terminators does indeed occur in mycobacteriophage.

Although they are less abundant, L-type terminators also appear to be effective in terminating transcription elongation by mycobacterial RNA polymerase (Table 2) (Figure 7). However, the L-type terminators identified in the Ukulele genome do not all comply with the characteristics of canonical L-type intrinsic terminators previously reported for *Mycobacteria*\(^30,31,32\). Of the four L-type terminators identified in this study, two (UT4 and
UT8) have ΔG values that are significantly more negative than the range of –15 to –25 kcal mol\(^{-1}\) reported by Mitra et al\(^{30}\). The GC-rich stems of these two also exceed the 17-bp stem length previously reported for mycobacterial intrinsic terminators (Table 2)\(^{30}\). These were also the only two terminators identified by RNA-seq that were not previously predicted (Table 2), likely because of these deviations from the canonical structure of intrinsic terminators. Transcript termination in the vicinity of these two mammoth L-type terminators was very strong however, heavily implicating their presence and function in the Ukulele genome. The identification of these two noncanonical L-type terminators through transcriptome analysis by RNA-seq demonstrates the validity of this approach in investigating novel transcriptional regulatory elements that cannot be predicted by computational analysis due to their deviation from the parameters of predictive algorithms.

By utilizing RNA-seq transcriptomic analysis of Ukulele to investigate transcriptional regulatory patterns, this study aimed to improve the understanding of intrinsic termination in mycobacteriophage as well as provide evidence for the validity of this method in complementing predictive strategies for identifying functional transcriptional regulatory elements. This was accomplished through the confirmation of 16 previous predicted intrinsic terminators and the identification of two unpredicted ones. However, there are some limitations to this approach. For example, regions of convergent transcription without a large gap are difficult to assess for termination due to the run-through of reads from either strand. This was a problem experienced for the intergenic gap between the forward-transcribed gp30 and the reverse-transcribed gp31. In order to deal with this, forward and reverse reads were visually separated by color and the region
was examined for a decrease of reads in either direction. As the expression level of gp30 was greater than that of gp31, a drop in forward reads indicating transcription termination was easier to discern and thus UT6 was identified but a reverse terminator was not. In order to determine whether a reverse terminator lies between gp30 and gp31, forward and reverse reads should be separated computationally rather than visually. Applying this approach in future studies can further improve the ability for transcription regulation to be investigated by transcriptome analysis. Further, even though transcriptome analysis by RNA-seq provides experimental evidence for regions of transcription termination, it does not prove the function of putative terminators. These putative terminators still must be experimentally proven to effectively terminate transcription elongation by RNA polymerase.

5.3 Cryptic expression of Ukulele gp53 is lethal to host cells

RNA-seq analysis of Ukulele gp53 has revealed that it is most highly expressed during early lytic growth, suggesting that it functions to help regulate the lytic lifecycle. However numerous attempts to test this by constructing a strain of *M. smegmatis* overexpressing Ukulele gp53 have been unsuccessful, suggesting that gp53 may be lethal to host cells. This effect was observed in the absence of the inducer anhydrotetracycline which indicates that if gp53 is in fact killing the cells, then it is likely being expressed from a promoter that does not originate from the Tet-inducible promoter of the pST-KT vector. Two weak promoters were identified in the 5′ region of gp53 that may be responsible for this cryptic expression. One of these promoters has a –35 motif that perfectly matches mycobacterial consensus sequence (5′–TTGACg/a–3′) and a –10 motif that weakly matches the mycobacterial consensus (5′–TATAaT(TGn)–3′) but
contains a strong −10 TGn extension motif (5′−GGTGCT(TGG)−3′) (Figure 9). It is possible that expression of gp53 is being driven by this promoter sequence in undetected transformants, and that the DNA-binding gene product has lethal effects on these cells.

If expression of gp53 is occurring from one of these cryptic promoters, then the gene product being produced has a translational start site even further downstream than the updated start codon identified in this study. An alternative start codon (ATG) at 39,663 is a strong candidate for the translational start site utilized by these promoters as it captures all typical coding potential (figure 9). RNA-seq data also shows a second increase in transcript reads that lines up with these promoters (Figure 9). Though it’s possible that this second rise in transcript reads is simply an artifact of the sequencing and/or read mapping process, this could indicate that expression from these promoters occurs in vivo.

It is possible that gp53 has multiple promoters and transcriptional start sites. The first promoter and translational start at 39,711 allows synthesis of a 116-amino acid sequence with an N-terminal DNA-binding domain. If the cryptic promoter is active, translation would begin at 39,663– the first start codon to include all coding potential (figure 9). However, this protein would lack the characteristic winged helix-turn-helix DNA-binding domain (Figure 9). Upon losing this feature, the gene product shows structural homology to the N-terminal helical bundle of the negative regulator of DNA-replication in Bacillus subtilis, YabA (figure9). YabA is a protein that binds both DnaA and DnaN at the replication fork and inhibit DnaA from unwinding the DNA and initiating replication, likely through stimulating the intrinsic ATPase activity of DnaA61,62. Because it functions to inhibit the initiation of DNA replication,
overexpression of YabA causes a significant decrease in replication\textsuperscript{61}. If the shortened gene product of Ukulele gp53 resulting from the start codon at 39,663 does truly encode a YabA-like protein, then overexpression driven by the identified cryptic promoters may be lethal to transformants by inhibiting host replication initiation. In order to test this hypothesis, future studies of Ukulele gp53 will aim to clone the shorter version the gene without capturing the cryptic promoters. If transformations are successful and not lethal in the absence of inducer, it is likely that cryptic overexpression of this gene product resulting from the identified weak promoter(s) was the cause of lethality in this study. Further, it would be expected that induction by anhydrotetracycline would result in the death of transformants. If induction by anhydrotetracycline is not lethal to transformants, superinfection immunity assays will be carried out to further investigate the role of gp53 in regulating phage lifecycles. If gp53 does encode two distinct proteins, it is unknown what role they might play in regulating lysogeny and further experiments would be required.

5.4 Conclusions and future directions

By using RNA-seq to improve the annotation of the model cluster E mycobacteriophage Ukulele, this study has both demonstrated the benefits of transcriptomic analysis for improving genome annotations and furthered the understanding of cluster E mycobacteriophage genomics. Putative translational start sites, gene functions, and transcriptional regulators have been updated and previously unpredicted genes have been identified based on transcriptomic experimental evidence. Though the function of Ukulele gp53 is still unknown, it likely does not encode the immunity repressor, and may even encode two distinct proteins.
Future experiments will be needed to confirm the function of Ukulele gp53 and to investigate the role of gp59 in lysogeny maintenance. To investigate the function of the cryptic promoters, gp53 will be cloned into *M. smegmatis* starting with the start codon at 39,663 that does not capture the cryptic promoters. If this is unsuccessful, gp53 will be cloned into *M. smegmatis* in halves to determine if a presumably non-functional version of the gene product is also lethal to transformants. To investigate the function of the newly added gp59, statistical analyses will be employed to determine during which growth phase it is most highly expressed, and it will be cloned into *M. smegmatis* for subsequent superinfection assays. Additionally, future experiments are needed to confirm the function of the intrinsic terminators identified in this study.
REFERENCES

11. Hendrix, RW, Hatfull, GF, Ford, ME & (Second …, M. Evolutionary relationships among diverse bacteriophages and prophages: all the world’s a phage. ... *Gene Transfer (Second ...* (2002). at <https://www.sciencedirect.com/science/article/pii/B9780126801262500165>


55. https://github.com/SEA-PHAGES/starterator


58. Molloy, S.D., personal communication.
AUTHOR’S BIOGRAPHY

Jackson R. Foley was born in Portsmouth, NH on December 25, 1995. He was raised in Eliot, ME and graduated from Marshwood High School in 2014. At the University of Maine, Jackson majored in biochemistry and had a minor in neuroscience. During his time at UMaine he earned the John M. Steinmetz Award, the Frederick H. Radke Memorial Award, the School of Biology and Ecology First and Second Year Academic Awards, and was awarded the INBRE Junior Year Research Fellowship in Functional Genomics and the INBRE Senior Honors Fellowship.

Upon graduation, Jackson plans to begin a job as a research technologist at Johns Hopkins University before returning to graduate school to earn an advanced degree in neuroscience.