SEQUENCE AND FUNCTIONAL ANALYSES OF DNA POLYMERASE A ENZYMES FROM BACTERIOPHAGES TO YEAST

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

Cecilia Ann Baumgardner

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

Committee Members: Dr. Indrani Bose, Biology Department
Dr. Maria Gainey, Chemistry and Physics Department

April 2021

ACKNOWLEGEMENTS

I would like to thank Dr. Jamie Wallen, for allowing me to join his lab and introducing me to the world of biochemistry and polymerases, Dr. Indrani Bose for assisting me with the *Cryptococcus neoformans* and Dr. Maria Gainey for introducing me to bacteriophages and bioinformatics. Without their expertise and patience, this would have not been possible. The entire Western Carolina Chemistry and Biology Departments have been incredibly welcoming and supportive, allowing me to not only deepen my love of chemistry but teaching as well.

I would also like to thank Nathan Folse and Sam Walter, for their support within and outside of the lab.

Without my family, especially Carrie, my sister, and my parents, I would not have been able to make it here. My appreciation and gratitude for them is exponential., as all of their support motivated me and paved the path for me to complete this program.

Finally, I would like to thank my amazing friends, including Josh York, Michael Bond, Daniel Free, Zay Watkins, Lori Neri, Vance Renaud, Jesse Ingham and so many others. I will always be grateful for the support, encouragement, and adventures we had.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

SEQUENCE AND FUNCTIONAL ANALYSES OF DNA POLYMERASE A ENZYMES FROM BACTERIOPHAGES TO YEAST

Cecilia Ann Baumgardner, M.S.

Western Carolina University (April 2021)

Director: Dr. Jamie Wallen

DNA polymerases are essential enzymes required to accurately and rapidly copy genetic material. Although they conserve a common function of duplicating DNA, DNA polymerases are incredibly diverse enzymes. The polymerase A (PolA) family of DNA polymerases perform a wide range of functions in diverse systems ranging from viruses to eukaryotic cells. PolAs at a minimum contain a polymerase domain that synthesizes DNA and a 3'-5' exonuclease domain that remove mistakes made during replication. PolAs have evolved to obtain additional protein domains that allow them to perform functions ranging from synthesizing short stretches of DNA during DNA repair to being the replicative polymerase tasked with duplicating the entire genome. The goals of this work are to compare PolA enzymes from both bacteriophages and fungi to better understand how these enzymes use additional protein domains for specialized function. Bacteriophages are viruses that only infect bacteria, and due to recent work by undergraduate students that are part of HHMI's Science Education Alliance, there are now 3,731 fully sequenced Actinobacteriophage genomes. Analysis of these genomes have revealed that many, but not all, of these phages contain a PolA enzyme, and we have become interested in the function of these enzymes in phage replication and repair. Our analysis identified 1,351 PolA sequences in viruses that infect several different bacterial hosts, and these proteins range in size

from as little as 582 to as large as 801 residues. These enzymes appear to conserve essential catalytic residues needed for polymerase and exonuclease activities, and the variation in protein sizes are due to additional novel domains present in these polymerases. These domains include glycosylase domains, which implicate these enzymes in DNA repair. The goals of this work are to compare these enzymes to the well-studied PolA from bacteriophage T7 to better understand how PolA enzymes have evolved in bacteriophages. Another member of the PolA family is DNA polymerase gamma (PolG), which is a replicative polymerase found in the mitochondria. The second part of this work focuses on the PolG enzyme from the pathogenic fungus *Cryptococcus neoformans*, which causes cryptococcosis of the lungs and central nervous system. The *C. neoformans* PolG enzyme has yet to be biochemically studied, and we have identified two novel domains that are not present in other PolGs. Due to their location and amino acid content, we predict that these domains either bind DNA or to partner proteins during mitochondrial DNA replication. The goals of this work are to perform extensive bioinformatics analysis of the novel domains to learn more about possible function and to confirm whether these novel domains bind partner proteins. The N-terminal domain has high sequence conservation and is only found in *Cryptococcus* species, while the internal domain, located between the spacer and the polymerase domains, lacks residue conservation and is found in a variety of species. We have successfully expressed and purified both the full-length enzyme as well as just the N-terminal domain, and pull-down assays were optimized to identify cryptococcal proteins that bind to PolG.

CHAPTER ONE: INTRODUCTION

DNA polymerases are tasked with not only replicating but also maintaining the integrity of a cell's genome.[1] To keep up with the constant demand of DNA replication, DNA polymerases must be very efficient. While replication times vary, *Escherichia Coli* replicates at a rate of 1000 nucleotides per second, while the human (*Homo sapiens*) DNA polymerase replicates at a rate of 50 nucleotides per second.[2] While they operate very quickly, they must balance efficiency with accuracy. Any mistakes during replication could cause genetic defects and diseases.

Since the first DNA polymerase (*E. coli*'s DNA polymerase I) was discovered in 1958, multiple types of DNA polymerases have been discovered. Some DNA polymerases are used for replication, some are used for repair, and some perform both functions. Generally, they are sorted into eight families (A, B, C, D, X, Y, RT and PrimPol) based on their sequences and domains.[3] This study focuses exclusively on family A DNA polymerases. Family A DNA polymerases contain a polymerase domain, which functions to synthesize new DNA, and a 3'-5' exonuclease domain, which corrects mistakes made by DNA polymerase during replication. The polymerases can be further classified as replicative or repair polymerases. The replicative polymerases have a working 3'-5' exonuclease and an overall error rate of 10^{-5}-10^{-7}. Alternatively, the repair polymerases have a 3'-5' exonuclease but are missing key residues, rendering the exonuclease domain inactive. The error rate of these polymerases is 10^{-3}-10^{-4}.[3]

Family A includes DNA polymerases I, gamma, theta and nu. These polymerases vary in species and roles within their respective hosts. Bacterial polymerases within this family have more minor roles in replication. For example, the primary function of *E.* coli's DNA polymerase

I is to assist in the repair of DNA.[4] T7, a bacteriophage, contains a replicative polymerase that performs all of its replication. There are three human family A DNA polymerases. DNA polymerase gamma replicates all of the mitochondrial genome, while DNA polymerases theta and nu are repair polymerases. These repair polymerases are the only family A DNA polymerases present in the nucleus.[1,5]

This study will go in-depth on two specific types of family A DNA polymerases, the bacteriophage DNA polymerases found in actinobacteriophages, and the DNA polymerase gamma of *Cryptococcus neoformans.* Bacteriophages, or phages, are viruses that infect bacteria, and they are the most abundant biological entities in nature.[6] Since they destroy their host, bacteriophages are being considered as potential therapeutic alternatives for antibiotic-resistant bacteria. The genomes of many actinobacteriophages have been sequenced (www.seaphages.org), and bioinformatic analyses of these genomes have shown that many of the bacteriophages contain a family A DNA polymerase. Currently, the database has these proteins annotated as "DNA polymerase" or "DNA polymerase I". It is currently unclear if these enzymes are replicative in nature (like polymerase gamma and T7 polymerase) and replicate the phage genomes, or if they have specialized roles in DNA repair. The goal for this portion of the study was to determine if these proteins are indeed DNA polymerase I enzymes, despite the different names, and to determine if these proteins were actually similar to the *E. coli* DNA polymerase I. While DNA polymerase I is a model family A polymerase, it contains an additional domain, a 5'-3' exonuclease, that is not expected to be in bacteriophage polymerases.[7]

Cryptococcus neoformans is an encapsulated, pathogenic fungus. This fungus infects humans through respiration, causing cryptococcosis in the central nervous system and lungs.[8,9] Traveling through the bloodstream, the infection can spread to the brain, causing cryptococcal

meningoencephalitis.[8] Once the fungus is inhaled, it can lie dormant for years until the host's immune system is weak enough to launch an infection. Symptoms are usually caused by this reactivation of the infection, instead of the primary infection.[8] Unfortunately, this fungus is very common and can be found in a wide variety of areas, including the soil, decomposing trees and bird feces.[8,10] Despite the common occurrence of the fungus, most people do not contract this fungal infection, as the fungus primarily infects people with suppressed immune systems. In fact, it is the most common fungal infection that attacks the central nervous system of immunocompromised patients, especially HIV positive patients.[8,9,10] Within the general population of the United States, the yearly cryptococcosis incidence is 0.4 to 1.3 cases per 100,000 people. Within the HIV positive community, the yearly incidence jumps to 200 to 700 cases per 100,000 people.[8]

Diagnostic tests for a cryptococcal infection include ELISA, latex agglutination or examining fresh cerebrospinal fluid with China Ink for the presence of the yeast's capsule and a cerebrospinal fluid culture.[9] There are some treatments for cryptococcus infections. However, the treatments take at least six months to work and can be very expensive.[10] This results in a very high mortality rate in developing countries. The CDC estimates that 220,000 cases of cryptococcal meningitis occur each year in developing countries, with nearly 181,000 fatalities.[10] Most of these cases occur in sub-Saharan Africa, where they cannot afford proper treatment or diagnostic screenings.[10,11]

C. neoformans is an obligate aerobe, meaning it requires oxygen to survive. [12,13] Mitochondria are the energy powerhouses of the cell, using oxygen as the final electron acceptor in the production of ATP. Within the mitochondria, DNA polymerase gamma is responsible for accurately replicating the mitochondrial DNA.[14] The enzyme has been proven to be essential for

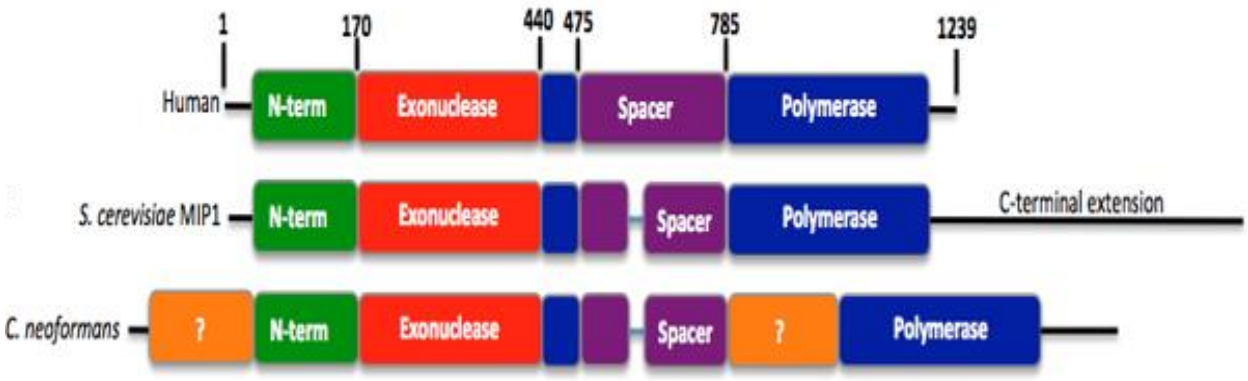


Figure 1. Block Diagram of DNA polymerase gamma in *H. sapiens* and *C. neoformans*, and MIP1 in *S. cerevisiae*. The sequences are aligned and color-coded according to the domain. The orange domains in the *C. neoformans* polymerase represent the two novel domains. The first domain(left) is referred to as the 1-135 domain, while the second domain (right) is referred to as the INT domain.

the fungus to survive (Dr. Indi Bose, unpublished result). Therefore, it could be possible to develop a drug that inhibits the polymerase gamma. However, it would need to be specific to the *C. neoformans* enzyme as humans also have an essential DNA polymerase gamma. [15]

The human DNA polymerase consists of four domains, a N-terminal domain (which includes the mitochondrial localization sequence), an exonuclease domain, a spacer domain and a polymerase domain. The spacer domain is further divided into an intrinsic processivity domain (IP) and an extended accessory-interacting determinant (AID) domain.[16] The IP domain provides the intrinsic processivity of the polymerase while the AID increases processivity through binding to the accessory subunits.[17] Through bioinformatics it has been determined that the *C. neoformans* polymerase gamma (CNPolG), human polymerase gamma and *Saccharomyces cerevisiae* MIP1 are homologous. This spacer is known to be conserved in other species but mostly missing in fungal species.[16] Also, CNPolG has two domains that are absent in the human and *Saccharomyces cerevisiae* (a model fungal species) DNA polymerase gamma. Block

diagrams of these architectures are shown in figure 1. These two domains, colored orange in figure 1, have no known function. Therefore, these two domains could serve as a target for a novel fungal treatment. The goal of the latter half of the study was to characterize the two novel domains using a mixture of bioinformatics and pull-down assays.

CHAPTER TWO: BACTERIOPHAGE BIOINFORMATICS

As new phages are discovered and sequenced, they are documented in the actinobacteriophage database on PhagesDB.org. Due to the constant discovery of new phages, a cutoff date for our data analysis had to be established. On January 25th, 2020, the sequences and phage information for this project were collected from the PhagesDB website.[18] Since the collection date, more phages have been added to the database and assigned to the phamilies used in this data set. These new phages were not included in this study. The database was searched using MySQL to identify all genes that had DNA polymerase 1 in the description. With most of these phage polymerases annotated as "DNA polymerase" or "DNA polymerase I", confusion can arise. Not only is the naming inconsistent, calling them "DNA polymerase I" indicates that they are similar to other DNA polymerase I. The model DNA polymerase I is found in *E. coli*. This polymerase contains a 5'-3' exonuclease, 3'-5' exonuclease and polymerase domains. While the phage polymerases have the 3'-5' exonuclease and polymerase domains, they are not known to have the 5'-3' exonuclease. If they do not have this exonuclease domain, the naming should be differentiated to alleviate any potential confusion.

When a phage is sequenced, the genes are given a pham number to group similar genes together. The MySQL search for DNA polymerase I produced a list of 13 phamilies, or phams, that contained a total of 1351 sequences (See Appendix). While the sequenced genes are grouped into phams, the phages are grouped into clusters based on similarities in their full nucleotide sequences.[19] Phages containing proteins annotated as DNA polymerase I are classified among 76 different clusters and are not exclusive to certain species of isolation host. They were isolated from a wide variety of hosts from genera *Mycobacterium, Gordonia, Arthrobacter, Streptomyces, Rhodococcus, Microbacterium, Corynebacterium, Tsukamurella* and

Brevibacterium. Most of the phages were isolated in *Mycobacterium smegmatis mc^{2}155*, however this is due to an increase of testing using *M. smegmatis* over other possible hosts. The proteins range in size from 582 amino acids to 891 amino acids, with an average of 626 amino acids.

Phylogenetic Analysis

Based on the polymerase sequences, UGENE was able to build the phylogenetic tree shown in figure 2. The actinophages separated into six clades. Interestingly, within the clades,

Table 1. Location of the bacteriophage clusters within the phylogenetic tree.

Color	Figure 2 Letter	Phams	Clusters
Red	A	6914	AQ
Orange	A	92929, 95411, 19223, 4548	EP, S, EK1, EK, EK2, EM, BG, BM
Yellow	A	91746	B1
	B	91746	B1
	C	91746	B1, S
	D	91746	B1, B13, B6, B11, B10, B7
	E	91746	B5, B12, B9
	F	91746	B4, B8
	G	91746	B2
	H	91746	B3, DR, S, CD
Green	A	22255, 106021, 58669	AV, AB, S, CT, GC
	B	106021	EA1
	C	106021	EA1
	D	106021	EA10, EA2, EA5, EA, EA6, EA11, EA4, EA9, EA8, EA3
Blue	A	99815, 30783	AK, EJ, DA, BH, BQ
	B	92105	DK, DS, FC, GD
Purple	A	95079	AZ, EB, BB1, BB2, BL
	B	95079	S
	C	95079	A1
	D	95079	A2, A17, A12, A6, A13, A11, A9, A14, A16, A15, A20, A5, A8

the phages are further divided by pham and again by cluster. This division is alluded to with the visual separation within the clades. In the tree, the clades are color-coded and the largest

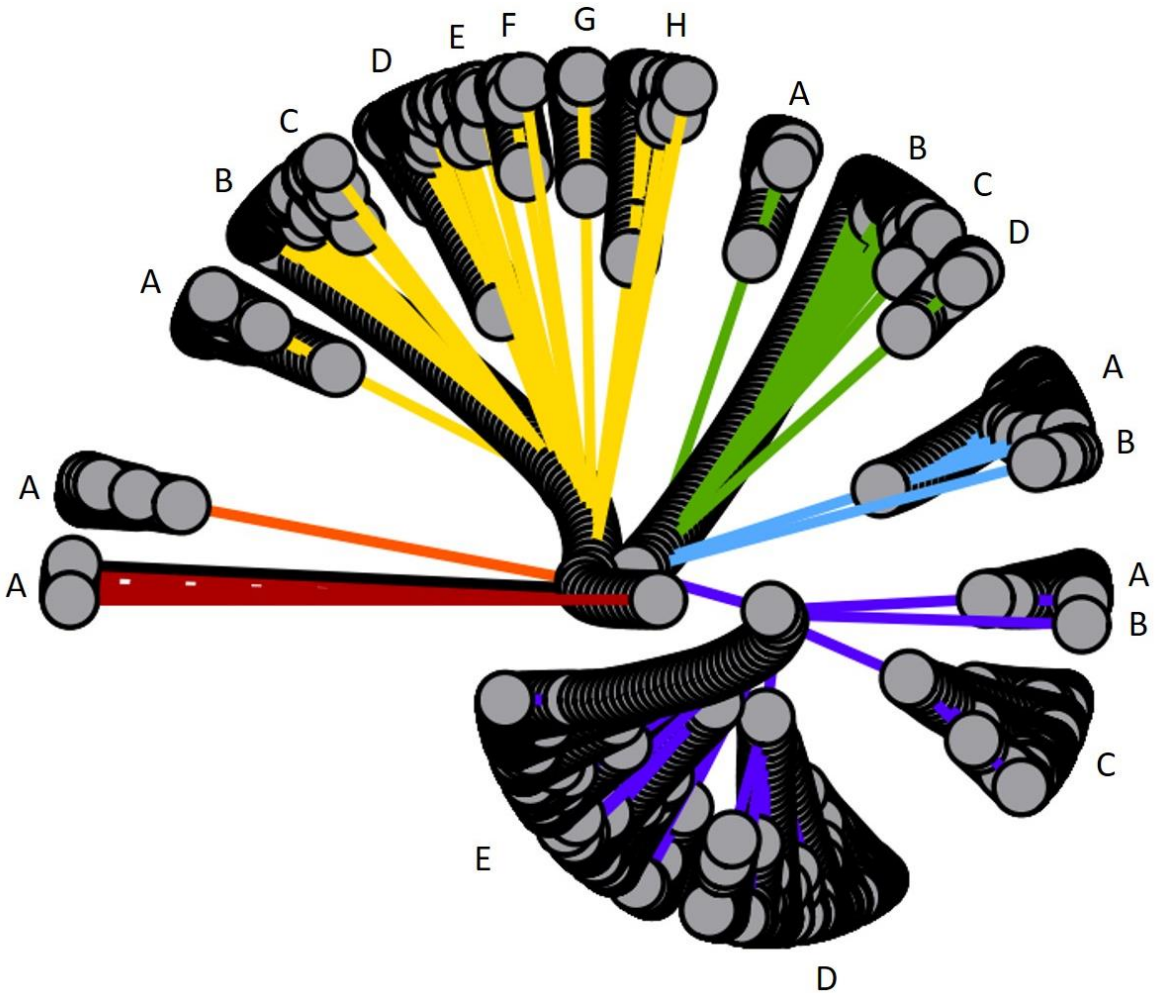


Figure 2. Phylogenetic Tree of the Actinophage DNA Polymerases. This phylogenetic tree is based on the amino acid sequences of the 1351 actinophage DNA polymerases as well as the T7 DNA polymerase. It is color-coded by clade. T7's DNA polymerase is denoted by a black branch. The red clade contains pham 6914. The orange clade contains phams 92929, 95411, 19223 and 4548. The yellow clade contains pham 91746. The green clade contains pham 22255, 106021, and 58669. The blue clade contains phams 99815, 58669 and 92105. Finally, the purple clade contains pham 95079.

separations of each clade is labeled. Table 1 summarizes the location of each cluster within the phylogenetic tree using these labels, where S refers to singletons.

Phyre^2

The Protein Homology/Analogy Recognition Engine V 2.0, or Phrye^2, is a website that predicts a 3D structure based on a protein's sequence and known protein structures.[20] 174 of the sequences were chosen as representatives of their respective phams. These representatives were chosen by first aligning the sequences within UGENE. Then each cluster was analyzed for outliers or variations. Based on the results of this analysis, at least two phages were chosen to be modeled as well as any outliers that had potentially unique domains. After performing an intensive Phyre^2 search for each sequence, the program produced a 3D structure and a list of homologues. These homologues can provide potential functions for proteins. Looking at the top 10 homologues for each of the Phyre^2 runs, there were many similarities to note (Table 1). At least the top 10 homologues for each sequence had 100% confidence that the template structures have true homology with the query. These results had protein sequence identities ranging from 14% to 34%. While these percentages seem low, identities above 15% are considered acceptable. High accuracy models are suggested to have 30-40% sequence identities. When high confidence scores are paired with lower identity percentages, it is more likely that these hits represent remote homology. Therefore, the server has provided a structure that adopts the overall fold and has a similar core, but it may be less accurate in modeling the shape of the surface loops.[20]

Out of the 174 phage proteins modeled, 169 had the same top 10 hits, which are detailed in Table 1. The other 5 phage proteins contained the same match list except for the 10th hit. The 10th hit was either d2hhva2 or d1qtma2. Both of these hits are domain matches to DNA polymerase 1. Out of the top 10 hits, all of the molecules were family A DNA polymerases, specifically T7's DNA polymerase, DNA polymerase I, DNA polymerase nu and DNA

Table 2. Top ten Phyre2 matches to the bacteriophage DNA polymerases.

Template	PDB Header	Chain	PBD Molecule	PBD Title
c6vdeA_	transferase	A	DNA polymerase i	full-length *M.smegmatis* pol1
c4xviA_	transferase/DNA	A	DNA polymerase nu	binary complex of human polymerase nu and DNA with the finger domain2 ajar
c4x0pB_	transferase/DNA	B	DNA polymerase theta	ternary complex of human DNA polymerase theta cterminal domain2 binding ddatp opposite a tetrahydrofuran ap site analog
c1njzA_	transferase/DNA	A	DNA polymerase i	cytosine-thymine mismatch at the polymerase active site
c6vddD_	transferase/DNA	D	DNA polymerase i	pol domain of pol1 from *M. smegmatis* complex with DNA primer-template2 and dntp
c4ktqA_	transferase/DNA	A	protein (large fragment of DNA polymerase i)	binary complex of the large fragment of DNA polymerase i2 from *T. aquaticus* bound to a primer/template DNA
c2kzzA_	transferase/DNA	A	protein (DNA polymerase i)	klenow fragment with normal substrate and zinc only
c1cmwA_	transferase	A	protein (DNA polymerase i)	crystal structure of taq DNA-polymerase shows a new orientation for2 the structure-specific nuclease domain
c5dkuB_	transferase	B	prex DNA polymerase	c-terminal his tagged appol exonuclease mutant
c1tk0A_	transferase/electron transport/DNA	A	DNA polymerase	T7 DNA polymerase ternary complex with 8 oxo guanosine and ddctp at2 the insertion site

polymerase theta. These four DNA polymerases were used as references to analyze the actinophage DNA polymerases.

Alignment and WinCoot Analysis

In order to confirm the type of polymerases found in the actinophages, UGENE and WinCoot were used to compare the modeled actinophage structures to the four reference DNA polymerases determined in the Phyre[2] results.[21,22] WinCoot was used to superimpose each of the modeled polymerases with the T7 DNA polymerase, *M. smegmatis* DNA polymerase I, human DNA polymerase nu, and DNA polymerase theta (PBD: 1t7p, 6vde, 4xvi, and 4x0p respectively).This resulted in RMSD (Root Mean Square Deviation) values ranging from 1.4770 to 2.9347Å, which shows that these predicted structures superimpose well. The majority (67%) of the phages had the lowest RMSD values when superimposed with *M. smegmatis* DNA polymerase I. Of the remaining phages, 30% had the lowest RMSD values with DNA polymerase theta, and 3.8% with DNA polymerase nu. There were no actinophages with T7's DNA polymerase as the lowest RMSD value. By analyzing the actinophage sequences for the presence of conserved active-site acidic residues that coordinate metals in both the exonuclease and polymerase active sites, and the absence of additional domains, the actinophage DNA polymerases were determined to be very similar to the cores of all four DNA polymerases.

T7 is one of the most studied models for bacteriophages.[23] Since the T7 DNA polymerase is known to contain robust DNA polymerase and exonuclease activities, T7's polymerase was used to locate the conserved active-site acidic residues. Figure 3 shows T7's DNA polymerase (pink) and Adahisdi's (green) DNA polymerase superimposed with an RMSD value of 2.1224Å. Adahisdi is an actinophage from cluster A1. Cluster A is one the largest and most diverse clusters. Therefore, this phage was chosen to represent the average modeled polymerase in figures 3, 4 and 5. It serves as a good model for these polymerases, as it does not contain

additional domains. In figure 3, the T7 structure includes the protein thioredoxin (blue) and DNA (yellow). There are five conserved acidic residues that coordinate metals in T7. They are Asp5, Glu7, and Asp174 in the exonuclease active site, and Asp475 and Asp654 in the polymerase active site. There is also another well-conserved acidic residue, Glu655, in the polymerase active site that is essential, but it does not coordinate metal ions.[24] As the corresponding acidic residues were identified in the actinophage polymerases, patterns began to emerge. The first two acidic residues in the exonuclease active site (equivalent to Asp5 and Glu7 in T7) were always located two amino acids apart. In Adahisdi, the exonuclease residues were found at Asp41, Glu43, and Asp195, while the polymerase residues were located at Asp391, Asp545 and Glu546. Figure 4 shows these residues aligned with the residues in T7. Within a pham, this numbering is typically conserved. Proteins whose residues were not consistent with its pham, often had additional domains or lacked domains. From this superposition, we can hypothesize that the Adahisdi DNA polymerase I, and other polymerases that conserve these residues, have functional exonuclease and polymerase domains.

Figure 3. DNA Polymerases from Adahisdi and T7 superimposed.

Within these actinophages, aspartic acid residues 174, 475 and 654 were completely conserved in all sequences studied. Only one out of the 1351 sequences did not have all of the conserved, metal-coordinating residues. Hiyaa, the only phage in cluster BQ, does not have acidic residues that correspond to T7's Asp5 and Glu7 in the exonuclease domain. According to a concise Conserved Domain Database (CDD) search, the protein only has a

polymerase domain from residues 428-761.[25] However, once you expand the search to a standard or full search, the CDD reports the presence of a "PRK14975" domain from residues 280-761 and a "PRK05755" domain from residues 299 to 767. These functions of these domains are a provisional DNA polymerase I, and a provisional bifunctional 3'-5' exonuclease/DNA polymerase. The glutamic acid residue, Glu655, in the polymerase active site was not as conserved. Only 82% of the actinophage DNA polymerases contained a glutamic acid residue that corresponded to Glu655. The remaining polymerases contained serine, glutamine, and alanine. As Glu655 in T7 plays an important role in the polymerase active site, the consequences of these mutations are unclear and will require further experimental work.

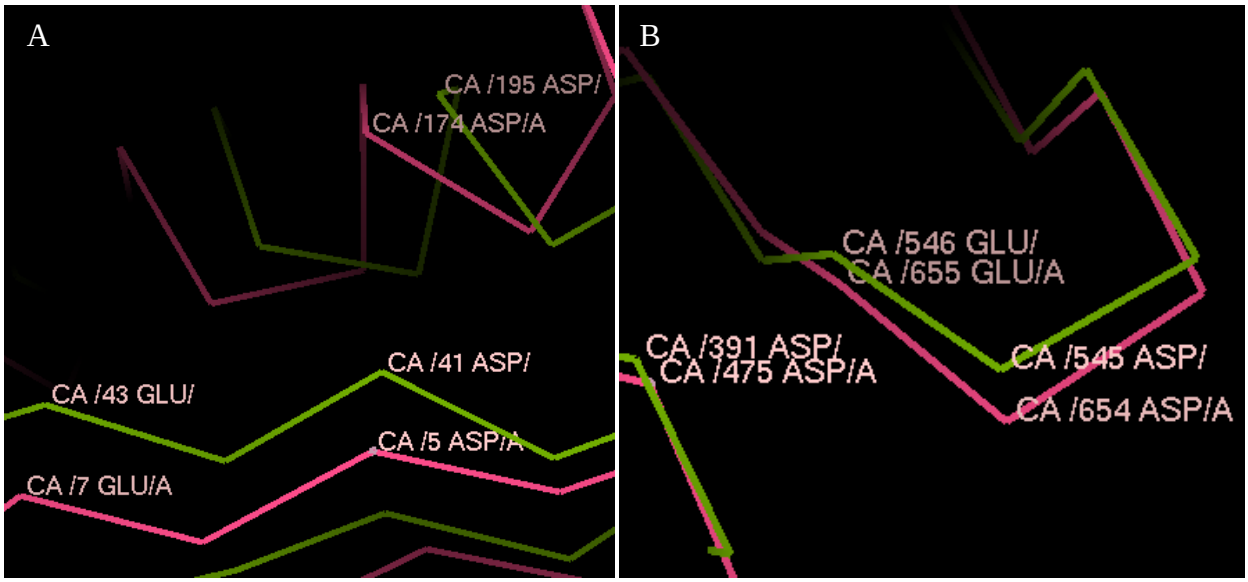


Figure 4. Exonuclease and Polymerase Active Sites of the T7 and Adahisdi DNA polymerases. T7's (green) and Adahisdi's (pink) DNA polymerases are superimposed using WinCoot with the residues of interested labeled in white. 2A shows the exonuclease residues while 2B shows the polymerase residues.

Each reference polymerase is a member of the DNA polymerase family A. This explains why these DNA polymerases appeared in the Phyre[2] search as well as the similarities in their structures, despite their different classifications. However, each reference polymerase has several unique characteristics, in the form of additional domains, that aid in determining the type of DNA polymerases contained in the actinophages. Each actinophage polymerase was examined for the presence of these characteristics.

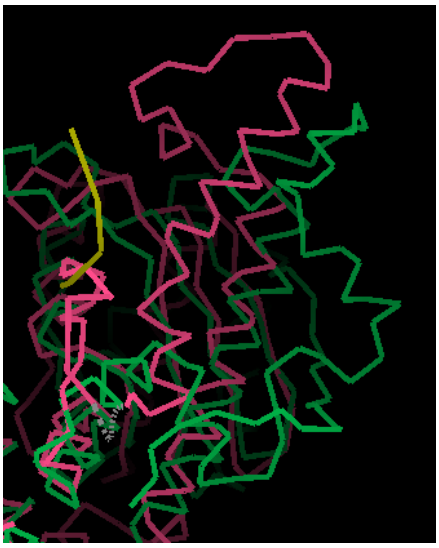


Figure 5. The exonuclease domain in T7 and Adahisdi's DNA polymerases.

Superpositions of the actinophage polymerases with T7's DNA polymerase had the highest average RMSD value of 2.3374Å. T7's DNA polymerase has a thioredoxin binding domain, or TBD, that is found between Gly262 and Val333.[24] Adahisdi's DNA polymerase, as well as other actinophage polymerases examined, do not contain this domain, which leads us to the conclusion that actinophages do not use thioredoxin as a processivity factor. All of the phages are not only missing the TBD, but a portion of their exonuclease domain also does not align with T7. This can be seen with Adahisdi in figure 5, ranging from Leu102 to Glu165 in T7 and Ala133 to Asp185 in Adahisdi.

M. smegmatis DNA polymerase I has the lowest average RMSD value (1.8807Å) when superimposed with the actinophage DNA polymerases. Despite belonging to family A, the *M. smegmatis* polymerase does not have an active 3'-5'exonuclease domain.[26] The domain is present, but it is lacking the essential acidic, metal-coordinating residues corresponding to Asp5,

Glu7 and Asp174 in T7. To compensate for this vestigial domain, it contains a flap endonuclease/5'exonuclease domain at its N-terminus.[26] This additional domain is not found in the actinophages; however, portions of this domain are found in a few of the modeled actinophage polymerases. This will be discussed more in the next section.

The human DNA polymerase nu and DNA polymerase theta are family A polymerases that specialize in DNA repair. They both contain a 3'-5' exonuclease domain and contain a homologous polymerase domain.[5] However, much like the *M. smegmatis* DNA polymerase I, the 3'-5' exonuclease is inactive due to missing acidic residues. Both polymerases can be further differentiated by their additional domains and inserts. DNA polymerase nu and DNA polymerase theta share three insertions. The first insertion, found at residues 499-509 in DNA polymerase nu and residues 2144-2177 in DNA polymerase theta, increases the processivity of the enzyme and is in a similar location to the TBD in T7's polymerase and the processivity factor in DNA polymerase gamma. These other insertions are found at residues 592-606 and 828-834 in DNA polymerase nu and at DNA polymerase theta residues 2254-2313 and 2503-2534, respectively. They allow the polymerases to perform TLS or translesion synthesis.[3,5,27] In addition to these inserts, DNA polymerase theta contains an N-terminal helicase domain, an unstructured region that separates the helicase and exonuclease/polymerase domains and two inserts within the exonuclease domain. Located within the exonuclease domain, these inserts are found at residues 1858-1899 and 1918-1936. The PBD model (4x0p) used in this study does not contain the helicase and unstructured region.[3,5] Some of the phages did contain minor inserts or folding anomalies in the same area as the inserts from polymerase nu and theta. This was most common in pham 95411. However, out of the actinophage polymerases with extensions or folding anomalies, the regions did not fully match the known inserts and the anomalies did not occur in

all of the insertion sites within the same protein. Also, almost all of the phage polymerases contained the exonuclease residues that are absent in polymerase nu and theta. Based on these results, the functions of the insert regions in the actinophages are unknown.

Additional Domains

Throughout the Coot analysis, several actinophage polymerases stood out due to additional domains. If these domains were less than 20 amino acids, they were not investigated. One of the most interesting cases is cluster CT and singletons Triscuit and Zuko. When the sequences were processed through the Conserved Domain Database, all of the polymerases within the cluster contained family 4 uracil-DNA glycosylase (UDG) domains. While the singletons Triscuit and Zuko do not belong to the cluster CT, they also have a UDG domain. Triscuit's and Zuko's DNA polymerases were classified in the same pham, 106021, as cluster CT. However, a UDG domain is not present in all members of this pham, and these actinophages were not isolated from the same hosts. Cluster CT was isolated from *Gordonia terrae* 3612 and *Gordonia rubripertincta* NRRL B-16540; where as Triscuit and Zuko were isolated from *Microbacterium foliorum* NRRL B-24224 SEA and *Streptomyces griseofuscus* ATCC 2391, respectively. The e-values for these hits range from 2.11E-15 to 0.009973. Triscuit and Zuko had the lowest e-values with 2.11E-15 and 9.22E-14, respectively. The cluster CT domains did not have an e-value less than 3.44E-6. Two cluster CT (Cleo and SketchMex) DNA polymerases and the two singletons were among the structures modeled by Phyre[2]. When superimposed, the UDG domains were not consistent in structure, as seen in figure 6. Cluster CT have large N-terminal domains that are modeled to reside above the active site. Figure 6A/B and 6C/D show

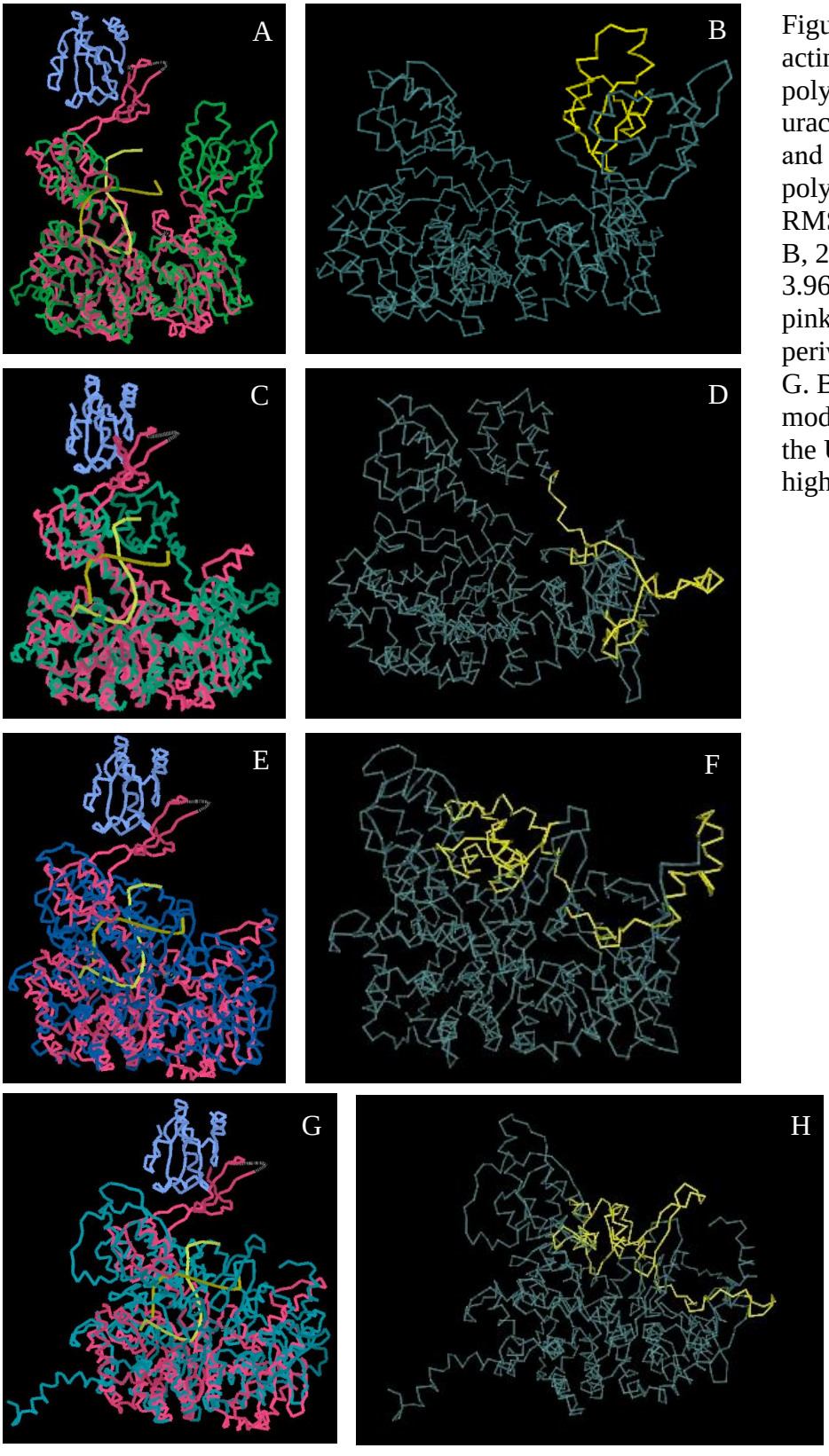


Figure 6. Superposition of actinophage DNA polymerases that contain uracil DNA glycosylases and T7's DNA polymerase (A,C,E,G). RMSD Values: A, 2.787; B, 2.2852; C, 4.1131; D, 3.9693. T7 is shown in pink and TBD in periwinkle for A,C,E and G. B,D,F and H show the models of the phages with the UDG domain highlighted in yellow.

the DNA polymerases from Cleo and SketchMex, respectively, superimposed with T7. Triscuit and Zuko appear to be more similar, as the polymerases are more compact and around the thumb domain (figures 6E/F and 6G/H, respectively); while Cleo and SketchMex appear to take on the traditional Family A polymerase core.

Cluster CD contains four actinophages, Gustav, Mahdia, Morrissey and Trine. The average sequence length is 742, which is 116 amino acids greater than the average actinophage polymerase. This large increase in sequence length can be explained by a N-terminal extension. According to the Conserved Domain Database, there is no known function found for these extensions. Based on the Phyre[2] models, there is not a consistent folding among the extensions. Figure 7 shows the cluster CD DNA polymerases superimposed with T7's polymerase. Gustav and Mahdia are the largest actinophages in this cluster and look the most similar. Their N-terminal extensions are primarily divided into two loops.

A loop is formed on both sides of the active site, above the exonuclease domain. Gustav's and Mahdia's extensions superimpose with part of the FEN/EXO domain, similar to SketchMex's N-terminal/UDG domain, roughly between residues 228 and 341 of the *M. smegmatis* polymerase. The extensions and *M. smegmatis*'s DNA polymerase 1 superimpose with 2.8566 and 1.5420Å RMSD values, respectively. Morrissey's N-terminal extension is rather compact and is located below the active site pocket. Trine does not have the N-terminal extension and is closer to the average actinophage length with 640 amino acids. In

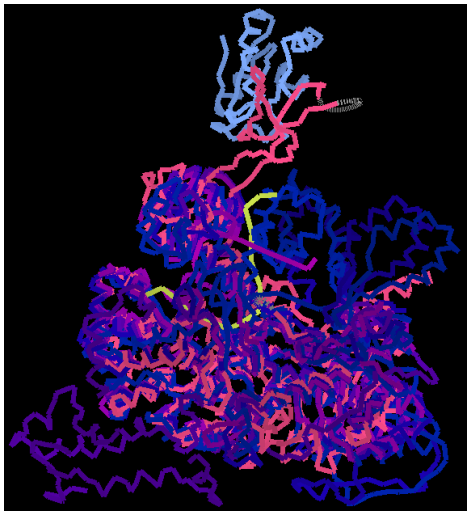


Figure 7. T7's DNA polymerase superimposed with DNA polymerases from the members of cluster CD. T7 is pink with a periwinkle TBD. Gustave is blue, Mahdia is grape, Morrissey is dark purple, and Trine is light purple.

addition, members of cluster AQ appear to contain large regions of the novel domains found in Gustav, Mahdia and SketchMex. The AQ proteins contain the two loops above the exonuclease domain, but the N-terminus is slightly shorter and the AQ proteins are missing a domain (Thr131-Ala168) that Mahdia and Gustav contain. These domains do not appear to be restricted to a specific host, as their respective actinophages were isolated in *Mycobacterium smegmatis*, *Gordonia terrae* and *Arthrobacter sp.*

There are other clusters whose proteins contain additional loops, as well as C-terminal or N-terminal extensions, but they do not have known functions according to the Conserved Domains Database. The representatives did not have a consistent location or shape for the domains. These clusters are summarized in table 3. Most of the modeled phages were determined to have short N-terminal extensions (20 to 50 aa), however the N-terminal extensions noted in the table are over 100 amino acids. Some of the C-terminal or N-terminal extensions may be due to mistakes in the gene call by the annotator. To ensure that a known domain was not identified, all of the sequences were searched through the CDD. All of the actinophage polymerases were called as containing family A DNA polymerase domains. There were only two other domains identified. The UDG domains were found in the CT, Triscuit and Zuko domains, as discussed previously and a domain listed as "DUF2779 superfamily". This domain is conserved in bacteria and has no known function.[28] Out of the 1351 actinophage polymerases, only two contained this domain. Cluster EK1, of pham 95411, contains eleven members, yet only TinyTimothy and Wesak contain the DUF2779 domain. It is located from residues 71 to 107 in both polymerases. TinyTimothy was modeled using Phyre[2]. This domain aligns with T7's polymerase in its exonuclease domain.

Discussion

Using various bioinformatic methods, almost all of these proteins were determined to have similar core features, not only amongst themselves but with the reference DNA polymerases as well. As shown by the extra domains in some of the actinophage DNA polymerases, there is still more to learn about these polymerases. However, despite being called as having the function of a DNA polymerase 1, they all lack a 5'-3' exonuclease domain. Due to this distinction, they could be separated into a separate category of DNA polymerases.

In the actinobacteriophage database, most of these proteins are called as a DNA polymerase I or a DNA polymerase, with the former being the most common called function. The reparative DNA polymerase I of *E. coli* is the prototype of this type of polymerase. *E. coli's* polymerase contains three main domains, a 3'-5' exonuclease, a 5'-3' exonuclease and a 5'-3' DNA polymerase.[29] These domains are found in the DNA polymerase I of *M. smegmatis* as well.[25] Despite belonging to the same polymerase family, T7's DNA polymerase only contain a 3'-5'exonucleus and a 5'-3' polymerase.[22] In order to accurately classify the actinophage DNA polymerases, a variety of bioinformatics methods were used. According to the Conserved Domain Database, all of the actinophage DNA polymerases belong to family A. The Phyre[2] analysis confirmed this with homology matches to T7's DNA polymerase, *M. smegmatis* DNA polymerase 1, and human DNA polymerase nu and DNA polymerase theta. While the hits to polymerase nu and theta were slightly unexpected, they are family A polymerases. As such, they have the same core, which leads to higher homology scores.

Each of the reference polymerases contain additional domains or inserts that are not found in the actinophage polymerases and despite the reference polymerases belonging to family A, the purpose of these polymerases vary. DNA polymerase nu and polymerase theta are repair polymerases, while T7's polymerase is responsible for replicating it's DNA.[3,5,29] Due to the

FEN/EXO and polymerase domains, the polymerase in *M. smegmatis* is a bifunctional protein.[26] T7's polymerase is the only reference protein that contains an active 3'-5' exonuclease domain. *M. smegmatis*'s DNA polymerase I, DNA polymerase nu and DNA polymerase theta are missing the conserved, acidic, metal-coordinating residues in their 3'-5' exonuclease domain, rendering them inactive.[3,26] Almost all of the actinophage polymerases contain these residues, thus it can be assumed that they have active exonucleases.

Based on the domains of the actinophage DNA polymerases, they are most similar to T7's DNA polymerase. However, the RMSD values found when the actinophages are superimposed with the reference polymerases, do not agree. With the highest RMSD values, the average RMSD value for super positions of the phage polymerase over T7's is 2.3374Å. The closer the RMSD value is to zero, the more similar the structures are. This increase in the RMSD value likely stems from two domains. First, T7's polymerase contains the TBD; whereas the actinophage polymerases do not. Within the exonuclease domain, the alignment between the polymerases is very poor from T7 residues 102 to 165. This is the region responsible for phosphodiester bond cleavage. Trp160 is essential to this cleavage; aligning the 3' terminal residue.[30] Proteins can have variations in folding depending on the presence of cofactors or substrates. The actinophage polymerases were modeled based on homologous proteins. It is probable that if the polymerases were isolated and crystalized with substrates present, that they would align better with T7.

Superimposing the structures reveal another possible hypothesis. When the actinophage polymerases were superimposed with *M. smegmatis*, most of them aligned with the lowest RMSD values of the reference polymerases. This suggests that some of the actinophages may have obtained their DNA polymerase gene, or at least part of it, from their host. Even with the

absence of *M.* smegmatis's FEN/EXO domain in the actinophage polymerases, the average RMSD score was 1.8807Å. This is a much better superimposition than the T7 RMSD score of 2.3374Å.

There is much more to learn about these proteins. This is especially the case of the polymerases where anomalies were discovered. Figure 8 shows the polymerase of actinophage Hiyaa superimposed with T7's polymerase. These polymerases appear to align well in the thumb and finger; however, they do not align in the exonuclease domain. The N-terminus of Hiyaa compacts around the thumb and the active site. The two polymerases begin to align around residue 169 in T7 (298 in Hiyaa). According to the CDD, Hiyaa's polymerase is still a family A polymerase, despite the missing exonuclease residues and structural differences. The e-values in for the CDD results range from 2.094E-18 to 5.939E-9. While these e-values are acceptable, they are much larger than many of the other CDD results. The CDD does not call a domain or function for the first 280 amino acids and the Phyre[2] server has very low confidence for the first 200 amino acids. This region may be disordered or there may not be a crystalized model that matches it, leading Phyre[2] to model it around the thumb. The latter may be the case for CD and AV polymerases as well. Each of the N-terminal domains are very different from one another, as seen in figure 6.

Figure 8. Actinophage Hiyaa's DNA polymerase (green) superimposed with T7's DNA polymerase (pink). RMSD value: 2.4963.

Uracil can occur in DNA due to the spontaneous deamination of cytosine or the misincorporation of dUMP during synthesis.[31] Uracil-DNA glycosylases are DNA repair enzymes that initiate the base excision repair pathway to remove uracil from DNA. They can be found in a variety of organisms, from archaea to eukaryotes. Based on substrate specificity, these proteins can be classified into six families.[31,32] The fourth family is of particular interest to this study, as the CDD indicated that the N-terminal domains of cluster CT and singletons Triscuit and Zuko are likely members of this family. Members of family four include eubacteria, archaea and some bacteriophages, and they are also often found in thermophiles.[32] The bacteriophage SPO1 of *Bacillus subtilis* contains a family four UDG domain. This domain spans the first 190 amino acids on the N-terminus of its DNA polymerase.[33] Typically, families one and four act as the main uracil remover.[32] This domain is useful to the actinophages because the deamination of cytosine can lead to pro-mutagenic events in DNA.[33] Initiating the base excision repair pathway allows the actinophage to reduce the occurrence and passage of these mutations. Having a UDG domain fused to the DNA polymerase allows for the UDG domain to be present and directly involved as DNA replication is occurring.

While some of these polymerases with additional domains are modeled to look similar, they are distant in the phylogenetic tree (figure 2). The locations of these polymerase clusters can be identified using table 1. The red clade contains pham 6914, which only contains one cluster, AQ. This clade is interesting as it is the closest to T7 (the black line). Despite having a N-terminal extension similar to cluster CD, they are separated by a considerable distance as the cluster CD is found in the yellow clade. In fact, the CD cluster is found closer to the UDG

domain containing polymerases in the green clade (pham 106021). The unique phage Hiyaa (pham 58669) resides in the blue clade. Aside from cluster AQ in the red clade, the tree does not exclusively group polymerases with extra domains by themselves. They are included with more T7-like polymerases. The purple, orange and yellow (excluding cluster CD) mostly contain the

Table 3. Summary of conserved acidic residues and additional domains found in the bacteriophage polymerases.

Clusters/Singletons	Description
A1, A2, A3, A4, A5, A6, A8, A9, A10, A11, A12, A13, A14, A15, A16, A18, A19, A20, AZ, B1, B2, B3, B4, B5, B6, B7, B8, B9, B10, B11, B12, B13, BB1, BJ, BL, CA, DK, DS, DR, EB, EH, FC, GC, GD, Anderson, Attoomi, EmiRose, Pine5, TPA2	• Contains all of the conserved acidic residues • Does not contain additional domains
AV, EA1, EA4, EA5, EA8	• Contains serine instead of Glu655 • Does not contain additional domains
AK, BH, DA, EA, EA2, EA3, EA6, EA7, EA9, EA10, EA11, EJ	• Contains serine instead of Glu655 • Contains uninvestigated inserts or small domains
A7, A17, AB, BB2, BG, BM, EJ, EK, EK1, EK2, EM, EP, Ibantik, LuckyBarnes, Zeta1847	• Contains all of the conserved acidic residues • Contains uninvestigated inserts or small domains
CD	• Contains all of the conserved acidic residues • Contains large N-terminal extensions (figure 6)
AQ	• Contains glutamine instead of Glu655 • Contains large N-terminal extensions similar to CD
AB, JacoRen57	• Contains alanine instead of Glu655 • Contains large N-terminal extensions
CT, Triscuit, Zuko	• Contains alanine instead of Glu655 • Contains UDG domain
BQ	• Does not align with the folding of the exonuclease of the T7 polymerase. • Does not have Asp5 or Glu7

polymerases that do not have additional domains. Table 3 sorts the phage clusters based on their acidic residue conservation and additional domains.

Many of the actinophage polymerases are annotated as DNA polymerase 1. This could be misleading as *E. coli* and *M. smegmatis* contain DNA polymerase I. However, despite being annotated as a DNA polymerase 1, they all lack the typical 5'-3' exonuclease domain found in these polymerases. Due to this distinction, they could be separated into a separate category of DNA polymerases.

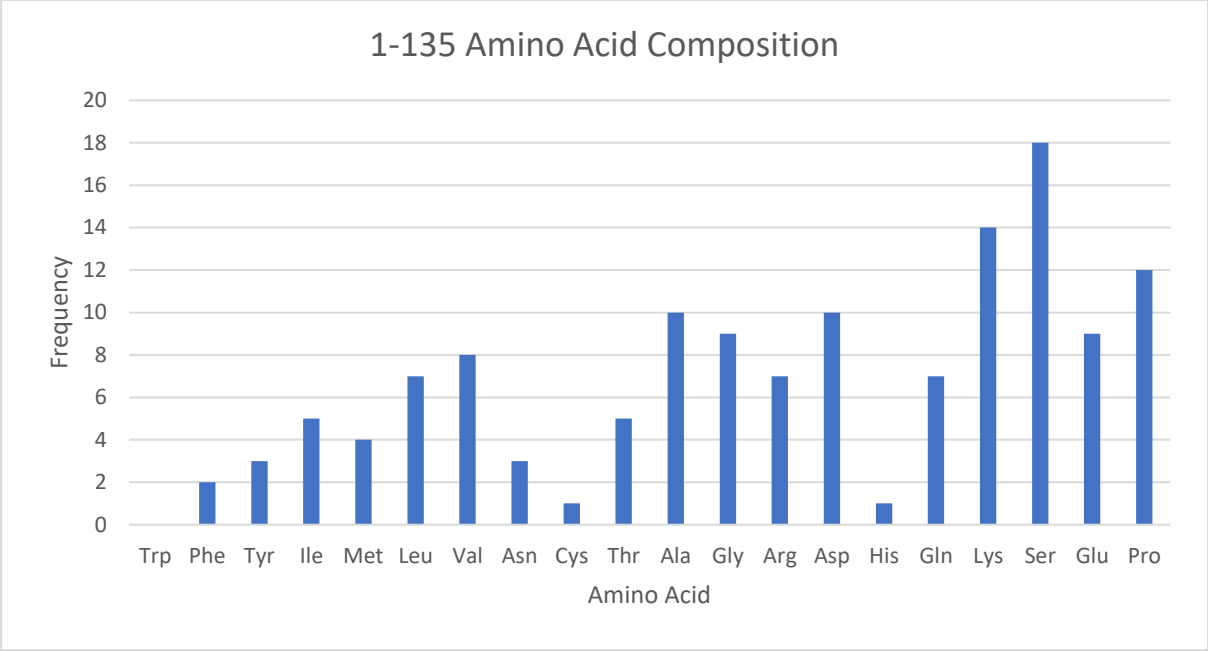
Future work will focus on the purification and characterization of these DNA polymerases to learn more about their function. Do they play a role in the replication of the DNA, the repair of DNA, or both? While our laboratory has tried to express and purify some of these polymerases, we have learned that they have poor solubility when expressed in *E.coli*. Current efforts are focused on the use of other expression systems in order to generate soluble enzyme for biochemical study.

CHAPTER 3: DNA POLYMERASE GAMMA IN *CRYPTOCOCCUS NEOFORMANS*

In order to predict the function of each novel domain, bioinformatic analyses were performed to study the primary, secondary and tertiary structure of each domain. Additionally, pull-down assays were optimized using the 1-135 domain for further investigation. For this study, *Cryptococcus neoformans var. grubii H99* (accession number: XP_012047111.1) was used for the residue numbering during the bioinformatics analyses, as well as the target protein in the pull-down assay. Using the amino acid sequence from the accession number above as the query sequence, the Conserved Domain Database server was searched to confirm known domains for the protein.[25] According to the CDD, there are only two known domains within the *C. neoformans* DNA polymerase gamma. Ranging from residues 222-543, the first domain received a specific hit from the human DNA mitochondrial polymerase exonuclease domain (accession: pfam18136) with an e-value of 3.89E-130. The second domain, located from residue 1052 to 1287, was identified as a DNA polymerase A domain (accession: smart00482) with an e-value of 5.78E-52.[28]

1-135 Domain Bioinformatics

The 1-135 domain resides at the N-terminus of the polymerase, directly after the mitochondrial localization sequence (MLS). The MLS spans from residue 1 to 35, leaving the 1-135 domain to cover residues 36 to 171. Examining the composition of the sequence, there is a trend in the frequencies of the amino acids. The graph in figure 9 shows the frequency of the amino acids within this domain. Most of the amino acids are small and polar, with serine as the most frequent amino acid at 18 and tryptophan being absent from the sequence. There are very few aromatic amino acids, comprising only 3.70% of the domain.



1-135:
MKPSDAPVKISDGKEEGVGKPLIPAFGARRAEMEDYILAMEMAKLEDGYGQPRVRKIR
KSKLPSLHDPQSFLCDSTQASSSKVTSSASPTSQPSRKGKEKEVVSNDYATNVPNQDVQT
LEDAKPIDSGQSKSGPR

Figure 9. Amino acid composition of the 1-135 domain. The bar graph shows the frequency of each essential amino acid within this domain. Below the graph, the amino acid sequence of the 1-135 domain is listed. The amino acids in red are disorder promoting, whereas the amino acids in black are order promoting.

Amino acids can be classified based on their chemical properties. These classifications can be very informative when attempting to deduce the structure, nature, and function of a protein. For example, amino acids can be separated into order-promoting amino acids and disorder-promoting amino acids. Generally, order-promoting amino acids are larger or hydrophobic, whereas disorder-promoting amino acids are small, charged, hydrophilic amino acids.[34] In the 2010 review "Understanding Protein Non-Folding", the amino acids are ranked based on their tendencies to promote order or disorder: Trp, Phe, Tyr, Ile, Met, Leu, Val, Asn,

Cys, Thr, Ala, Gly, Arg, Asp, His, Gln, Lys, Ser, Gly, Pro.[35] Despite this ranking, the order or disorder-promoting tendencies are disputed for some amino acids, specifically Met, His, Thr, and Asp. For this study, amino acids that promote order are Trp, Phe, Tyr, Ile, Met, Leu, Val, Asn, Cys, and His; while amino acids that promote disorder are Thr, Ala, Gly, Arg, Asp, Gln, Lys, Ser, Glu, and Pro.[34,35,36,37,38,39] The graph in figure 9 is arranged to follow the order to disorder ranking. Arranging the frequencies in this order reveals a pattern. Within this region, the most frequent amino acids promote disorder. In fact, 74.8% of the amino acids promote disorder. Despite the abundance of disorder-promoting amino acids, a region is only disordered if these amino acids are clustered together. Below the graph in figure 9, the sequence for the 1-135 domain is displayed. The amino acids are color-coded with black letters representing order-promoting amino acids and red letters representing disorder-promoting amino acids. The sequence view emphasizes that these hydrophilic amino acids are typically found in larger sections, separated by one to three hydrophobic amino acids.

Using various servers, the secondary structure of the 1-135 domain was predicted. The 1-135 domain was processed through four servers: JPred4, RaptorX, PredictProtein and PROTEUS.[40,41,42,43] These servers can predict whether an amino acid will be part of a helix, strand or coil. Each of the servers predicted a helix beginning around residue 33. They also agreed on long coiled regions for much of the latter half of the proteins. RaptorX and PredictProtein provide the ability to predict disordered regions based on the amino acids and their interactions with each other. PredictProtein predicts that the entire domain (aside from residue 37) is disordered. RaptorX agrees with most of the domain being disordered, but it predicts it to be ordered from residue 21 to 44. This contains the region where the servers predict the domain to have a helix or beta strands.[41] PredictProtein also provides potential protein and

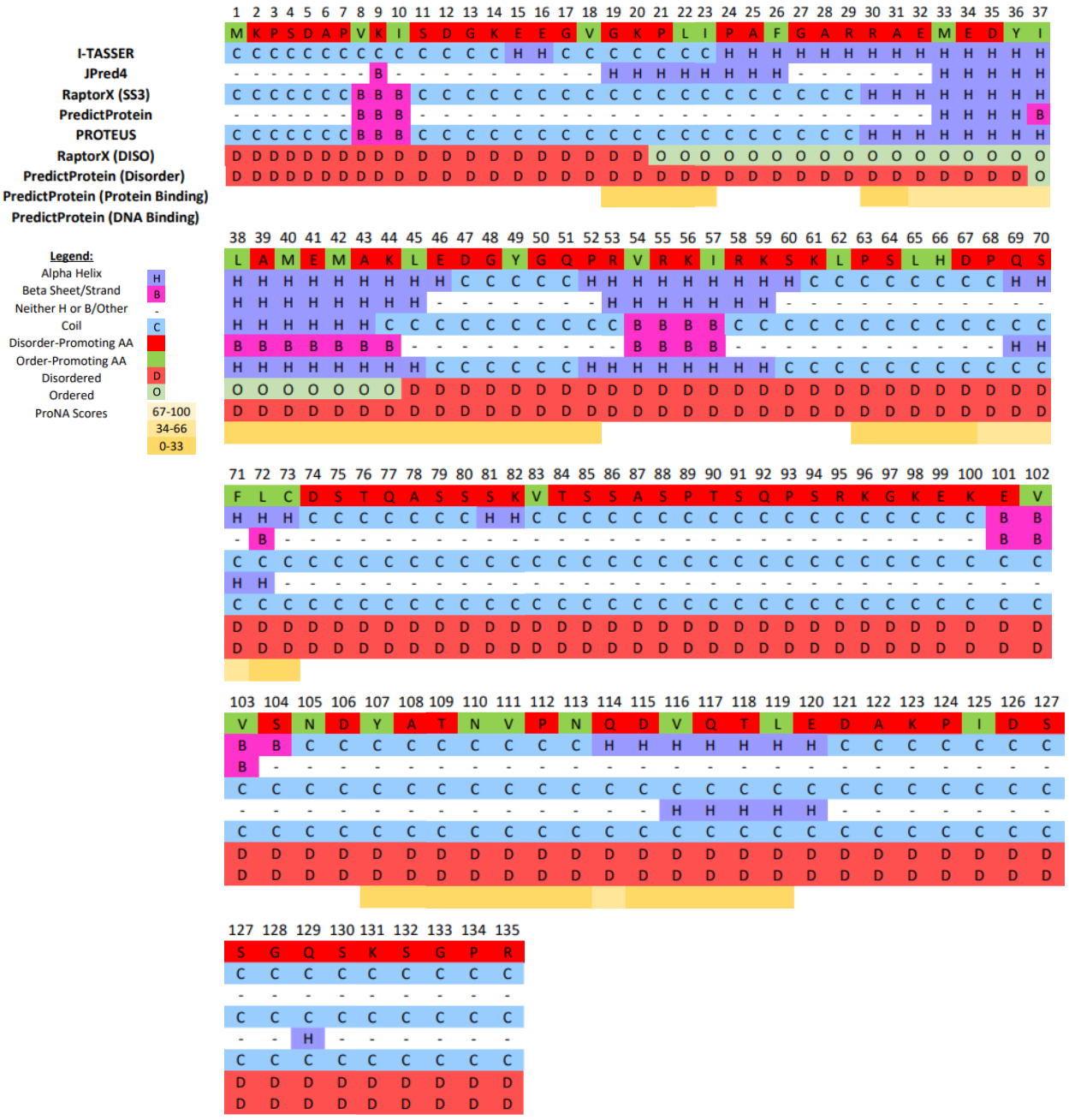


Figure 10. Secondary structure predictions for the 1-135 domain. I-TASSER, JPred4, RaptorX, PredictProtein and Proteus were used to predict the secondary structure for the 1-135 domain. This sequence view summarizes the predictions based on the legend above.

binding at residues 19 to 23, 30 to 52 and 63 to 73.[42] These predictions are summarized in figure 10.

Phyre[2] and I-TASSER.

I-TASSER (Iterative Threading ASSEmbly Refinement) is an online protein structure and function prediction service. This server provides a wide range of information including the top ten threading templates used to model the protein, and five possible models of the protein.[44] The top ten threading templates for the 1-135 domain have normalized Z-scores ranging from 0.42 to 1.17. A good alignment is considered as any template with a normalized Z-score over 1.

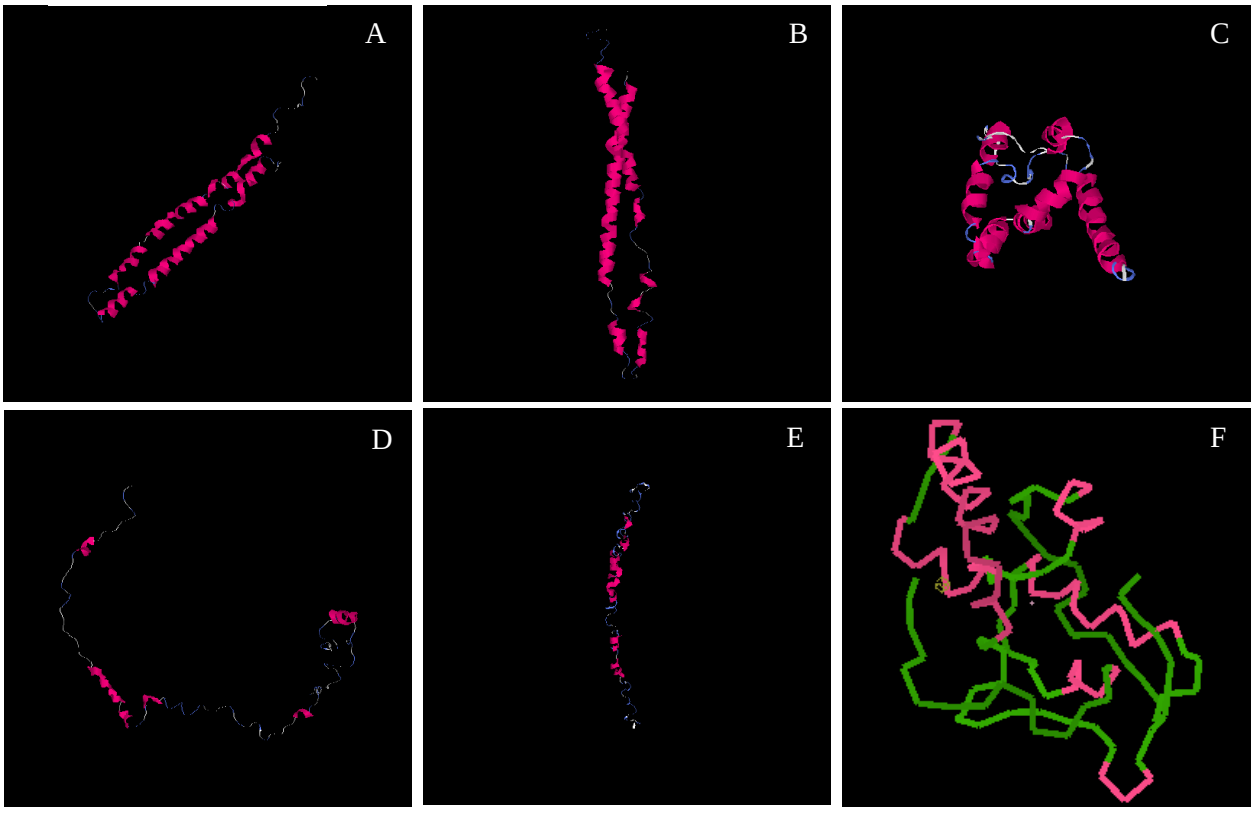


Figure 11. I-TASSER (A-E) and Phyre2 (F) predictions of the 1-135 domain. A shows the most confident model proposed by I-TASSER with a C-value of -2.60 and an estimated RMSD value of 10.3±4.6Å. The other models (B-E) have the following C-values: -3.39, - 4.07, -3.15 and -4.27 respectively. F shows the model proposed by Phyre[2]. This model is most like C, but they are not the identical.

Seven of the templates were over the Z-score threshold, but they were not high Z-scores, and the highest percentage sequence identity between the threading aligned region and query sequence was 0.27. The server provides five models, which are given a C-score to indicate the server's confidence in the model. C-scores range from -5 to 2, with a higher score signifying higher confidence. For these models, the C-scores are low, ranging from -4.27 to -2.60. Figure 11A-11E shows these models and their respective C-scores. I-TASSER provided an estimated RMSD score for the best model, based on the C-score and protein length. For this domain, the estimated RMSD value is 10.3±4.6Å. According to the Phyre2 server, 59% of the domain was predicted to be disordered. The model and template hits had very low confidence. Only 11 template hits were retrieved with very low confidence numbers between 5.6% and 14.6%.[20] Figure 11F shows the model produced by the Phyre^{2}server.

While the previous servers provided insight to the folding of the domain, blastp and PSI-BLAST tools were used to investigate potential functions and similar domains in other proteins. First, the sequence for the 1-135 domain was analyzed with the protein-protein Basic Local Alignment Search Tool (blastp).[45] With the initial blastp analysis there were 34 hits. Of the hits, 33 of the proteins belong to members of the *Cryptococcus* species. There was also a hit to an uncharacterized protein found in *Kwoniella shandongensis*, a fungus in the same class (Tremellomycetes) as the *Cryptococcus* species.[46] 56% are *C. neoformans var. grubii*, 35% are *C. gattii*, 3% are *C. neoformans var. neoformans*, 3% are *C. depauperatus*, and 3% are *Kwoniella shandongensis*. Most of the proteins were called as DNA polymerase gamma 1. Aside from the protein from *K. shandongensis,* only four proteins did not have a known function, thus being referred to as hypothetical proteins.

In this region, most of the sequences are very similar among the *Cryptococcus neoformans var. grubii*, *Cryptococcus neoformans var. neoformans* and *Cryptococcus gattii*. However, the sequences for *C. depauperatus* and *K. shandongensis* differ greatly, as shown in figure 12. This is most notable in the gaps produced by the *K. shandongensis* polymerase. Due to the sequence of its polymerase, there are two major gaps added at 1-135's residues 106 and 131.

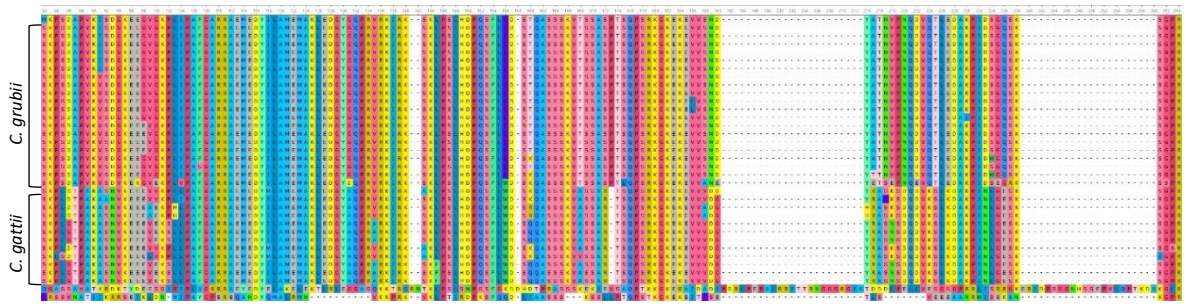


Figure 12. Clustal W alignment of the 1-135 blastp hits. To the left of the alignment, most of the sequences have been labeled as *C. grubii* or *C. gattii*. There are four unlabeled sequences. The 1-135 sequence is above the *C. grubii* sequences. Between the *C. grubii* and *C. gattii* sequences, is the *C. neoformans* sequence. Finally, the last two sequences are the *Kwoniella shandogensis* and *C. depauperatus*, respectively. Each amino acid was color-coded using UGENE's default color scheme, which gives each amino acid its own color. This helps to visually identify consensus trends.

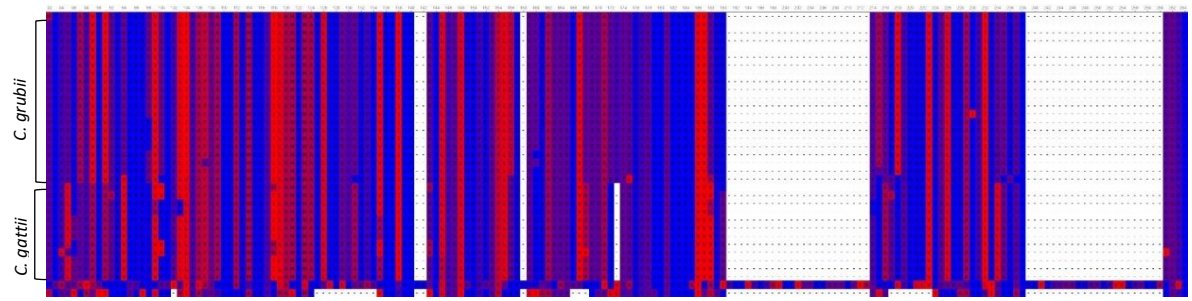


Figure 12. Hydrophobicity of the 1-135 domain. Using the alignment from figure 12, the alignment has been color-coded based on the hydrophobic or hydrophilic properties of the amino acids. The colors range from blue(hydrophilic) to red(hydrophobic).

Despite these differences, 96 of the 135 amino acids were over 90% conserved, with 22 amino acids fully conserved. Most of these amino acids are polar and are located between residues 22 and 99. The longest region with consecutive, conserved amino acids occurs from amino acids 96 to 99 and includes the conserved sequence KGKE. Figure 13 uses the same alignment, but the amino acids are color-coded according to their hydrophobicity, with blue being hydrophilic and red being hydrophobic. While the majority of the amino acids are hydrophilic, there are a few smaller patches of hydrophobic amino acids. This further emphasizes the trends determined by the amino acid concentration analysis and sequence view in figure 9.

Position-Specific Iteration Basic Local Alignment Search Tool, also known as PSI-BLAST, is used to detect distant relationships between proteins.[47] Initially the PSI-BLAST performs a blastp, which is referred to as the first iteration. After the initial search, or iteration, additional iterations are performed until no new hits are discovered with an E-value less than 0.005.[47] Since the 1-135 blastp only returned similar species, a PSI-BLAST search was performed to look for distant relatives. When the 1-135 domain was analyzed using PSI-BLAST, the database retrieved 41 hits after 4 iterations. These additional results came from proteins found in *Cryptococcus depauperatus*, *Kwoniella mangrovenis*, *Kwoniella dejecticola*, *Kwoniella pini* and *Kwoniella bestiolae*. All the hits are fungi, specifically fungi of the order Tremellales. 83% of the proteins belong to *Cryptococcus* species and the remaining 17% belong to *Kwoniella* species. To further examine the species, they were organized into a phylogenetic tree based on their polymerase sequence alignments as shown in figure 14. The tree is divided into five sections, which is represented by black, red, golden, green, and blue branches. These five sections are primarily separated by species, except for the blue branched section, as this section contains all of the *Kwoniella* species. The division in the tree supports the division seen in the

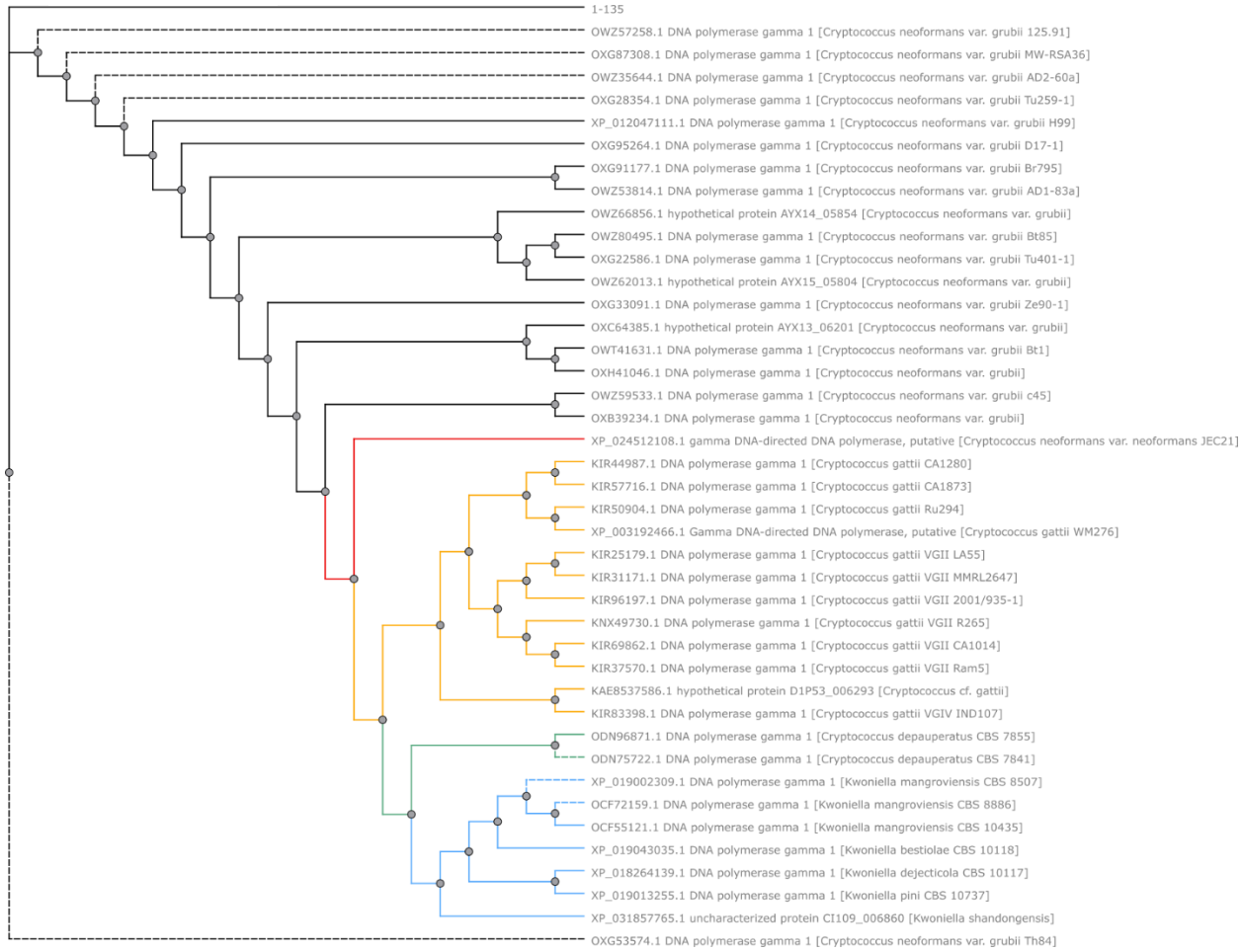


Figure 13. A Phylogenetic Tree of the 1-135 PSI-BLAST Hits. The tree is color-coded based on the genus and species. *Cryptococcus neoformans var. grubii* are connected by black branches. *Cryptococcus neoformans var. neoformans* is connected by red branches, *Cryptococcus gattii* are connected by golden branches, *Cryptococcus depauperatus* are connected by green branches and *Kwoniella* species are connected by blue.

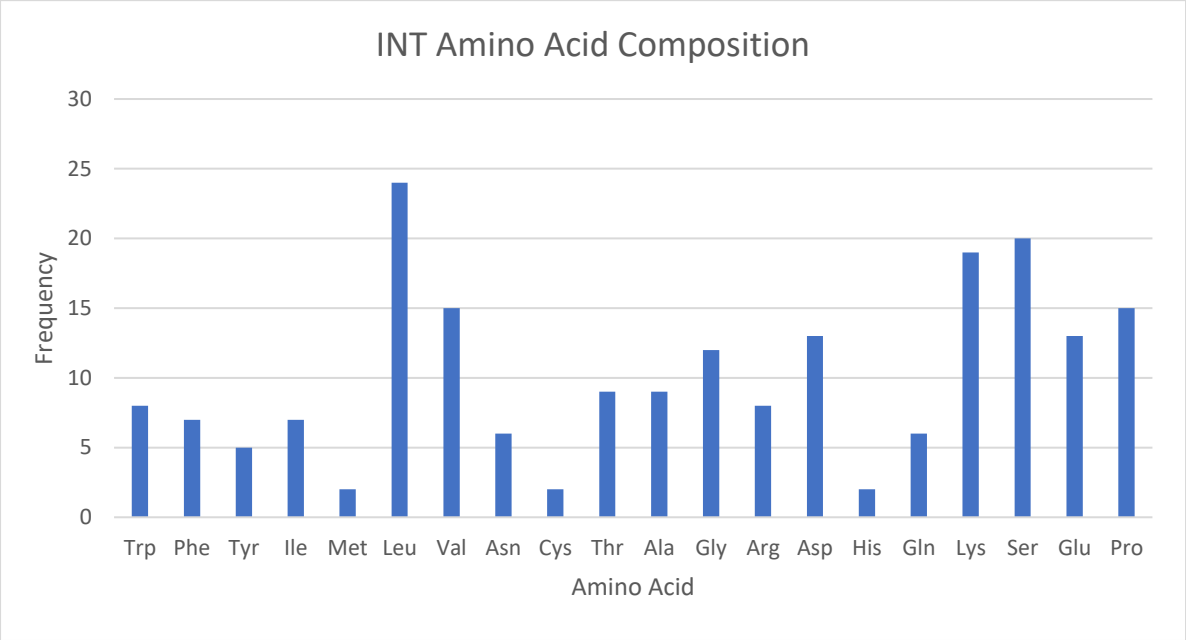
alignments. Looking at figure 12, there is a noticeable change in amino acid content between the species. Based on the phylogenetic tree, the *Kwoniella* species are most related to *Cryptococcus depauperatus*, which explains why the *C. depauperatus* and *Kwoniella* polymerases greatly differ from the *C. neoformans* and *C. gattii* polymerase.

While most of the proteins were called as DNA polymerase gamma 1, five of them did not have a function. With the addition of the PSI-BLAST sequences, the number of amino acids with consensus over 90% dropped dramatically from 96 in the blastp alignment to 27 in the PSI-BLAST alignment. One of the regions with the most consecutive conservation occurs from amino acids 93 to 100. This region contains five of the fully conserved amino acids, including the KGKE region. The other amino acids in this region have 92.9% or more consensus.

INT Bioinformatics

Located between the spacer and the polymerase domains, the internal domain (INT) begins at Lys732 and has a length of 202 amino acids. Figure 15 shows the amino acid composition of this domain. The frequencies for the amino acids range from methionine, cysteine and histidine with 2 residues, to leucine with 24 residues. The distribution of the frequencies does not immediately reveal a clear pattern with order-promoting versus disorder-promoting amino acids. Out of the 202 amino acids, 124 amino acids or 61.4% were disorder-promoting. As seen in the sequence view, the order-promoting residues are dispersed throughout the domain, preventing more than eight disorder-promoting residues to bond in a row.

The secondary structure of the INT domain was predicted by the I-TASSER, JPred4, RaptorX, PredictProtein and PROTEUS servers. While each server varies in their predictions, they agree that the domain has 3 helices and two beta strands. These structures are predicted to occur around residues 767, 852, 875, 891 and 898 respectively. RaptorX and PredictProtein predict the domain to be mostly ordered with the exception of a disordered region at the N-terminus and another disordered region from residues 817 to 835. PredictProtein also proposes possible DNA and protein binding sites within the domain. Four DNA binding sites are proposed. They occur from residues 743 to 765, 844 to 856, 869 to 882, and 891 to 904. These



INT:

KLEESYSFFRIGNAGSPKKTKLVGPSIKPFVNSGDLTSAYPELLVKVMKTDLNDVVED
LWECVVDMGNLKESEWGQQLDWTPTTQDITSSNDVPLFSSSSSLRPSSIKKSKANLGI
WPKWYWDLTGPVSRLPVGELDLTCKKAIAPLLLRLQWQGFPLVHSKEHKWLYRLPR

Figure 14. Amino Acid Composition of the INT Domain. The amino acid composition of the INT domain are shown graphically as well as in a sequence view. The graph displays the amino acids from left to right, increasing in probability of being disorder-promoting. The sequence view is color-coded according to the order-promoting (black) or disorder-promoting (red) properties of the amino acids.

predictions are provided with a confidence level for each residue, ranging from 0 to 100. Each prediction has lower confidence for the outer residues but increases confidence in the middle of each site. Only one residue (17) had a confidence level of 67 to 100. Overall, there were 7 individual proposed sites for protein binding. However, only one of the sites spanned more than two residues. This site ranges from residues 879 to 895. Figure 16 summarizes the secondary structures for the INT domain

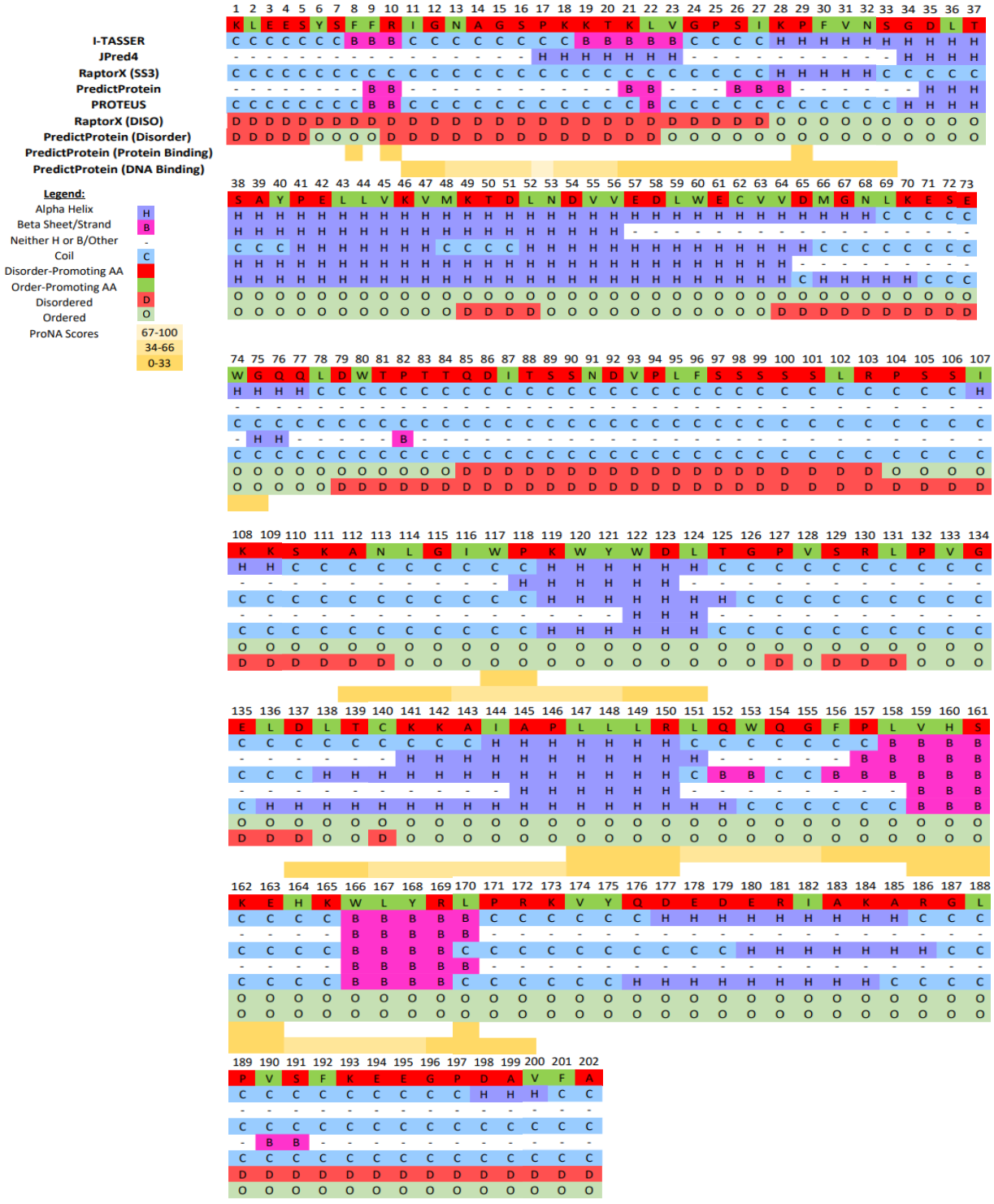


Figure 15. Secondary Structure Predictions for the INT domain. I-TASSER, JPred4, RaptorX, PredictProtein and Proteus were used to predict the secondary structure for the INT domain. This sequence view summarizes the predictions based on the legend above. The prediction servers were fairly consistent for the INT domain.

.

Using I-TASSER and Phyre[2] to model the structure provided the models in figure 17. Figure 17A-E shows the models proposed by I-TASSER. Out of the top ten threading templates I-TASSER used to create these models, there were only two unique PBD files 3ikm and 1ldj. Nine of the top ten templates were listed as 3ikm, 3ikmA or 3ikmD. These PBD files are crystal structures of the human DNA polymerase gamma-1 holoenzyme.[16] The ninth template, 1ldjA, is the crystal structure of the cullin homolog 1 within the Cul1-Rbx1-Skp1-F boxSkp2 SCF Ubiquitin Ligase Complex.[49] The normalized Z-scores of these templates were very high, ranging from 0.88 to 8.09. With a Z-score higher than 1 signifying a good alignment, the only template below the cutoff was the cullin homolog. The lowest 3ikm Z-score was 1.30. C-scores

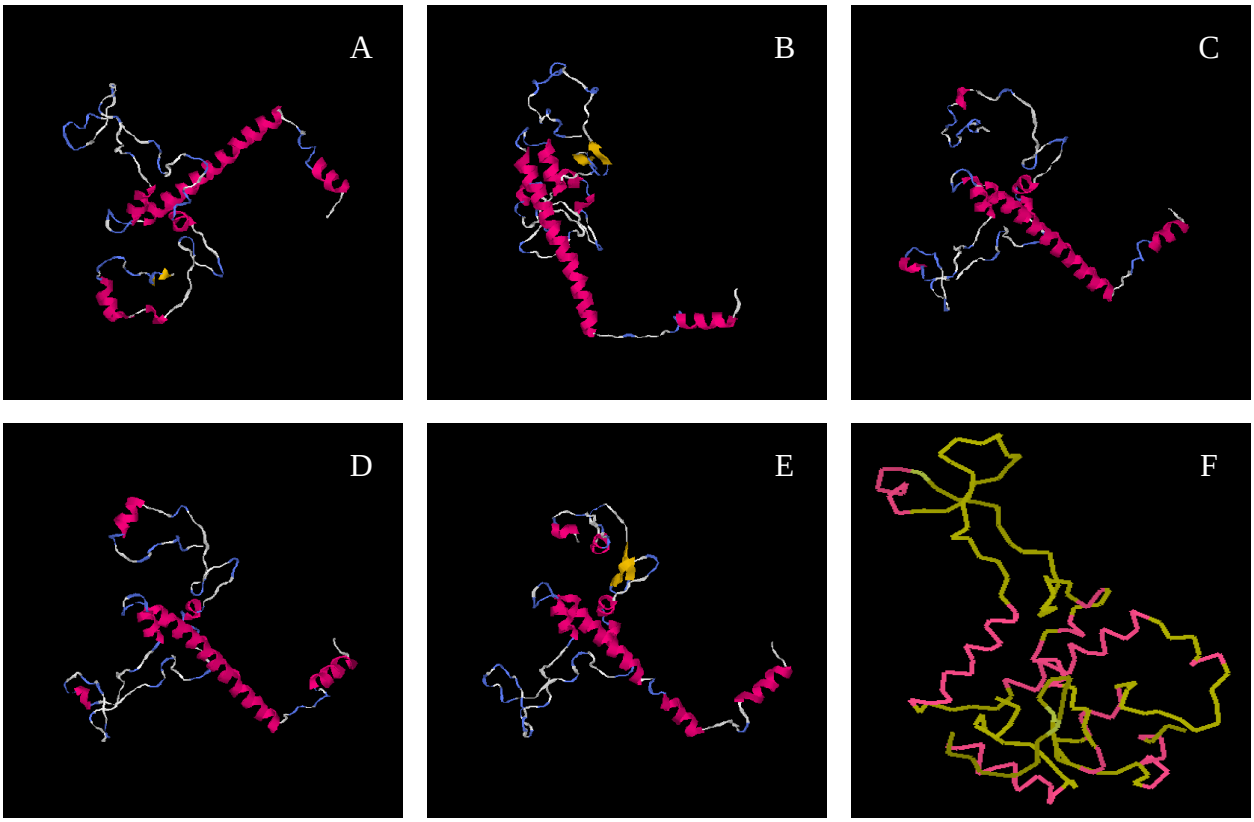


Figure 16. Proposed models of the INT domain from I-TASSER (A-E) and Phyre2 (F). A shows the most confident model proposed by I-TASSER with a C-value of -0.17. The other models (B-E) have the following C-values: -1.15, -1.66, -2.66 and -3.62, respectively.

for the models ranged from -3.62 to -0.17 and are listed with their models in figure 17. Figure 17A is the most confident model from I-TASSER. The estimated TM-score and RMSD are 0.69±0.12 and 5.7±3.6Å, respectively. Comparing the I-TASSER models, they all look very similar.

Table 4. Phyre[2] Results for the INT Domain.

Template	Confidence (%)	% i.d.	PDB Header/Superfamily	PDB Molecule/Family
c3ikmD	100.0	26	transferase	DNA polymerase subunit gamma-1
c4ztuA	100.0	29	DNA binding protein/DNA	DNA polymerase subunit gamma-1
c6mcjA	40.7	21	Protein binding	Orange carotenoid-binding protein
d1knya1	36.2	37	Nucleotidyltransferase substrate binding subunit/domain	Kanamycin nucleotidyltransferase (KNTase), C-terminal domain
c5fcxB	34.9	26	Cartenoid binding protein	Red carotenoid protein (rcp)

When sent through the Phyre[2] server, 78% of the residues were able to be modeled with over 90% confidence. The first two hits on the template list have 100% confidence, both of which are the human DNA polymerase gamma. After the first two hits, the confidence drops to 40.7% and continues to drop rapidly. The top five results do not all relate to DNA polymerase gamma. Template hits three and five are proteins that bind to orange and red carotenoids, respectively, while the fourth hit is the c-terminal domain of a kanamycin nucleotidyltransferase.[50,51,52] This information is summarized in table 4. The template hits that start with a "c" are chains and have PDB headers and molecule names. However, the fourth hit,

d1knya1 is a domain and is classified using superfamilies and families. The proposed model from the Phyre[2] server is shown in figure 17F. The high confidence scores of the top Phyre[2] results, in combination with the I-TASSER hits suggest that the INT domain is very similar to the human DNA polymerase gamma.

Compared to the 1-135 domain, the INT domain retrieved more blastp and PSI-BLAST hits. When a blastp was performed, 1124 hits were retrieved. However, only 867 of the hits were above the e-value threshold of 0.005. Without the threshold, e-values ranged from 2.00E-134 to 0.05. Unlike the 1-135 domain, the INT domain blastp produced hits from a wider variety of species, with only 38 of the 867 hits belonged to the cryptococcus species. Most of the remaining species were various types of fungi, however there were five interesting hits above the threshold. *Carpinus fangiana,* better known as a hornbeam tree, had the highest e-value of the non-fungi hits with 4.00E-5. *Acropora millepora, Acropora digitifera and Orbicella faveolate* are stony corals (e-values of 0.001, 0.001 and 0.004, respectively). The final interesting species was *Scapholeberis mucronata,* a species of water flea with an e-value of 0.004. The proteins from *A. millepora, A. digitifera, O. faveolate and S. mucronate* are called as DNA polymerase gamma while the protein from *C. fangiana* is called as a hypothetical protein.

Most of the proteins were called as DNA polymerase gamma. Aside from hits labeled as DNA polymerase gamma, there were seven other functions or labels: hypothetical protein, uncharacterized protein, predicted protein, unnamed protein product, alpha-beta hydrolase, putative septin AspA and pentatricopeptide repeat-containing protein. Using the CDD, the putative septin AspA was determined to only have DNA polymerase gamma domains. The alpha-beta hydrolase and pentatricopeptide repeat-containing protein were found in both the

blastp and the PSI-BLAST for the INT domain and will be discussed after the PSI-BLAST results.

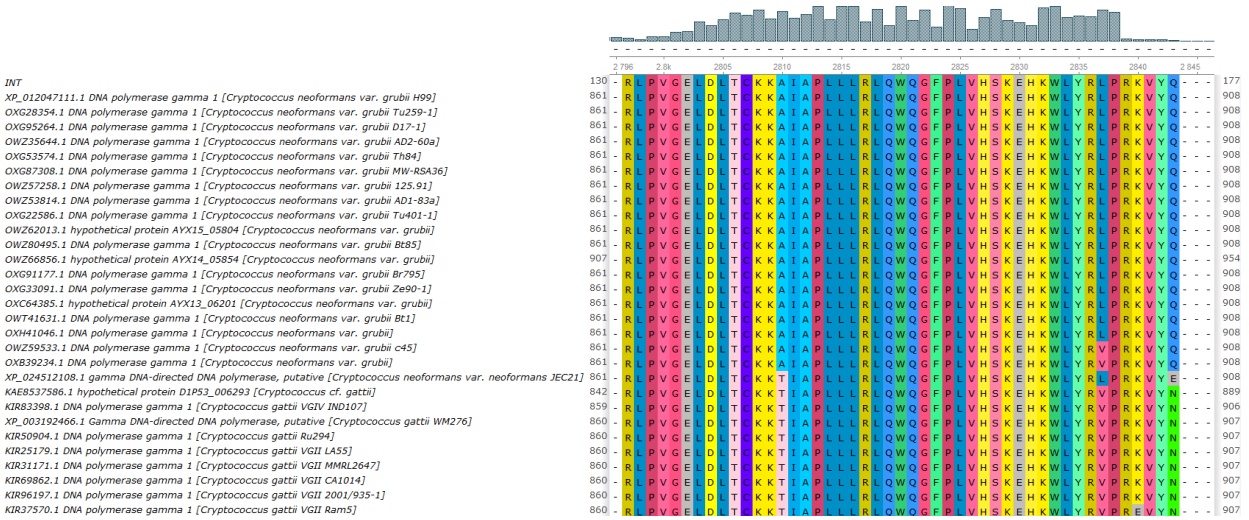


Figure 17. Partial blastp alignment of the INT domain. This portion of the alignment is the most conserved region within the blastp of the INT region. While it does not contain any fully conserved residues, which would be represented as a capital letter under the consensus bar, it contains most of the highly conserved residues.

While the alignment of these sequences did not show any fully conserved residues, there were 13 residues with over 90% conservation. These residues include Pro850, Trp852, Tyr903, Leu856, Pro878, Leu880, Leu881, Leu883, Trp885, Pro889, Leu890, Ser893 and Trp898. Due to the large sample set and the variance between the sequences, the INT region is split into several smaller sections in the alignment. Unfortunately, due to these gaps, showing the entire alignment is impossible. Including the gaps stretches the 202-residue domain to span 541 residues in the alignment. Therefore, figure 18 only shows the most conserved region with the top *Cryptococcus* protein hits. This region contains nine of the thirteen highly conserved residues.

PSI-BLAST provided roughly 2.5 times more results than the blastp with 2851 total hits. These results were much more varied in their host species ranged from plants, fungi, insects,

viruses and bacteria to chordates and humans. Only 1429 of the proteins belonged to fungal species. Some of the hits in this list are considered hypothetical proteins, or do not have a defined function. However, the majority of the proteins are called as DNA polymerase gamma or similar proteins. Two of the protein hits do not fall into either of these categories. These outliers are the alpha-beta hydrolase and the pentatricopeptide repeat-containing protein originally found in the blastp. The pentatricopeptide repeat-containing protein was found in a fungus, *Gigaspora margarita*. However, the protein was very close to the cutoff with an E-value of 3.00E-6.

The alpha-beta hydrolase hit is found within the genome of *Coniophora puteana,* a fungus and has an E-value of 3.00E-42. When the sequence for the alpha-beta hydrolase was searched within the CDD, the protein was shown to have the five domains. Domains belonging to the alpha-beta hydrolase superfamily cap each end of the protein, with a DNA mitochondrial polymerase exonuclease and polymerase domains, and an uncharacterized domain (accession number cI23818) in the middle. Figure 19 shows the map from the CDD of the alpha beta-hydrolase.[2] Using the alignment of the PSI-BLAST hits, the INT domain aligns to the region between the exonuclease and the polymerase domain. This was confirmed by an alignment of the alpha beta-hydrolase and the INT domain. An individual blastp of the alpha beta-hydrolase was performed. However, the only hits hypothetical proteins and DNA polymerase gamma.

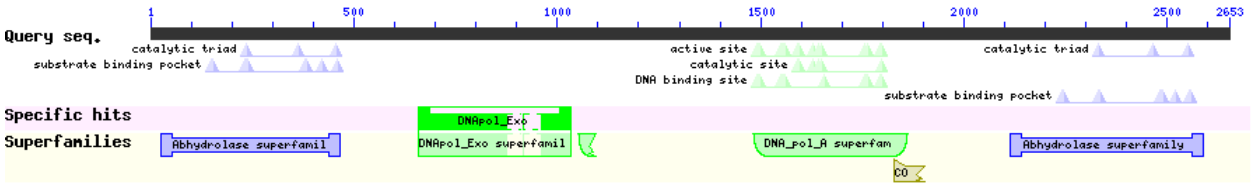


Figure 18. CDD Map of the Alpha Beta-Hydrolase from Coniophora puteana.

An alignment of the PSI-BLAST results did not provide any fully conserved residues, which is to be expected as PSI-BLAST builds off of an initial blastp. However, despite the large increase in the number of results, there are still highly conserved residues. These residues are Pro146, Leu148, Trp153, Pro157 and Trp166. Pro157 is the most conserved with 97.4% consensus.

Homo sapiens PolG and CNPolG

By aligning the full-length *Cryptococcus neoformans* DNA polymerase to the *Homo sapien* DNA polymerase (HSPolG; PBD: 3IKM_A), the relative locations of the 1-135 and INT domains were determined. The 1-135 region extends before the HSPolG, with the first residue of the HSPolG aligned with the last residue of the 1-135 domain. This confirms that the 1-135 region is novel as compared to the human enzyme. The INT region was determined to begin at residue 498 and end at 676 in the HSPolG, a region within the HSPolG spacer domain. The spacer divides into two smaller subdomains, the intrinsic processivity domain (residues 629 to 783) and the extended accessory-interacting determinant domain (AID, residues 477 to 579), and each subdomain increases the processivity of the polymerase.[16] The intrinsic processivity domain (IP) provides a site for the upstream primer-template DNA duplex to bind during replication, increasing the intrinsic processivity of the catalytic subunit, HSPolGA. To further increase processivity, the AID domain interacts with the accessory subunit (HSPolGB), similar to the use of thioredoxin in T7. In the absence of the accessory subunit the enzyme is an inefficient polymerase. Similar spacer regions have been found in polymerases from other species, including mouse and *Drosophila*.[16]

Despite being found in other species, most of the AID subdomain is often missing in fungal DNA polymerase gamma. To confirm the absence of the AID subdomain and to look for the IP subdomain, an alignment was performed on the DNA polymerase gamma catalytic

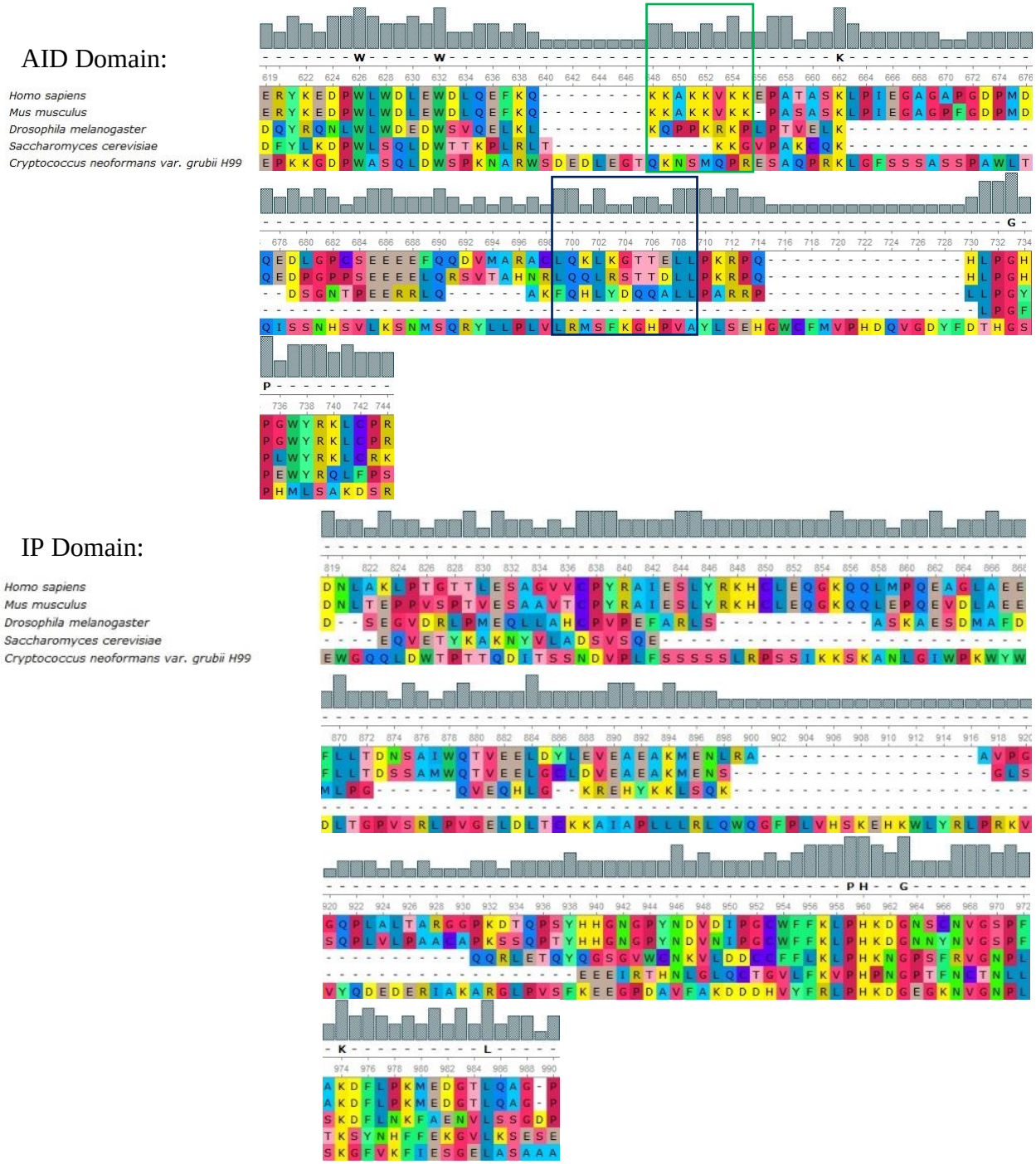


Figure 19. Alignment of the HSPolG's AID and IP subdomains. Using the ClustalW alignment tool in UGENE, the DNA polymerase gamma in *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Cryptococcus neoformans.* The K-tract is found in the green box, while the conserved hydrophobic residues are found in the navy box.

subunits from *Homo sapiens* (PBD: 3IKM), *Mus musculus* (accession: NP_001347025.1), *Drosophila melanogaster* (accession: AAC47658.1), *Saccharomyces cerevisiae* (accession: CAA89977.1) and *Cryptococcus neoformans*. This alignment is shown in figure 20. Interestingly, the *Cryptococcus neoformans* sequence appears to be present within these domains. The INT domain is aligned to start between the two subdomains and it ends over half-way through the IP domain. If the CNPolG contains an AID domain, it should contain the conserved or essential residues. Within the AID, there are four conserved residues that are responsible for the hydrophobic interactions with HSPolGB, L482, L485, L491 and L492. There is also a series of K residues in the beginning of the AID domain, known as the K-tract.[16] These residues are mostly conserved in the *H. sapiens, M. musculus* and *D. melanogaster* polymerases, but they are absent in the *S. cerevisiae.* Of the four hydrophobic conserved residues, CNPolG has one of the conserved residues (L482) and has a conservative substitution at L491. The K-tract is also only partly conserved. Despite the CNPolG covering all of this region, its sequence causes three gaps in the HSPolG, indicating that the HSPolG does not contain these residues.

Pull-Down Assay

PredictProtein predicted that both the 1-135 and INT domains could bind proteins and identifying potential binding partners can help deduce the function of a protein or domain. Therefore, pull-down assays attempted were on both domains. Before the optimization of the pull-down assays could begin, the domains and lysates had to be prepared. To prepare the lysates, a mutant strain (Δcap59) of *C. neoformans* was used. This strain contains a mutation in the gene *CAP59,* one of the four essential genes for capsule formation. When the gene is mutated, this protein is not produced, and the capsule is either very thin or missing from the

cell.[52] Making the cells easier to bust open, the cells can be busted without destroying the proteins inside of the cells. Bead-beating these cells until 80% of them are busted open ensures that the cells were freshly busted, and the proteins had suffered little degradation. Figure 21 shows an example of the cells before and after they were sufficiently broken.

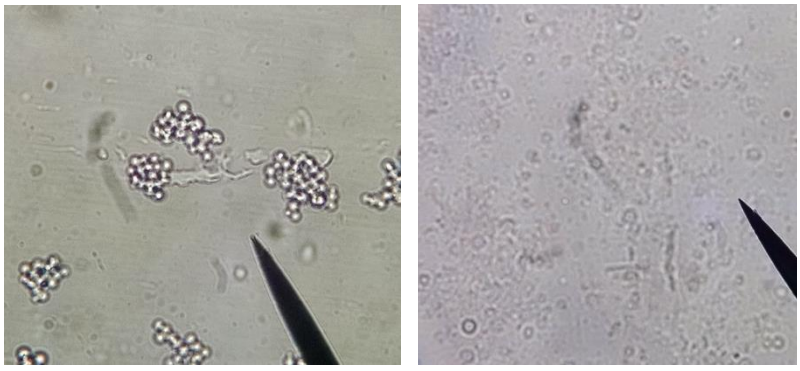


Figure 20. Normal (Left) Versus Broken (Right) *C. neoformans* Δ*cap59* Cells.

The 1-135 and INT domains were expressed and purified in E. coli. Due to the insoluble nature of the INT domain, the domain was unable to be purified at a high enough concentration for the assay. Therefore, the pull-down assays were optimized using Gp4A, SSB and 1-135. Gp4A is a bacteriophage T7 primase-helicase that does not contain a His tag. Without the tag, the Gp4A should not bind to the nickel resin, allowing it to serve as the negative control. SSB, or the T7 single stranded binding protein, contains a tag and serves as the positive control.

Initial runs of the assay did not use lysates. These runs were used to determine if the domains or proteins were binding to the column correctly. The initial run provided mixed results and is shown in Figure 22. In this run, only the 1-135 domain and the negative control, Gp4A, were used. As expected, the 1-135 domain bound to the column, however the Gp4A did as well. This indicates that there is non-specific binding occurring. Due to the non-specific binding, any

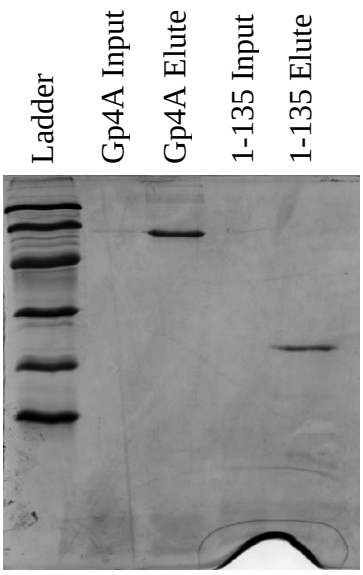


Figure 21. Initial Pull-down Assay.

proteins could stick to the resin, rendering it impossible to determine if any of the proteins were binding to the bait protein or domain. In an attempt to remedy this issue, triton was added to the buffer in varying concentrations (0.01% to 0.5%). However, this also did not prevent the non-specific binding. Another way to prevent non-specific binding is to add imidazole into the buffer. 10 mM imidazole was able to decrease Gp4A binding with the resin, with 40 mM imidazole preventing the binding almost entirely. Unfortunately, with SSB, imidazole concentrations over 10 mM prevented the protein from binding to the resin.

In a final attempt to eliminate non-specific binding, the assay's methods shifted away from the traditional procedure. Instead of using the nickel resin as a suspension in the buffer, the nickel resin was packed into a column. Initially only a 10 mM imidazole wash was used to wash the column. This caused an abundance of proteins coming off during the final elution, causing it to be difficult to analyze the gels. Using a 10 mM imidazole wash followed by a 40 mM imidazole wash decreased the amount of protein in the final elution. After the 40 mM imidazole wash was added to the protocol, the assay began to produce a possible binding partner. However, the size of these proteins was never consistent. Figure 23 shows the elution in multiple 1-135 containing samples, as well as samples without the 1-135 domain. Each of these assays were ran under similar conditions, except A-C did not include a 40 mM imidazole wash. The gel images C and F are inverted images of B and E, respectively. Comparing inverted and noninverted images can allow for easier identification of bands. For each assay, the lane on the left contains the 1-135 domain, while the lane on the right does not

contain the 1-135 domain. Runs B/C and E/F appear to have an extra protein band in the elution with the 1-135 domain versus the elution without the 1-135 domain. However, it is not the same band each time and it is not consistently present in all of the runs.

Two types of staining were attempted for these gels. The Coomassie stain worked well and is featured in Figure 23A-F. Silver staining is known to have greater sensitivity for lower protein concentrations. This experiment was attempted twice with silver stain. One of these attempts is shown in Figure 23G. While it was more sensitive, the protein bands were lighter and harder to compare. Therefore, further attempts were stained with Coomassie.

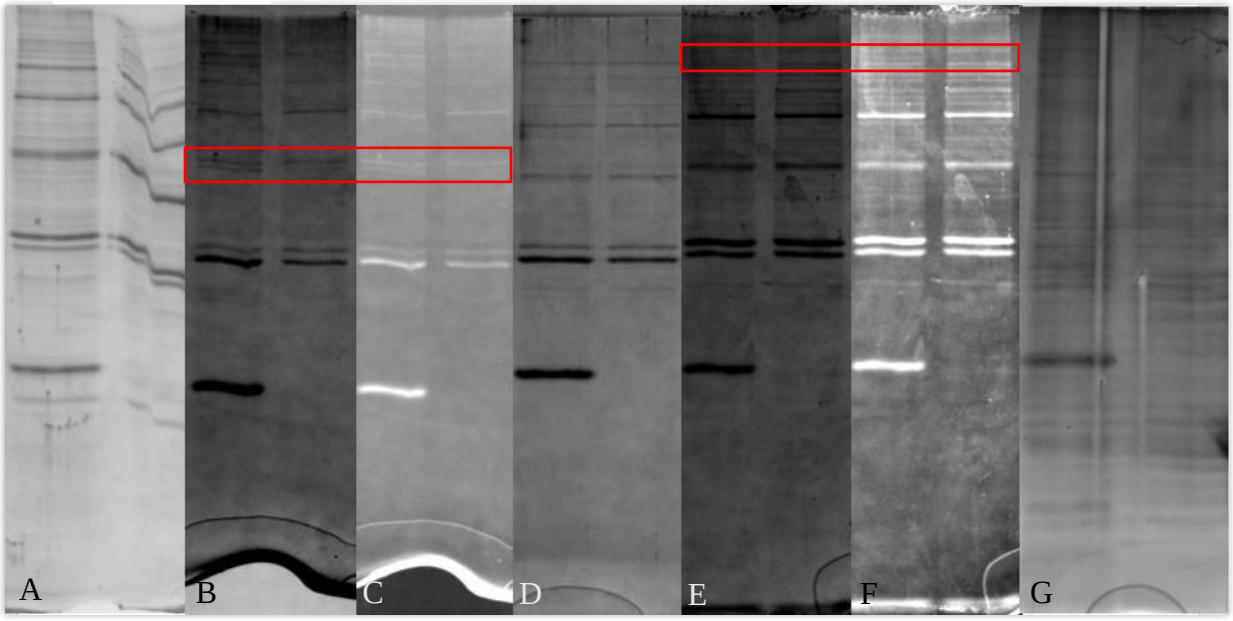


Figure 22. Pull-down Assay Elution Bands. The results of five assays are shown in this figure, as C and F are the inverted images of B and E, respectively. Each figure shows the input (left) and elution (right) of the assay. A-C did not include the 40 mM wash, thus they have more protein bands. The possible binding partners have been highlighted with a red square Coomassie was used to stain images A through F; whereas G underwent silver staining.

Discussion

Based on the bioinformatic analysis of 1-135 domain, the domain is mostly, if not completely disordered. The amino acid sequence consists mostly of disorder-promoting, small, polar amino acids. There are large regions of hydrophilic amino acids with very few aromatics or hydrophobic amino acids to disrupt these regions. The hydrophobic amino acids that are present are fairly spread out across the domain. Therefore, the domain likely does not need to fold to build a hydrophobic core. Looking at the proposed 3-D models from I-TASSER and Phyre[2], all of the models are vastly different and had very low confidence scores. This contributes to the idea of the domain being disordered. Intrinsically disordered regions (IDR) adopt their structure based on the binding of a substrate and are flexible. Instead of the lock and key method of binding, the regions can adapt to different substrates as they are available. Due to their flexible nature and adaptability, they often function in cell signaling hubs and regulation of transcription, translation and the cell cycle and protein transport.[34,38]

The BLAST analyses show that the 1-135 domain is exclusive to fungi within the order Tremellas. Coupled with the possibility of the domain being disordered, this domain could make a good candidate for a drug target. IDRs are of interest in drug targeting. These drugs can target the binding partner or the IDR.[38] If the drug targeted the 1-135 domain, only fungi would be affected. If the CNPolG was hindered, the cells would not be able to survive, as the fungus is an obligate aerobe.

The pull-down assay has shown promise as a method to identify the binding partners of the 1-135 domain, and potentially the INT domain. Using a column and adding a 40 mM imdazole wash greatly reduced the amount of nonspecific binding and background peaks. Initially, it appears as though there is still nonspecific binding occurring due to the change in isolated protein size. While this is possible, there is another possible explanation. If the 1-135

domain is indeed intrinsically disordered, it could have multiple binding partners. Depending on the diffusion of the proteins within the lysate, the domain could be retrieving proteins at different concentrations.

The INT domain was predicted by secondary structure prediction servers to be much less disordered than the 1-135 domain. In fact, there was only one small region that was predicted to be disordered. This is further supported by the I-TASSER and Phyre^2 models, as the models look very similar. The most important difference between the I-TASSER and Phyre^2 models is the gap central helix. Spanning residues 44 to 82, the helix in the Phyre^2 model contains a gap from residues 61 to 65. Model 5 is the only other model to have a gap in its helix. The rest of the models propose a straight helix without any gaps, but the overall shape is the same.

The BLAST results reveal that the INT region is primarily found in DNA polymerase gammas. Surprisingly, there was a wide variety of species found in the BLAST hits. Based on the 1-135 BLAST results, it was expected to have fungal results, but the hits extended past fungi and into species like birds, humans and viruses. The INT domain aligned with the spacer regions of other species, including *H. sapiens, M. musculus* and *D. melanogaster.* Each of these spacer regions are comparable to the spacer in the *H. sapiens* DNA polymerase gamma. According to the alignments of these polymerases, both the CNPolG and the *S. cerevisiae* polymerases do not contain the most of the residues for the AID subdomain. CNPolG only contains the first residue while *S. cerevisiae* is completely absent in this area.

The lack of an AID subdomain would explain why the fungal polymerases are monomeric and do not have accessory proteins similar to other DNA polymerase gammas. However, PredictProtein predicted that the INT region could bind proteins. This could indicate that the region binds to a partner that has yet to be discovered but doesn't function like the

accessory protein in the human polymerase gamma. This could be remedied by optimizing the pull-down assay for the INT domain, but the INT would need to be stabilized during production and produced at higher concentrations.

CHAPTER 4: CONCLUSIONS AND FURTHER WORK

It is crucial to continue to study DNA polymerases, especially family A DNA polymerases, as many of them are essential to life for their respective organisms, including the bacteriophages and *C. neoformans* studied here. Calling the bacteriophage polymerases "DNA polymerase I" insinuates that these polymerases would be very similar to the DNA polymerase I from *E. coli*. However, this study shows that they do not contain the 5'-3' exonuclease like its *E. coli* counterpart. Due to this distinction, the bacteriophage family A DNA polymerases should have a different name, such as DNA polymerase A. This would differentiate between the two types of polymerases. It would also represent the family that the polymerases belong to. However, it is critical that the name is written with a capital 'A'. If a lower-case 'a' is used, it could easily be confused with DNA polymerase α, a family B polymerase.

A larger investigation should attempt to decipher the unique domains in some of the actinophage polymerases, such as the UDG domains and cluster CD. First, the phages with unique domains should be 3D modeled using Phyre[2] and at least one other modeling server. As seen with the UDG domains, they can look drastically different, even within clusters. Modeling with PredictProtein or RaptorX can provide information on possible protein or DNA binding sites. Determining the structure can act as a good starting point for further analysis. Supposing the actinophage polymerases can be produced in the lab, the unique domains could be studied using pull-down assays. If the pull-down assays do not provide results, it would be interesting to see the effects on the processivity of the polymerase if these unique domains are removed.

The novel domains in the CNPolG have interesting implications for its DNA polymerization mechanism. HSPolG is known to be heterotrimeric, while the *Drosophila* DNA polymerase gamma is dimeric, and the *S. cerevisiae* DNA polymerase is known to be

monomeric. CNPolG has no known binding partners. However, both 1-135 and INT domains are predicted to bind proteins. Pull-down assays could help find these potential partners, but the assay would have to be optimized. First, the INT domain would have to be kept stable and it would need to be produced in a high concentration to run the assay. There is also a lot of background noise (other proteins) in the gels, even with the 40 mM imidazole wash. One way to remedy this is through mitochondrial enrichment. Instead of using the whole lysates, only the lysates from the mitochondria would be used. Since the CNPolG is a mitochondrial protein, it would be safe to assume that its binding partners would reside in the mitochondria as well. Once the binding partners are observed, they can be removed from the gel and ran on an LCMS for identification. Further study on this 1-135 domain could provide a possible drug target that would direct the drug to only impact the *C. neoformans* cells. With the results in this study, the INT domain is not exclusive to the Tremellas order. Further study on this domain, could provide insight into the polymerase's mechanism.

CHAPTER 5: MATERIALS AND METHODS

Bacteriophage Bioinformatics

Utilizing MYSQL, Phagesdb was searched for any phams that contained genes called as a DNA polymerase 1, DNA polymerase A or a similar function. 1351 genes spread across 13 phams were retrieved from the database on January 25th, 2020. Along with the sequences, the following information was retrieved for each bacteriophage: cluster, subcluster, bacterial host, accession number, gene number, gene length and notes.[18] The amino acid sequence for T7's DNA polymerase (accession number: NP_041982.1) was retrieved from NCBI's database (ncbi.nlm.nih.gov).

The algorithm ClustalW within UGENE was used to perform the multi-sequence alignments.[21] An alignment was performed for each pham, as well as an alignment of all 1351 sequences. Each of the alignments contained the T7 sequence as a reference. To confirm the sequences were correctly identified as family A DNA polymerases, the alignments were analyzed for the presence of acidic residues that coordinate divalent cations or interact with water and have been proven essential for exonuclease and polymerase activity. The following residues were identified as coordinating divalent cations in the T7 DNA polymerase: Asp5, Glu7, and Asp174 of the exonuclease active site, and Asp475, Asp654 and Glu655 of the polymerase active site. In addition to determining the presence of these residues, the alignments were studied for sequences that lacked domains or contained extra domains aside from the thioredoxin binding domain in T7.

Phyre2 analysis was performed on two sequences from each subcluster. If a sequence was determined to lack or contain additional domains, it was analyzed with the Phyre2 program as well.[20] Once hypothetical structures of the proteins were produced, they were uploaded

individually into WinCoot and superimposed with the T7 structure (PDB accession code: 1t79), *Mycobacterium smegmatis* DNA polymerase I (PDB: 6vde), human DNA polymerase nu (PDB: 4xvi) and DNA polymerase theta (PDB: 4x0p).[22] With the proteins superimposed, the presence of both active site residues and conserved domains was determined. The RMSDs were also recorded to determine the best superimposition for each polymerase. By superimposing the phage polymerases with polymerase nu and polymerase theta, it could be confirmed that the phages did not contain other types of family A polymerases.

Cryptococcus Neoformans DNA Polymerase Gamma

The bioinformatics analysis uses the *Cryptococcus neoformans var. grubii H99* sequence (accession: XP_012047111.1), which was retrieved from the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein). The bioinformatic analysis of the CNPolG was performed using multiple prediction servers to investigate the domains 1-135 and INT individually. JPred[4] and PROTEUS were used exclusively to predict the secondary structure of the domains while RaptorX and PredictProtein were able to provide potential binding sites with the secondary structures.[40,41,42,43] I-TASSER and Phyre2 provided the tertiary structure models and secondary structure predictions.[44]

To gain insight on a potential function, the domains and full-length protein were sent through the CDD server.[28] Then, similar proteins were found via blastp and PSI-BLAST.[45,47] The e-value threshold for each of these searches was 0.005 and the protein result cap was set at 20000. PSI-BLAST was allowed to continue performing new iterations until it no longer found new results. Once the sequences were acquired, they were imported into UGENE and aligned using the ClustalW algorithm.

For pull-down assays, lysates were prepared by growing Δcap59 cells with ampicillin at 30° C. The cells were then harvested using PBS as the resuspension buffer. To bust open the cells and retrieve the lysates, beads were added to the resuspension. Initially a bead-beating machine was used. This was later changed to vortexing, as the bead-beating machine was too harsh on the cells. The cells were placed in the bead-beater or vortexed for one minute and then chilled in ice for two minutes. This process would repeat until about 80% of the cells were lysed, as observed via a microscope.

pET-28 vectors containing the domains were transformed into BL21 (*E. coli*) cells. Using LB media and 50 mg/mL kanamycin, the cells were grown at 37 °C until the OD600 reached 0.6. Then the cells were put on ice, induced with 0.5 M IPTG and left to shake overnight at 16 °C. The cells were harvested through centrifugation at 3000 RPM and 4 °C. About 80 mL of nickel buffer A (50 mM Tris 8.0 pH, 0.5 M NaCl, 5% glycerol, 0.1 mM EDTA, 1 mM BME) was used to resuspend the pellets, which were then stored at -80 °C. To prepare for purification, the cells were thawed on ice and 1mg/mL lysozyme and 0.1 mM PMSF were added. A Branson sonifier was used to sonicate the cells in 1 minute intervals for 5 minutes total. Once the cells were centrifuged at 17000 RPM for 1.5 hours, the supernatant was collected and ran through a Q column using an AKTA FPLC.

Pull-down assays were based on the method used by Yong-Zheng W. et. al, where 1 µM protein was incubated in PBS with nickel-agarose beads for 30 minutes.[53] The liquid was removed and the proteins were eluted off the beads with PBS and 250 mM imidazole. Samples would be taken before and after elution for comparison. These samples were ran on a 15% SDS-PAGE gel. The method was optimized using SSB and Gp4 from bacteriophage T7. Unfortunately, this method had non-specific binding of the proteins to the beads. Various levels

of triton (0.01% to 0.5%) were added in an effort to prevent the non-specific binding. The detergent did not help the non-specific binding; therefore the pull-down assay procedure shifted from nickel-agarose beads to a nickel column. A similar process was followed, except the incubation time was increased to 1 hour, and the concentration of the protein was increased to 5 µM. With the introduction of the column, the cryptococcal lysates were also added to the experiment. The lysates were incubated (1 hour, 4 °C) with the domain prior to being put on the column. Two columns were ran simultaneously. One contained the lysates and the 1-135 domain, while the other contained lysate only to serve as a control. A 10 mM and a 40 mM imidazole wash would be added to help wash off non-specifically bound protein that was making it difficult to analyze the protein bands. The pull-downs were performed with cold reagents.

Samples were taken of the input before the washes, of the10 mM and 40 mM imidazole washes, and of the final elution. This would be ran on a 15% SDS-PAGE gel and stained with either Coomassie or Silver Stain.

REFERENCES

(1) Garcia-Diaz, M.; Bebenek, K. Multiple Functions of DNA Polymerases. *CRC. Crit. Rev. Plant Sci*. **2007**, *26* (2), 105–122.

(2) Cells Can Replicate Their DNA Precisely. https://www.nature.com/scitable/topicpage/cells-can-replicate-their-dna-precisely-6524830.

(3) Yang, W.; Gao, Y. Translesion and Repair DNA Polymerases: Diverse Structure and Mechanism. *Annu. Rev. Biochem.* **2018**, *87* (1), 239–261.

(4) Makiela-Dzbenska, K.; Jaszczur, M.; Banach-Orlowska, M.; Jonczyk, P.; Schaaper, R. M.; Fijalkowska, I. J. Role of Escherichia Coli DNA Polymerase I in Chromosomal DNA Replication Fidelity. *Mol. Microbiol.* **2009**, *74* (5), 1114–1127.

(5) Zahn, K. E.; Averill, A. M.; Aller, P.; Wood, R. D.; Doublié, S. Human DNA Polymerase θ Grasps the Primer Terminus to Mediate DNA Repair. *Nat. Struct. Mol. Biol.* **2015**, *22* (4), 304–311.

(6) Clokie, M. R. J.; Millard, A. D.; Letarov, A. V.; Heaphy, S. Phages in Nature. *Bacteriophage* **2011**, *1* (1), 31–45.

(7) Beese, L. S.; Derbyshire, V.; Steitz, T. A. Structure of DNA Polymerase I Klenow Fragment Bound to Duplex DNA. *American Association for the Advancement of Science*. **2009**, *260* (5106), 352–355.

(8) Barragan, N. C.; Sorvillo, F.; Kuo, T. Cryptococcosis-Related Deaths and Associated Medical Conditions in the United States, 2000-2010. *Mycoses* **2014**, *57* (12), 741–746.

(9) Gaona-Flores, V. Central Nervous System and Cryptococcus Neoformans. *N. Am. J. Med. Sci.* **2013**, *5* (8), 492.

(10) *C. neoformans* Infection https://www.cdc.gov/fungal/diseases/cryptococcosis-neoformans/index.html.

(11) Park, B. J.; Wannemuehler, K. A.; Marston, B. J.; Govender, N.; Pappas, P. G.; Chiller, T. M. Estimation of the Current Global Burden of Cryptococcal Meningitis among Persons Living with HIV/AIDS. *AIDS* **2009**, *23* (4), 525–530.

(12) Chang, Y. C.; Khanal Lamichhane, A.; Garraffo, H. M.; Walter, P. J.; Leerkes, M.; Kwon-Chung, K. J. Molecular Mechanisms of Hypoxic Responses via Unique Roles of Ras1, Cdc24 and Ptp3 in a Human Fungal Pathogen Cryptococcus Neoformans. *PLoS Genet.* **2014**, *10* (4).

(13) Chun, C. D.; Liu, O. W.; Madhani, H. D. A Link between Virulence and Homeostatic Responses to Hypoxia during Infection by the Human Fungal Pathogen Cryptococcus Neoformans. *PLoS Pathog.* **2007**, *3* (2), e22.

(14) Genga, A.; Bianchi, L.; Foury, F. A Nuclear Mutant of Saccharomyces Cerevisiae Deficient in Mitochondrial DNA Replication and Polymerase Activity. *J. Biol. Chem.* **1986**, *261* (20), 9328–9332.

(15) Chan, S. S. L.; Copeland, W. C. DNA Polymerase Gamma and Mitochondrial Disease: Understanding the Consequence of POLG Mutations. *Biochim. Biophys. Acta - Bioenerg.* **2009**, *1787* (5), 312–319.

(16) Lee, Y. S.; Kennedy, W. D.; Yin, Y. W. Structural Insight into Processive Human Mitochondrial DNA Synthesis and Disease-Related Polymerase Mutations. *Cell* **2009**, *139* (2), 312–324.

(17) Lodi, T.; Dallabona, C.; Nolli, C.; Goffrini, P.; Donnini, C.; Baruffini, E. DNA Polymerase γ and Disease: What We Have Learned from Yeast. *Front. Genet.* **2015**, *6* (MAR).

(18) Russell, D. A.; Hatfull, G. F. PhagesDB: The Actinobacteriophage Database. *Bioinformatics* **2017**, *33* (5), 784–786.

(19) Hatfull, G. F.; Jacobs-Sera, D.; Lawrence, J. G.; Pope, W. H.; Russell, D. A.; Ko, C.; Weber, R. J.; Patel, M. C.; Germane, K. L.; Edgar, R. H.; et al. Comparative Genomic Analysis of 60 Mycobacteriophage Genomes: Genome Clustering, Gene Acquisition, and Gene Size. *J. Mol. Biol.* **2010**, *397* (1), 119–143.

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

(21) Golosova, O.; Henderson, R.; Vaskin, Y.; Gabrielian, A.; Grekhov, G.; Nagarajan, V.; Oler, A. J.; Quiñones, M.; Hurt, D.; Fursov, M.; et al. Unipro UGENE NGS Pipelines and Components for Variant Calling, RNA-Seq and ChIP-Seq Data Analyses. *PeerJ* **2014**, *2014* (1), 1–15.

(22) 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.

(23) Magill, D. J.; McGrath, J. W.; O'Flaherty, V.; Quinn, J. P.; Kulakov, L. A. Insights into the Structural Dynamics of the Bacteriophage T7 DNA Polymerase and Its Complexes. *J. Mol. Model.* **2018**, *24* (7).

(24) Doublié, S.; Tabor, S.; Long, A. M.; Richardson, C. C.; Ellenberger, T. Crystal Structure of a Bacteriophage T7 DNA Replication Complex at 2.2 Å Resolution. *Nature* **1998**, *391* (6664), 251–258.

(25) Marchler-Bauer, A.; Bryant, S. H. CD-Search: Protein Domain Annotations on the Fly. *Nucleic Acids Res.* **2004**, *32* (WEB SERVER ISS.), 327–331.

(26) Ghosh, S.; Goldgur, Y.; Shuman, S. Mycobacterial DNA Polymerase I: Activities and Crystal Structures of the POL Domain as Apoenzyme and in Complex with a DNA Primer-Template and of the Full-Length FEN/EXO-POL Enzyme. *Nucleic Acids Res.* **2020**, *48* (6), 3165–3180.

(27) Lee, Y. S.; Gao, Y.; Yang, W. How a Homolog of High-Fidelity Replicases Conducts Mutagenic DNA Synthesis. *Nat. Struct. Mol. Biol.* **2015**, *22* (4), 298–303.

(28) 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.

(29) Bedford, E.; Tabor, S.; Richardson, C. C. The Thioredoxin Binding Domain of Bacteriophage T7 DNA Polymerase Confers Processivity on Escherichia Coli DNA Polymerase I. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94* (2), 479–484.

(30) Foster, B. M.; Rosenberg, D.; Salvo, H.; Stephens, K. L.; Bintz, B. J.; Hammel, M.; Ellenberger, T.; Gainey, M. D.; Wallen, J. R. Combined Solution and Crystal Methods Reveal the Electrostatic Tethers That Provide a Flexible Platform for Replication Activities in the Bacteriophage T7 Replisome. *Biochemistry* **2019**, *58* (45), 4466–4479.

(31) Schormann, N.; Ricciardi, R.; Chattopadhyay, D. Uracil-DNA Glycosylases - Structural and Functional Perspectives on an Essential Family of DNA Repair Enzymes. *Protein Sci.* **2014**, *23* (12), 1667–1685.

(32) Lucas-Lledó, J. I.; Maddamsetti, R.; Lynch, M. Phylogenomic Analysis of the Uracil-DNA Glycosylase Superfamily. *Mol. Biol. Evol.* **2011**, *28* (3), 1307–1317.

(33) Pearl, L. H. Structure and Function in the Uracil-DNA Glycosylase Superfamily. *Mutat. Res. - DNA Repair* **2000**, *460* (3–4), 165–181.

(34) Wright, P. E.; Dyson, H. J. Intrinsically Disordered Proteins in Cellular Signalling and Regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16* (1), 18–29.

(35) Uversky, V. N.; Dunker, A. K. Understanding Protein Non-Folding. *Biochim. Biophys. Acta* **2010**, *1804* (6), 1231–1264.

(36) Uversky, V. N. The Alphabet of Intrinsic Disorder II . Various Roles of Glutamic Acid in Ordered and Intrinsically Disordered Proteins. *Intrinsically Disord. Proteins* **2013**, *1* (1), 1–23.

(37) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; et al. Intrinsically Disordered Protein. *J. Mol. Graph. Model.* **2001**, *19* (1), 26–59.

(38) Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V. N. Introducing Protein Intrinsic Disorder. *Chem. Rev.* **2014**, *114* (13), 6561–6588.

(39) Campen, A.; Williams, R.; Brown, C.; Meng, J.; Uversky, V.; Dunker, A. TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. *Protein Pept. Lett.* **2008**, *15* (9), 956–963.

(40) Drozdetskiy, A.; Cole, C.; Procter, J.; Barton, G. J. JPred4: A Protein Secondary Structure Prediction Server. *Nucleic Acids Res.* **2015**, *43* (W1), W389–W394.

(41) Källberg, M.; Margaryan, G.; Wang, S.; Ma, J.; Xu, J. RaptorX Server: A Resource for Template-Based Protein Structure Modeling. In *Protein Structure Prediction*; Kihara, D., Ed.; Springer New York: New York, NY, **2014**; pp 17–27.

(42) Bernhofer, M.; Dallago, C.; Karl, T.; Satagopam, V.; Littmann, M.; Olenyi, T.; Qiu, J.; Schütze, K.; Ashkenazy, H.; Ben-tal, N.; et al. PredictProtein – Predicting Protein Structure and Function for 29 Years. *bioRxiv*. **2021**, 1–5.

(43) Montgomerie, S.; Cruz, J. A.; Shrivastava, S.; Arndt, D.; Berjanskii, M.; Wishart, D. S. PROTEUS2: A Web Server for Comprehensive Protein Structure Prediction and Structure-Based Annotation. *Nucleic Acids Res.* **2008**, *36* (Web Server issue), 202–209.

(44) Yang, J.; Zhang, Y. I-TASSER Server: New Development for Protein Structure and Function Predictions. *Nucleic Acids Res.* **2015**, *43* (W1), W174–W181.

(45) 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.

(46) Chen, R.; Jiang, Y. M.; Wei, S. C.; Wang, Q. M. Kwoniella Shandongensis Sp. Nov., a Basidiomycetous Yeast Isolated from Soil and Bark from an Apple Orchard. *Int. J. Syst. Evol. Microbiol.* **2012**, *62* (11), 2774–2777.

(47) Aravind, L.; Bhagwat, M. PSI-BLAST Tutorial. In *Comparative Genomics: Volumes 1 and 2*; Bergman NH, Ed.; Humana Press: Totowa NJ, 2007.

(48) Zheng, N.; Schulman, B. A.; Song, L.; Miller, J. J.; Jeffrey, P. D.; Wang, P.; Chu, C.; Koepp, D. M.; Elledge, S. J.; Pagano, M.; et al. Structure of the Cul1-Rbx1-Skp1-F BoxSkp2 SCF Ubiquitin Ligase Complex. *Nature* **2002**, *416* (6882), 703–709.

(49) Dominguez-Martin, M. A.; Polívka, T.; Sutter, M.; Ferlez, B.; Lechno-Yossef, S.; Montgomery, B. L.; Kerfeld, C. A. Structural and Spectroscopic Characterization of HCP2. *Biochim. Biophys. Acta - Bioenerg.* **2019**, *1860* (5), 414–424.

(50) Melnicki, M. R.; Leverenz, R. L.; Sutter, M.; López-Igual, R.; Wilson, A.; Pawlowski, E. G.; Perreau, F.; Kirilovsky, D.; Kerfeld, C. A. Structure, Diversity, and Evolution of a New Family of Soluble Carotenoid-Binding Proteins in Cyanobacteria. *Mol. Plant* **2016**, *9* (10), 1379–1394.

(51) Pedersen, L. C.; Benning, M. M.; Holden, H. M. Structural Investigation of the Antibiotic and ATP-Binding Sites in Kanamycin Nucleotidyltransferase. *Biochemistry* **1995**, *34* (41), 13305–13311.

(52) García-Rivera, J.; Chang, Y. C.; Kwon-Chung, K. J.; Casadevall, A. Cryptococcus Neoformans CAP59 (or Cap59p) Is Involved in the Extracellular Trafficking of Capsular Glucuronoxylomannan. *Eukaryot. Cell* **2004**, *3* (2), 385–392.

(53) Wu, Y. Z.; Manevich, Y.; Baldwin, J. L.; Dodia, C.; Yu, K.; Feinstein, S. I.; Fisher, A. B. Interaction of Surfactant Protein A with Peroxiredoxin 6 Regulates Phospholipase A2 Activity. *J. Biol. Chem.* **2006**, *281* (11), 7515–7525.

APPENDIX A

A. Bacteriophage Sequences:

NP_041982.1 DNA polymerase [*Escherichia phage T7*]

MIVSDIEANALLESVTKFHCGVIYDYSTAEYVSYRPSDFGAYLDALEAEVARGGLIVFHN
GHKYDVPALTKLAKLQLNREFHLPRENCIDTLVLSRLIHSNLKDTDMGLLRSGKLPGKR
FGSHALEAWGYRLGEMKGEYKDDFKRMLEEQGEEYVDGMEWWNFNEEMMDYNVQD
VVVTKALLEKLLSDKHYFPPEIDFTDVGYTTFWSESLEAVDIEHRAAWLLAKQERNGFP
FDTKAIEELYVELAARRSELLRKLTETFGSWYQPKGGTEMFCHPRTGKPLPKYPRIKTPK
VGGIFKKPKNKAQREGREPCELDTREYVAGAPYTPVEHVVFNPSSRDHIQKKLQEAGW
VPTKYTDKGAPVVDDEVLEGVRVDDPEKQAAIDLIKEYLMIQKRIGQSAEGDKAWLRY
VAEDGKIHGSVNPNGAVTGRATHAFPNLAQIPGVRSPYGEQCRAAFGAEHHLDGITGKP
WVQAGIDASGLELRCLAHFMARFDNGEYAHEILNGDIHTKNQIAAELPTRDNAKTFIYG
FLYGAGDEKIGQIVGAGKERGKELKKKFLENTPAIAALRESIQQTLVESSQWVAGEQQV
KWKRRWIKGLDGRKVHVRSPHAALNTLLQSAGALICKLWIIKTEEMLVEKGLKHGWD
GDFAYMAWVHDEIQVGCRTEEIAQVVIETAQEAMRWVGDHWNFRCLLDTEGKMGPN
WAICH

6VDE_1|Chains A,B|DNA polymerase I|*Mycolicibacterium smegmatis* (1772)

MSPAKTATKKTPAKAADDTPTLMLLDGNSLAFRAFYALPAENFKTQSGLTTNAVYGFT
AMLINLLRDEQPTHIAAAFDVSRQTFRKDKYPEYKEGRSATPDEFRGQIDITKEVLGALG
ITVLAEPGFEADDIIATLATQAEQEGYRVLVVTGDRDSLQLVSDQVTVLYPRKGVSELTR
FTPDAVVEKYGLTPQQYPDFAALRGDPSDNLPGIPGVGEKTATKWIVEYGSLQALVDNV
DAVKGKVGDALRANLSSVILNRELTDLIRDVPLPQTPDTLRMQPWNRDQIHRLFDDLEF
RVLRDRLFETLVAVEPEVEHGFDVRGRALEPGELAAWLSEHSLGSRFGVAVVGTHKAY

DADATALAIVAADGDGRYIDTSTLTPEDEAALASWLADPGPPKALHEAKLAMHDLAGR
GWTLRGVTSDTALAAYLVRPGQRSFTLDDLAVRYLHRELRAETPEQQQLSLLDDSDGV
DEQAVQTVILRACAVLDLADALDQELARIDSLSLLSRMELPVQRTLAEMEHAGIAVDLG
MLEQLQSEFADQIRDAAEAAYSVIGKQINLGSPKQLQAVLFDELEMPKTKKTKTGYTTD
ADALQSLFEKTGHPFLQHLLAHRDATRLKVTVDGLLNSVASDGRIHTTFNQTIAATGRL
SSTEPNLQNIPIRTEAGRRIRDAFVVGEGYAELMTADYSQIEMRIMAHLSRDAGLIEAFN
TGEDLHSFVASRAFSVPIDEVTPELRRRVKAMSYGLAYGLSAYGLAQQLKISTEEAKVQ
MEQYFDRFGGVRDYLRDVVDQARKDGYTSTVLGRRRYLPELDSSNRQVREAAERAAL
NAPIQGSAADIIKVAMINVDQAIKDAGLRSRILLQVHDELLFEVSEGEQGELEQLVREHM
GNAYPLDVPLEVSVGYGRSWDAAAH

4XVI_1|Chain A|DNA polymerase nu|*Homo sapiens* (9606)

KKHFCDIRHLDDWAKSQLIEMLKQAAALVITVMYTDGSTQLGADQTPVSSVRGIVVLV
KRQAEGGHGCPDAPACGPVLEGFVSDDPCIYIQIEHSAIWDQEQEAHQQFARNVLFQTM
KCKCPVICFNAKDFVRIVLQFFGNDGSWKHVADFIGLDPRIAAWLIDPSDATPSFEDLVE
KYCEKSITVKVNSTYGNSSRNIVNQNVRENLKTLYRLTMDLCSKLKDYGLWQLFRTLEL
PLIPILAVMESHAIQVNKEEMEKTSALLGARLKELEQEAHFVAGERFLITSNNQLREILFG
KLKLHLLSQRNSLPRTGLQKYPSTSEAVLNALRDLHPLPKIILEYRQVHKIKSTFVDGLLA
CMKKGSISSTWNQTGTVTGRLSAKHPNIQGISKHPIQITTPKNFKGKEDKILTISPRAMFV
SSKGHTFLAADFSQIELRILTHLSGDPELLKLFQESERDDVFSTLTSQWKDVPVEQVTHA
DREQTKKVVYAVVYGAGKERLAACLGVPIQEAAQFLESFLQKYKKIKDFARAAIAQCH
QTGCVVSIMGRRRPLPRIHAHDQQLRAQAERQAVNFVVQGSAADLCKLAMIHVFTAVA

ASHTLTARLVAQIHDELLFEVEDPQIPECAALVRRTMESLEQVQALELQLQVPLKVSLSA
GRSWGHLVPLQ

4X0P_1|Chains A,B,C,D|DNA polymerase theta|*Homo sapiens* (9606)

GFKDNSPISDTSFSLQLSQDGLQLTPASSSSESLSIIDVASDQNLFQTFIKEWRCKKRFSISL
ACEKIRSLTSSKTATIGSRFKQASSPQEIPIRDDGFPIKGCDDTLVVGLAVCWGGRDAYYF
SLQKEQKHSEISASLVPPSLDPSLTLKDRMWYLQSCLRKESDKECSVVIYDFIQSYKILLL
SCGISLEQSYEDPKVACWLLDPDSQEPTLHSIVTSFLPHELPLLEGMETSQGIQSLGLNAG
SEHSGRYRASVESILIFNSMNQLNSLLQKENLQDVFRKVEMPSQYCLALLELNGIGFSTA
ECESQKHIMQAKLDAIETQAYQLAGHSFSFTSSDDIAEVLFLELKLPPNREMKNQGSKKT
LGSTRRGIDNGRKLRLGRQFSTSKDVLNKLKALHPLPGLILEWRRITNAITKVVFPLQRE
KCLNPFLGMERIYPVSQSHTATGRITFTEPNIQNVPRDFEIKMPTLVGESPPSQAVGKGLL
PMGRGKYKKGFSVNPRCQAQMEERAADRGMPFSISMRHAFVPFPGGSILAADYSQLEL
RILAHLSHDRRLIQVLNTGADVFRSIAAEWKMIEPESVGDDLRQQAKQICYGIIYGMGAK
SLGEQMGIKENDAACYIDSFKSRYTGINQFMTETVKNCKRDGFVQTILGRRRYLPGIKD
NNPYRKAHAERQAINTIVQGSAADIVKIATVNIQKQLETFHSTFKSHGHREGMLQSDQT
GLSRKRKLQGMFCPIRGGFFILQLHDELLYEVAEEDVVQVAQIVKNEMESAVKLSVKLK
VKVKIGASWGELKDFDV

Actinophage sequences available upon request. Contact Dr. Jamie Wallen (jamiewallen@email.wcu.edu) or Cecilia Baumgardner (ceciliabaumgardner@gmail.com).

B. *Cryptococcus neoformans* Analysis Sequences

XP_012047111.1 DNA polymerase gamma 1 [*Cryptococcus neoformans var. grubii H99*]

MRKALDISRLTRPARIRCRPSLFLRNRSLSSSASQSKPSDAPVKISDGKEEGVGKPLIPAFG

ARRAEMEDYILAMEMAKLEDGYGQPRVRKIRKSKLPSLHDPQSFLCDSTQASSSKVTSS

ASPTSQPSRKGKEKEVVSNDYATNVPNQDVQTLEDAKPIDSGQSKSGPRRNPVGVQML

SSSLHSQLFPGQPLPKPPQALLDISKRHLKDNDLFPEGAAVLPEISFNLPSLRGNNIRDHFH

TLGQYTAEPYASMAREFAATKLPAKPDRWEMGRPGWTKYYSDGRMEAVDDLGDETL

VSFDVEVLYKLSRFPVMATAVTPNAWYSWLSPVIFQSPPAEIPKPLPPWEASIPIYHPNEL

IPLFNNESSIPRIVIGHNVGYDRARVKEEYSIERTQTRWLDTLSLHVSTRGITSVQRPAWM

AYRKNKKAKKLREQENLSILQEMAEKSGDGTIMESLQEFGAASETEEAEALQSRWEDV

TSMNSLAEVAALHCGYPVDKSVRDRFGDDSIKHASQIHSELHQLLSYCADDVRVTHDV

YAKVFPLFLESCPHPATLSGILSMGSSFLPIDQSWKEYLRNAEETYREMDVAVKKALRLL

AEKLRAEGEPKKGDPWASQLDWSPKNARWSDEDLEGTQKNSMQPRESAQPRKLGFSSS

ASSPAWLTQISSNHSVLKSNMSQRYLLPLVLRMSFKGHPVAYLSEHGWCFMVPHDQVG

DYFDTHGSPHMLSAKDSRLEKLEESYSFFRIGNAGSPKKTKLVGPSIKPFVNSGDLTSA

YPELLVKVMKTDLNDVVEDLWECVVDMGNLKESEWGQQLDWTPTTQDITSSNDVPLF

SSSSSLRPSSIKKSKANLGIWPKWYWDLTGPVSRLPVGELDLTCKKAIAPLLLRLQWQGF

PLVHSKEHKWLYRLPRKVYQDEDERIAKARGLPVSFKEEGPDAVFAKDDDHVYFRLPH

KDGEGKNVGNPLSKGFVKFIESGELASAAAESGDDVAAKAAADATNMNAFCSYWISSR

ERIMDQMVVYRDQEFGMILPQVITMGTVTRRAVEATWLTASNAKKNRVGSELKAMVR

APPGYSIVGADVDSEELWISSVMGDSQFGMHGATAIGWMTLEGTKSAGTDLHSKTANIL

GISRDAAKVFNYSRIYGAGKKHAVQLLLQGDSKLTKETAGKLADNLYKSTKGAKAVRA

RNLPVASIPSLWHGGSESYLFNTLEAIALSDRPTTPALGCGVTRALRKSYLEENASYLPSR

VNWVVQSSGVDYLHLLIVSMEYLIKKYNIQARYLISVHDEVRYLAKEEDRYRTALALQI

ANAWTRALFCFNLGIDDMPQGITFFSAVDIDHVLRKEVFLTCETPSHPKVIPAGESLDIIS
LLEKIPRGDLGTPVPDDLQPPTDIKPPVALFPNIQSAQHRQFLQAQASKGGMGAKKWLD
NLPPVQYIDEVNEGNEKPYQKSHKKAVLSSSKKFQ

3IKM_1|Chains A,D|DNA polymerase subunit gamma-1|*Homo sapiens* (9606)

RHNPLDIQMLSRGLHEQIFGQGGEMPGEAAVRRSVEHLQKHGLWGQPAVPLPDVELRL
PPLYGDNLDQHFRLLAQKQSLPYLEAANLLLQAQLPPKPPAWAWAEGWTRYGPEGEA
VPVAIPEERALVFDVEVCLAEGTCPTLAVAISPSAWYSWCSQRLVEERYSWTSQLSPAD
LIPLEVPTGASSPTQRDWQEQLVVGHNVSFDRAHIREQYLIQGSRMRFLDTMSMHMAIS
GLSSFQRSLWIAAKQGKHKVQPPTKQGQKSQRKARRGPAISSWDWLDISSVNSNSLAEV
HRLYVGGPPLEKEPRELFVKGTMKDIRENFQDLMQYCAQDVWATHEVFQQQLPLFLER
CPHPVTLAGMLEMGVSYLPVNQNWERYLAEAQGTYEELQREMKKSLMDLANDACQL
LSGERYKEDPWLWDLEWDLQEFKQKKAKKVKKEPATASKLPIEGAGAPGDPMDQEDL
GPCSEEEEFQQDVMARACLQKLKGTTELLPKRPQHLPGHPGWYRKLCPRLDDPAWTPG
PSLLSLQMRVTPKLMALTWDGFPLHYSERHGWGYLVPGRRDNLAKLPTGTTLESAGVV
CPYRAIESLYRKHCLEQGKQQLMPQEAGLAEEFLLTDNSAIWQTVEELDYLEVEAEAK
MENLRAAVPGQPLALTARGGPKDTQPSYHHGNGPYNDVDIPGCWFFKLPHKDGNSCN
VGSPFAKDFLPKMEDGTLQAGPGGASGPRALEINKMISFWRNAHKRISSQMVVWLPRS
ALPRAVIRHPDYDEEGLYGAILPQVVTAGTITRRAVEPTWLTASNARPDRVGSELKAMV
QAPPGYTLVGADVDSQELWIAAVLGDAHFAGMHGCTAFGWMTLQGRKSRGTDLHSKT
ATTVGISREHAKIFNYGRIYGAGQPFAERLLMQFNHRLTQQEAAEKAQQMYAATKGLR
WYRLSDEGEWLVRELNLPVDRTEGGWISLQDLRKVQRETARKSQWKKWEVVAERAW
KGGTESEMFNKLESIATSDIPRTPVLGCCISRALEPSAVQEEFMTSRVNWVVQSSAVDYL

HLMLVAMKWLFEEFAIDGRFCISIHDEVRYLVREEDRYRAALALQITNLLTRCMFAYKL
GLNDLPQSVAFFSAVDIDRCLRKEVTMDCKTPSNPTGMERRYGIPQGEALDIYQIIELTK
GSLEKRSQPGP

CAA89977.1 MIP1 [*Saccharomyces cerevisiae*]

MTKLMVRSECMLRMVRRRPLRVQFCARWFSTKKNTAEAPRINPVGIQYLGESLQRQVF
GSCGGKDEVEQSDKLMELSKKSLKDHGLWGKKTLITDPISFPLPPLQGRSLDEHFQKIGR
FNSEPYKSFCEDKFTEMVARPAEWLRKPGWVKYVPGMAPVEVAYPDEELVVFDVETL
YNVSDYPTLATALSSTAWYLWCSPFICGGDDPAALIPLNTLNKEQVIIGHNVAYDRARV
LEEYNFRDSKAFFLDTQSLHIASFGLCSRQRPMFMKNNKKKEAEVESEVHPEISIEDYDD
PWLNVSALNSLKDVAKFHCKIDLDKTDRDFFASTDKSTIIENFQKLVNYCATDVTATSQ
VFDEIFPVFLKKCPHPVSFAGLKSLSKCILPTKLNDWNDYLNSSESLYQQSKVQIESKIVQ
IIKDIVLLKDKPDFYLKDPWLSQLDWTTKPLRLTKKGVPAKCQKLPGFPEWYRQLFPSK
DTVEPKITIKSRIIPILFKLSWENSPVIWSKESGWCFNVPHEQVETYKAKNYVLADSVSQE
EEEIRTHNLGLQCTGVLFKVPHPNGPTFNCTNLLTKSYNHFFEKGVLKSESELAHQALQI
NSSGSYWMSARERIQSQFVVPSCKFPNEFQSLSAKSSLNNEKTNDLAIIIPKIVPMGTITRR
AVENAWLTASNAKANRIGSELKTQVKAPPGYCFVGADVDSEELWIASLVGDSIFNVHG
GTAIGWMCLEGTKNEGTDLHTKTAQILGCSRNEAKIFNYGRIYGAGAKFASQLLKRFNP
SLTDEETKKIANKLYENTKGKTKRSKLFKKFWYGGSESILFNKLESIAEQETPKTPVLGC
GITYSLMKKNLRANSFLPSRINWAIQSSGVDYLHLLCCSMEYIIKKYNLEARLCISIHDEI
RFLVSEKDKYRAAMALQISNIWTRAMFCQQMGINELPQNCAFFSQVDIDSVIRKEVNMD
CITPSNKTAIPHGEALDINQLLDKSNSKLGKPNLDIDSKVSQYAYNYREPVFEEYNKSYT
PEFLKYFLAMQVQSDKRDVNRLEDEYLRECTSKEYARDGNTAEYSLLDYIKDVEKGKR

TKVRIMGSNFLDGTKNAKADQRIRLPVNMPDYPTLHKIANDSAIPEKQLLENRRKKENR
IDDENKKKLTRKKNTTPMERKYKRVYGGRKAFEAFYECANKPLDYTLETEKQFFNIPID
GVIDDVLNDKSNYKKKPSQARTASSSPIRKTAKAVHSKKLPARKSSTTNRNLVELERDIT
ISREY

NP_001347025.1 DNA polymerase subunit gamma-1 isoform 3 [*Mus musculus*]
MRFLDTMSMHMAISGLSSFQRSLWMGAKQGKHKTQQSTKRGQKSPRKANGPAISSWD
WMDISSANNLADVHNLYVGGPPLEKEPRELFVKGSMRDIRENFQDLMQYCARDVWAT
FEVFQQQLPLFLERCPHPVTLAGMLEMGVSYLPVNQNWERYLTEAQNTYEELQREMKK
SLMDLANDACQLLSGERYKEDPWLWDLEWDLQEFKQKKAKKVKKPASASKLPIEGAG
PFGDPMDQEDPGPPSEEEELQRSVTAHNRLQQLRSTTDLLPKRPQHLPGHPGWYRKLCP
RLDDPAWAPGPSLLSLQMRVTPKLMALTWDGFPLHYSDSHGWGYLVPGRRDNLTEPP
VSPTVESAAVTCPYRAIESLYRKHCLEQGKQQLEPQEVDLAEEFLLTDSSAMWQTVEEL
GCLDVEAEAKMENSGLSQPLVLPAACAPKSSQPTYHHGNGPYNDVNIPGCWFFKLPHK
DGNNYNVGSPFAKDFLPKMEDGTLQAGPGGASGPRALEINKMISFWRNAHKRISSQMV
VWLPRSALPRVVTRHPSFDEEGHYGAILPQVVTAGTITRRAVEPTWLTASNARPDRVGS
ELKAMVQAPPGYVLVGADVDSQELWIAAVLGDAHFAGMHGCTAFGWMTLQGRKSRG
TDLHSKTAATVGISREHAKIFNYGRIYGAGQSFAERLLMQFNHRLTRQEAAEKAQQMY
AVTKGLRRYRLSADGEWLVKQLNLPVDRTEDGWVSLQDLRMIRREASRKSRWKKWE
VASERAWTGGTESEMFNKLESIAMSDTPRTPVLGCCISRALEPSVVQGEFITSRVNWVV
QSSAVDYLHLMLVAMKWLFEEFAIDGRFCISIHDEVRYLVREEDRYRAALALQITNLLT
RCMFAYKLGLNDLPQSVAFFSAVDIDQCLRKEVTMDCKTPSNPTGMERRYGIPQGEAL
DIYQIIELTKGSLEKRSQPGP

AAC47290.1 DNA polymerase gamma [*Drosophila melanogaster*]

MQFHLIRKYASKVSREHYASSSVKIFRRVKPPQKVNKPKKPENVENGPTEYAENLVKVQ
MISRNLHAQLFPQAPRSISEQQVASAKVYKDELRRHGVDIESSAPVSDVQLKLPALRGA
NIEEHFHNIAKEQVQPYEELLLPLVQCEQLPKRPKRWAFHTGWTAYDPEDGTATPVDHP
LEKGLVFDVEVCVSEGQAPVLATAVSTKRWYSWVSSKLTKHRLSVEKLEPLDVDTDSE
RPHYTTDELIPLGTTGPGLVVGHNVSYDRARLKEQYLTEDTGTRFVDTMSLHMCVSGV
TSYQRAMLKSKKEPAAEDLGWLEQSSLNSLVELHRLYCGGDTLSKEPRNIFVEGTLE
QVRQSFQSLTNYCASDVEATHRILRVLYPLYAERFPHPASLAGMLEMGSAYLPVNSNW
ERYIREAQLTYEDLSIEAKYHLGRRAEEACSLLLDDQYRQNLWLWDEDWSVQELKLKQ
PPKRKPLPTVELKDSGNTPEERRLQAKFQHLYDQQALLPARRPLLPGYPLWYRKLCRKP
PAKRADEILEDDEEPWSPGASEISTGMQIAPKLLSLCWEGYPLHYEREQGWGFLVPFRSD
SEGVDRLPMEQLLAHCPVPEFARLSASKAESDMAFDMLPGQVEQHLGKREHYKKLSQK
QQRLETQYQGSGVWCNKVLDDCCFFLKLPHKNGPSFRVGNPLSKDFLNKFAENVLSSG
DPSCQAAARVIDIARMMSYWRNNRDRIMGQMVVWLDSQQLPNEFTGEKCQPIAYGAIC
PQVVACGTLTRRAMEPTWMTASNSRPDRLGSELRSMVQAPPGYRLVGADVDSQELWI
ASVLGDAYACGEHGATPLGWMTLSGSKSNGSDMHSITAKAVGISRDHAKVINYARIYG
AGQLFAETLLRQFNPTFSASEAKAKAMKMFSITKGKRVYRLREEFHDELEDRAYSSYEA
SRLAIQRNRTLAEVFHRPKWQGGTESAMFNRLEEIATGSQPRTPFLGGRLSRALEADTGP
EQEQRFLPTRINWVVQSGAVDFLHLMLVSMRWLMGSHVRFCLSFHDELRYLVKEELSP
KAALAMHITNLMTRSFCVSRIGLQDLPMSVAFFSSVEVDTVLRKECTMDCKTPSNPHGL
RIGYGIQPGQSLSVAEAIEKAGGNDVSQWDWIKKS