Investigating the Unusual Gene Structure of DNA Primase in Mycobacteriophage Subclusters

A Major Qualifying Project Submitted to the Faculty of Worcester Polytechnic Institute in partial fulfillment of the requirements for the Degree in Bachelor of Science by:

_____

Daniel Champlin

_____

Eoin O'Connell

_____

Jennifer Payano

Date: 4/25/19

Approved by:

_____

Professor Elizabeth Ryder

_____

Professor JoAnn Whitefleet-Smith

Table of Contents

1. Introduction……8
2. Background……9
2.1 Structure and Function of Bacteriophages……..…....9
2.2 DNA Primase……..11
2.3 A4 DNA Primase……11
2.4 Proposed Mechanism as Solutions……...12
2.5 One Transcript vs Two Transcripts……13
2.6 Transcriptional and Translational Slippage……...13
2.7 Protein Intein Splicing……...14
2.8 Utilization of Comparative Genomics Tools……17
2.9 Sequence Analysis Tools……18
2.10 Protein Modeling for A4 DNA Primase……18
2.11 Phylogenetic Tools……...19
3. Project Goals……23
4. Methods……24
4.1 Sequence Collection……...24
4.2 Sequence Alignments……24
4.3 Cocoaberry mRNA Secondary Structure Modeling……...26
4.4 Protein Modeling……26
4.5 Phylogenetic Networks……..27
4.6 Bacterial Cultures……..28
4.7 Phage Stock……28
4.8 Titering Cocoaberry……..28
4.9 RT-PCR Primer Design……28
4.10 Growth of phage infected cultures……28
4.11 RNA Extraction……29
4.12 DNA Removal……...29
4.13 cDNA Synthesis……30
4.14 RNA Digestion……..30
4.15 cDNA Cleanup……..30
4.16 PCR……...30
4.17 Gel……..31
5. Results and Discussion……32
5.1 DNA Primase in Actinobacteriophages……...32
5.1.1 Introduction and Nomenclature……32
5.1.2 A-Cluster DNA Primase……33
5.1.3 Actinobacteriophage Database Split DNA Primase……36
5.2 Plausible Locations for Motifs within Cocoaberry……39
5.3 Transcriptional slippery sequence……...41

5.4 Translational slippery sequence……...41
5.4.1 Overall Approach……41
5.4.2 There was no Conservation of Translational Slippery Sequence……...41
5.4.3 Reverse search of Conserved region in the ZBD……..44
5.5 Intein……...47
5.5.1 Oppositely Charge Strands were not found……48
5.5.2 Most Amino Acids were not Located at the precise location needed….50
5.5.3 The Zinc binding domain intein fails to meet the minimum required length……51
5.6 Number of Transcripts……..52
6. Conclusion……55
Appendix A……...60
Appendix B……...61
Appendix C……...62

Acknowledgements

We would like to thank Professor Elizabeth Ryder and Professor JoAnn Whitefleet-Smith for advising and guiding us throughout this project. We would like to thank Professor Buckholt for helping us with phage establishment. Finally, we would like to thank the Shell Lab especially Diego Vargas, and Louis Roberts for their help with different aspects of the RT-PCR.

Abstract

In A4 and other subclusters of mycobacteriophages, DNA primase is a two-domain protein made from two genes with significant overlap that are read in different open reading frames. These genes create a functioning protein through an unknown mechanism. Our investigations using multiple sequence alignments and protein modelling do not support intein splicing, ribosomal frameshifting, or RNA polymerase slippage as mechanisms of forming a functional DNA primase. Preliminary RT-PCR results suggest both domains are transcribed on one transcript.

Table of Figures

Figure 1: Bacteriophage structure……8
Figure 2: Bacteriophage lytic life cycle……...9
Figure 3: General DNA primase structure……10
Figure 4: Genomic layout of split primase feature……11
Figure 5: Intro vs Intein splicing comparison……14
Figure 6: cis- and trans-splicing comparison……14
Figure 7: Overall layout of a conventional split intein……..15
Figure 8: Phylogenetic network of phage subcluster……20
Figure 9: Classification of A-cluster bacteriophage DNA primase……...31
Figure 10: Conserved domains of Cocoaberry primase genes……...33
Figure 11: Conserved domain in A-cluster single gene primase……...33
Figure 12: Pham diversity of ZBD and RPD……34
Figure 13: MEGA alignment of conserved cystine of ZBD……..34
Figure 14: Phylogenetic network of split primase ZBD……...36
Figure 15: Translational slippery candidate gene coordinates……...38
Figure 16: Cocoaberry FS finder results……42
Figure 17: Slippery sequence candidate conservation……...43
Figure 18: mRNA secondary structure 'TTTGGCG'……44
Figure 19: Predicted primase structure from slippery sequence compared to crystalized T7 primase……...45
Figure 20: Overall layout of a conventional split intein……...45
Figure 21: Split inteins candidate gene coordinates……..46
Figure 22: MSA alignment of N-terminal intein charged region……..47
Figure 23: MSA alignment of C-terminal intein charged region……..47
Figure 24: MSA alignment of N-terminal required amino acids……...48
Figure 25: MSA alignment of C-terminal required amino acids……...49
Figure 26: MSA alignment of labeled N-terminal blocks……49
Figure 27: MSA alignment of labeled C-termianl block……...50
Figure 28: Predicted primase structure from split intein sequence compared to crystalized T7 primase……...50
Figure 29: Primer location and overall lengths of expected transcripts……51
Figure 30: Gel results of DNA extracted from RT-PCR……...52

Table of Tables

Table 1: Canonical translational slippery sequences from literature…… 41

1. Introduction

Over the last decade, treating bacterial illness has become a major public health problem because of an increase of antibiotic resistance. The increased use, misuse, and improper disposal of antibiotics has allowed bacteria to develop into various antibiotic resistant strains (Teng et al., 2016). To combat these antibiotic resistant bacteria, microbiologists have looked into potential alternatives. One approach to removing bacteria resistance involves bacteriophages, the most abundant and diverse organism on earth. Bacteriophages, or "phages" for short, are viruses that can infect and lyse specific bacterial hosts in microbial communities. Because phages infect a precise host, they hold a huge advantage over antibiotics. This advantage can be used to target specific pathogenic bacteria while having no effect on the normal microbial community. However, the lack of research on bacteriophages is a major limitation on the use of this novel treatment.

In order to safely use phage as a weapon against bacteria, we need to understand their physiology, and in particular, how they replicate. To replicate their genome and pass it on, DNA primase is required to create short RNA primers to allow further DNA synthesis. It accomplishes this by using two domains to first identify certain sequences with one domain and synthesize the RNA primer with the other domain. Some phage hijack the host machinery to accomplish this, but many contain their own copy of DNA primase in their genome. In the majority of mycobacteriophages, the coding regions for the two domains are next to each other, lack any overlap, and are read in the same frame. However, the A and B subclasses of mycobacteriophage contain a DNA primase that breaks two of these normalities: The domains overlap significantly, and the domains are read in different frames. While gene compression is common in bacteriophage genomes, such a large section of overlap is unusual and leads to questions about the production of a functional gene. This project sought to understand the mechanism of how DNA primase in A4 actinobacteriophage is translated to help broaden our knowledge of bacteriophage before it can be used as an alternative method to antibiotics.

2. Background

2.1 Structure and Function of Bacteriophages

Bacteriophages, like other types of viruses, vary in shapes, sizes, and genetic material, but are primarily composed of a head, a tail, and tail fibers (Poxleitner et al., 2017, Figure 1). Each component plays an essential role in the infection of a host bacterium. The head, otherwise called a capsid, contains the linear DNA or RNA used for DNA replication and protein synthesis. Attached to the head, a tail along with the tail fibers work to bind to specific receptors on the surface of bacterial cells. Although the process of bacteriophage infection is the same for all, bacteriophages have various life cycles that dictate their role in the host bacterium.

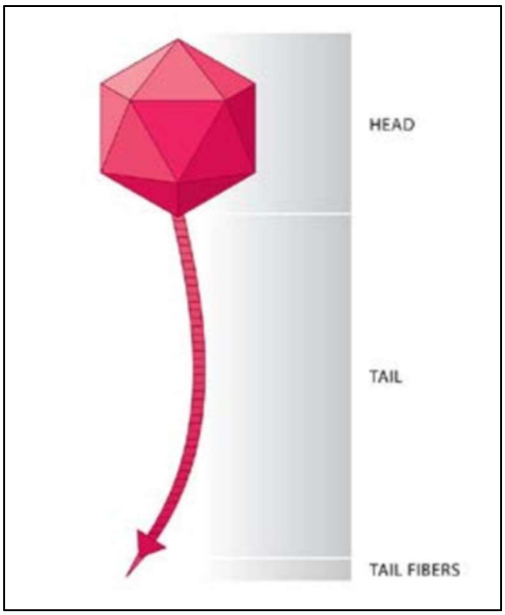


Figure 1: Diagram of bacteriophage major structural components (Poxleitner et al, 2017).

There are two primary life cycle for bacteriophages; lytic life cycle, and lysogenic life cycles. The lytic life cycle occurs when there is a rapid replication of the bacteriophage inside the host which causes the host bacterium to lyse (Poxleitner et al., 2017, Figure 2). While some phages can only reproduce through the lytic cycle, other phages can alternate between the lytic and lysogenic cycle. The decision of a phage to enter the lysogenic cycle is solely controlled through a set of genes and regulatory sequences that are called a "genetic switch". While in the lysogenic life cycle bacteriophages alter the genotype of the bacterium by integrating its DNA into the host DNA. However, the phage's DNA is not expressed immediately, instead the phage genome becomes stably maintained within the cell and remains dormant through many generations (Poxleitner et al., 2017). Under the right conditions, the phage genome within the cell becomes active, triggering the remaining steps of the lytic cycle. Ultimately, both cycles end with a lytic phase that requires the DNA primase to produce primers.

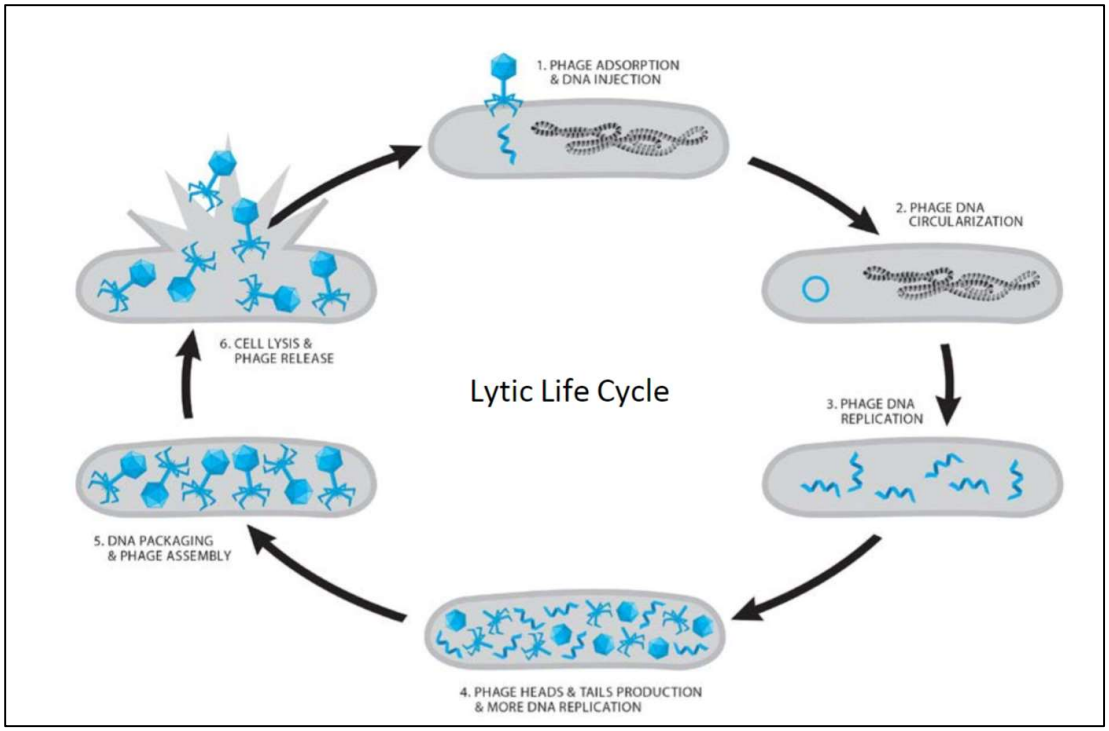


Figure 2: The lytic life cycle of phage. (1) A phage attaches and absorbs to a host cell using its tail fibers and injects its linear chromosome into bacterium. (2) The phage chromosome circularizes inside the host cell. (3) DNA is replicated. (4) Phage heads and tails are produced. (5) New virions are assembled. (6) The bacterial host cell is lysed, and the virions are released (Poxleitner et al. 2017).

There are many stages involved in the reproduction of bacteriophages, but this report will specifically focus on the phage genome replication process. As shown in Figure 2, once the linear phage DNA has entered the cytoplasm, the host's normal synthesis of protein and nucleic acids is disrupted and forced to manufacture the viral phage gene. Within minutes the phage's DNA is replicated. During replication, the double stranded DNA (dsDNA) is unwound by the helicase enzyme, which breaks the hydrogen bonds holding the complementary bases of the DNA together thus creating a replication fork. The two separated strands will then act as templates for making the new strands of DNA. One of the strands is called the leading strand, which is oriented in the 3' to 5' direction, towards the replication fork. While, the other strand is called the lagging strand, which is oriented in the 5' to 3' direction, away from the replication fork. Because of their orientation, the two strands are replicated differently. DNA replication begins when a short piece of RNA called a primer, which is produced by an enzyme called DNA primase, binds to the end of the end of the leading strand and to various point along the lagging strand. The primer serves as the starting point for DNA replication.

2.2 DNA Primase

The replication of the double helix DNA molecule is a complex process requiring numerous enzymes. DNA primase is an enzyme that catalyzes the synthesis of short RNA nucleotide primers complementary to the DNA template strand. The nucleotide primers are used to initiate the process of DNA replication. DNA primase is made up of two or three conserved domains; Zinc-binding, RNA polymerase domain and potentially a helicase domain, such as in bacteriophage T7 (Figure 3, Lee et al., 2010). The zinc binding domain, located in the N-terminal region of the protein, functions to recognize specific sequences in the DNA. The RNA polymerase domain, located in the C terminal, works to synthesize RNA nucleotides for the formation of a four-nucleotide primer. Both domains are connected by a linker, which is preserved in length and sequence (Figure 3, Lee et al., 2010).

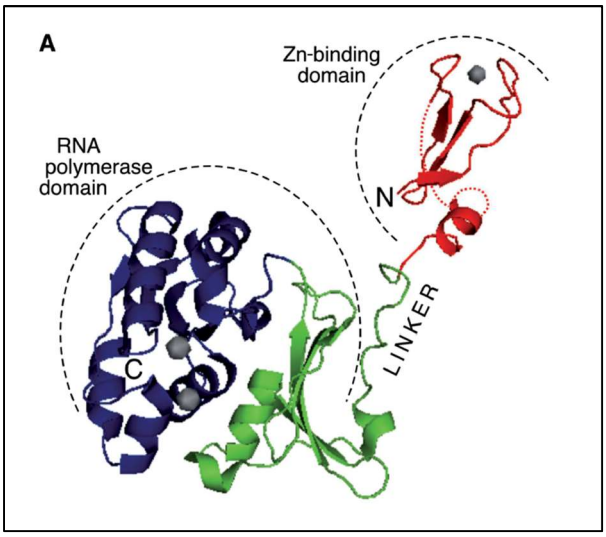


Figure 3: Tertiary structure of phage T7 DNA primase with labeled structural domains.

The linker, although small, has been found to play a vital role in the function and to be fairly conserved in length between bacteriophage and bacterial DNA primases (Lee et al., 2012). The linker works as a hinge for the protein, swinging the RNA polymerase domain into the position once the zinc binding domain binds to a specific sequence so a primer can be synthesized properly. An example of the importance of the linker is seen in T7 phage, where two proteins are translated, one containing all three domains, and one lacking the zinc binding domain. When the protein lacks the zinc binding domain, the RNA polymerase domain cannot produce tetranucleotide primers that are needed for transcription to occur. If the primer is less than four nucleotides long, polymerase isn't able to work properly (Lee et al., 2010). This shows the importance of the linker between the zinc binding domain and RNA polymerase domain. When the linker is shortened, lengthened or changed significantly, this function is lost, and the protein loses its function due to not properly being able to find specific sequences and produce primers long enough (Lee et al., 2012).

2.3 A4 DNA Primase

The A4 class mycobacteriophage DNA primase has similarities to the DNA primase in T7 phages. There are two protein coding genes which contain conserved domains for both a catalytic subunit (RPD) and a zinc binding subunit (ZBD). The A4 helicase domain is contained in a gene downstream of the DNA primase domains and is contained in its own protein, unlike the T7 three domain primase. Since a crystalized structure of the A4 DNA primase is unavailable, it is unknown whether or not a linker region between the components is conserved,

but it is likely to be conserved based on the necessary geometric flexibility for a functional DNA primase. The uncommon characteristic of the A4 primase has to do with the amount of overlap between the genes for its functional domains. Viral genomes have evolved mechanisms to streamline the replication process by shortening the space between genes by a process known as gene compression. Normally, this involves overlapping open reading frames by a 1-3 bp margin in a viral genome to decrease its overall length, with a maximum threshold of overlap occurring at 30 bp. It is believed that minor sections of overlap are primarily influenced by a finite capsid size, but regulation of gene expression and decreased replication time are additional positively selected factors (Chirico et al., 2010). The process of gene compression also allows for a greater degree of horizontal gene transfer between phages in the same geographic area because of flexibility in genome structure. However, sections of significant overlap tend to increase the effect of negative mutations, since they can impact more coding regions of the genes that overlap, affecting the structure of multiple gene products.

The A4 subclass of actinobacteriophage DNA primase violates many of the typical standards seen in phage genomes. With an overlap of 185 base pairs (bp) between two protein domains, the A4 bacteriophages greatly exceed normal thresholds of overlap (Jacob-Sera et al., 2011) (Figure 4). This isn't the only novelty encountered; upon closer inspection, the first domain is translated in one frame while the other is translated in a different frame, complicating how the gene is translated and made into a protein product.

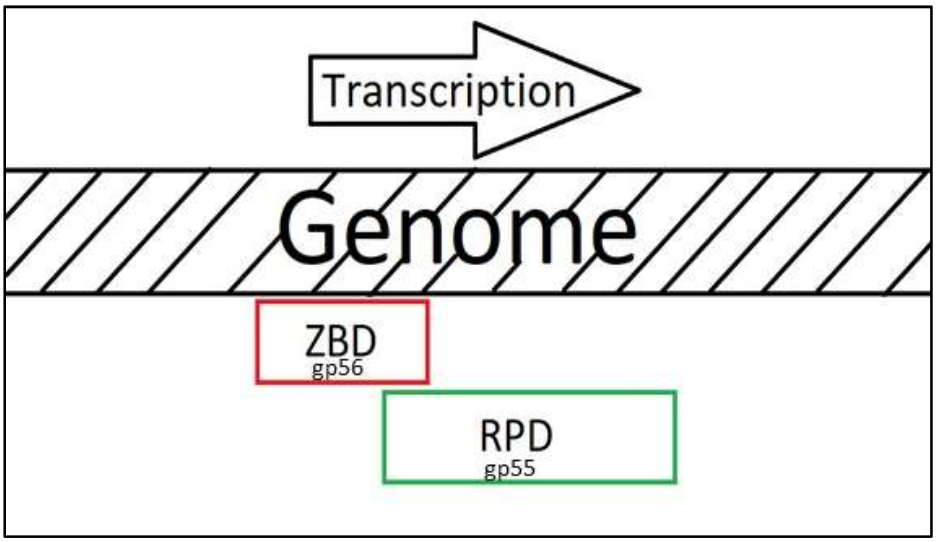


Figure 4: A basic drawing of the overlap of the two genes of DNA primase in mycobacteriophage. The overlap in Cocoaberry is 185 bp.ZBD, zinc-binding domain; RPD, RNA polymerase domain; gp56, gp55 gene product numbers in Phamerator database

2.4 Proposed Mechanisms as Solutions

To answer the question of how a functional DNA Primase is made while overcoming the overlapping region and reading frame problems, we have developed three hypotheses: use of a transcriptional slippery sequence, use of a translational slippery sequence, and use of intein splicing. Each of these hypotheses will be explained further, as they require specific sets of

circumstances to be viable solutions. In addition to the specific circumstances needed for each method, the viability of our proposed solutions can be tested by figuring out whether the two genes that compose DNA primase for Cocoaberry are transcribed on the same transcript or on different ones.

2.5 One Transcript vs Two Transcripts

There are several ways that the protein could be made from mRNA into a functional protein. The mRNA that is transcribed to cover this area could be two shorter transcripts with each transcript containing one of the two coding regions or one long transcript that has both coding regions on the transcript. One thing to note with this is that the genome of subclass A phage, among other phage, seem to be transcribed in two major sections, typically the front half of the genome is transcribed in the forward direction and the other half is transcribed in the reverse direction (Halleran et al., 2015). However, looking through *Transcriptomic Characterization of an Infection of Mycobacterium smegmatis by the Cluster A4 Mycobacteriophage Kampy*, there are two parts of the genome that appear to be transcribed at greater rates than the rest of the arms (Halleran et al., 2015). Upon further inspection of these areas, they contain proteins for DNA replication in the first half and the proteins required for protein coat assembly in the second half. This suggests that there might be more than the two main transcripts that is typical convention for phage transcripts, one transcript for the forward direction genes and one for the reverse direction genes. The evidence presented by Halleran, Clamons, and Saha for the presence of multiple transcripts, as opposed to two main transcripts as dictated by convention, provides us with an interesting dilemma for hypotheses on formation of DNA primase. In the former case, the ZBD and RPD genes could be on two different transcripts. This would rule out use of a slippery sequence, which requires a single transcript. In the latter case, both the ZBD and RPD genes would be on the same transcript, which would still allow for use of a slippery sequence. Even though the intein splicing method could occur in the one transcript or two transcript scenarios, it would be useful to determine which of the two mechanisms is used by Cocoaberry since it could possibly rule out two of our hypotheses. This idea will be explored in later experiments attempting to answer the one transcript vs two transcript question.

2.6 Transcriptional and Translational Slippage

Slippery sequences have been established as a way for protein synthesis to occur using a single transcript to produce a protein that is read in two different frames (Xu et al., 2004). A translational slippery sequence is a set of nucleotides, most commonly found in the form YYYXXXZ (Y, X and Z are nucleotides, Y and Z can be the same) that cause the ribosome to slip backwards or forwards when reading the mRNA, causing a change of frame. A common example of translational slippery sequence is in the assembly of phage tail proteins in ƛ phage. The tail assembly coding regions are the same in ƛ as the DNA primase regions are in A4 mycobacteriophage in the way that they contain an overlap of two regions that need to be read in

different frames, like in Figure 4. What happens in ƛ is that the ribosome slips back one nucleotide near the c terminus of the first domain, causing a frameshift that skips the first stop codon. (Levin et al., 1993). The slip doesn't happen often, only occurring 4% of the time. However, when the slip does occur, this results in a functional protein. The exact mechanism that causes the ribosome to slip is unknown.

There is also the potential for transcriptional slippery sequence, where when reading a poly A or T or a CCCUUUUUUUU sequence, the polymerase slips causing a change in frame, either forwards or backwards, by a select number of bases (Atkins et al 2016) (Baranov et al 2005). This slippage occurs in a wide variety of organisms but occurs in a low percentage of transcripts with many effects (Baranov et al 2005). Slippery sequence at either the transcriptional or translational level are a potential candidate because it fits the basic criteria of allowing overlapping frames that need to be read in different frames to be made into one, whole amino acid chain.

2.7 Protein intein splicing

Protein intein splicing is another well-known and studied potential mechanism that could be used to explain the creation of the overlapping protein domains in DNA primase of A4 mycobacteriophage. Protein splicing is the excision of an intervening sequence (intein) out of a large polypeptide chain that, in the process, ligates together the two flanking protein fragments producing a single mature host protein (Shah & Muir, 2014). Inteins can be thought of as a 'protein intron'. Introns are the intervening sequences that get spliced out of the RNA before the mature RNA (mRNA) is translated into a protein, as can be seen in Figure 5. However, in protein splicing the intein is present in both the mRNA and in the translated protein precursor. Once in the translated protein precursor, the intein gets spliced out, making a mature spliced protein (Figure 5).

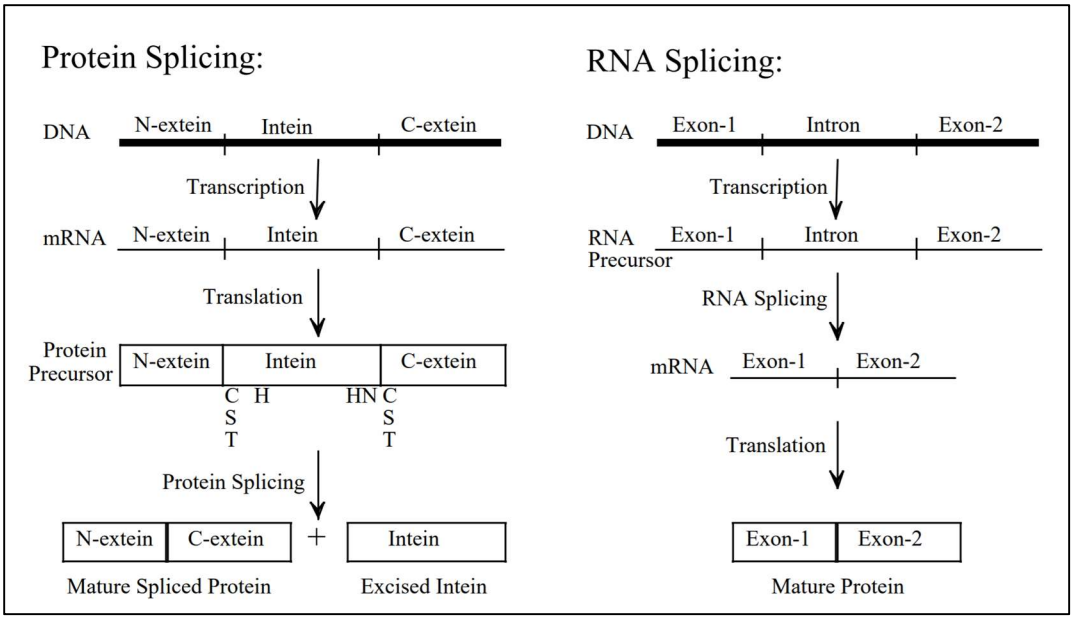


Figure 5: Comparison of RNA Splicing and Protein splicing (Perler, 2002).

The two different ways in which protein splicing can work are protein cis-splicing (standard inteins) and protein trans-splicing (split inteins) (Figure 6). During the standard protein cis-splicing, one single gene is translated and transcribed into one protein precursor which is used to excise the intein (Figure 6a). But, unlike the standard inteins, split inteins are transcribed and translated as two separate polypeptides and joined together by salt bridges in oppositely charged strands (Figure 6b). This process forms a single precursor protein from two transcripts instead of one. Once in the protein precursor, both cis- and trans-splicing use an identical mechanism to do a multi-step process. This process requires the help of certain amino acids for the detachment of the inteins and the ligation the two flanking exteins (Figure 6) (Shah and Muir, 2014). Therefore, the only method of intein splicing that could be used to explain why two translational frames drastically overlap in the DNA primase of A4 bacteriophages is protein trans-splicing.

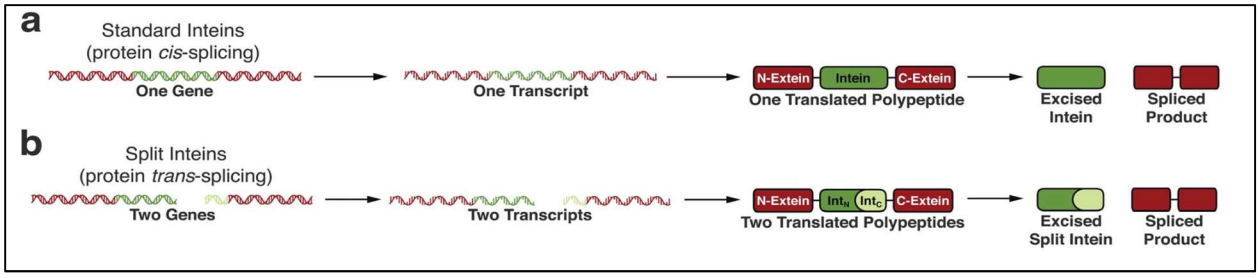


Figure 6: Comparison of protein cis-splicing and protein trans-splicing (Shah and Muir, 2014).

In order for split intein trans-splicing to be the mechanism occurring for the DNA primase in A4 bacteriophages, there are three criteria that must be met. First, the charged strand of the C-terminal end of the N-terminal intein must be negative and the N-terminal end of the C-terminal intein must be positive or vice versa (Shah et al., 2011). These oppositely charged

strands facilitate the binding of the inteins. Second, certain amino acids must be present at a precise location within the N-terminal and C-terminal inteins. These amino acids are responsible for the cleavage of the inteins and ligation of the exteins. Third, the length of the excised split intein should be close to the minimal amount of amino acids required. This length will prevent the splicing of functional domains within the ZBD and RPD by this mechanism while also leaving enough room for a functional linker and protein.

The ionic interactions of the oppositely charged regions in each half of the intein facilitates the joining of the split intein. These charged regions must have a negative charge C'end of the N-terminal intein and a positive charge at the N'end of the C-terminal intein or vice versa, for splicing to occur (Dassa et al., 2007). For this to work, the two oppositely charged strands must be sufficiently close to one another to be able to form a salt bridge between the two protein (Dassa et al., 2007). As suggested by the data presented by Shah and Muir, without these ionic interactions of the opposite charged strands there would not be any intein splicing activity for split inteins.

Two different nomenclatures are used to describe the position of conserved amino acids within the split intein. The first nomenclature uses motifs to demonstrate the amino acids position within the N-terminal intein and C-terminal intein. There are four conserved motifs that make up the N-terminal intein. These motifs are presented as N1, N2, N3, and N4 (see Figure 7) (Pietrokovski, 1994). Likewise, the C-terminal intein is composed of two motifs, C1 and C2 (Figure 7). However, the second nomenclature uses blocks and motifs to illustrate where the conserved amino acid sequences are positioned. The splicing domains consist of block A, motif N2, block B, block F, and block G (see Figure 7) (Perler, 2002). Figure 7, shows how these two nomenclatures match up. For the purpose of this paper, the first nomenclature will be used to describe the positions of consensus sequences.

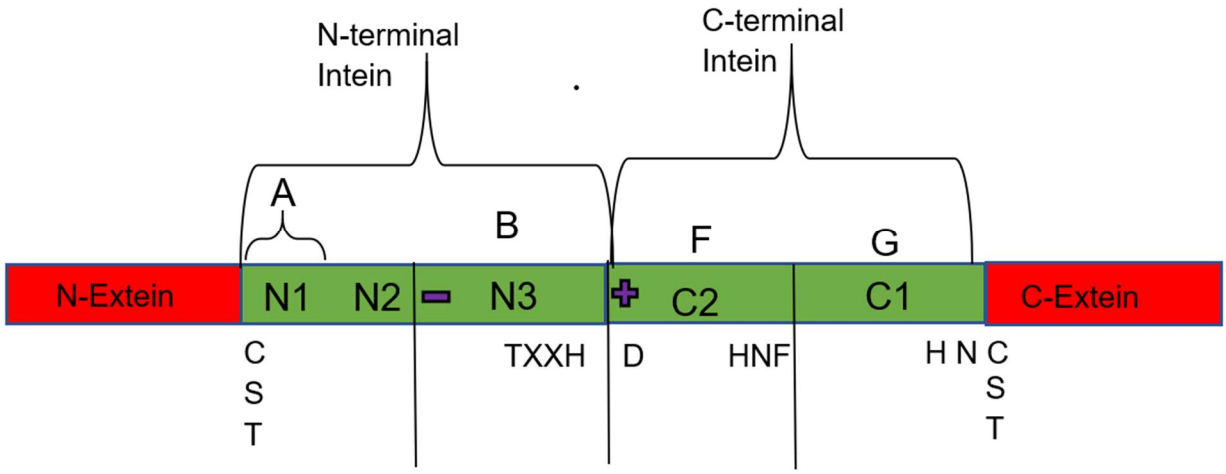


Figure 7: Position of amino acids required for the split intein mechanism.

Some amino acids sequences are conserved within motifs. The conservation of these amino acids is used to determine consensus sequences found for protein splicing. These consensus sequences are responsible for the splicing of the split intein and the joining of the flanking extein. To initiate the process of protein splicing, the junction points located at the N' end of motif N1 and at the C' end of motif C1 requires a cysteine, serine, or threonine (Wang et al., 2018) (Figure 7). These amino acids coupled with a highly conserved asparagine (N) in motif N2 are responsible for the autocatalytic reactions needed for the cleavage of the intein and ligation of the extein. This reaction, although necessary, is not sufficient for the split intein mechanism. For this mechanism to work, additional amino acids are needed to help facilitate the reaction. The threonine and histidine found at the C' end of motif N3 help with the cleavage of the N-terminal intein from the N-extein (Figure 7) (Shah & Mui, 2014). Similarly, the N3 histidine (H) is highly conserved along with the histidine found in the motif C1, asparagine (N), and phenylalanine (F) facilitate the excision of the intein from the extein complex at the C' end of the C-terminal intein (Figure 7). This interaction leaves the exteins attached through a non-peptidic linkage (Shah & Mui, 2014). After this process is completed, the structure rearranges into a favorable and stable peptide bond thus forming a single protein from two proteins.

Once the spliced products are united and the split inteins is excised out of the proteins, the length of the excised split intein should be close to the minimum amount of amino acid required for this mechanism to occur. The N-terminal intein contains four motifs with a span of 100-150 amino acids. The last motif, N4, is the longest and most conserved motif within the N-terminal intein (Pietrokovski, 1997). But despite the length and conservation, it is not as essential as the others. In fact, the last motif could not be identified in three out of four of the eukaryotic inteins, while the rest of the motifs have been found in all inteins (Pietrokovski, 1997). Therefore, if the N4 motif was not present in inteins the length required for the N-terminal intein to encompass the remaining 3 motifs should range between 60-64 aa (Pietrokovski, 1997, see table 1 in Figure 2). Likewise, the C-terminal intein that includes the two required motifs, C1 and C2, has a combined length ranging from 25-40 aa (Pietrokovski, 1997).

2.8 Utilization of Comparative Genomics Tools

Recognizable patterns, conserved DNA or protein sequences, tertiary structure comparisons, and evolutionary history can be explored using a selection of bioinformatics tools and could help narrow the focus of hypotheses about the creation of a functional DNA primase in A4 bacteriophages. The bioinformatics tools we will utilize can be divided into categories of sequence analysis, protein modelling, and phylogenetics. They will be primarily responsible for answering the questions about the origin of the 185 bp overlapping region in A class phage DNA primase genes. They will also be used to formulate new hypotheses and models based on in vitro experimental results.

2.9 Sequence Analysis Tools

The majority of sequence data used in this study comes from the bacteriophage data stored in Phamerator and PhagesDB, which are both tools affiliated with the HHMI Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science program, or SEA-PHAGES. SEA-PHAGES is an educational program designed to give high school and undergraduate college students laboratory experience by isolating DNA from a bacteriophage and annotating the phage's genome (Jordan et al., 2013). The amount of data in Phamerator and PhagesDB has increased dramatically due to the crowdsourcing of data collection to university students, and the current database includes data from 151 institutions worldwide. In order to maintain the annotation standards for GenBank, sequence annotations are proofread by annotation experts who have looked at phage genomes before. They reformat and correct any incorrect gene calls before submitting the final sequences to GenBank. Because annotation guidelines prohibit regions of significant overlap between genes, draft sequences usually do not annotate the overlapping DNA primase region, even though verified sequences contain the overlap in their final annotations. For that reason, only non-draft sequences that have been verified by expert reviewers will be used in multiple sequence alignments.

Tools that can align phage DNA primase DNA or residue sequences in multiple sequence alignments (MSAs) will be crucial in identifying structural gene patterns and conserved residues. NCBI's MUSCLE MSA tool and COBALT MSA tool are frequently used for protein sequence alignment and recognition of conserved sequence regions respectively. NCBI BLAST will be used to search for similar sequences and to verify sequence similarity of two or more aligned fasta sequences. Since the overlapping of open reading frames in the DNA Primase genes is not limited to the A4 or A cluster phages, the diversity of phage evolution will make protein alignments useful for characterizing the feature. Comparison of DNA sequences after protein alignments in MEGA (Molecular Evolutionary Genetics Analysis) can show whether or not conserved motifs for inteins or slippery sequences are present.

2.10 Protein Modelling for A4 DNA Primase

Protein modelling tools will be useful for this study in formulating hypotheses about the structure and function of the functional DNA primase. Distantly related proteins in sequence between organisms can have similar functions or even geometric similarities depending on the function of the protein and its complexity (Koonin et. al, 2002). PhagesDB contains groups of phages collected from a variety of bacterial hosts, and even between DNA Primases that have the characteristics similar to primases of the A4 subcluster, there may be a divergence in sequence similarity but less of a divergence in geometry. Since crystallography and protein isolation are outside the scope of this study, protein modelling can provide a predicted model of our unusual DNA primase.

There are three types of modelling that can be used depending on the percentage of identity that a protein sequence has to crystalized structures in PDB: homology modeling, protein threading, and de novo modeling. Based on thresholds of similarity between a template amino

acid sequence and a query sequence, homology modeling is recommended if the template has between 30% and 50% sequence identity or more; protein threading and de novo modelling are recommended if the template has less than 30% sequence identity (Baker & Sali, 2001). While more accurate crystalized models of proteins and high sequence identity homology models can be used to visualize docking of ligands or improving of ligands and active sites, protein threading models and de novo models can be used to show functional relationships between structures of proteins (Baker & Sali, 2001). Homology modeling involves four steps: template selection, model generation, side chain generation, and energy minimization, in which the template acts as a baseline in calculations done by the algorithms responsible for these steps (Schwede et al., 2003). Protein threading uses algorithms to find models in PDB that have patterns of supersecondary structure folds similar to those in the query sequence, and builds a model using the same steps of template selection, model generation, side chain generation, and energy minimization to create a model (Jones et al., 1992). Tools such as MUSTER access PDB in order to obtain patterns of folds based on alignments of a query sequence to the sequences of models (Wu & Zhang, 2008). De novo ("from the beginning") modeling tools are often criticized for lack of accuracy, but annual competitions such as CASP are allowing the method to gain more accuracy in its predictions when paired with threading or homology modeling.

Tools like MODELLER and RaptorX use homology modelling to predict models of protein tertiary structure, while tools such as I-TASSER can be used to predict models using protein threading. Our choice of modelling software will rely on percentage of query cover to the protein sequence of the model and the percentage of identities between the two sequences, as low thresholds of either sequence identity or query coverage will make modelling less predictable (Chakravarty et al., 2008). MODELLER uses database connections to PDB to compare primary protein sequence inputs to models in order to perform homology modelling, where those models with a high sequence identity to the query and long query coverage will be used as a template for modelling (Webb & Sali, 2016). I-TASSER uses a combination of supersecondary structure comparisons, or fold recognition, to act as partial templates for modelling the query sequence and de novo modelling to model those sections of the query sequence without matches to fold patterns (Zhang, 2008). The best models of the functional DNA primase can be compared to other DNA primase molecules in geometric alignments using UCSF Chimera to hypothesize whether or not function would be preserved (Pettersen, et al., 2004).

2.11 Phylogenetics Tools

In order to investigate patterns in genes found by annotation and multiple sequence alignment tools, we plan on developing phylogenetic trees and networks of the A4 DNA primase and of related DNA primase subunits of phages and more distant homologs collected via PSI-BLAST from NCBI. Phylogenetic trees are useful tools for understanding the evolutionary relatedness of organisms and entities. They can be modified to show time between evolution of specific features and can be used to group species into higher classifications based on their attributes. Phylogenetic networks are commonly used to measure conservation of features within

genes while accounting for horizontal gene transfer, insertions, deletions, and multiple paths of ancestry based on conservation for different regions of the sequence.

There is a wide selection of tools that can be used to create phylogenetic networks based on different graphs and predetermined tree structures or alignments. For this study, we will be utilizing MEGA as a tool for protein sequence alignments and unrooted maximum likelihood or maximum parsimony trees. The unrooted character tree results are being used because there is no indication that the sequences involved will have a representative common ancestor, and because the trees will be more representative of the evolutionary history of phages that originate from different taxa (Hall, 2018). These results will be compared to a statistical parsimony phylogenetic network created by SplitsTree that will use hybrid nodes and minimum distances calculated between nodes to represent a complex network with horizontal gene transfer. This will try to minimize the number of assumptions between sequence comparison distances while allowing for multiple parents to each node to not exclude relationships between phage DNA primase sequences. This is the best method for generating a phylogenetic network since the evolution of phages is non-linear and can include insertions and deletions from other related phages.

Phylogenetic tools will be utilized in this study to look for evidence of evolutionary conservation between the A cluster actinobacteriophage DNA primase ZBD and RPD to other related actinobacteriophages or species. Networks will also be analyzed to look for evidence supporting a mechanism for creating functional DNA primase in the A4 phages.

Phage genetic mosaicism, a characteristic of phage genomes defined by high amounts of gene loss and horizontal gene transfer, makes the choice of which sequences to include outside of normally divided clusters harder when creating a phylogenetic network (Pope et al., 2015). SEA-PHAGES divides their annotated bacteriophages into classes and subclasses based on sequence similarity, and overall gene pattern similarity (Figure 8). Class A phages were separated from other actinobacteriophages because of their overall size and transcriptional organization, while other classes were separated due to a large number of tRNA genes, characteristic repeated sequences, or unusual capsid geometries (Pope et al., 2015). Metrics such as CLASP (The Cluster Averaged Shared Phamilies), CAP (Cluster Associated Phamilies), and CCI (Cluster Cohesion Index) were used to separate phage genomes into clusters. For the A cluster, the CCI values were low compared to other clusters, meaning it had a high level of diversity, but the A cluster did not share many of its phamilies with other clusters (Pope et al., 2015). The same study showed degrees of genetic mosaicism between phages within classes based on a whole genome alignment, but that individual genes can evolve independently of the clusters to which they belong, since subclusters of phages have access to different gene pools based on their geographic locations (Figure 8) (Pope et al., 2015). The categorization of genes into specific phams, genes with high amino acid sequence similarities grouped together, based on sequence and function similarity also allowed for a greater amount of classification, since phams may be shared between many classes of phages due to horizontal gene transfer.

Trees will be built to explore the possibility of horizontal gene transfer causing the overlapping region to form from a different phage by comparing geographic data and seqence alignments. Clusters have been related between different species of enterobacteriophages showing that it may be possible to track gene evolution between various clusters, even though horizontal gene transfer is much more common within clusters than between them (Grose, & Casjens, 2014). Trees will also be built staying within phage groups with related phams or similar species, as some gene co-occurrence networks have shown a correlation between gene presence and specific phage hosts (Shapiro, & Putonti, 2018).

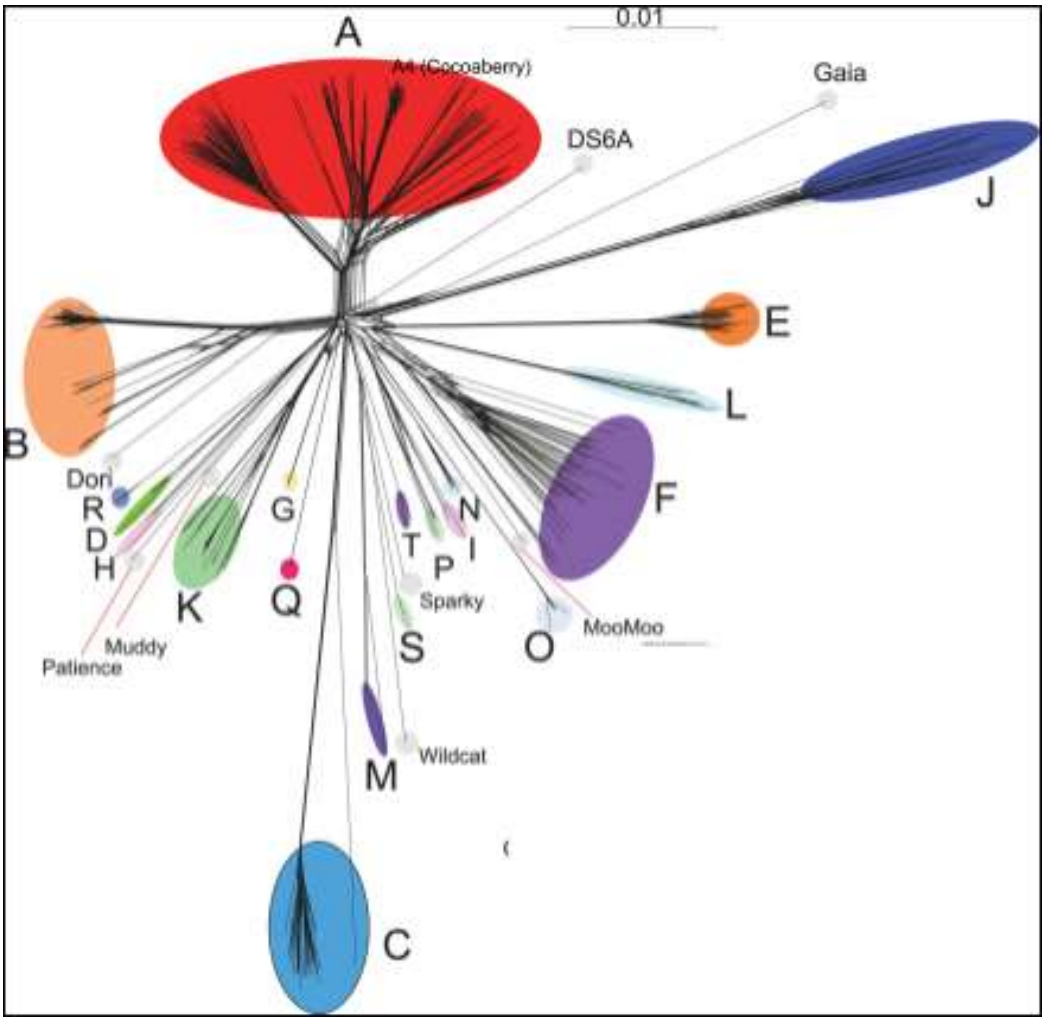


Figure 8: Phylogenetic network of 627 mycobacterium bacteriophages created using SplitsTree (Pope, W. H. et al, 2015). Letters correspond to the cluster designations for groups of phages, which are circled in colors. Singletons are shown using their phage names, and Cocoaberry is marked in the section with A4 phages.

The hope with phylogenetics is to understand the changes that occurred in the distant relatives of A4 mycobacteriophages that lead to the modern mechanism that is being used to translate DNA primase as there is no apparent advantage to having such a complex mechanism. We hope to better characterize both the evolution of the split primase compared to the sequence and structure diversity of DNA Primase genes in the actinobacteriophage database. Phage have

slowly adapted to use the current novel mechanism over time. By understanding the steps that were taken throughout evolution, this might shed some light onto how the current mechanism works.

3. Project Goals

This project attempted to answer several questions about the unique yet problematic split primase feature in the DNA primse of Cocoaberry. Computational modelling and sequence alignments were used to find conserved sequences in split primases to determine if our hypotheses of transcriptional slippage, translational slippage, or intein splicing explained the mechanism for creating DNA primase. RT-PCR protocols were followed to identify if Cocoaberry used one or two transcripts to produce DNA primase. Data mining and creation of phylogenetic networks was accomplished to try and understand the origin of the split primase feature and whether or not other bacteriophages also used this mechanism. We hoped that by better characterizing the split primase identified in A4 mycobacteriophages, we could begin to understand how it forms a functional protein.

.

4. Methods

4.1 Sequence Collection

DNA Master version 5.0.2 was downloaded from (http://cobamide2.bio.pitt.edu/) under the DNA Master link opening in a new tab. It is a 32-bit program that is compatible with versions of Windows XP and later. The .dnam5 file for viewing Cocoaberry in DNA Master can be found on PhagesDB at (https://phagesdb.org/phages/Cocoaberry/) under the "Final DNAMaster File" section. Any phages that needed to be viewed in DNA Master had to be formatted as a .dnam5 file. Those phage .dnam5 files were collected by searching by the phage name at (https://phagesdb.org) and downloading the file from the "Final DNAMaster File" section.

To collect the sequences necessary for multiple sequence alignments, the online tool Phamerator was used (phamerator.org). Two categories of files were created for the A4 subcluster and the A cluster respectively, as the A cluster phages were our initial focus for split primases in bacteriophages in PhagesDB. Later on after a thorough analysis of the sub clusters in the Actinobacteriophage database, a file was created with split primases in the A, B, C clusters phages where a significant overlapping region between their DNA primase genes was found. For both files, only non-draft phages were used, as the draft phages have not been reviewed and submitted to GenBank. In Phamerator, draft phages can be identified by the "_Draft" added to the end of their names in the Select Phages screen. For the A4 fasta files, all A4 non-draft phages were first selected on the Select Phages screen. Since not all of the phages loaded on the web server due to the high volume of selected phages in the map, two columns were loaded at a time into the map, and the final file was checked to ensure all A4 phages were collected. The ZBD DNA sequences were collected by left-clicking on the feature number that corresponded to the ZBD for each phage, then copying and pasting the FASTA formatted DNA sequence into "A4ZBDAllNonDraft.fasta". The RPD DNA sequences were collected the same way and were stored in "A4RPDAllNonDraft.fasta".

The A cluster fasta files were created on September 27, 2018 using the same criteria as the A4 fasta sequences and the same methods of collection from Phamerator. Instead of collecting all phages from all clusters, a representative selection was collected from all 20 A subclusters counting at most 5 phages per subcluster, depending on how many non-draft phages were available for each subcluster. Subcluster A20 was not included, because there were no non-draft phages in the subcluster. Representatives were chosen because they were the most recently submitted non-draft phages as of October 8th, 2018. The dates that each phage was sequenced can be found by going to (https://phagesdb.org/phages/), filtering by subcluster in the dropdown menu, and searching by subcluster. A cluster ZBD DNA sequences were stored in "AClassZBD.fasta" and A cluster RPD DNA sequences were stored in "AClassRPD.fasta".

Protein sequences for the A cluster and A4 cluster fasta files were created by copying the contents of the DNA files and translating them using an ExPASy translation tool (https://web.expasy.org/translate/). The A cluster protein sequences are stored in "AClassZBDProtein.fasta" and "AClassRPDProtein.fasta" for the ZBD and RPD sequences

respectively. All of the multiple sequence alignments for A clusters and A4 subcluster DNA and protein sequences were done using these files.

To create the files necessary for the intein sequence comparison, an online website called InBase (http://www.biocenter.helsinki.fi/bi/iwai/InBase/tools.neb.com/inbase/index.html) was used. Two categories of files were created for the inteins found in the database, on January 10,2019. The first file, "Inteins.fasta" collected all the sequences given by the InBase database. The second file, "SplitInteinsv2.fasta", only contained inteins that were categorized as split inteins. Within InBase, these split inteins were separated by the name id. Inteins that were split would have two entries, '.n' and '.c'. This indicated that '.n' was the N-terminal intein, while the '.c' was the C-terminal intein. Some sequences like "Neq_Pol-n", "Neq_Pol-c", and "Hut_MCM-2" were removed from the fasta file because their short length prevented the cobalt tool from aligning the conserved of amino acids of other inteins. A new file called "SplitInteinsw.oShortsqw.Cocoaberry.fasta" was created on January 25th, 2019. This file included the given intiens without the short sequences from InBase, along with the annotated Actionobacteriophage Cocoaberry. A full README file in the supplementary information is provided describing the ".fasta" files collected, and what alignments they were used to create.

4.2 Sequence Alignments

A multiple sequence alignment for both the Zinc binding domain (ZBD) and the RNA polymerase domain (RPD) was done through a web portal called NCBI COBALT (https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi). COBALT is a tool that can be used to align multiple protein sequences, construct molecular phylogenies, and identify functionally important (evolutionarily conserved) sites. For our project, the FASTA files "AClassRPDProtein.fasta" and "AClassZBDProtein.fasta" were attached in separate tabs of COBALT. Before aligning, all advanced parameters were kept in their default setting. Once the multiple alignment finished, the conservation setting further down the page was changed from 2 Bits to 3 Bits, this was done to increase the threshold for determining which were colored in red. The red color indicated highly conserved columns, while the blue indicated less conserved ones. The Bit coloring in this section of the results page specified conserved regions of the alignment by residue to identify motifs commonly associated with intein splicing and slippery sequences.

To do a multiple alignment of the both the inteins from InBase and Cocoaberry, the web portal NCBI was used (https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi). FASTA file "SplitInteinsw.oShortsqw.Cocoaberry.fasta" was attached. Before aligning, all advanced parameters were kept the in their default setting. Once the multiple alignment was finished, to get more of an accurate comparison of the inteins collected from InBase to Cocoaberry's intein the colors were changed to rasmol amino acid coloring through the tools setting. Rasmol amino acid coloring is a scheme of colors for amino acid that follows the traditional amino acid properties.

MSAs with nucleotide sequences of A cluster and A4 subcluster phages were done using the EMBL-EBI MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/), as well as the MUSCLE algorithm extension in MEGA software for creating phylogenetic trees.

The primary tool used for multiple sequence alignments and identification of conserved sequences was the Windows 64-bit version of MEGA (Molecular Evolutionary Genetics Analysis) downloaded from (https://megasoftware.net/). After DNA sequences were collected from Phamerator and placed into fasta files prepared for alignments, MEGA was opened and new DNA alignment files were created by clicking Align>Edit/Build Alignment>Create New Alignment. Sequences copied and pasted into the new file, and aligned by clicking Alignment>Align by MUSCLE (Codons). All sequences were selected into each alignment, and default parameters were used including a gap penalty of -2.9 and a UPGMA clustering method. The views of the resulting alignments could be adjusted by percent conservation of residues at specific sites by clicking Display>Toggle Conserved Sites and selecting the desired percentage conservation. Random samplings of sequences from A subclusters attempted to prevent biases in conservation percentages, but lack of representation from smaller subclusters in the A and B clusters should be noted.

4.3 Cocoaberry mRNA Secondary Structure Modelling

In the event that supporting evidence for a translational slippery sequence was found, additional supporting figures for the secondary structure of the mRNA molecule from Cocoaberry ZBD gene were created to identify the presence of additional requirements for this hypothesis. The secondary structure prediction could be analyzed to identify both a spacer region and a secondary loop or pseudoknot downstream of any canonical frameshift sequences, both of which are required for a translational frameshift (Napthine et al, 2017). The program RNAstructure was used from the Matthew's Lab and follows a four step process for predicting RNA secondary structure from a given sequence. A partition function first calculates base pair probabilities between nucleotides of a given sequence, then a Fold algorithm predicts a lowest free-energy structure based on the sequence. MaxExpect modifies the structure based on the most likely base pairings, and ProbKnot predicts any possible pseudoknots or stem-loops before modifying the model accordingly (Matthews et al., 2004). Cocoaberry ZBD RNA secondary structures were calculated using the DNA sequence collected from Phamerator and saved in fasta format. This file was imported to the web server for RNAstructure at (https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html) and default parameters were not modified before running the software.

4.4 Protein Modelling

Protein modelling was done using sequences from bacteriophages that were modified based on sections in conserved areas containing motifs that supported our hypotheses. The two primary tools used for modelling were MODELLER v9.21 for homology modelling, and I-TASSER v5.1 for protein threading. MODELLER can be downloaded with an educational

license from (https://salilab.org/modeller/download_installation.html), and I-TASSER can be downloaded with an educational license from (https://zhanglab.ccmb.med.umich.edu/I-TASSER/download/). MODELLER input requires a primary protein sequence and there needs to be a known template in the Protein Database (PDB) with a high threshold of sequence identity and query cover. MODELLER is a Windows compatible program that can be run on a local machine, and the resulting models are ranked based on best matches to the templates in PDB.

Primary protein sequences with query coverage and sequence identity below the threshold to known models in PDB were modelled using I-TASSER. I-TASSER input requires the primary protein sequence and jobs are run through both the web server and on a local server for expedited service. Results from I-TASSER jobs through both the web server and local installation produced reports with both alignments to sequences in PDB that represent supersecondary protein structure matches based on sequence for areas that have strong matches, and de novo prediction sites. Faster cluster computing of I-TASSER simulations was done using the WPI Research Cluster with modifications made to "runI-TASSER.pl" to allow it to run on a Slurm based job scheduler versus a PBS job scheduler. Model PDB files are generated to be viewed through the web server, and to be accessed through model viewing software. I-TASSER analyses returned predicted functionality of the final tertiary structure based on gene ontology terms, enzyme classification numbers, and identified ligand binding sites.

Models produced by either MODELLER or I-TASSER were viewed using UCSF Chimera v1.11. Saved .pdb files of the models were opened in Chimera and can be viewed and manipulated. For geometric alignments to known DNA primases, model PDB IDs were taken from RSCB PDB for DNA primases of several bacterial, viral, and eukaryotic varieties. They were opened using the open by PDB ID option in Chimera, which will automatically align the two proteins optimally for comparison. Comparison of side-by-side images were also used to show similarities.

4.5 Phylogenetic Networks

Phylogenetic networks were created using a combination of files from MEGA and SplitsTree. First, the protocol for aligning desired sequences was followed for MEGA. After selecting the translated proteins tab, a NEXUS file was exported by clicking Data>Export Alignment>NEXUS/PAUP Format. Note: FASTA files were edited to remove any spaces from comment lines, to create unique identifiers in the NEXUS file when submitted to SplitsTree (Huson and Bryant, 2006). The new nexus files were opend in SplitsTree by clicking File>Open then changing the file type to .nexus files being searched, then clicking Ok. Networks were calculated by using the NeighborNet algorithm.

4.6 Bacterial Cultures

M*ycobacterium smegmatis* stock was inoculated from a freezer stock in glycerol received from the SEA-PHAGES group that was stored in a -80 C freezer. A sterile wand was used to pick a sample of the frozen stock and streak it across an LB plate in a single long streak. Resulting colonies were cultured in a 0.2 um filter sterilized Middlebrook 7H9 liquid medium made per SEA-PHAGES instructor guide. 25 mL medium was aliquoted to a 125 mL Erlenmeyer baffle flask, inoculated with bacteria, and shaken at 30 C at 250 rpm for 36 hours or until the medium appeared cloudy.

4.7 Phage Stock

200 uL aliquot of *M. Smegmatis* culture was mixed with a glass bead from frozen Cocoaberry stock. This sat for 15 minutes before adding 3 mL of PY6A top agar and pipetting onto a LB plate. This sat for 5 minutes before being placed in the incubator at 30 C for 72 hours. 4 mL of phage buffer was pipetted on the plate and the plate was placed in a 4 C fridge for 24 hours. The buffer was pipetted off the plate to establish a solution with Cocoaberry.

4.8 Titering Cocoaberry

The stock Cocoaberry solution was diluted by a 1:10 serial dilution. 100 uL of stock Cocoaberry was added to 900 uL of phage buffer to create a 10^{-1} stock. This was repeated to generate dilutions down to 10^{-12}, and 10^{-3} to 10^{-12} were plated, to find a plate with the most plaques that didn't overlap to calculate plaque forming units (PFUs) per uL.

4.9 RT-PCR Primer Design

Four rules were followed and applied throughout the design process. They are as follows: not having the same nucleotide three times in a row, end in a C or G, the last three bp contain two C or G nucleotides, and the last five bp contain three C or G nucleotides. Two primer sets were designed one to flank each end of coding region of the primase gene and the other to just promote synthesis of the overlap region. After an initial design phase, each primer was blasted through nucleotide blast on phagesdb.org against the *M. smegmatis* mc2 155 database to check for similar sequences in the *M. Smegmatis* genome. All settings were left unchanged except the matrix used, which was switched to PAM30 to look for greater sequence identity. To reduce the amount of hits from the *M. Smegmatis* genome, the primers were moved around following the initial set of rules. All primers had only one or two High score pairings with the genome that were greater than 10 without a gap.

4.10 Growth of phage infected cultures

A 50 mL *M. Smegmatis* culture was cultured to an $O.D._{600}$ of 0.7 A to 1.0 A in a 250 mL Erlenmeyer baffle flask at 30 C shaking at 250 rpm. The culture was centrifuged for 4 minutes at 4170g. The supernatant was removed and kept. The pellet was resuspended in 1 mL of 'stock' Cocoaberry solution for a MOI greater than 10. If more liquid was required to resuspend the

pellet, a small portion of the supernatant that was removed was used. This was shaken for 5 minutes at 250 rpm and 30 C. It was then then centrifuged for 4 minutes at 4170 g. The supernatant was aspirated off and 2 mL of phage buffer was added to rinse the cell pellet of phage. Centrifuged for 4 minutes at 4170 g. Rinsed and centrifuged two more times. The rest of the supernatant was added back after the third wash. This was put in a shaker at 30 C and 250 rpm. At time points of interest an aliquot of 7 mL was taken and spread evenly in 14 1.5 mL eppendorf tubes (500 uL into each eppendorf tube). 1 mL of RNAprotect reagent was added eppendorf tube and it was vortexed for 5 seconds. The samples were incubated at room temperature for 5 minutes then they were centrifuged for 1 minute at 5000 g. The supernatant was removed, and the tube was inverted for 10 seconds on a towel to collect residual moisture. They were placed in a - 80 C freezer for storage.

4.11 RNA Extraction - RNeasy Mini Kit (Qiagen 74104)

Added 700 uL of RLT to a 500 uL pellet and resuspend pellet. Removed 700 uL solution and added to the next tube. Passed this 700 uL solution to all 14 1.5 mL eppendorf tubes. Placed 700 uL solution into 1 100 mircometer zirconium lysing tube and bead beat for 30 seconds at 7.0 ng/sec used quick prep setting on FastPrep-24 5G from MP Biomedical., Incubated on ice for 2 minutes. The cycle of bead beating and icing for 2 minutes was repeated twice more. Then centrifuged for 2 minutes at 14500 g and transferred supernatant to new tube that contained 560 uL of 80% EtOH. This was shook vigorously for 10 seconds. Then transfer 700 uL into RNeasy minispin column. Centrifuged for 15 seconds at 10,000 rpm and flow through was discarded. Added the rest of the supernatant and centrifuged for 15 seconds at 10,000. Discarded flow through again and added 700 uL of RW1 buffer. Centrifuged for 15 seconds at 10,000 rpm. Discarded flow through and placed in new column. Added 500 uL of RPE buffer and centrifuged for 15 seconds at 10,000 rpm. Discarded flow through and added another 500 uL of RPE buffer and centrifuged for 2 minutes at 10,000 rpm. Check if cloudy, if so it needs to be centrifuged for another minute at 10,000 rpm. Discarded flow through and placed in 1.5 mL collection tube. Added 50 uL of RNase-free water to column. Centrifuge for 1 minute at 10,000 rpm. Added another 50 uL of RNase-free water and centrifuge for another minute at 10,000 rpm.

4.12 DNA Removal - Turbo DNA-free Kit (Thermo-Fischer AM1907)

Added 10% (10 uL) of total volume of 10x Turbo DNase buffer and 1 uL Turbo DNase per 50 uL of RNA (2 uL). Incubated at 37 C for 30 minutes. Added another 1 uL of Turbo DNase per 50 uL of RNA (2 uL). Incubated again at 37 C for 30 minutes. Added 20% (20 uL) total volume of DNase inactivation reagent and incubated for 5 minutes at room temperature. Centrifuged for 90 seconds at 10,000g and transferred supernatant to RNAse free eppendorf tube. Make sure that no white beads are dislodged, if some RNA is left behind to prevent the collection of white beads that is ok.

A second run was conducted following slightly different protocol. 1.5 uL of Turbo DNase was used in each addition instead of 1 uL and this was incubated for 45 minutes instead of 30 minutes after adding the DNase.

4.13 cDNA Synthesis

Each time point was split in half, 45 uL, in each eppendorf tube. 1 uL of random primers. Incubated at 70 C for 10 minutes. Then placed in an ice/water bath for 5 minutes. Created two master mixes each contained 4 uL of NEB Protoscript II buffer, 1 uL of 10 mM dNTPs, 1 uL of 100 mM DTT and 0.5 uL RNAse inhibitor at 40,000 units/mL. In one mix, added 1 uL of NEB Protoscript II Reverse Transcriptase at 200,000 units/mL and the other had 1 uL of double distilled H_2O. After the cooling step, 3.75 uL was added to each eppendorf tube. Then incubated at 25 C for 10 minutes. Then 42 C for 2.5 hours.

4.14 RNA Digestion

Created a master mix that contained 20 uL of 0.5 mM EDTA and 20 uL of 1 N NaOH. 10 uL was added to each tube. Incubated for 15 minutes at 65 C. Added 12.5 uL of 1 M Tris pH 7.5 to each eppendorf tube. Then cooled at 4 C for 2 minutes.

4.15 cDNA Clean Up NuceloSpin Plasmid kit (Macherey-Nagel 740499.250)

Transferred contents of eppendorf tube to new eppendorf tube with 400 uL of binding buffer. Inverted 3 times to mix. Then transferred to a MinElute column. Next it was centrifuged for 90 seconds at 14,500 g. Flow through was discarded by aspiration. 750 uL of wash buffer was pipetted to the column. Spun at 14,500 g for 90 seconds. Flow through was discarded by aspiration. Another 750 uL of wash buffer was added and then it was centrifuged at 14,500 g for 90 seconds. The flow through was discarded before being centrifuged again at 14,500 g for 90 seconds. The flow through was discarded again and 30 uL of RNAse free H_2O was added. This was spun for 90 seconds at 14,500 g. The MinElute column was discarded and the flow through was kept.

4.16 PCR

PCR was performed to amplify the target strand by adding 12.5 uL of Q5 NEB master mix, 1 uL of 10 uM forward primer, 1 uL of 10 uM reverse primer, 1 uL of sample DNA, and 9.5 uL of ddH_2O. For the sigma primers, the thermocycler up to run at 95 C for 20 seconds, then 32 cycles that consisted of 95 C for 20 seconds, 62 C for 20 seconds and then 72 C for 30 seconds. Then it was run at 72 C for 2 minutes before being chilled to 4 C for storage. For the overlap primers, the thermocycler up to run at 94 C for 30 seconds, then 35 cycles that consisted of 94 C for 1 minute, 55 C for 45 seconds and then 72 C for 1 minute. Then it was run at 72 C for 1 minute before being chilled to 4 C for storage. The whole primers followed this same protocol, but instead of using 55 C for 45 seconds, it gradually increased by 0.3 C each cycle starting at 50 C for 45 seconds for the first cycle.

4.17 Gel

Added 6.5 uL of Lonza FlashGel 5X loading dye (Cat # 50463) to 25 uL of PCR product. Loaded 4 uL of Flashgel DNA Marker 100-4000 bp (Cat # 50473) ladder into the first well. Then 4 uL of dye + PCR product was added to each well of a 1.2% Lonza Flashgel cassette. Once each well was filled, the cassette was placed in the Lonza FlashGel dock and ran at 275 V for 8 minutes. An image was taken of the gel in the Lonza FlashGel dock using the Lonza camera.

5. Results & Discussion

Several different analyses were performed to accomplish the goals laid out for this project, and to discover the mechanism for creating a functional DNA primase in A4 mycobacteriophages. Initial results describe the extent of the split primase feature as well as the variation of the split primase among the various subclusters in which it is present. Multiple sequence alignments (MSAs) were created to identify conserved regions between split primase genes and to search for conserved motifs that correspond to our different hypotheses. In alignments where exact or close matches were found to different hypotheses, models were created using I-TASSER due to a lack of a template in PDB with enough homology to collected sequences to perform homology modelling. The overall geometry of the produced models was compared to known structures of DNA primase in other phages to determine plausability of both the model and the hypothesis based on the model. Finally, the results of our RNA extraction experiments with RT-PCR were recorded to determine whether or not the DNA primase in Cocoaberry was transcribed using one transcript or two.

5.1 DNA Primase in Actinobacteriophages:

5.1.1 *Introduction and Nomenclature*

The sequences used in the MSAs and phylogenetic networks were collected from Phamerator, a web based platform for the PhagesDB database that organizes both the phages and genes within them into groups for easier comparisons. Based on overall genome similarities, each phage in the database is sorted into a cluster and subcluster of other similar phages. If there are no other similar phages, it is placed into a Singleton cluster by itself. Predicted genes within phages are also sorted into Phamilies, or other groups of genes with a similar sequence. When viewing a genome map in phamerator, each gene can have both its Pham number listed to the left, and the number of other genes in that Pham in parenthesis to the right. For ease of identification, each gene is also given a number in phamerator, which can be listed as the 'gp' of that gene. In the case of Cocoaberry, gp56 will refer to the ZBD gene and gp55 will refer to the RPD gene in later context (view Figure 4 or Figure 10 for reference). We will also refer to specific "conserved domains" within the genes, and these refer to the horizontal layered yellow bars shown in Figure 10 inside each gp. The conserved domains are sections of the gene that match to a specific family of genes in NCBI that have a particular function. They were used as reference points when exploring our hypotheses, because if more of a conserved domain is left out of a final protein sequence due to a slippery sequence or intein splicing, there is a good chance that the protein will not function normally. When referring to specific phages and gps within those phages without a figure, the format will include an underscore after the phage name indicating its cluster and underscores after the gps indicating which one is the ZBD gene or RPD gene (example: Cocoaberry_A4, gp56_ZBD and gp55_RPD).

5.1.2 *A-cluster DNA Primase:*

Several stages of data collection provided insight into the spread of the split primase feature among actinobacteriophages9. Initial collections focused on the extent of the split primase feature within the A cluster following criteria of not including draft sequences, and including at most 5 randomly selected representatives from each subcluster. From this subsection based on data collected from January 2019, there were 449 non-draft A cluster phages (subclusters A1-A20) with 345 phages from the A1, A2, A3, and A4 subclusters. Subclusters A5-A19 were much smaller in size, and the A20 subcluster was not included because it only contained draft phages. The primases for these phages were divided into five categories: Single gene DNA primase (ex. Bigfoot_gp50), Long primase gene plus ZBD gene (ex. Fascinus_A1 gp51 and gp52), two gene primase (< 50 bp overlap) (ex. Bethlehem_A1 gp54 and gp55), two gene primase *analogous to the A4 split primase feature (> 50 bp overlap) (ex. Cocoaberry gp55 and gp56), and short RPD with no ZBD called (ex. MissWhite_A2 gp58) (Appendix A). The majority of phages (378) had the split primase with an overlap greater than 50 bp, or the characteristics of our split primase feature. Even so, there was a wide range of overlap from 150 bp (BeardedLady_BD2, gp39_ZBD and gp40_RPD) to 282 bp (Steamy_A12, gp56_ZBD and gp55_RPD). However, all of the genes in this category had the same types of conserved domains in the ZBD genes and RPD genes in about the same locations, and a list of phams associated with split primase phages is given in “01_27_2019PhamCollectionA1-BX”.

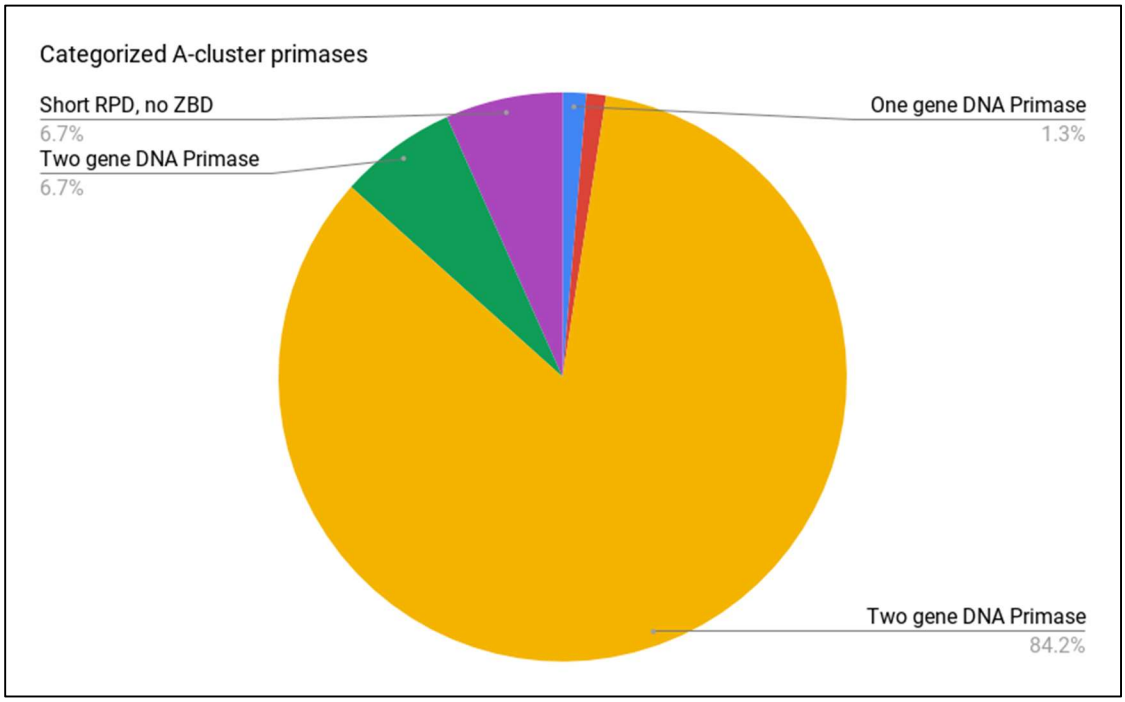


Figure 9: Classification of A-cluster bacteriophage DNA primases. Yellow- two gene primase (overlap >50bp), Green- two gene primase (overlap <50bp), Purple- short RPD, no ZBD called, Blue- One gene primases, Red- Long primase with ZBD.

ZBD genes usually had a region of slightly under 100 bp before the start of the conserved domain, which continued until just under the last 100 bp of the ZBD gene (Figure 10, gp56). This conserved domain was usually labeled as a DnaG, zinc finger, or DNA primase domain. Upon further examination, proteins produced from genes with DnaG domains have matches in sequence to both zinc binding and RNA polymerase functionality, as the DnaG domains map to the complete DNA Primase function. In the case of our collected phages with these domains, we can infer that the conserved domain region represents the functional portion of the gene associated with binding to a zinc atom and attaching to the backbone of a DNA molecule.

RPD genes for the majority of collected phages had conserved domains that spanned a much wider section of the entire gene, leaving only a gap of ~30 bp at the beginning of the gene that is not covered by the conserved domain (Figure 10, gp55). Additionally, there was a distinction in conserved domain label for the three longer domains spanning the entire gene and the five shorter domains covering the latter portion of the gene. The longer RPD domains mapped to a DnaG domain, like the domains found in the ZBD genes, while the shorter RPD domains mapped to a TOPRIM_primase, which is a much broader function including both primases and topoisomerases. We can interpret this data to mean that the majority of the RPD should be conserved in order to retain the more specific DNA primase functionality of the gene, rather than a more generic portion. We can also infer that almost the entire structure of the RPD is needed in this case to perform the function of laying down RNA primers to the DNA template.

An additional note that must be taken into account refers to the gap at the end of the ZBD gene after the end of the conserved domain (Figure 10, gp56). This region, while still conserved for DNA primase function, could be the location for a linker region between the ZBD and RPD protein subunits. When looking for slippery sequences, we compared the sections we found to this gap region to determine its viability as a candidate because it would not exclude too much of the ZBD or RPD genes if the slip occurred in that region.

The phages that had two gene DNA primases with overlap < 50bp still had the same conserved domains as the two gene primases with greater overlap, and the phages that called only the RPD gene had the same conserved domains as other RPD genes.

An anomaly was also found in the A cluster that could show how this trait evolved. Six phages with single gene DNA primases were found exclusively in the A1 subcluster with a shared Pham 1445. When referring to Figure 11, Pham 1445 appears to have two distinct regions with conserved domains. The earlier conserved domains had a HINT (Hedgehog/intein N-terminal region domain) function and spanned from residues 113-189 in Bigfoot_A1 (Figure 11). The domains in the right of the figure had a combination of TOPRIM_primase and DnaG functions.

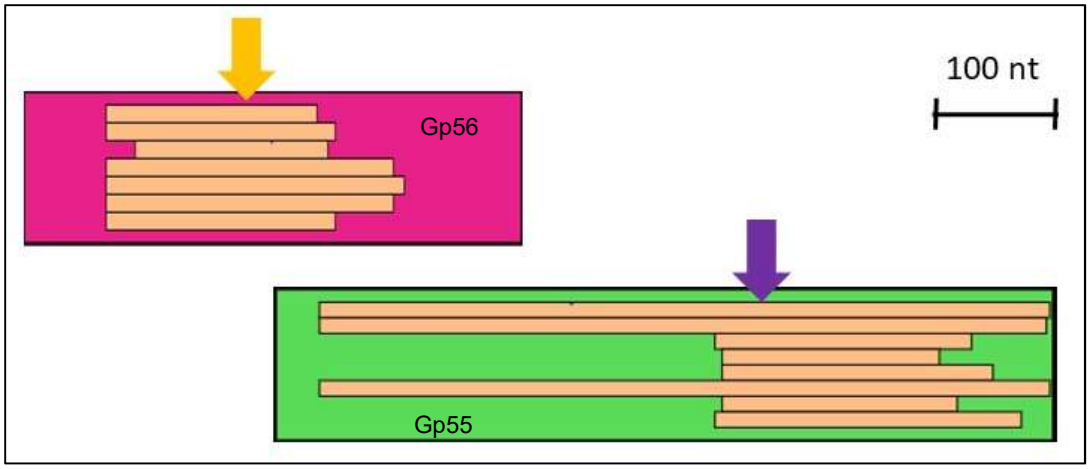


Figure 10: Phamerator image of ZBD gene (pink-gp56) and RPD gene (green-gp55) with conserved domains shown. Yellow arrow points to the DnaG conserved domain, and the purple arrow points to the conserved TOPRIM domains.

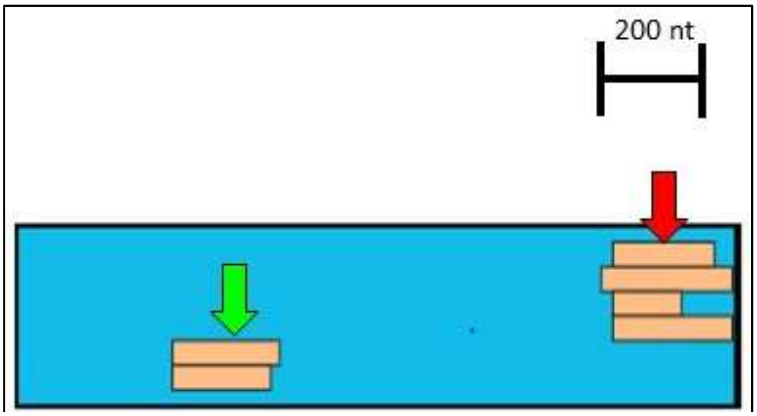


Figure 11: Image of Bigfoot_A1 gp50_ZBD, one of the six A1 subcluster phages with a single gene DNA primase. The conserved domains in this gene are marked by the green arrow (Hedgehog/intein domain), and the red arrow (TOPRIM domain).

In the multiple sequence alignments of the A-cluster split primases with the single gene primases from the A cluster, there did not appear to be any conservation of sequence (PrimaseZBDPlusBigfootBones.mas). A number of single primase A1 subcluster phages were modeled using I-TASSER to examine their unique structure. Bigfoot_A1_gp50 was one model that showed a similar geometry to known DNA primase structures. It maintained a L-shaped structure with a ZBD and RPD connected by a linker. However, the ZBD appeared to be much shorter in the model (only 36 residues) when compared to split primase ZBDs such as Cocoaberry (137 residues), and Bigfoot's RPD appeared to be much longer than the RPD of Cocoaberry (369 residues vs 219 residues) (Appendix B. Bigfoot_Model1). This may be due to inaccuracy of the modelling process, but confirmation of a crystalized structure would be needed to solve the discrepancy.

5.1.3 *Actinobacteriophage Database Split DNA Primase:*

In the second collection of data as of January 27, 2019, a survey of random samplings from all actinobacteriophage clusters by using a keyword "primase" was done to examine the extent of the split primase feature throughout other clusters. In this search, a categorization of single gene primases was also done to explore the diversity of sequence and conserved domain layout. The split primase subclusters included the A1-A19, BF, BD1-4, BD6, BU, CA, and singleton (Ibantik) subclusters. Measures of gene similarities were calculated roughly by using distribution of phams, and through multiple sequence alignments. There were 3 RPD phams among the distributed split primase RPDs, and 7 ZBD phams arguing that there is a greater amount of conservation in sequence between RPD genes of the split primase than ZBD genes (Figure 12). Multiple sequence alignments of all the split primase zinc binding domains revealed minimal conservation of sequence, notably a conserved cysteine residue aligning with residue 56 of Cocoaberry_gp56 that could be relevant to our intein splicing hypothesis, as this conserved cysteine was present in all split primase ZBDs (Figure 13, the full supplemental file can be viewed using MEGA: DNAPrimaseZBDAlignmentNoDupsClusterLabeled.mas).

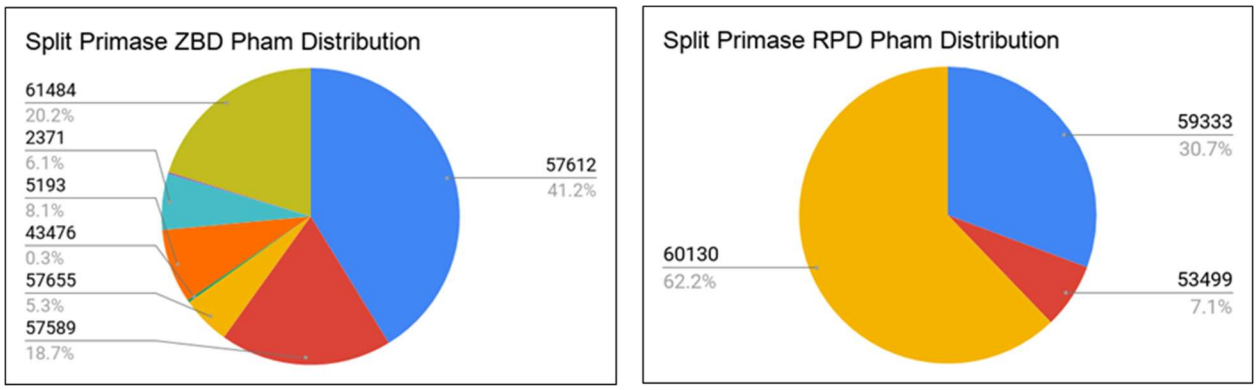


Figure 12: Amount of pham diversity among split primase phages divided by gene. Intial results showed a greater amount of diversity among ZBD gene sequences than RPD sequences. Phages were from all sub-clusters containing the split primase (A1-A19, BF, BD1-4, BD6, BU, CA, and singleton (Ibantik) subclusters).

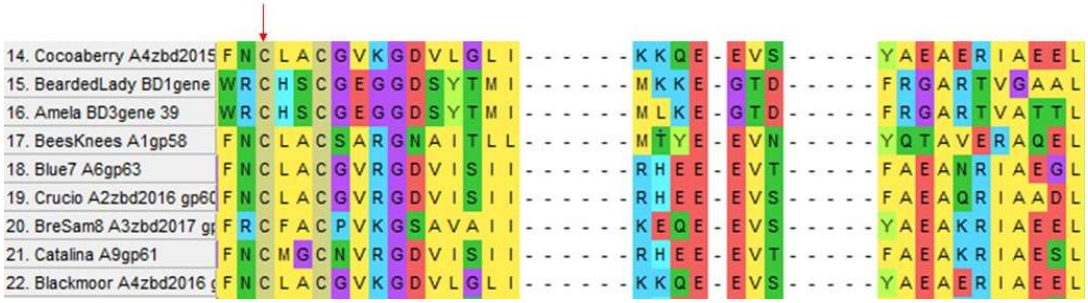


Figure 13: Subsection of split primase ZBD alignment showing conservation of cystine between Cocoaberry, two BD cluster phages, and other A cluster phages (marked by the red arrow).

Alignments of the RPD sequences collected from split primase families showed a greater degree of conservation between residues when compared to the total alignment of the ZBD

genes. This is apparent due to both the number of positives increasing, less gap insertions needed to create the alignment, and a greater number of residues being at least 80% conserved when viewed in MEGA. Additionally, the minimal conservation of residues in the A cluster for the conserved cystine at between a section of the A cluster phages served as the end of the C-terminal split intein for Cocoaberry that was used in the MSAs.

For single gene primases, over 30 individual phams existed to categorize the genes, and the distribution of domain patterns was also diverse. Conserved domains of long single gene primases typically had a primase/polymerase function and the same conserved domains as found in the A class for primase sections. Ranges in length for single gene primases ranged from similar lengths to the combined ZBD + RPD domains in Cocoaberry (~900 nt), to just under 3kB in pham 6612 of the EF phages.

It is also important to notice that the number of subclusters represented by our data collection of both single gene primases and split primases may under-represent the total diversity of DNA primases in actinobacteriophages. A large number of subclusters surveyed did not have annotated genes with DNA primase as a listed function. This may be due to lack of knowledge in the annotation process to categorize a gene as having a DNA primase function, or due to the use of host bacteria DNA primases as in *Analysis of Cis and Trans acting elements required for the initiation of DNA replication in the Bacillus subtilis bacteriophage SPP 1* (Pedre et al., 1994.)

After multiple alignments were created for both the split primase proteins and the single gene primases, several phylogenetic networks were created to show conservation of sequence both within the split primase subclusters and between those subclusters and single gene primases. The phylogenetic network of the ZBD genes for the A1 through CA subclusters was based on the MSA of the same sequences which was created in MEGA using the MUSCLE algorithm aligning by codons. As expected due to the distribution of ZBD genes into 7 distinct phams, the relative divergence between sequences is wide. The SplitsTree network created from the ZBD alignment showed that divergence (Figure 14). The split primase ZBDs showed correlations by

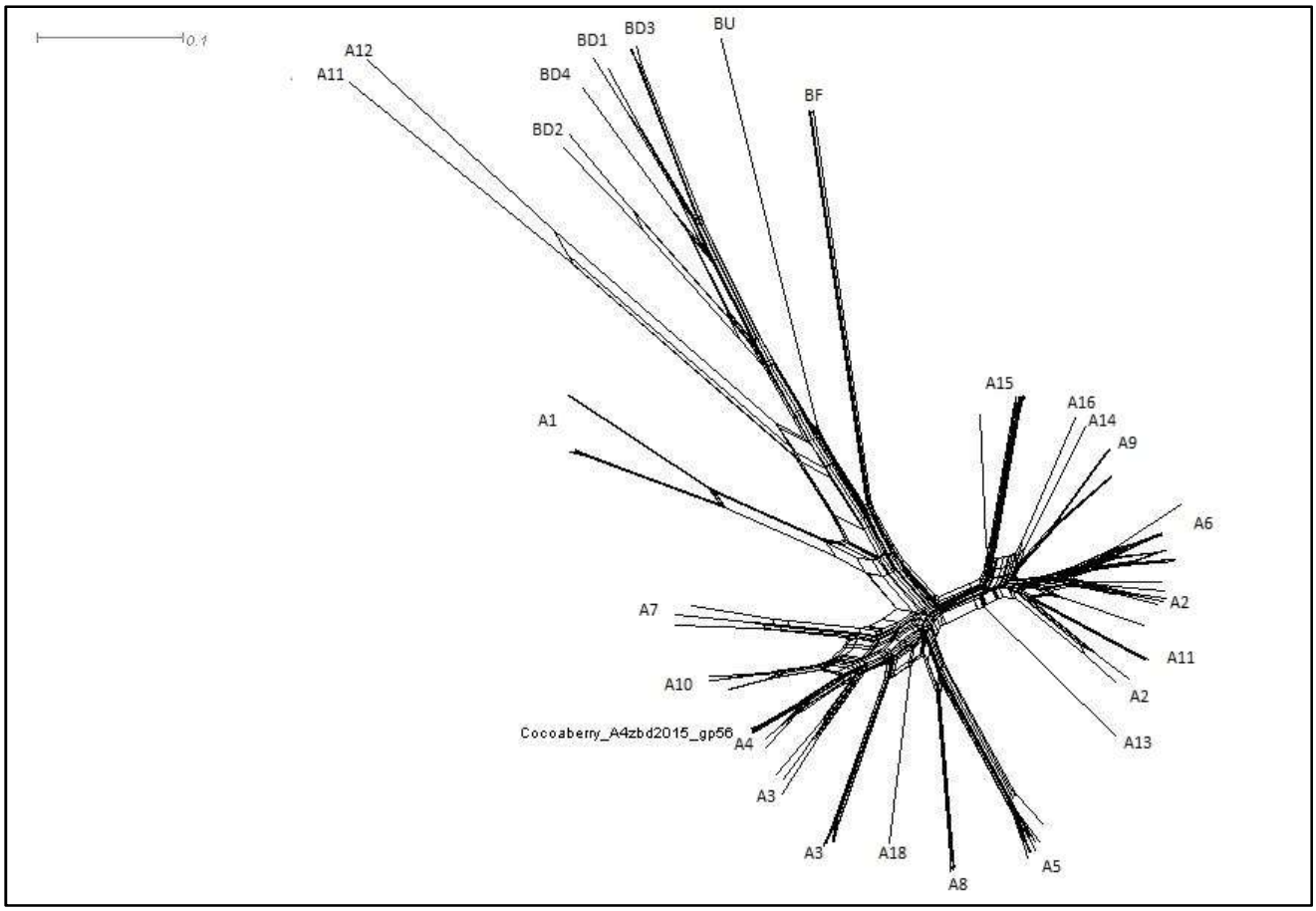


Figure 14: Phylogenetic network created by the MSA from MEGA of the split primase ZBDs, labeled by overall subcluster in each area of the tree.

subcluster for the random sampling that was provided, but the distribution of subclusters did not correspond to the distribution of phams for these genes. While the majority of A subcluster phages are distributed between two distinct branches of the network, subclusters A11 and A12 remain outliers with even more sequence diversity than B cluster split primases. While the overall length of the tree shows that there may be enough similarity between sequences of the split primases to expect conserved motifs such as translational or transcriptional slippery sequences or intein motifs, there is a large amount of diversity even among a relatively small group of phams.

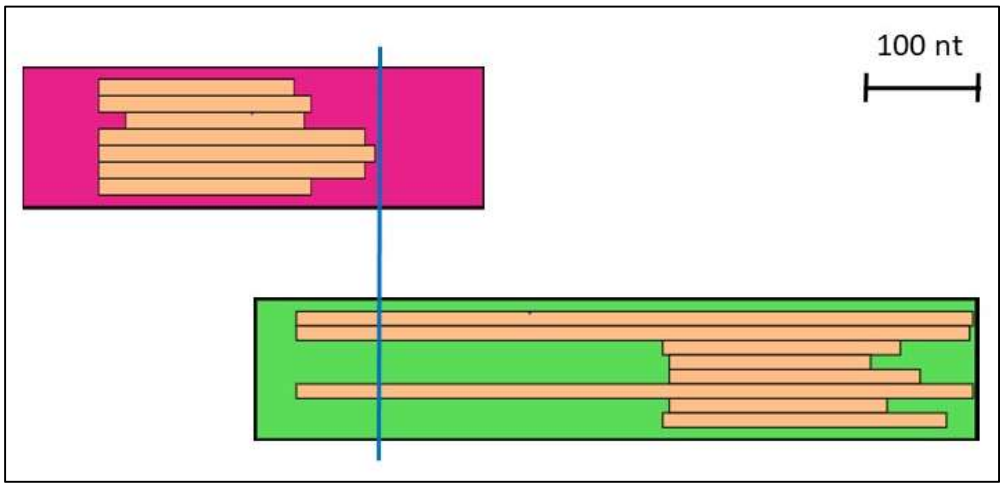
The phylogenetic network for the RPDs of the split primases showed a greater degree of conservation between sequences as well due to the relative short length of branches extending to final taxa (Appendix C). The distribution of sequences was still diverse and appeared to branch into several groups for the A cluster phages and with greater length of substitutions per site for the B cluster phages from the A cluster phages. However, the greater sum of the sequence distances was found in internal edges of the tree as opposed to at the edges.

"OnesZBDCombined.nexus", a network in the supplementary data created in SplitsTree, shows the combined network of split primase ZBD and single gene primases in order to determine the distinction of the split primases as a group from the single gene primases. Most of the sequence similarities between these groups should come from the conserved domains shared between the groups, however a distinct branch from the center of the network shows closer sequence relatedness between split primase ZBDs when compared to the single gene DNA primases. This network also shows the diversity of the single gene primases, as the majority of those sequences do not branch into distinct groups like in Figure 14 (Supplemental File: "OnesZBDCombined.nexus").

5.2 Plausible Locations for Motifs within Cocoaberry

When searching the sequences of the ZBD and RPD in Cocoaberry for conserved motifs in translational or transcriptional slippage or for intein motifs, there were several distinct factors we needed to take into account. In addition to conservation of the motifs throughout the split primases of phages in our MSAs, the location of the motifs could not affect the function of the protein too drastically. Thus, motifs needed to be located so as to exclude as little of the functional domains of both genes as possible, through either slippage too early or intein splicing that cuts too much of the functional domain out. The coordinates of the ZBD and RPD genes had an overlap region of 185 bp, with a total length of 888 nt for the entire genome region (Figure 15.a). In addition to the gene coordinates, both genes had mapped functional domains that were marked in Phamerator. The ZBD functional domain extends from 72-318 nt for the longest domain, and from 72-246 nt for the shortest domain, while the RPD extends from 42-657 nt for the longest domain, and from 381-564 for the shortest domain (Figure 15.b, Figure 15.c) Ideally, a slippery sequence motif would occur in the region after the end of the longest ZBD domain and before the start of the longest RPD conserved domain.

A



B.

(38372 (1))-
ATGCCATCAACGGATGAGCCGCTGATAGTCCAGGCGATCCACCGGTACCACCCGGACTGGGAGCCAC
CGAAGGACACCGGCAAGGACTGGATCAAGTGTCTGTGCCCGTTCCACGCCGAGGAGGTTCCCTCGGC
TGCCGTCTCGTTCGTGCGCCAGGCGTTCAACTGCCTGGCCTGCGGGGTGAAGGGCGACGTGCTCGG
GCTGATCAAGAAGCAAGAGGAGGTGAGTTATGCAGAGGCTGAGCGAATCGCAGAAGAGCTTTCTCCG
GGAGGCAACCGAGCGGTACCGGCGAAGTCTGAACGGCAGTCCAGCCGAAGAGTATTTGGCGACAAGG
GGACTGATGTTCGACAGCGTCAGGGACGAAGTCGACCGGTTCATGCTCGGGTACGTGGACGACCCAC
TCCCTGGTCATGA- (37959 (413))

C.

(38143 (229))-
ATGCAGAGGCTGAGCGAATCGCAGAAGAGCTTTCTCCGGGAGGCAACCGAGCGGTACCGGCGAAGTC
TGAACGGCAGTCCAGCCGAAGAGTATTTGGCGACAAGGGGACTGATGTTCGACAGCGTCAGGGACGA
AGTCGACCGGTTCATGCTCGGGTACGTGGACGACCCACTCCCTGGTCATGAGATGTTCCGGGGCTTCA
TGGCGATCCCGTACCTGCGCTGGTCGCGGGAACACGGCTGGATCGTCGTCGCGATCCGCTACCGCTG
CATCCAGGACCACGACCACCGAGGGCATGGCAAGTACATGACCGCGCCGGGGGATCAACCGTGGCT
GTACAACACTCTCGCGCTGCTGCGTGAGGTCCCCGATGTCGCGATCACCGAAGGTGAGATCGACGCC
ATCACAGCGCAAGTCTGCGGCCTCCCGGCCGTGGGGGTGCCTGGGGCCAACATGTGGAAGCCCTACA
TGCGAGAGCTGTTCATCGGATACCGAACCGTCTACGTCCTCGCGGACGGCGACGAGCCCGGAGCCGA
GTTCGCCAACCGCGTGGCACTGACCCTGCCGAACAGCAGGGTGATCCCGATGCCACCTGGTGAGGAT
GTCAACTCGCTAGTCATCAGCAAGGGCAAATCCGCTCTGCTGGAAAGGATGTCATGA- (37484 (888))

Figure 15: To scale map of ZBD (gp56- pink) and RPD (gp55- green) genes in Cocoaberry with conserved domains shown for each. The veritcal line in blue shows the site of the candidate translational slippery sequence 'TTTGGCG' in relation to those conserved domains. B. Separated genome coordinate region containing the ZBD gene (blue caps on ends), with the longest (pink) conserved domain, and overlap between the shortest and longest domains (purple) marked. Position of the candidate translational slippery sequence is highlighted in green. C. Separated genome coordinate region containing the RPD gene, with the longest (pink) conserved domain and overlap between the shortest and longest domains (purple) marked. Position of the candidate translational slippery sequence is highlighted in green. Left numbers are genome coordinates while right numbers in parenthesis are genome coordinates normalized to the length of the whole region.

5.3 Transcriptional slippery sequence

The conventional sequences required for a transcriptional slippage to occur is a poly A or T ('AAAAAA' or 'TTTTTT') or a sequence of three C's and a minimum of 8 U's (3'CCCUUUUUUUU'5) (Baranov et al., 2015). To determine whether such a 'slippery sequence' could result in a single transcript that would create a single frame for the DNA primase, we created a multiple sequence alignment (MSA) consisting of a random selection of A and B phage selected by using the sequence collection method that contain the split primase gene. The MSA only includes the area of overlap because the switch of reading frame needs to occur here to allow both domains to form and function properly. A manual search of this alignment of the overlap region for any poly sequence greater than five nucleotides in length of a single nucleotide yielded no results, which does not support the hypothesis of having a conventional transcriptional slippery sequence.

5.4 Translational slippery sequence:

5.4.1 *Overall approach.*

In order to conclude that the DNA primase of Cocoaberry is created by a translational slippery sequence mechanism, several criteria needed to be satisfied. First, there must be a match to a canonical slippery sequence from literature in the zinc binding domain of Cocoaberry, or gp56. Second, the matching sequence in Cocoaberry must also be a conserved sequence among the ZBDs of other split primases in the A-cluster. We determined this criterion because we found it to be unlikely that several different conserved motifs and mechanisms would be present in similar phages to form functional DNA primases. Third, the nucleotides remaining upstream from the site of the matching slippery sequence must be able to form a functional ZBD. To fulfil this criterion, we determined that sequences that meet the first two points must also be at least downstream of the shortest predicted functional domain in the ZBD by Phamerator. This would allow enough nucleotides upstream of the slippery sequence to form a functional ZBD that can bind to a DNA molecule. Since a translational slippery sequence would need to be in the overlapping region between Cocoaberry_gp56 and Cocoaberry_gp55, part of the beginning of the RPD gene would be cut out of the final sequence of residues. We needed to make sure the slippery sequence left enough of the conserved domain in the RPD gene to allow a functional RPD subunit to form in the protein. As an additional caveat for this point, we determined that there must be enough space between the start of the slippery sequence in the ZBD and the beginning of the functional domain of the RPD to accommodate a linker, as the linker region connecting the two domains was found to be highly conserved.

5.4.2 *There was no conservation of translational slippery sequences from literature in bacteriophages with split primase.*

Following one direction of the approach in searching the multiple sequence alignment of zinc-binding domains of split primase families for conserved translational frameshift sequences from the literature, a collection of known translational frameshift motifs was gathered (Table 1,

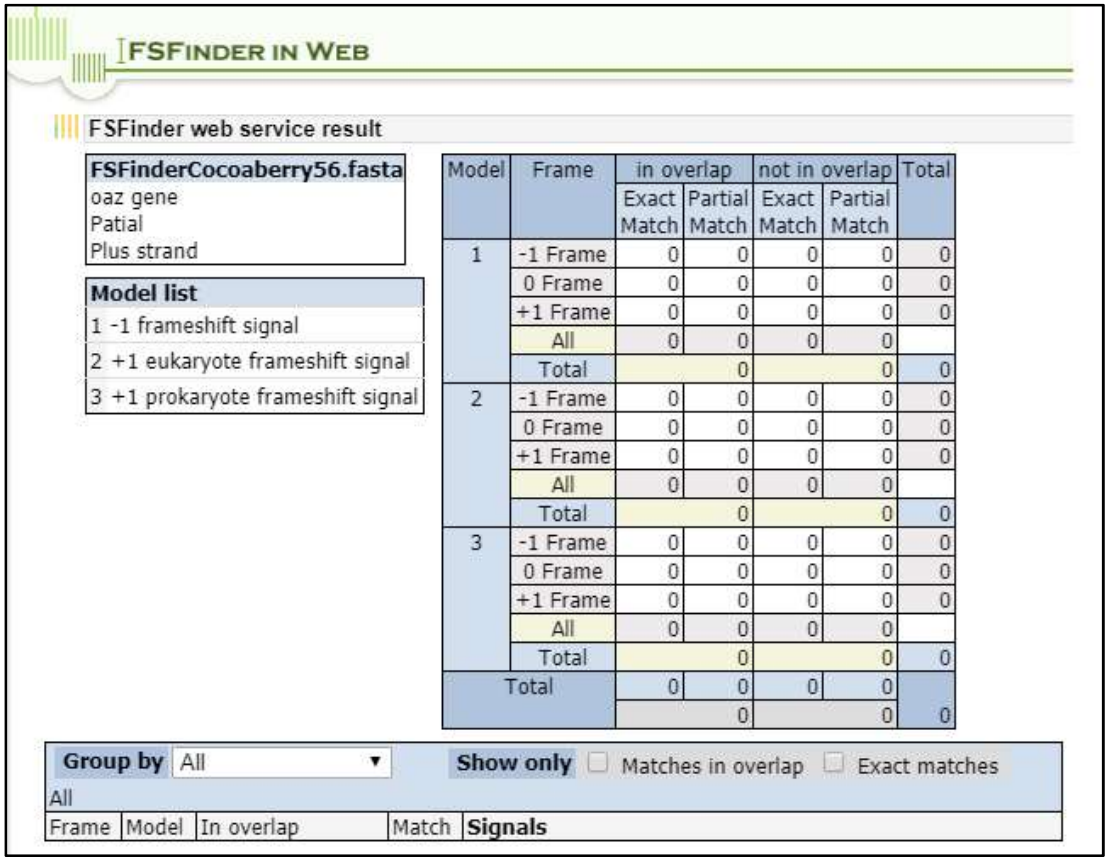
Atkins et al., 2016). Patterns for a translational slippery sequence varied in both length and number of repeated nucleotides as they originated from a wide variety of taxa. Generally, sequences contained at least a three or four nucleotide repeat in the canonical slippery sequence while using either a stem-loop or pseudoknot mechanism further downstream separated by a spacer region between 6-9 nt long (Napthine et al., 2017). However, the amount of diversity in slippery sequence motifs required manual comparisons to the multiple sequence alignment of split primase ZBDs using the search function in MEGA. The most common pattern among the motifs was the 'XXXYYYZ' pattern with a stem loop, and this pattern was additionally queried using the FSFinder program on the ZBD in Cocoaberry.

Table 1: Collection of translational slippery sequence motifs from Atkins et al., paper sorted by overall nucleotide pattern. These sequences were used as queries in the multiple sequence alignment of both the 60 A-cluster bacteriophage ZBDs, and the combined A, B, and C cluster MSAs.

Translational Slippery Sequence Motif	**Sequences**
X_XXX_XXY	A_AAA_AAC, U_UUU_UUA, U_UUU_UUC
X_XYY_YYZ	G_GAA_AAC
X_YYY_YYZ	G_AAA_AAC
X_YYY_YYY	G_UUU_UUU
X_XXY_YYY	A_AAU_UUU, G_GGU_UUU, G_GGC_CCC, G_GGA_AAA, C_CCU_UUU
X_YYX_XXX	U_CCU_UUU
W_WXY_YYZ	G_GAU_UUC
X_XYY_YYY	G_GUU_UUU
X_XXY_YYX	U_UUA_AAU, A_AAU_UUA
X_XYZ_ZZZ	G_GAU_UUU
X_XXY_ZZ*	C_CCG_AA*
X_XXX_XYY	G_GGG_GAA
X_XXY_YYZ	U_UUA_AAC, G_GGA_AAC, G_GGA_AAU, G_GGC_CCU, G_GGU_UUA, C_CCU_UUA, A_AAU_UUC
X_XYZ_ZZY	G_GAU_UUA
W_XXY_YYZ	G_UUA_AAC
W_XYZ_ZZY	G_UCA_AAC
Unique	CCC_UUU_CGA, CCC_UUU_CGU, UCU_UUU_CGU, ACU_UUU_CGC, UUU_UUU_CGA, AG_GUU_UUU, AG_AUU_UUU, CG_GUU_UUC, GG_GUU_UUU, GU_GUU_UUU, UG_GUC_UCU, AU_UUC_UCU, UCU_UUU_CGA, UCC_UUU_CGU, GUU_CGG, UUU_CGA, GUU_stop_C

The results from both the manual search of motifs and the FS finder results found no conserved translational frameshift motifs. FS Finder searched for both the ‘XXXYYYZ’ pattern and a pseudoknot or stem loop downstream from any predicted match or partial match, but it was unable to find any matches in the Cocoaberry gp56 (Figure 16). Manual searches of the multiple

sequence alignment found one exact match to any canonical sequence. The ‘C_CCG_AA*’ motif was only found in Attacne, a BU cluster phage. It started at 86 nt, and since the beginning of the first conserved ZBD in Attacne is at 108 nt, it failed the last two criteria.



FSFINDER IN WEB

FSFinder web service result

FSFinderCocoaberry56.fasta
oaz gene
Patial
Plus strand

Model list
1 -1 frameshift signal
2 +1 eukaryote frameshift signal
3 +1 prokaryote frameshift signal

Model	Frame	in overlap		not in overlap		Total
		Exact Match	Partial Match	Exact Match	Partial Match	
1	-1 Frame	0	0	0	0	0
	0 Frame	0	0	0	0	0
	+1 Frame	0	0	0	0	0
	All	0	0	0	0	
	Total		0		0	0
2	-1 Frame	0	0	0	0	0
	0 Frame	0	0	0	0	0
	+1 Frame	0	0	0	0	0
	All	0	0	0	0	
	Total		0		0	0
3	-1 Frame	0	0	0	0	0
	0 Frame	0	0	0	0	0
	+1 Frame	0	0	0	0	0
	All	0	0	0	0	
	Total		0		0	0
Total		0	0	0	0	
			0		0	0

Group by All ▾ Show only Matches in overlap Exact matches

All

Frame	Model	In overlap	Match	Signals

Figure 16: Image of web server results from FS finder in searching for both exact and close matches to the 'XXXYYYZ' motif and for pseudoknots in the given fasta sequence for Cocoaberry gp56.

5.4.3 *Reverse search of conserved regions in the ZBD partially matched to a canonical motif.*

Since canonical motifs were not found in the MSA of the ZBDs, a reverse approach was tried by looking for conserved sequences among the ZBDs that were downstream of the functional domain. These sequences were compared to slippery sequence motifs for accuracy. Using this approach, one conserved section of the MSA among A-cluster phages (about 80% conservation when searched in MEGA for all split primase ZBDs), was found to have a partial match to the canonical motif ‘XXXYYYZ’. Since it was selected as a conserved sequence, it automatically satisfies the second criterion. This motif started at 321 nt, and since the end of the longest ZBD conserved domain is at 318 nt, it satisfies the third criterion. It contains the sequence ‘TTTGGCG’, which follows an ‘XXXYYZY’ pattern (Figure 17)

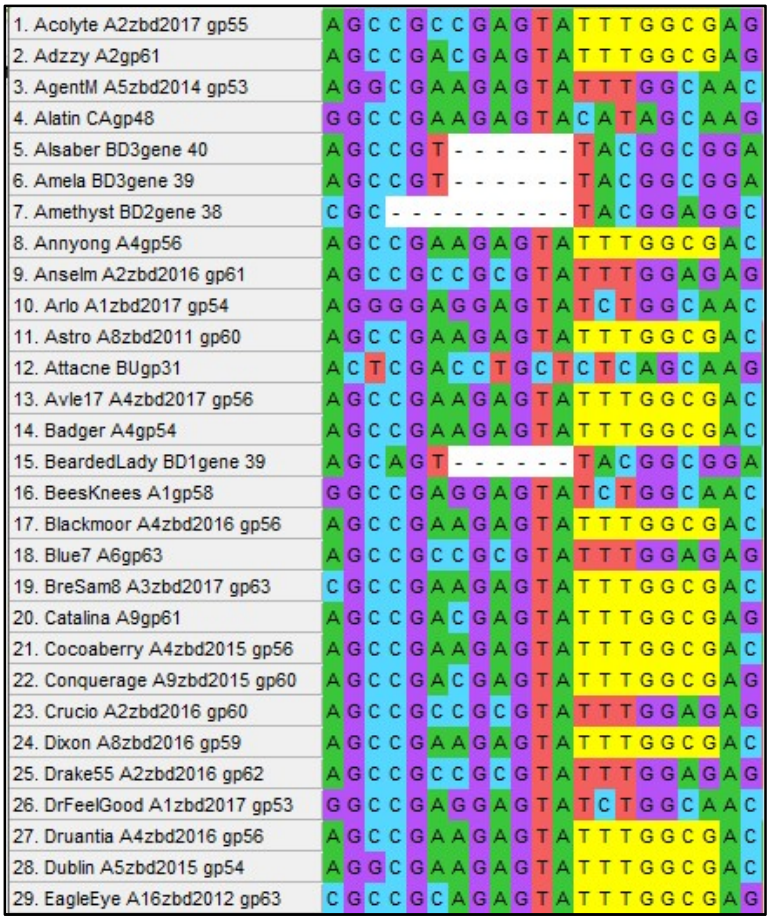


Figure 17: Section of A, B, and C cluster ZBD MSA showing partial conservation of a possible translational frameshift sequence 'XXXYYZY' pattern. Conservation was strong among A cluster phages but did not persist in B or C cluster ZBDs.

However, when this sequence was compared for conservation between all split primase families, including those B cluster and CA subcluster ZBDs, there was no conservation of the ‘TTTGGCG’ sequence among the B or CA clusters. As the selected sequence did show conservation among A cluster phages, additional exploration of the mRNA structure of the ZBD in Cocoaberry was performed to determine the presence of a spacer region downstream from the site, as well as an RNA secondary loop or pseudoknot. Results from RNAstructure revealed a longer spacer region of 21 nucleotides before a stem-like structure from position 350-363. This could provide supporting evidence for the use of this region as a slippery sequence, however the length of the spacer region is longer than most conserved spacers between 6-9 nt (Matthews et al., 2004, Figure 18). While the slippery sequence candidate is not conserved in the B-cluster and CA cluster split primase ZBDs, the presence of a spacer and linker downstream from the

candidate site could indicate use of a slippery sequence. This “candidate” slippery sequence, while not matching any current canonical slippery sequences, could warrant further exploration.

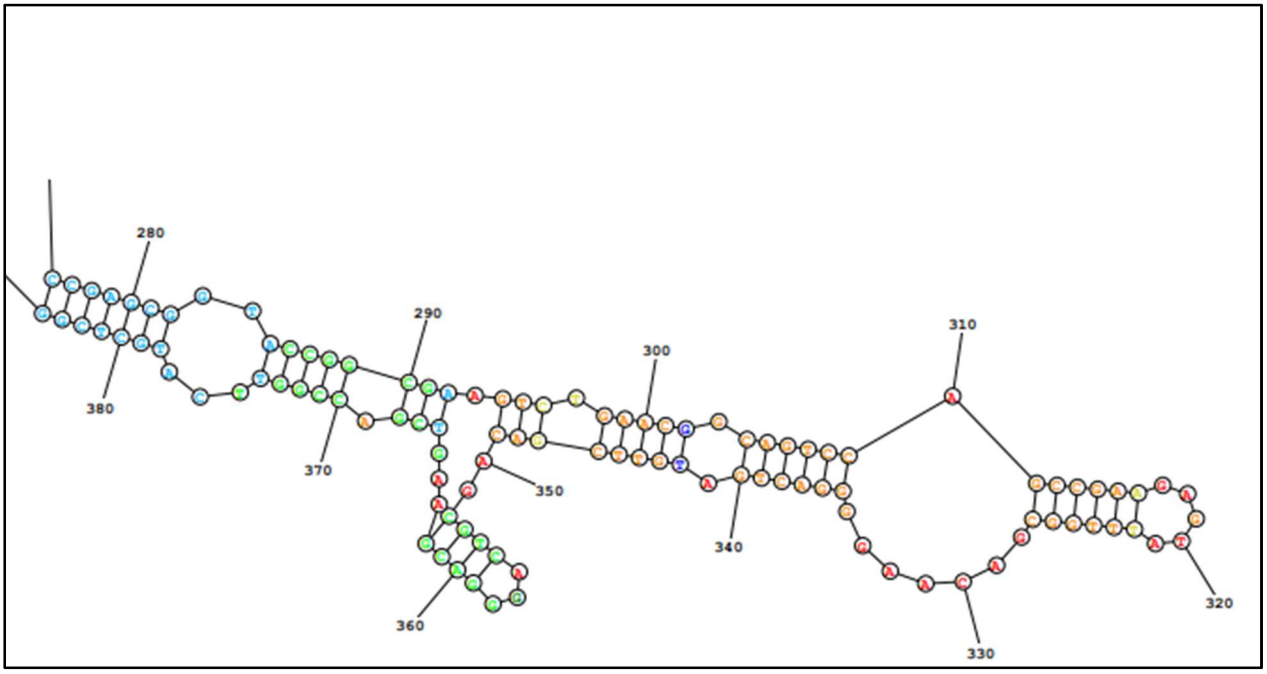


Figure 18: Computationally predicted secondary structure of Cocoaberry ZBD RNA. A longer spacer region (329-349 nt) exists between the end of the 'TTTGGCG' sequence and the beginning of a stem structure (350-366 nt). Colors from red to blue indicate confidence levels for structural components (Red >99%, Orange 95-99%, Yellow 90-95%, Dark Green 80-90%, Light Green 70-80%, Light Blue 60-70%, Dark Blue 50-60%, Pink < 50%).

The predicted model for the candidate ‘TTTGGCG’ sequence was created using I-TASSER to predict the general structure of the protein based on the combined genes including the slip into the RPD gene frame. When compared to the crystalized image for T7 primase (1NUI), there appears to be the same general domains in the resulting protein (ZBD + linker + RPD). The length of the RPD and ZBD was longer in the Cocoaberry I-TASSER model than in the T7 model, while the length of the linker was comparable. Future experiments could try to isolate the structure of the Cocoaberry DNA primase, but the model provided a good overall structure showing the plausability of the slippery sequence model (Figure 19).

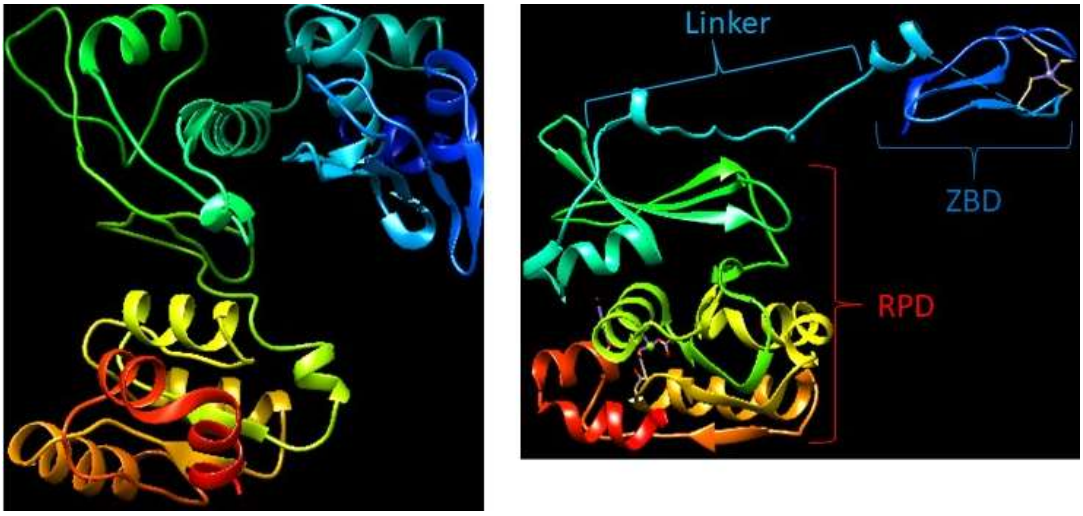


Figure 19: Right: Crystalized model of T7 DNA Primase from PDB (1NUI) with labeled ZBD, linker, and RPD sections compared to the I-TASSER generated model for Cocoaberry from combined sequence using slippery sequence 'TTTGGCG'. Results were viewed using UCSF Chimera.

5.5 Inteins

There are three criteria that must be met for trans-splicing to occur. First, the charge strands of the C-terminal end of the N-terminal intein must be negative and the N-terminal end of the C-terminal intein must be positive or vice versa (Figure 20). Second, certain amino acids must be present at a precise location within the N-terminal and C-terminal inteins. Third, the length of the excised split intein should be close to the minimal amount of amino acids required for splicing, in order to leave intact the ZBD and RPD, along with space for a linker region. Each criterion contributes to the excision of the intein and ligation of the extein.

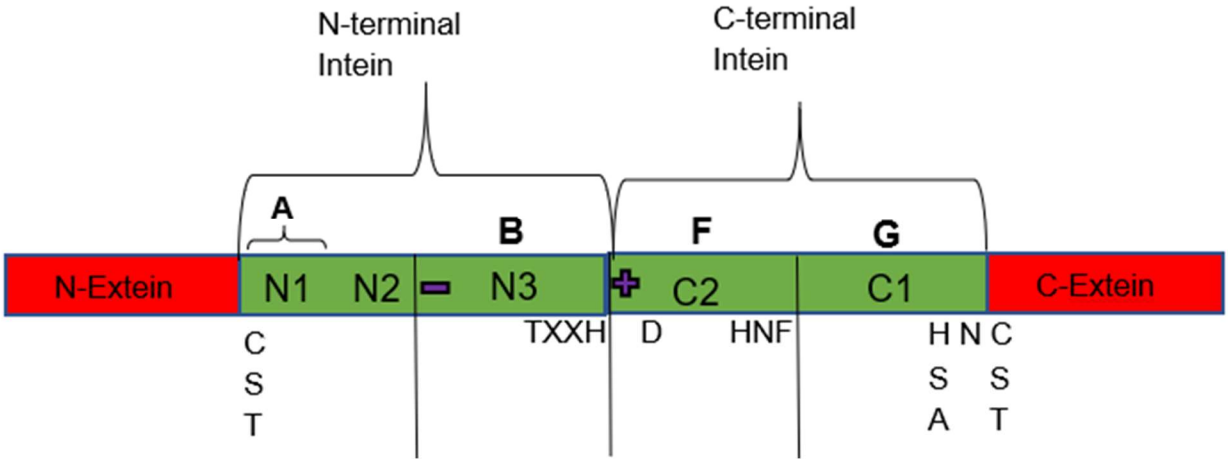


Figure 20: Structure of both halves of a split intein with required residues and charged regions marked.

Thirteen cyanobacterial split-intein pairs, formerly identified by Perler et al., were examined. Through primary literatures, many of the split-inteins researched were from archaebacteria organism. However, only one archaebacterial split-intein was found and

documented in the intein database; therefore, the cyanobacterial split-inteins were used to provide enough evidence for this mechanism. Despite that these are not archaebacterial split-inteins, they include the four conserved motifs needed for this mechanism to occur. Moreover, the selected split-inteins aligned in this paper only share 51-57% protein sequence identity with each other, thus containing natural variability in both intein halves. The only difference found between cyanobacterial split inteins and archaebacterial split-inteins is the change of amino acid histidine at the C' end of the C-terminal intein to a serine or alanine. The serine and alanine, like the histidine, help with the cleavage of the C-terminal intein from the C-extein.

The sequence that best matched the overall pattern of conserved residues and blocks in Cocoaberry was greatly determined by the presence of the conserved cystines in the Cocoaberry_gp56 and Cocoaberry_gp55. The interior region between the two yellow lines in represents the intein region for the C-terminal and N-terminal halfs of the split intein. The effect of the cut into the bulk of the RPD conserved domain will be shown later in the results (Figure 21).

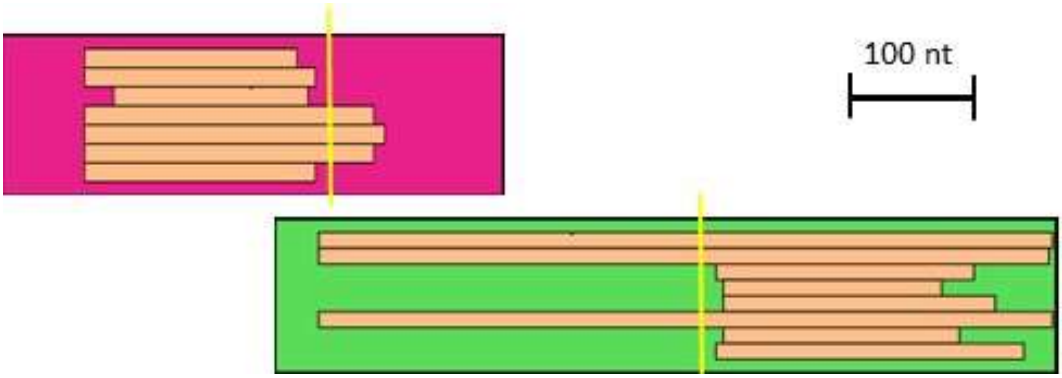


Figure 21: Graph of Cocoaberry ZBD and RPD showing conserved domains in their respective genes. The region after the yellow line in Cocoaberry_ZBD (pink) will represent the Cocoaberry N-terimal intein, while the region before the yellow line in Cocoaberry_RPD (green) will represent the C-terminal intein. The fasta file used for modelling will be a combined portion of the outer halves (extein region).

5.5.1 *Oppositely charged strands were not found.*

To demonstrate the protein splicing activity of the thirteen examined split inteins, their N- and C- terminal halves were aligned. There are two negatively charged amino acids: aspartic acid (D) and glutamic acid (E), and three positively charged: lysine (K), arginine (R) and histidine (H) that are used to identify the charge of the N- and C-terminal strands. As suggest by the data presented by Dassa et al.,, the fourteen amino acid long N-terminal charge strands starts eight amino acid before the conserved TXXH found in motif N3 (see black box, Figure 22). The sequence alignment of the split-inteins found in cyanobacteria were found to all have a negatively charged strand at the N-terminal charged region. In contrast, the charged strand for Cocoaberry's intein resulted in one negative amino acid and three positive amino acid, making the candidate N-terminal charged region positive. But, to form a salt-bridge between the two proteins there needs to be two strands with opposite charged strands, therefore we additionally looked at the charge present at the C-terminal charged region.

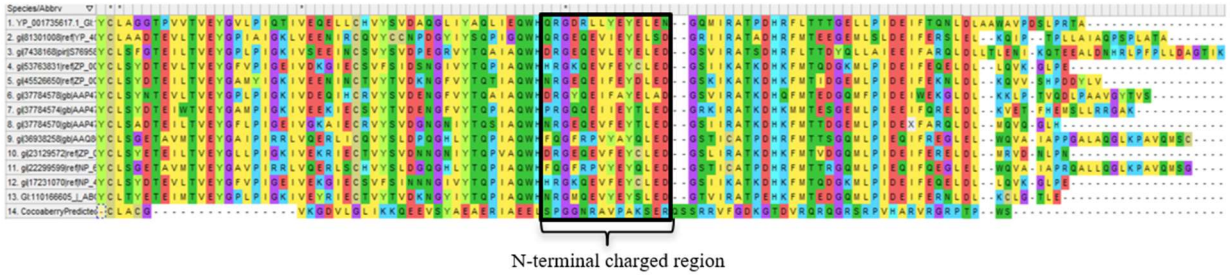


Figure 22: Section of MSA aligned by MUSCLE for collected InBase split intein N-terminal sequences compared to the predicted possible N-teriminal split intein of Cocoaberry.

As can be seen in Figure 23, the alignment between the split inteins found in the intein database and our candidate intein had a lot of gaps due to Cocoaberry's long sequence. This sequence length was chosen because it prevented the splicing of the last subdomain in the RPD. In addition, there was only one cysteine present in the entire RPD, therefore it was assumed that it must be where the C-terminal intein excises. Nevertheless, we still compared the split inteins from the database to the candidate intein. As noted by Dassa et al., the charged region of the C-terminal intein of cyanobacteria is 12 amino acids long, starting from the first positive charged amino acid. As can be counted in Figure 23, the split-inteins from the intein database had more positively charged amino acid, thus making the C-terminal charged strand positive. This evidence suggests that the split-inteins from the intein database forms a salt bridge between the two proteins. However, when the first 12 amino acids were counted in the candidate intein for Cocoaberry, two positively charged amino acids and two negatively charged amino acids were found, consequently making the candidate intein C-terminal charged strand also neutral., Therefore, without the ionic interactions of the opposite charged strands within Cocoaberry's intein, there would be no salt bridge formed between the two protein, thus there would not be any intein splicing activity for Cocoaberry intein.

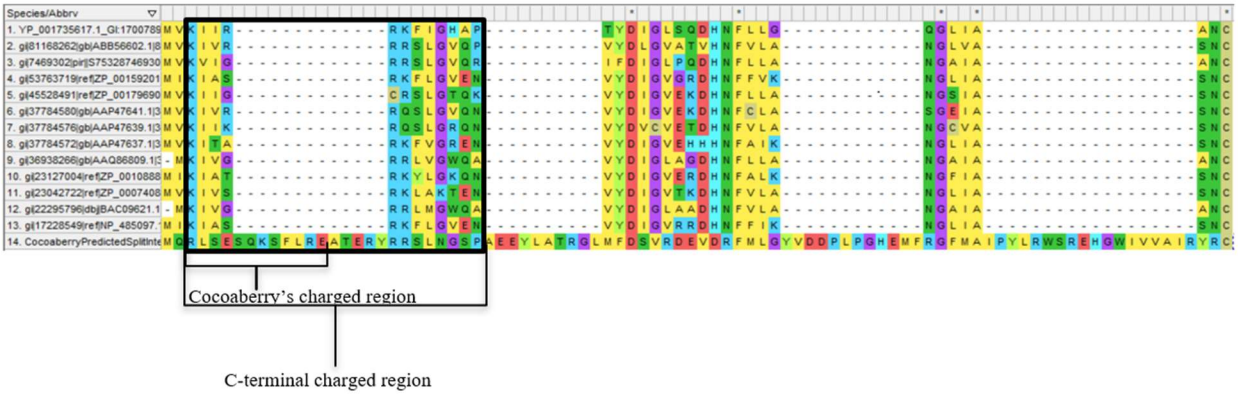


Figure 23: Section of MSA aligned by MUSCLE by codons for collected InBase split intein C-terminal regions compared to the predicted possible C-teriminal split intein of Cocoaberry.

5.5.2 *Most amino acids were not located at the precise location needed.*

Having certain amino acids at a precise location is the second requirement for intein trans-splicing. As previously mentioned, these amino acids are required for the excision of the split intein and the joining of the two flanking exteins. To initiate the process of protein splicing, a cysteine, serine, or threonine needs to be one of the first amino acids in the alignment of the N-terminal intein. In Figure 24, our sequence alignment shows a conserved cysteine as one of the first amino acids for both the collected inteins and Cocoaberry. Although this cysteine is needed, is it not sufficient for the split intein mechanism to occur. Therefore, we took a further look down the alignment as an attempt to find the highly conserved threonine and histidine that help with the cleavage of the N-terminal intein from the N-extein. As highlighted in red in Figure 24, the conserved TXXH sequence was found in all inteins from the database. Yet Cocoaberry's intein did not have the conserved threonine and histidine at that position.

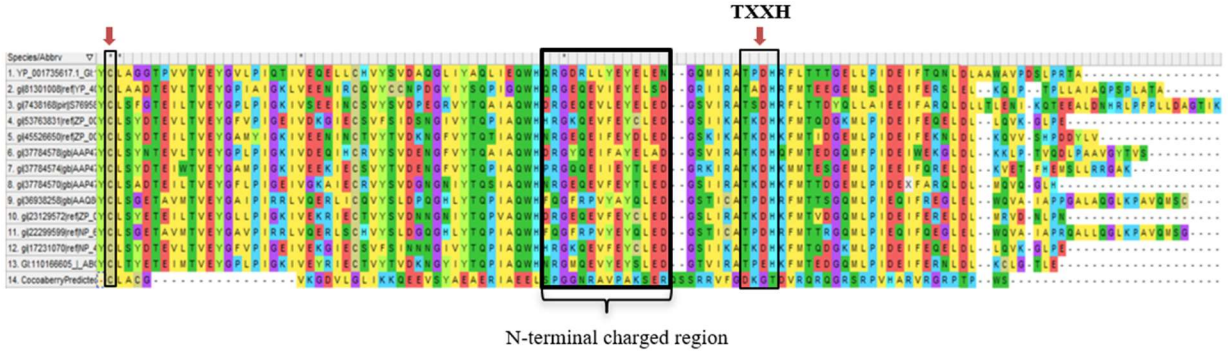


Figure 24: N-terminal inteins alignment with labeled motifs for conserved cystine, negative charged region, and TXXH conserved residues.

To further examine the sequences for split inteins a multiple sequence alignment of the C-terminal intein was performed. The intein alignment showed a highly conserved aspartate (D) from both our database inteins and Cocoaberry (first red arrow in figure 25). This aspartate helps the cystine at the N'end of the N-terminal intein to excise the N-terminal intein from the N-extein (Shah and Muir, 2014). Subsequently we proceeded to look for the other amino acids required for this mechanism. The second amino acid sequence needed is for intein trans-splicing is HNF. Highlighted as the second red arrow in Figure 25, the HNF sequence could be found in all inteins from the database but not in Cocoaberry. In addition, the (A/S)N(C/S/T) conventional sequence was only found in the inteins from the database (third red arrow in Figure 25). Given that only two conserved amino acids were found from our multiple alignment, we can suggest that this is not a probable mechanism happening for Cocoaberry.

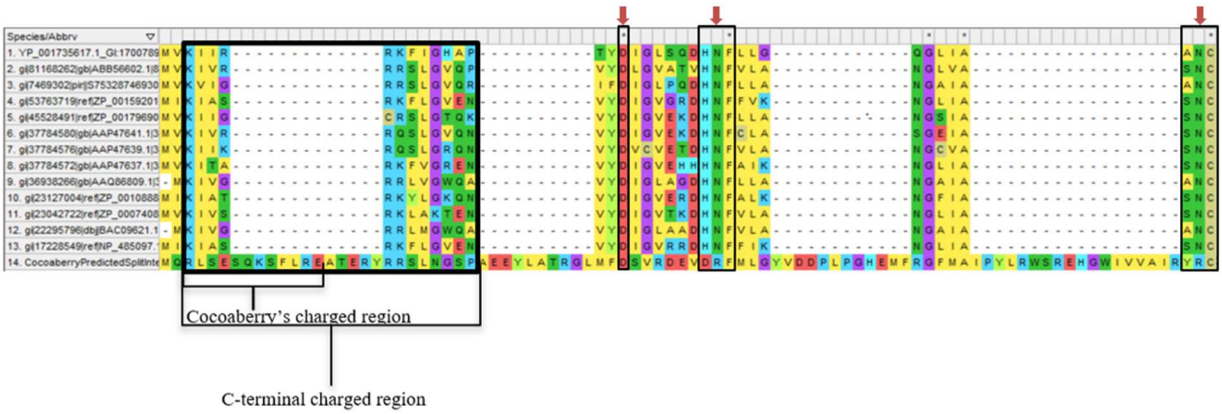


Figure 25: C-terminal split inteins alignment compared to Cocoaberry predicted C-terminal intein region with labeled conserved residues and the positively charged region.

5.5.3 *The zinc-binding domain intein fails to meet the minimal required length.*

Lastly, we calculated the length of the excised split intein. This length will be compared to the length given to us by Pietrokovski. From his data, we expected the minimum N-terminal intein length to range between 60-64 amino acids, while the C-terminal intein must have a maximum length ranging between 25-40 amino acids (Pietrokovski, 1997). Once the multiple alignment was completed for the N-terminal intein, the amount of amino acids present between motif N1 and N3 was counted for all inteins (Figure 26). Results showed that inteins from the database were 72 amino acid long, while cocoaberry's intein only had 58 amino acids.

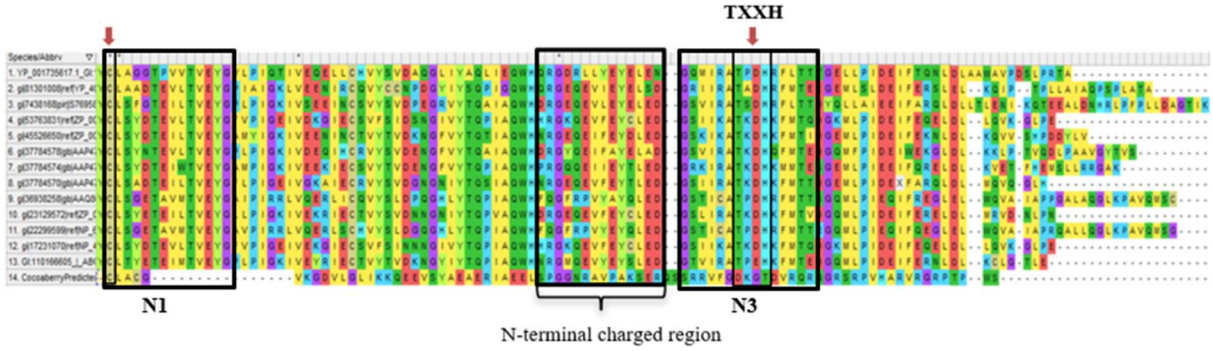


Figure 26: N-terminal inteins alignment with labeled N1 and N3 blocks in addition to motifs for conserved cystine, negatively charged region, and TXXH conserved residues.

In addition to counting the length for the N-terminal intein, we also compared the expected length of the C-terminal intein to our database inteins and Cocoaberry. Alignment showed cocoaberry's intein to be 50 amino acid long, while the intein's from the database were 20 amino acid. This extended length for the C-terminal intein of Cocoaberry does not leave enough room for a functional linker or RPD (Figure 27).

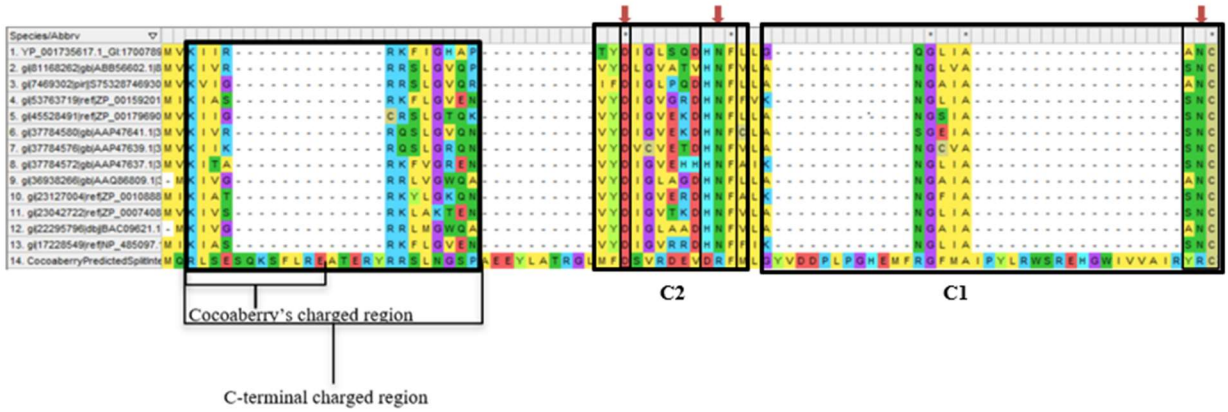


Figure 27: C-terminal split inteins alignment compared to Cocoaberry predicted C-terminal intein region with labeled conserved C2 and C1 blocks as well as residues marked by red arrows and the positively charged region.

Additional supporting evidence of the amount of sequence cut away by the N-terminal split intein region in the ZBD of Cocoaberry is provided through the use of a model from I-TASSER of the resulting combined exteins product. The model showed a decrease in specificity of function from a DNA primase prediction for the translational slippery sequence model, to a DNA binding protein for the exteins predicted model (Figure 28).

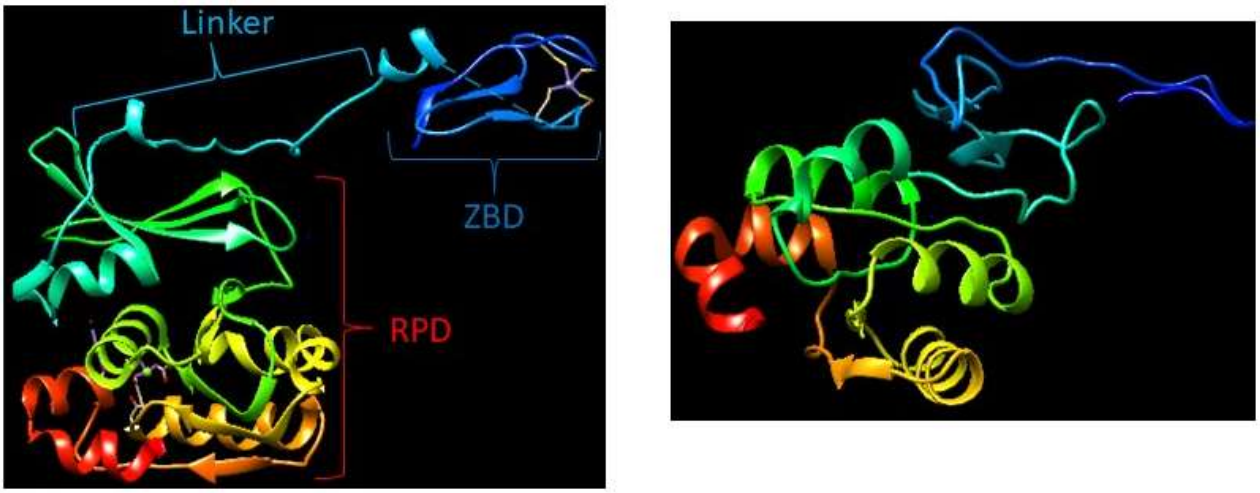


Figure 28: Left: Crystalized T7 bacteriophage DNA Primase (1NUI) with labeled domains from PDB. Right: I-TASSER predicted model for tertiary structure of Cocoaberry DNA primase given the best possible fit intein splicing region flanked by conserved cystines.

5.6 Number of Transcripts

Slippery sequence requires that the entire DNA primase protein be on one RNA strand to allow the slippage to occur. For trans-intein splicing, the conventional model has two RNA strands producing two protein products that can associate and splice themselves into one long functional DNA primase. Therefore, knowing whether there are one or two transcripts can help direct future work. To try and distinguish the difference, an RT-PCR was designed to isolate and amplify the entire coding region for DNA primase using the green and black primers in Figure

29. If a single transcript is made, we would predict that primers for before and after the coding region of the gene would produce a band, while if two transcripts are made, we would predict that those same primers would not produce a band. To verify that phage RNA is isolated, the blue and red primers were used to act as a positive control, since any transcript for either domain will most likely have the area of overlap between the ZBD and the RPD. In addition a set of primers for sigma a, a housekeeping gene in the host, *mycobacterium smegmatis*, were also used, in order to confirm that RNA was being isolated and the experiment was working as expected. Finally, to control for genomic DNA, some samples were run without the use of reverse transcriptase. Once the RNA had been isolated, converted to DNA and amplified through PCR, the resulting DNA was run through a gel to visualize bands for the whole region (929 bp) and overlap region (264 bp) (Figure 29)

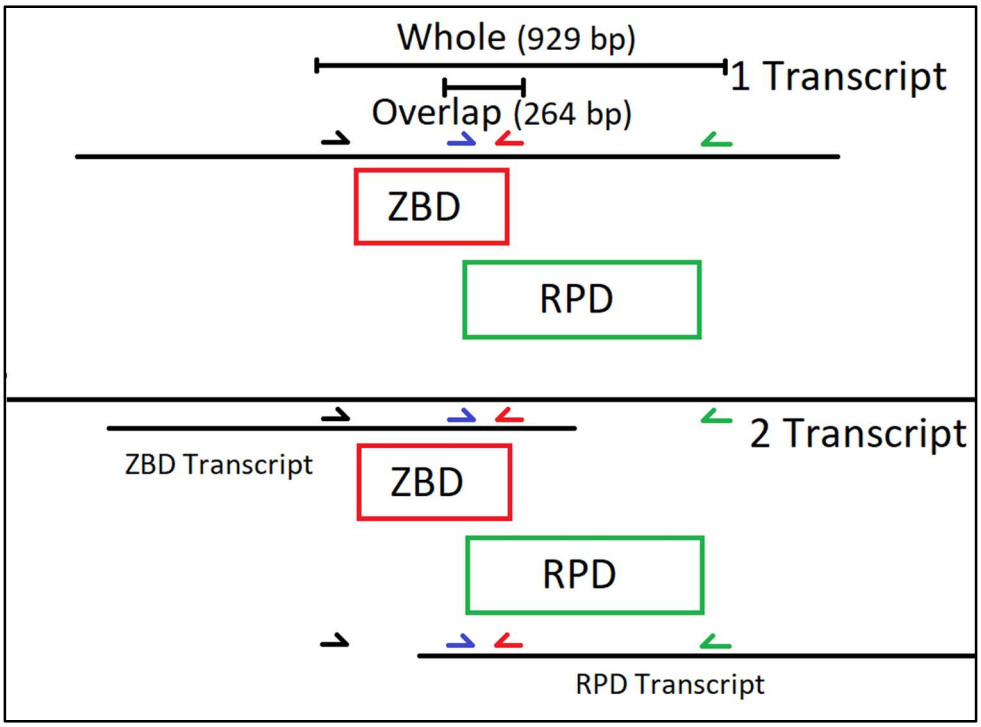


Figure 29: A basic drawing of the area of the genome of Cocoaberry containing DNA primase where different primers (represented by multi-colored arrows) are in reference to the domains of the proteins. The whole and overlap depict the primer groups used for the PCR.

The first attempt at this isolation showed that the primers worked properly, as seen in Figure 30. There are all strong bands at the appropriate length for every primer in the gel. For the Sigma primers, the band is just above 100. The bands for the whole region of the gene are above the 800 bp ladder which is to be expected. The bands for the overlap region are just above the 200 bp ladder rung which is to be expected. However, even though the bands are at the appropriate spot, there are bands in the negative controls. In the lanes that were not treated with reverse transcriptase, there shouldn't be any band as the genomic DNA was lysed and the RNA was degraded. A possible explanation of this is that the DNA lysis step didn't work as intended,

leaving the genomic DNA in these samples. When PCR was run, the primers were able to bind and amplify to create that bands in all the NO lanes when none were expected.

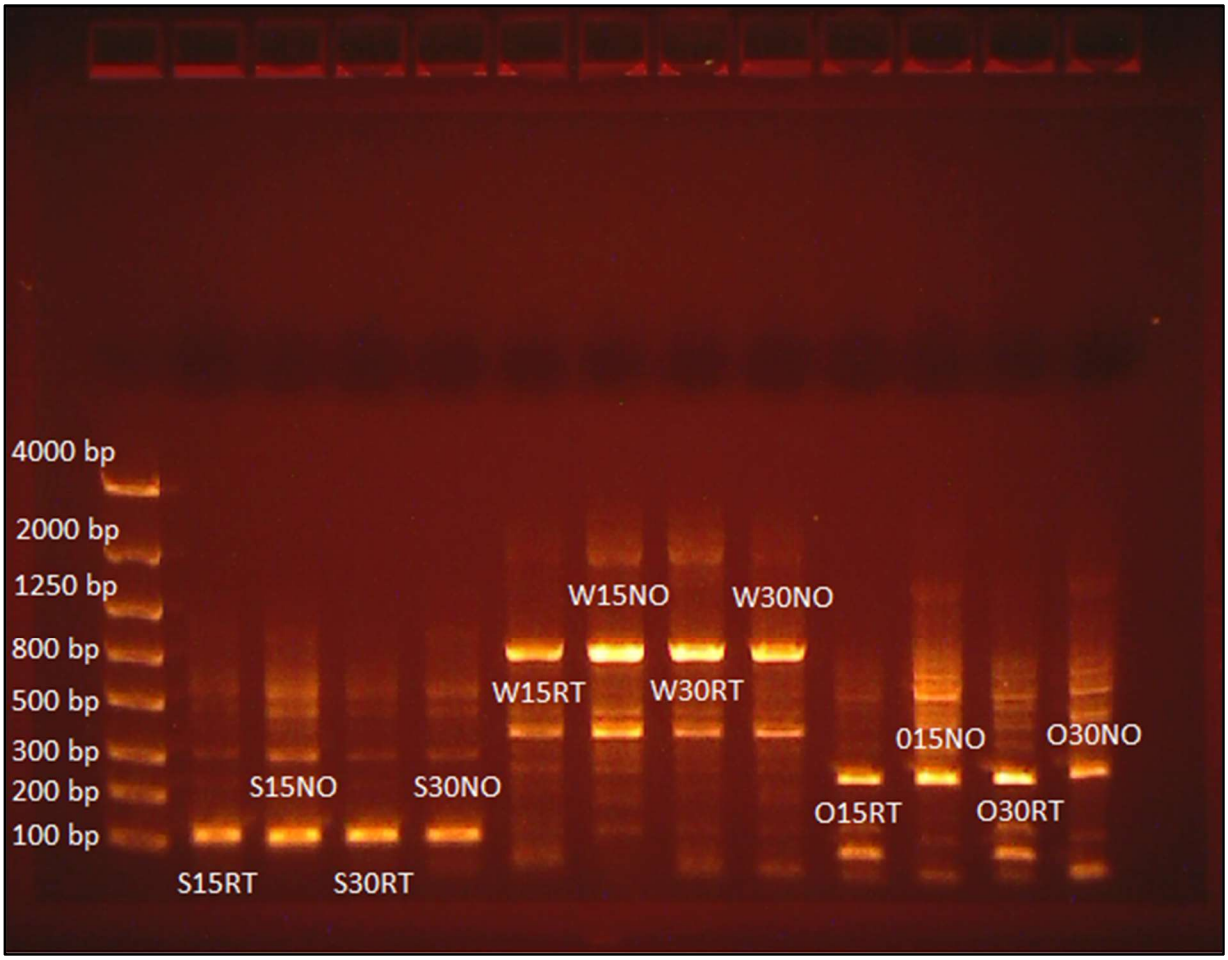


Figure 30: A gel run from the results of the RT-PCR. S denotes Sigma primers, W denotes the primer for the entire DNA Primase gene and O denotes the overlap region set of primers. 15 or 30 refers to the length of time the phage was allowed to infect the M. Smegmatis host. RT means the sample was treated with reverse transcriptase and NO means that it received sterile water instead of reverse transcriptase.

The RNA extraction was rerun using more DNase for longer times with 4 mL of total cell pellet instead of 7 mL of cell pellet to reduce the amount of genomic DNA present. However, the gel negative control lanes still showed bands at the expected lengths, and at just under 500 bp for the whole primer set. This band appeared in every gel run and leads to question where this band is coming from. If this sequence is appearing from the RNA, it could potentially be intron splicing which is exceedingly rare in phage (Pope et al., 2013). The band would need to be isolated and sequenced to determine this and further optimization of the RT-PCR and RNA isolation must be completed to support the phage either having one or two transcripts for this region of the genome.

6. Conclusion

With both computational and experimental methodologies, the three hypotheses, transcriptional slippery sequence, translational slippery sequence and intein splicing, that could explain the formation of a functional DNA primase protein in A4 bacteriophage Cocoaberry were explored. The goals of the project to quantify the extent of this split primase feature were also met to create a foundation of data for use in future projects. The extent of the overlapping region was found to spread among the A, B, and CA clusters of phages in PhagesDB. The translational and transcriptional slippery sequence hypotheses were explored by using multiple sequence alignments of the ZBD genes for phages that have the split primase feature. Conserved regions in the section of the gene overlapping with the RPD gene were examined to see if they matched any canonical slippery sequences from literature. While no exact matches were found, a close sequence of "TTTGGCG" was found to be conserved in A-cluster ZBD genes while meeting our other criteria. Additionally, suitable secondary mRNA structures of a linker and loop structure were found downstream of the transcriptional slippery sequence candidate. Even though this does not support the existence of a slippery sequence since it does not match previously identified sequences, it created a possible avenue for future research to explore this sequence as a possible mechanism.

Additional motifs for the existence of a possible split intein mechanism were also searched for in the sequence alignments. The presence of conserved cystine residues near the end of the ZBD gene and near the middle of the RPD gene met some of the criteria for forming a split intein. Upon further examination however, the lack of conservation of the cystine in the RPD combined with inaccuracies in the charged region motifs, residue conservation, and mismatched lengths of the predicted split inteins provide strong evidence against the mechanism of intein splicing. The strong conservation of the conserved cystine at the end of the ZBD genes found between all clusters in the alignment could warrant further exploration as a part of a mechanism for forming a functional protein. Modelling results for a potential protein following the present criteria produced a protein that bore a resemblance to a crystalized DNA Primase from a T7 phage and was given the description of having DNA binding capabilities.

Finally, the laboratory results from using RT-PCR to determine whether the DNA Primase protein was formed using a single transcript found mixed results. The protocol appears to have isolated the correct length bands from Cocoaberry and the host *Mycobacterium Smegmatis* but the consistent appearance of a band in the gels at the 500 bp mark was unexpected. While these results did not definitively predict whether the DNA primase in Cocoaberry was formed using one or two transcripts due to the negative controls most likely containing genomic DNA, the 500 bp band could be explored in future research to collect definitive supporting evidence for the number of transcripts used. The findings from this study can be used to further explore the functional origin and mechanism of the split DNA primase genes in mycobacteriophages.

References

Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov PV (2016). Ribosomal frameshifting and transcriptional slippage:From genetic steganography and cryptography to adventitious use. *Nucleic Acids Research* 44 (15), 7007-7078. doi: 10.1093/nar/gkw530

Baker D, Sali A (2001). Protein structure prediction and structural genomics. *Science*, 294(5540), 93-96. doi: 10.1126/science.1065659

Baranov PV, Hammer AW, Zhou J, Gesteland RF, Atkins JF (2005). Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. *Genome Biology* 6 (3), 1-9.

Dassa B, Amitai G, Caspi J, Schueler-Furman O, Pietrokovski S, Dassa B (2007). Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations. *Biochemistry (Washington), 46*(1), 322-330. doi:10.1021/bi0611762

Chakravarty S., Godbole S, Zhang B, Berger S. Sanchez R (2008). Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. *BMC structural biology*, 8(1), p.31.

Chirico N, Vianelli A, Belshaw R (2010) Why genes overlap in viruses *Proc Biol Sci*. 277(1701): 3809–3817. doi: 10.1098/rspb.2010.1052

Grose JH, Casjens SR, (2014) Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. *Virology 0* (468-470). 421–443. doi: 10.1016/j.virol.2014.08.024

Hall BG (2018). Phylogenetic trees made easy: A how-to manual (Vol. 5). New York: *Oxford University Press*.

Halleran A, Clamons S, Saha M (2015) Transcriptomic Characterization of an Infection of Mycobacterium smegmatis by the Cluster A4 Mycobacteriophage Kampy. *PLOS ONE* 10(10): e0141100. doi.org/10.1371/journal.,pone.0141100

Huson D. H. and Bryant D. (2006) Application of Phylogenetic Networks in Evolutionary Studies, Mol. Biol. Evol., 23(2):254-267.

Jones DT, Taylort, WR, Thornton JM (1992). A new approach to protein fold recognition. *Nature*, 358(6381), 86.

Jordan TC, Burnett SH, Carson S, Caruso SM, Clase K, DeJong RJ, Dennehy JJ, Denver DR, Dunbar D, Elgin SCR, Findley AM, Gissendanner CR, Golebiewska UP, Guild N, Hartzog GA, Grillo WH, Hollowell GP, Hughes LE, Johnson A, King RA, Lewis LO, Li

W, Rosenzweig F, Rubin MR, Saha MS, Sandoz J, Shaffer CD, Taylor B, Temple L, Vazquez E, Ware VC, Barker LP, Bradley KW, Jacobs-Sera D, Pope WH, Russell DA, Cresawn SG, Lopatto D, Bailey CP, Hatfull GF, Losick R (2013) A Broadly Implementable Research Course in Phage Discovery and Genomics for First-Year Undergraduate Students. *mBio* 5(1):e01051-13. doi:10.1128/mBio.01051-13.

Koonin EV, Wolf YI, Karev GP (2002). The structure of the protein universe and genome evolution. *Nature*, 420(6912), 218. doi: 10.1038/nature01256

Lee S-J, Zhu B, Hamdan SM, Richardson CC, (2010). Mechanism of Sequence-Specific Template Binding by the DNA Primase of Bacteriophage T7. *Nucleic Acids Research,* vol. 38, no. 13, pp. 4372–4383. doi:10.1093/nar/gkq205.

Lee S, Zhu B, Akabayov B, Richardson CC (2012). Zinc-binding domain of the bacteriophage T7 DNA primase modulates binding to the DNA template. *The Journal of Biological Chemistry, 287*(46), 39030-39040. doi:10.1074/jbc.M112.414151

Levin ME, Hendrix RW, Casjens SR (1993). A Programmed Translational Frameshift is Required for the Synthesis of a Bacteriophage lambda Tail Assembly Protein. *Journal of Molecular Biology* 234(1): 124-139: doi:10.1006/jmbi.1993.1568

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuke, M, & Turner DH, (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19), 7287-7292.

Napthine S, Ling R, Finch LK, Jones JD, Bell S, Brierley I, & Firth AE, (2017). Protein-directed ribosomal frameshifting temporally regulates gene expression. *Nature communications*, 8, 15582.

Pedré, X., Weise, F., Chai, S., Lüder, G., & Alonso, J. C. (1994). Analysis of Cis and Trans acting elements required for the initiation of DNA replication in the Bacillus subtilis bacteriophage SPP 1. *Journal of molecular biology*, 236(5), 1324-1340.

Perler FB (2002) InBase, the Intein Database. *Nucleic Acids Res*., 28, 344–345.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE, (2004), UCSF Chimera--a visualization system for exploratory research and analysis. *Comput Chem.* 25(13):1605-12.

Pietrokovski, Shmuel (1994) Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein Science*. 3 (12), 2340–2350.

Pietrokovski S (1997). Modular organization of inteins and c-terminal autocatalytic domains. *Protein Science, 7*(1), 64-71. doi:10.1002/pro.5560070106

Pope WH, Bowman CA, Russell DA, Jacobs-Sera D, Asai DJ, Cresawn SG, Jacobs Jr, W. R., Hendrix RW, Lawrence JG, Hatfull GF (2015) Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *eLife* 4. e06416. doi: 10.7554/eLife.06416

Pope WH, Jacobs-Sera D, Best AA, Broussard GW, Connerly PL, Dedrick RM, Kremer TA, Offner S, Ogiefo AH, Pizzorno MC, Rockenbach K, Russell DA, Stowe EL, Stukey J, Thibault SA, Conway JF, Hendrix RW, Hatfull GF (2013). Cluster J mycobacteriophages: Intron splicing in capsid and tail genes. *Plos One*, 8(7)

Poxleitner M, Pope W, Jacobs-Sera D, Silvanathan V, Hatfull G (2017). Phage Discovery Guide Howard Hughes Medical Institute.

Schwede Torsten, et al., (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research* 31.13 3381-3385. doi: 10.1093/nar/gkg520

Shah NH, Muir TW (2014). Inteins: Nature's gift to protein chemists. *Chemical Science;Chem.Sci., 5*(2), 446-461. doi:10.1039/c3sc52951g

Shah NH, Vila-Perello M, Muir TW (2011). Kinetic control of one-pot trans-splicing reactions by using a wild-type and designed split intein. *Angewandte Chemie, 50*(29), 6511-5. doi:10.1002/anie.201102909

Shapiro JW, Putonti C (2018) Gene Co-occurrence Networks Reflect Bacteriophage Ecology and Evolution. *American Society for Microbiology* 9(2). e01870-17. doi: 10.1128/mBio.01870-17

Teng CP, Zhou T, Ye E, Liu S, Koh LD, Low M, . . . Han M. (2016). Effective targeted photothermal ablation of multidrug resistant bacteria and their biofilms with NIR-absorbing gold nanocrosses. *Advanced Healthcare Materials*, 5(16), 2122-2130. doi:10.1002/adhm.201600346

Wang H, Liu J, Yuet KP, Hill AJ, Sternberg PW (2018). Split cGAL, an intersectional strategy using a split intein for refined spatiotemporal transgene control in caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States, 115*(15), 3900. doi:10.1073/pnas.1720063115/-/DCSupplemental

Webb B, Sali A (2016) Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics* 54, John Wiley & Sons, Inc., 5.6.1-5.6.37.

Wu S, Zhang Y (2008). MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics*, 72(2), 547-556.

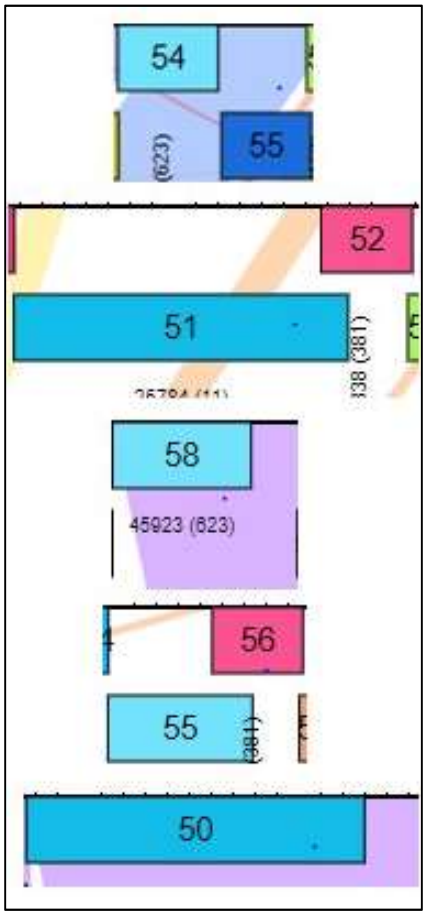
Xu J, Hendrix RW, Duda RL (2004). Conserved Translational Frameshift in dsDNA Bacteriophage Tail Assembly Genes. *Molecular Cell*, 16, 11-21.

Zhang Y (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9: 40

Appendix

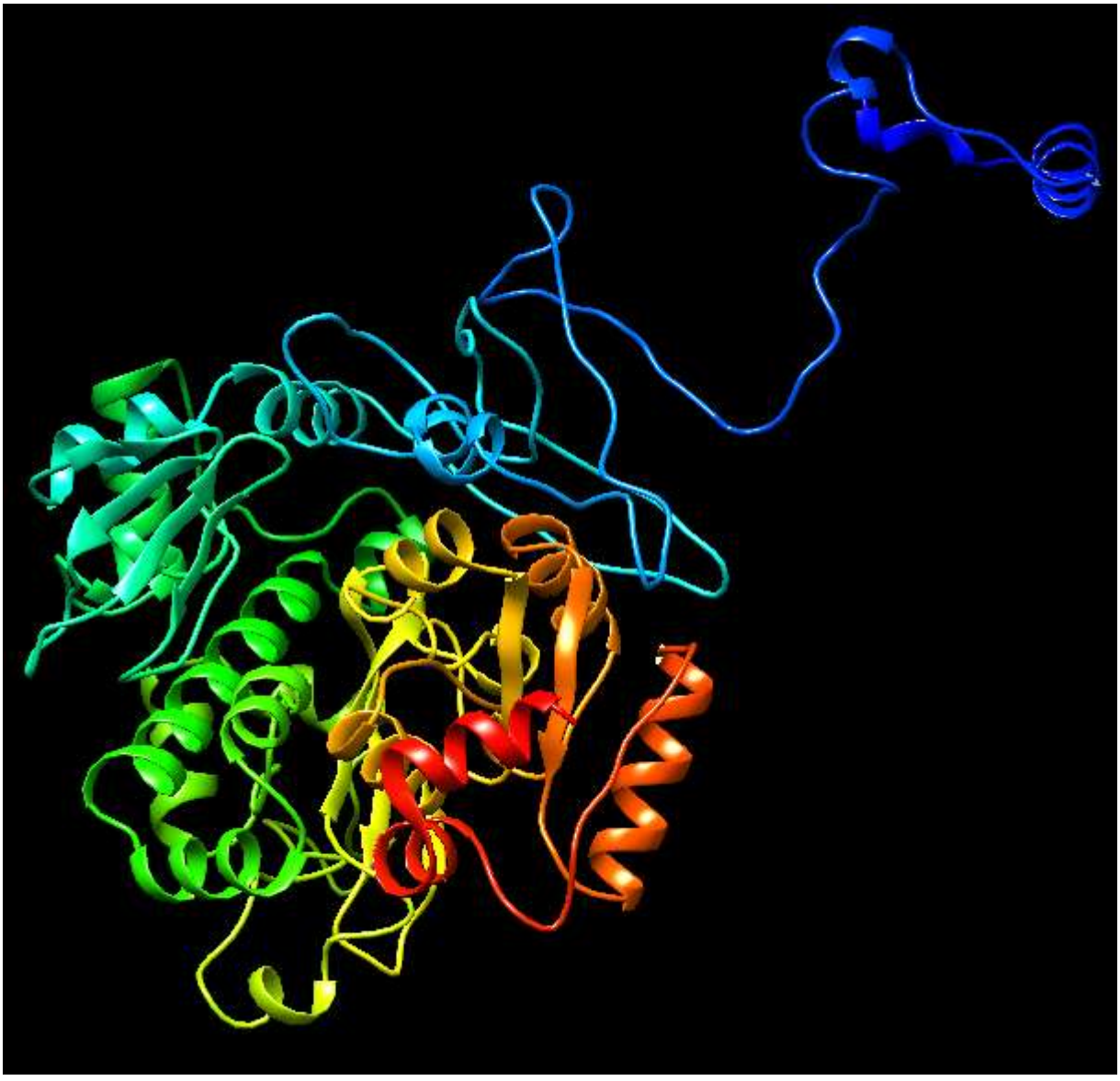
A:

Gene maps image generated from sections of a Phamerator map indicating the five distinct categories that were used to organize the A-cluster phage DNA Primases. From top to bottom: Two gene primase (< 50 bp overlap) (ex. Bethlehem_A1 gp54 and gp55), Long primase gene plus ZBD gene (ex. Fascinus_A1 gp51 and gp52), short RPD with no ZBD called (ex. MissWhite_A2 gp58), two gene primase *analogous to the A4 split primase feature (> 50 bp overlap) (ex. Cocoaberry gp55 and gp56), and Single gene DNA primase (ex. Bigfoot_gp50)



B:

I-TASSER generated structure of Bigfoot_gp50 single gene DNA primase. ZBD region begins in the blue colored region until the beginning of the cyan color. The predicted linker extends from the cyan color to the beginning of the turquoise. The RPD region begins at the turquoise region and extends to the end of the protein marked in red.



C:

Results of A1_CA cluster alignment used to create a phylogenetic network in SplitsTree. Full results of the details for this tree can be shown by viewing the tree in SplitsTree using the Supplementary File: "RPDsA_CA.nexus".