# BIOINFORMATIC AND *IN VITRO* CHARACTERIZATION OF PRIMASE-POLYMERASE ENZYMES FROM VIRUSES THAT INFECT ACTINOBACTERIAL HOSTS

A thesis presented to the faculty of the Graduate School of
Western Carolina University in partial fulfillment of the
requirements for the degree of Master of Science in Chemistry.

By

Nathan Benjamin Folse

Director: Dr. Jamie Wallen
Assistant Professor of Biochemistry
Chemistry and Physics Department

Committee Members: Dr. Maria Gainey, Biochemistry
Dr. Carmen Huffman, Chemistry

April 2020

ACKNOWLEDGEMENTS

I would like to thank Dr. Jamie Wallen, whose enthusiasm for discovery is pure and infectious.

> *If he is indeed wise he does not bid you enter the house of his wisdom, but rather leads you to the threshold of your own mind.*
>
> - Kahlil Gibran, 'The Prophet'

I would also like to express my sincerest gratitude to Dr. Maria Gainey for diligence, to Dr. Carmen Huffman for guidance, to Dr. Bill Kwochka for playfulness, to Dr. Alesia Jennings for joy, to Dr. Al Fischer for patience, to Wes Bintz for gloves and things, to the Western Carolina University Chemistry and Physics Department for community, to the incredible faculty for inspiration, to all my fellow students for friendship, to the mountains for a home, to my twelve year-old MacBook for everything.

And to Ashley and Leo. For doing the most important work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| prim-pol | DNA primase-polymerase enzyme |
| AEP | archaeo-eukaryotic primase |
| dsDNA | double-stranded DNA |
| ssDNA | single-stranded DNA |
| NTP | ribonucleoside triphosphate |
| dNTP | deoxyribonucleoside triphosphate |
| ORF904 | prim-pol gene orf904 from *Sulfolobus islandicus* archaeal plasmid pRN1 |
| NrSp | prim-pol gene 28 from Nitratiruptor phage NrS-1 |
| PrimPol | prim-pol encoded by human gene |
| aa | amino acid |
| AAA | ATPases Associated with various cellular Activites |
| ATP | adenosine triphosphate |
| SF3 | superfamily 3 (helicases) |
| SF1 | superfamily 1 (helicases) |
| SF2 | superfamily 2 (helicases) |
| Larva64 | prim-pol gene 64 from mycobacteriophage Larva |
| kb | DNA kilobase |
| pham | family of closely-related phage proteins |
| CDD | Conserved Domain Database |
| DUF | domain of unknown function |
| MCM | minichromosome maintenance protein |
| BLAST | Basic Local Alignment and Search Tool |
| N280 | Larva64 model comprises first 280 N-terminal residues |
| ssM13 | single-stranded M13 DNA |
| SD | strand-displacement synthesis |
| dsM13 | double-stranded M13 DNA |
| bp | DNA base pair |
| FAM | fluorescein amidite |
| PPI | protein-protein interaction |
| 2-IPM | 2-isopropylmalate synthase |
| CRISPRi | CRISPR interference |
| dCas9 | ‘dead’ Cas9 protein |
| SAXS | small-angle X-ray scattering |
| PRS | primase recognition site |
| ddNTP | dideoxynucleoside triphosphate |

# ABSTRACT

BIOINFORMATIC AND *IN VITRO* CHARACTERIZATION OF PRIMASE-POLYMERASE ENZYMES FROM VIRUSES THAT INFECT ACTINOBACTERIAL HOSTS

Nathan Benjamin Folse

Western Carolina University (April 2020)

Director: Dr. Jamie Wallen

Primase-polymerases (prim-pols) are enzymes that exhibit primase, polymerase, and potential helicase-like activities by way of a bifunctional N-terminal prim-pol domain and a C-terminal ATPase domain. Presented is a multifaceted analysis of prim-pols encoded by actinophages, or viruses that infect Actinobacterial hosts. The aims of this study are to bioinformatically characterize all identifiable actinophage-encoded prim-pols and to biochemically characterize the prim-pol encoded by mycobacteriophage Larva, including determining protein-protein interactions between Larva's prim-pol and host *Mycobacterium smegmatis* proteins using a bacterial two-hybrid system. Bioinformatic analyses reveal nearly 600 actinophages encoding prim-pols that span a variety of host types, genome sizes, conserved domains, and encoded genetic metabolism proteins. A novel class of truncated prim-pols that contain an intact prim-pol domain but lack any additional C-terminal domain has been identified. Most phages encoding truncated prim-pols also encode a separate protein resembling the ATPase domain of full-length prim-pols. Interestingly, the C-terminal functional domains of full-length prim-pols vary from phage to phage aside from a conserved nucleotide binding motif. The C-terminal domains of prim-pols allow them to be grouped more narrowly than by their N-terminal prim-pol domain alone. To begin to understand the importance of prim-pols in actinophage DNA replication, a

detailed characterization has been performed of a prim-pol from mycobacteriophage Larva, which is the virus' only encoded DNA polymerase. Larva's prim-pol exhibits $Mg^{2+}$-dependent primase/polymerase activity on an unprimed ssDNA substrate in the presence of dNTPs. It also binds multiple DNA substrates and translocates on ssDNA in the presence of ATP. CRISPR interference silencing of prim-pol is lethal to Larva, indicating it is essential for viral survival. Bacterial two-hybrid analysis reveals interactions between Larva's prim-pol and at least five *M. smegmatis* proteins, including transcription and nucleic acid synthesis proteins. Based on these results, it is hypothesized that despite sharing a conserved prim-pol domain, actinophage-encoded prim-pols fulfill a variety of functions in the replication of phage DNA, some critical to viral survival, depending on the domain organization of the proteins' C-termini and on the phages' individual genomic architectures.

# CHAPTER ONE: INTRODUCTION

Primase-polymerases (prim-pols) are enzymes that exhibit primase, polymerase and ATPase activities by way of a bifunctional N-terminal prim-pol domain and a C-terminal ATPase domain. They are considered part of the archaeo-eukaryotic primase (AEP) superfamily due to the presence of a distinctive palm-domain composed of antiparallel β-strands[1] that houses the prim-pol active site. DNA replication commonly requires three separate enzymes: a helicase, a primase, and a polymerase (Figure 1). Double-stranded DNA (dsDNA) is separated by the helicase to yield two strands of single-stranded DNA (ssDNA). Primases will then incorporate ribonucleoside triphosphates (NTPs) to synthesize a short complement strand of RNA, called an RNA primer, on each ssDNA strand. Finally, DNA polymerases will incorporate deoxyribonucleoside triphosphates (dNTPs), starting at the 3'-end of the RNA primers, to extend them into a full-length DNA complement and yield two identical strands of dsDNA. Most DNA polymerases require the free 3' hydroxyl group provided by an RNA primer to initiate synthesis of a new DNA strand.[1]

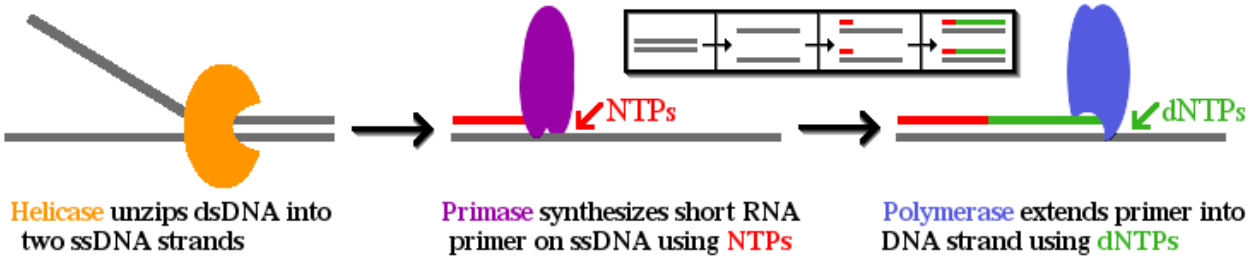


Figure 1. Schematic of typical DNA replication.

An important note about prim-pol proteins is that, despite containing a bifunctional prim-pol domain at their N-terminus, they also possess varied domains at the C-terminus that are not

distinguishable by the term "prim-pol." This current method of nomenclature imposes limitations on our ability to describe different and unique enzymes that contain the prim-pol domain.

Prim-pols are unique among polymerases in that they are able to initiate *de novo* DNA synthesis; that is, they are capable of first assembling a primer on an unprimed substrate and then synthesizing a complementary strand to yield dsDNA without the aid of additional replicative enzymes. They are also distinct among primases in that they preferentially incorporate dNTPs, rather than NTPs, to synthesize a DNA primer directly. The first described prim-pol originates from the open reading frame 904 gene (ORF904) of the archaeal plasmid pRN1 from *Sulfolobus islandicus* and was characterized in 2003.[2] ORF904 contains an N-terminal prim-pol domain and C-terminal D5N, primase-like, and winged-helix DNA binding domains and has demonstrated limited helicase activity at its C-terminus.[3] ORF904 is part of the minimal replicon of pRN1 and is responsible for recognizing the plasmid's origin of replication.[4]

Like many AEP family proteins, ORF904 binds zinc, and it contains a stem of zinc binding residues that protrudes above its prim-pol active site. The primase and polymerase activities of ORF904 observed *in vitro* are $Mg^{2+}$-dependent, and full primase activity requires the prim-pol domain, the zinc stem, and an additional helix bundle domain. Two zinc ions remain tightly bound to the prim-pol domain through purification and crystallizaiton.[5] In contrast, zinc binding domains are not found in prim-pols encoded by bacteriophages, which are viruses that infect bacteria. The first crystal structure of a phage-encoded prim-pol—from gene 28 of the deep-sea Nitratiruptor phage NrS-1 (NrSp)—does not contain any identifiable zinc binding domain but exhibits $Mg^{2+}$-dependent prim-pol activity nonetheless. It also contains similar conserved domains to ORF904.[6,7]

Interest in prim-pols has grown considerably since the 2013 discovery of a human-encoded prim-pol (PrimPol) occupying both the nucleus and mitochondrion. PrimPol was the second primase identified in human cells, as well as the second polymerase to be found within the mitochondrion after DNA Polγ.[8] It is thought to work *in vivo* chiefly in tandem with Polγ (mitochondria) and Polε (nucleus) to restart stalled replication forks by repriming the template DNA strand past the point of a lesion or other obstruction.[9] PrimPol has been observed *in vitro* to exhibit a wide variety of activities including DNA and RNA priming, DNA polymerization, trans-lesion synthesis, lesion bypass by template scrunching, and even non-template terminal transferase activites.[10] Although a fascinating study, the human enzyme shares little overall homology with ORF904 or phage prim-pols and operates by somewhat different mechanisms. PrimPol is 560 amino acids (aa) long and does not contain any additional C-terminal domains. Its prim-pol domain is roughly twice the length of ORF904's despite containing the same types of functional motifs.[8] However, like ORF904, PrimPol depends on zinc. Its zinc binding domain adopts a finger-like structure, rather than the type of rigid zinc stem found in ORF904, and forms a flexible protrusion that is important for substrate recognition and stabilization of PrimPol during DNA priming.[11] Still, most prim-pols exist as multidomain enzymes. The bifunctional prim-pol domain invariably sits at the N-terminus but is fused to a C-terminus containing varying ATPase domains from one prim-pol to the next. The variety of C-terminal domains, and the interactions between the N- and C-termini, of prim-pols is a fascinating but under-researched topic.

Prim-pol C-termini most broadly classify as P-loop NTPases. P-loop proteins are named as such for a characteristic α-helix/β-strand fold where hydrolysis of the phosphate groups in NTPs—most often ATP—occurs. The Walker A and B motifs are short amino acid sequences

linked to phosphate binding by proteins and are ubiquitous among active P-loop NTPases. Most proteins that hydrolyze nucleotides contain a Walker A motif, represented canonically as [G/A]xxxxGK[S/T] where 'x' represents any amino acid. The lysine is present in all known permutations of the Walker A and is responsible for binding the α and γ phosphate groups of ATP. Walker A motifs have been located in the C-terminus of every full-length prim-pol in which one has been sought. Walker B motifs follow the pattern hhhhDE, where 'h' represents any hydrophobic residue, but exhibit much greater sequence variation than the Walker A and can be difficult to identify without structural data. The aspartic acid is generally responsible for coordination of the $Mg^{2+}$-ATP complex, while the glutamic acid acts in a catalytic role.[12]

AAA+ (ATPases Associated with various cellular Activities) superfamily proteins are P-loop-containing proteins that catalyze the hydrolysis of ATP in order to perform various types of mechanical work.[13] The superfamily 3 (SF3) helicases are a subset of AAA+ proteins, typified by the papillomavirus E1 helicase, that are found almost exclusively in small DNA and RNA viruses.[14] SF3 domains are found fused to the prim-pol domain of ORF904 and other prim-pols.[2] SF3 helicases are diverse in their domain structure and in sequence, and translocate 3' to 5' along DNA. Their primary function is the binding and melting of dsDNA, which is accomplished through the formation of a hexameric ring through which DNA strands are passed as the hydrolysis of ATP drives movement of adjacent monomers relative to one another. Many SF3 helicases have an α-helical domain at their N-terminus that appears to be important for oligomerization. All contain Walker A and B motifs and they tend to exist as part of larger origin-binding proteins.[14] The C-terminal structure of NrSp was recently solved as a homohexamer, with multiple NTPase and unwinding activities that resembles SF3 helicases.[15]

Conversely, RecA is an ATP-dependent homologous DNA repair protein, typified by the extensively studied RecA protein of *Escherichia coli*. RecA domains are also commonly found fused to prim-pols. Proteins in the RecA class are P-loop NTPases but are not helicases and do not belong to the AAA superfamily. However, superfamily 1 and 2 (SF1 and SF2) helicases share significant homology with RecA's core parallel β-sheet domain that contains the Walker A and B motifs.[16] The Walker B motif marks a point of divergence between the cores of SF1/SF2 helicases and RecA. While the former two have an intact DE motif, the latter contains the aspartic and glutamic acid residues on neighboring β-strands. The small, DnaB helicase family shares this distinctive Walker B arrangement with RecA.[16] DnaB proteins translocate 5' to 3' on DNA and encompass recombinases and hexameric helicases such as RepA.[17] RecA proteins are generally involved in DNA repair processes,[18] but have also been implicated in DNA-dependent ATP hydrolysis, regulating their own expression,[19] and even shielding of dsDNA from damage by ionizing radiation in *Deinococcus radiodurans*.[20]

The most broadly significant role of RecA is its ability to catalyze DNA repair by recognizing and joining complementary ssDNA overhangs of damaged dsDNA to initiate homologous recombination. This function is achieved largely by 5' to 3' directional polymerization along a DNA strand, rather than translocation;[18] RecA proteins have the interesting property of self-assembling on DNA, in the presence of ATP, into lengthy helical nucleoprotein filaments composed of tightly packed monomer units.[21] RecA's propensity for self-assembly is highly sensitive to type and concentration of salt and has led to the observation of numerous oligomeric states *in vitro*, both on and off of DNA, including rings, rods, filaments, and disorderly aggregated masses.[19] In a notable similarity to the role of human PrimPol, RecA has also been found to aid in the reinstatement of stalled replication forks.[18]

Prim-pols are found in a number of bacteriophages. Bacteriophages are the world's most abundant biological entities and constitute an immensely rich genomic landscape filled with genetic diversity and novel proteins.[22] Research into phages has gained increasing relevance with the advent of genomics and, in particular, as a means to combat the increasingly urgent public health threat of antibiotic-resistant bacteria.[23] Insight into virus-host interactions and the mechanisms by which phages combat or circumvent endogenous bacterial immune responses afford the scientific community a deeper understanding of the physiology of bacteria themselves. The logic of utilizing the biological tools employed by the natural predators of bacteria to combat infections, rather than depending on the development of new and inherently short-term pharmaceutical solutions, has become increasingly evident as our understanding of phages deepens.[23]

Actinophages are viruses that infect bacteria belonging to the phylum Actinobacteria. Despite the fact that many actinophages contain genes annotated as prim-pols based on domain conservation, the prim-pols from this massively variegated grouping of viruses have yet to be studied in detail en masse. To date, the only actinophage prim-pol to be biochemically characterized is the truncated N-terminus of the prim-pol encoded by corynephage BFK20.[24] A full-length actinophage-encoded prim-pol has not previously been expressed and characterized. Mycobacteriophages are of particular interest among actinophages because they infect Mycobacterial hosts such as *Mycobacterium tuberculosis* and *Mycobacterium leprae*, the causative agents of tuberculosis and leprosy, respectively.[25] Mycobacteriophage Larva was isolated from *Mycobacterium smegmatis* $mc^2155$ in Williamsburg, VA and is a temperate phage of the Siphoviridae morphotype that belongs to cluster K5 (www.phagesdb.org).[26] Larva's gene 64 (Larva64) encodes a putative prim-pol. Among phage-encoded prim-pols, Larva64 is an

intriguing candidate for *in vitro* characterization because it is the virus' only encoded DNA polymerase and therefore is the sole source of Larva's endogenous genetic replication abilities. Larva64 contains an N-terminal prim-pol domain and a C-terminal ATPase domain with homology to RecA and similar recombinatorial proteins.

The aims of this study are to bioinformatically characterize all identifiable actinophage-encoded prim-pols and to biochemically characterize the full-length prim-pol encoded by mycobacteriophage Larva. To those ends, presented is a comprehensive bioinformatic exploration of the sprawling array of actinophage-encoded prim-pols, which emerge as diverse in residue conservation, domain organization, genomic environment, and even putative function. Also presented is an *in vitro* characterization of the purified primase-polymerase from phage Larva that reveals an active and multifunctional enzyme. These results provide the first biochemical analysis of a full-length actinophage-encoded prim-pol.

# CHAPTER TWO: RESULTS OF BIOINFORMATIC ANALYSIS

## Sequence-Based Investigation of Actinophage Prim-Pols

The Phamerator database used to extract sequences is constantly updated, and the bioinformatic data used for this analysis was extracted on 05/14/2019. Actinophage clusters grow as new specimens are discovered and characterized. As a result, the present number of phage-encoded prim-pols is likely larger than what is reported here. Cluster names of phages that appear have been updated to reflect any reclassification since extraction. Gordonia phages Ghobes and Reyja were both classified as singletons when data were collected but have since been assigned to clusters DA and DY, respectively. Cluster B has seen the addition of subclusters B9 - B13, containing phages that were classified only as cluster B before. Phages that now belong to subcluster B9 were not yet characterized when data were collected, but they contain a prim-pol gene belonging to the same pham—family of closely related phage proteins—as other cluster B phages. At the time of writing all phages that have been newly characterized since data was pulled, and that belong to a cluster included in this study, contain a prim-pol gene belonging to the same pham as other cluster members.

### Meta-Analysis Findings

Bioinformatic meta-analysis of putative actinophage-encoded prim-pols reveals a highly diverse class of enzymes. 569 total prim-pol genes were identified from actinophages spanning 28 clusters (47 subclusters), including 5 singletons. Phages encoding prim-pols infect hosts belonging to the genera *Arthrobacter*, *Corynebacterium*, *Gordonia*, *Microbacterium*, *Mycobacterium*, *Rhodococcus*, *Streptomyces*, and *Tsukamurella*, with the majority (400) infecting *Mycobacterium*. The imbalance of host types, however, is the direct result of sampling

bias; bacteriophage discovery and sequencing efforts focused almost exclusively on mycobacteriophages until recently. Protein lengths range from 198-1085 aa, while overall genome sizes of the phages range from 40-122 kilobases (kb). No apparent correlation exists between the length of prim-pols and the overall genome size of phages that code for them. Prim-pols are ubiquitous in almost all clusters in which they are found. The only exceptions to this trend are clusters EA7 and K5; prim-pols are not known to be encoded by any other EA or K subclusters.

The general domain architecture of actinophage prim-pols agrees with that of previously characterized prim-pols, with a bifunctional prim-pol domain at the N-terminus and an ATPase domain at the C-terminus. In contrast to the prim-pol domain, which is reliably conserved according to Conserved Domain Database (CDD)[27] predictions, the C-terminus exhibits substantial diversity from phage to phage. Significantly, close examination of the CDD and Phyre2[28] homology results for each subcluster reveals that actinophage prim-pol genes can be organized into four distinct classes based on the homology of their C-terminal domains: RecA-like, SF3-like, truncated, and unknown/special cases. Truncated prim-pols are proteins that entirely lack a C-terminus beyond the prim-pol domain, and the unknown classification is a catch-all for those prim-pols with unusual homology or no predictable domain at their C-termini. Overall findings in Table 1 are organized by C-terminal class and further divided by cluster and subcluster. The identification of truncated prim-pols in phages is of special interest; although human PrimPol also lacks an additional C-terminal domain, it is twice the size of truncated phage prim-pols, Zn-dependent, and shares virtually no homology with the prim-pols examined here.

Table 1. Overall findings of actinophage sequence meta-analysis.

| C-Term Class | Phage Subcluster | # Encoding Prim-Pol | Average Protein Length (aa) | Host Genus |
|---|---|---|---|---|
| RecA-like | B1 | 198 | 926 | Mycobacterium |
| | B2 | 27 | 914 | |
| | B3 | 31 | 867 | |
| | B4 | 18 | 927 | |
| | B5 | 7 | 968 | |
| | B6 | 5 | 970 | |
| | B7 | 1 | 922 | |
| | B8 | 1 | 896 | |
| | B10 | 1 | 958 | |
| | B11 | 1 | 952 | |
| | B12 | 1 | 948 | |
| | B13 | 1 | 998 | |
| | DR | 6 | 860 | Gordonia |
| | K5 | 14 | 720 | Mycobacterium |
| | Singleton | 3 | 850 | Gordonia Tsukamurella |
| | **4 Clusters** | **315 Total** | **910 Average** | **3 Host Types** |
| Unknown / Special Cases | AK | 71 | 877 | Arthrobacter |
| | BH | 8 | 887 | Streptomyces |
| | DA | 1 | 852 | Gordonia |
| | DY | 1 | 772 | Gordonia |
| | EA7 | 1 | 887 | Microbacterium |
| | ED1 | 10 | 820 | Microbacterium |
| | ED2 | 3 | 867 | |
| | EJ | 4 | 901 | Microbacterium |
| | O | 15 | 787 | Mycobacterium |
| | Unknown* | 1 | 890 | Microbacterium |
| | **8 Clusters** | **115 Total** | **860 Average** | **5 Host Types** |

| C-Term Class | Phage Subcluster | # Encoding Prim-Pol | Average Protein Length (aa) | Host Genus |
|---|---|---|---|---|
| SF3-like | AC | 3 | 908 | Mycobacterium |
| | CR1 | 4 | 853 | Gordonia |
| | CR2 | 11 | 852 | |
| | CR3 | 3 | 853 | |
| | CR4 | 7 | 849 | |
| | CR5 | 2 | 844 | |
| | D1 | 18 | 980 | Mycobacterium |
| | D2 | 1 | 973 | |
| | DG | 4 | 955 | Gordonia |
| | EN | 8 | 847 | Corynebacterium |
| | FA | 5 | 837 | Arthrobacter |
| | H1 | 7 | 983 | Mycobacterium |
| | H2 | 2 | 917 | |
| | R | 7 | 916 | Mycobacterium |
| | U | 2 | 902 | Mycobacterium |
| | Singleton | 2 | 849 | Rhodococcus Corynebacterium |
| | **10 Clusters** | **86 Total** | **904 Average** | **5 Host Types** |
| Truncated | DK | 2 | 439 | Gordonia |
| | DS | 2 | 435 | Gordonia |
| | DU | 1 | 238 | Gordonia |
| | EK1 | 5 | 273 | Microbacterium |
| | EK2 | 2 | 274 | |
| | EM | 2 | 256 | Microbacterium |
| | J | 37 | 231 | Mycobacterium |
| | X | 2 | 235 | Mycobacterium |
| | **7 Clusters** | **53 Total** | **253 Average** | **3 Host Types** |

*Not tallied as individual cluster

The relative levels of sequence agreement shown in Figure 2 demonstrate the value of dividing these prim-pols into separate classes. The farther the colored region reaches toward the upper boundary the more conserved the residues are at that point, and points of contact with the upper boundary represent 100% consensus. Individual alignments were generated for each C-terminal class and calibrated to the meta-alignment at the DxD prim-pol active site motif and the Walker A lysine, due to the high level of agreement at these sites in all alignments. The truncated prim-pols were oriented using the DxD motif and a highly conserved proline downstream of the prim-pol active site, since they lack a Walker A.

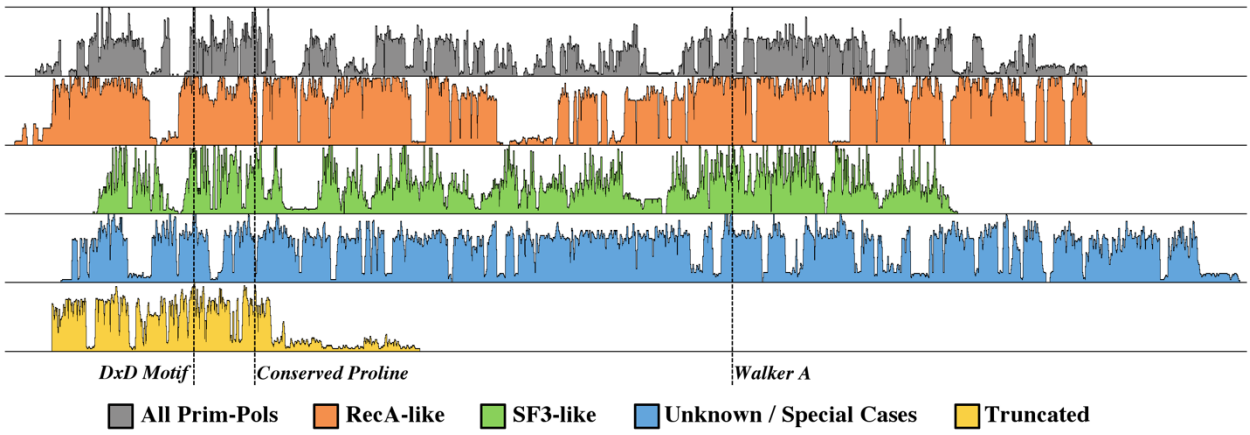


Figure 2. Sequence agreement in meta-alignment of prim-pols vs. individual C-terminal classes.

A few critical residues emerge as consistently present when all 569 prim-pol sequences are compared in a single meta-alignment. Overall consensus is low though, and the Walker A lysine is virtually the only point of agreement at the C-terminal end. When the prim-pols of each C-terminal class are aligned only with one another, the level of agreement increases dramatically across the length of the protein. Unsurprisingly, RecA-like prim-pols, the vast majority of which are present in the closely related cluster B phages, show the highest levels of consensus from start to finish. SF3-like prim-pols demonstrate a significantly higher number of residues with total and near-total consensus compared to the meta-alignment, and the overall levels of consensus among truncated and unknown prim-pols are also greatly improved. The extended region of low consensus between the conserved proline and Walker A in the meta-alignment is also essentially absent from class-based alignments.

**Four Classes of Actinophage Prim-Pols**

RecA-like prim-pols have AAA_25 (pfam13481), RepA (cd1125), and radB (TIGR02237 and PRK09361) putative conserved domains, and their top Phyre2 matches are to the ATPase domain of *E. coli* plasmid RSF1010 RepA (a DnaB hexameric helicase),[17] *M. smegmatis* RecA,[29] and *D. radiodurans* RecA.[20] This class contains singletons Tsukamurella

phage TPA2 and Gordonia phages Pine5 and REQ1. The conserved AAA domain is fairly non-specific, hinting only at the presence of a P-loop, while RepA/radB are RadB (RecA-like SF2) helicase domains. Outliers among RecA-like prim-pols are cluster K5, which matches typical RecA-like homologs by Phyre2 modelling but contains only the AAA_25 conserved domain, and REQ1. REQ1 prim-pol is the only actinophage prim-pol in a pham by itself, and some of its top Phyre2 matches are to the RecA homologs RadA (archaea)[30] and Rad51 (eukaryotes)[31] as helical homoheptamers. The C-termini of RecA-like prim-pols are expected to bind DNA and participate in DNA repair/recombination or to act as a replicative helicase. They likely translocate or polymerize along DNA in a 5' to 3' direction. They may also hexamerize as is seen with RSF1010 RepA.[17]

SF3-like prim-pols have primase_Cterm (TIGR01613), COG3378, and D5N (pfam 08706) conserved domains, and their top Phyre2 matches are to the AAA+ domain of Rep40,[32] multiple matches to papillomavirus E1 helicase,[33,34] the winged-helix protein hRFX1 DNA binding domain,[35] and the C-terminus of NrS-1 prim-pol.[15] This class contains singletons Rhodococcus phage ChewyVIII and Corynebacterium phage P1201. The primase_Cterm domain is implicated in DNA replication, recombination, and/or repair. COG3378 and D5_N are both domains found in primase-helicases with 3' to 5' directionality, as represented by *E. coli* phage P4 gpα.[36] SF3-like prim-pols are expected to have helicase-like properties, form a hexamer around a DNA strand, and translocate on DNA with 3' to 5' directionality.

The conserved domain predictions and Phyre2 homologs of RecA-like and SF3-like prim-pols are more or less mutually exclusive, leading to a surprisingly definitive separation of the two. Because so many similarities exist between the core domains of RecA and SF3 type proteins, the structural and biochemical differences between the classes are difficult to predict

without more rigorous individual analyses. Unknown and truncated prim-pols are less stringently defined and so lower levels of similarity are expected between them.

Prim-pols classified as unknown or special cases are full-length prim-pols that did not have any putative conserved domains at their C-termini, or only domains of unknown function (DUF) without a clear pattern. Most also failed to return any high-confidence (>90%) homologs from Phyre2 beyond the crystal structures of ORF904 and NrSp confirming the presence of a prim-pol. Some internal consistencies emerge among these prim-pols. Clusters AK, BH, and DA contain PriCT_1 (pfam08708), a primase C-terminal-like conserved domain, immediately following their prim-pol domain at the N-terminus. This domain is unique to these 3 clusters among the prim-pols analyzed. Clusters DA and EA7 share homology to the conserved DUF927 (pfam06048), which putatively belongs to replicative helicases. Cluster EJ shows no additional homology beyond the prim-pol domain, but the prim-pol gene itself is closely related to those of clusters AK and EA7.

Clusters DY, ED, and O are the only prim-pols of unknown class with high confidence homology matches. Although no conserved C-terminal domains were found for clusters DY and O, both clusters—O in particular—had Phyre2 matches with minichromosome maintenance (MCM) proteins from *Saccharomyces cerevisiae* and *Pyrococcus furiosus*. These proteins belong to the AAA+ family, similar to SF3 helicases, and are involved in eukaryotic genetic replication.[37] Cluster ED prim-pols show no C-terminal homology through Phyre2 but have a CDD conserved DUF3987 (pfam13148) domain. NCBI BLAST (Basic Local Alignment Search Tool)[38] alignment reveals that the C-terminus of cluster O and DY prim-pols also exhibits pronounced homology with bacterial proteins containing DUF3987. Proteins with this domain are generally MCM-like helicases associated with both AEP and DnaG primases.[37] This class

also contains a phage of unknown cluster, Microbacterium phage vB_MoxS-ISF9. Due to its high level of similarity to ED phages, ISF9 is herein considered in concert with cluster ED.

Based on these findings, the unknown class of prim-pols may be loosely organized into PriCT-like (clusters AK, BH, DA, EA7, EJ) and MCM-like (clusters DY, ED, O) subclasses. Interestingly, cluster K5 prim-pols also show some similarity to the DUF3987-containing protein. The expectation values of predicted domains and levels of homology used in these subclassifications are substantially less favorable than those used to classify RecA-like and SF3-like, not to mention that PriCT homology occurs near the prim-pol domain, rather than at the C-terminus. The sample size of each subclass is small enough that alignments cannot elucidate much beyond intra-cluster agreements, but alignment of only PriCT-like prim-pols shows very high levels of agreement across the entire length of the protein. The differentiation between PriCT-like and MCM-like prim-pols may prove useful in future research when examining the unknown C-terminal type of prim-pol more closely.

Truncated prim-pols are simply proteins consisting of only a conserved prim-pol domain. They lack any additional C-terminal region but are predicted to be functional prim-pols based on sequence analysis. While most are around 250 aa in length, roughly the size of the minimal prim-pol domain of previously characterized prim-pols,[6,39] truncated prim-pols from clusters DK and DS are over 400 aa long. The function of this extended protein region is unclear. Interestingly, the size of the prim-pol domain of truncated prim-pols is homologous to ORF904 and thus roughly half the size of the human prim-pol domain.[8] Because these prim-pols lack a C-terminal domain, and actinophage prim-pols all share the same homology and domain conservation at their N-terminus, truncated prim-pol proteins cannot definitively be further classified as of yet.

Under this classification method, ORF904 and NrSp both belong to the SF3-like class of prim-pol. ORF904 also contains a pRN1-specific helicase domain in its internal region, but many of its top Phyre2 homologies align neatly with other SF3-like prim-pols. NrSp lacks conservation of the D5_N domain seen in other SF3-like prim-pols. Its prim-pol domain belongs to the COG4983 superfamily, which has not been observed in actinophage prim-pols. The crystal structure of the NrSp C-terminal domain was recently obtained as a homohexameric ring. The structure confirms that it is indeed an SF3 helicase but also contains novel head and tail domains as well.[15] Human PrimPol, on the other hand, does not share any specific homology or conserved domains with phage-encoded prim-pols.

**Protein Complements to Truncated Prim-Pols**

The functions of the N- and C-termini of full-length prim-pols are likely intertwined given the pervasiveness of this multi-domain arrangement in prim-pols spanning numerous species and kingdoms. Therefore, it is important to consider the possibility that truncated prim-pols are one part of a modularized version of larger prim-pols that work collaboratively with a separately coded protein to achieve the activity of the full-length version of the enzyme. The genomes of phages encoding truncated prim-pols were examined for separate proteins with similar conserved domains to the C-termini of other prim-pols. Candidates were analyzed by BLAST and Phyre2 to confirm the presence of relevant domains and homologs. Results are presented in Table 2, which shows a representative putative complementary phage protein from each cluster and the domains used to identify them.

Phages of cluster EK and EM are very similar, as are gene 2 (prim-pol) and gene 3 (complementary protein) of all their members. Cluster EK and EM complements fit the characteristics of an SF3-like C-terminus according to both domain conservation and homology

matches. Two different complement proteins are found among members of clusters J and X. These two groups of proteins both resemble the RecA-like C-terminus with homology results to match, albeit with lower confidence than EK/EM to the SF3-like class. Results from clusters DK, DS, and DU are far less convincing. The complement from cluster DU showed some similarities to RecA-like prim-pols, but its conserved domains and homologs point more specifically to a DnaB helicase. Interestingly, clusters DK and DS both contain a cluster-exclusive protein with a DUF3987 domain and very high confidence homology to MCM proteins. But as with cluster DU, suspected complements from DK and DS do not share the characteristics of other prim-pol C-terminal regions. These predictions are not definitive, and represent only preliminary exploration of the concept of modular expression of prim-pols by actinophages.

Table 2. Predicted protein complements to truncated prim-pols based on domain conservation and structural homology.

| Cluster | Phage Name | Prim-Pol Gene # | Complement Gene # | Conserved Domain(s) | Predicted Class |
|---|---|---|---|---|---|
| DK | GodonK | 201 | 225 | DUF3987 | Unknown (MCM-like) |
| DS | Boopy | 190 | 161 | DUF3987 | Unknown (MCM-like) |
| DU | Neville | 93 | 71 | COG0305 (DnaB), DnaB_C, PRK09165 | Unknown |
| EK | TinyTimothy, Akoni | 2 | 3 | COG3378 (P4 primase), primase_Cterm | SF3-like |
| EM | Burro | 2 | 3 | COG3378 (P4 primase), primase_Cterm | SF3-like |
| J | Courthouse | 133 | 142 | AAA_25, RepA, PriCT_1 | RecA-like |
| X (J) | Gaia | 94 | 109 | AAA_25, RepA | RecA-like |

**Residue Conservation in the Prim-Pol Domain**

Vital catalytic prim-pol and nucleotide-binding residues are largely conserved across the prim-pol genes examined. The essential prim-pol active-site residues of pRN1 ORF904 (K65, D111, E113, H145, D171) are intact across most phage-encoded prim-pols based on a ClustalW protein sequence alignment of all 569 sequences. Mutation of any of these residues in ORF904 leads to a massive decrease in both primase and polymerase activities.[39]

Table 3 shows the amino acid composition in actinophages at these functionally important sites (highlighted in green where conserved) as a percentage of conserved residues, as well as corresponding residues for Larva64 (highlighted in red where conserved). The double-acidic motif [D/E]x[D/E] found in the catalytic center of ORF904 (D111, E113) and other prim-pols' active sites,[8,40] where 'x' represents any residue, is 100.0% present in actinophages. This feature presents as DxD in all actinophages aside from the 5 members of the anomalous cluster FA that contain an ExE motif instead. The basic histidine (H145), which is responsible for stabilizing the first acidic active site residue (D111) that binds $Mg^{2+}$ to coordinate with incoming deoxynucleotides,[1] is almost entirely conserved by sequence alignment. This is also true of the N-terminal lysine (K65).

Table 3. Conservation of functional ORF904 residues in actinophage meta-alignment.

| Function | ORF904 Residue | Consensus Sequence Residue | % Identical (% Similar) | Larva64 Residue |
|---|---|---|---|---|
| Prim-Pol Active Site | K65 | **K** | 96.3% (0.4% R) | **K** |
| | D111 | **D** | 99.1% (0.9% E) | **D** |
| | E113 | **D** | 0.9% (99.1% D) | **D** |
| | H145 | **H** | 81.4% (15.4% R) | **H** |
| | D171 | V | - | **E** |
| Zinc Stem | H141 | S | - | P |
| | H188 | gap | - | gap |
| | C191 | P | - | P |
| | C196 | Y | - | Y |
| Substrate Binding | K66 | A | - | H |
| | K70 | L | - | I |
| | N176 | K | - | **N** |
| | Y178 | **Y** | 90.7% (4.7% F) | V |
| | S184 | **S** | 83.7% (12.5% T) | **S** |
| | H190 | **R** | 27.0% (52.6% R) | **H** |

No identifiable consensus corresponding to the third acidic catalytic residue (D171) in ORF904 emerged based on sequence alignment. However, all phage prim-pols contain a glutamic or aspartic acid within five residues of the site of alignment with ORF904's D171 and the residue is aligned in Larva64. Additionally, the NrSp residue E139 has been shown to coincide in its crystal structure with ORF904 D171; but in this and previous sequence

alignments, E139 is offset from D171 by one residue.[6] Larva64 is one of the few actinophage genes that shows sequence conservation at D171. 3D structural modelling also shows that Larva's E137 aligns with high confidence with D171. It is likely that the primary sequences surrounding the third acidic active site residue in prim-pols are simply diverse enough that predicting their specific location is difficult based on sequence alignment alone, and structural modelling has thus far rectified the misalignment in all cases.

ORF904 contains at least two DNA binding domains, comprising six residues at the N-terminus[39] and a combination α-helix/β-hairpin at the C-terminus.[2] Overall, three of the six N-terminal DNA binding residues found in ORF904 are conserved in actinophages, including Y178, but the C-terminal DNA binding domain could not be identified. In particular, the tyrosine at position 178 has been found in both human PrimPol and NrSp to be critical for nucleotide binding and sugar selectivity.[7,41] Larva64 and other K5 phages lack a tyrosine at the Y178 position, instead containing a valine in both meta- and class-specific alignments.

Finally, the zinc stem found in ORF904 is notably absent from actinophage prim-pols. Primases often contain a transition metal chelation site, and human prim-pol has been shown to depend on zinc-binding to initiate priming and to stabilize itself on ssDNA.[11] Primase-polymerase enzymes from actinophages, on the other hand, lack any semblance of a zinc-binding domain despite apparent similarities in the overall shape of their active sites to proteins that do bind zinc.[39,40] This is also the case for the phage prim-pol NrSp.[6] Whether phage-encoded prim-pols utilize zinc or any other transition metal is not yet known—even without adding them to reaction buffers, metal ions have been observed to remain tightly bound to human PrimPol through the protein purification process[9]—but the crystal structure of NrSp bound to dGTP does

not contain any zinc.[7] If actinophage prim-pols utilize zinc to initiate priming, they are achieving chelation via a hitherto uncharacterized type of metal binding domain.

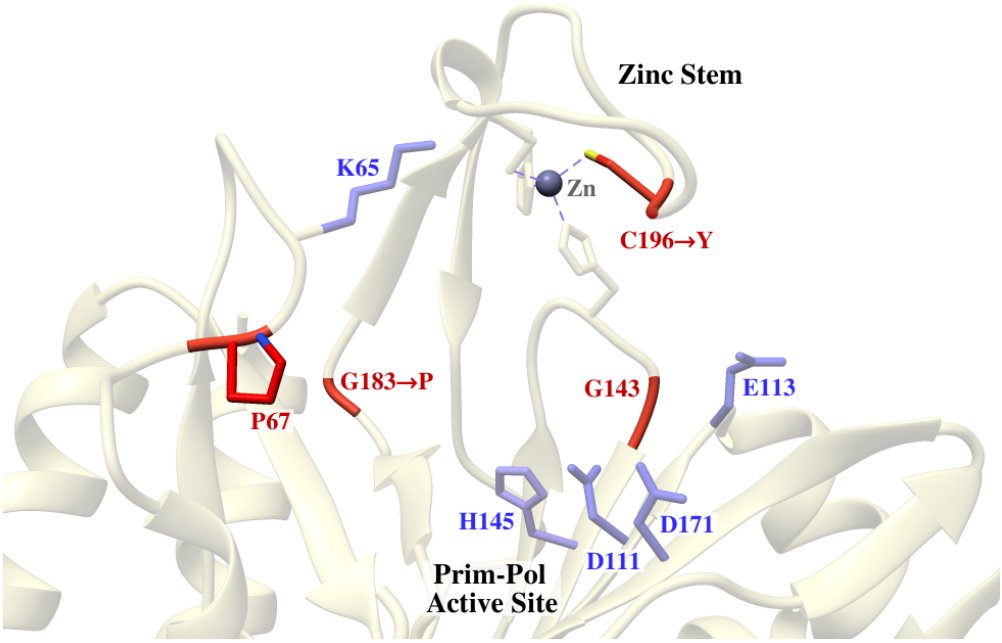


Figure 3. Orientation of ORF904 catalytic residues in relation to zinc stem and additional actinophage-conserved residues.

Actinophages show high levels of consensus at several ORF904 residues that appear to be structurally significant (A47, G54, P67, G143, G183) as well as at the site of a zinc-binding residue (C196). P67, G143, G183, and C196 are pictured in Figure 3 (red) in the context of the crystal structure of ORF904, as are 5 active site residues (blue). Due to the lack of zinc binding residues, actinophage prim-pols likely contain an entirely different structure in place of the zinc stem. The glycine that sits at the base of the zinc stem in ORF904 (G183) has been replaced by the more rigid proline, which is likely involved in stabilizing a critical structural juncture not present in pRN1. In place of the zinc binding residue C196 is a tyrosine that, due to its presence in almost all actinophages, may have functional importance. The conservation of G143 is likely due to its placement in the active site β-hairpin turn.

Finally, the highly conserved P67 lies at the base of the lysine-containing loop (K65 in ORF904). The presence of a rigid residue there likely aids in maintaining proper placement of the essential lysine. The vertical portion of that same loop on the other side of the lysine contains a second proline in ORF904 (P62). Nearly every actinophage prim-pol contains a second proline within 5 residues of their P67 alignment, indicating that the lysine-containing loop as a general feature is conserved. In fact, prim-pols from clusters CR, D, DG, EN, H, O and R have 2 - 3 prolines in a row at their point of alignment with P67 and many contain 5 or 6 prolines total in the space of 13 residues. Most house several basic residues within this proline-rich area. While the structure housing important active-site lysine(s) is clearly rigorously maintained, it appears that the specific composition of the lysine loop varies with phage cluster.

**C-Terminal Nucleotide Binding Motifs**

All full-length actinophage prim-pols contain some form of Walker A sequence, as seen in Table 4. Non-canonical features are underlined. Despite some variation in motif arrangements, the key Walker A binding residue lysine[42] is 100% conserved in full-length prim-pols. In fact, the vast majority of prim-pols contain a canonical Walker A. K5 is the only subcluster whose Walker A does not align with those of other prim-pols in the meta-alignment, although this anomaly is rectified by a RecA-like specific alignment and it contains a canonical Walker A. Non-canonical Walker A types found include AKT, SKS, and GKG.

Table 4. Walker A motifs associated with actinophage prim-pols.

| C-Terminal Class | Cluster(s) | Walker A Type |
|---|---|---|
| RecA-like | B, DR, K5, Singleton (2) | GxxxxGKS |
| | Singleton (1) | GxxxxGKT |
| Unknown / Special Cases | AK, BH, DA, ED | AxxxxGKT |
| | ED | AxxxxGKS |
| | EJ, EA7 | AxxxxAKT |
| | O | AxxxxGKG |
| | DY | GxxxxGKG |
| | *Predicted complementary proteins | |

| C-Terminal Class | Cluster(s) | Walker A Type |
|---|---|---|
| SF3-like | AC, D, DG, H, R, U | GxxxxGKS |
| | CR, EN, Singleton (2) | GxxxxGKT |
| | FA | GxxxxGKG |
| Truncated* | DU, EK, J, X | GxxxxGKS |
| | J | AxxxxGKS |
| | J | GxxxxSKS |
| | EM | GxxxxGKG |
| | DK, DS | Unknown |

Walker A motifs found in putative complementary proteins to truncated prim-pols are also included. However, cluster DK and DS complement proteins present a unique challenge (Figure 4). Among DK complements, one contains a Walker A-like sequence (GtgtGKG) with only 3 internal residues instead of 4 and the other has a serine in place of the critical lysine residue. Two members of cluster DS contain a fathomable Walker A permutation (GgsggGKK), while DS phage GMA2 specifically lacks this portion of the gene and contains no other semblance of a Walker A motif.

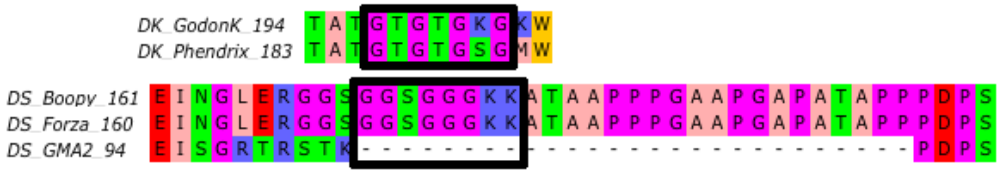


Figure 4. Unusual Walker A-type patterns present in cluster DK and DS phages.

Conserved Walker B sequences could not be definitively identified but a double-acidic motif can be found in phage prim-pol sequences just C-terminal to the Walker A in nearly all clusters, and regions of high hydrophobicity precede such motifs in clusters AK, DR, EA7, ED, EJ, EN, K5, OB, several B subclusters, and some singletons. RecA-like prim-pols may also contain the two acidic residues of their Walker B motif on separate β-strands, rather than as adjacent residues. Based on the presence of Walker motifs, all actinophage-encoded prim-pols likely bind and hydrolyze ATP. This may be experimentally determinable on an individual basis using a fluorescent nucleotide triphosphate such as MANT-ATP as well as translocation studies.[43] Identification of some of the remaining signature nucleotide binding features (Sensor 1, Sensor 2, Arginine finger)[42] would likely require additional structural data to orient specific residues and secondary structural elements in relation to the ATPase active site, since their location is not apparent based on met-alignment of actinophage prim-pols.

**Genomic Environment**

Actinophage-encoded prim-pols inhabit diverse genomic environments. Many share the genome with DEAD-box helicases—a type of SF2 helicase—and PD(D/E)XK superfamily nucleases, as well as other types of DNA polymerases. Others stand more or less isolated within the genome from additional replicative enzymes. The sheer diversity of endogenous genetic metabolism proteins surrounding prim-pols hints at their multivariate and multifunctional nature. They may participate in completely different replicative mechanisms from phage to phage or even perform multiple functions within a single virus.

SF3-like prim-pols belong almost exclusively to lytic phages and exist within an interesting genomic pattern. In clusters AC, CR, EN, and both singletons of this class, the prim-pol gene comes just after the structural proteins, followed by a DNA polymerase III, a AAA protein, a PD(D/E)XK nuclease, and finally a DEAD-box helicase. Clusters D, DG, H, R, and U, on the other hand, all contain the same sequence of genes but in the exact opposite order, placing the prim-pol at or near the end of the genome. The temperate cluster FA phages are unique among SF3-like encoders. At only 43 kb, their genome is 20 - 30 kb smaller than other phages of this class. It is worth noting that cluster FA phages encode the only actinophage prim-pols to carry an ExE active site motif, rather than DxD. It is also worth noting that FA phages share an exonuclease gene with cluster DY and O, the only other clusters with prim-pols genes containing a GKG Walker A motif, as well as cluster K5. Clusters FA, DY, and O also appear to be closely related to one another in a phylogenetic analysis (see Figure 10 below). Aside from this exonuclease and prim-pol, the only other relevant proteins in FA phages are a handful of putative DNA biding proteins. Excepting cluster FA, phages containing this class of prim-pol are remarkably consistent despite spanning 9 clusters.

RecA-like prim-pols are usually found sandwiched between a DEAD-box helicase and a DNA pol I as a conserved replicative cassette in lytic phages. The frequency of this pattern's occurrence indicates RecA-like prim-pols likely cooperate with the helicase and pol I in the process of normal genetic metabolism. The exceptions to these trends are the temperate cluster K5 phages, which are relatively anomalous among phages encoding RecA-like prim-pols. They have a beta clamp and exonuclease, then prim-pol immediately followed by a DEAD-box helicase. They do not contain any other DNA polymerases, but have an RNA ligase and contain a tRNA. K5 prim-pols are also smaller than other RecA-like proteins—just over 700 aa, compared to a typical 860 - 1000 aa. As noted, K5 phages share an exonuclease with clusters DY, FA, and O, among which K5 is the only cluster lacking a GKG Walker A. K5 is also unusual in that the only CDD prediction at its prim-pol C-terminus is a AAA domain. For most RecA-like prim-pols, CDD predicts a large number of DNA repair and recombination domains. REQ1, a singleton, also stands out among phages containing RecA-like prim-pols. REQ1 codes for the only known actinophage prim-pol without any other pham members. Its only DNA metabolism-related protein aside from prim-pol is a serine recombinase, well downstream from prim-pol.

The unknown class of prim-pols are substantially less predictable overall. This is to be expected, since their classification relies not on the presence of a conserved domain but on its absence. However, PriCT-like clusters (AK, BH, EA7, EJ, DA) do share notable homology. The former 4 all present a helicase, PD(D/E)XK nuclease, AAA protein, prim-pol, DNA polymerase I, and then adenylosuccinate synthase. Cluster DA contains the same genetic metabolism proteins, but in an entirely different order.

The MCM-like prim-pols do not show the same internal consistency. Cluster DY contains only a PD(D/E)XK nuclease and MCM-like prim-pol. Phages from cluster O have an endonuclease, exonuclease, and prim-pol at the start of their genomes but—in contrast to other instances where a prim-pol lies near the genome start—are not circularly permuted. After the cluster O structural genes lies series of replicative proteins comprising a beta clamp, a Ku-like DNA breaking protein, and a ParB genetic partitioning protein, with a AAA protein at the far end of the genome. Cluster ED phages have their prim-pol reliably located after structural genes, followed by a pol III ε exonuclease domain, RNA ligase, and a SNF helicase.

The environments surrounding truncated prim-pols are highly inconsistent from cluster to cluster. They range from the tidy consistency and elegance of clusters EK and EM to the chaos of the massive DK, DS, and J genomes. DS, DK, EK, and EM phages all encode a DNA pol I in addition to prim-pol. In EK and EM, prim-pol lies at the start of the genome and is followed directly by a primase, helicase, DNA pol I, and PD(D/E)XK nuclease. The nuclease is followed by a AAA protein in EM phages. Clusters EK and EM also contain an enigmatic and truly enormous (approximately 13 kb) gene with no known function. DK, DS, DU, J, and X all contain a sprawling array of poorly conserved DNA metabolism and processivity-related proteins following their structural genes and demonstrate little in the way of reliable replicative protein groupings.

## Structural Analysis of Larva64

The broad domain organization of Larva64 is typical of prim-pols: it contains conserved N-terminal prim-pol and C-terminal ATPase domains. It has a RecA-like C-terminus but conserves only a AAA domain. To gain further insight into the structure of Larva64 and to complement biochemical characterization, both the full-length protein and the first 280 N-

terminal residues (N280) were modelled. N280 comprises the isolated prim-pol domain. The full-length Larva64 protein was only modelled with high confidence over 65% of its residues, residues 10-170 of the prim-pol domain and 370-680 of the ATPase domain.

Both models are pictured in Figure 5. The upper diagram displays the modeling confidence throughout the full-length Larva64 sequence (722 aa), with regions of high modelling confidence highlighted in blue (prim-pol domain) and red (ATPase domain), and regions of low confidence in gray. Some significant residues are shown in yellow and indicated on the confidence diagram: the essential prim-pol lysine and its structural proline (K36, P38), five active site residues (D81, D83, H113, E137, V144), the actinophage-conserved proline and tyrosine residues where ORF904 has a zinc stem (P149, Y158), and the Walker A lysine (K415). Modelling shows an antiparallel β-sheet core at the prim-pol active site consistent with previously characterized prim-pols[39] and the AEP superfamily more generally,[1] as well as a parallel β-sheet core in the ATPase domain consistent with RecA.[18]

The full-length model lacks the helix bundle domain, found C-terminal to the active site in ORF904, that comprises part of that enzyme's minimal primase domain.[5] However, the residues that would make up the helix bundle domain in Larva64, if it has one, lie in the low confidence internal protein region. Both models yield highly similar active site architectures. The inclusion of the C-terminal region does not appear to affect how the prim-pol domain is modelled, highlighting the separation of homology between the two domains. The unknown internal region of Larva64 represents a significant gap in our understanding of the global architecture of full-length prim-pols in actinophages.

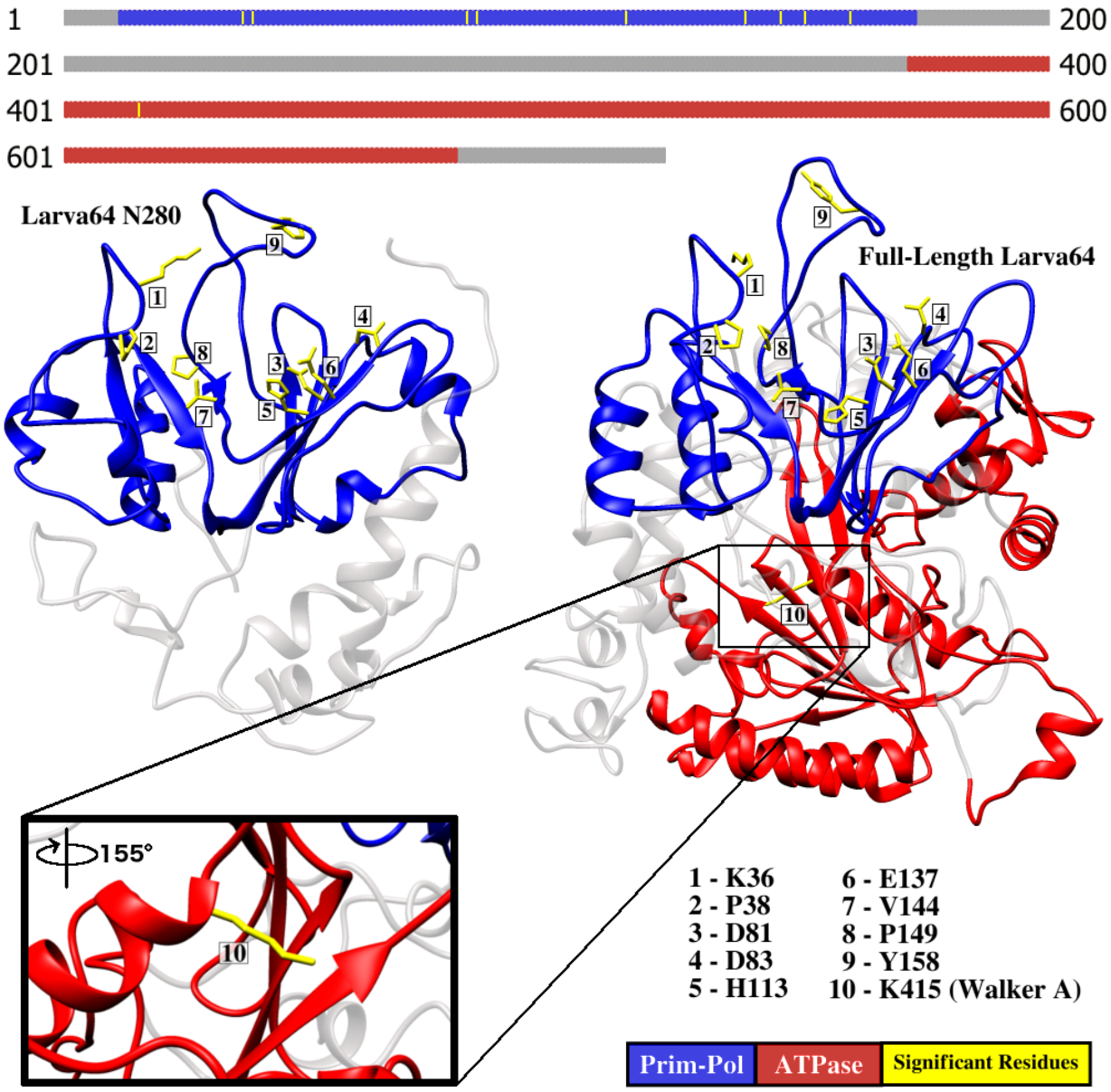


Figure 5. Models of N280 and full-length Larva64 highlighting regions of high confidence in the prim-pol and ATPase domains, as well as significant residues.

ORF904's zinc binding stem is situated above the prim-pol active site and is critical to its primase activity.[5] As seen in Table 3, Larva64 and other actinophage-encoded prim-pols do not conserve this zinc binding domain. Figure 6a shows an alignment of prim-pol sequences from pRN1, NrS-1, and Larva. No conservation of zinc binding residues is seen in the phage-encoded sequences. ORF904's zinc binding residues are indicated by gray circles and numbering is based

on pRN1. Larva64 residues R105-H113 and S150-G164 (blue) are superimposed with ORF904 residues T137-H145 and C185-T204 (orange) in Figure 6b. When ORF904 is used as a model template, Larva64 shows preservation of the stem-like feature above its active site but not of critical zinc binding residues in that region. NrS-1 enitrely lacks a stem in the region of interest and is not pictured.

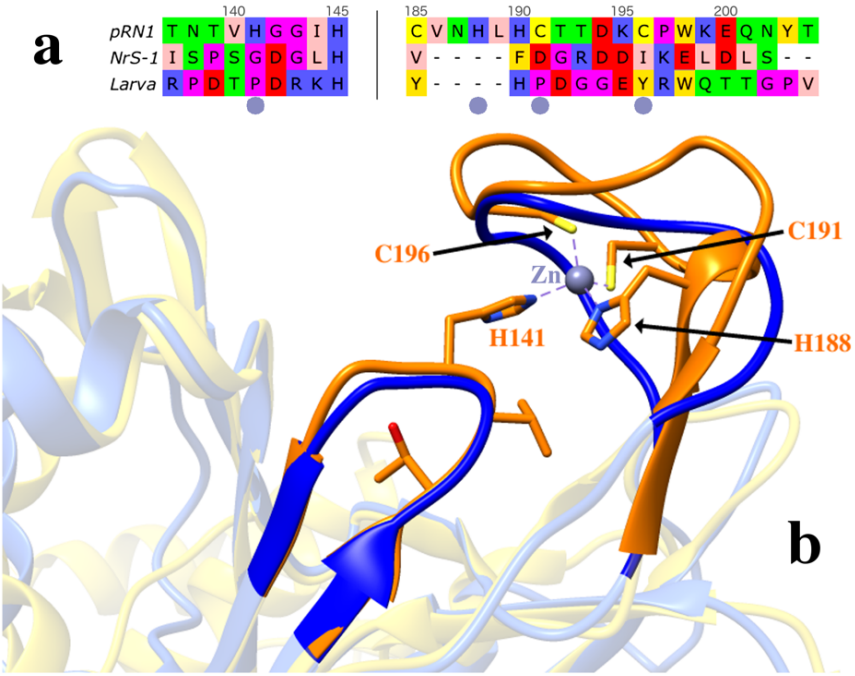


Figure 6. (a) Zinc binding residues in the zinc stem of ORF904 not conserved in NrS-1 or actinophage prim-pols. (b) Structural alignment of Larva64 with ORF904 zinc stem.

The prim-pol active site of Larva64 aligns structurally with those found in ORF904 and NrSp. Figure 7 shows a superposition of the N280 model (blue) with the crystal structures of ORF904 (orange) and NrSp (magenta), with residues labelled accordingly. ORF904's critical prim-pol active site residues D111, E113, H145, and D171[39] are all conserved. The tyrosine at position 178 in ORF904 also appears in NrSp as Y146, but it is replaced by V144 in the Larva model. While all three sequences also contain a lysine at the location of the essential ORF904

residue K65, the orientation of the lysine in NrSp differs substantially from pRN1. In ORF904, this lysine stretches toward H190 on the zinc stem, forming a basic bridge over an open groove near the active site. NrSp, like other phage-encoded prim-pols, lacks zinc binding residues and a zinc stem making this structural interaction impossible.

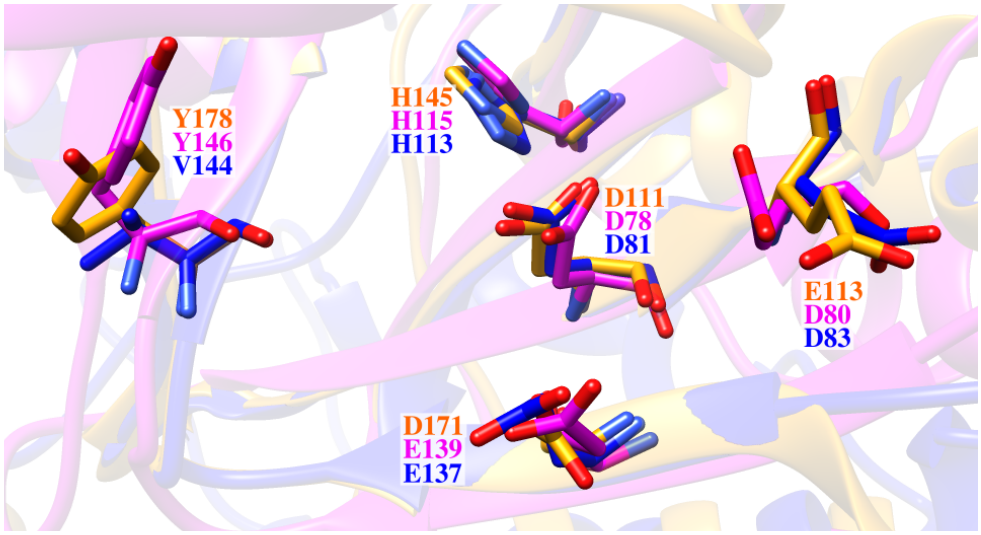


Figure 7. Conservation of prim-pol active site architecture between pRN1, NrS-1, and Larva.

Although ORF904's K65 and H190 are both broadly conserved in actinophages by sequence alignment, the exact relationship they have outside of ORF904, if any, is impossible to determine by modelling alone. Different modelling templates place these two residues in Larva64 anywhere from 3 - 18 Å apart from one another. Mutation of K65 has been shown to eliminate polymerase activity in ORF904, though the reasons why are unclear, and mutation of H190 causes a substantial weakening of DNA binding affinity.[39] Perhaps the role of the lysine is simply to provide a basic environment and structural integrity to the upper reaches of the active site. In fact, the corresponding lysine in the NrSp crystal structure (K27) stretches downward to interact almost directly with the phosphate groups on the loaded dGTP and encloses the back side of an electrostatically positive groove in which the phosphates rest.[7] Perhaps this is the role

the active site lysine plays in actinophages, rather than interacting with a zinc binding domain as in ORF904.

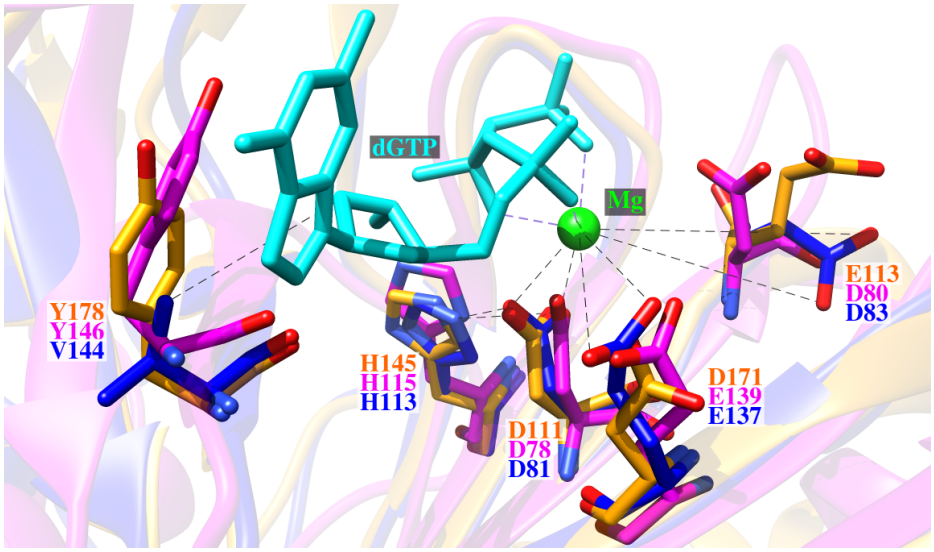


Figure 8. Relationship between Larva64 active site residues and a bound $Mg^{2+}$-dGTP complex

In Figure 8, the view of the active site in Figure 7 is rotated downwards approximately 90°, and a dGTP (cyan) and $Mg^{2+}$ ion (green) are modeled in a docked position. Some probable interactions of Larva64 residues are indicated by dashed gray lines. Upon binding of a dNTP, Larva64 residue D83 would likely rotate about its $C_\alpha$-$C_\beta$ bond to allow closer coordination of its carboxyl group with the $Mg^{2+}$ ion, as seen in NrSp.[7] Additionally, H113 is properly oriented to coordinate with D81 as it interacts with $Mg^{2+}$. The steric difference between Larva64's V144 and the corresponding tyrosine residues in ORF904 and NrSp is particularly evident here. Any interactions that tyrosine's aromatic ring or hydroxyl group have with the incoming dNTP would be entirely absent in Larva64. The exact interactions that allow Larva to bind dNTPs will require empirical structural studies to determine.

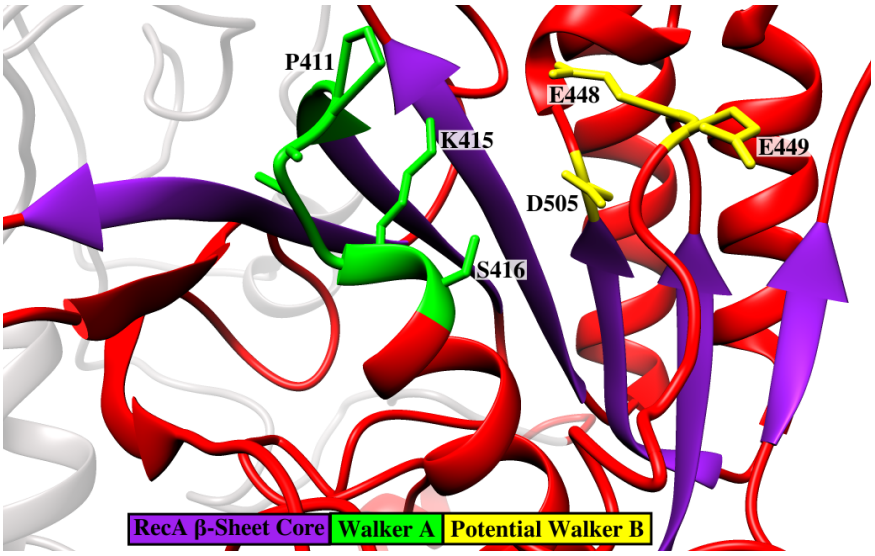


Figure 9. RecA parallel β-sheet core, Walker A motif, and possible Walker B motif in Larva64 structural model.

The five parallel β-strands that comprise the RecA core appear to be present in Larva64. Figure 9 shows the predicted RecA-like ATPase domain in Larva64's C-terminal. A series of β-strands (purple) run in parallel and the Walker A-containing P-loop (green) is entirely intact. Larva64 appears to also possess a structural proline (P411) that is commonly conserved in the Walker A motif of RecA proteins.[18] Possible Walker B residues (yellow) are shown as immediately adjacent to one another (E448, E449) and on neighboring β-strands (D505), as is seen in RecA.[16] The structural similarity of Larva64 to the RecA ATPase domain supports its classification as a RecA-like prim-pol, and therefore it is entirely feasible that Larva64 possesses recombinatorial activity similar to RecA.

## Discussion

What is clear from analysis of actinophage prim-pols is that they constitute a diverse grouping of proteins that cannot be fully understood by interpreting them as a monolithic entity. Prim-pols are not a single type of enzyme, but rather various types of enzymes with a shared prim-pol functional domain. Additionally, while most follow the general domain organization of

other prim-pols, some are truncated and appear to have the primpol and helicase-like activities expressed as separate proteins. Protein and genome size vary widely from phage to phage, and C-terminal domain conservation further divides these enzymes. Prim-pols certainly carry some amount of evolutionary favorability—they are nearly ubiquitous in the clusters where they occur. By their very nature, prim-pols at large are multifaceted enzymes, harboring at minimum DNA primase and DNA polymerase activities. Yet the breadth of their physiological roles also extends into more specialized functions and are still ill understood. Recall the variety of activities human prim-pol is capable of *in vitro* (Chapter One: Introduction). The presence of a prim-pol domain within an enzyme does not appear to dictate the broader functional possibilities of the protein. These findings suggest that the presence of a bifunctional prim-pol domain may be only one component of a replicative enzyme's overall purpose, rather than its defining feature.

Unsurprisingly, the most reliably conserved region among actinophages is the antiparallel β-sheet in the primpol domain typical to AEP family proteins.[1] This region houses the catalytic prim-pol residues and thus its structural conservation is a defining feature of prim-pols. More surprising is how little else remains congruent across the meta-sequence alignment. In fact, the catalytic prim-pol [D/E]x[D/E] motif and the Walker A lysine are the only residues that are 100% conserved among actinophages. This highlights the need to classify these enzymes more selectively than simply by looking at their shared prim-pol domain.

Grouping actinophage prim-pols into RecA-like, SF3-like, truncated, and unknown C-terminal types, and understanding the features that are common to each type, eliminates a great deal of the informational noise encountered when studying a large number of diverse sequences such as these. RecA-like and SF3-like prim-pols in particular may be understood, for the most part, by examining a few representative members. Systematic cluster-by-cluster analysis reveals

that in almost all cases, multiple clusters come together to form common threads that behave congruently.

Exceptions to this trend are numerous. Larva64 was chosen for biochemical characterization because it is Larva's only encoded DNA polymerase and because it expresses and purifies well, but not much additional information about it was available. As it turns out cluster K5 is somewhat anomalous in the RecA-like class, so much of the information learned about Larva64's C-terminus will be unique to K5 prim-pols. On the other hand, when studying a single prim-pol from, say, cluster CR3, the knowledge gained is directly pertinent to prim-pols from clusters AC, CR, D, DG, EN, H, R, U, and 2 singletons. The BLAST and Phyre2 results for the predicted complement to truncated DU prim-pols would seem to suggest meaningful similarity to RecA-like prim-pols. While that may or may not be the case, its specific domains and homology models were distinct from what was found with other prim-pols in the RecA-like class to the point that one cannot assume it's a match. Properly classifying prim-pols allows one to make a more educated guess as to what can and—more importantly—what cannot be presumed based on individual results. Although the actinophage prim-pol classes presented here are not very rigorously defined, further refinement based on what makes groups of prim-pols distinct from one another will allow for the planning of progressively more focused and meaningful investigations.

Another approach to classification is to organize sequences in a phylogenetic tree to compare how closely different clusters of sequences are related. This was done based on the same ClustalW meta-alignment used to compare sequences of all 569 prim-pols and is shown in Figure 10. Branches belonging to prim-pols from each C-terminal class are colored accordingly. The separation of classes is notably consistent, with RecA-like concentrated at bottom left, SF3-

like concentrated in the middle, and unknown type concentrated to the right. Among the unknown type prim-pols, PriCT-like (clusters AK, BH, DA, EA7, EJ) appear on a single branch, while MCM-like (clusters ED, DY, O) are spread out. This adds validity to the classification of some prim-pols as PriCT-like and suggests that, despite the lack of a predictable C-terminal domain, these proteins are in fact closely related.

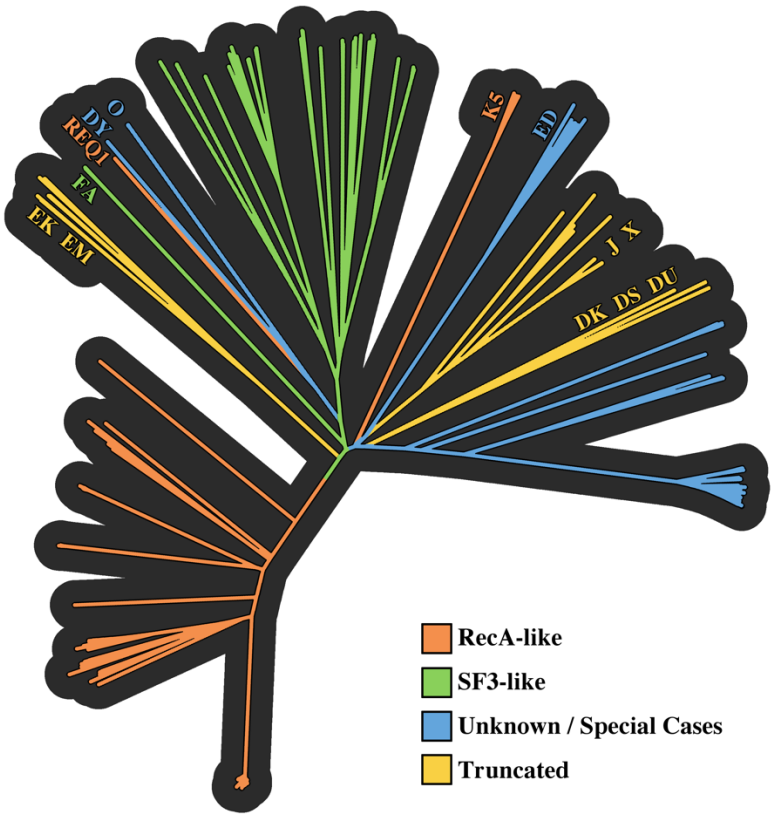


Figure 10. Phylogenetic tree of actinophage prim-pols color-coded by C-terminal type with outlying clusters labelled.

As could be expected, the truncated prim-pols—which are impossible to classify based on C-termini other than by their lack of one—do not self-segregate. The status of clusters K5 and FA and singleton REQ1 as outliers in their respective prim-pol types is corroborated by their placement in the tree. The level of separation between clusters DY and O and cluster ED is

surprising given their shared similarity to MCM proteins. It is likely that the homology they share is not as straightforward as it first appeared.

This type of visualization helps to predict where truncated prim-pols may lie in relation to the other 3 classes, which is not possible based solely on N-terminal domains. Clusters EK and EM appear to be closely related to the SF3-like class as predicted. Interestingly, clusters J and X do not show up anywhere near the main RecA-like branch despite similarity in their predicted protein complements. Clusters J, X, DK, DS, and DU may potentially be understood by taking a close look at the cluster ED and/or PriCT-like prim-pols to which they are closely related. It is important to remember that truncated prim-pols are being aligned with other prim-pols based on similarity in the N-terminus only. The validity of putative complementary proteins may be further tested by performing a series of alignments and phylogeny studies that include them.

The existence of truncated prim-pols is especially interesting in light of the hypothesis that prim-pols represent somewhat of a relic from early evolutionary history, a multifaceted solution to the manifold issues encountered during early genetic replication.[10] Perhaps evolutionary forces pushed some viruses to express full-length prim-pol as two individual proteins in order to better regulate discrete functions. Or, perhaps the C-terminus was shed from the genome entirely as phages adapted to eliminate redundancy with host proteins; any means of genomic reduction is highly advantageous to compact, efficient entities such as phages. Biochemical characterization of truncated prim-pols will be required to gain insight into their precise functions and interactions. At present they have yet to be expressed and purified and even their efficacy as prim-pols is undetermined, making them excellent subjects for further study.

Lastly, some aspects of the Larva64 structural comparisons are worth expounding upon. The presence of a valine in Larva64 (V144) at the position of the broadly conserved tyrosine

(Y178 in ORF904) is highly unusual, even among actinophage prim-pols. Both sequence alignment and structural modelling place valine where a tyrosine is usually present, at the back side of Larva's prim-pol substrate binding pocket. The significance of this residue has become increasingly clear as its role in the active site is better understood. Tyrosine's aromatic ring is thought to act as the main steric gate to NTP incorporation in human prim-pol by clashing with the 2' hydroxyl group of an incoming NTP. Mutation of tyrosine to a less bulky histidine largely abolishes sugar selectivity, and the Y100H mutation even allows human prim-pol to process NTPs, instead of exclusively dNTPs, with minimal impact on catalytic efficiency and fidelity.[41] NrSp, on the other hand, was found to contain a tyrosine in a similar location but different orientation relative to its binding pocket. It has been proposed that it is the main chain carbonyl group of this residue—rather than its aromatic ring—that prevents NTP binding in the prim-pol site.[7]

If this valine's orientation in Larva64 resembles that of the PrimPol's tyrosine, it should be able to incorporate NTPs since the sterically bulky aromatic ring is absent. On the other hand, if the residue is oriented as seen in NrSp's tyrosine, the absence of a bulky side chain should not affect its sugar selectivity except insofar as it affects the active site architecture at large. Of course, it is also possible that Larva and other K5 phages adopt a different configuration altogether. For this and other reasons, Larva64 is an excellent candidate for crystallization. A crystal structure would also shed light on the structure of the lysine loop that overhangs the catalytic prim-pol depression and the structure, if any, that stands in the place of the zinc stem of ORF904.

# CHAPTER THREE: RESULTS OF BIOCHEMICAL ANALYSIS

## *In Vitro* Characterization of Larva64

Biochemical studies of Larva64 confirm the presence of an active and robust primase-polymerase domain. Larva64 is also essential to viral survival, binds multiple DNA substrates, and translocates along ssDNA. It is soluble and expresses in high yield, and elutes off of an anion exchange column with two distinct retention times. These separate fractions, dubbed Peak 1 and Peak 2, were kept as separate protein stocks. Peak 1 was used in all biochemical studies of Larva64, though both fractions contain active prim-pol. The reasons for their chromatographic separation have not yet been determined.

### *De Novo* DNA Synthesis

Primase and polymerase activities were demonstrated simultaneously by a rolling circle polymerization assay using genomic M13 DNA. M13 is a filamentous *E. coli* bacteriophage with a circular, ssDNA genome approximately 6400 bases long. Rolling circle assays are typically used to measures the ability of a polymerase to extend a pre-annealed DNA primer on circular single-stranded M13 DNA (ssM13) and generate a complementary DNA strand. The enzyme's DNA synthesis progress is monitored by stopping fractions of the reaction at various timepoints following initiation using a quench buffer. The dsDNA product for each timepoint is then measured to show how large of a complement was synthesized for each timepoint. Strand-displacement (SD) synthesis is when a polymerase is capable of detaching an annealed DNA strand from the template it is reading, in a helicase-like fashion, using the forward motility of its polymerase activity. In a rolling circle assay if a polymerase is incapable of SD synthesis, progress will stop at the size of fully complemented M13 DNA (dsM13), 6400 base pairs (bp), at

the site where synthesis of a complement began. If the enzyme is capable of SD, it will continue to synthesize after a full complement has been generated, displacing the original strand as it goes and yielding a product larger than 6400 bp. *In vitro*, small DNA primers are annealed to a ssDNA substrate ahead of testing. *In vivo*, DNA polymerases require a DNA primase to synthesize a short RNA complement on ssDNA from which synthesis of a new DNA strand can proceed.[44] Figure 11 shows an agarose gel with DNA products of the Larva64 rolling circle assay.

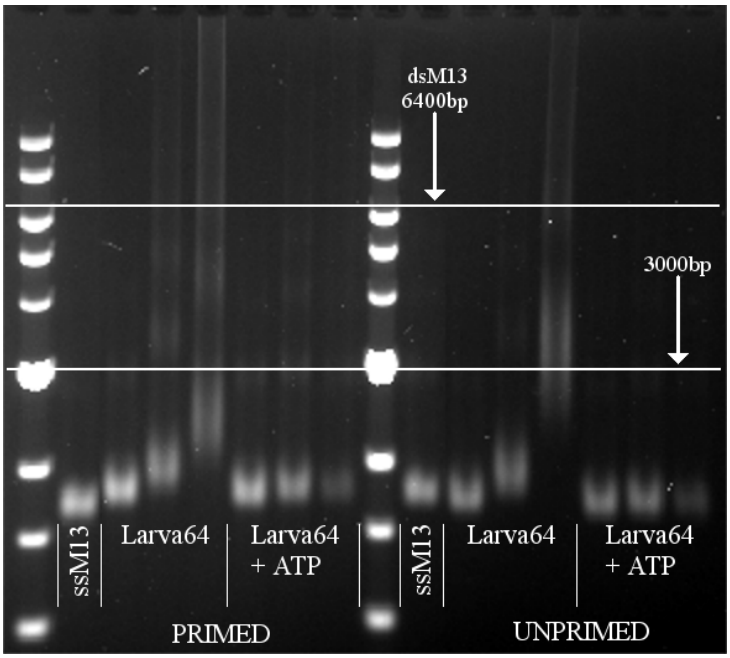


Figure 11. Agarose gel results of M13 rolling circle assay. Larva64 does not require a primer to initiate synthesis of a complement on ssDNA and acts as both a dNTP-dependent DNA primase and a DNA polymerase.

The darkening seen on the right side of the image is an artifact of imaging—bands on both sides of the gel are of roughly equal intensity. Unreacted ssM13 is included as a 0-minute timepoint ("ssM13") for each substrate type. Reaction progress is shown for Larva64 in the presence of dNTPs and $Mg^{2+}$ ("Larva64" lanes) and in the presence of dNTPs, $Mg^{2+}$ and ATP

("Larva64 + ATP" lanes) at 5, 20, and 60-minute timepoints. The size of fully complemented dsM13 is indicated just past the 6 kb marker.

Larva64's polymerization activity was measured on both primed and unprimed ssM13 substrates. It shows robust and virtually identical activity on both DNA types. The inclusion of a primer does not reduce Larva64's polymerization ability, and in fact appears to cause a concentration of product at a smaller size when compared to the results on unprimed M13—see bright clusters near 3 kb for each substrate. It is not known whether Larva64 begins synthesis at the primer on primed M13 substrate, or whether it performs ordinary prim-pol activity despite the presence of a pre-annealed primer. If the latter is true, prim-pol would need to perform SD when it reaches the primer. While relatively slow-acting when compared to highly efficient enzymes such as the T7 DNA polymerase, Larva64 demonstrates the ability to bind and prime ssM13 and to reliably synthesize a complementary DNA strand *de novo*. Figure 12 shows that eliminating $Mg^{2+}$ from the reaction buffer or replacing it with $Mn^{2+}$ entirely abolishes prim-pol activity. Oddly, increasing $Mg^{2+}$ concentration from 10 to 15 mM causes a reduction in activity. All conditions in Figure 12 are on unprimed ssM13, include dNTPs, and show product at 20 and 60 minutes of activity. As mentioned previously, RecA proteins can be highly sensitive to salt concentration. If Larva64 is acting as a RecA-like enzyme, an increase in $Mg^{2+}$ might stimulate some level of self-assembly behavior and interrupt polymerization. Clustering also occurs here but at a larger product size than in other rolling circle assays that were performed—around 6 kb.

Two peculiar features of Larva64's activity in this assay are the apparent stalling above 3 kb and the streaking of DNA products up to the top of the gel. The streaking phenomenon is particularly visible in 60 minute timepoints. Whether varying rates of SD synthesis are behind the DNA products larger than 6400 bp has not been determined; Larva64's SD capabilities, or

lack thereof, have not been proven. The streaking observed may be the result of robust polymerization and SD, the sum of multiple smaller stretches of dsDNA, or an artifact of some unknown variable.

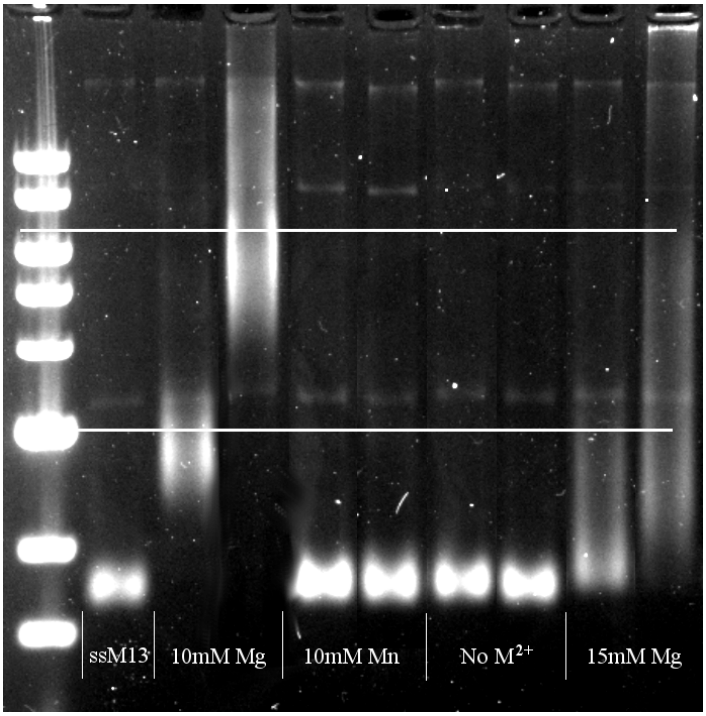


Figure 12. Divalent cation ($M^{2+}$) dependence of Larva64 prim-pol activity, $Mg^{2+}$ vs. $Mn^{2+}$.

These results also show that Larva64 acts as a dNTP-dependent DNA primase. A typical primase will exclusively incorporate NTPs when synthesizing a primer strand, but prim-pols demonstrate the ability to incorporate dNTPs directly. Indeed, Larva64 successfully synthesizes and elongates a primer in the presence of dNTPs alone but shows a remarkable inhibition of its prim-pol activity when ATP is added to the reaction mixture. Larva64's ability to use NTPs to synthesize a primer has not been tested. Preliminary rolling circle assay data indicate that the inclusion of $Zn^{2+}$ in the reaction may have an inhibitory effect similar to that of ATP, although this result will need to be replicated to determine its validity.

**DNA Binding**

The DNA binding affinity of Larva64 was measured on ssDNA, forked dsDNA, 5'-tail dsDNA, and 3'-tail dsDNA substrates by fluorescence polarization spectroscopy. The fluorescein amidite (FAM) label was used to tag short DNA substrates. When FAM is attached to a small molecule, in this case a short DNA segment, it can rotate freely in solution and polarizes the emitted light. But when something bulky like a protein is attached to the small molecule, the fluorophore's movement in solution slows considerably and the amount of polarization of the emitted light is altered in a quantifiable way.

Larva64 binds ssDNA, forked dsDNA, and 5'-tail dsDNA in the presence of ATP with 0.048, 1.610, and 0.979 μM affinity, respectively (Figure 13). Binding affinities are reported in terms of $K_d$, or equilibrium dissociation constant. $K_d$ values for each substrate type are indicated in blue along the protein concentration axis. Interestingly, the enzyme precipitates in the presence of 3'-tail dsDNA under the exact same buffer conditions that it binds the other substrates tested, including 5'-tail dsDNA. The reasons for this phenomenon are unclear, however the single-stranded tail of the 3'-tail substrate is capped with a FAM label that may interfere with binding. Blunt-end dsDNA binding was also tested and while preliminary results seemed to show binding activity, the results could not be replicated in more carefully controlled follow-up experiments. Notably, no binding was observed for any type of DNA substrate in the presence of dNTPs alone and ATP is absolutely required for binding.

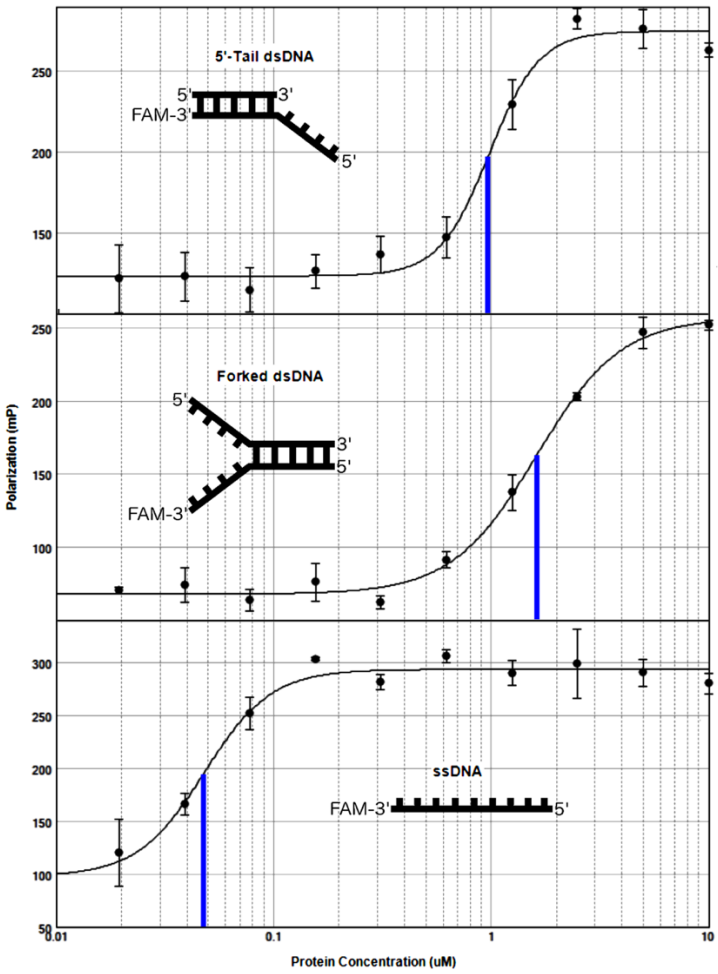


Figure 13. Binding of Larva64 to ssDNA, forked dsDNA, and 5'-tail dsDNA substrates in the presence of ATP.

Larva64's 48 nM affinity for ssDNA is more than 20x its affinity for 5'-tail substrate and more than 30x its affinity for the forked substrate. Therefore, based on this data it is unlikely that its C-terminus prefers binding at the junction between single- and double-stranded DNA, which is a binding property observed in many helicases. Given the ATP-dependency of the binding studies, the ssDNA binding activity shown is very likely occurring at Larva64's C-terminal ATPase domain.

## Other Work

### Bacterial Two-Hybrid Analysis

Preliminary data regarding Larva64's interactions with host *M. smegmatis* proteins were acquired via a bacterial two-hybrid system developed specifically for *M. smegmatis*. The two-hybrid method is designed to illuminate protein-protein interactions (PPIs) between a protein of interest and a library of bacterial protein fragments. Sequencing of vectors obtained through the two-hybrid selection process reveal PPIs between Larva64 and five *M. smegmatis* proteins: transketolase, 2-isopropylmalate synthase (2-IPM), transcription termination/antitermination protein NusA, chaperonin GroEL, and a fifth protein of unknown function.

Preliminary two-hybrid data, though far from conclusive, raises a number of interesting questions in regard to Larva64's role and mechanism during infection of the bacterial host. In addition to its role in bacterial shock responses, GroEL has also been shown to be essential for biofilm formation by *M. smegmatis*.[45] Larva64's interaction with transketolase may relate to that protein's role in the pentose phosphate pathway which, among other functions, is necessary for nucleic acid synthesis.[46] The 2-IPM analogue α-isopropylmalate synthase from *M. tuberculosis* has been explored as a possible anti-tuberculosis drug target due to its initiating role in the leucine biosynthesis pathway, an essential pathway for the bacteria's survival.[47] *E. coli*-encoded NusA can halt or pause transcription, and NusA in *M. smegmatis* has been linked exclusively to cells in a dormant, rather than active, state.[48] Viral interaction with NusA may play a role in allowing the infecting body to modulate protein production by the host cell as a means of dampening host immune response. BLAST and structural prediction data indicate that the final protein of unknown function may also be transcription-related. Additional two-hybrid screening

is necessary to confirm PPIs indicated by the initial screening and to reveal any further PPIs between Larva64 and *M. smegmatis* proteins.

**CRISPRi Silencing**

A CRISPR interference (CRISPRi) platform was used to render Larva64 transcriptionally silent. CRISPRi uses a modified version of the CRISPR-Cas9 system in which an inactive or 'dead' version of Cas9 (dCas9) binds to a target sequence with extraordinary specificity, but instead of facilitating a gene edit it simply remains bound to DNA at the chosen sequence. In doing so, dCas9 is able to form a non-negotiable steric blockade that prevents transcription of the chosen gene and effectively silences it.[49] An inherent trait of CRISPRi is that silencing a target gene will also prevent transcription of any downstream genes belonging to the same operon, or group of genes whose expression is controlled by a single promoter. Figure 14 shows the results of CRISPRi silencing of the Larva tapemeasure gene (positive control), prim-pol gene 64, and helicase gene without (left) and with (right) tetracycline induction of the CRISPR mechanism. Plaques are serially diluted 10-fold from left to right across each row. The embedded partial genome map shows the genetic context surrounding Larva64. The tapemeasure gene is known to be essential. Its knockout visibly eliminates most plaques, and the same result is achieved by silencing prim-pol. Silencing gene 64 is lethal to Larva, but because gene 64 lies near the start of a larger operon, genes after Larva64 were also silenced separately. While muting transcription of the operon from gene 64 onwards killed the virus, muting the operon starting at gene 65 was not lethal. This indicates that Larva64 is the gene responsible for viral takedown when the operon is silenced, making it an essential protein for viral survival.

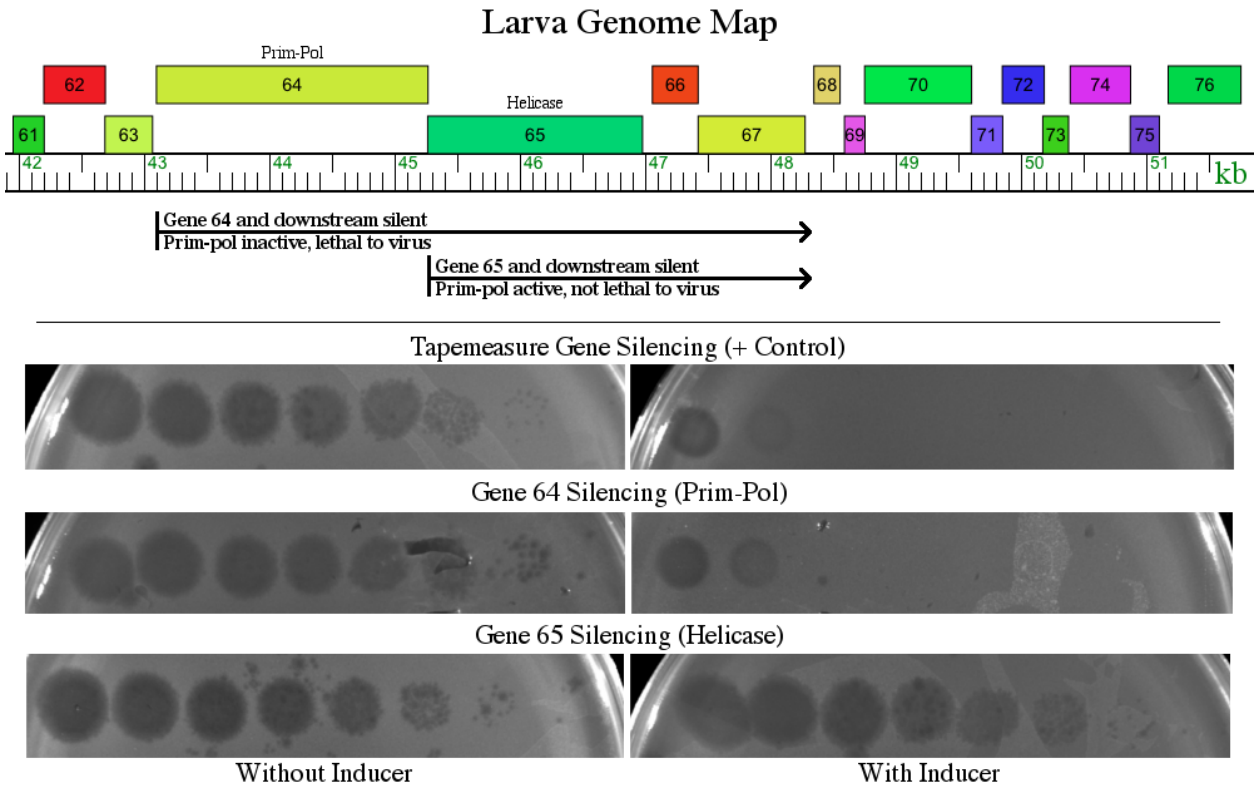


Figure 14. Results of CRISPRi gene silencing. Silencing of Larva gene 64 is lethal to virus, while silencing of gene 65 is not

**Peak 1 vs. Peak 2**

Peak1 and Peak 2 eluted with distinct retention times during ion exchange and were kept separate. Peak 1 was used in all biochemical studies because it was recovered in greater quantity than Peak 2. The differences in structure and chemical properties between these two fractions are yet to be determined. Portions of recovered protein from each fraction were also diluted with 1 volume of glycerol to create a 50% glycerol stock for stable liquid storage at -20°C. In order to determine whether both fractions are capable of primase and polymerase activity and to compare the fidelity of pure protein stored at -80°C to the 50% glycerol stock stored at -20°C, the *de novo* synthesis assay was repeated with Peak 1, Peak 1 glycerol stock, Peak 2, and Peak 2 glycerol stock. Results are presented in Figure 15. 20 and 60-minute timepoints were taken for each of the

4 protein stocks tested. All protein samples demonstrated *de novo* DNA synthesis capabilities on unprimed ssM13, confirming that Peak 1 and Peak 2 both contain active prim-pol despite eluting separately. All samples yielded product that streaked to the top of the gel and product clustered around 3500 bp.

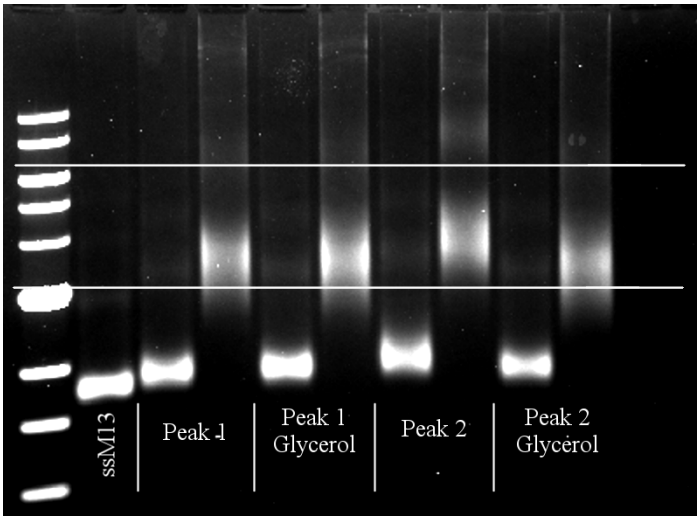


Figure 15. *De novo* DNA synthesis capabilities of Larva64 Peak 1 vs. Peak 2.

**DNA Translocation**

Stopped-flow DNA translocation studies reveal that Larva64's helicase domain indeed has ATPase activity, and translocates along ssDNA in the presence of ATP and $Mg^{2+}$. In Figure 16, translocation rates are compared on 30-mer 3'- and 5'-FAM labelled substrates at injection concentrations of 1 and 10 mM ATP (0.5 and 5 mM final), as well as on 60-mer 5'-FAM labelled ssDNA at injection concentrations of 1 and 10 mM ATP. Translocation took roughly 20 seconds under all conditions on a 30-mer substrate and signals were anticorrelated between 3'-FAM and 5'-FAM substrates. Interestingly, on a 60-mer DNA substrate Larva64 exhibits several seconds of lag time and takes approximately 2 minutes to translocate along the DNA in the presence of 0.5 mM ATP. When the ATP concentration is increased to 5 mM, the lag time is

eliminated and translocation is complete in approximately 2 seconds. This dichotomy suggests a multi-step mechanism of translocation requiring multiple phases of ATP binding and hydrolysis. At low ATP concentration, the rate limiting step in translocation involves this ATP turnover and is limited by ATP availability. When ATP concentration is increased, the enzyme is saturated and the rate limiting step is instead likely a conformational change, substrate release, or some other process not tied as directly to ATP turnover.

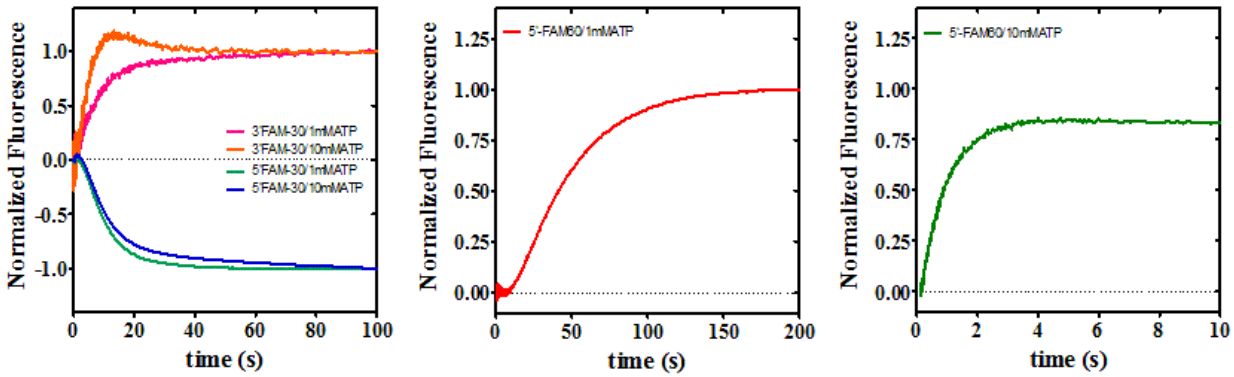


Figure 16. Stopped-flow fluorescence studies revealing lag kinetics of ATP-dependent ssDNA translocation by Larva64.

In light of these data, the *de novo* DNA synthesis assay was repeated exactly as above but in the presence of both 1 mM and 10 mM ATP. The impetus was that the inhibition of prim-pol activity in the M13 assay seen in the presence of 1 mM ATP (Figure 11) may correlate with the lag kinetics seen in the translocation assay. Increasing ATP concentration to 10 mM would allow the translocation mechanism to move at full speed and, if the translocation and prim-pol activities can operate collaboratively, may even result in a bolstering of polymerization speeds. As before, however, no prim-pol activity occurred upon the inclusion of ATP even at high concentration (data not shown), which suggests that Larva64's prim-pol and translocation mechanisms are not cooperative in nature.

## Additional Work

All of Larva64's observed primase, polymerase, and translocation activities depend specifically on the presence of $Mg^{2+}$. Larva64 not only lacks primase and polymerase activity with $Mn^{2+}$, but also aggregates more easily in its presence. Vigorous precipitation of protein is observed when $Mg^{2+}$ in stable buffer conditions is replaced with $Mn^{2+}$. The effect of $Mn^{2+}$ on translocation activity has not been tested. Despite Larva64's apparently negative reaction to manganese, it is possible that it employs $Mn^{2+}$ as a cofactor for different functions under appropriate conditions.

A Zincon-based assay for the detection of $Zn^{2+}$ in metalloproteins[50] was employed in an unsuccessful attempt to determine the zinc content, if any, of Larva64. Whether Larva64 binds zinc is as of yet unknown and will need to be determined by different methods. A protein crystal structure would help to determine if zinc is present in the prim-pol active site.

Samples of Larva64 with and without ssDNA present were analyzed by small-angle X-ray scattering (SAXS) in order to elucidate the conformation and oligomeric state of the protein, as well as to probe any changes that occur in the presence of DNA. From the SAXS data there is some evidence to support that Larva64 exists as a trimer in the absence of DNA; however, protein aggregation was apparent in these samples, so the oligomeric state cannot be determined with confidence. Buffer conditions will need to be optimized to allow Larva64 to remain soluble at the high concentrations required for SAXS analysis. SAXS will be used as a guide to drive protein crystallization studies.

# Discussion

The fact that Larva64 binds DNA is unsurprising. First and foremost, execution of both primase and polymerase activities requires that the enzyme load onto ssDNA. Additionally, the

RecA-like ATPase C-terminus likely interacts with DNA to perform some form of recombinatorial or repair-type function. The observed DNA binding activity is likely occurring at the C-terminal ATPase domain of Larva's prim-pol, since binding is ATP-dependent and experimental results in the *de novo* synthesis assay offer no indication that ATP is used by Larva64's catalytic prim-pol domain. Binding of ssDNA by the prim-pol domain occurs in the presence of dNTPs; however, no ssDNA binding activity was observed *in vitro* under dNTP-containing conditions in the equilibrium DNA binding assay.

The most likely explanation for the lack of binding with dNTPs is that ssDNA binding by the prim-pol domain occurs exclusively at a specific primase recognition site (PRS), as is commonly the case for primases. A PRS is a short nucleotide sequence to which a primase preferentially binds to initiate primer synthesis. ORF904 was found to be highly selective as a primase in its preferred DNA binding site, which is a GTG motif.[3] Larva64 is expected to bind ssDNA containing its preferred PRS with high affinity in the presence of dNTPs, but the PRS for Larva's prim-pol is yet to be determined and it is highly unlikely that the 30-mer DNA substrate used would contain a recognizable motif purely by chance. On the other hand, the isolated prim-pol domain of corynephage BFK20's prim-pol enzyme has been shown to not rely on a specific nucleotide motif to initiate priming.[24] While it is possible that Larva64 shares this lack of specificity, it is unlikely given the lack of binding activity observed when dNTPs are present.

It is worth noting the possibility that Larva64's ATPase domain adopts a hexameric configuration when binding DNA, as is seen in ORF904[51] and SF3 helicases more generally,[14] despite its homology with RecA. A hexameric state is predicted by SWISS protein modelling of Larva64, and the top Phyre2 result for Larva64 and other RecA-like prim-pols is the crystal structure of the *E. coli* plasmid RSF1010 RepA replicative helicase as a homohexamer (PDB ID

1NLF). Because protein concentrations are calculated based on a monomeric configuration, binding affinities of the hexamer would be six times those reported above. However, Larva64's homology to RecA and related proteins introduce the possibility that it interacts with DNA monomerically within nucleoprotein filaments. Filament self-assembly could also account for the type of protein precipitation seen in the presence of $Mn^{2+}$ and 3'-tail dsDNA, as well as the SAXS studies, were these conditions to stimulate filament formation.

Larva64's tendency to stall at particular product sizes during prim-pol activity is readily observable in repeated *de novo* synthesis experiments. Products are generally concentrated just below 3 kb on primed M13 and just above 3 kb on unprimed M13. The exact reasons for this phenomenon have not been determined, but one possibility is that Larva64 is only synthesizing a complement to small, specific portions of ssM13. If, like human PrimPol, Larva64 functions as a polymerase *in vivo* primarily for the purpose of DNA repair, it could be loading onto ssM13 just long enough to yield discrete stretches of complementarity. In other words, the DNA product clustered around 3500 bp may be M13 DNA with a number of short dsDNA segments totaling 3500 bp, rather than a single 3500 bp segment of dsDNA. If the site of an exogenous primer lies within one of those segments and Larva64 is *not* capable of SD, polymerization would be blocked at that location and fewer stretches of dsDNA would be produced resulting in smaller apparent product size on primed M13 DNA. Another, albeit less likely, possibility is that Larva64 initiates priming at a specific location and dissociates from the DNA after synthesizing a single dsDNA stretch of a particular length.

In Figure 12, product clustering is seen around 6 kb rather than 3 kb. The reason for the difference is not clear, but there does not seem to be any sign of product being held up on dsM13 at 6400 bp. This could mean that Larva64's polymerization leads seamlessly into SD synthesis

once it has generated a full complement, but more likely is a particularly robust example of the type of fragmentary dsDNA synthesis described above. Nonetheless, an agarose gel-based assay cannot convey much in the way of mechanistic detail. Real-time kinetic measurements of dNTP incorporation are needed to learn granular information about the fidelity and patterns of Larva64's polymerase activity.

The inhibitory effect imposed on Larva's prim-pol activity by ATP is very interesting as well. One plausible explanation is that ATP activates Larva64's C-terminus, causing it compete with N-terminal primase and polymerase activity. At the very least, it is certain that the protein binds ssDNA when ATP is present, that that binding is likely occurring outside of the N-terminal prim-pol domain, and that Larva64 translocates on ssDNA when both ATP and $Mg^{2+}$ are present. Therefore, one potential mechanism of inhibition is that the C-terminus of at least some of the protein is binding to and possibly translocating along ssM13 when ATP is added to the reaction, and is sterically blocking the polymerase activity of other prim-pol molecules. Having the C-terminus bound to DNA could also impact the prim-pol domain's motility and its flexibility to perform primer synthesis and extension, or more likely would cause a conformational change that deactivates the prim-pol domain entirely. Lastly, the presence of ATP may stimulate the formation of a hexameric ring or other oligomeric state that is unfavorable to N-terminal prim-pol activity.

Another possible mechanism is that the ribonucleotide is acting as a competitive inhibitor in the prim-pol active site. While the tyrosine nested in the binding pocket of most prim-pols has been shown to prevent exactly this form of inhibition by promoting ribose sugar selectivity, Larva64 is one of the few prim-pols that does not appear to conserve that residue. If the NTP gate is functionally absent in Larva64, it would be expected to be able to utilize NTPs in its

primase activity. Human PrimPol is specifically a DNA primase and polymerase, but remarkably, mutation of the essential tyrosine selector residue to a histidine allows it to act as an almost equally effective RNA primase and polymerase.[41] Larva64's activity in the presence of mixed NTPs still needs to be examined.

ATP's ability to inhibit one of Larva64's major domains while activating the other opens the door for some fascinating means of enzymatic regulation. Perhaps Larva64's ATPase activity is most advantageous to the virus in ATP-rich environments, while prim-pol activity is more advantageous in ATP-deficient environments. Although perplexing to observe in an isolated laboratory environment, the ebb and flow of substrate-dependent regulation is extremely common in biological signaling pathways. An abundance of ATP could signal that certain other viral or host proteins have been up- or down-regulated. Further examination of this phenomenon will provide significant insight into the relationship between this Larva64's N- and C-terminal domains. It is also important to remember that this is an essential protein to Larva. In its absence, the virus is unable to successfully infect its host, avert (or subvert) the bacterial immune response, and propagate itself for future infections.

# CHAPTER FOUR: CONCLUSIONS AND FUTURE DIRECTIONS

Bioinformatic characterization of actinophage prim-pols reveals a variegated collection of 569 enzymes. Rather than a cohesive category, actinophage prim-pols represent at least four classes of proteins based chiefly on differences in their C-terminal domains: RecA-like prim-pols, SF3-like prim-pols, truncated prim-pols, and full-length prim-pols with a C-terminus of unknown function. Within the last class, prim-pol enzymes encoded by several clusters contain a conserved PriCT domain adjacent to the N-terminal prim-pol domain and appear to be closely related to one another. The size of prim-pol genes ranges from roughly 200 to 1000 aa, and they inhabit very different genomic environments from phage to phage. They are sometimes found as part of a widely conserved genetic replication cassette, and sometimes appear as a solitary genetic metabolism gene. Actinophages encoding prim-pols are equally diverse, spanning 28 clusters, 8 host types, and with genomes ranging from 40 to 122 kb. The trends identified—and, of equal importance, their exceptions—present a strong case for the classification of these enzymes more specifically than by the blanket term "prim-pol." Rather, the bifunctional prim-pol center is only a single functional domain found in an array of enzymes that do not fit neatly under a single title.

Larva64, a predicted RecA-like prim-pol from cluster K5 phage Larva, exhibits dNTP-dependent DNA primase and polymerase activity, ATP-dependent DNA binding on three substrates (with strongest affinity for ssDNA), and ATP-dependent translocation along ssDNA. Larva64 is an essential protein for Larva and interacts with at least five Mycobacterial host proteins. It is an interesting study in that it contains fewer conserved C-terminal domains and is smaller than other RecA-like prim-pols and lacks a critical tyrosine residue found in the N-

terminal domain of nearly all other actinophage prim-pols. Larva64's prim-pol activity is dNTP- and $Mg^{2+}$-dependent and inhibited by ATP. Larva64 elutes off of an anion exchange column as two distinct, equally active fractions.

The results obtained thus far present an array of avenues for future investigation. Bioinformatically, future work will consist of ever more precise examination of what makes each class or cluster or individual prim-pol interesting and different. In particular, the truncated and unknown classes of prim-pol require further bioinformatic characterization. The same is true of outliers such as cluster K5, cluster O, and phage REQ1, among others. Indeed, the PriCT-like prim-pols may represent a class of their own based on sequence similarity and phylogeny, although their similarity is not based on C-terminal domains. Moving forward, the different classes of actinophage-encoded prim-pols can be examined individually rather than monolithically. While obvious patterns of similarity exist broadly within each class, they are not yet understood and are worthy of further study. As new examples of prim-pols are unearthed, possibly from new phage clusters, they can be classified relatively quickly based on homology and domain conservation. This will allow investigations to begin from a basic understanding of an individual prim-pol's distinctive characteristics, rather than from scratch, and will also make outliers and interesting studies easier to spot.

In the laboratory, there are a number of aspects of Larva64's biochemistry that require further study. Many of the studies performed on this enzyme so far have been exploratory in nature in an attempt to ascertain its general properties and behaviors. Its ease of expression and purification, its identity as an active prim-pol, its ability to bind and translocate on DNA, and its essentiality to the virus have been established. The next step in the characterization of Larva64 will be to specifically and rigorously test the unknowns surrounding its activity. Repetition of the

*de novo* synthesis assay under a number of different conditions can be used to test the limits and fidelity of the prim-pol active site and answer some important questions. These include testing the ability of Larva64 to synthesize with NTPs instead of dNTPS; more firmly establishing the relationship between ATP inclusion and prim-pol inhibition; and including $Zn^{2+}$ in reactions to determine what, if any, effect it has on activity. It may also be informative to perform a similar assay on different DNA substrates. Will activity be the same on linear, rather than circular, ssDNA? What about a DNA strand with a substantial portion already complemented?

Similar questions surround Larva64's DNA binding properties. Does increasing ATP concentration increase binding capabilities? What is the stoichiometric ratio of enzyme to DNA when bound? Does zinc affect DNA binding? The issue of aggregation in the presence of 3'-tail DNA should also be addressed. It is entirely possible that, since the single-stranded 3'-tail in this case contained the FAM label, the unusual behavior arose from an interaction of Larva64 with the fluorescent dye rather than a conflict with the substrate itself. Blunt-end dsDNA should be re-examined to find out if there are any conditions under which prim-pol will bind it.

It is important to firmly establish Larva64's preference for $Mg^{2+}$ over $Mn^{2+}$ for different types of activity. If it is absolutely proven that Larva64's prim-pol domain is inactive with $Mn^{2+}$ then it may be useful for understanding situations where $Mg^{2+}$ is required for normal activity, for instance to freeze the enzyme in an inactive state at the substrate binding step. Such a method could be helpful in addressing the significant issue of Larva64's PRS, if it requires one. Thus far, there is evidence suggesting that Larva64's primpol domain will not bind ssDNA with dNTPs alone. This either means that a necessary cofactor is absent, that the enzyme selectively binds only a certain nucleotide motif, or both. Based on the *de novo* synthesis results, ssDNA binding and prim-pol activity should begin once dNTPs and $Mg^{2+}$ are both present. Therefore, even if

Larva64 did bind the ssDNA it would be processive and dissociate quickly from the short DNA strand due to prim-pol activity. The reaction would not reach an equilibrium of bound/free DNA to measure. Some enzymes are capable of recognizing dideoxynucleoside triphosphates (ddNTPs), which lack both the 2' and 3' hydroxyl group found on NTPs. If an enzyme can recognize them, ddNTPs will be successfully incorporated but lack the 3' OH necessary for phosphodiester bond formation and thus cannot be extended. A combination of ddNTPs and $Mg^{2+}$ would halt prim-pol activity once the first ddNTP was incorporated. Additionally, if conditions can be optimized to avoid aggregation with $Mn^{2+}$ then it may be used as a cofactor to allow the protein to coordinate with nucleotides without incorporating them into a DNA strand. If no activity is seen under either of these conditions, it is possible that a PRS is also required. A number of methods have been used to identify PRS motifs, but a relatively simple (and affordable) method to narrow down the possibilities is depicted in Figure 17 below.

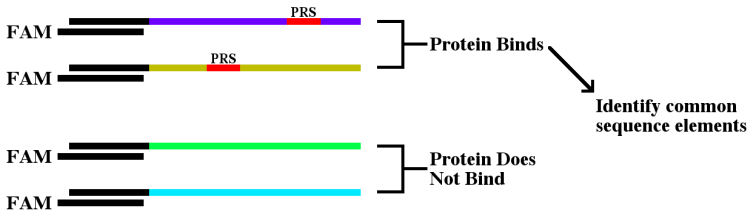


Figure 17. Simple schematic of PRS identification.

A single fluorescent-labeled substrate may be used. By keeping black regions identical and varying the sequence of colored regions, any number of small test sequences can be generated. With careful design and a sufficient number of distinct sequences, the PRS may be narrowed down to a few possibilities or even determined with certainty. Overhangs with identical sequences to both black regions can be designed and tested to ensure that loose, un-annealed DNA fragments don't already contain a PRS and won't create a false positive.

Identification of the PRS will allow for focused probing of the mechanisms that govern the prim-pol domain's loading and priming on ssDNA, as well as determination of its ssDNA binding affinity and whether it can bind dsDNA. It would also provide insight on the size of product expected in the *de novo* synthesis assay on M13 or other ssDNA substrates and allow for more enlightened interpretation of results.

The question of whether Larva's prim-pol binds zinc can be answered by any method that can sufficiently denature the protein to facilitate ion release, and that is sensitive enough to detect zinc at low concentrations. A Zincon-based colorimetric method[50] may be successful if modified, but previous attempts to use Zincon resulted in false positives that were difficult to resolve.

Expressing and purifying Larva64's N- and C-termini separately from one another will add several new layers to our understanding of each domain's purpose. Would the isolated N-terminal domain still exhibit prim-pol activity? Would it still be inhibited by ATP? Is DNA binding affinity of either domain affected by the absence of the other? If altered, would its behavior return to normal when both protein fragments are in solution together? And importantly, would the absence of only one domain or the other still be lethal to Larva? Truncated protein fragments would also be more plausible crystallization targets individually than the full-length protein and could possibly even be co-crystallized if predicted regions of disorder are removed from each.

Larva64 is an excellent candidate for crystallization. The three prim-pols for which crystal structures are available—human, pRN1, and NrS-1—differ enormously. Larva's prim-pol possesses its own distinctive features that will remain enigmatic until sufficiently high-resolution structural data is obtained. Residue V144 is of particular interest. If Larva contains a valine where other prim-pols contain a tyrosine, what is its mechanism for selectivity? How does it

discriminate between ribo- and deoxyribonucleotides? Crystallization would also show the shape and positioning of the highly conserved lysine loop that overhangs the prim-pol active site, as well as the structure that stands in the position occupied by the zinc stem in ORF904. Zinc itself could be located if present, and the residues and protein architecture that bind it could be identified. The most desirable structures to obtain are Larva64 bound to DNA, ATP, or a dNTP. Ideal targets for crystallization would include the full-length wild-type protein, as these would show the spatial relationship between the two domains and how the conformation of one end of the protein affects the other on different substrates. However full-length prim-pol has yet to be crystallized, so one can assume it is a fairly difficult task to achieve. Once buffer conditions are optimized, SAXS analysis will also provide valuable insight into the overall shape of Larva64 as a living protein in solution and as it interacts with DNA and individual nucleotides.

Additional *in vitro* studies should revolve around expressing and purifying prim-pols that represent as diverse of a selection of bioinformatically defined groupings as possible. This includes a prototypical RecA-like and SF3-like prim-pol, as well as an endogenously truncated prim-pol. Putative complementary proteins to a truncated prim-pol could also be expressed, and used to examine the relationship between the two proteins as described with individual Larva64 domains above. Unknown-type prim-pols should be more thoroughly studied before candidates for biochemical characterization are chosen. The more unique a prim-pol is, the more can be learned from each new piece of information about it. Scientific research has barely begun to scratch the surface of the novelty, diversity, and evolutionary ingenuity that is the bacteriophage. Research into prim-pols will benefit from continuing to biochemically and bioinformatically explore the characteristics that make these enzymes, and the entities that encode them, both analogous to and distinct from one other.

# CHAPTER FIVE: MATERIALS AND METHODS

## Bioinformatic Characterization

### Bioinformatics Tools

Data for actinophage genes encoding prim-pols were extracted from the "Actino_Draft" Phamerator database (www.phamerator.org © Cresawn, Hatfull, Bogel, Mavrich, Gauthier, HHMI SEA-PHAGES) on 05/14/2019 using the MySQL monitor via the 2017 SEA VM package made available by the HHMI SEA-PHAGES program. Phams containing possible prim-pol genes were isolated by identifying all phams containing genes whose functional notes contained "prim" or "Prim". All phams containing genes with notes suggesting a possible bifunctional primase-polymerase were selected for further review. Data were organized for processing in Microsoft Excel. Data extracted includes pham name, phage cluster, phage name, phage genome size, phage host type, gene number, gene protein sequence, and gene notes. The amino acid sequences encoded by one member of each subcluster containing a putative prim-pol were modelled using the Phyre2 protein fold recognition server.[28] Any phams containing proteins that match with known prim-pol structures through Phyre2 were retained for analysis and any that did not were considered false positives and removed from analysis. 569 sequences were retained as putative prim-pols. Protein sequences were aligned by ClustalW alignment using Clustal Omega.[52] Sequences were analyzed and visualized in Unipro UGENE Version 33.[53] Conserved domain predictions were obtained from NCBI BLAST,[38] Phamerator, and the CDD.[27] Phylogenetic tree was generated using Clustal Omega and visualized using iTOL.[54] Tables were created in Microsoft Excel. Additional editing and preparation of figures was performed using GIMP open-source freeware.

### Structural Analysis

N280 and full-length structural models of Larva64 were generated using Phyre2[28] intensive mode modeling. Models were also generated for additional reference using SWISS-MODEL.[55] Crystal structures of pRN1 ORF904,[5] NrS-1 prim-pol,[6,7] and *D. radiodurans* RecA[20] were downloaded from the Protein Data Bank (PDB ID 3M1M, 6A9W/6JOQ, and 1XP8, respectively). Structural comparisons were performed in *Coot* Version 0.8.9.1[56] using secondary-structure matching (SSM) superposition. Structural alignments and molecular graphics were generated with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH.[57]

## *In Vitro* Protein Characterization

### Expression and Purification of Larva64

The entirety of Larva gene 64 was amplified by PCR and cloned into a pET21PPS vector between NdeI and XhoI restriction digest sites, giving it an N-terminal His-tag. The construct was assembled by Mike Sands and Morgan Cheek in the Biochemistry Laboratory course at Western Carolina University. The plasmid was transformed into chemically competent *E. coli* BL21(DE3) cells and grown in LB media containing 100 ug/mL ampicillin. Cells were grown at an OD600 of 0.6-1.0, then induced using 0.5 mM IPTG overnight at 16 °C. Cells were harvested by centrifugation at 3,000 RPM, and the pellets were resuspended in ~100 mL of Nickel Buffer A (50 mM Tris pH 8.0, 0.5 M NaCl, 0.1 mM EDTA, 1 mM betamercaptoethanol (BME), and 10% glycerol) and stored at -80 °C.

For purification, cells were thawed on ice and lysozyme was added at a concentration of 0.2 mg/mL. Cells were sonicated in 1 minute intervals for a total of 5 minutes using a Branson sonifier. Lysed cells were centrifuged at 18,000 RPM for 1 hour, and the supernatant was

collected and passed through a nickel affinity chromatography column equilibrated in Nickel Buffer A containing 10 mM imidazole. The column was washed with 10 mM and then 40 mM imidazole and then protein was eluted with Nickel Buffer B, which is identical to Nickel Buffer A but contains 250 mM imidazole. Fractions containing Larva64 were identified by SDS-PAGE, then combined and diluted to 100 mM NaCl. When the solution was diluted, a significant amount of protein precipitated out of solution. The aggregated protein was gently pelleted and the supernatant was recovered for further purification.

A theoretical pI of 5.89 for Larva64 was calculated using ProtParam.[58] Therefore, the supernatant was passed over a Q anion exchange column equilibrated in 100 mM NaCl. The column was washed and protein was eluted with a 100 – 800 mM NaCl gradient over a volume of 200 mL. Ion Exchange A buffer contains 25 mM Tris pH 7.5, 1 mM EDTA, 1 mM BME, and 10% glycerol. Ion Exchange B buffer is identical but contains 1 M NaCl. Larva64 eluted off of the Q column at two distinct retention times ("Peak 1" and "Peak 2"). Peaks 1 and 2 were kept separate, and Peak 1 is used in all biochemical studies unless otherwise noted. Protein was dialyzed overnight in 2 L of storage buffer containing 20 mM Tris pH 7.5, 0.2 M NaCl, 0.5 mM EDTA, 1 mM DTT, and 10% glycerol. Final concentrations for Peak 1 and Peak 2 were 117.30 and 114.27 μM, respectively. Portions of both Peak 1 and 2 were aliquoted, flash frozen in liquid nitrogen, and stored at -80°C. Each of the remaining protein solutions was diluted with 1 volume glycerol and stored at -20°C. Protein purification was performed on a GE Healthcare AKTA FPLC, using chromatography columns from GE Healthcare.

**ssM13 Preparation and *De Novo* DNA Synthesis Study**

Pelleted M13 virus was suspended in 4 mL TE buffer, then combined with 2 volumes of 0.2 M NaOH, 1% SDS and mixed by swirling gently. Solution was combined with 1.5 volumes

potassium acetate (pH 5.05), gently swirled, and incubated in an ice water bath for 15 minutes. Solution was transferred to a sterile Oak Ridge tube and centrifuged for 10 minutes at 13,000 RPM at 4°C. Supernatant was transferred to a sterile Falcon tube, combined with 1 volume of 100% ethanol at -20°C, gently mixed, and incubated on ice for 30 minutes.

Solution was transferred to a sterile Oak Ridge tube and centrifuged for 20 minutes at 13,000 RPM at 4°C, after which a white DNA pellet was observed. Supernatant was poured off and the DNA was washed with one half-volume of 75% ethanol at -20°C, inverted to mix, and spun for 10 minutes at 13,000 RPM at 4°C. Supernatant was poured off and DNA was allowed to air dry. Aliquots of ssM13 were resuspended in 100 μL TE buffer overnight at 4°C without mixing. Purity was confirmed by 0.8% agarose gel.

Purified ssM13 DNA was reconcentrated by ethanol precipitation. 4 M NaCl was added to each aliquot to a final concentration of 0.2 M NaCl. Two volumes of 100% ethanol at -20°C were added to each and then aliquots were combined into a single 1.5 mL microcentrifuge tube. The sample was centrifuged for 14 minutes at 14,500 RPM at 4°C. The supernatant was poured off, and the DNA was washed by adding 500 μL of 70% ethanol at -20°C and spun an additional 4 minutes at 14,500 RPM at 4°C. A white DNA pellet was observed. Tube was inverted to dry and ssM13 DNA was resuspended in 25 μL TE buffer.

The *de novo* DNA synthesis assay was performed with the purified ssM13 DNA. Studies were performed in a 37°C water bath in 1.5 mL microcentrifuge tubes at a protein (Larva64) concentration of 2.5 μM. 40μL reactions were carried out under the following conditions: 25 mM Tris pH 7.5, 50 mM NaCl, 1 mM DTT, 100 μg/mL BSA, and 20 nM ssM13. Variable components included 10/15 mM $(Mg/Mn)Cl_2$ as indicated, 1 mM dNTPs, 1 mM ATP, 10 mM ATP and 10 μM $ZnSO_4$ as indicated. Reactions were quenched with a 2X buffer containing

100 mM EDTA, 0.4% SDS, and 10% glycerol. Samples were analyzed by running 18 μL of each timepoint overnight on a 0.8% agarose gel in a buffer-recirculating gel rig at 20V in TAE buffer before imaging. Ultraviolet imaging of gels was performed on a Biorad Chemi-Doc.

**DNA Binding Study**

DNA binding studies were carried out at a mean temperature of 25°C with 0.02 μM - 10.00 μM protein under the following conditions: 20 mM Tris pH 7.5, 200 mM NaCl, 1 mM DTT, 100 μg/mL BSA, 10 nM DNA, and 1 mM nucleotide. Substrates were annealed onto a 3'-FAM-labelled 30mer oligonucleotide (5'-GCTGGCTGGGCTAGGGGATTTCTCCAGGGT/36-FAM/-3'). Reaction components were mixed and allowed to equilibrate at room temperature for 30 minutes before data was collected. Binding strength was measured by fluorescence polarization at $\lambda_{ex}$ = 490 nm and $\lambda_{em}$ = 530 nm.

Data were collected in triplicate on a Molecular Devices Spectramax iD5 96-well plate reader in a 96-well Corning half-area opaque plate (14.2 mm height), read by row, with a 400 ms integration time, 1 mm read height, and automatic PMT gain. SoftMax Pro 7.1.0 build 246936 was used to visualize and fit binding data. $K_d$ (binding affinity) was calculated based on a 4-parameter logistic curve fit,

$$y = D + \frac{A - D}{1 + (\frac{x}{C})^B}$$

where $y$ is the polarization in mP, $x$ is the protein concentration in μM, $A$ is the lower limit of $y$, $D$ is the upper limit of $y$, $C$ is the EC50 ($K_d$) in μM, and $B$ control the rate of growth or decay of exponential increase on either side of $x = C$. The equation parameters ($A$, $B$, $C$, $D$, $R^2$) were estimated at (97.81, 2.835, *0.048*, 293.3, 0.986) for ssDNA, (123.3, 3.777, *0.979*, 274.9, 0.990)

for 5'-tail dsDNA, and (67.88, 2.235, *1.610*, 257.1, 0.995) for forked-end dsDNA. $K_d$ values are italicized.

**Bacterial Two-Hybrid Analysis**

The *M. smegmatis* two-hybrid protocol was created and provided by Dr. Danielle Heller at the HHMI SEA-PHAGES program and is included as a supplemental document. A modified version of the two-hybrid protocol has been developed by Dr. Heller since this study was performed.

**Other Work**

CRISPRi work was performed by Shelby Watson in the Wallen research group at Western Carolina University. DNA translocation results and figures were provided by Dr. Justin Miller at Middle Tennessee State University as part of a collaboration with Western Carolina University (unpublished). Divalent cation preference was tested at room temperature in the presence of 20 nM ssDNA 30-mer in a buffer containing 25 mM Tris pH 7.5, 50 mM NaCl, 1 mM DTT, 500 μg/mL BSA, 1 mM ATP, and 15 mM of either $Mg^{2+}$ or $Mn^{2+}$ at a protein concentration of 2.5 μM.

REFERENCES


(1) Iyer, L. M.; Koonin, E. V.; Leipe, D. D.; Aravind, L. Origin and Evolution of the Archaeo-Eukaryotic Primase Superfamily and Related Palm-Domain Proteins: Structural Insights and New Members. *Nucleic Acids Res.* **2005**, *33* (12), 3875–3896.

(2) Lipps, G.; Röther, S.; Hart, C.; Krauss, G. A Novel Type of Replicative Enzyme Harbouring ATPase, Primase and DNA Polymerase Activity. *EMBO J.* **2003**, *22* (10), 2516–2525.

(3) Lipps, G. Molecular Biology of the PRN1 Plasmid from Sulfolobus Islandicus . *Biochem. Soc. Trans.* **2009**, *37* (1), 42–45.

(4) Berkner, S.; Hinojosa, M. P.; Prangishvili, D.; Lipps, G. Identification of the Minimal Replicon and the Origin of Replication of the Crenarchaeal Plasmid PRN1. *Microbiologyopen* **2014**, *3* (5), 688–701.

(5) Beck, K.; Vannini, A.; Cramer, P.; Lipps, G. The Archaeo-Eukaryotic Primase of Plasmid PRN1 Requires a Helix Bundle Domain for Faithful Primer Synthesis. *Nucleic Acids Res.* **2010**, *38* (19), 6707–6718.

(6) Guo, H.; Li, M.; Wang, T.; Wu, H.; Zhou, H.; Xu, C.; Yu, F.; Liu, X.; He, J. Crystal Structure and Biochemical Studies of the Bifunctional DNA Primase-Polymerase from Phage NrS-1. *Biochem. Biophys. Res. Commun.* **2019**, *510* (4), 573–579.

(7) Guo, H.; Li, M.; Wu, H.; Wang, W.; Yu, F.; He, J. Crystal Structures of Phage NrS-1 N300-DNTPs-Mg2+ Complex Provide Molecular Mechanisms for Substrate Specificity. *Biochem. Biophys. Res. Commun.* **2019**, *515* (4), 551–557.

(8) García-Gómez, S.; Reyes, A.; Martínez-Jiménez, M. I.; Chocrón, E. S.; Mourón, S.;

Terrados, G.; Powell, C.; Salido, E.; Méndez, J.; Holt, I. J.; et al. PrimPol, an Archaic Primase/Polymerase Operating in Human Cells. *Mol. Cell* **2013**, *52* (4), 541–553.

(9) Keen, B. A.; Jozwiakowski, S. K.; Bailey, L. J.; Bianchi, J.; Doherty, A. J. Molecular Dissection of the Domain Architecture and Catalytic Activities of Human PrimPol. *Nucleic Acids Res.* **2014**, *42* (9), 5830–5845.

(10) Rudd, S. G.; Bianchi, J.; Doherty, A. J. PrimPol—A New Polymerase on the Block. *Mol. Cell. Oncol.* **2014**, *1* (2), 1–10.

(11) Martínez-Jiménez, M. I.; Calvo, P. A.; García-Gómez, S.; Guerra-González, S.; Blanco, L. The Zn-Finger Domain of Human PrimPol Is Required to Stabilize the Initiating Nucleotide during DNA Priming. *Nucleic Acids Res.* **2018**, *46* (8), 4138–4151.

(12) Iyer, L. M.; Leipe, D. D.; Koonin, E. V.; Aravind, L. Evolutionary History and Higher Order Classification of AAA+ ATPases. *J. Struct. Biol.* **2004**, *146* (1–2), 11–31.

(13) Miller, J. M.; Enemark, E. J. Fundamental Characteristics of AAA+ Protein Family Structure and Function. *Archaea* **2016**, *2016*.

(14) Hickman, A. B.; Dyda, F. Binding and Unwinding: SF3 Viral Helicases. *Curr. Opin. Struct. Biol.* **2005**, *15* (1 SPEC. ISS.), 77–85.

(15) Chen, X.; Su, S.; Chen, Y.; Gao, Y.; Li, Y.; Shao, Z.; Zhang, Y.; Shao, Q.; Liu, H.; Li, J.; et al. Structural Studies Reveal a Ring-Shaped Architecture of Deep-Sea Vent Phage NrS-1 Polymerase. *Nucleic Acids Res.* **2020**, *48* (6), 3343–3355.

(16) Caruthers, J. M.; McKay, D. B. Helicase Structure and Mechanism. *Curr. Opin. Struct. Biol.* **2002**, *12* (1), 123–133.

(17) Xu, H.; Sträter, N.; Schröder, W.; Böttcher, C.; Ludwig, K.; Saenger, W. Structure of DNA Helicase RepA in Complex with Sulfate at 1.95 Å Resolution Implicates Structural

Changes to an “open” Form. *Acta Crystallogr. - Sect. D Biol. Crystallogr.* **2003**, *59* (5), 815–822.

(18) McGrew, D. A.; Knight, K. L. Molecular Design and Functional Organization of the RecA Protein. *Crit. Rev. Biochem. Mol. Biol.* **2003**, *38* (5), 385–432.

(19) Masui, R.; Mikawa, T.; Kato, R.; Kuramitsu, S. Characterization of the Oligomeric States of RecA Protein: Monomeric RecA Protein Can Form a Nucleoprotein Filament. *Biochemistry* **1998**, *37* (42), 14788–14797.

(20) Rajan, R.; Bell, C. E. Crystal Structure of RecA from Deinococcus Radiodurans: Insights into the Structural Basis of Extreme Radioresistance. *J. Mol. Biol.* **2004**, *344* (4), 951–963.

(21) Datta, S.; Krishna, R.; Ganesh, N.; Chandra, N. R.; Muniyappa, K.; Vijayan, M. Crystal Structures of Mycobacterium Smegmatis RecA and Its Nucleotide Complexes. *J. Bacteriol.* **2003**, *185* (14), 4280–4284.

(22) Hatfull, G. F. *The Secret Lives of Mycobacteriophages*, 1st ed.; Elsevier Inc., 2012; Vol. 82.

(23) Roach, D. R.; Debarbieux, L. Phage Therapy: Awakening a Sleeping Giant. *Emerg. Top. Life Sci.* **2017**, *1* (1), 93–103.

(24) Halgasova, N.; Mesarosova, I.; Bukovska, G. Identification of a Bifunctional Primase-Polymerase Domain of Corynephage BFK20 Replication Protein Gp43. *Virus Res.* **2012**, *163* (2), 454–460.

(25) Hatfull, G. F. Mycobacteriophages. *Microbiol. Spectr.* **2018**, *6* (5).

(26) Dana, R.; Gray, V. The Actinobacteriophage Database: Larva https://phagesdb.org/phages/Larva/ (accessed Oct 29, 2019).

(27) Lu, S.; Wang, J.; Chitsaz, F.; Derbyshire, M. K.; Geer, R. C.; Gonzales, N. R.; Gwadz,

M.; Hurwitz, D. I.; Marchler, G. H.; Song, J. S.; et al. CDD/SPARCLE: The Conserved Domain Database in 2020. *Nucleic Acids Res.* **2020**, *48* (D1), D265–D268.

(28) Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. E. The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis. *Nat. Protoc.* **2015**, *10* (6), 845–858.

(29) Prabu, J. R.; Manjunath, G. P.; Chandra, N. R.; Muniyappa, K.; Vijayan, M. Functionally Important Movements in RecA Molecules and Filaments: Studies Involving Mutation and Environmental Changes. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2008**, *64* (11), 1146–1157.

(30) Shin, D. S.; Pellegrini, L.; Daniels, D. S.; Yelent, B.; Craig, L.; Bates, D.; Yu, D. S.; Shivji, M. K.; Hitomi, C.; Arvai, A. S.; et al. Full-Length Archaeal Rad51 Structure and Mutants: Mechanisms for RAD51 Assembly and Control by BRCA2. *EMBO J.* **2003**, *22* (17), 4566–4576.

(31) Short, J. M.; Liu, Y.; Chen, S.; Soni, N.; Madhusudhan, M. S.; Shivji, M. K. K.; Venkitaraman, A. R. High-Resolution Structure of the Presynaptic RAD51 Filament on Single-Stranded DNA by Electron Cryo-Microscopy. *Nucleic Acids Res.* **2016**, *44* (19), 9017–9030.

(32) James, J. A.; Aggarwal, A. K.; Linden, R. M.; Escalante, C. R. Structure of Adeno-Associated Virus Type 2 Rep40-ADP Complex: Insight into Nucleotide Recognition and Catalysis by Superfamily 3 Helicases. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (34), 12455–12460.

(33) Enemark, E. J.; Joshua-Tor, L. Mechanism of DNA Translocation in a Replicative Hexameric Helicase. *Nature* **2006**, *442* (7100), 270–275.

(34) Sanders, C. M.; Kovalevskiy, O. V; Sizov, D.; Lebedev, A. A.; Isupov, M. N.; Antson, A. A. Papillomavirus E1 Helicase Assembly Maintains an Asymmetric State in the Absence of DNA and Nucleotide Cofactors. *Nucleic Acids Res.* **2007**, *35* (19), 6451–6457.

(35) Gajiwala, K. S.; Chen, H.; Cornille, F.; Roques, B. P.; Reith, W.; Mach, B.; Burley, S. K. Structure of the Winged-Helix Protein HRFX1 Reveals a New Mode of DNA Binding. *Nature* **2000**, *403* (6772), 916–921.

(36) Ziegelin, G.; Lanka, E. Bacteriophage P4 DNA Replication. *FEMS Microbiol. Rev.* **1995**, *17* (1–2), 99–107.

(37) Kazlauskas, D.; Krupovic, M.; Venclovas, C. The Logic of DNA Replication in Double-Stranded DNA Viruses: Insights from Global Analysis of Viral Genomes. *Nucleic Acids Res.* **2016**, *44* (10), 4551–4564.

(38) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410.

(39) Lipps, G.; Weinzierl, A. O.; von Scheven, G.; Buchen, C.; Cramer, P. Structure of a Bifunctional DNA Primase-Polymerase. *Nat Struct Mol Biol* **2004**, *11* (2), 157–162.

(40) Zhu, B.; Wang, L.; Mitsunobu, H.; Lu, X.; Hernandez, A. J.; Yoshida-Takashima, Y.; Nunoura, T.; Tabor, S.; Richardson, C. C. Deep-Sea Vent Phage DNA Polymerase Specifically Initiates DNA Synthesis in the Absence of Primers. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (2), E2310–E2318.

(41) Díaz-Talavera, A.; Calvo, P. A.; González-Acosta, D.; Díaz, M.; Sastre-Moreno, G.; Blanco-Franco, L.; Guerra, S.; Martínez-Jiménez, M. I.; Méndez, J.; Blanco, L. A Cancer-Associated Point Mutation Disables the Steric Gate of Human PrimPol. *Sci. Rep.* **2019**, *9* (1), 1–13.

(42) Hanson, P. I.; Whiteheart, S. W. AAA+ Proteins: Have Engine, Will Work. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (7), 519–529.

(43) Brambley, C. A.; Marsee, J. D.; Halper, N.; Miller, J. M. Characterization of Mitochondrial YME1L Protease Oxidative Stress-Induced Conformational State. *J. Mol. Biol.* **2019**, *431* (6), 1250–1266.

(44) Lovett, S. T. Polymerase Switching in DNA Replication. *Mol. Cell* **2007**, *27* (4), 523–526.

(45) Yang, Y.; Thomas, J.; Li, Y.; Vilchèze, C.; Derbyshire, K. M.; Jacobs, W. R.; Ojha, A. K. Defining a Temporal Order of Genetic Requirements for Development of Mycobacterial Biofilms. *Mol. Microbiol.* **2017**, *105* (5), 794–809.

(46) Alexander-Kaufman, K.; Harper, C. Transketolase: Observations in Alcohol-Related Brain Damage Research. *Int. J. Biochem. Cell Biol.* **2009**, *41* (4), 717–720.

(47) Koon, N.; Squire, C. J.; Baker, E. N. Crystal Structure of LeuA from Mycobacterium Tuberculosis, a Key Enzyme in Leucine Biosynthesis. *Proc. Natl. Acad. Sci.* **2004**, *101* (22), 8295–8300.

(48) Trutneva, K.; Shleeva, M.; Nikitushkin, V.; Demina, G.; Kaprelyants, A. Protein Composition of Mycobacterium Smegmatis Differs Significantly between Active Cells and Dormant Cells with Ovoid Morphology. *Front. Microbiol.* **2018**, *9* (SEP), 1–14.

(49) Rock, J. M.; Hopkins, F. F.; Chavez, A.; Diallo, M.; Chase, M. R.; Gerrick, E. R.; Pritchard, J. R.; Church, G. M.; Rubin, E. J.; Sassetti, C. M.; et al. Programmable Transcriptional Repression in Mycobacteria Using an Orthogonal CRISPR Interference Platform. *Nat. Microbiol.* **2017**, *2* (February), 1–9.

(50) Säbel, C. E.; Neureuther, J. M.; Siemann, S. A Spectrophotometric Method for the Determination of Zinc, Copper, and Cobalt Ions in Metalloproteins Using Zincon. *Anal.*

*Biochem.* **2010**, *397* (2), 218–226.

(51) Sanchez, M.; Drechsler, M.; Stark, H.; Lipps, G. DNA Translocation Activity of the Multifunctional Replication Protein ORF904 from the Archaeal Plasmid PRN1. *Nucleic Acids Res.* **2009**, *37* (20), 6831–6848.

(52) Madeira, F.; Park, Y. M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A. R. N.; Potter, S. C.; Finn, R. D.; et al. The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019. *Nucleic Acids Res.* **2019**, *47* (W1), W636–W641.

(53) Okonechnikov, K.; Golosova, O.; Fursov, M.; Varlamov, A.; Vaskin, Y.; Efremov, I.; German Grehov, O. G.; Kandrov, D.; Rasputin, K.; Syabro, M.; et al. Unipro UGENE: A Unified Bioinformatics Toolkit. *Bioinformatics* **2012**, *28* (8), 1166–1167.

(54) Letunic, I.; Bork, P. Interactive Tree Of Life (ITOL) v4: Recent Updates and New Developments. *Nucleic Acids Res.* **2019**, *47* (W1), W256–W259.

(55) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303.

(56) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (4), 486–501.

(57) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.

(58) Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A. Protein Identification and Analysis Tools on the ExPASy Server. In *The*

*Proteomics Protocols Handbook*; Walker, J. M., Ed.; Humana Press Inc.: Totowa, NJ, 2005; pp 571–608.