Spring 5-2016

Characterization of Transcriptional Control Elements in Cluster E Mycobacteriophage Ukulele

Campbell Belisle Haley
University of Maine

CHARACTERIZATION OF TRANSCRIPTIONAL CONTROL ELEMENTS IN CLUSTER E MYCOBACTERIOPHAGE UKULELE

By

Campbell Belisle Haley

Thesis Submitted in Partial Fulfillment
of the Requirements for a Degree in Honors
(Biochemistry and Spanish)

The Honors College

University of Maine

May 2016

Advisory Committee:

Sally D. Molloy, Advisor, Assistant Professor, Honors College and Department of Molecular and Biomedical Sciences
Keith W. Hutchison, Professor Emeritus, Department of Molecular and Biomedical Sciences
Melissa Maginnis, Assistant Professor, Department of Molecular and Biomedical Sciences
Dorothy Croall, Professor, Department of Molecular and Biomedical Sciences
David Gross, Adjunct Associate Professor in Honors

Abstract

Mycobacteriophage (phage) are a diverse group of viruses that infect *Mycobacterium*. Their study allows further understanding of viral evolution and genetics. Phage tightly control gene expression and transcribe their genes using host RNA polymerases. This project identifies potential transcriptional control elements in the genome of mycobacteriophage Ukulele. Promoters are sequences of the genome that allow binding of RNA polymerase and initiation of transcription. 21 putative promoters were identified in the Ukulele genome. To confirm transcriptional activity from putative promoters, a GFP reporter system was developed in mycobacterial cells. Intrinsic terminators are mRNA sequences that form secondary structure during transcription and stall the RNA polymerase. In Ukulele, 19 terminators were identified computationally. Future research includes confirmation of these terminators. Identification of elements that control transcription allows better understanding of phage gene expression.

ACKNOWLEDGMENTS

- Dr. Sally Molloy and Dr. Keith Hutchison, for dedicating a remarkable amount of time to all of the students in our laboratory, and for supporting me from my first day of college until my last.
- Dr. Dorothy Croall, Dr. David Gross, and Dr. Melissa Maginnis for their guidance on this thesis journey.
- Emily Whitaker, Emily Illingworth, Katrina Harris, Gwendolyn Beacham, and all other past and present members of the Hutchison-Molloy laboratory.
- The University of Maine Honors College for supporting undergraduate research education
- The University of Maine Department of Molecular and Biomedical Sciences for challenging me and demanding excellence from all of their students
- SEA-PHAGES program, Dr. Graham Hatfull, and all other professors committed to mycobacteriophage research.

A thanks to all of the following organizations for funding this research:

- The University of Maine Center for Undergraduate Research
- NIH-INBRE Grant 8P20GM1003423-12
- The INBRE Junior Year Research Award

Table of Contents

1. Introduction

Bacteriophage, or viruses that infect bacteria, are the most abundant biological entity on earth. This group of viruses is remarkably diverse, and many aspects of phage biology have been successful models for describing other biological systems. Mycobacteriophage, which infect mycobacteria, have been studied extensively to learn about the lifestyle of their mycobacterial hosts and about viral evolution [1] . Mycobacteria are of specific interest because this genus contains the infectious agents that cause tuberculosis and leprosy [2]. Mycobacteriophage are valuable tools for studying molecular biology and provide insight into the diversity of systems that control phage infection and gene expression. Despite these advances, many aspects of mycobacteriophage gene expression, including transcriptional regulation, remain poorly understood.

The diversity of these phage represents a high potential for discovery of novel transcriptional control mechanisms [1,3] Many models that describe transcriptional regulation have been developed in organisms such as *Escherichia coli* and mycobacteria [2,4,5]. Other studies of transcription have taught us about the possible complexities in these systems, and about the DNA sequence elements that appear to control transcription. Promoters and terminators are well-studied transcriptional control elements in other organisms. However, these haven't been fully characterized in mycobacteriophage and there is limited knowledge related to identification and characterization of these sequences.

Gene expression begins by initiating transcription or messenger RNA (mRNA) synthesis at DNA sequences, called promoters. Promoter sequences are located 10 and 35 bases before the start of transcription (TSP) [4]. The sigma protein bound to RNA polymerase recognizes these sequences and allows transcription initiation at the TSP. Promoters have been identified and confirmed in *E. coli* and many strains of mycobacteria. Promoter search in mycobacteriophage remains biased towards canonical *E. coli* promoters recognized by sigma protein 70, and very little has been done to identify and characterize mycobacteriophage-specific promoters.

Transcription is terminated in the presence of terminator sequences downstream of stop codons of protein coding regions. These sequences consist of GC-rich hairpins that form secondary structure and U-rich sections of DNA that follow these hairpins [6]. Promoters with poly-U tracts are known as L-type terminators. In addition to L-type terminators, there is evidence of the ability of terminators to function without a poly-U tract in some phage [7]. Promoters without poly-U tracts are known as I-type terminators. There is also evidence of tandem, U-type terminators following *E. coli* rRNA genes [5]. There have been limited studies focusing on mycobacteriophage terminators, and in these studies only L-type terminators were considered.

This study aims to optimize computational analyses for terminator and promoter identification in the mycobacteriophage Ukulele and closely related cluster E phage. Putative promoters were identified in the Ukulele genome based on the consensus sequences of promoter elements from multiple mycobacterial species. Terminaors were predicted using an algorithm that searches for GC-rich hairpin structures downstream of predicted genes. These computational methods were applied to preexisting knowledge of mycobacteriophage biology and resulted in a clearer

picture of phage transcriptional control. In addition, it has developed an experimental system for confirming promoter efficiency in mycobacterial cells using a fluorescent reporter gene.

2. Literature Review

2.1 Mycobacteriophage Research: Past and Present.

Mycobacteriophage (phage) are viruses that infect bacteria of genus *Mycobacteria* and have been studied extensively since their initial isolation in 1947 [1]. Mycobacterial hosts are of interest due to the pathogenic nature of certain mycobacteria, a genus that includes the causative agents of tuberculosis and leprosy. Phage have been studied to explore the lifestyles of their mycobacterial hosts, but have more recently become a powerful model for learning about viral evolution and diversity [1].

Following the development of modern genomics, the characterization of phage has rapidly expanded and demonstrated remarkable diversity [1]. Different clusters of phage demonstrate both varied infection cycles and sets of functional gene products [8]. Researchers use phage to learn about viral diversity and evolution through genomic analysis combined with molecular genetic approaches in the laboratory [1].

There has been a substantial increase in phage research in recent years due to discoveries related to phage evolution and functional genomics [1]. This research includes elucidating phage lifestyles through the study of repressor systems [9] [10], exploring phage-host interactions [11], and exploring genomic differences between related groups of phage [12]. There is great potential for future discoveries due to the breadth of phage diversity.

2.2 Clustering and Genome Mosaicism

Bacteriophage are considered the most abundant biological entity on earth, with a population exceeding 10^{30} [13]. Their evolution has spanned 3 to 4 billion years, and the diversity of phage is robust [14]. The main source of this variation is horizontal transfer of genetic material by recombination events between phage through homologous recombination at short conserved boundary sequences and sequence-independent recombination between phage genomes and foreign DNA [15]. These recombination events result in genetic mosaicism, where similar sequences between phage are interspersed with other completely unrelated sequences [16].

The result of mosaicism is that each phage genome is typically composed of a unique set of interchangeable modules, which are groups of genes with related function [15,16]. Modules that were exchanged recently in evolutionary history show strong nucleotide similarity. Distant exchanges typically have low nucleotide similarity but some level of significant amino acid similarity [1]. Mycobacteriophage strongly follow this pattern, and the similarities between genomes range from being nearly identical to having barely discernable relationships [17]. This genomic diversity makes it useful to assign mycobacteriophage to groups showing significant similarity, which are known as clusters.

Phage genomes are divided into clusters based on nucleotide sequence similarity [17]. Phage are assigned to the same cluster if they show significant nucleotide similarity for more than 50% of the genome [17]. The methods for determining this similarity include comparison using pairwise dotplots in Gipard, pairwise average nucleotide identities (ANI), analysis of gene phams, and whole genome map comparisons [17]. Through these analyses and subsequent formation of

clusters, we can begin to discern both close and distant evolutionary relationships between mycobacteriophage. There are currently 1141 mycobacteriophage classified into 26 clusters, and this list is rapidly expanding due to intensive sequencing efforts [18].

Despite the advantages of clustering, it is clear that the process has its limitations. The degree of diversity within clusters widely varies and delineation can at times be somewhat arbitrary [11]. However, the formation of clusters based on nucleotide sequence similarity allows researchers to represent heterogeneity of phage and search for the effects of this diversity on their phenotypes [8]. Levels of characterization varies greatly between clusters of phage, and many clusters have yet to be studied intensively [8].

2.3 Establishment and Maintenance of Lysogeny

When phage infect their bacterial host, they typically monitor the health of the cell to determine the infection pathway that best suits these conditions. If the bacteria are growing rapidly the phage enter lytic cycle. This pathway involves using host metabolic machinery to rapidly produce progeny and eventually lyse the host cell [19]. Temperate phage are able to utilize an alternate infection pathway known as lysogeny. Lysogeny involves integration into the host genome and maintenance of this integration for subsequent replication cycles in the bacteria.

Lysogenic pathways vary greatly among mycobacteriophage. Most of these systems involve production of integrase proteins that allow integration into the host genome and repressor proteins that prevent lytic gene expression and allow expression of genes necessary to enter into and maintain lysogeny [8]. In Ukulele, it is hypothesized that gp52 is the immunity repressor due to its location relative to

integrase, its presence in a region of divergent transcription, and the fact that it is the only gene in this region to contain a predicted DNA-binding domain [20]. Lysogeny remains poorly characterized in the majority of bacteriophage, but several well-studied systems are useful models for learning about lysogeny in other bacteriophage.

Lysogeny in the *E. coli* phage Lambda is the most well understood system. Lysogeny in lambda involves a complex decision that begins with transcription of CII and Cro from promoter Pr and CIII from promoter Pl [21]. CII is the key protein involved in the lytic-lysogeny decision and is used to monitor the health of the cell [22]. The predominance of Cro and the lytic pathway occurs when CII is degraded by host proteases within the cell, normally during periods of stress in the cell [23]. If cellular conditions are conducive to growth, CII is protected by CIII and activates both CI from the promoter P_{RE} and integrase from the promoter P_I [24] . If CII prevails, CI repressor is produced which binds to operators Or (instead of Cro) and leads to repression of early lytic genes [25] [26]. The phage then uses its integrase gene for site-specific recombination and inserts the phage genome into the host chromosome. Active production of CI continues and the phage will maintain lysogeny through repression of lytic genes [27].

Another system for establishing lysogeny has been identified in mycobacteriophage: integration-dependent immunity in Clusters P, N, I, and G [9]. The three key proteins to this genetic switch are integrase (Int), repressor protein (Rep), and Cro protein [1,9]. In this instance, the genetic switch deciding between lytic growth and lysogeny depends on the Int expression levels and its ability to carry out site-specific recombination into the host genome [2,9]. If sufficient integrase is produced, the phage integrates and causes production of active Rep protein [1,3,9]. Rep down regulates Cro by acting on Promoter Pr and allows the phage to maintain

lysogeny effectively [2,4,5,9]. When protease levels in the host cell are high, Int is rapidly degraded and unable to integrate [4,9]. The inability to integrate means that the active form of the repressor (Rep) isn't produced and Cro production ensues [6,9]. Cro production leads to the lytic pathway by antagonizing Rep action [7,9]. This way of establishing lysogeny is similar to lambda, but integration is the key decision-making step and no additional proteins such as CII and CIII are required [5,28].

Lysogeny of Cluster A2 phage L5 is another well-characterized example of lysogeny in mycobacteriophage [1,8]. It is simpler than integration-dependent immunity and relies mainly on the active production of the gp71 repressor within the host cell [1,29]. Three promoters, P1, P2, and P3, control transcription of this gene [8,29]. Once produced, gp71 acts on promoter P_{left}, which prevents transcription of lytic genes in L5. It also binds to 28 other protein-binding sites within the genome to downregulate transcription of other genes in the L5 genome [1,10]. L5 then integrates into the mycobacterial genome via integrase-mediated recombination at specific regions in the bacterial and phage genomes, and maintains lysogeny through active production of gp71 [1,10,30]. This establishment of lysogeny differs from integration-dependent immunity because it only requires transcription of gp71 to establish lysogeny [9].

These diverse systems for establishing and maintaining lysogeny demonstrate the variance of genetic switches that are important in the phage decision between active replication and lysogeny [10,28].

2.4 Phage Transcription and Gene Expression

Genome length and gene organization in mycobacteriophage vary greatly [1,11]. In some cases single genes are expressed from a single mRNA but more often

genes with similar functions are grouped in operons. An operon is a group of genes, often with related functions, that are expressed from a single mRNA. This mRNA is produced by a host-encoded RNA polymerase through recognition of promoter sequences upstream of operons. In mycobacteriophage Giles, the structure of these operons is closely related to the function of the genes within them as either structural genes, genes involved in DNA replication, or other genes required for lytic growth [12,31]. The relationship between operon structure and the function of genes within operons appears essential in many cases.

The transcription of these genes is tightly controlled based on their role in the phage life cycle. In lambda, genes are regulated by protein-protein interactions and the production of phage proteins such as repressor C1 [13,19]. Repressor proteins are a well-characterized example of transcriptional control in mycobacteriophage [10,14]. In lytic growth, phage appear to separate transcription into two distinct phases temporally: early and late genes [3,15,32]. These phases demonstrate the importance of transcriptional control during phage infection.

Studies completed using L5 as a model mycobacteriophage suggest that most phage transcription is completed using host RNA polymerase, due to similarities between L5 and canonical mycobacterial promoters[16,32]. This is contrasted by studies of singleton mycobacteriophage Giles, which suggest the presence of promoters not recognized by the most common mycobacterial host sigma factor: sigma-70 [15,16,31]. Study of promoters illustrates a diversity of mechanisms for initiating transcription of the phage genome.

2.5 Control of Transcription Through Promoters

Expression of genes begins by initiating transcription or mRNA synthesis at DNA sequences, called promoters. Promoter sequences are located 10 and 35 bases before the start of transcription (TSP) [1,4]. The sigma protein bound to RNA polymerase recognizes these sequences and allows transcription initiation at the TSP. The important components of the promoter are two hexmeric sequences, the −35 element and −10 element, and the space between these elements (interhexameric space) [4,17]. The rate-limiting step of RNA synthesis in prokaryotes is normally related to the strength of the promoter and the ability of RNA polymerase to recognize and bind to it [17,33]. Study of promoters is therefore important for analyzing patterns of gene expression.

In *E. coli,* Sigma-70 specifically recognizes promoters of housekeeping genes of the cell. *E. coli* sigma-70 consensus promoter sequence is a −35 sequence of TTGACA and a −10 sequence of TATAAT. The space between these elements is between 16-19 base pairs [17,34]. The mycobacterial sigma-A promoter has an analogous architecture and function to sigma-70 in *E. coli*, as both the −35 box and interhexameric space are conserved [17,34]. However, mycobacteria appear to have a slightly different −10 consensus sequence (TAYgAT) and appear more able to withstand mutations of the −35 box without loss of function [4,18]. In addition to these elements, an extension of the −10 box (known as a TGN motif) may also play a role in efficiency of mycobacterial promoters [11,35].

An RNA polymerase has yet to be identified in a mycobacteriophage genome, which indicates they rely on host transcriptional machinery [8,36]. Many mycobacteriophage contain SigA like promoters, indicating they rely on

mycobacterial RNA polymerase and sigma factors [8,29,37,38]. These SigA like promoters in mycobacteriophage have been characterized extensively to determine their role in lysogeny, lysis, and context-dependent transcription.

One example of these promoters is in L5, which contains four well characterized SigA type promoters that drive expression of the repressor gene, gp71 [19,29]. P1, P2, and P3 are upstream of gp71 and are important in transcriptional initiation of gp71. P_{left} is an early lytic promoter that is depressed by the presence of repressor. The structure of P1, P2, and P_{left} are consistent with that of other SigA promoters [8,29]. Interestingly, P3 shows a strikingly different structure that appears to be recognized by a different mycobacterial sigma factor and not by SigA [20,29].

An analogous promoter sequence to Pl_{eft} has been identified in mycobacteriophage Bxb1 [21,29,38]. SigA type promoters have also been identified in lysis genes of mycobacteriophage Ms6, perhaps playing a similar role to P_{left} in L5 [22,37]. This conservation of SigA promoter structure supports the use of host transcriptional machinery.

Another SigA associated promoter, mycobacteriophage BPs promoter Pr, was studied to learn about the importance of context to promoter efficiency [23,36]. Through mutational analysis, the researchers discovered that changes to each base in the –10 sequence was deleterious to promoter function and that consensus sequence TATAMT shows maximal activity [24,36]. In addition, it was determined that 17 bp is the optimal interhexameric space for Pr. Interestingly, mutation of the –35 sequence did not significantly change efficiency of the promoter [25,36]. Thus, mutations of the promoter elements are dependent on the context in which they are carried out, demonstrating complexity in control of mycobacteriophage transcription

[26,36]. These discoveries point to general characteristics of optimal mycobacteriophage promoters.

No late promoters, or promoters that don't depend on SigA, have been identified and characterized in mycobacteriophage. Current methods for promoter search also depend on *E. coli* Sig70 structure, which may be insufficient to describe all possible mycobacteriophage promoters. This gap in knowledge and inconsistencies in promoter search methods require further research.

2.6 Termination of Transcription via Intrinsic Termination

Transcription of operons by RNA polymerase is terminated in the presence of terminators downstream of stop codons: GC-rich hairpins that form secondary structure and U-rich sections of DNA that follow these hairpins[6,27]. Terminators with these features are known as L-type terminators [9,39]. In addition to canonical terminators containing hairpins and poly-U tracts, there is evidence of I-type terminators that function without a poly-U tract in some actinophage [7]. In *E. coli*, there is also evidence of U-type tandem termination which are multiple GC-rich inverted repeat structures that occur in tandem[5]. An additional study refutes the findings of these studies in mycobacteria, claiming that a poly-U tract is required for termination [6]. These studies represent the multiple ideas about how termination occurs in prokaryotic systems and the necessity for further study in mycobacterial systems.

None of these studies deal specifically with mycobacteriophage terminators, although study of mycobacterial transcription is frequently utilized as a model for mycobacteriophage [3]. Therefore, there are multiple possible types of terminators within mycobacteriophage. A web program, named WebGEster, uses an algorithm

that searches different types of terminators within genomic sequence [39] WebGEster identifies four types of terminators: L-shaped terminators containing hairpin structures followed by poly-U tails, I-shaped terminators containing only potential hairpin structures, U-shaped terminators containing tandem termination structures, and X-shaped bidirectional terminators containing that may terminate transcription on both strands [39]. This algorithm is based on the identification of potential hairpin structures with highly negative ΔG, limited mismatches within the structure, and the contextual significance of these sequences following open reading frames (ORFs) [39]..

Another program, named ArNOLD, uses a similar algorithm but only identifies terminators containing poly-U tails [40]. ARNold also identifies hairpin sequences intragenically as well as intergenically [40]. This program was previously used in our lab to identify 4 terminators containing poly-U tails within the Ukulele genome [20]. Terminators are not sufficient to explain transcriptional termination in the entire Ukulele genome, and that other mechanisms must function to catalyze termination of transcription.

2.7 Cluster E: A Relative Unknown

Cluster E is a group of poorly characterized mycobacteriophage [20]. There are currently 68 Cluster E phage, with an average genome length of 75,512 base pairs and a GC content of 63.0% (PhagesDB.org). There are no subclusters in Cluster E and the organization of genes in Cluster E is generally conserved [8].

All Cluster E phage are likely to be temperate mycobacteriophage, as they form plaques with slight turbidity, but no stable lysogens of Cluster E phage have been isolated [8]. Despite the lack of published evidence for stable lysogens, our data

(unpublished) suggest that Ukulele forms lysogens at an efficiency of approximately 7.5% [20].

There are conserved genomic features that have been discovered in Cluster E phage that are not present in most other clusters. Cjw1 is the Cluster E representative phage, and has a number of genes that are rare among mycobacteriophage and therefore of interest [8]. For example, Cjw1 gene 39 encodes a regulatory protein important in expression of many host genes and gene 102 encodes a single-stranded binding protein only found elsewhere in Cluster L phage and singleton Wildcat [8]. The function of these proteins within Cjw1 and other Cluster E phage is yet to be elucidated.

Previous studies in our lab have uncovered several characteristics of Cluster E phage Ukulele. Four factor-independent terminators were identified by using ARNold and by inspecting regions of convergent transcription in the genome [20]. There are also two conserved repeats within the genome. Repeat CR1 (5'–CTTCACTGAACTg/aAA) is of particular interest because it is oriented in the direction of transcription in four of its five locations [20]. CR1-2 and CR1-3 are almost a set of inverted repeats and are oriented in the direction of gp52 and gp53 respectively [20]. gp52 and gp53 are divergently transcribed and located close to the integrase, which signifies a potential role in establishment of lysogeny [20]. As mentioned previously, it is hypothesized that Ukulel gp52 acts as the repressor protein [20].

3. Materials and Methods

3.1 Bacterial and Viral Strains and Plasmids

M. smegmatis mc²155 was grown at 37°C with shaking in liquid 7H9 media supplemented with 50 μg ml^{-1} carbenicillin (CB) and 10 μg ml^{-1} cyclohexamide (CX) or on 7H10 agar plates. When required, hygromycin (Sigma, St Louis, MO) was added to media at a final concentration of 50 mg L^{-1}. *E. coli* XL1 Blue was grown at 37°C in L-broth with shaking or on L-agar plates [41]. When required, hygromycin was added to a concentration of 200 mg L^{-1}.

Mycobacteriophage Ukulele was isolated in 2011 from a soil sample in Old Orchard Beach, Maine by the Honors 150 Phage Genomics Course. Ukulele was isolated by an enrichment culture prepared with the soil sample and the bacterial host *M. smegmatis* mc²155 in 7H9 complete media. The culture was incubated overnight with shaking at 37°C. After centrifugation at 2,000 x *g* for 10 min, the supernatant of the culture was filtered on a 0.2 μM filter, serially diluted in phage buffer (10 mM Tris pH 7.5; 10 mM $MgSO_4$; 68 mM NaCl; 1 mM $CaCl_2$) and plated in 7H9 top agar onto a lawn of *M. smegmatis* mc²155. After multiple rounds of purification, high-titer stocks of Ukulele were prepared.

pUV15tetORm was a gift from Sabine Ehrt (Addgene plasmid # 17975). pUV15tetORm has a pBR322 origin that has a high copy number in *E. coli* and low copy number in mycobacteria. pUV15tetORm encodes a hygromycin resistance gene, and a green fluorescent protein (GFP) gene under the control of a strong mycobacterial promoter, $P_{imyc\text{-}tetO}$. The mycobacterial promoter is flanked by restriction sites *Pac*I and *Spe*I [42]. To construct recombinant plasmids, Ukulele genomic sequences and plasmid pUV15tetORm were digested with *Pac*I and *Spe*I according to manufacturers protocol. The restriction digest products were subsequently ligated, and all plasmids were sequenced to ensure proper insertion of the insert.

3.2 Promoter Identification

A mycobacterial-specific matrix was developed by Keith Hutchison that determined likelihood of each nucleotide (A, T, C, G) at each promoter position based on previous identification of promoter sequences in multiple species of *Mycobacterium* [4]. With this matrix, the PhiSite Promoter Search tool was used to identify candidate promoter sequences that contain consensus −10 sequences, and subsequently both −10 and −35 sequences [43] Special consideration was given to −10 boxes containing a TGN motif directly preceding them due to previous study of this feature [4]. The promoter analysis was performed on intergenic sequences of the Ukulele genome with a particular focus on regions of divergent transcription and sequences downstream of terminators. Following identification of candidate promoter sequences, the promoter with highest consensus −10 strength compared to the mycobacteria-specific promoter matrix was chosen if a viable −35 sequence was present.

3.3 Primer design

Primers were designed to amplify predicted promoter sequences of interest in the Ukulele genome and in the L5 genome. Ukulele gp53 and L5 gp71 promoter primers were designed on Primer3Plus with the parameters Tm Min=40°C, Tm Max=50°C, Min Length=16, Max Length=24. In order to clone sequences into the reporter splasmid pUV15tetORm, restriction enzyme sites for *Pac*I and *Spe*I (NEB) were included in the five primer ends of the forward and reverse primer, respectively

(Table 1) [44]. In order to construct control plasmids with promoter sequences inserted in the reverse orientation, primers were also designed with *Pac*I and *Spe*I sites on the five prime ends of the reverse and forward primer, respectively. All primers were checked for potential primer-dimer formation using ThermoFisher Scientific Multiple Primer Analyzer (ThermoFisher Scientific, Waltham, MA).

3.4 Polymerase Chain Reaction (PCR)

PCR was performed in 25-µl reactions containing 1 ng of Ukulele genomic DNA, 0.5μM of each primer, and Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA) according to the manufacturer's recommendations. Reactions were incubated at 95 °C for 2 min, then cycled 35 times through 98°C for 10 s, 65°C for 30 s, and 72°C for 30 s.. L5 gp71 promoter was amplified using PROMEGA Taq polymerase (Promega, Madison, WI) and reactions were incubated at 98°C for 30 s then subjected to 35 cycles of 98°C for 10 s, 61°C for 20 s, and 72°C for 30 s. PCR products were analyzed on 2% SeaKem LE agarose (Lonza, Rockland, ME) gels or 2% SeaPlaque GTG gel agarose (Lonza) and purified using the Qiagen PCR Purification kit (QIAGEN, Valencia, CA) according to the manufacturer's recommendation.

3.5 Restriction Endonuclease Digestion

10 units of *Spe*I and *Pac*I restriction enzymes per 5 μg of plasmid DNA were used for restriction digests according to NEB protocol (NEB). Restriction digests of insert and plasmid DNA were performed in 100μL volume reactions according to manufacturer's recommendations (New England Biolabs).

3.6 DNA Sequencing

To confirm the orientation and sequence of promoter inserts in pUV15_GFP, all recombinant plasmid DNA were sequenced at University of Maine DNA Sequencing Facility (Orono, ME).

3.7 Agarose Gel Electrophoresis

DNA fragments were separated on Seakem LE agarose (Lonza) using 80-150V of electrical current in TAE buffer (40 mM Tris, 20 mM acetic acid, 2 mM EDTA). These gels were stained using ethidium bromide (0.5 μg ml^{-1}) and visualized using UV transilumination.

Fragments that needed to be recovered for cloning were separated on 1.5% SeaPlaque GTG low melt agarose gels (Lonza) prepared in TAE (40 mM Tris, 20 mM acetic acid, 2 mM EDTA) and containing 0.5 μg ml^{-1} of ethidium bromide. Fragments were purified using a Qiagen Gel Extraction Kit according to manufacturer's recommendations (QIAGEN, Valencia, CA).

3.8 DNA Ligations and Transformations

Ligations were performed in 20-μl reactions containing 500 ng total of digested PCR product and vector and T4 DNA ligase according to manufacturer's instructions (New England Biolabs). Competent *E. coli* XLI Blue cells were transformed with 10 μl of the ligation reaction (250 ng of DNA) [45]. Hygromycin-resistant transformants were selected for on L-agar plates containing 200 μg ml^{-1} of

hygromycin (Sigma-Aldrich, St. Louis, MO). Colonies were inoculated in L-broth containin 200 μg ml^{-1} hygromycin. Plasmid DNA was isolated from 3 ml of culture using Qiagen MiniPrep Spin Kit (QIAGEN) according to the manufacturer's recommendations. Plasmid was then electroporated into competent *M. smegmatis* mc^2155 cells. Transformants on 7H10 plates containing 50 μg ml^{-1} hygromycin [46] were inoculated in 7H9 broth containing 50 μg ml^{-1} hygromycin.

3.9 Terminator Identification

Intrinsic terminators were identified globally in Ukulele genome using WebGester [47]. WebGeSTer uses an algorithm to search regions downstream of stop codons for palindromic DNA sequences. Following WebGeSTer identification of terminators, data was processed using Excel and TextWrangler to confirm validity of terminator predictions. Strong terminators in Cluster E were those likely capable of independent transcriptional termination, based on low ΔG, location in intergenic regions following coding region, and a maximum of 3 mismatched base pairs in the hairpin structure. WebGeSTer-identified sequences were accepted when there was sufficient contextual evidence of the genome structure to suggest termination. All strong terminators were mapped on Ukulele genome map.

Tandem terminators with low combined ΔG were found by changing the WebGeSTer parameters to allow for higher ΔG [47]. This setting allowed WebGeSTer to identify tandem terminators that had a combined ΔG lower than -16.03 kcal/mol and a distance of less than 50 bp between structures [39]. These terminators were subsequently mapped on Ukulele genome with the strong terminators.

3.10 Terminator BLAST

Terminators identified by WebGeSTer were aligned on PhagesDB.org BLASTn to search for conservation in cluster E phage. Location of sequences showing high identity were mapped on other Cluster E genomes to determine if they were likely terminator sequences.

3.11 Genome Alignment of Ukulele and Identification of Cluster E Highly-Variable Regions

The Ukulele genome was aligned with other mycobacteriophage genomes using local BLAST on PhagesDB.org [18]. Highly variable regions that appeared conserved in many other cluster E genomes were identified and labeled HVR1-16. Representative Cluster E phage that as a group contained these HVR were chosen to complete terminator analysis: 244, Willez, DrDrey. Some HVR were represented by more than one of the Cluster E representative phage, but all gaps were contained by at least one of these phage. Phamerator maps were generated between Ukulele and each phage and the HVR were identified in these alignments [48]). These maps and identification of HVR were used to determine their potential relevance to terminator search.

3.12 GFP Fluorescence Assay

To determine if promoter sequences in the pUV15 recombinant plasmids drive expression of GFP in mycobacterial cells, cells were grown for four d at 37°C with shaking. Cultures were sub-cultured to an optical density at 600 nm (OD600) of 0.02 and grown overnight at 37°C. Once cultures reached an OD600 of approximately 0.5, cultures were dispensed in 200-μL volumes in replicates of 8 into a 96-well plate.

Fluorescence of each sample was measured at 528 nm and OD measured at 600 nm using a Biotek multiplate Reader. Fluorescence values were normalized to the OD600 values to determine approximate fluorescence per number of cells. Values are reported as average relative fluorescence units (RFUs) plus or minus the standard error of the mean with n=8.

4. Results

4.1 Thirty-One Potential-Promoter Regions Were Identified Upstream of Ukulele Genes

Loci of potential Ukulele promoters were determined by identification of long intergenic regions, areas of divergent transcription, and regions directly following potential terminators (Figure 1). PhiSite promoter search identified 22 potential promoters on the forward strand and 9 potential promoters were identified on the reverse strand (Table 2). The maximum PhiSite score is 1, and the strength of Ukulele promoters ranged from PhiSite −10 scores of 0.56 – 0.98 (Table 2). Seven promoters contain TGN-motif immediately upstream of the −10 box. Only one of these TGN-type promoter sites had a recognizable −35 box. For some genes, there were multiple candidates for promoters (Table 2).

4.2 Intergenic Region Between gp52 and gp53 Contains Evidence of Strong Promoter Activity and Potentially Important Conserved Repeated Sequences

The intergenic region between gp52 and gp53 is a region of divergently transcribed genes, and was therefore a target for promoter identification using PhiSite Promoter Search. There are two rightward promoters in the direction of gp53, both

containing a –35 element and one containing a TGN-motif. These promoters both occur upstream of a previously identified conserved repeated sequence within the Ukulele genome (CR1-3) (Figure 2A) [20]. There are two leftward promoters in this region in the direction of gp52, and one of these contains a TGN- motif (Figure 2A). One of the gp52 potential promoters overlaps with a related repeat motif to that upstream of gp53 (CR1-2).

4.3 Complete gp52-gp53 Intergenic Region Drives Expression of Reporter Gene in M. smegmatis Cells.

In order to determine the activity of potential promoters within mycobacterial cells, sequences upstream of gp52 (P1, P2, P3) and gp53 (P2rev, P3rev, gp53) were cloned into the pUV15tetORm reporter plasmid containing GFP and tested for the ability to drive expression of GFP in *M. smegmatis* cells (Figure 2B). Recombinant plasmid gp53 produced a significant amount of fluorescence in *M. smegmatis* compared to *M. smegmatis* controls without reporter plasmid (Figure 3A) (Figure 2B). The gp53 promoter showed low efficiency compared to known mycobacteriophage promoter from L5 gp71, (Figure 3A) (Figure 2B). The gp53 plasmid also demonstrated higher efficiency than all other recombinant plasmids. P2rev contained putative gp53 promoter 2, and P3rev contained putative gp53 promoter 1, but these plasmids failed to drive expression of GFP (Figure 2B) (Figure 3A).

Recombinant plasmids P2 and P3 each contained a putative promoter region upstream of gp52 (Figure 2B). Recombinant plasmid P1 contains a sequence immediately upstream of gp52 initially thought to have a promoter sequence in our first analysis of the sequence but that we decided in subsequent analysis wasn't likely

to have a promoter region. All three plasmids generated fluorescence levels significantly lower than that of *M. smegmatis* cells without plasmid (Figure 3B).

4.4 Nineteen Terminators of Different Structure Types Were Identified in Ukulele Genome

WebGeSTer identified L, I, U, and X-type terminator structures in the Ukulele genome. These terminators were selected based on their containing low ΔG (less than -16.03 kcal mol $^{-1}$), location in intergenic regions following coding region, and a maximum of 3 mismatched base pairs in the hairpin structure. Eight strong terminators were identified on the forward strand of the Ukulele genome and 11 strong terminators were identified on the reverse strand (Table 3). Of the 19 terminators in the Ukulele genome, 15 were I-shaped terminators and 4 were L-shaped terminators containing poly-U tracts. U-shaped tandem terminators were identified downstream of gp124 and gp51. One X-type convergent terminator was identified between gp49 and gp51 (Figure 4).

The forward strand of the Ukulele genome contains terminators following gp5, gp12, gp18, gp28, gp29, gp31, gp49, gp50 (Table 3). The reverse strand contains terminators following gp6, gp38, gp48, gp51, gp122, and gp124 (Table 3). The maximum ΔG of these terminators is −16.02 kcal mol^{-1}, they are all located in intergenic regions, and they all contain 3 or fewer mismatches in the stem. A number of these terminators are highly conserved within Cluster E genomes, while others do not show identity to sequences in other phage.

4.5 Ukulele Terminators Frequently Occur in Tandem, Exhibiting Characteristics of U-Type Terminators

Intergenic regions in the Ukulele genome were also analyzed for multiple structures that don't meet the ΔG criteria of a strong terminator but may act in tandem to terminate transcription. WebGeSTer predicted 6 putative tandem terminator sequences on the forward strand. gp20 contains three potential hairpin structures downstream of the stop codon that have ΔG values that add to −16.38 kcal mol^{-1} (Figure 5A). gp29 contains two potential terminator structures with a combined ΔG of −18.65kcal/mol and a 22-bp gap between them (Figure 5B). gp37 contains 4 total structures with combined ΔG of −35.57 kcal mol^{-1} and gaps of 4-, 4-, and 19 bp between them. (Figure 5C). gp82 contains two structures separated by 12 bp that have combined ΔG of −25.09 kcal mol^{-1} (Figure 5D). gp100 contains two structures with combined ΔG of −25.1 kcal mol^{-1} (Figure 5E). gp120 occurs at region of convergent transcription and has combined ΔG of −20.27 kcal mol^{-1}. On the complementary strand, one tandem terminator was identified following gp30, with three structures totaling −32.65 kcal mol^{-1}.

Because terminator structures often appear in multiples, we analyzed regions containing strong terminators to identify potential additional terminator structures that may assist strong terminators. Supporting structures were identified for previously accepted strong terminators following gp49 (−8.97 kcal mol^{-1}) on the forward strand and gp6 (−14.34 kcal mol^{-1}, −6.79 kcal mol^{-1}, −9.05 kcal mol^{-1}), gp38 (−12.45 kcal mol^{-1}), gp48 (−11.53 kcal mol^{-1}), and gp121 (−9.97 kcal mol^{-1}) on the complementary strand (Figure 6). There are 7 total potential termination structures downstream of gp51, totaling −106.54 kcal mol^{-1} (Figure 6E).

4.6 Sixteen Highly-Variable Regions Were Identified in Cluster E Mycobacteriophage

Due to the presence of terminators in Ukulele's genome that are not conserved among other Cluster E mycobacteriophage, the Ukulele genome was aligned using BLAST on PhagesDB.org to identify regions of variation within Cluster E phage (Figure 7). All 75 Cluster E phage contained at least 97% nucleotide identity to Ukulele, however there were 16 regions in the genome that appeared highly variable among the Cluster E genomes (Figure 7). These 16 highly variable regions (HVR) were identified among multiple Cluster E genomes (HVR1-16) and were chosen to represent the potential sequence variation among these phage (Figure 7). Cluster E phage 244, Willez, and DrDrey were chosen to analyze for terminators using WebGeSTer because all 16 HVRs are represented among the four genomes. Phamerator alignment of Ukulele with 244, Willez, and Dr. Drey was used to determine if these variations occurred at intergenic regions where terminators occur (Figure 8). HVR1, HVR2, HVR6-11, HVR13-G16 are present within genes and therefore are unlikely to contain terminators (Figure 8). HVR3, HVR4, HVR5, and HVR12 are found in Willez within intergenic regions relevant to terminator search (Figure 8). These results demonstrate the similarity of Cluster E genomes and the use of Willez, Dr. Drey, 244, and Ukulele to represent all significant sequences differences amongst the Cluster E phage.

4.7 Seven Additional Terminators Were Identified Through WebGeSTer Analysis of Other Cluster E Genomes

In order to identify additional Cluster E terminators, 244, Willez, and DrDrey were chosen as phage that represented sequence variation among the cluster. These genomes were analyzed using WebGeSTer identification of intrinsic terminators, and seven additional terminators that do not show identity to Ukulele terminators were identified (Table 5).

Six unique terminators were identified within the Willez genome, named WT1-5 on the forward strand and WRT1 on the reverse strand. Willez contained highly variable regions HVR3, HVR4, HVR5, and HVR12 that occur intergenically near Ukulele-type terminators (Figure 8). HVR3 is in the region of unique Willez-type terminator (WT1), HVR4 is in the region of WT2, and HVR12 occurs near both WT4 and WT5 on the forward strand and a Willez terminator on the reverse strand (WRT1) (Figure 8) (Figure 9). WT3 does not correspond to a Cluster E HVR, but was identified in a region where Willez and Ukulele gene structure differ (Figure 9). In addition to unique terminators, Willez contained UT1 and UT4-8 on forward strand and URT1, URT3-8, URT10a-b on complementary strand (Table 5).

One unique terminator was identified in DrDrey, which did not correspond to any regions of high variation in Cluster E genomes (Table 5) (Figure 8). DrDrey also contains UT1-8 and all URT terminators with the exception of URT2.

244 didn't contain unique terminators, despite containing Cluster E HVR. These HVR all occurred within genes (Figure 8). 244 contained all 19 Ukulele-type strong terminators with the exception of URT2, which is a region of different gene structure (Figure 8).

4.8 Cluster E Class of Terminators Contains 26 Novel Terminators from Ukulele, Willez, DrDrey

Strong terminators in Cluster E were those chosen by WebGeSTer as potential terminators based on low ΔG, location in intergenic regions following coding regions, and a maximum of 3 mismatched base pairs in the hairpin structure. An initial search of the Ukulele genome located eight strong terminators on the forward strand (UT1-UT8), nine strong terminators on the reverse strand (URT1-9), and 1 tandem terminator (U-shaped) on the reverse strand (URT10a, URT10b). Terminators with significant identity to these terminators were also identified by WebGeSTer in representative Cluster E phage 244, Willez, and DrDrey (Table 4). BLAST results indicate the presence of identical sequences to UT terminators in many other Cluster E genomes. Another Cluster E phage chosen at random, Dumbo, showed only Ukulele-type terminators (Table 5).

In addition to Ukulele-type terminators, six novel strong terminators were identified in the Willez genome. Five of these terminators are on the forward strand (WT1-5) and 1 is on the reverse strand (WRT1). 1 novel strong terminator was identified on the forward strand of the DrDrey genome (DT1) (Table 4). The majority of these non-Ukulele type terminators occur in regions of conserved gaps in Willez contained by many other Cluster E phage (Figure 7). This class of 25 strong terminators constitutes all currently identified terminators in Cluster E mycobacteriophage.

4.9 Cluster E Mycobacteriophage Have One of Two Different Terminator Structures Downstream of the Major Capsid Protein

WebGeSTer analysis combined with BLAST of terminators indicates there is consistently a terminator located downstream of the major capsid protein in Cluster E phage. However, this terminator appears to occur in at least three different

morphologies (Figure 10). Ukulele contains a putative terminator (UT2) following the major capsid protein (gp12) with a ΔG of −24.17 kcal/mole (Table 3). This terminator contains 100% nucleotide identity with 33 Cluster E phage and an identical terminator is predicted in representative Cluster E phage 244 and DrDrey (Table 4).

UT2 aligns with a conserved gap HVR3 in the Willez genome and a number of other cluster E genomes (Figure 7, Figure 3). A terminator is predicted in Willez in this region (WT1) with a ΔG of -34.17 kcal/mole (Table 4). This terminator shows close identity with 40 other Cluster E phage according to BLAST alignment. Cluster E phage Manda and Eureka have a terminator in this region however it differs in sequence from that of Ukulele and Willez (Figure 10). This terminator has a similar sequence to Willez, but doesn't include an additional region found in the WT1 terminator.

The 33 Cluster E phage with identity to UT2 and the 42 related to WT1 demonstrate that all 75 of the Cluster E phage currently sequenced appear to have a terminator following the major capsid protein, and that they exist in three distinct types [18].

4.10 Twenty-One Potential Operons in Ukulele Were Identified by Combining Promoter and Terminator Identification

The location of promoters and terminators in the Ukulele genome was used to identify potential operons. On the forward strand of the Ukulele genome there were 16 operons and 5 operons were defined on complementary strand (Figure 9). Potential operons range in length from single gene modules to operons containing 24 gene products (gp58 – gp82) (Figure 11). For some operons, multiple putative promoters or

terminators appear to control initiation and termination of transcription respectively (Table 2)(Table 3). For others, a single promoter and terminator has been identified (Table 2)(Table 3).

5. Discussion

Mycobacteriophage continue to be studied extensively due to their propensity for teaching us about the lifestyle of their mycobacterial hosts and about viral evolution [1]. Transcriptional control has been studied extensively within mycobacteriophage, with studies focusing primarily on control elements such as repressor proteins, context-dependent transcription, and early vs. late transcription [10] [3,36]. This study aimed to elucidate potential transcriptional control DNA sequences in mycobacteriophage Ukulele and closely related Cluster E phage, to determine if recognizable patterns exist. To do this, potential promoter sequences and intrinsic terminator sequences were identified. Following prediction, experimental work attempted to confirm promoter activity of Ukulele promoters in mycobacterial cells. This study contributes to the computational analysis of mycobacteriophage genomes. The methods in this study have already been used to further characterize other Cluster E phage and may be applied to computational analysis of other mycobacteriophage in future studies.

5.1 Ukulele Contains a Variety of Potential Strong Intrinsic Terminators and Weak Tandem Intrinsic Terminators

The presence of I-type, U-type, L-type and X-type terminators in the Ukulele genome demonstrates the importance of using multiple analyses to account for all possible terminator structures. Ukulele contains 19 strong terminators, most of which

are I-type terminators lacking poly-U tails (Table 3). I-type terminators are also the dominant terminator in host *M. smegmatis* [47]. There is also evidence of both convergent X-type terminators and U-type tandem terminators within Ukulele (Figure 6, 7). Some studies claim that only L-type intrinsic terminators with a poly-U tail can direct termination in mycobacteria [6], while others report confirmed functional I-type terminators that lack poly-U tracts in actinophage [7]. Nevertheless, the potential Ukulele terminators structures have statistically significant free energy values compared to the GC content of the Ukulele genome and are highly conserved among other Cluster E mycobacteriophage. This evidence strongly suggests that I-type terminators play a large role in Ukulele transcriptional termination, and should not be ignored in analysis of intrinsic terminators.

Terminator structures in tandem, U-type terminators, occur frequently in the Ukulele genome and may play a role in termination (Figure 1, 7). Tandem termination through canonical U-type terminators has been confirmed downstream of an *E. coli* rRNA gene, which supports the possibility of multiple terminators controlling transcription [5]. Therefore, an attempt to identify tandem structures may further contribute to characterization of Ukulele intrinsic termination.

This method for determining terminators needs to be confirmed experimentally, but this computational analysis provides evidence of intrinsic termination in Ukulele via U-type, L-type, but mostly I-type terminator structures. The 25 terminators in Ukulele are found in nearly all regions of convergent transcription and other likely transcriptional termination sites in the genome (Figure 1). The identification of these terminators creates increased clarity in the definition of mycobacteriophage transcriptional termination, and contributes to the process of characterizing newly identified phage. The diversity of terminator structure

demonstrates the importance of identifying all possible terminator regions within mycobacteriophage.

5.2 A Class of Potential Strong Terminators in Cluster E Mycobacteriophage Appears Conserved Throughout the Cluster

Using identification of highly variable regions (HVR) in Cluster E genomes, we were able to identify the majority of different terminator structures that exist in Cluster E mycobacteriophage. These HVR were identified using local BLAST analysis to determine potential sequence variation between Ukulele and other Cluster E genomes (Figure 7) [18]. The result of this search yielded 16 highly variable regions (HVR) among the Cluster E phage (Figure 7). Cluster E phage 244, Willez, and DrDrey were chosen for additional terminator analysis because these phage as a group represented each of these 16 HVRs (Figure 7). These variable regions represent other genomic regions in Cluster E phage that are not present in Ukulele but may contain intrinsic termination signals.

WebGeSTer analysis of these phage genomes indicated the presence of six novel strong terminators in Willez and 1 in DrDrey. Of the six terminators in Willez, five were located near Cluster E HVR (Figure 9). The correlation between novel terminators and HVR provides justification for the use of BLAST to determine sequence variation relevant to terminator search. This set of 26 strong terminators in Cluster E mycobacteriophage represents a group of terminators that may describe the majority of intrinsic termination signals in Cluster E genomes because all of the major sequence variation between Cluster E phage has been analyzed in this analysis.

This global analysis of Cluster E termination signals is important when characterizing mycobacteriophage due to the mosaic nature of phage. The implication

of this analysis is that computational identification of termination signals can be applied to phage besides Ukulele and that characterization of terminators can be extended to entire clusters of phage.

5.3 A Terminator Following the Major Capsid Protein Shows Two Distinct Classes of Terminators that Are Present in all Cluster E Genomes

The terminator following the major capsid protein is an example of genetic mosacism resulting in homologous function [1,16]. WebGeSTer analysis identified two distinct potential terminator structures following the stop codon of gp12 in Ukulele (UT2) and Willez (WT1) (Figure 10). These potential terminators are located near HVR3 identified during Phamerator alignment of these genomes (Figure 8A). BLAST alignment of each of these sequences on PhagesDB.org showed that 40 other Cluster E phage contain sequences identical to WT1 terminators and the remaining 33 phage contain sequences identical UT2 terminators. The complete conservation of this terminator in Cluster E phage may indicate its significance to transcriptional control within this cluster. The distinct terminator morphology following the major capsid protein also highlights the mosaicism of phage genomes and demonstrates the potential for terminator identification to teach us about viral diversity.

5.4 Ukulele and the Problem with Promoter Search in Mycobacteriophage

The use of a mycobacterial-specific matrix for promoter search represents a novel approach that accounts for the potential complexity of promoter structure within mycobacteriophage. This analysis applied to Ukulele identified 22 candidate promoter sequences on the forward strand and 9 candidate promoter sequences on the reverse

strand (Figure 1). Eight of these promoters contained an extended −10 box (TGN– motif), that increases promoter function in mycobacterial cells [4] (Table 2). The promoter search matrix was based on identified promoter regions in multiple species of mycobacteria [4]. The process for selecting promoters also took into account functional requirements for promoters in mycobacteria, such as the centrality of the −10 box to promoting transcription, a potentially unnecessary −35 box, and the optimal bases for the transcriptional start point (TSP) [4].

This research is a novel and more specific approach to identifying promoters in mycobacteriophage. A common tool used in mycobacteriophage promoter analysis is the promoter search tool in DNA Master. This tool identifies sequences similar to *E. coli* sigma-70 promoters, which shows significant homology to housekeeping promoter SigA in mycobacteria [49]. While this method is effective at identifying SigA-type promoters, it excludes promoters that do not have the SigA structure [4]. One recent study demonstrated a difference in promoter efficiency between a canonical SigA promoter and a mycobacteriophage promoter with optimal strength in mycobacterial cells [36]. This approach indicated that non-SigA promoters in mycobacteriophage are not only likely, but can have increased promoter efficiency. There are also many other sigma factors identified in mycobacteria that likely interact with different DNA sequences, as well as evidence of sigma factors within mycobacteriophage genomes [4,20]. These studies highlight the complexity of promoter search, and the limitations of current approaches for addressing this complexity.

This study increased specificity for discovering promoter elements within mycobacteriophage that may differ from the canonical SigA promoter structure. The

variety of methods and conflicting results of promoter search demonstrate the need for a more streamlined approach to identifying and confirming promoters. Computational analysis in our lab aims to develop the optimal method of determining promoters, but the complexities of transcriptional initiation in mycobacteriophage indicate the need for a process that incorporates more factors into promoter search. This study begins to accomplish this aim by looking for candidate promoters homologous to a variety of mycobacterial promoters and incorporating relevant knowledge of optimal promoter function in mycobacteria.

5.5 Promoter Confirmation Using GFP Reporter System Yields Inconclusive Results

The pUV15tetORm plasmid is a promising GFP reporter system for confirming mycobacteriophage promoter function. The complete intergenic region between gp52-gp53 on the forward strand was able to direct expression of a GFP reporter gene, as was a previously identified strong L5 promoter. The L5 sequence contains three strong, previously identified promoters and showed a significantly stronger fluorescent signal than gp53 (Figure 2B, 3B). The relative fluorescence from the L5 gp71 promoter and the gp53 promoter indicate the validity of this reporter system as a test for promoter efficiency. Despite these positive results, none of the other recombinant plasmids that contained candidate promoters appeared to express GFP in *M. smegmatis* (Figure 3). This may be due to these sequences lacking viable promoter elements, but it is possible that expression from these plasmids could be detected with optimization of the assay.

Use of a negative control strain of *M. smegmatis* carrying a promoter-less pUV15tetORm plasmid could improve the sensitivity of the GFP expression assay.

The negative control used in this experiment, an *M. smegmatis* strain without any plasmid, had higher fluorescence than P1, P2, P3, P2rev, and P3rev recombinant plasmids. This may be due to the lack of hygromycin in the control culture, or have to do with a decrease in fluorescence caused by the presence of pUV15tetORm plasmid. The *M. smegmatis* strain with a promoter-less plasmid would provide an optimal control for background fluorescence. Mutagenesis of specific bases within identified promoter elements would also be an important improvement for future confirmation of promoter regions. Quantification of promoter efficiency by plasmid constructs containing single-point mutations would allow more accurate determination of the key bases involved in promoter activity.

To conclude, there are many improvements necessary to increase the efficacy of our promoter confirmation assay, but our current results provide a proof-of-concept that demonstrates the potential of this assay.

5.6 Conclusions

This research has supported the hypothesis that we can further genomic characterization of Cluster E mycobacteriophage through identification of potential terminator and promoter sequences. Our novel approach for identifying promoters using a mycobacteria-specific matrix has applied relevant knowledge of mycobacteria biology to the search for mycobacteriophage promoters. In addition, the application of a preexisting web resource WebGeSTER to find intrinsic terminators has characterized a class of strong terminators in Cluster E phage based on analysis of sequence variation. Experimental confirmation of promoter sequences has also increased knowledge of Ukulele transcription of potentially important genes gp52 and

gp53. The discoveries of this research demonstrate the necessity of further steps to characterize Cluster E mycobacteriophage transcription.

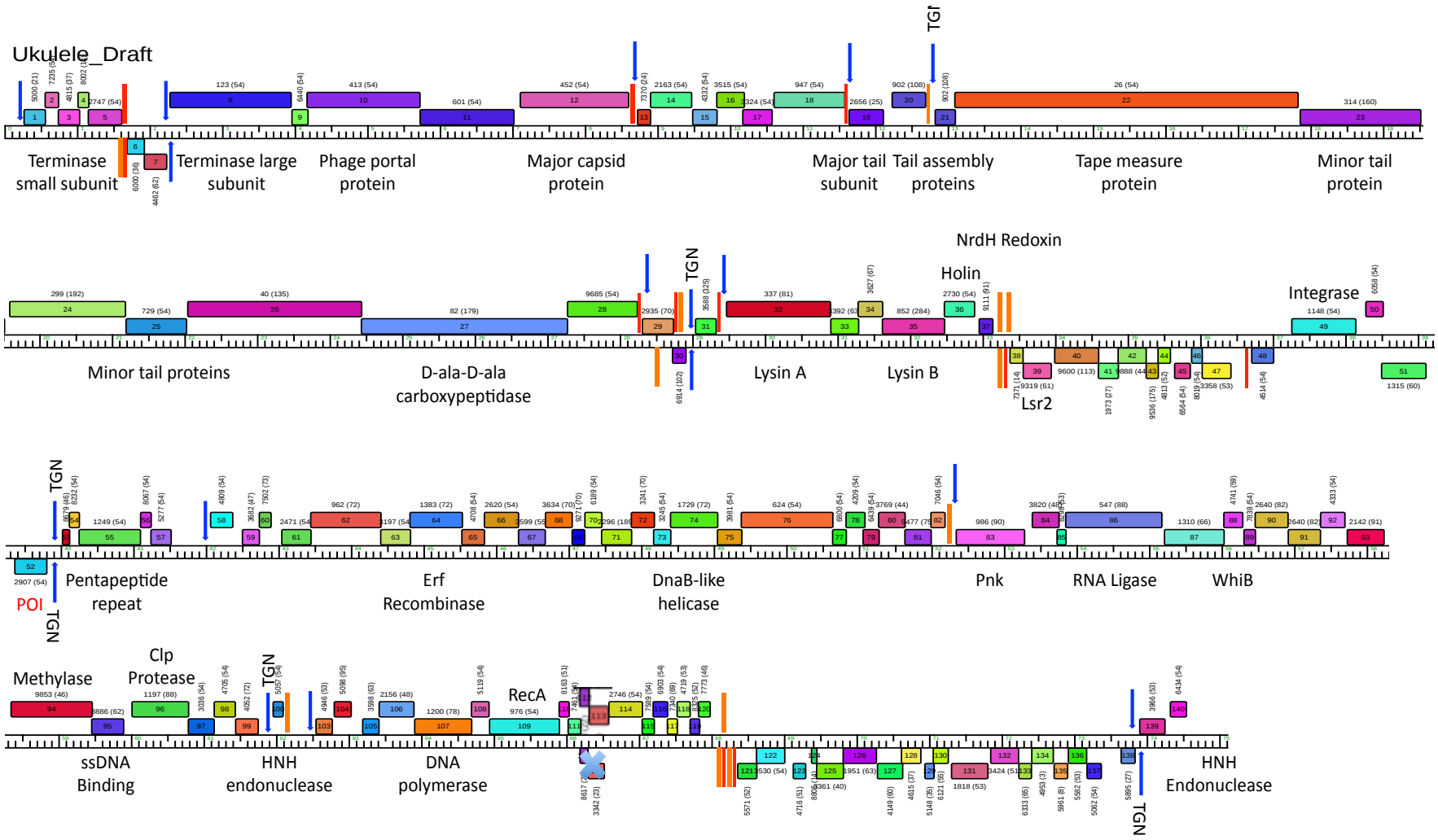
5.7 Future Research

It is clear that further optimization is required for both identification and confirmation of transcriptional control elements in Cluster E mycobacteriophage. A number of novel approaches used in this research demonstrate potential for characterization of mycobacteriophage transcription, but further experiments are required.

One further step in promoter identification is to streamline confirmation of mycobacteriophage promoters and determine the best way to identify them. Alternate methods use DNA Master Promoter Search tool, while ours utilizes a mycobacterial-specific matrix and PhiSite promoter search to identify promoters. The reconciliation or combination of these approaches is a necessary future step, and relies on further experimental confirmation of promoter sequences. Once promoters are confirmed experimentally, these sequences can both validate our identification approach and improve it by increasing the sequences that contribute to our matrix.

A class of strong Cluster E terminators has been identified through WebGEster analysis of potential hairpin structures in intergenic regions. In addition, weak terminators with tandem structures and contextual significance have been identified in Ukulele. Both strong and weak terminators must be confirmed experimentally using *in vitro* termination assays that place potential hairpin sequences at the conclusion of genes. These experiments would resolve the conflict about viable terminator sequences and demonstrate the possible intrinsic termination signals in mycobacteriophage.

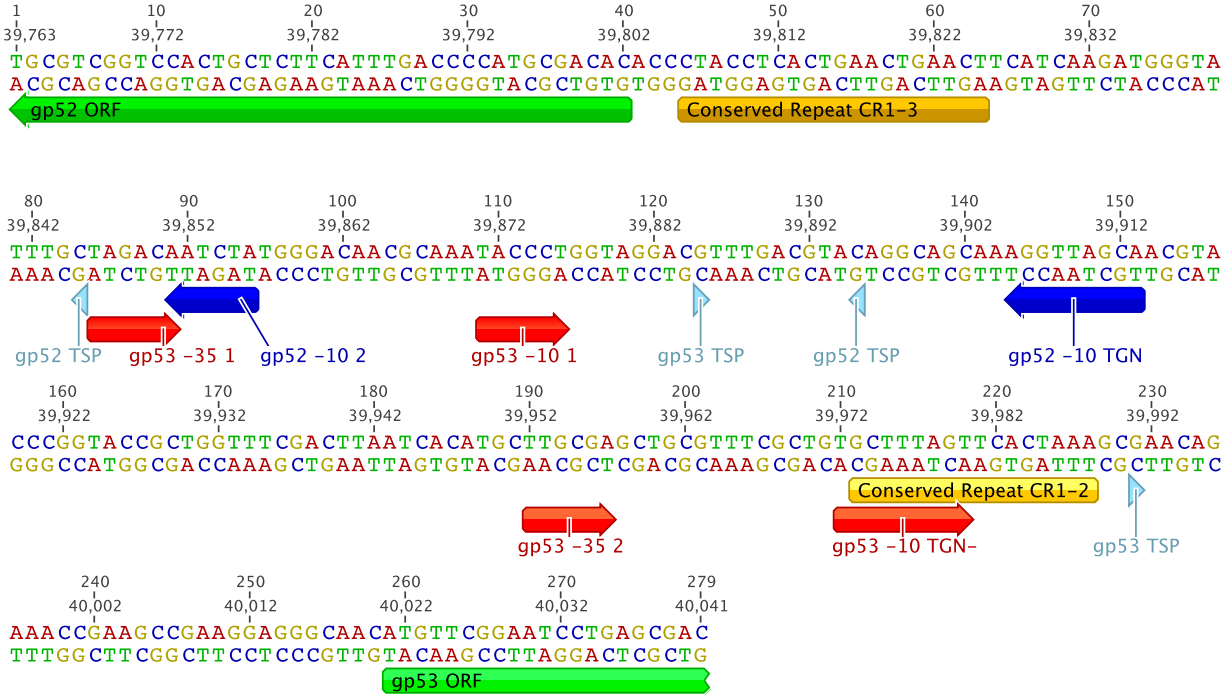
FIGURES


Ukulele_Draft
TGI
Terminase small subunit
Terminase large subunit
Phage portal protein
Major capsid protein
Major tail subunit
Tail assembly proteins
Tape measure protein
Minor tail protein
NrdH Redoxin
Holin
Integrase
TGN
Minor tail proteins
D-ala-D-ala carboxypeptidase
Lysin A
Lysin B
Lsr2
TGN
POI
TGN
Pentapeptide repeat
Erf Recombinase
DnaB-like helicase
Pnk
RNA Ligase
WhiB
Methylase
Clp Protease
TGN
RecA
ssDNA Binding
HNH endonuclease
DNA polymerase
HNH Endonuclease
TGN

= Ukulele-Type Strong Terminators
= Ukulele-Type Weak Terminators
= Site of potential Ukulele Promoters

Figure 1. Phamerator Map of the Ukulele Genome Showing All Predicted Promoter and Terminator Loci. The ruler represents linear genome, in units of 1kb. Colored boxes above and below ruler represent genes transcribed leftwards and rightwards, respectively. Strong terminators are indicated by a red box, weak tandem terminators are indicated by an orange box, and the blue arrows indicate a potential promoter. TGN– indicates the presence of an extended –10 box attached to the promoter.

(a)


gp52 ORF
Conserved Repeat CR1-3
gp52 TSP
gp53 -35 1
gp52 -10 2
gp53 -10 1
gp53 TSP
gp52 TSP
gp52 -10 TGN
Conserved Repeat CR1-2
gp53 -35 2
gp53 -10 TGN-
gp53 TSP
gp53 ORF

(b)

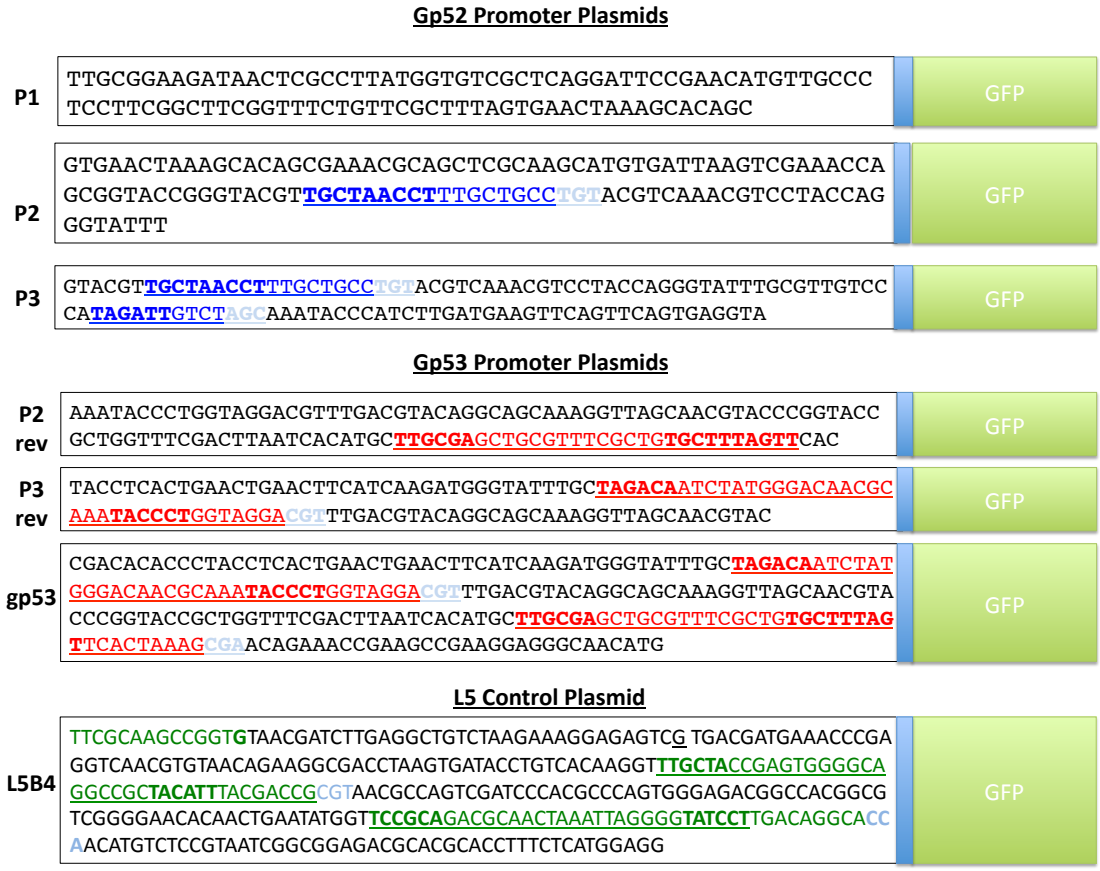


Figure 2. Putative Promoter Locations in Region of Divergent Transcription Between gp52 and gp53 Mapped in Geneious. (a) Green boxes indicate the beginning of ORF for gp52 and gp53, dark blue arrows indicate potential promoter elements for gp52, red arrows indicate potential promoter elements for gp53, yellow boxes indicate conserved repeat sequences CR1-2 and CR1-3 [20], and light blue arrows indicate transcriptional start points (TSP) for each putative promoter. (b) Sequences cloned into reporter plasmid pUV15tetORm [44]. Red sequence represents putative gp53 promoters, dark blue sequence represents putative gp52 promoters, green sequence represents known L5 gp71 promoters, and light blue sequence represents TSP for each promoter.

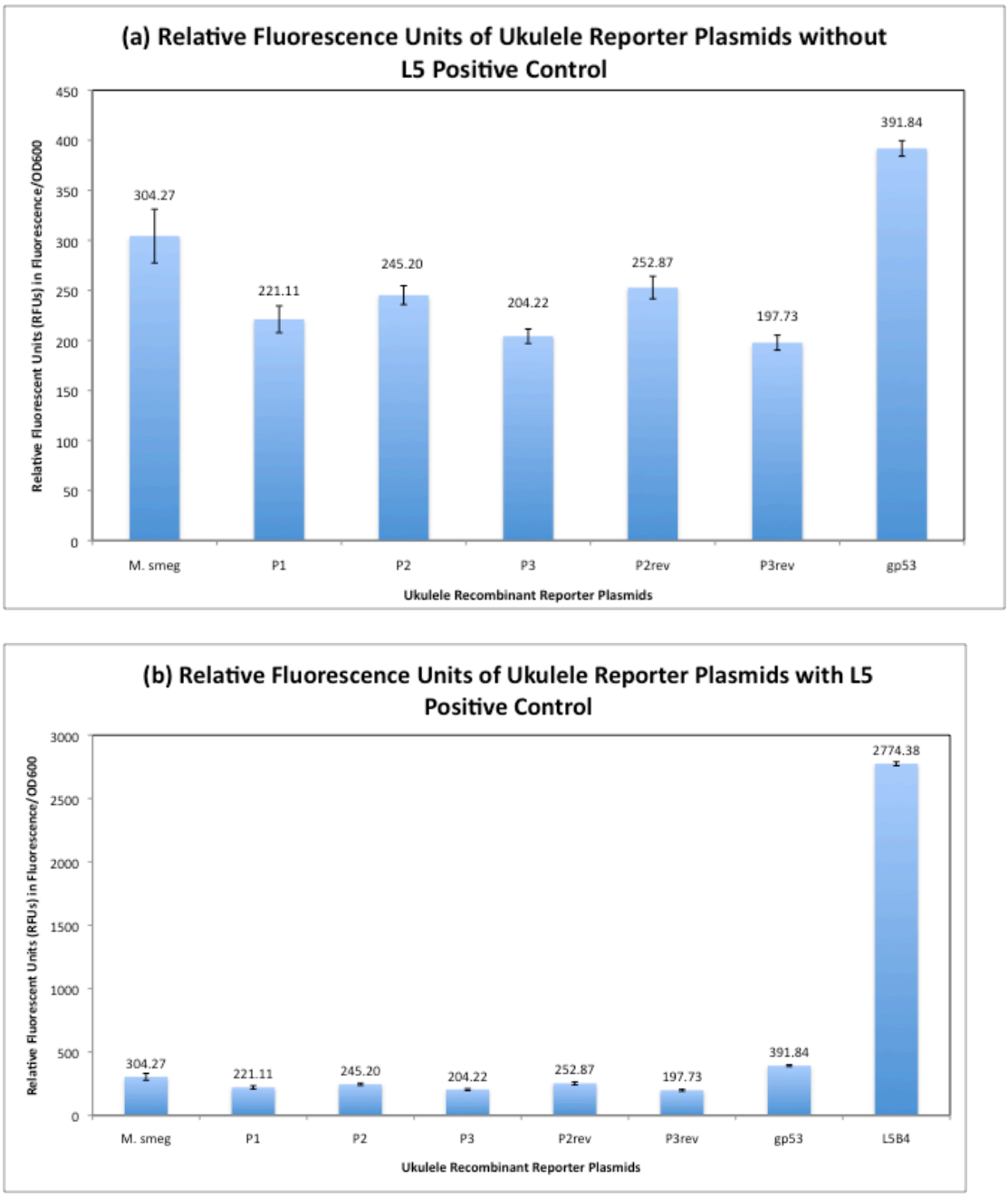
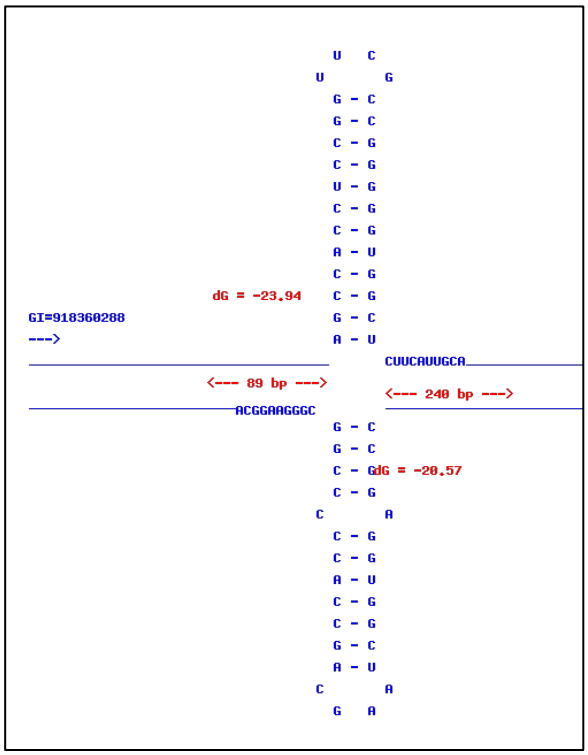


Figure 3. Average Relative Fluorescence of M. Smegmatis Cells Carrying a Recombinant Reporter Plasmid with Varying Putative Promoters Upstream of the GFP Gene (n=8). The GFP fluorescence of *M. smegmatis* cells was normalized by dividing relative fluorescence by the optical density at 600 nm. The error bars represent the standard error of the mean (n=8). (a) Ukulele recombinant plasmids and their average relative fluorescence compared with *M. smegmatis* cells without

plasmid. (b) Ukulele recombinant plasmid average fluorescence compared with confirmed L5 promoter for gp71.

(a)

(b)



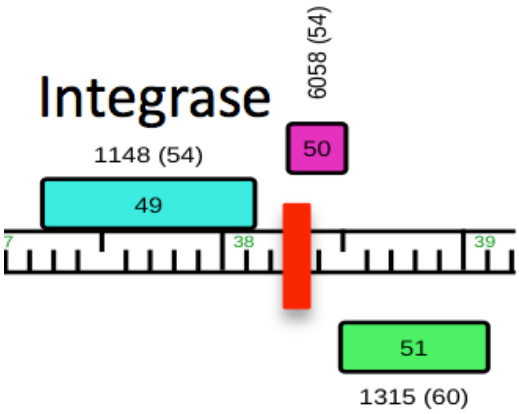
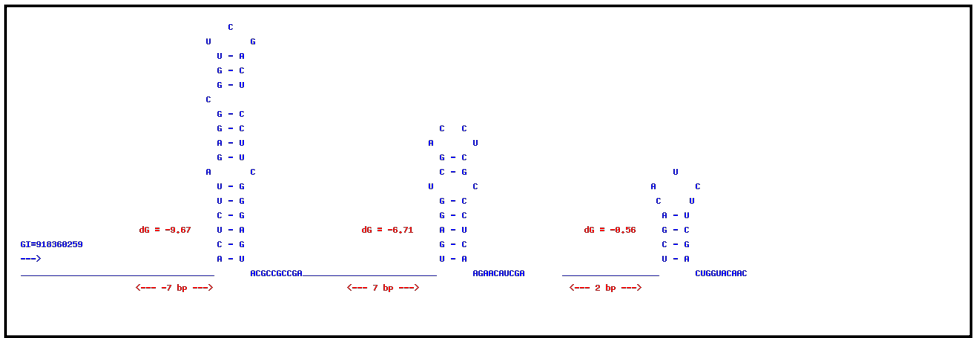


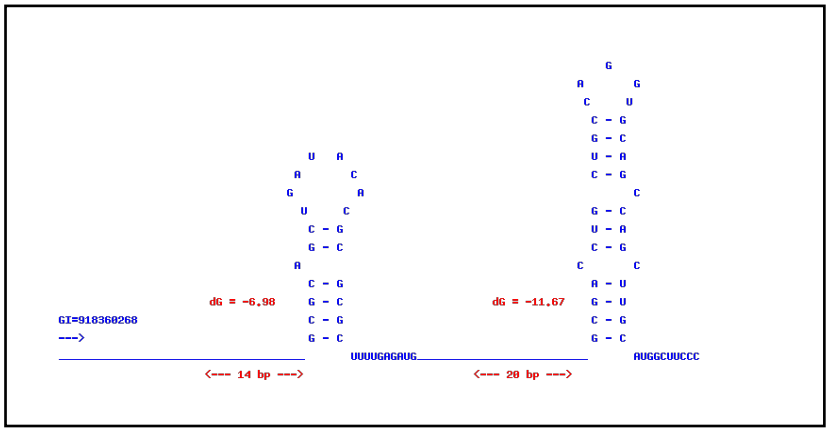
Figure 4. Predicted Terminators Between Converging Ukulele Genes, gp49 and gp51. (a) Predicted RNA secondary structure of putative terminator leading to termination of both gene products as predicted by WebGEsTer [39]. (b) Phamerator map showing region containing potential putative terminator with terminator

indicated by red box. Features of Phamerator map as described in Figure 1.

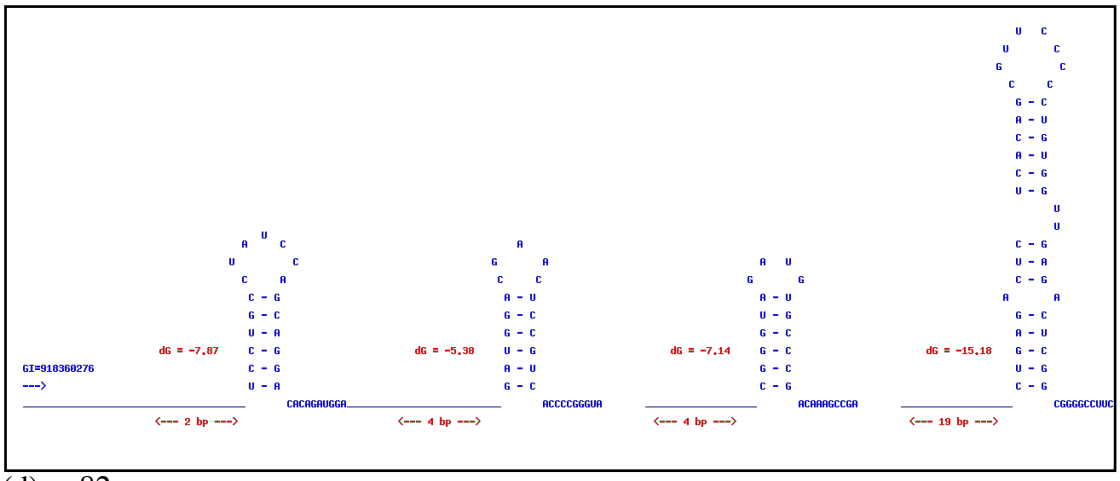
(a) gp20



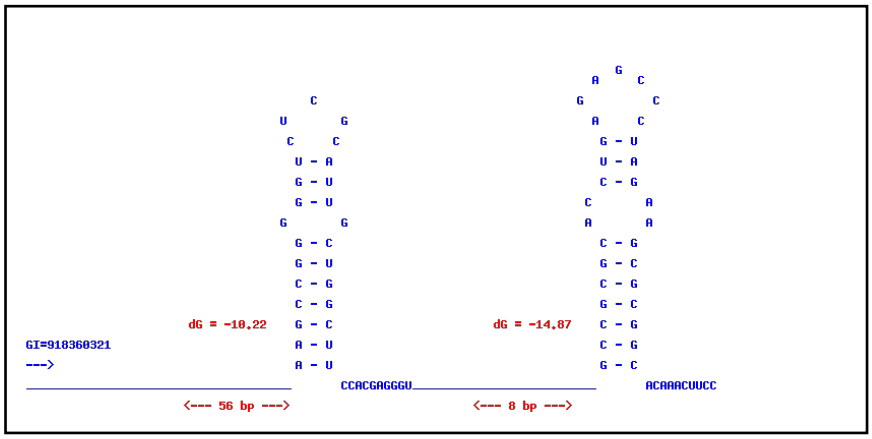
(b) gp29



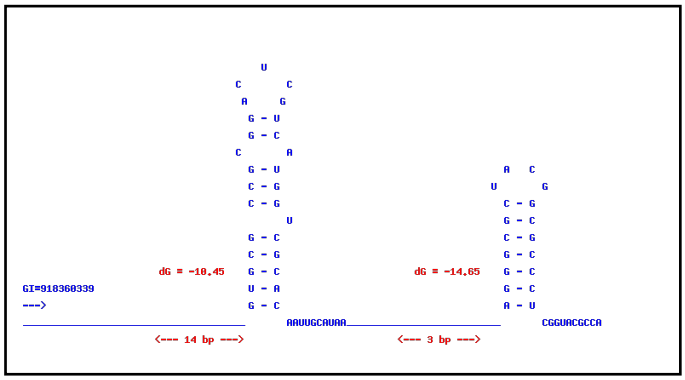
(c) gp37



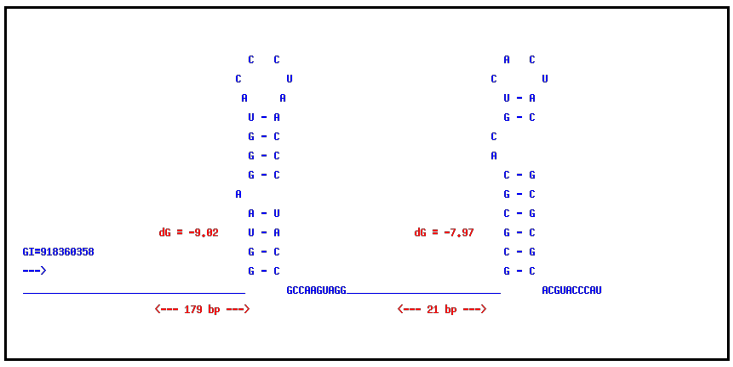
(d) gp82



(e) gp100



(f) gp120



(g) gp30

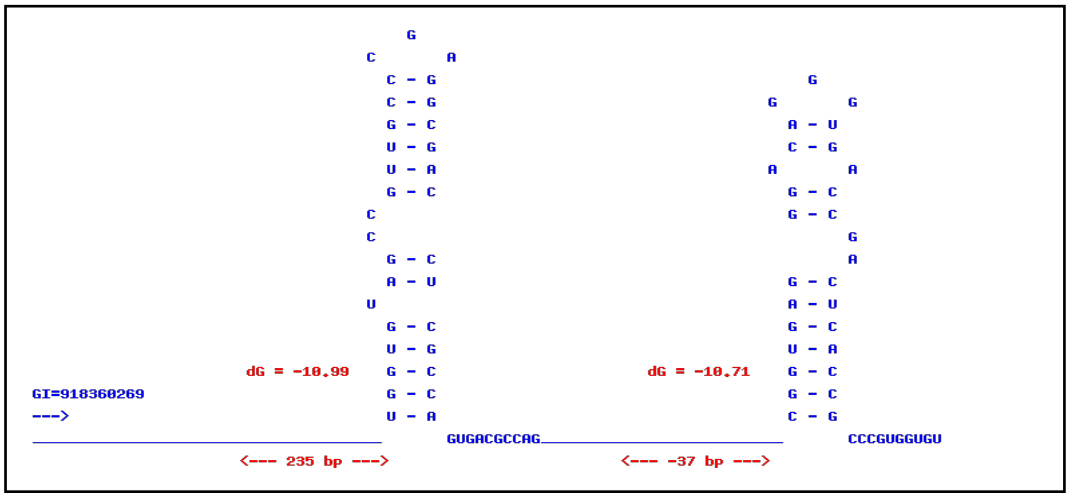
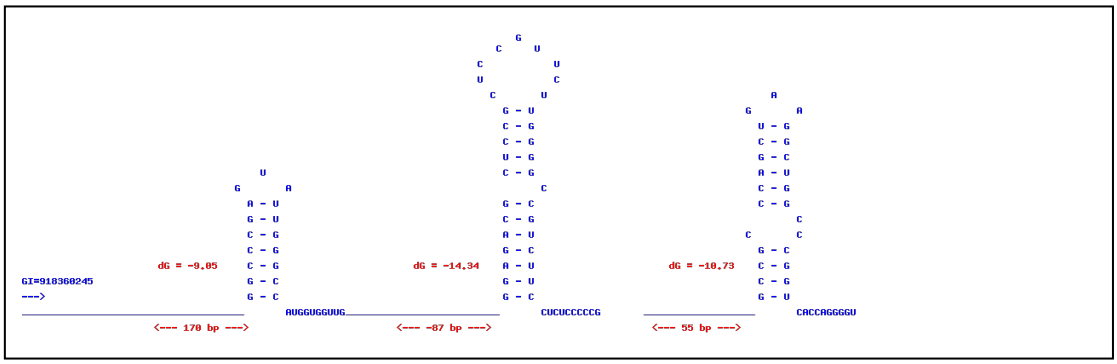
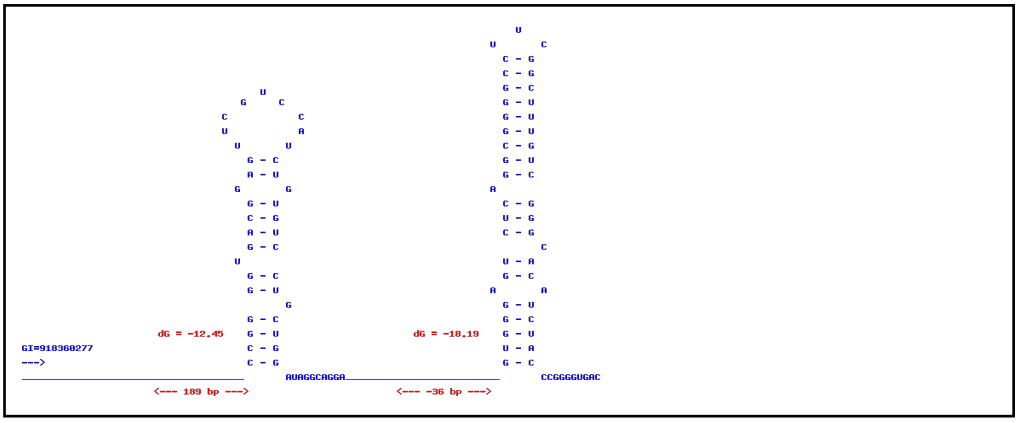


Figure 5. Predicted RNA Secondary Structure of Putative Tandem Terminators. Individual ΔG greater than ΔG threshold -16.02 kcal mol^{-1} established by WebGeSTer. Tandem secondary structures were identified in Ukulele genome downstream of (a) gp20 (tail assembly protein), (b) gp29, (c) gp37 (NrdH-redoxin), (d) gp82, (e) gp100 (HNH endonuclease), (f) gp 120, and (g) gp30. Secondary structure was predicted using WebGEsTer [39].

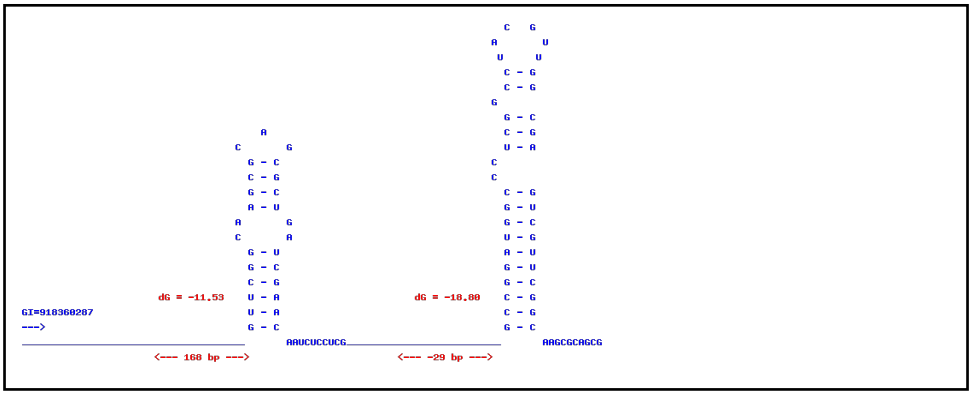
(a) gp6



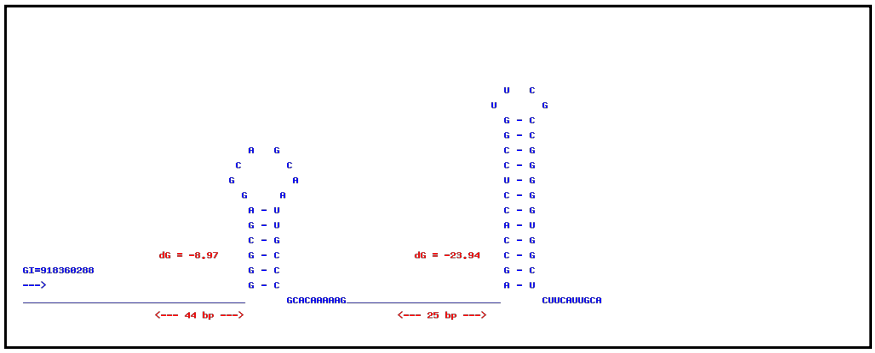
(b) gp38



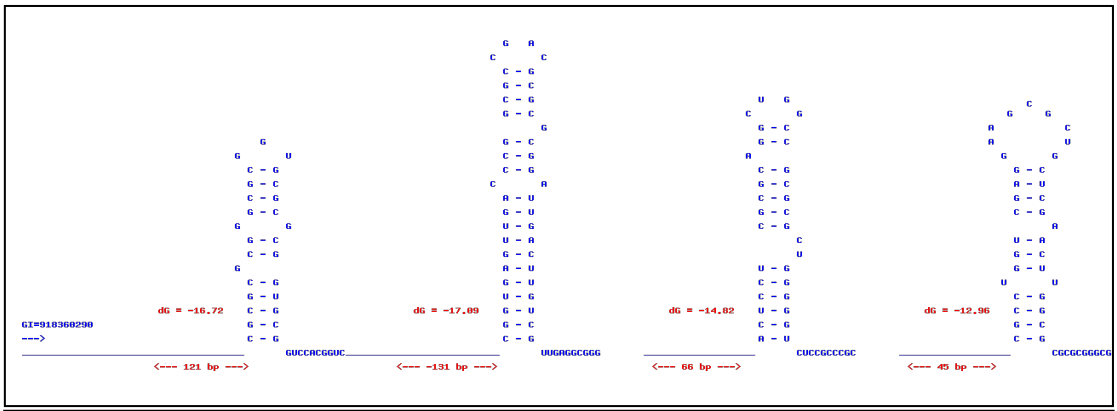
(c) gp48

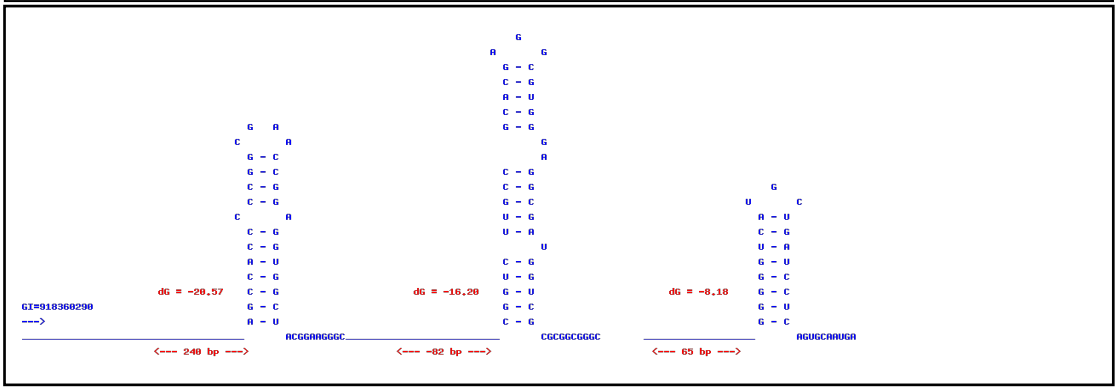


(d) gp49


dG = -8.97
dG = -23.94
GI=910360280
GCACAAAAAG
CUUCAUUGCA
<--- 44 bp --->
<--- 25 bp --->

(e) gp51


dG = -16.72
dG = -17.09
dG = -14.82
dG = -12.96
GI=910360290
GUCCACGGUC
UUGAGGCGGG
CUCCGCCCGC
CGCGCGGGCG
<--- 121 bp --->
<--- 131 bp --->
<--- 66 bp --->
<--- 45 bp --->


dG = -20.57
dG = -16.20
dG = -8.18
GI=910360290
ACGGAAGGGC
CGCGGCGGGC
AGUGCAAUGA
<--- 240 bp --->
<--- 82 bp --->
<--- 65 bp --->

(f) gp120

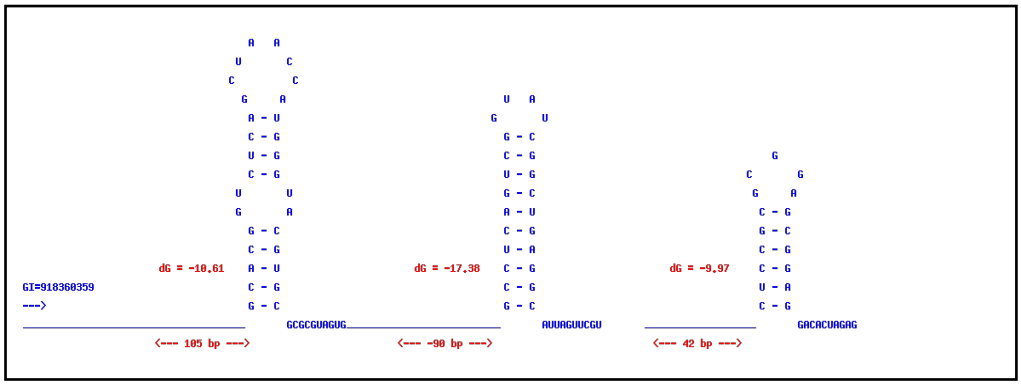


Figure 6. Supporting Terminator Structures Adjacent to Strong Ukulele Terminators that may Assist in Termination of Transcription. Each terminator contains one strong terminator with a ΔG lower than the threshold and other supporting structures with significant ΔG. Supporting secondary structures identified in Ukulele genome downstream of (a) gp6, (b) gp38, (c) gp48, (d) gp49 (tyrosine integrase), (e) gp51, (f) gp120. Secondary structure was predicted using WebGEsTer [39].

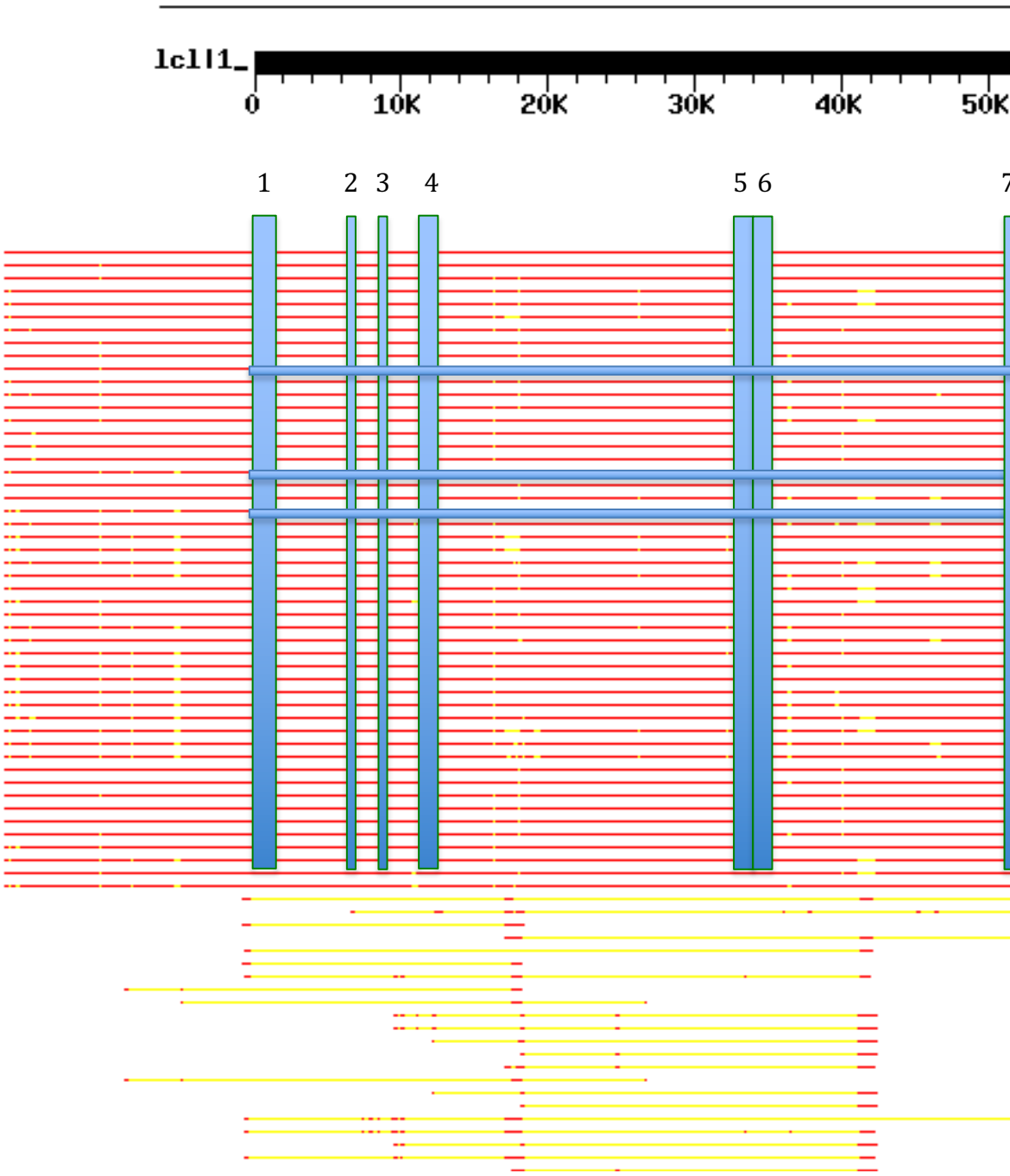
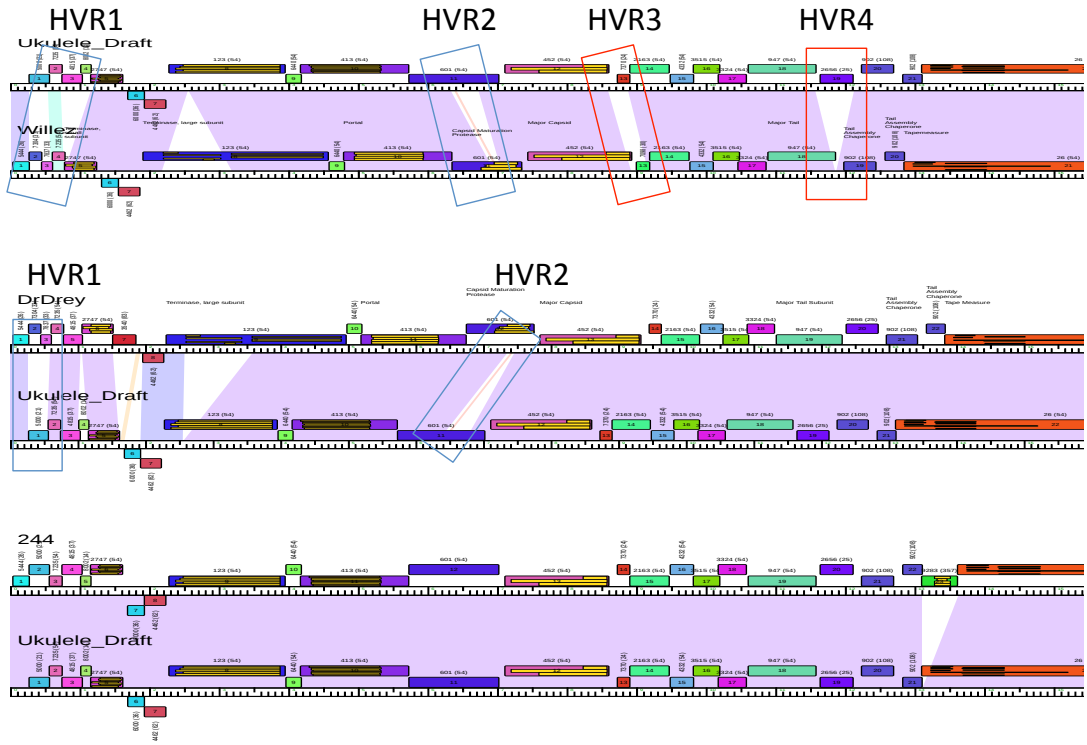
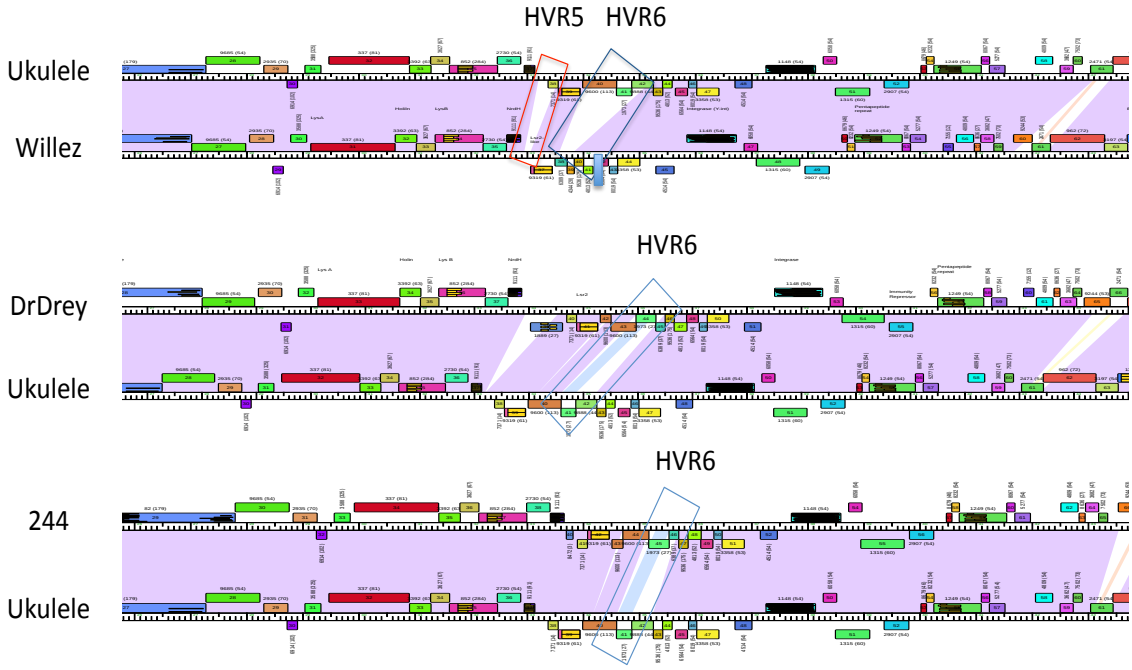


Figure 7. BLAST Alignment on PhagesDB.org of Ukulele Genome. Green boxes indicate 16 regions in cluster E genomes that are highly variable. Blue boxes show phage genomes containing these highly variable regions and chosen for terminator analysis.

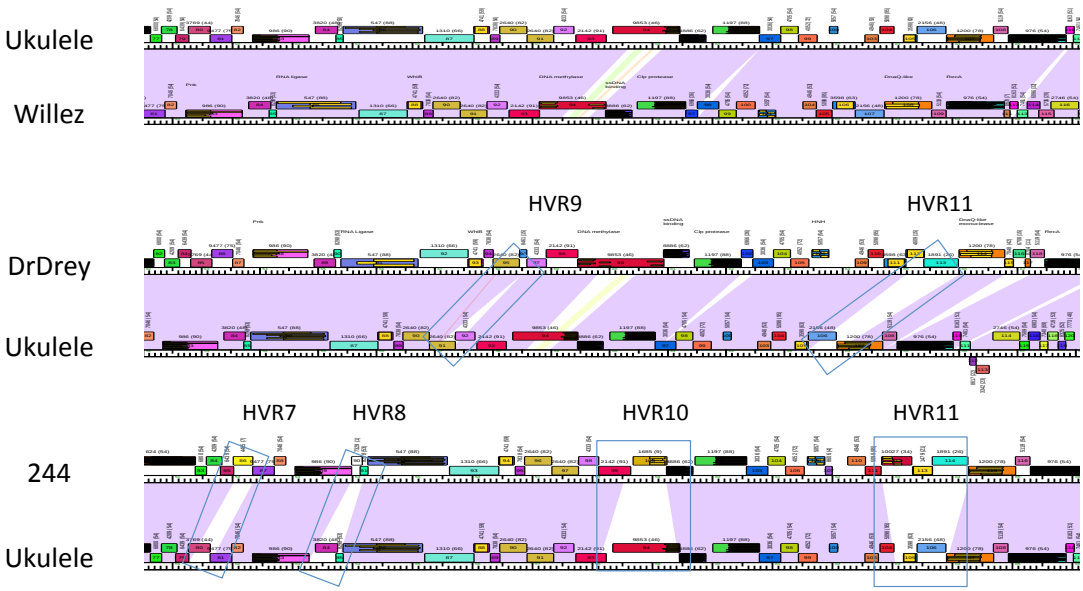
(a)


HVR1
HVR2
HVR3
HVR4
Ukulele_Draft
Willez
HVR1
HVR2
DrDrey
Ukulele_Draft
244
Ukulele_Draft

(b)


HVR5
HVR6
Ukulele
Willez
HVR6
DrDrey
Ukulele
HVR6
244
Ukulele

(c)


Ukulele
Willez
HVR9
HVR11
DrDrey
Ukulele
HVR7
HVR8
HVR10
HVR11
244
Ukulele

(d)

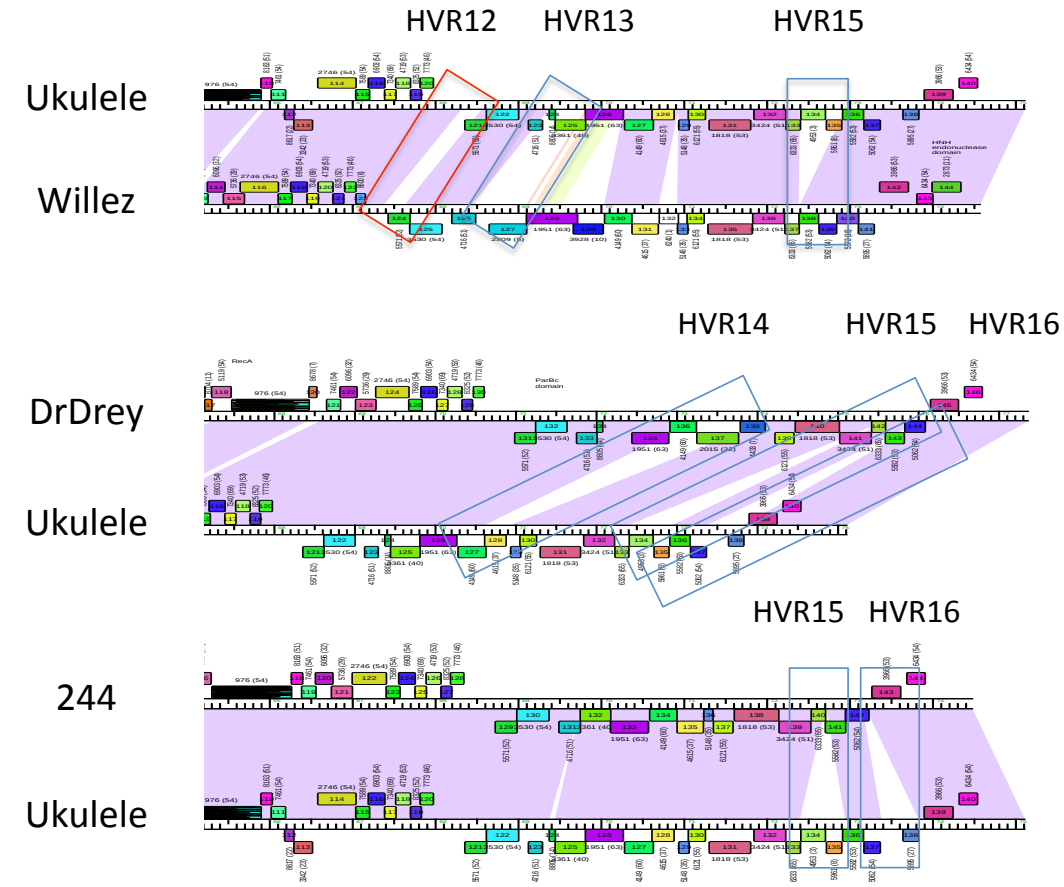

HVR12
HVR13
HVR15
Ukulele
Willez
HVR14
HVR15
HVR16
DrDrey
Ukulele
HVR15
HVR16
244
Ukulele

Figure 8. Phamerator Alignment of Ukulele Variable Regions with Cluster E Phage Genomes Willez, DrDrey, and 244. These genomes collectively contain all of the highly variable regions (HVR) within this cluster. Organization of map is the same as indicated in Figure 1. Boxes indicate presence of Cluster E HVR 1-16. Red boxes indicate the presence of an HVR that may be relevant to terminator identification, while blue boxes indicate intragenic HVRs. (a) HVR1-4, (b) HVR 5-6, (c) HVR 7-11, (d) HVR 12-16

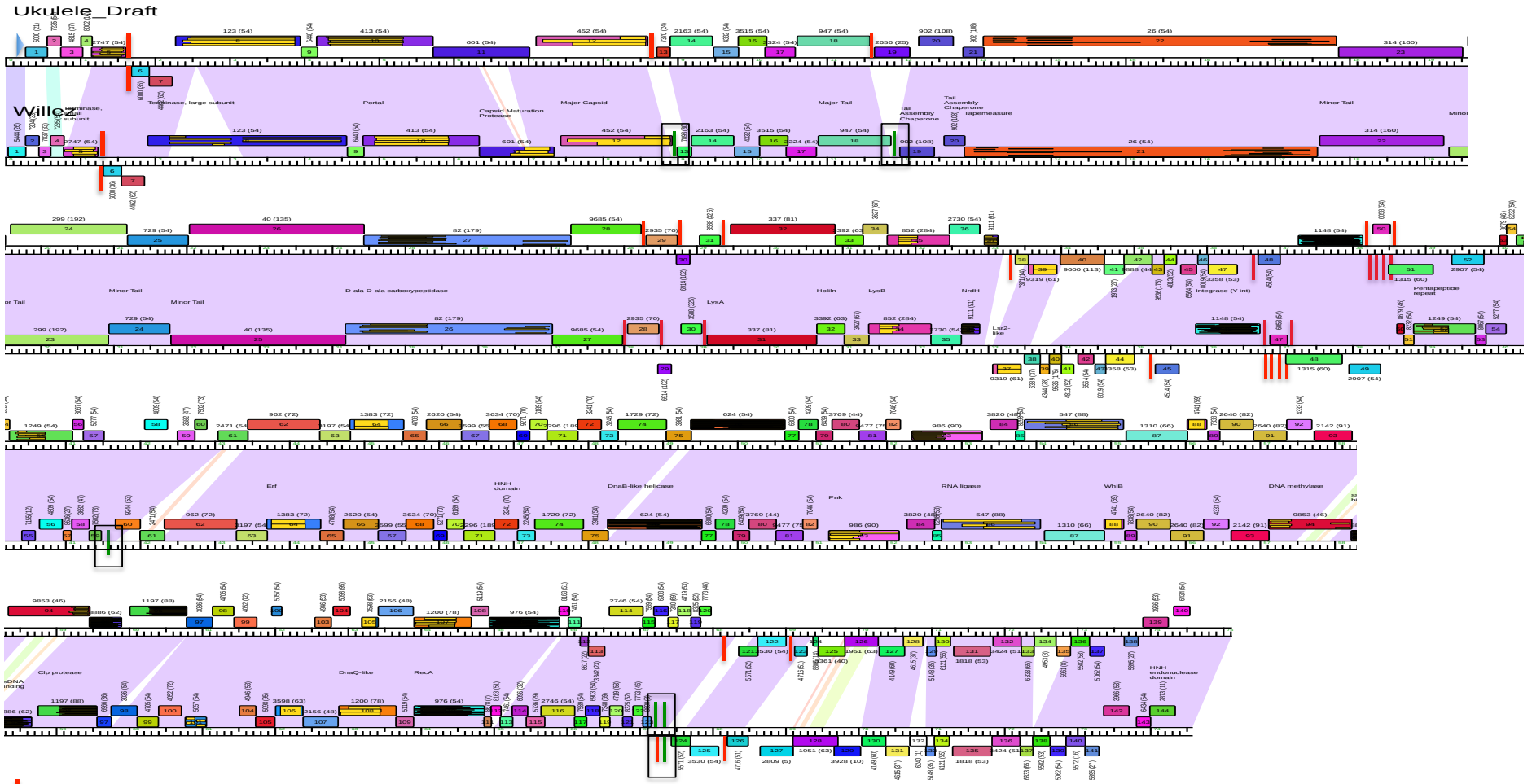

Ukulele_Draft
Willez
= Ukulele-Type Terminators
= Willez-Type Terminators

Figure 9. Phamerator Alignment of Ukulele Genome with Willez Genome. Organization of map is the same as indicated in Figure 1. Purple regions indicate regions of high genomic identity between the genomes. Red boxes indicate terminators in Ukulele that show identity to terminators identified in Willez genome. Green bars indicate unique Willez-type terminators, which show no identity to Ukulele terminators.

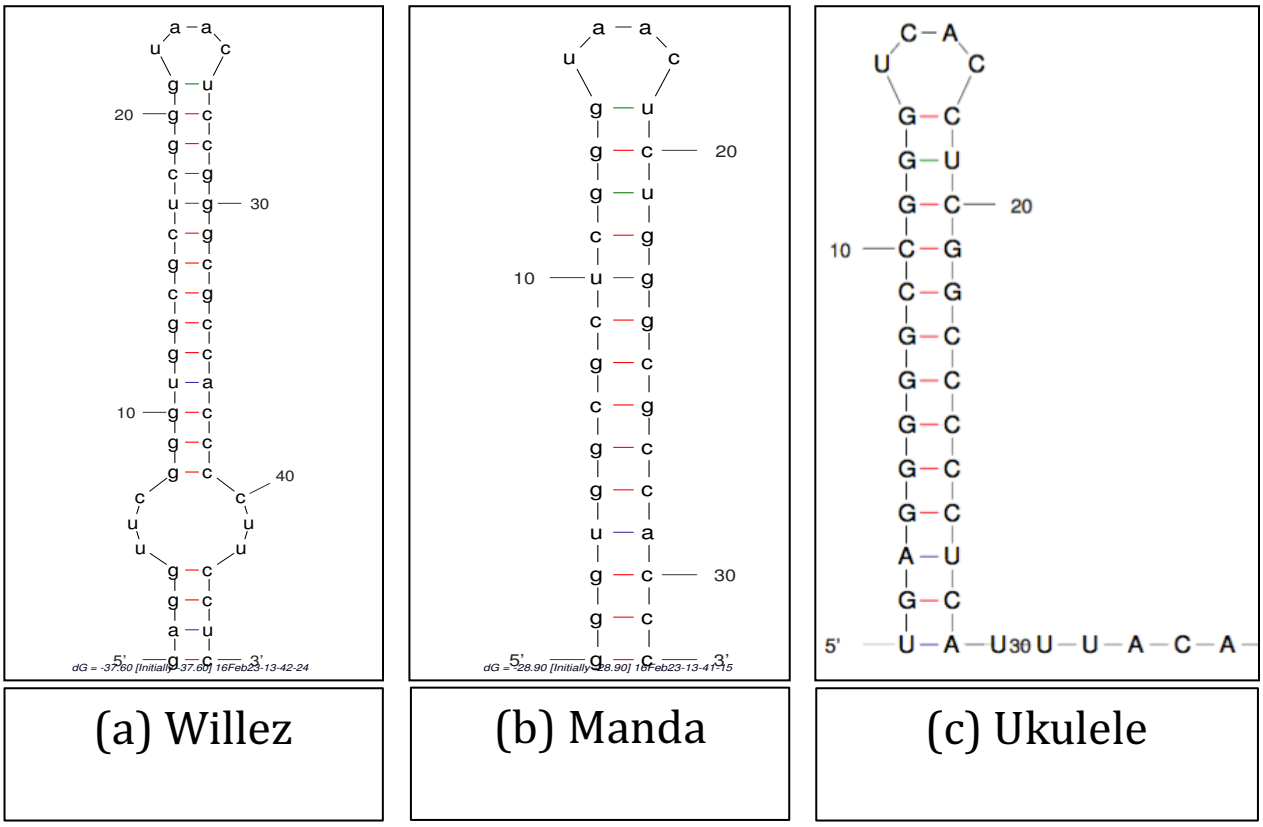
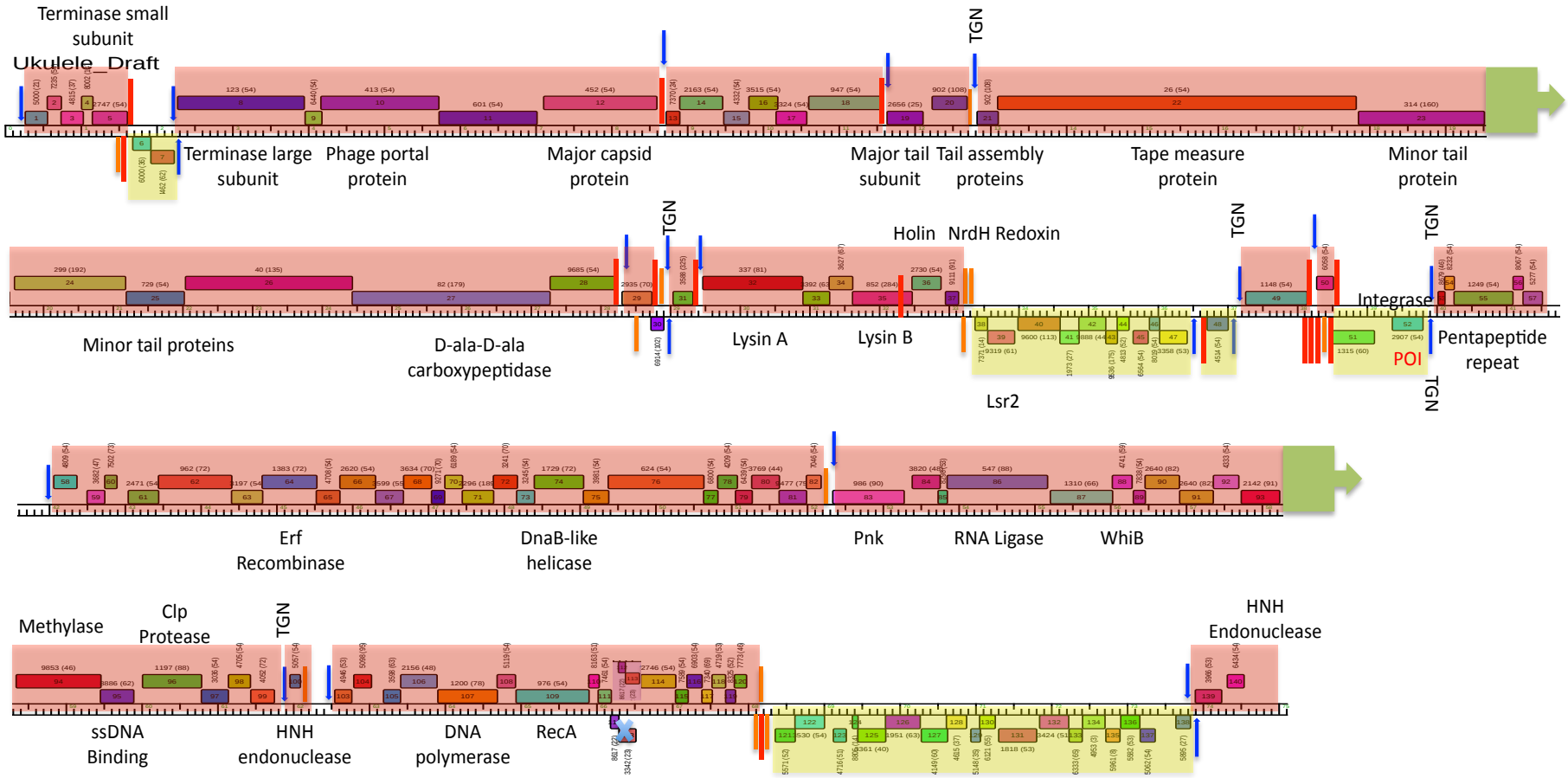


Figure 10. Predicted Secondary Structure of Three Types of Terminators Found Downstream of the Major Capsid Protein Gene in Phage Genomes (a) Willez (b) Manda, and (c) Ukulele. Secondary structure was predicted using the RNA mFOLD program [50].


Terminase small subunit
Ukulele_Draft
Terminase large subunit
Phage portal protein
Major capsid protein
Major tail subunit
Tail assembly proteins
Tape measure protein
Minor tail protein
TGN
Minor tail proteins
D-ala-D-ala carboxypeptidase
Lysin A
Lysin B
Holin
NrdH Redoxin
Integrase
POI
Pentapeptide repeat
Lsr2
Erf Recombinase
DnaB-like helicase
Pnk
RNA Ligase
WhiB
Methylase
Clp Protease
ssDNA Binding
HNH endonuclease
DNA polymerase
RecA
HNH Endonuclease

= Ukulele-Type Strong Terminators
= Ukulele-Type Weak Terminators
= Site of potential Ukulele Promoters

Figure 11. Phamerator Map of Ukulele Genome with Predicted Operons. Map is organized as described in Figure 1. Operons are defined as regions of potential tandem regulation by promoters and terminators. Red boxes indicate operons on forward strand and yellow boxes indicate operons on the reverse strand. Green arrows indicate a line break in which an operon continu

TABLES

Table 1. Primers used in this study

Primer name	Sequence 5' - 3'
gp53_P_F	GTCGACTAGTACCCTACCTCACTGAAC
gp53_P_R	GCCATTAATTAAGTACGTTGCTAACCTTT
L5gp71_P_F	GTAAACTAGTGAGGAGTACATCAAGAGC
L5gp71_P_R	GTCTTTAATTAACCTCCATGAGAAAGCT
gp52_P3_F_RD	AAGTACTAGTGTACGTTGCTAACCTTT
gp52_P3_R_RD	ACCTTAATTAATACCTCACTGAACTGAA
gp52_P2_F_RD	CTCTACTAGTGTCAACTAAACCACAGC
gp52_P2_R_RD	TCGTTAATTAAAAATACCCTGGTAGGAC
gp52_P1_F_RD	TAGACTAGTTTGCGGAAGATACTC
gp52_P1_R_RD	AAATTAATTAAGCGTCAGATTAGATCAAGC
gp52_P2Rev_F_RD	TCGAAATCGTAGTGACGTTTAGTCATACCATAGC
gp52_P2Rev_R_RD	GCAGTACTAGTAAATACCTCCCGTAGGAC
gp52_P3Rev_R_RD	TTGACTTCTGGTCTAGTTCAACTAG
gp52_P3Rev_F_RD	CCCTTAATTAAGTACCTTCCTAACCTTT

Table 2. Putative promoter sequences in Ukulele genome, as identified by PhiSite promoter search. (a) Promoter sequences on forward strand and (b) Promoter sequences on reverse strand.

(a)

Gene Downstream	Sequence	PhiSite -10 Score	Best -10 (Y/N)	Viable -35 (Y/N)	TGN Motif (Y/N)
gp1	TGCACCGAGGGGAACTCCGATCACTGAA**TATCAT**GCACTGC	0.93	Y	N	N
gp3	**TGCACC**GAGGGGAACTCCGATCACTG**GAATAT**CATGCACTGC	0.70	N	Y	N
gp8	TCGCTCATGAAAGCGAGGC**TAAGCT**TAGTCGAC	0.82	Y	N	N
gp13	**TGACGT**TCGTGAGGGGCGCGGCG**TCAGGT**GGGCCCCT	0.56	Y	Y	N
gp19	**GTCAAA**GAGCTTCCCGCGCAGCGC**TGTCTT**CCCAGGAGGT	0.72	Y	Y	N
gp21	CATCGAGATTTCAC**TGCTAAT**TTAGATCAAGC	0.66	Y	N	Y
gp29	TCGAAATCGTAGTGACGTTTAGTCATACCATAGC	0.67	Y	N	N
gp31	**TTGACT**TCTGGTCTAGTTCAACTAGGTGAGCTAACGC	0.84	Y	Y	N
gp32	**CTCGGG**GGCTTTCGCATTGGGGCAA**CATTTT**GAGTGT	0.71	Y	Y	N
gp49	**TTTAGT**TGTCCTGGTGTCGATAA**TACACT**AGGACAACGC	0.86	Y	Y	N
	CCGTAAAAGC**TGGTCTAAT**GTTCCAT	0.84	N	N	Y
gp50	**TCCGGT**TCGCCGGGGGTGGCTC**TTCATT**GCAC	0.64	N	Y	N
gp53	**TAGACA**ATCTATGGGACAACGCAAA**TACCCT**GGTAGGACGT	0.79	Y	Y	N
	TTGCGAGCTGCGTTTCGCTG**TGCTTTAGT**TCACTAAAGCGA	0.74	N	Y	Y
gp58	TCTTACA**TCTTCT**TCCGGGGACGT	0.75	Y	N	N
	GTCATTGTCCATCACGGTCCCGC**TCACAT**TGGGTTGAGCGT	0.71	N	Y	N
gp84	**CTGAGA**GCCCTAGAAGCGCGGCA**CAAACT**TCCAACCAT	0.68	Y	Y	N
gp100	AATTCGCGTG**TGATAACAT**CAAAGGTGC	0.86	Y	N	Y
gp103	**TCGATA**TCCCCAGTGGTCCCGGG**TATAGG**GGATTGG	0.72	Y	Y	N
gp139	**CACAGT**GTCACCTGCTCATCCACGG**TGTAGT**GCTTCAC	0.70	Y	Y	N
	TTCACCGGGCTATTCGGCTTGCG**GAAATT**CTTGATGC	0.69	N	Y	N

Table 3. Strong Ukulele terminators determined by WebGEsTer Analysis

Forward Terminators	Gene Product	Function	Distance from Stop Codon	Terminator or Start Position (bp)	Right Leg of Stem	ΔG (kcal/mole)
UT1	5	terminase small subunit	35	1648	ACTGACCTCCCCGGGTGAATGTACCGGGGAGGCCCAGT	-24.17
UT2	12	major capsid protein	66	8665	TGAGGGGGCCGGGTCACCTCGGCCCCCTCA	-27.21
UT3	18	NKF	4	11581	CCGTGCCCGGGTCAAACACCTTCCCGGGCACGG	-20.52
UT4	28	NKF	16	28254	AAGGCCCCTCGAAATCCGAGGGGCCTT	-21.71
UT5	29	NKF	178	28914	GCAGTCTAGCAGGGCTTGACTTCTGGTCTAGTTCAACTAGACTGT	-21.3
UT6	31	NKF	28	29353	AAGCCCCTGAGTTTCGGCTCGGGGGCTT	-21.07
UT7	49	integrase	89	38229	AGCCACCTCCGGTTCGCCGGGGGTGGCT	-23.94
UT8	50	NKF	110	38628	CCCGGGGGTCGCCAGGTCGGTGACCGCCCGGG	-25.79
Ukulele Reverse Terminators						
URT1	6	NKF	-6	1691	GGACTGACTGGGCCTCCCCGGTACATTCACCCGGGGAGGTCAGTTTCGTCT	-25.94
URT2	38	NKF	144	33231	GTGGGAGTCTCAGGCGGGGCCTTCGGCTTTGTCGGGCACATCTAC	-18.19
URT3	48	NKF	136	36570	GCCGGATGGCCCTCGGCCTACGTTGGCGAGTCGTTCGGC	-18.8
URT4	51	NKF	240	38256	AGCCACCCCCGGCGAACCGGAGGTGGCT	-20.57
URT5	51	NKF	154	38342	CGGTCTTGCCCCACGAGGCGTGGGAGGCGATGGTCG	-16.2
URT6	51	NKF	121	38375	CGCGCGCGGGCGCGGTGCGCGCGGTGCG	-16.72
URT7	51	NKF	-10	38506	CGGTGAGTTGACCCGGCGCCGACGCGCGCGGATTGACTTGTCG	-17.09
URT8	122	NKF	130	68228	CGTGCGCGCGTAGTGACGTGCGCGCG	-21.28
URT9	122	NKF	7	68351	GCCTCAGTCGGTATCGGCTGAGGC	-17.38
U-shaped Terminators						
URT10a	124	NKF	68	69057	CAGAGTCCTCAGCCATGCGCTGAGCGCTCTG	-15.86
URT10b	124	NKF	17	69108	GCCCGCACGGGATATGCCGTGCGGGC	-21.89
X-shaped Terminators						
8	49	integrase	89	38229	AGCCACCTCCGGTTCGCCGGGGGTGGCT	-23.94
4	51	NKF	240	38256	AGCCACCCCCGGCGAACCGGAGGTGGCT	-20.57

(b)

Gene Downstream	Sequence	PhiSite -10 Score	Best -10 (Y/N)	Viable -35 (Y/N)	TGN Motif (Y/N)
gp7	TTCGATGTCATCGAGGGTGTAGCCGCATACTTCCATGA	0.78	Y	Y	N
gp30	CCCGTCAGGTCGGTGCGTTAGCTACAGTCTAGTTGA	0.84	Y	N	N
	GTCAGGTCGGTGCGTTAGCTACAGTCTAGTTGAACTAG	0.68	N	Y	N
gp47	GCCGGATGGCCCTCGGCCTACGTTGGCGAGTCGT	0.77	Y	N	N

Table 4. Class E Representative Phage and their putative strong terminators. Terminators of (a) 244 (b) DrDrey and (c) Willez

(b)

(a)

244 Initial Terminators	DrDrey Initial Terminators	Gene Product	Distance from Stop Codon	Terminator Position on Genome	Terminator Sequence	ΔG (kcal/mole)	Ukulele T-Type
1	1	6	35	1645	ACTGACCTCCCCGGGTGAATGTACCGGGGAGGCCCAGT	-24.17	UT1
2	2	17	25	1669	AGAGGGTGCCGGTGTGACATGTACCGGTGAGGCTCAGT	-22.86	UT1
3	3	13	64	9593	TGAGGCGCGCGGTCATCACCTCGGCCGGCTCA	-27.81	UT3
4	4	30	16	28770	AAGGGCCCCGGGAAATCAGCGGGGCGTT	-20.38	UT4
5	5	31	178	29436	GCAGTCTAGCAGGGCTTGACTTCTGGTCTAGTTCAACTAGACTGT	-21.3	UT5
6	6	32	28	39087	AAGCACCTGAGTTTCGGCTCGGGGGGCTT	-23.94	UT6
7	7	52	89	39483	AGCCACCTGCGGTTCGCCGGGGGTGGCT	-23.94	UT7
8	8	53	110	39882	CCCGGGGGTCGCCAGGTCGGTGACCGCCCGGG	-25.79	UT8
Reverse							
1	9	129	17	68543	GGACTGACTGGGCCTCCTCGGTACATTCACCCGGGGAGGTCAGTTTCGTCT	-25.94	URT1
2	10	52	136	37022	GCCGGATGGCCCTCGGCCTACGTTGGCGAGTCGTTCGGC	-18.8	URT3
3	11	55	240	38708	AGCCACCCCCGGGCGAACCGGAGGTGGCT	-20.57	URT4
4	12	51	136	37824	GCCCGATGGCCCCTCGGCCTACGTTGGCGAGTCGTTCGGC	-18.8	URT3
5	13	55	121	38827	CGCGCGCGGGCGCGGTGCGCGCGGTGCG	-16.72	URT6
6	14	54	154	39296	CGGTGAGTTGACCCGGCGCCGACGCGCGCGGATTGACTTGTCG	-17.09	URT7
7	15	129	130	68595	CGTGCGCGCGTAGTGACGTGCGCGCG	-21.28	URT8
8	16	131	68	69425	CAGAGTGCTCAGCCATGAGGCTGAGCGCTCTG	-25.44	URT10a
9	17	133	68	71689	CAGAGTGCTCAGCCATGAGGCTGAGCACTCTG	-25.87	URT10a
10	18	133	17	71740	GCCCGCGCGGAATATGCCGCGCGGGC	-24.72	URT10b

(c)

Willez Initial Terminators	Gene Product	Distance from Stop Codon	Terminator Position on Genome	Terminator Sequence	ΔG (kcal/ mole)	Ukulele T-Type
1	5	35	1272	ACTGACCTCCCCGGGTGAATGTACCGGGGAGGCCCAGT	-24.17	UT1
2	**12**	**7**	**8935**	**GAGGTTCGGGTGGCGCTCGGGTAACTCCGGGCGCCACCCCTTCCTC**	**-37.46**	**WT1**
3	**18**	**25**	**11882**	**CCCGCTGCCCGCCACCGCCCTTCGTGGTACGGCGGGTACGGG**	**-27.16**	**WT2**
4	27	16	28000	AAGGCCCCTCGAAATCCGAGGGGCCTT	-21.71	UT4
5	28	178	28660	GCAGTCTAGCAGGGCTTGACTTCTGGTCTAGTTCAACTAGACTGT	-21.3	UT5
6	30	28	29099	AGGCCCCTGAGTTTCGGCTCGGGGGCTT	-20.72	UT6
7	46	89	36819	AGCCACCTCCGGTTCGCCGGGGGTGGCT	-23.94	UT7
8	47	110	37218	CCCGGGGGTCGCCAGGTCGGTGACCGCCCGGG	-25.79	UT8
9	**59**	**49**	**41530**	**GGGACGAGCGCGCCCAGCGTGGCTGGCTGGGGTCTTTCCT**	**-16.9**	**WT3**
10	**123**	**110**	**67268**	**GCGCGCGCGCCACCCCTGCGCGCACGTGC**	**-17.97**	**WT4**
11	**123**	**241**	**67399**	**GCCCTCAGCCCATGAACGGACTGAGGGC**	**-17.52**	**WT5**
Reverse						
1	6	-6	1315	GGACTGACTGGGCCTCCCCGGTACATTCACCGGGGAGGTCAGTTTCGTCT	-25.94	URT1
2	45	136	35160	GCCGGATGGCCCTCGGCCTACGTTGGCGAGTCGTTCGGC	-18.8	URT3
3	48	240	36846	AGCCACCCCCGGCGAACCGGAGGTGGCT	-20.57	URT4
4	48	154	36932	CGGTCTTGCCCCACGAGGCGTGGGAGGCGATGGTCG	-16.2	URT5
5	48	121	36965	CGCGCGCGGGCGCGGTGCGCGCGGTGCG	-16.72	URT6
6	48	-10	37096	CGGTGAGTTGACCCGGCGCCGACGCGCGCGGATCGACTTGTCG	-22.42	URT7
7	124	140	67293	CGTGCGCGCAGGGGTGGCGCGCGCG	-18.56	URT8
8	**124**	**3**	**67430**	**GCATGCCCTCAGTCCGTTCATGGGCTGAGGGCATGC**	**-30.11**	**WRT1**
9	126	68	68133	CAGAGTCCTCAGCCATGAGGCTGAGCGCTCTG	-19.52	URT10a
10	126	17	68184	GCCCGCGCGGAATATGCCGCGCGGGC	-24.72	URT10b

Table 5. Predicted Dumbo terminators with their corresponding Ukulele-type terminator sequences

Ukulele Type	Gene Product	ΔG	Terminator Start Position (bp)	Full Terminator Sequence
UT1	5	-24.17	1648	ACTGACCTCCCCGGGTGAATGTACCGGGGAGGCCCAGT
UT2	12	-27.21	8374	TGAGGGGGCCGGGTCACCTCGGCCCCCTCA
UT3	18	-20.52	11290	CCGTGCCCGGGTCAAACACCTTCCCGGGCACGG
UT4	28	-21.71	27963	AAGGCCCCTCGAAATCCGAGGGGCCTT
UT5	29	-21.3	28623	GCAGTCTAGCAGGGCTTGACTTCTGGTCTAGTTCAACTAGACTGT
UT6	31	-21.07	29062	AAGCCCCTGAGTTTCGGCTCGGGGGCTT
UT7	52	-23.94	38259	AGCCACCTCCGGTTCGCCGGGGGTGGCT
UT8	53	-25.79	38658	CCCGGGGGTCGCCAGGTCGGTGACCGCCCGGG
URT1	6	-25.94	1691	GGACTGACTGGGCCTCCCCGGTACATTCACCCGGGGAGGTCAGTTTCGTCT
URT2	38	-18.19	32940	GTGGGAGTCTCAGGCGGGGCCTTCGGCTTTGTCGGGCACATCTAC
URT3	51	-18.8	36600	GCCGGATGGCCCTCGGCCTACGTTGGCGAGTCGTTCGGC
URT4	54	-20.57	38286	AGCCACCCCCGGCGAACCGGAGGTGGCT
URT7	54	-22.42	38536	CGGTGAGTTGACCCGGCGCCGACGCGCGCGGATCGACTTGTCG
URT8	129	-21.28	68855	CGTGCGCGCGTAGTGACGTGCGCGCG
URT10b	131	-21.89	69735	GCCCGCACGGGATATGCCGTGCGGGC

REFERENCES

1. Hatfull GF. Mycobacteriophages: Genes and Genomes. Annu Rev Microbiol. 2010;64: 331–356. doi:10.1146/annurev.micro.112408.134233

2. Cosma CL, Sherman DR, Ramakrishnan L. The secret lives of the pathogenic mycobacteria. Annu Rev Microbiol. 2003;57: 641–676. doi:10.1146/annurev.micro.57.030502.091033

3. Ford ME, Stenstrom C, Hendrix RW, Hatfull GF. Mycobacteriophage TM4: genome structure and gene expression. Tuber Lung Dis. 1998;79: 63–73. doi:10.1054/tuld.1998.0007

4. Newton-Foot M, Gey van Pittius NC. The complex architecture of mycobacterial promoters. Tuberculosis. 2013;93: 60–74. doi:10.1016/j.tube.2012.08.003

5. Orosz A, Boros I, Venetianer P. Analysis of the complex transcription termination region of the Escherichia coli rrnB gene. Eur J Biochem. 2004;201: 653–659.

6. Czyz A, Mooney RA, Iaconi A, Landick R. Mycobacterial RNA Polymerase Requires a U-Tract at Intrinsic Terminators and Is Aided by NusG at Suboptimal Terminators. mBio. 2014;5: e00931–14–e00931–14. doi:10.1128/mBio.00931-14

7. Ingham CJ, Hunter IS, Smith MC. Rho-independent terminators without

3' poly-U tails from the early region of actinophage øC31. Nucleic Acids Research. 1995;23: 370–376.

8. Hatfull GF. The Secret Lives of. 1st ed. Bacteriophages, Part A. Elsevier Inc; 2012. pp. 179–288. doi:10.1016/B978-0-12-394621-8.00015-7

9. Broussard GW, Oldfield LM, Villanueva VM, Lunt BL, Shine EE, Hatfull GF. Integration-Dependent Bacteriophage Immunity Provides Insights into the Evolution of Genetic Switches. Molecular Cell. 2013;49: 237–248. doi:10.1016/j.molcel.2012.11.012

10. Brown KL, Sarkis GJ, Wadsworth C, Hatfull GF. Transcriptional silencing by the mycobacteriophage L5 repressor. EMBO J. 1997;16: 5914–5921. doi:10.1093/emboj/16.19.5914

11. Jacobs-Sera D, Marinelli LJ, Bowman C, Broussard GW, Bustamante CG, Boyle MM, et al. On the nature of mycobacteriophage diversity and host preference. Virology. Elsevier; 2012;434: 187–201. doi:10.1016/j.virol.2012.09.026

12. Pope WH, Jacobs-Sera D, Russell DA, Peebles CL, Al-Atrache Z, Alcoser TA, et al. Expanding the Diversity of Mycobacteriophages: Insights into Genome Architecture and Evolution. Aziz R, editor. PLoS ONE. 2011;6: e16329. doi:10.1371/journal.pone.0016329.s007

13. Hendrix RW. Bacteriophages: Evolution of the Majority. Theoretical Population Biology. 2002;61: 471–480. doi:10.1006/tpbi.2002.1590

14. Hendrix RW, Hatfull GF, Smith MCM. Bacteriophages with tails:

chasing their origins and evolution. Research in Microbiology. 2003;154: 253–257. doi:10.1016/S0923-2508(03)00068-8

15. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, et al. Origins of highly mosaic mycobacteriophage genomes. Cell. 2003;113: 171–182.

16. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci USA. 1999;96: 2192–2197.

17. Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko C-C, et al. Comparative Genomic Analysis of 60 Mycobacteriophage Genomes: Genome Clustering, Gene Acquisition,and Gene Size. Journal of Molecular Biology. Elsevier Ltd; 2010;397: 119–143. doi:10.1016/j.jmb.2010.01.011

18. The Actinobacteriophage Database | Home. In: phagesdb.org [Internet]. [cited 15 Apr 2016]. Available: http://phagesdb.org

19. Dodd IB, Shearwin KE, Egan JB. Revisited gene regulation in bacteriophage λ. Current Opinion in Genetics & Development. 2005;15: 145–152. doi:10.1016/j.gde.2005.02.001

20. Beacham G. Complete Annotation of the Cluster E Mycobacteriophage Ukulele Genome and Characterization of Cluster E Lysogeny Regulation. 2015 May pp. 1–97.

21. Herskowitz I, Hagen D. The lysis-lysogeny decision of phage lambda:

explicit programming and responsiveness. Annu Rev Genet. 1980;14: 399–445. doi:10.1146/annurev.ge.14.120180.002151

22. KAISER AD. Mutations in a temperate bacteriophage affecting its ability to lysogenize Escherichia coli. Virology. 1957;3: 42–61. doi:10.1016/0042-6822(57)90022-3

23. Reichardt LF. Control of bacteriophage lambda repressor synthesis after phage infection: the role of the N, cII, cIII and cro products. Journal of Molecular Biology. 1975.

24. Kobiler O, Rokney A, Oppenheim AB. Phage Lambda CIII: A Protease Inhibitor Regulating the Lysis-Lysogeny Decision. Herman C, editor. PLoS ONE. 2007;2: e363. doi:10.1371/journal.pone.0000363.t002

25. Echols H, Green L, Oppenheim AB, Oppenheim A, Honigman A. Role of the cro gene in bacteriophage lambda development. Journal of Molecular Biology. 1973;80: 203–216. doi:10.1016/0022-2836(73)90167-8

26. Dodd IB, Shearwin KE, Perkins AJ, Burr T, Hochschild A, Egan JB. Cooperativity in long-range gene regulation by the lambda CI repressor. Genes Dev. 2004;18: 344–354. doi:10.1101/gad.1167904

27. Kourilsky P. Lysogenization by bacteriophage lambda and the regulation of lambda repressor synthesis. Virology. 1971;45: 853–857.

28. Broussard GW, Hatfull GF. Evolution of genetic switch complexity. Bacteriophage. 2014;3: e24186. doi:10.4161/bact.24186

29. Nesbit CE, Levin ME, Donnelly-Wu MK, Hatfull GF. Transcriptional regulation of repressor synthesis in mycobacteriophage L5. Mol Microbiol. 2006;17: 1045–1056.

30. Lee MH, Hatfull GF. Mycobacteriophage L5 integrase-mediated site-specific integration in vitro. Journal of Bacteriology. 1993;175: 6836–6841.

31. Dedrick RM, Marinelli LJ, Newton GL, Pogliano K, Pogliano J, Hatfull GF. Functional requirements for bacteriophage growth: gene essentiality and expression in mycobacteriophage Giles. Mol Microbiol. 2013;88: 577–589. doi:10.1111/mmi.12210

32. Hatfull GF, Sarkis GJ. DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics. Mol Microbiol. 1993;7: 395–405.

33. Mulder MA, Zappe H, Steyn LM. Mycobacterial promoters. Tuber Lung Dis. 1997;78: 211–223.

34. Bashyam MD, Kaushal D, Dasgupta SK, Tyagi AK. A study of mycobacterial transcriptional apparatus: identification of novel features in promoter elements. Journal of Bacteriology. 1996;178: 4847–4853.

35. Bashyam MD, Tyagi AK. Identification and analysis of “extended -10” promoters from mycobacteria. Journal of Bacteriology. 1998;180: 2568–2573.

36. Oldfield LM, Hatfull GF. Mutational Analysis of the Mycobacteriophage

BPs Promoter PR Reveals Context-Dependent Sequences for Mycobacterial Gene Expression. Journal of Bacteriology. 2014;196: 3589–3597. doi:10.1128/JB.01801-14

37. Garcia M, Pimentel M, Moniz-Pereira J. Expression of Mycobacteriophage Ms6 Lysis Genes Is Driven by Two 70-Like Promoters and Is Dependent on a Transcription Termination Signal Present in the Leader RNA. Journal of Bacteriology. 2002;184: 3034–3043. doi:10.1128/JB.184.11.3034-3043.2002

38. Jain S, Hatfull GF. Transcriptional regulation and immunity in mycobacteriophage Bxb1. Mol Microbiol. 1910;38: 971–985.

39. Mitra A, Kesarwani AK, Pal D, Nagaraja V. WebGeSTer DB--a transcription terminator database. Nucleic Acids Research. 2010;39: D129–D135. doi:10.1093/nar/gkq971

40. Naville M, Ghuillot-Gaudeffroy A, Marchais A, Gautheret D. ARNold: A web tool for the prediction of Rho-independent transcription terminators. RNA Biology. 2014;8: 11–13. doi:10.4161/rna.8.1.13346

41. BULLOCK WO. XL1-Blue: a high efficiency plasmid transforming recA Escherichia coli strain with beta-galactosidase selection. Bio Tecchniques. 1987;5: 376–379.

42. Guo XV, Monteleone M, Klotzsche M, Kamionka A, Hillen W, Braunstein M, et al. Silencing Essential Protein Secretion in Mycobacterium smegmatis by Using Tetracycline Repressors. Journal of

Bacteriology. 2007;189: 4614–4623. doi:10.1128/JB.00216-07

43. Klucar L, Stano M, Hajduk M. phiSITE: database of gene regulation in bacteriophages. Nucleic Acids Research. 2009;38: D366–D370. doi:10.1093/nar/gkp911

44. Ehrt S. Controlling gene expression in mycobacteria with anhydrotetracycline and Tet repressor. 2005;33: e21–e21. doi:10.1093/nar/gni013

45. Cohen SN, Chang AC, Hsu L. Nonchromosomal antibiotic resistance in bacteria: genetic transformation of Escherichia coli by R-factor DNA. Proc Natl Acad Sci USA. 1972;69: 2110–2114.

46. van Kessel JC, Marinelli LJ, Hatfull GF. Recombineering mycobacteria and their phages. Nat Rev Micro. 2008;6: 851–857. doi:10.1038/nrmicro2014

47. Unniraman S. Alternate Paradigm for Intrinsic Transcription Termination in Eubacteria. Journal of Biological Chemistry. 2001;276: 41850–41855. doi:10.1074/jbc.M106252200

48. Cresawn SG, Bogel M, Day N, Jacobs-Sera D, Hendrix RW, Hatfull GF. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinformatics. BioMed Central Ltd; 2011;12: 395. doi:10.1186/1471-2105-12-395

49. Cresawn SG, Pope WH, Jacobs-Sera D, Bowman CA, Russell DA, Dedrick RM, et al. Comparative Genomics of Cluster O

Mycobacteriophages. van Raaij MJ, editor. PLoS ONE. 2015;10: e0118725. doi:10.1371/journal.pone.0118725.s001

50. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research. 2003;31: 3406–3415. doi:10.1093/nar/gkg595

Author's Biography

Campbell Belisle Haley was raised in Yarmouth, ME and graduated from Yarmouth High School in 2011. Following high school, he travelled to Guatemala City to volunteer at an educational reinforcement center named Safe Passage. This experience continues to motivate and inspire him as he travels the long road towards becoming a physician. Campbell will graduate from the University of Maine in May 2016 with a B.S. in Biochemistry and a B.A. in Spanish. At the University of Maine, he has been a member of the Sophomore Owls, a participant on the Club Soccer Team, and has worked as a teaching assistant in Organic Chemistry. In the Fall of 2017, he will enter the Tufts School of Medicine Maine Track program and pursue his MD.