THE UNIVERSITY OF CHICAGO

BACTERIOPHAGE N4 COMPARATIVE GENOMICS AND THE MECHANISM OF N4 RNAPII TRANSCRIPTION INITIATION

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY

BRYAN RICHARD LENNEMAN

CHICAGO, ILLINOIS

JUNE 2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

5-IdU	5-iododeoxyuracil
A_{260}	absorbance at 260 nm
A_{280}	absorbance at 280 nm
aa	amino acid
AcN	acetonitrile
am	amber mutation
ANI	average nucleotide identity
APS	ammonium persulfate
ATP	adenosine triphosphate
BC	binary complex
BLAST	basic local alignment search tool
BLASTn	nucleotide BLAST
BLASTp	protein BLAST
bp	base pair
BSA	bovine serum albumin
CAO	cluster associated ORFams
Cas	CRISPR-associated
CII	cluster isolation index
CLASO	cluster average shared ORFams
CRISPR	clustered regularly interspaced palindromic repeats
cryo-EM	cryo-electron microscopy
CTP	cytidine triphosphate

dCTP	deoxycytidine triphosphate
DELTA-BLAST	domain enhanced lookup time accelerated BLAST
DNA	deoxyribonucleic acid
DNAP	DNA polymerase
dNTP	deoxyribonucleotide triphosphate
dsDNA	double-stranded DNA
dTMP	deoxythymidine monophosphate
DTT	dithiothreitol
dTTP	deoxythymidine triphosphate
dUMP	deoxyuridine monophosphate
dUTP	deoxyuridine triphosphate
E. coli	*Escherichia coli*
EcoSSB	*E. coli* single-stranded DNA-binding protein
EDTA	ethylenediaminetetraacetic acid
EMSA	electrophoretic mobility shift assay
FRET	fluorescence resonance energy transfer
GMPCPP	guanylyl 5′-α,β-methylenediphosphonate
gp	gene product
GTP	guanosine triphosphate
H-D	hydrogen-deuterium
His_6	hexahistidine tag
HMG-box	high-mobility-group-box
hr	hour

IMAC	immobilized metal affinity chromatography
kbp	kilobase pair
k_{cat}	rate constant for limiting step at saturation
K_d	dissociation constant
kDa	kiloDalton
K_m	Michaelis constant
LB	Luria-Bertani broth
LC-MS/MS	liquid chromatography-tandem mass spectrometry
MCL	Markov Cluster Algorithm
min	minute
MOI	multiplicity of infection
mRNA	messenger RNA
mtDNA	mitochondrial DNA
mtRNAP	mitochondrial RNA polymerase
MWCO	molecular weight cutoff
m/z	mass to charge ratio
N4SSB	N4 single-stranded DNA-binding protein
NCBI	National Center for Biotechnology Information
NMP	nucleotide monophosphate
NSCU	normalized synonymous codon usage
N-site	NTP-binding site
nt	nucleotide
NT	non-promoter template

NTP	nucleotide triphosphate
OD_{600}	optical density at 600 nm
ORF	open reading frame
ORFam	open reading frame family
ORFan	single-ORF ORFam
PAGE	polyacrylamide gel electrophoresis
*p*Bpa	*p*-benzoyl-L-phenylalanine
PCR	polymerase chain reaction
PDB	Protein Data Bank
PFU	plaque-forming unit
PNK	polynucleotide kinase
PPi	pyrophosphate
ppm	parts per million
PPR	pentatricopeptide repeat
PSI-BLAST	position-specific iterated BLAST
P-site	priming site
PVDF	polyvinylidene fluoride
RCF	relative centrifugal force
RI	refractive index
RLM	RNA ligase mediated
RNA	ribonucleic acid
RNAP	RNA polymerase
RNAPII	N4 RNA polymerase II

rNTP	ribonucleotide triphosphate
RPM	rotations per minute
rRNA	ribosomal RNA
RT	retention time
RT-PCR	reverse transcription PCR
SAM	S-adenosyl-L-methionine
SD	standard deviation
SDS	sodium dodecyl sulfate
sec	second
SEM	standard error of the mean
SSB	single-stranded DNA-binding protein
ssDNA	single-stranded DNA
Suppl	supplement
TAP	tobacco acid pyrophosphatase
TEMED	tetramethylethylenediamine
tRNA	transfer RNA
UTP	uridine triphosphate
UV	ultraviolet light
V_{max}	maximum velocity
vRNAP	N4 virion RNA polymerase
WT	wild-type
β-IH	β-intercalating hairpin

ABSTRACT

Bacteriophage N4, an *Escherichia coli* (*E. coli*) K-12 strain-specific podovirus, is the founding member of the N4-like phage subfamily, which contains phage species prevalent in the human gut microbiome. Although numerous phage species have been identified as N4-like in available databases recently, no disciplined methodology for the classification of phages to the N4-like phage subfamily currently exists. Therefore, I defined a universal framework for the classification of new N4-like phage species using N4 virion-encapsidated RNA polymerase (vRNAP) as a marker due to its activity across numerous N4 physiological processes and ease of detection in genomic sequences and isolated virions. 54 phage species encoding N4 vRNAP homologs were detected in current databases through DELTA-BLAST homology search. These species share similar genomic and morphological properties as N4, but infect a broad range of Proteobacteria occupying diverse ecological niches across five continents. I created a reticulate phylogeny of N4-like phages through shared ORFs, which revealed greater genetic similarity among phages infecting closely related hosts and identified the N4 transcriptional machinery (vRNAP, N4 RNAPII, and N4SSB) as a hallmark of N4-like phages due to its complete conservation across the subfamily. I also identified ORFs uniquely conserved within clusters of phages that encode putative host specificity factors, which are excellent targets for the design of novel antibiotics and engineering phages to infect new hosts.

Comparative genomics studies demonstrated that the N4 transcriptional strategy utilizing the sequential activity of a virion-encapsidated RNAP for early transcription, a heterodimeric RNAP for middle transcription, and single-stranded DNA-binding protein (SSB) required for coupling late transcription with DNA replication is a unique feature of N4 and its relatives. The phage-encoded heterodimeric N4 RNA polymerase II (N4 RNAPII), responsible for the

transcription of N4 middle genes, is a member of the T7-like RNA polymerase family. Unlike T7 RNAP, N4 RNAPII cannot initiate transcription from double-stranded templates and requires the additional factor N4 gp2 for transcription *in vivo*. Gp2 is an SSB that activates N4 RNAPII transcription through direct interaction with N4 RNAPII. In this work, I define the requirements for N4 RNAPII promoter recognition and elucidate the molecular mechanism of transcription activation by its transcription factor gp2. *In vitro* transcription, DNA binding, and crosslinking assays show that the N4 RNAPII specificity loop directly interacts with bases in the template strand within short, AT-rich sequences for promoter recognition and transcription initiation. I also used crosslinking and mass spectrometry techniques to show that the gp2 N-terminus localizes to the N4 RNAPII active site, where it activates N4 RNAPII transcription by increasing the catalytic efficiency of first phosphodiester bond formation 1.5-fold. I propose a model for N4 RNAPII transcription activation where interaction between the N-terminus of ssDNA-bound gp2 and the N4 RNAPII active site recruits N4 RNAPII to single-stranded templates and coordinates the N4 RNAPII active site to increase the catalytic efficiency of transcription initiation.

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Lucia Rothman-Denes for her guidance, mentorship, and support throughout the course of this work. Lucia has been wonderful example of passion and dedication and I will carry many of the lessons she has taught me for the rest of my life. I would also like to thank the members of my thesis committee, Jon Staley, Sean Crosson, and Alex Ruthenburg, for their valuable scientific insights and support.

The past members of the Rothman-Denes lab have been instrumental in driving this work forward and have provided me invaluable guidance and comradery. I would first like to thank Abby Strayart and Mike Hammer for their excellent work to lay the foundation for these projects and to thank Jacob Waldbauer for a wonderful collaboration to help me with the crosslinking and mass spectrometry data. I am also particularly grateful to Wolf Epstein, Laura Satkamp, Eda Koculi, Abby Strayart, and Nhi Khuong for their guidance. Their technical expertise and passion for science were instrumental to me as I began this journey. I have also been fortunate enough to work alongside many wonderful scientists in laboratories throughout the 8th floor of CLSC. They have fostered a wonderfully collegiate and collaborative atmosphere and I thank them for their generosity and for their friendship.

Finally, I would also like to express my unceasing gratitude to my family. My parents have instilled a passion for learning and curiosity in me from an early age and have provided me the perfect example of hard work, dedication, and compassion. They have been a fountain of love and support throughout my entire life and I am eternally grateful to them. My wife Megan has been my constant source of love and joy. Words cannot express my gratitude for her patience and support and I am truly blessed to spend a lifetime with her.

CHAPTER I: INTRODUCTION

OVERVIEW OF THE BACTERIOPHAGE N4 INFECTIOUS CYCLE

Bacteriophage N4,a lytic *E. coli* K-12 strain-specific bacteriophage, is a member of the *Podoviridae* family of tailed phages with an icosahedral head ~70 nm in diameter and a short non-contractile tail ~35 nm in length (1). To initiate infection, N4 recognizes the outer membrane protein NfrA, present in 3-5 copies per cell, through its tail sheath protein (2–4). Receptor-sheath interaction must induce a conformational change that leads to portal opening to allow the 4±1 copies of vRNAP, a 3,500-residue polypeptide located just above the portal, to be injected in an unfolded conformation through the 25 Å diameter tail tube (1). Upon injection, vRNAP localizes to the cytoplasmic membrane and is responsible for the injection of the ~ 500 bp of the left end of the N4 genome (5, 6). At this stage, the introduction of negative supercoils by *E. coli* DNA gyrase induces the extrusion of early promoter hairpins that, stabilized by *E. coli* single-stranded DNA-binding protein (*Eco*SSB), yield the "activated promoter" (7–12). vRNAP specifically recognizes the template strand of extruded hairpin promoters to continue genome injection and transcription of early genes encoding the middle transcription machinery (13, 14). The N4-encoded heterodimeric T7-like N4 RNAPII, with transcription factor gp2, recognize AT-rich promoter sequences to complete genome injection and transcribe the middle genes (15–19), which encode the replicative functions (20–22). The N4 single-stranded DNA-binding protein (N4SSB), required for N4 genome replication, activates the host σ^{70}-RNAP at N4 late promoters, leading to synthesis of N4 late transcripts that encode morphogenetic components and proteins required for cell lysis (20–23).

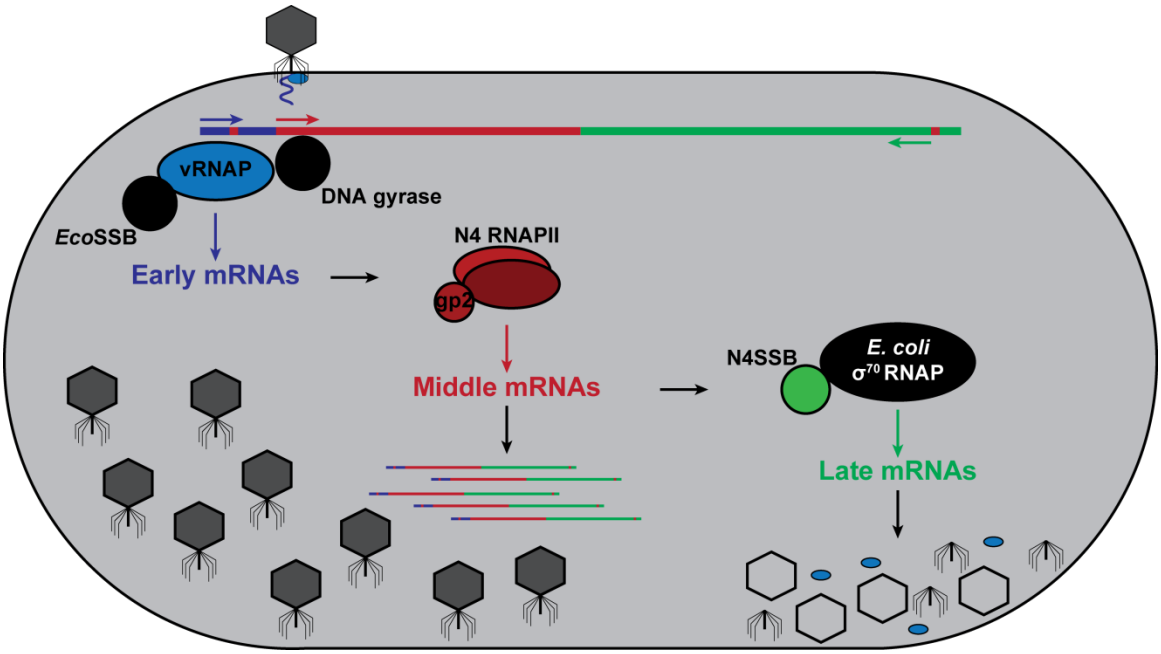

Figure I.1. N4 transcriptional program. Upon interaction with its receptor, NfrA, N4 injects vRNAP and genomic DNA into the host (4, 24). Host DNA gyrase introduces negative supercoils into the phage genome, driving early promoter hairpin extrusion (7, 10). *Eco*SSB stabilizes the open DNA conformation by binding the non-template strand, allowing vRNAP to initiate transcription from early promoters to transcribe early genes required for middle transcription (11, 12). The early gene product gp2 acts as a transcription factor for the heterodimeric polymerase N4 RNAPII (gp15 and gp16) to carry out middle transcription (15, 16, 18). These transcripts encode functions required for N4 DNA replication (20–22). One such protein, N4SSB, interacts with host σ^{70}-RNAP, redirecting it to N4 late promoters (21–23). The late genes encode morphogenetic proteins involved in virion assembly, DNA and vRNAP packaging, and host lysis. Host proteins depicted in black, phage-encoded early transcriptional machinery in blue, phage-encoded middle transcriptional machinery in shades of red, and phage-encoded late transcriptional machinery in green. Arrows indicate polarity of transcription (25).

This transcriptional architecture, unique to N4, is reversed from strategies commonly employed by other DNA phages whereby the host RNAP recognizes early promoters and is subsequently hijacked through antitermination (*e.g.*, λN, λQ, HKO22 Put RNA), sigma factor remodeling (*e.g.*, T4 AsiA-MotA), sigma factor replacement (*e.g.*, SPO1 gp28, gp34) (26–28), or the early synthesis of a phage-encoded RNAP (*e.g.*, T7 RNAP, Xp10 RNAP (29, 30). In contrast, N4's transcriptional strategy enables host-independent early and middle transcription while still

hijacking the host RNAP to divert resources toward N4 development late in the infection cycle and allows for the coupling of DNA replication to the synthesis of morphogenetic proteins (31).

Unlike most phages, N4 does not lyse the host until approximately 3 hrs after infection, leading to enlarged host cells harboring phage progeny at near-crystalline density and, upon release, a burst size of 3,000 plaque-forming units (PFUs) per infected cell (32). N4 neither degrades the host chromosome nor hinders host transcription and translation, allowing host cells to grow continuously, but employs three strategies to disrupt and co-opt other host functions. Host cell division and host DNA replication are inhibited by two early gene products, gp6 and gp8, respectively. Gp6, a cell division inhibitor, binds to both FtsZ and FtsA to disrupt the FtsZ ring at the midcell, while gp8, a host DNA replication inhibitor, binds to the β clamp loader to shut down host DNA synthesis (33, 34). Although not essential, deletion of either ORF6 or ORF8 leads to a significant decrease in burst size, indicating that inhibition of host DNA replication and cell division contribute to N4 fitness. The third contributor to large burst size is delayed lysis, which is accomplished through the properties of the N4-encoded N-acetylmuramidase that is localized to the inner membrane through its positively charged N-terminal signal sequence and is released from the inner membrane to adopt an active conformation (35).

OVERVIEW OF BACTERIOPHAGE GENOMICS AND TAXONOMY

The recognition that tailed dsDNA bacteriophages are the most abundant biological entities on Earth (36, 37) has raised interest in understanding the role that phages play in biogeochemical processes (38). This, along with the advances in sequencing technologies and increasing cost-effectiveness, has led to an explosion in sequencing viral species from

environmental and clinical samples (39–43). Despite the accumulation of DNA sequence data, bacteriophage genomics is still in its infancy. Full genome sequences are only available for approximately 5,000 tailed dsDNA phages (National Center for Biotechnology Information database as of 2018) out the 100 million bacteriophage species predicted to exist (36, 44).

The vast majority of bacteriophage genome sequences available to date were either individually isolated on bacterial hosts in the laboratory or identified through viral metagenomics analysis where DNA is concentrated and sequenced from complex environmental or clinical samples (44, 45). For over 100 years, new phage species have been primarily detected and isolated through culture-based methods using laboratory strains of bacteria. The isolation of bacteriophage species in culture remains the gold standard for microbiology and bacteriophage genomics since: i) bacteriophage genome assembly is substantially simplified when starting from purified, isogenic phage populations, and ii) the study of isolated phages provides the capability to perform downstream genetic, biochemical, and structural studies to test hypotheses generated from comparative genomics (45). However, isolation of phage species through culture-based methods is a slow process and limits the diversity of sequenced phages to species that infected hosts that could be easily cultured in the laboratory, which only represent approximately 1% of the extant bacterial diversity (46). Therefore, a comprehensive assessment of bacteriophage diversity throughout the biosphere has required the use of culture-independent methods of bacteriophage identification, made possible through the advent of next generation sequencing techniques (46, 47).

Identification of phage species from environmental samples and metagenomics datasets is difficult due to the lack of a universal marker gene, the presence of phages as extrachromosomal species, and the challenges in differentiating functional integrated prophage from bacterial

sequences and decayed or inactive prophage (42, 46, 48). Methods in phage identification from metagenomics datasets, which relies on DNA sequence alignment to known reference sequences (49), have been improved through enrichment of viral DNA from metagenomics samples and refinement of the databases and algorithms used for alignment (46, 49). Although techniques such as viral propagation in cultures and virus-like particle purification have been effective, these techniques are limited due to their inability to propagate phages with unidentified hosts and inability to capture giant dsDNA viruses and phages that do no form stable extracellular particles (49, 50). Numerous computational methods are also being developed to improve the efficiency and accuracy of sequence alignment and refine annotations within reference viral sequence databases. Although the majority of sequences are aligned using the Basic Local Alignment Search Tool (BLAST), new alignment algorithms using K-mer based classification (e.g. VirFinder and NBC) have recently been shown to improve the speed and fidelity of sequence alignment to viral sequences (51, 52). Furthermore, curated databases of viral sequences to search against (e.g. ACLAME, VirSorter, MetaPhinder) are currently in development to enable streamlined homology search and annotation of viral sequences (53–55). Recent studies have shown that VirSorter has been a useful tool to identify viral sequences in metagenomics reads in ocean samples, find prophages in microbial genomes, and characterize viral sequences with known hosts from bacterial and archaeal genomic datasets (43, 48, 53).

Upon identification of phage sequences, phylogenetic classifications through canonical methodologies make comparative genomics difficult due to the modular nature of phage evolution. The mosaicism of bacteriophage genomes was first appreciated through the study of regions of similarity in lambdoid genomes by DNA heteroduplex mapping (56, 57). These studies indicated that although lambdoid phage genomes largely share a common organization of

genetic elements, DNA sequence similarity was largely mosaic and contained rapid transitions between areas of high and low sequence similarity. Further analysis revealed that these variable regions were not randomly distributed throughout the genome, but rather occurred between individual or clusters of ORFs (58). Although it was originally hypothesized that phage mosaicism occurred through recombination at specialized intergenic sequences, subsequent studies of fully sequenced phage genomes have been unable to identify conserved linker sequences responsible for targeted recombination events (58–61). Therefore, mosaicism likely results from non-homologous recombination events randomly distributed throughout the genome that are subject to purifying selection to enrich for recombination events that do not disrupt essential genes and maintain genomes of similar architecture and size (60, 62). Through this process, incredibly diverse phage genomes may arise to create novel combinations of genes and protein domains. As a consequence, different genes or groups of genes contain separate evolutionary histories and will vary dramatically even between closely related phage species.

Classical definitions of hierarchical classification fail to capture the reticulate properties of phage genomes, relationships that have recently been more accurately characterized through networks of shared protein families in various datasets (59, 62, 63). Canonical viral taxonomy, defined by the International Committee on Taxonomy of Viruses, has relied on genome type (dsDNA, ssDNA, etc.) and the morphologies of phage capsids and tails to determine evolutionary relatedness. Subsequent studies have shown that morphological similarity does not correlate with genetic similarity, suggesting that a new method of viral classification was necessary (59, 63, 64). Therefore, a new methodological framework for the taxonomic assessment of bacteriophages relying on shared gene cassettes across the entire genome was proposed where pairwise relationships of phage genomes are characterized as a weighted

network of shared gene content to create a reticulate phylogeny of phages (59). This methodological framework was first used to create evolutionary cohesive modules of 306 fully sequenced phage genomes as a proof of principle and has since been used to determine the evolutionary relationships of newly identified phages such as ΦJM-2012 (63, 65).

As of 2017, half of the approximately 2,000 sequenced phage genomes in the NCBI database were derived from only four bacterial genera (*Mycobacterium*, *Enterobacteria*, *Pseudomonas*, and *Staphylococcus*) (46). Even among this small sample of bacteriophage genomes, comparative genomics analyses have revealed tremendous diversity in gene content, genome organization, and inferred diversity of regulatory and life cycle strategies among the tailed phage population, even among presumably closely related phages capable of infecting the same hosts (44, 66, 67). The most comprehensive comparative genomics studies to ascertain the genetic diversity of phages that infect the same host have been conducted on the mycobacteriophages (phages that infect *Mycobacterium smegmatis*). Currently, over 850 different mycobacteriophages have been identified and a panel of 627 have recently been subjected to comparative genomics analyses (67). These phages separate into 20 clusters by nucleotide similarity with considerable variation in cluster size: 50% of mycobacteriophages are grouped within two large clusters (67). Although phages within each cluster share higher degrees of sequence similarity, considerable gene flow is observed between clusters of phages, suggesting that the mycobacteriophages all have access to a shared gene pool available for horizontal gene transfer. Surprisingly, rarefaction analyses of the mycobacteriophages show that this is not a closed system: there is an influx of novel gene sequences to the population of mycobacteriophages (44, 67). These results suggest that gene flow is rampant between phages

infecting the same host and that a continuum of diversity can be established within a population by the rapid rate of change for host specificity by phages in natural populations (44).

Despite the recent increase in fully-sequenced phage genomes available for analysis, the biological meaning that can be inferred has been limited due to the large proportion (60-95%) of phage sequences that lack homology to known sequences within the current databases (42, 47, 68, 69). These ORFs of unknown function, the "dark matter of the biosphere", provide an opportunity for the discovery of novel genetic sequences and new biological mechanisms (44, 48, 69). The analysis of gene conservation through families of related phages has suggested that a subset of genes, frequently including those encoding morphological proteins, are more likely evolutionarily conserved due to strong co-evolutionary forces maintaining these interactions (45, 70). Unlike these so-called "core genes," the vast majority of bacteriophage genes are "non-core" or "auxiliary" genes that are often small in size, poorly conserved, and non-essential for phage viability. These genes are highly mobile and may represent a "gene nursery" where novel combinations of genes and protein domains are created without purifying selection (45, 69). Auxiliary genes are frequently picked up by phages from the host genome through non-homologous recombination or sloppy excision of prophages and are retained in the phage genome if they provide a fitness advantage that allows the phage to occupy a new ecological niche (45, 71).

Indeed, comparative genomics studies have characterized numerous auxiliary metabolic genes in marine phages that augment the metabolism of host species in nutrient poor conditions. For example, cyanophages residing in regions of low phosphate concentration frequently contain the *phoA* gene encoding an alkaline phosphatase, while some *Prochlorococcus* phage isolates encode a gene involved in biosynthesis of photosynthetic pigments (72–74). Additionally, phage

auxiliary genes have been shown to provide evolutionary advantages by encoding proteins responsible for evasion of host immune systems. Surprisingly, an active phage-encoded CRISPR/Cas system has been identified that counteracts a phage inhibitory chromosomal island to enable phage infection of *Vibrio cholera* hosts with an active immune system (75), while other phages have also been shown to encode proteins (termed anti-CRISPRs) that inhibitor host CRISPR-Cas systems (76). Continued research into bacteriophage genomes are likely to reveal additional genes of interest, as a significant portion of the viral dark matter is yet to be discovered and characterized.

REVIEW OF T7-LIKE RNA POLYMERASES AND THEIR TRANSCRIPTION FACTORS

Evolutionary relationships of polymerases

The efficient and accurate synthesis of DNA and RNA is essential for all living organisms. The nucleic acid polymerase superfamily of enzymes responsible for these processes are generally grouped by their specificity for nucleic acid templates and products, then are further subdivided into protein families by sequence homology (77). The vast majority of polymerases within the superfamily are single-subunit enzymes that share a conserved polymerase domain structure and mechanism of catalysis despite differing template and substrate nucleotide preferences (77–80). Structural, genetic, biochemical, and sequence data have revealed that single-subunit polymerases all share a common polymerase domain structure resembling a cupped right hand. The thumb and fingers subdomains flank the palm subdomain at the base of the catalytic cleft to create the "polymerase fold," characteristic of these enzymes (78–81). Primary sequence alignment of single-subunit polymerases identified two motifs conserved across the entire superfamily. These motifs, termed motifs A and C, both contain

invariant aspartates that coordinate magnesium ions for catalysis and are located in the catalytic cleft of the palm subdomain (82, 83). A third motif, motif B, was found only in DNA-directed polymerases. Motif B is located in the fingers subdomain and includes an invariant lysine residue and is involved in binding the NTP substrate (82, 84). Additionally, the T/DxxGR motif is commonly found in DNA-directed, but not RNA-directed RNAPs (80). This motif is located in the palm subdomain and plays a role in stabilizing the RNA:DNA hybrid during transcription initiation of single-subunit RNAPs (85).

The two N4-encoded RNAPs (N4 vRNAP and N4 RNAPII) belong to the T7-like (or "single-subunit") RNAP family, which is comprised of three distinct groups: mitochondrial and chloroplast RNAPs, linear plasmid-encoded RNAPs, and bacteriophage RNAPs (86, 87). These enzymes have evolved from a single common ancestor, likely via duplication and divergence of a DNA polymerase or reverse transcriptase, and are evolutionarily distinct from the multi-subunit RNAPs of bacteria, archaea, and eukaryotes (86, 88). Surprisingly, phylogenetic analysis of the catalytic domain of N4 vRNAP (mini-vRNAP) and N4 RNAPII show that these polymerases share greatest sequence similarity with the linear plasmid-encoded enzymes and are highly divergent members of the T7-like RNAP family (87). Overall, T7-like RNAPs contain all conserved structural motifs outlined above and a total of 12 conserved blocks of sequence motifs (86, 89). T7-like RNAPs display a higher degree of sequence and structural conservation within their C-terminal regions, which contain motifs responsible for template binding and catalysis, but display considerable length and sequence variation within the N-terminal regions responsible for promoter recognition (86). These results suggest that the differences within the N-terminal domains may be responsible for differing promoter sequence preference and transcription factor dependence among T7-like RNAPs.

T7 RNAP

The T7 RNAP is a 98 kDa single-polypeptide RNAP capable of promoter recognition, promoter melting, transcription initiation, and transition to the elongation phase without additional factors (90). The simplicity of T7 RNAP relative to the much larger multi-subunit RNAPs has made T7 RNAP an attractive model system for structure-function studies of RNAPs and an ideal tool for biotechnological purposes such as *in vitro* RNA synthesis and protein expression in bacterial systems (91, 92).

During bacteriophage T7 infection, T7 proteins are synthesized in three temporal classes: class I (early), class II (middle), and class III (late) (29). Class I proteins are synthesized by the host σ^{70}-RNAP initiating transcription from four early promoters (29, 93). T7 achieves transcriptional independence from the host through the activity of the product of gene 1, which encodes a single-subunit RNAP (T7 RNAP) responsible for transcription of T7 middle and late genes and the product of gene 2 (gp2), which inhibits σ^{70}-RNAP transcription initiation to prevent unregulated early transcription during the late stages of infection (29, 90, 94–96). Synthesis of the class II product T7 lysozyme is responsible for shutting off T7 RNAP transcription of class II genes. T7 lysozyme binds to T7 RNAP directly and acts as an allosteric inhibitor to prevent the transition to the elongation phase, facilitating the switch from inefficient class II promoter to strong class III promoter expression during the late stage of infection (84, 97, 98).

The crystal structure of T7 RNAP provided the first evidence that single-subunit RNAPs were structurally related to the nucleic acid polymerase superfamily of enzymes despite the lack of sequence homology (82, 99). Structural comparisons with the Klenow fragment of DNAP I and HIV-1 reverse transcriptase revealed a conserved polymerase core domain that resembles a

cupped right hand with a deep DNA-binding cleft surrounded by the thumb, palm, and fingers subdomains (81, 99, 100). The structural homology was limited to the T7 RNAP C-terminal polymerase domain (residues 326-883), as T7 RNAP has an N-terminal domain (residues 1-325) appended opposite the palm subdomain that closes off the catalytic cleft. Early biochemical and mutational analyses showed that disruptions to the N-terminal domain reduce RNA binding and processivity, suggesting that this domain may interact with promoter DNA and nascent RNA in processes specific for RNAPs (84, 99, 101, 102).

The structure of T7 RNAP in complex with the transcriptional inhibitor T7 lysozyme was determined to significantly higher resolution and allowed for a more detailed view of the orientation of T7 RNAP subdomains (84). The palm subdomain (residues 412-553 and 785-879), located at the base of the DNA-binding cleft, is comprised of three β-sheets containing the conserved motifs A and C. These motifs contain invariant aspartate residues (D537 and D812) required for the coordination of the catalytic Mg^{2+} ions to catalyze the phosphoryl transfer reaction (83, 84, 103–105). The palm subdomain also contains the DxxGR motif (residues 421-425), which interacts with the 3' end of the nascent transcript at the active site to stabilize the RNA:DNA hybrid during transcription initiation (85). The thumb subdomain (residues 326-411), located to the right of the catalytic cleft, is required for processive elongation. Substitutions to or deletion of this region led to decreased processivity and reduced stability of T7 RNAP ternary complexes (106–108). The fingers subdomain (residues 554-784) is located to the left of the catalytic cleft and includes motif B (residues 627-640), responsible for binding to the incoming NTP substrate and T7 RNAP translocation (103, 109, 110). Apart from the previously mentioned N-terminus, T7 RNAP contains several insertions that differentiate the enzyme from the pol I family of polymerases. The T7 RNAP contains an additional palm insertion module (residues

450-527) between the palm and thumb subdomains, while the fingers subdomain contains an extended foot module (residues 838-883) (84). A third insertion, the specificity loop (residues 739-770), is a flexible β-hairpin within the fingers subdomain that extends into the catalytic cleft and has been shown to be involved in promoter recognition (111, 112).

T7 RNAP recognizes 17 distinct promoter sequences in the T7 genome. These promoter sequences consist of a 23 bp consensus sequence that extends from -17 to +6 (relative to the site of transcription initiation at +1) (93). Class III promoters match the consensus sequence, while class II promoters are weaker and differ from the consensus sequence at two or more positions (93). The promoter sequence comprises two functional regions: the upstream binding region spanning -17 to -5 and the initiation region spanning -4 to +6. Nucleotide substitutions in the upstream binding region reduce polymerase affinity, while substitutions to the downstream region reduce the efficiency of transcription initiation (113–116). Promoter sequences in bacteriophage T7-like RNAPs (e.g. T7, T3, SP6, and K11) have a high degree of sequence conservation from -7 to -3, but differ significantly from -12 to -8 (117, 118). This implies that the -7 to -3 core region has a conserved function in related phages, while the upstream region confers species-specific recognition (119).

The crystal structure of T7 RNAP in complex with a 17 bp promoter provided a definitive model for sequence-specific promoter recognition for this enzyme (see Figure IV.1) (120). In this structure, 13 bp of upstream duplex DNA (-17 to -5) makes contact with the T7 RNAP N-terminus, while the two DNA strands are melted downstream of the -4 position and the template strand is directed towards the T7 RNAP catalytic cleft. T7 RNAP promoter recognition involves three structural elements: the AT-rich recognition loop, β-intercalating hairpin (β-IH), and the specificity loop. The AT-rich recognition loop (residues 93-101) located in the T7 RNAP

N-terminal domain forms a flexible loop that inserts into the minor groove of the AT-rich sequence between -17 to -13 positions of T7 RNAP promoters. Insertion of the AT-rich recognition loop distorts the DNA helix by widening the minor groove and bending the upstream DNA (120). Specific recognition of promoter sequences is achieved solely through the insertion of the antiparallel β-hairpin specificity loop (residues 739-770) into the major groove from -7 to -11 (116, 120–122). Specific residues on one face of the specificity loop interact with individual bases in both the template and non-template strand of promoters. The β-IH (residues 227-246) in the T7 RNAP N-terminus is required for promoter melting and maintenance of the upstream edge of the transcription bubble during elongation. The V237 residue intercalates between the -5 and -4 bases, inducing promoter melting from -4 to +3 through base stacking interactions with the -5 base and redirection of the template strand towards the active site (120, 123–125). The open DNA conformation is stabilized through phosphodiester backbone contacts between the specificity loop and the template strand.

The orientation of nascent RNA products within the a transcribing T7 RNAP was observed in the structure of T7 RNAP bound to a duplex promoter transcribing a trinucleotide RNA product (109). In this structure, T7 RNAP maintains contacts with promoter DNA while the A-form DNA:RNA heteroduplex accumulates in the active site to position the +4 template base in the active site to interact with the incoming NTP. The Y639 residue is responsible for the preference for transcription initiation with GTP and stacks against the template base at the NTP-binding site (N-site) (109, 126, 127), while the R425 residue in the DxxGR motif hydrogen bonds to the 2’-OH group of rNTPs in the minor groove of the A-form DNA:RNA hybrid to stabilize the nascent transcript (80, 85, 109).

During transcription initiation, T7 RNAP goes through multiple rounds of abortive initiation, where small (2-8 nt) RNAs are synthesized and subsequently released (128). Once an RNA product of eight or more nucleotides has been synthesized, the enzyme transitions into the elongation conformation, where transcription proceeds with increased processivity (128). During abortive initiation, T7 RNAP maintains contacts with the promoter sequence; the upstream T7 RNAP footprint is maintained (129) and the template DNA accumulates in the active site in a process called "DNA scrunching"(109). DNA scrunching and rotation during nucleotide addition destabilizes the T7 RNAP initiation complex through the accumulation of steric and electrostatic strain (130). The resolution of these high energy intermediates through promoter release or abortive synthesis is determined by a competition between the energetic cost of breaking contacts with upstream promoter DNA and peeling apart the RNA:DNA heteroduplex (130, 131). The energetic cost of breaking these upstream contacts is directly responsible for release of nascent RNA during abortive initiation, as demonstrated by the fact that the efficiency of promoter clearance is inversely correlated with T7 RNAP affinity for promoter sequences (132, 133). In contrast, the base-pairing energy of longer RNA:DNA hybrids (longer than seven base pairs) favors promoter clearance, which involves disruption of contacts between the specificity loop and AT-rich recognition loop with promoter DNA to facilitate the transition to the elongation complex (134, 135).

Upon transition to the elongation complex, T7 RNAP undergoes drastic conformational changes in the N-terminal domain, which rearranges to break all sequence-specific contacts with upstream DNA and forms a channel to stabilize the RNA:DNA heteroduplex (136, 137). Upstream DNA migrates from its location in the initiation complex to form non-specific interactions with a cleft between the helix C2 and thumb, and allow the formation of a seven

base pair RNA:DNA heteroduplex (134). In addition to the N-terminal rearrangements, the T7 RNAP thumb subdomain rotates to create a binding pocket for the non-template strand DNA, while the specificity loop rotates to form the RNA exit tunnel (106–108, 138, 139). The refolded subdomain H and the specificity loop form a positively charged tunnel for RNA exit and make non-specific contacts through the phosphodiester backbone to enhance the stability of the elongation complex and increase T7 RNAP processivity (102, 138).

Mitochondrial RNA polymerases

Transcription of the mitochondrial genomes in yeast and humans is performed by Rpo41 and POLRMT, respectively; these nuclear-encoded polymerases share sequence similarity to the T7/T3 class of RNAPs (88, 140–142). The mitochondrial RNAPs (mtRNAPs) consist of three domains: the C-terminal domain, the N-terminal domain, and the N-terminal extension. The C- and N-terminal domains share sequence and structural similarity to the corresponding domains in T7 RNAP, while the N-terminal extension is absent in the bacteriophage cluster of T7-like RNAPs and is unique to mtRNAPs (86, 141–145). The N-terminal extensions in Rpo41 and POLRMT are required for catalysis *in vitro* and contain an N-terminal targeting signal that is cleaved upon mitochondrial import (144, 146). In contrast to the well-characterized bacteriophage T7-like RNAPs, the mtRNAPs are unable to catalyze all the steps necessary for transcription without additional protein factors (147–149). Since mtRNAPs serve crucial functions in the transcription of required mitochondrial genes and maintenance of mitochondrial genomes, numerous studies have been undertaken to understand the biochemical and structural properties of the yeast and human mtRNAPs and the role of their required accessory factors in transcription regulation.

Rpo41

The yeast mitochondrial genome encodes two rRNAs, 25 tRNAs, and eight proteins that, together with additional nuclear-encoded proteins, comprise the oxidative phosphorylation and electron transport chain complexes (150, 151). Evaluation of the 5' end of yeast mitochondrial transcripts has revealed 14 promoters actively utilized by Rpo41 *in vivo*, although as many as 28 yeast promoters have been reported to be active *in vitro* (143, 152, 153). Calculation of the yeast mitochondrial consensus promoter sequence revealed a relatively small nine base pair promoter sequence represented by the TATATTCAT (-8 to +1) template-strand sequence (152, 154, 155). Interestingly, five exact matches of the nonanucleotide consensus promoter were found elsewhere in the yeast mitochondrial genome, however, no evidence of transcription was found for these loci (152). Yeast promoters have a wide variation in activity; strong promoters with purine residues at the +2 position of the non-template strand direct transcription with 15-20 times higher activity than weak promoters with pyrimidine residues at +2 (156–158). Substitutions to the conserved template-strand thymidine at +1 are tolerated without dramatically reducing transcription. This is uncommon among T7-like RNAPs, which have a strong preference for incorporation of purines as the initiating nucleotide (159). Substitutions at all upstream bases reduce Rpo41 transcription. Substitutions at -3, -1, +1, and +2 positions could be alleviated by initiation from dinucleotide primers, indicating that these positions are required for transcription initiation of yeast mitochondrial promoters (154, 155, 157).

The yeast mtRNAP, Rpo41, was the first single-subunit mtRNAP identified. The RPO41 nuclear gene was shown to encode a ~150 kDa single-subunit RNAP with extensive homology to members of the T7-like RNAP family (140, 141). The high degree of C-terminal sequence similarity between T7 RNAP and Rpo41 suggests that these two enzymes have structural

homology and share conserved mechanisms of catalysis, while the lack of N-terminal homology and dependence on the additional factor Mtf1 suggests that Rpo41 has alternative mechanisms for promoter recognition (145). Homology modeling and cryo-EM reconstructions have confirmed that Rpo41 exhibits the characteristic cupped right hand architecture and identified Rpo41 structural analogs of the AT-rich recognition loop, β-IH, and specificity loop elements required for promoter recognition in T7 RNAP (109, 145, 160, 161). The Rpo41 specificity loop (amino acids 1127-1149) was identified and shown to directly contact the yeast mitochondrial promoters between -1 and -8 and substitutions within the specificity loop drastically reduce promoter specificity. These results confirm previous reports that Rpo41 contained the determinants of promoter specificity in the core enzyme and suggest that specificity loop interaction with upstream promoter sequences may be a conserved mechanism of promoter recognition for T7-like RNAPs (145, 162).

Although Rpo41 is capable of nonspecific transcription from dsDNA and specific transcription from pre-melted bubbled templates, the additional factor Mtf1 is required for specific transcription initiation from dsDNA promoters *in vivo* and *in vitro*, suggesting that Mtf1 is required for promoter melting (162–164). Indeed, Rpo41 and Mtf1 bind to dsDNA with very high affinity at 1:1 stoichiometric ratio and are required for melting promoters from -4 to +2 (165), but Mtf1 is released upon transition to the elongation phase (166).

The crystal structure of Mtf1 revealed two domains separated by a basic cleft at the interface: an N-terminal domain with an extended α/β-structure and a smaller C-terminal domain composed of four helices and a flexible C-terminal tail (167). The N-terminal domain contains the S-adenosyl-L-methionine (SAM)-binding site characteristic of this family of

methyltransferases, although SAM is not essential for binding to Rpo41 or for transcriptional activation (167).

Rpo41 and Mtf1 form a 1:1 stoichiometric complex both in the presence and absence of promoter DNA (146, 166). In the absence of a crystal structure of Rpo41 in complex with Mtf1, deletion mutagenesis, crosslinking, and genetic screening methods were used to map the site of interaction between these two proteins. Deletions of the Rpo41 N-terminal extension suggested that residues 270-380 are required for Mtf1 interaction, while deletion of the Mtf1 C-terminal tail inhibits transcription activation and binding to Rpo41 (146, 168). Genetic screens identified multiple point mutants throughout Mtf1 that disrupt interaction with Rpo41, while complementary studies showed that residues within the Rpo41 specificity loop, β-IH, and extended foot module are involved in Mtf1 interaction (168, 169). These studies suggest that interaction between Mtf1 and Rpo41 promoter recognition elements may be required to reorient these structures into the proper conformation for interaction with promoter DNA. Indeed, low resolution cryo-EM studies of the Rpo41-Mtf1 initiation complex show that Mtf1 induces conformational changes within Rpo41 to facilitate formation of the open complex (161). In the absence of Mtf1, Rpo41 binds to dsDNA non-specifically and assumes a clenched conformation as observed in the POLRMT apo structure. Binding of Mtf1 to the Rpo41-DNA complex widens the Rpo41 catalytic cleft and reorients upstream duplex DNA in close proximity to the AT-rich recognition loop (161).

Through a series of crosslinking studies, Mtf1 has been shown to make extensive contacts with DNA throughout the promoter sequence. Mtf1 crosslinks to the melted -2 to -4 non-template strand bases and makes additional contacts with -8 to -10 bases within duplex DNA, which suggests that Mtf1 is directly involved in the formation and stabilization of the open

complex (160). Additional contacts between the Mtf1 C-terminal tail and template-strand bases at -3 and -4 suggest that this domain localizes within close proximity to the Rpo41 active site (170). Rpo41 itself crosslinks to template-strand bases in the melted promoter, confirming that Rpo41 does contribute sequence recognition, possibly through the specificity loop (145, 160, 162).

The mechanism of Rpo41 promoter recognition is vastly different than the mechanism utilized by T7 RNAP. T7 RNAP discriminates promoter from non-promoter DNA sequences through differential affinity at the initial binding step. T7 RNAP binds to promoter DNA with 10^5-fold greater affinity than to non-promoter DNA (123). In contrast, Rpo41 binds to promoter and non-promoter DNA with similar affinities. Although Mtf1 does not bind dsDNA on its own, the Rpo41-Mtf1 complex binds dsDNA with increased affinity, but only distinguishes promoter from non-promoter DNA with a 3-6-fold difference in affinity (171). Therefore, the Rpo41-Mtf1 complex does not utilize differential affinity for promoter selection, but utilizes an induced fit mechanism where transcription initiation is regulated by sequence-dependent bending and melting of promoter DNA (123, 143, 172). Although Rpo41 is capable of non-specific DNA bending and melting on its own, the polymerase forms dynamic complexes with a fast promoter closing rate constant (171, 173). The Rpo41-Mtf1 complex severely bends promoter DNA in a sequence-specific fashion to a 90° angle, which is required for DNA melting (171). Furthermore, Mtf1 significantly decreases the Rpo41-DNA off rate and the rate constant for promoter closing; resulting in a 500-fold increase in the rate constant of promoter opening (171, 173). These results confirm that Mtf1 aids Rpo41 promoter unwinding by lowering the activation energy of melting DNA bps and stabilizes the open conformation of promoter DNA (143).

In many RNAPs, stabilization of the open complex and transcription initiation are regulated by binding to the initiating nucleotides. Studies of the effect of nucleotides at the +1 and +2 positions of yeast mitochondrial promoters suggests that the Rpo41-Mtf1 complex has greater affinity for the +1 NTP than the +2 NTP and that these nucleotides are required for stabilization of the open complex (158, 174). Interestingly, Mtf1 has been shown to increase the affinity for the +2 NTP, suggesting that Mtf1 may contribute to promoter specificity directly and increase the catalytic efficiency of transcription initiation (158, 174, 175). These results explain the fact that strong promoters incorporate ATP as the +2 NTP and suggest that yeast mtRNAP transcription acts as an *in vivo* ATP sensor, regulating the rate of transcription from different mitochondrial promoters in response to shifting concentrations of intracellular ATP (176).

POLRMT

The 16.6 kbp circular dsDNA human mitochondrial genome encodes 22 tRNAs, two rRNAs, and 13 proteins comprising key components of the electron transport chain (177–179). Transcription of the human mitochondrial genome is performed by the human mitochondrial RNA polymerase (POLRMT) and its required transcription factors mitochondrial transcription factor B2 (TFB2M) and mitochondrial transcription factor A (TFAM) (142, 148, 149).

Transcription initiates from three mitochondrial promoters clustered together in the displacement loop (D-loop) noncoding region of human mitochondrial DNA: the light-strand promoter (LSP) and heavy-strand promoters 1 and 2 (HSP1 and HSP2) (180, 181). Transcription from each promoter produces a long, polycistronic RNA molecule that is processed to produce all mRNA, rRNA, and tRNAs required for mitochondrial function (178, 181–183). LSP, HSP1, and HSP2 promoters are contained within a DNA fragment spanning -28 to +16 (180, 181). Alignment of these three promoters show that they all contain polyG tracks in the template

strand immediately upstream of the transcription start site and initiate transcription with ATP (180, 184, 185). Substitutions to nucleotides within the polyG tracks significantly reduce transcription *in vitro* (185). Further *in vitro* characterization of human mitochondrial promoters identified two regions required for transcription initiation: the promoter distal (-17 to -20) and promoter proximal (-1 to -4) regions. Substitutions to the promoter distal region, which shares sequence conservation between human and mouse mtRNAP promoters, reduce TFAM function, while substitutions to the promoter proximal region reduce transcription in a POLRMT-dependent fashion. These results suggest that sequence-specific binding of TFAM upstream of HSPs and LSP is required to introduce architectural changes to promoter DNA and partially explains why the distance between TFAM binding and the transcription start site is crucial for transcription initiation (186, 187). POLRMT was unable to recognize promoters with mouse mtRNAP promoter sequences from -1 to -4, suggesting that POLRMT makes sequence-specific contacts with the polyG sequences within the promoter at these positions (187).

In vivo, transcription from HSP1 is greater than transcription from LSP, while transcription initiating from HSP2 is significantly lower (184). HSP2 promoters do contain the template strand polyG track characteristic of other human mitochondrial promoters and substitutions within this region also significantly reduce transcriptional activity (184). However, this polyG track is located further upstream from the transcription start site in HSP2, perhaps accounting for the relatively poor activity from this promoter relative to HSP1 and LSP. Recent evidence has shown that TFAM is an important modulator of mitochondrial transcription *in vivo* (188). Low concentrations of TFAM promote expression of mRNAs encoding components of the electron transport chain, moderate concentrations of TFAM promote the synthesis of RNAs required for mitochondrial biogenesis and mtDNA replication, while very high concentrations of

TFAM reduce transcription and mtDNA replication (188). Furthermore, the relative utilization of human mitochondrial promoters is also controlled by intracellular ATP concentrations (184, 189, 190). Increasing ATP concentrations inhibited HSP1 and LSP transcription, while HSP2 transcription became favored (184). Therefore, intracellular ATP concentration provides another mechanism for differential regulation of human mitochondrial promoters to down-regulate mitochondrial translation under conditions where increased electron transport chain activity is not needed.

The polymerase responsible for transcription of the human mitochondrial DNA, POLRMT, is a 1,230 aa (140 kDa) protein that belongs to the T7-like family of RNAPs (142). The apo structure of POLRMT confirms that POLRMT shares strong structural homology to T7 RNAP and provides insight into the POLRMT dependence on transcription factors for promoter recognition (144). Overall, POLRMT displays the cupped right hand architecture characteristic of this family of polymerases and shares significant structural similarity with T7 RNAP throughout their C-terminal domains, suggesting that these enzymes share similar mechanisms of RNA catalysis (77, 144, 177). Despite limited sequence similarity, the POLRMT N-terminal domain shares structural similarity to the T7 RNAP N-terminal domain and contains structural analogs of the AT-rich recognition loop and β-IH promoter binding elements. In POLRMT, however, these elements are poorly positioned to bind promoter DNA (144). Furthermore, POLRMT is in a partially closed ("clenched") conformation due to a rotation of the fingers domain that blocks ssDNA access to the active site. (144). The poor positioning of the AT-rich recognition loop and clenched conformation is also observed in the Rpo41 cryo-EM structure, which may explain both enzymes' requirement for additional transcription factors for transcription initiation from dsDNA promoters (144, 161).

In contrast to T7 RNAP, which undergoes extensive conformational changes upon the transition between the transcription initiation and elongation complexes, POLRMT adopts an intermediary elongation conformation that does not require significant refolding (137, 191). POLRMT transition to the elongation phase is dependent on dissociation of transcription factors TFAM and TFB2M (191) rather than large conformational changes to the promoter binding domain as observed in T7 RNAP (137). These data suggest that the mechanisms of POLRMT promoter recognition and transcription initiation are unique (191).

In vivo, POLRMT is incapable of interacting with promoter DNA and initiating transcription without its essential factors TFAM and TFB2M. TFAM, a member of the high-mobility-group-box (HMG-box) family of DNA-binding proteins, contains two DNA-binding domains (HMG box A and B), a 27 aa linker, and a 25 aa C-terminal tail (192–194). Like other HMG-box proteins, TFAM binds to specific DNA sequences with its two HMG-box domains to induce DNA bending. At high concentrations, TFAM binds to DNA nonspecifically, inducing less dramatic bends in the DNA, but resulting in genome packaging into nucleoids (195–197). The crystal structures of TFAM in complex with LSP templates revealed that each TFAM HMG-box domain induces a 90° bend, resulting in 180° wrapping of DNA upstream of POLRMT promoters and positioning the C-terminal tail near the transcription initiation site (198, 199). The C-terminal tail is responsible for both differential bending of promoter and non-promoter DNA and interacts directly with the N-terminal extension of POLRMT (195, 200–202).

POLRMT is capable of initiating transcription from bubbled templates, but requires the yeast Mtf1 homolog TFB2M for transcription initiation from fully double-stranded templates, suggesting that TFB2M is required for promoter melting (203–205). DNA footprinting, Förster resonance energy transfer, and 2-aminopurine assays have shown that TFB2M binds to promoter

DNA at the transcription start site overlapping the POLRMT footprint and that TFB2M interaction with POLRMT is required for promoter melting from -4 to +1, which is the rate limiting step in transcription initiation (203, 204, 206). Additional crosslinking studies have shown that the TFB2M N-terminus makes extensive contacts with template-strand DNA, while the TFB2M C-terminus makes direct contacts to the POLRMT B-loop motif (residues 588-604), which is present in mitochondrial but not in bacteriophage RNAPs (201, 204). Furthermore, TFB2M was shown to crosslink to the initiating nucleotide in catalytic autolabeling reactions (204). These data suggest that POLRMT-TFB2M interactions are required for promoter recognition and that TFB2M acts to coordinate the reactants in the POLRMT active site to facilitate first phosphodiester bond formation and transcription initiation.

The recently solved crystal structures of human mitochondrial transcription initiation complexes have elucidated the roles of TFAM and TFB2M in POLRMT promoter recognition (207). These structures show that TFAM binding 16-39 nt upstream of the transcription start site induces a sharp bend to the dsDNA (198, 199, 207) and confirms crosslinking data suggesting that TFAM directly recruits POLRMT to promoters (200–202). TFAM anchors POLRMT to the promoter through interactions between its distal HMG Box domain and the newly observed POLRMT "tether" helix, while the TFAM C-terminal tail interacts with the POLRMT PPR domain and D-helix to position the polymerase active site (207). Upon POLRMT recruitment, interaction between the TFB2M C-terminus and POLRMT induces a rotation of the POLRMT promoter binding domain to reposition the β-IH between DNA strands for promoter opening. Furthermore, TFB2M aids in the stabilization of the open complex by stabilizing the β-IH and by trapping the non-template strand of melted promoters through interaction with a positively charged surface in a manner that is topologically related to that employed by sigma factors in

multi-subunit RNAP initiation (207). Surprisingly, few contacts were observed between POLRMT and promoter DNA. The specificity loop appears to localize to the major groove of upstream DNA between -7 and -9; limited density observed in the crystal structure and a lack of sequence conservation among the three promoters at these positions suggest that these contacts only minimally contribute to promoter recognition (207).

In summary, these studies indicate that TFAM acts as an architectural transcription factor to introduce site-specific DNA bending required for promoter melting. TFAM also directly recruits the POLRMT-TFB2M complex to activated promoters and relieves the autoinhibitory effect of the POLRMT N-terminus (146, 207). TFB2M stimulates promoter melting from by inducing conformational changes to POLRMT to reposition the β-IH between upstream DNA strands and sequestering the non-template strand of open promoters (207). TFB2M is also present at the site of catalysis; crosslinking to the initiating nucleotide in catalytic autolabeling reactions and the template strand at positions +1 to +3 (203, 204, 206). This model clearly demonstrates that TFAM recruitment of POLRMT to mitochondrial promoters compensates for the lack of sequence-specific interactions upstream of the site of DNA opening and that TFB2M assists in promoter opening by positioning key structural elements in mtRNAP (208). Additional evidence suggests that TFAM may have functions downstream of initiation complex formation, although further studies are required to study these interactions (206).

N4-encoded RNA polymerases

Phylogenetic analyses have shown that both N4-encoded RNAPs are highly divergent members of the T7-like RNAP family, sharing greatest sequence similarity to the linear plasmid-encoded group (87). Like other T7-like RNAPs outside of the bacteriophage cluster, vRNAP and N4 RNAPII are incapable of initiating transcription on linear, dsDNA without the activity of

additional protein factors. Numerous genetic, biochemical, and structural studies of these N4-encoded RNAPs have been performed to elucidate the biochemical properties of these enzymes and understand the diversity of mechanisms of transcription initiation utilized by T7-like RNAPs and their transcription factors.

N4 vRNAP

Unlike most phages, where the host RNAP holoenzyme is responsible for early transcription, N4 establishes transcriptional independence of the host immediately upon infection through the injection of a virion-encapsidated RNAP responsible for transcription of the N4 early genes (24, 209, 210). The presence of a virion-encapsidated RNAP was first postulated upon the observation of a burst of RNA synthesis immediately after N4 infection under conditions where the host RNAP and host protein synthesis are both inhibited (209, 211) and confirmed through the purification of a 320 kDa polypeptide with RNAP activity from disrupted N4 virions (24, 210).

Limited trypsin digestion of the full-length vRNAP polypeptide (3,500 aa, 320 kDa) revealed a 1,106 aa domain (residues 998-2,103) that possesses the same transcription initiation, elongation, and termination properties as the full-length protein (87). A BLAST search with the minimally active transcriptional domain of vRNAP (mini-vRNAP) showed that this protein is a highly diverged member of the T7-like RNAP family with little sequence homology outside of the conserved sequence blocks that have been implicated in catalysis (86, 87). Mini-vRNAP contains sequence homologs of motifs A, B, and C, along with the DxxGR (TxxGR in mini-vRNAP) motif and their catalytic functions were confirmed in mini-vRNAP through mutational analyses (87).

In vitro analysis of this large polypeptide showed that vRNAP is inactive on native, double-stranded N4 DNA, but could initiate transcription only on denatured N4 DNA with high efficiency and *in vivo* specificity (7, 24, 210). Host DNA gyrase, which introduces negative superhelical turns to circular DNA, was shown to be required for N4 vRNAP transcription *in vivo*, indicating that negative supercoiling is required for vRNAP transcription initiation (7, 212).

Mapping the sites of vRNAP transcription initiation *in vitro* and *in vivo* showed that vRNAP recognizes three promoter sequences (P1, P2, and P3) located within the leftmost 10% of the genome (25, 213, 214). N4 early promoters share sequence conservation spanning -17 to +1, containing a 3' AA/GG sequence centered at -11 and flanked by 5 nt inverted repeats that form a highly stable 3 base loop, 5 bp stem hairpin (214). Conserved bases within promoter hairpin stems are required for both vRNAP binding and transcription initiation, while the identity of the non-conserved bases is interchangeable as long as they maintain the promoter hairpin (9). These observations suggest that hairpin formation and direct contacts with vRNAP are required for N4 early promoter recognition.

Subsequent genetic and biochemical studies confirmed these hypotheses and elucidated the molecular mechanism of vRNAP promoter recognition. The extrusion of ssDNA hairpins was detected in N4 early promoter sequences through enzymatic and chemic probes *in vitro* (9), while hairpin extrusion under physiological conditions *in vivo* was dependent on the activity of host DNA gyrase and presence of Mg^{2+} (10). Surprisingly, the non-template strand was sensitive to single-stranded probes, while the template strand was resistant, indicating that the two strands adopt different conformations and that these structural differences are determined by the sequence of the 3 nt loop and loop closing base pair (9, 10, 215). Results of runoff transcription assays on mutant promoters and crosslinking experiments with templates containing the

photocrosslinking nucleotide analog 5-iododeoxyuracil (5-IdU) at specific positions showed that vRNAP specifically recognizes −11A/G, −10G, −8G, and +1C (13, 14, 215). Contact with the −8 base was shown to occur through the major groove of the hairpin stem, while the site of crosslinking to the −11 purine residue was mapped to W129 in vRNAP (14).

Although transcription by vRNAP is dependent on the activity of host DNA gyrase to introduce negative supercoils into N4 DNA, supercoiled promoter-containing plasmids did not support vRNAP activity *in vitro* (7, 11); suggesting that another factor was required for N4 early promoter activation. Genetic analysis revealed *E. coli* single-stranded DNA-binding protein (*Eco*SSB) to be required for N4 early transcription *in vivo* (11). *In vitro*, *Eco*SSB activates vRNAP transcription initiation 40-fold from supercoiled templates with *in vivo* specificity by melting the non-template strand and stabilizing the template-strand DNA hairpin (11, 12). In this context, *Eco*SSB acts as an architectural transcription factor to provide an active promoter conformation for vRNAP binding (12, 13). Based on these data, a new model of promoter activation was proposed for N4 early transcription: the introduction of negative supercoils into N4 genomic DNA during genome injection by host DNA gyrase facilitates the extrusion of N4 early promoter template-strand hairpins, which are stabilized by *Eco*SSB to provide an active promoter for vRNAP binding and transcription initiation (13).

Along with its role in promoter activation, *Eco*SSB plays a role in vRNAP template recycling. *Eco*SSB activates vRNAP transcription *in vitro* at limiting ssDNA template concentrations by binding the transcript as it exits vRNAP and preventing the formation of inert RNA:DNA hybrids (11, 216). Comparison of the T7 RNAP and N4 vRNAP primary sequences indicated that part of the T7 RNAP N-terminal domain responsible for RNA separation and exit is missing from vRNAP. Therefore, *Eco*SSB binding to nascent RNA exiting the vRNAP active

site fulfills this function and activates vRNAP transcription on ssDNA templates through template recycling (137, 216).

Although mini-vRNAP shares limited sequence similarity with other T7-like RNAPs, the crystal structure of the apo form mini-vRNAP shows that it displays the characteristic cupped right hand architecture shared by all related DNAPs and RNAPs (217). Mini-vRNAP shares strong structural similarity with the T7 RNAP palm subdomain, confirming the results of mutational analyses identifying the vRNAP active site (84, 87, 217). In contrast to T7 RNAP, the crystal structure of apo mini-vRNAP reveals two structures, the "plug module" and the "motif B loop," that block the pathway of the DNA to the active site, suggesting that the apo mini-vRNAP structure is in an inactive conformation and must undergo structural rearrangements upon DNA-binding (217). Indeed, the binary complex (BC) of mini-vRNAP with an ssDNA hairpin promoter showed a rotation of the plug module and β-IH, allowing for the translocation and rearrangement of the motif B loop into the O helix upon promoter binding to grant ssDNA access to the active site (217, 218). These data indicate that the ssDNA hairpin promoter acts as an allosteric effector to reconfigure the vRNAP active site into an active conformation and explain this enzyme's remarkable specificity for its cognate promoters (218).

Despite drastic differences in consensus promoter sequence and architecture, mini-vRNAP and T7 RNAP surprisingly utilize the same structural motifs for promoter recognition (109, 120). In mini-vRNAP, the N-terminal −11 recognition element (analogous to the T7 RNAP AT-rich recognition loop) contacts the -11G base through base stacking with residue W129 (14, 217, 218), the β-IH defines and enforces the junction between dsDNA and ssDNA between bases −5 and −4 by melting the two base pairs at the base of the 7 bp stem, and the specificity loop makes sequence-specific contacts through the major groove of the hairpin stem. Interestingly,

there are no base contacts with the four As near the initiation site, suggesting that these bases act as a molecular ruler to start transcription eight nucleotides downstream of the −8 position (218).

Compared to the related DNAPs, DNA-dependent RNAPs are uniquely capable of carrying out first dinucleotide bond formation. Four structures of mini-vRNAP in complex with promoter hairpin DNA and various NTP substrates were solved to elucidate the molecular and structural mechanism of transcript initiation (219). These structures show that residues in the palm subdomain stabilize the initiating nucleotide, which base-stacks with -1 template-strand purine, explaining the conservation of purines at the −1 position of many T7-like RNAP promoters (119, 219). Furthermore, these structures show that the Mg^{2+} brought in with the nucleotide substrates causes a conformational change in the catalytic aspartates, suggesting that binding of the catalytic Mg^{2+} is the last step before catalysis of the dinucleotide bond (219), which is supported by direct observation through time-resolved X-ray crystallography (220).

N4 RNAPII

The existence of a second N4-encoded transcriptional activity was postulated based on the 100-fold decrease in the rate of post-infection RNA synthesis when cells were pretreated with chloramphenicol, indicating that N4 middle transcription requires the synthesis of N4 early gene products (5, 209, 211). Sites of N4 middle transcription initiation were mapped to a series of overlapping transcripts confined to the leftmost 50% of the N4 genome (25, 221). The early products responsible for this second transcriptional activity were identified through genetic analysis. Infection with N4 phages containing mutations in ORF15 (N4*am*15) or ORF16 (N4*am*23) significantly reduced the amount of post-infection RNA synthesis relative to wild-type (WT) N4, indicating that the products of these genes (gp15 and gp16) encode a second transcriptional activity responsible for middle transcription (5, 15, 209, 211). Additional

mutagenesis screens facilitated the discovery of a third N4 gene product (gp2) required for middle transcription. N4*am*126S mutant phages, which contain a mutation in ORF2, were defective in middle transcription but could complement infections by either N4ORF15*am* or N4ORF16*am* mutants, indicating the involvement of gp2 in N4 middle transcription (16, 222).

Gp15 and gp16, which comprise the heterodimeric, rifampicin-resistant RNAP responsible for N4 middle transcription (N4 RNAPII), were purified to homogeneity from infected cell extracts and tested for transcriptional activity *in vitro* (16, 222). N4 RNAPII does not bind or initiate transcription from native, double-stranded N4 DNA on its own, but transcribes single-stranded N4 DNA templates with low efficiency and little specificity (16, 17). N4 RNAPII displayed increased activity and specificity upon the addition of gp2 contained within the DNA-membrane complex of infected cell extracts, suggesting that gp2 may have a role in providing specificity to N4 RNAPII either by direct interaction or providing a specific secondary structure for promoter recognition in a sequence-dependent fashion (16, 17).

Gp2 is essential for middle transcription *in vivo*, but is insufficient to activate N4 RNAPII transcription from linear dsDNA templates *in vitro* (18, 222). To elucidate the role of gp2 in middle transcription, ORF2 was cloned and sequenced, revealing no homology to proteins of known function (18). Purified gp2 was tested for DNA-binding and transcription-enhancing properties *in vitro*, revealing that it: i) is a non-sequence specific ssDNA-binding protein; ii) stimulates N4 RNAPII transcription on ssDNA templates containing weak promoters; iii) binds cooperatively with N4 RNAPII to ssDNA; iv) directly interacts with N4 RNAPII with equimolar stoichiometry (18). These results suggest that gp2 activates N4 middle transcription by recruiting N4 RNAPII to melted ssDNA promoter sequences.

Sequence analysis of ORF15 and ORF16 (gp15 and gp16) showed that these proteins align to non-overlapping portions of T7 RNAP (883 aa), confirming that N4 RNAPII is a heterodimer and identifying the enzyme as a member of the T7-like RNAP family (15, 16, 86). N4 RNAPII contains 13 blocks of conserved sequence shared among T7-like RNAPs and all four conserved motifs required for DNA-dependent RNAP activity (DxxGR, A, B, and C) (15, 80). Gp15 (269 aa) aligns to the N-terminal domain, thumb subdomain, and DxxGR motif, while gp16 (404 aa) aligns to the palm and fingers subdomains along with motifs A, B, and C (15, 80). N4 RNAPII (673 aa) represents a minimal RNAP, as it is one of the smallest members of this family, with a truncated thumb subdomain and a truncated N-terminal domain shown to be involved in promoter recognition in vRNAP and T7 RNAP (15). However, N4 RNAPII shows much greater sequence homology to T7 RNAP than vRNAP and likely shares a similar architecture (15, 87).

The recently reported crystal structures of N4 RNAPII binary and elongation complexes largely confirm the sequence homology data (15, 223), indicating that N4 RNAPII displays the characteristic cupped right hand architecture of T7-like RNAPs and contains structural analogs of conserved motifs and promoter binding elements (see Figure IV.1) (80, 223). The palm subdomain forms a deep cleft for DNA binding and contains the catalytic aspartate residues in motifs A and C as well as the DxxGR motif. The palm subdomain is flanked by the fingers, which contains motif B, and the thumb subdomains. Compared to other T7-like RNAPs, N4 RNAPII fingers and thumb subdomains are truncated, resulting in a relatively open catalytic cleft (84, 191, 217, 223). Furthermore, N4 RNAPII structural analogs of the specificity loop, AT-rich recognition loop, and β-IH are truncated and poorly positioned to contact upstream DNA, which localizes to a basic patch between the thumb subdomain and N-terminal domain (15, 223) (see

Figure IV.1). The majority of N4 RNAPII-promoter interactions are non-sequence specific and the authors suggest that the N-terminal domain may primarily function as a platform for assembly with additional protein factors on pre-melted promoter DNA (223).

Surprisingly, the N4 RNAPII conformation in the elongation complex is largely unchanged from the conformation observed in the binary complex (223). Unlike T7 RNAP, which undergoes dramatic N-terminal rearrangements upon transition to the elongation phase, the N4 RNAPII N-terminal domain does not refold to accommodate the growing RNA:DNA heteroduplex and the RNA exit pore, formed by the specificity loop and N-terminal domain, is apparent prior to RNA synthesis (137, 139, 223). Therefore, N4 RNAPII promoter escape may be facilitated by the release of its transcription factor gp2, as occurs during the transition to transcription elongation in the mtRNAPs (166, 208).

These structures suggest that N4 RNAPII neither makes sequence-specific contacts with upstream promoter DNA during transcription initiation nor undergoes significant conformational changes upon transition to the elongation phase (223). Since N4 RNAPII does not bind to dsDNA, the use of DNA templates with dsDNA upstream of the -3 position likely interferes with N4 RNAPII interaction with upstream promoter sequences in the binary complex (17, 223). Thus, N4 RNAPII likely binds to the ssDNA region, which does not contain a consensus N4 RNAPII promoter sequence, in a sequence-independent manner (221, 223). Furthermore, the omission of the essential transcription factor gp2 means that these structures are not indicative of the pre-initiation complex *in vivo* and do not capture any potential protein or DNA conformational changes induced upon gp2 binding. Therefore, the reported N4 RNAPII binary complex more likely represents the N4 RNAPII elongation complex conformation, while the

mechanisms of N4 RNAPII promoter recognition and gp2 activation of transcription are yet to be determined.

CHAPTER II: MATERIALS AND METHODS

MATERIALS

Bacterial strains and phages

Bacterial strains used in this study are listed in Table II.1. All bacterial strains were obtained from the Rothman-Denes laboratory freezer stocks. WT N4 and N4 ORF15/16*am* phages were grown from the Rothman-Denes laboratory stocks.

Plasmids

All plasmids used in this study are listed in Table II.5. Expression plasmid pBAD-HisB was purchased from Invitrogen (Carlsbad, CA). Plasmid pSupT/BpF was obtained courtesy of the Rick Gourse laboratory (University of Wisconsin-Madison).

Media

Luria-Bertani (LB) broth was purchased from Research Products International (Mt. Prospect, IL) and prepared according to manufacturer's instructions.

Buffers

All buffers used in this study are listed in Table II.2.

Radiochemicals

UTP [α-^{32}P] (6,000 Ci/mmol), ATP [α-^{32}P] (3,000 Ci/mmol), ATP [γ-^{32}P] (6,000 Ci/mmol), and GTP [γ-^{32}P] (6,000 Ci/mmol) were purchased from Perkin Elmer (Waltham, MA).

Enzymes

Restriction enzymes, T4 polynucleotide kinase (PNK), DpnI, T4 DNA ligase, and DNaseI were purchased from New England Biolabs (Ipswich, MA). Lysozyme from chicken egg

white, trypsin, and chymotrypsin were purchased from Sigma-Aldrich (St. Louis, MO). Phusion Hotstart DNA polymerase and Proteinase K were purchased from Thermo Fisher Scientific (Ashville, NC). PFU Ultra II Fusion Hotstart DNA polymerase was purchased from Agilent Technologies (Wood Dale, IL).

Chemicals

Tween-20 was purchased from Acros Organics (Morris Plains, NJ). *p*-benzoyl-L-phenylalanine (*p*Bpa) was purchased from Bachem (Switzerland). Acrylamide solution, agarose, bis solution, bromophenol blue, oriole stain, and xylene cyanol were purchased from Bio-Rad Laboratories (Des Plaines, IL). Ethanol was purchased from Decon Laboratories (King of Prussia, PA). Cesium chloride, Coomassie brilliant blue G250, dithiothreitol (DTT), and L(+)-arabinose were purchased from Gold Biotechnology (Olivette, MO). Formic acid was purchased from J.T. Baker (Pillipsburgh, NJ). Ammonium bicarbonate, guanidinium chloride, and Triton X-100 were purchased from MP Biomedicals (Solon, OH). Bis-tris-propane-HCl, bovine serum albumin (BSA), and nucleotide triphosphates (NTPs) were purchased from New England Biolabs. Deoxynucleotide triphosphates (dNTPs) were purchased from Pharmacia Biotech (Sweden). Ampicillin and glycerol were purchased from Research Products International. Acetonitrile (AcN) was purchased from Riedel-deHaen (Germany). Ammonium persulfate (APS), boric acid, chloramphenicol, D(+)-glucose, ethidium bromide, formamide, iodoacetamide, magnesium chloride, phenol, chloroform, potassium hydroxide, SIGMAFAST ethylenediaminetetraacetic acid (EDTA)-free protease inhibitor, and β-mercaptoethanol were purchased from Sigma-Aldrich. Acetic acid, agar, chloroform, EDTA, glycine, hydrochloric acid, imidazole, methanol, orange G, sodium chloride, sodium dodecyl sulfate (SDS),

tetramethylethylenediamine (TEMED), and tris base were purchased from Thermo Fisher Scientific. Urea was purchased from VWR International (Batavia, IL).

Chromatographic resins and supplies

Empty 20 ml chromatography columns were purchased from Bio-Rad Laboratories. Amicon Ultra 30K centrifugal filters and 0.22 µm Millex GV polyvinylidene fluoride (PVDF) syringe filters were purchased from EMD Millipore (Ireland). Illustra NAP-10 columns, Illustra MicroSpin columns, and Heparin Sepharose 6 Fast Flow affinity resin were purchased from GE Healthcare (Piscataway, NJ). Ni-NTA agarose resin was purchased from Qiagen (Germantown, MD). SnakeSkin 3,500 molecular weight cutoff (MWCO) dialysis tubing was purchased from Thermo Fisher Scientific.

Immunoblot analysis

0.45µm PVDF membranes were purchased from EMD Millipore. ECL prime blocking agent and 0.1 µm nitrocellulose membranes were purchased from GE Healthcare. Mouse α HisG antibody was purchased from Invitrogen. To create rabbit α N4 gp2 antibody, gp2 was purified from inclusion bodies in strain RC9, New Zealand White rabbits were immunized, and three separate post-injection samples were obtained. Goat α rabbit Licor 680 and goat α mouse Licor 800 antibodies were obtained from LI-COR Biosciences (Lincoln, NE). Whatman 3MM chromatography paper was purchased from Whatman plc (United Kingdom).

Oligonucleotides

All oligonucleotides used in this study were purchased from Invitrogen. Oligonucleotides used for PCR amplification, cloning, site-directed mutagenesis, and DNA sequencing are listed in Table II.3. Oligonucleotides used as N4 middle promoter templates in *in vitro* runoff

transcription, electrophoretic mobility shift assays (EMSAs), filter binding assays, and crosslinking are listed in Table II.4.

Nucleotide purification kits

QIAprep spin miniprep, QIAquick gel extraction, QIAquick PCR purification, and MinElute PCR purification kits were purchased from Qiagen.

Protein and DNA standards

1 kb plus DNA ladder was purchased from Invitrogen. 10 bp DNA ladder was purchased from Life Technologies (Carlsbad, CA). microRNA marker was purchased from New England Biolabs. Precision plus protein unstained standards and Precision plus protein dual xtra standards were purchased from Bio-Rad.

METHODS

Bacterial strain growth conditions

E. coli K-12 strains W3350 and W3350*supF* were used for WT N4 and N4 ORF15/16*am* mutant phage infections, respectively. *E. coli* BL21 was used for protein expression. *E. coli* DH5α was used for plasmid screening and propagation. Bacteria were grown at 37°C with shaking at 250 rpm in LB medium supplemented with ampicillin (100 μg ml^{-1}) and chloramphenicol (20 μg ml^{-1}) as indicated.

Growth of N4 phages

Overnight cultures of *E. coli* W3350 or W3350*supF* were diluted 1:100 in 50 ml LB and grown at 37°C with shaking at 250 rpm until cells reached OD_{600} 0.2 ($7.7e^7$ cell ml^{-1}). Cultures were then infected with WT or ORF15/16*am* N4 phage at a multiplicity of infection (MOI) of 10 followed by incubation at 37°C with shaking at 250 rpm for 3 hrs. Cell lysis was induced by

addition of 250 µl chloroform followed by15 min incubation at 37°C with shaking at 250 rpm. Cellular debris was pelleted by centrifugation (15 min, 16,000 rpm, 4°C, Sorvall SM24). The supernatant was removed and filtered through 0.45 µm PVDF filter membrane and applied to a two-step glycerol gradient of 3 ml 40% (v/v) glycerol in TM and 3 ml 5% (v/v) glycerol in TM layered into a 14 x 89 polyallomer tube (Beckman Coulter; Brea, CA). Phages were pelleted through the glycerol gradient through ultracentrifugation at 178,305 rcf for 2 hrs at 4°C. The supernatant was discarded and phage pellet was resuspended in 50 µl TM buffer. Resuspended phages were then applied to three-step cesium chloride gradient of 1.15 ml of RI = 1.3955 cesium chloride in TM, 3.45 ml of RI = 1.3808 in TM, and 4.6 ml RI = 1.3704 in TM layered into a 14 x 89 ultra-clear tube (Beckman Coulter). Phages were spun at 151,263 rcf for 1 hr at 20°C. The opalescent phage band was extracted using 21.5 gauge needle and syringe. Extracted phages were applied to 3,500 MWCO SnakeSkin dialysis tubing and dialyzed into 1.8 L TM buffer for 2 hr at 4°C thrice. Phages were titered and stored at 4°C.

Isolation of N4 phage DNA

10^{10} plaque forming units (PFU) of WT N4 phage stocks were treated with Proteinase K (50 µg ml^{-1}) and SDS (10% (w/v)) for 2 hrs at 55°C. Sample was diluted to 300 µl total volume, an equal volume (300 µl) phenol was added, the sample was vortexed rigorously for 1 min, and then spun at 14,000 rcf for 3 min. The top aqueous DNA-containing layer was removed and phenol extraction was repeated two additional times. An equal volume of chloroform (300 µl) was added, the sample was vortexed rigorously for 1 min, and then spun at 14,000 rcf for 3 min. The top aqueous DNA-containing layer was removed and chloroform extraction was repeated two additional times. 1:10 volume (30 µl) 3 M sodium acetate pH 5.2 and 3 volumes (900 µl) 100% ethanol was added to the final chloroform-extracted layer and samples were incubated

overnight at -20°C. DNA was precipitated by centrifugation at 16,000 rcf for 30 min at 4°C. The supernatant was aspirated, the pellet air-dried, and the DNA pellet resuspended in 15 µl TE buffer. DNA concentration was measured by A_{260} measurement on Nanodrop 1000 (Thermo Fisher Scientific).

Cloning of gp2 and N4 RNAPII into inducible plasmids

Gp2 and N4 RNAPII were cloned into pBAD-HisB expression vectors by Oleksander Demidenko as previously described (18, 19). Briefly, WT untagged gp2 was cloned into the NcoI-PstI site of pBAD-HisB to create plasmid pOD9 and WT N4 RNAPII containing a 36 aa (MGGSHHHHHHGMASMTGGQQMGRDLYDDDDKDPSSR) N-terminal leader sequence, including a hexahistidine (His_6) tag and Xpress epitope, was cloned into pBAD-HisB vector to create plasmid pAD1.

Polymerase chain reaction (PCR) amplification of DNA

PCR amplification of DNA sequences was used for sequencing, plasmid screening, and cloning. PCR amplification was performed using Phusion Hotstart DNA polymerase according to manufacturer's instructions.

Site-directed mutagenesis of plasmid sequences

Individual aa substitutions were introduced to N4 RNAPII and gp2 sequences through Quikchange PCR mutagenesis of pAD1 and pOD9, respectively. Complementary primers containing desired point mutations (see Table II.3) were used to prime PCR amplification performed using PFU Ultra II Fusion Hotstart DNA polymerase according to manufacturer's instructions. PCR products were digested with DpnI for 3 hrs at 37°C and mutagenized products were screened through electroporation into DH5α cells (see below).

Isolation and purification of DNA

Small scale purification of plasmid DNA from bacterial cells was performed using the QIAprep spin miniprep columns according to manufacturer's instructions. Digested DNA fragments for restriction enzyme cloning were purified by QIAquick gel extraction according to manufacturer's instructions. PCR-generated DNA fragments were purified and isolated by either MinElute or QIAquick PCR purification columns according to manufacturer's instructions.

Transformation of plasmid DNA

Plasmid DNA was electroporated into electrocompetent W3350, DH5α, or BL21 *E. coli* strains for protein expression, propagation of plasmid stocks, and screening of cloned plasmid sequences. 1 or 2 µl of plasmid DNA was added to a 50 µl aliquot of electrocompetent cells and incubated on ice for 3 min. Solutions were transferred to cold 1 mm electroporation cuvette and electroporation was performed using Gene Pulser Xcell (Bio-Rad Laboratories) (25 µF, 1.8 kV). For plasmids containing ampicillin resistance markers, cells were resuspended in 200 µl LB media and immediately plated on LB plates supplemented with ampicillin (100 µg ml^{-1}). For all other plasmids, cells were immediately resuspended in 1 ml LB media, incubated at 37°C for 1 hr, then plated on LB plates supplemented with the corresponding antibiotic. Transformants were incubated overnight at 37°C and individual colonies were restreaked onto fresh LB plates supplemented with the corresponding antibiotic. Plasmids were isolated and sequenced to confirm identify of the mutations.

Electrocompetent cells were prepared by 1:100 dilution of overnight culture of W3350, BL21 or DH5α cells into 1 L LB. Cells were incubated at 37°C with 250 rpm rotation until cells reached OD_{600} 0.5. Cells were harvested by centrifugation (10 min at 6,000 rcf) and resuspended in 200 ml H_2O. Cells were harvested by centrifugation (10 min at 6,000 rcf) and resuspended in

25 ml 10% (v/v) glycerol solution. Cell harvesting by centrifugation (10 min at 6,000 rcf) and resuspended in 25 ml 10% (v/v) glycerol was repeated two additional times. Upon resuspension in 0.5 ml 10% (v/v) glycerol, 50 µl aliquots were distributed into individual Eppendorf tubes, frozen in liquid nitrogen, and stored at -80°C.

DNA sequencing

The recommended amount and concentration of sequencing primers (see Table II.3) and plasmid or PCR-amplified DNA samples were submitted to the University of Chicago Comprehensive Cancer Center DNA Sequencing and Genotyping Facility for automated sequencing. Chromatograms were compared to N4 genomic DNA and plasmid constructs using SnapGene (GSL Biotech; Chicago, IL) (v. 4.2.3) to verify sequences of interest.

Oligonucleotide annealing

Oligonucleotides were annealed in 20 µl reactions containing 1 µM template strand and 1 µM non-template strand Pm5 templates (see Table II.4) in Oligonucleotide annealing buffer for 5 min at 95°C followed by slow cooling to room temperature over 3 hrs.

Polyacrylamide gel electrophoresis (PAGE)

PAGE was used for high-resolution separation of RNA, DNA, and protein-DNA complexes. Polyacrylamide gel composition and running conditions are individually listed for each experiment.

Agarose gel electrophoresis

Agarose gel electrophoresis was used for the identification and separation of digested plasmids and PCR products. 5 µl of DNA samples were mixed with 1 µl 6x DNA loading buffer and loaded into 0.8% or 2% agarose gels (0.8% or 2% (w/v) agarose, 0.5x TBE, 0.5 µg ml^{-1} ethidium bromide). Gels were run for 30 min at 100 mA in 0.5x TBE supplemented with 0.5 µg

ml^{-1} ethidium bromide. DNA bands were visualized by ultraviolet light (UV) illumination at 300 nm.

SDS-PAGE analysis of proteins

SDS-PAGE was used for the separation of proteins. SDS-PAGE gels contained 5 ml resolving portion (10, 12, or 15% (v/v) 29:1 acrylamide:bisacrylamide, 25 mM Tris-HCl pH 6.8, 1% (w/v) SDS) and 3 ml stacking portion (4% (v/v) 29:1 acrylamide:bisacrylamide, 25 mM Tris-HCl pH 6.8, 1% (w/v) SDS). 8 µl of protein samples were mixed with 2 µl 5x SDS sample buffer and loaded into SDS-PAGE gels. Gels were run in Laemmli running buffer for 40 min at 100 mA. Proteins were visualized by either Coomassie stain (0.1% (w/v) Coomassie Blue R-250, 40% (v/v) methanol, 10% (v/v) acetic acid) or Oriole staining visualized by UV illumination at 300 nm.

Expression of *p*Bpa-substituted N4 RNAPII alleles

Overnight culture of BL21 cells containing N4 RNAPII overexpression plasmids and tRNA/tRNA synthetase plasmids (pAD1 derivatives and pSupT/BpF) were used to inoculate 1 L LB containing *p*Bpa (1 mM), ampicillin (100 µg ml^{-1}), and chloramphenicol (20 µg ml^{-1}). Cells were incubated at 37°C with 250 rpm rotation until cells reached OD_{600} 0.5 and protein expression was induced with 0.2% (w/v) L(+)-arabinose for 4 hrs at 37°C with 250 rpm rotation. Cells were harvested by centrifugation (10 min, 7,000 rpm, 4°C, Sorvall SH3000), resuspended in 3 ml Ni-NTA binding buffer, and stored at -80°C.

Expression of *p*Bpa-substituted gp2 alleles

Overnight culture of BL21 cells containing gp2 overexpression plasmids and tRNA/tRNA synthetase plasmids (pOD9 derivatives and pSupT/BpF) were used to inoculate 200 ml LB containing *p*Bpa (1 mM), ampicillin (100 µg ml^{-1}), and chloramphenicol (20 µg ml^{-1}).

Cells were incubated at 37°C with 250 rpm rotation until cells reached OD_{600} 0.5 and protein expression was induced with 0.2% (w/v) L(+)-arabinose for 4 hrs at 37°C with 250 rpm rotation. Cells were harvested by centrifugation (10 min, 7,000 rpm, 4°C, Sorvall SH3000) and resuspended in 6 ml gp2 lysis buffer. Resuspended cells were stored at -80°C in 1 ml aliquots.

N4 RNAPII purification

N4 RNAPII was purified as previously described with modifications (224). Overnight culture of BL21 cells containing N4 RNAPII overexpression plasmids (pAD1 and derivatives) were used to inoculate 1 L LB containing ampicillin (100 μg ml^{-1}). Cells were incubated at 37°C with 250 rpm rotation until cells reached OD_{600} 0.5 and protein expression was induced with 0.2% (w/v) L(+)-arabinose for 3 hrs at 37°C with 250 rpm rotation. Cells were harvested by centrifugation (10 min, 7,000 rpm, 4°C, Sorvall SH3000), resuspended in 3 ml Ni-NTA binding buffer and stored at -80°C. N4 RNAPII was purified by immobilized metal affinity chromatography (IMAC) followed by heparin affinity chromatography. Frozen cell pellets were lysed by addition of lysozyme from chicken egg white to a final concentration of 500 μg ml^{-1} and 4 freeze-thaw cycles. Further lysis was induced by 4x 30 sec sonication. Cellular debris was cleared by ultracentrifugation at 100,000 rcf for 1 hr at 4°C. Cleared lysate was then added to 0.5 ml pre-equilibrated Ni-NTA resin. Resin-lysate mix was incubated at 4°C with rotation for 1 hr. Lysate-resin mix was then added to column and washed with 100 ml Ni-NTA wash buffer, 2.5 ml of Ni-NTA wash buffer II, 2.5 ml of Ni-NTA wash buffer III, 2.5 ml of Ni-NTA wash buffer IV, and 0.5 ml Ni-NTA wash buffer V. N4 RNAPII was eluted in 3 ml Ni-NTA elution buffer. N4 RNAPII-containing fractions were pooled and dialyzed into TMGE + 50 mM NaCl buffer. Dialyzed sample was added to 0.25 ml pre-equilibrated Heparin Sepharose 6 Fast Flow resin. Resin-sample mix was incubated at 4°C with rotation for 1 hr. Sample-resin mix was added to

column and washed with 6 ml TMGE+50 mM NaCl buffer and 0.75 ml TMGE+100 mM NaCl buffer. Sample was eluted with 1.25 ml TMGE+500 mM NaCl buffer. Eluted sample was pooled and concentrated four-fold with Amicon Ultra 30K centrifugal filter unit, dialyzed into N4 RNAPII storage buffer, aliquoted into 250 µl aliquots, flash frozen with liquid nitrogen, and stored at -80°C. Protein concentrations were measured by A_{280} measurement on Nanodrop 1000 using extinction coefficient and molecular weight estimates.

Gp2 purification

Gp2 was purified according to a denaturing protocol as previously described (224). Overnight culture of BL21 cells containing gp2 overexpression plasmids (pOD9 and derivatives) were used to inoculate 1 L LB containing ampicillin (100 µg ml^{-1}). Cells were incubated at 37°C with 250 rpm rotation until cells reached OD_{600} 0.5 and protein expression was induced with 0.2% (w/v) L(+)-arabinose for 3 hrs at 37°C with 250 rpm rotation. Cells were harvested by centrifugation (10 min, 7000 rpm, 4°C, Sorvall SH3000), resuspended in 30 ml gp2 lysis buffer, and stored as 1 ml aliquots at -80°C. Gp2 was purified from inclusion bodies by a denaturing protocol followed by protein refolding. Frozen cell pellets were lysed by addition of lysozyme from chicken egg white to a final concentration of 500 µg ml^{-1} and 5 freeze-thaw cycles. Further lysis and DNA fragmentation was induced by 2x 15 sec sonication. Cellular debris, DNA, and inclusion bodies were pelleted by centrifugation at 18,000 rcf for 30 min at 4°C. Soluble material was removed and insoluble material was resuspended in 1 ml gp2 1 M NaCl buffer by sonication for 30 sec. Insoluble material was pelleted by centrifugation at 18,000 rcf for 30 min at 4°C. Soluble material was removed and insoluble material was resuspended in 1 ml gp2 detergent buffer by sonication for 30 sec. Insoluble material was pelleted by centrifugation at 18,000 rcf for 30 min at 4°C. Soluble material was removed and insoluble material was resuspended in 1 ml

gp2 50 mM NaCl buffer by sonication for 30 sec. Insoluble material was pelleted by centrifugation at 18,000 rcf for 30 min at 4°C. Soluble material was removed and insoluble material was resuspended in 0.5 ml gp2 denaturing buffer followed by overnight incubation at 4°C with rotation. Insoluble material was pelleted by centrifugation at 18,000 rcf for 30 min at 4°C. Soluble material was removed, filtered through 0.22 μm PVDF membrane, diluted 20-fold with gp2 50 mM NaCl buffer, and incubated at 4°C for 2 hrs. Insoluble material was pelleted by centrifugation at 10,000 rcf for 15 min at 4°C. Soluble material was removed and insoluble material was resuspended in 1 ml gp2 denaturing buffer followed by incubation at 4°C for 2 hrs with rotation. Soluble material was removed, filtered through 0.22 μm PVDF membrane, diluted with gp2 denaturing buffer II to approximately 0.25 mg ml^{-1} as determined by A_{280} measurement on Nanodrop 1000 using extinction coefficient and molecular weight estimates. NAP-10 column was equilibrated with 15 ml gp2 exchange buffer and 1 ml diluted protein sample was added to column and buffer exchanged by addition of 1.5 ml gp2 exchange buffer. Buffer exchange was repeated for entire diluted sample. Buffer exchanged gp2 was dialyzed into gp2 storage buffer, aliquoted into 500 μl aliquots, flash frozen with liquid nitrogen, and stored at -80°C. Protein concentrations were calculated by 15% SDS-PAGE and Oriole staining with lysozyme concentration standards.

End-labeling oligonucleotides

Oligonucleotides (see Table II.4) were 5' end-labeled with [γ^{32}P] ATP in 20 μl reactions containing 100 nM oligonucleotide, 1x PNK buffer (New England Biolabs), 10 U T4 PNK, and 100 nM [γ^{32}P] ATP for 3 hrs at 37°C. Enzymes were heat inactivated by incubation at 95°C for 5 min and products were stored at -20°C.

Runoff transcription

N4 RNAPII runoff transcription was performed as previously described with modifications (15, 18). Reactions (10 µl) contained the indicated concentrations of N4 RNAPII, gp2, and ssDNA oligonucleotide templates (see Table II.4), 1 mM each of ATP, CTP, and GTP, 0.1 mM UTP, 5 µCi [α^{32}P] UTP (unless otherwise specified) in transcription buffer. Reactions were prepared on ice and transcription was initiated upon addition of NTPs and transferred to 37°C for 5 min. Reactions were terminated upon addition of equal volume (10 µl) transcription stop buffer. Reactions were incubated at 95°C for 3 min and RNA products resolved by electrophoresis on denaturing 8% (v/v) polyacrylamide gels (19:1 acrylamide:bisacrylamide, 8 M urea, 1x TBE) for 1 hr at 60 W. Gels were dried and analyzed by phosphorimaging. Runoff transcription products were quantified and normalized to 5' end-labeled ssDNA loading control (TotalLab v. 12.2).

Kinetics of first phosphodiester bond formation

Transcription reactions (10 µl) contained 1 µM ssDNA oligonucleotide template (BL-121, see Table II.4), 1 mM ATP, 1 µCi [α^{32}P] ATP, gradient of GTP (11.72, 23.44, 46.88, 91.75, 187.5, 375, 750 µM GTP), 250 nM WT N4 RNAPII, and indicated amounts of gp2 in transcription buffer. Reactions were prepared on ice and transcription was initiated upon addition of 1 mM ATP and transfer to 37°C for 5 min. Reactions were terminated upon addition of equal volume (10 µl) transcription stop buffer. Reactions were incubated at 95°C for 3 min and RNA products resolved by electrophoresis on denaturing 24% (v/v) polyacrylamide gels (19:1 acrylamide:bisacrylamide, 4 M urea, 1x TBE) for 2 hr at 60 W. Gels were dried and analyzed by phosphorimaging. Dinucleotide product and free ATP bands were quantified (TotalLab) and molar amounts of RNA synthesized were calculated according to Equation 1:

$$\text{RNA dinucleotide } (\mu M) = \frac{R}{R+A} * [\text{ATP}] \ (\mu M)$$

R and A are the band intensities of RNA dinucleotide and free ATP, respectively. To estimate the kinetic parameters of first phosphodiester bond formation, the rate of dinucleotide synthesis was calculated according to Equation 2:

$$V\left(\text{min}^{-1}\right) = \frac{D\ (\mu M)}{t\ (\text{min}) * E\ (\mu M)}$$

V is the rate of dinucleotide synthesis, D is the molar amount of RNA synthesized from Equation 1, t is the reaction time, and E is the concentration of N4 RNAPII. The Michaelis constant (K_m) and rate constant for the limiting step at saturation (k_{cat}) were calculated by fitting data through non-linear least squares regression (nlstools, R statistical package v. 3.3.0) (225) to Equation 3:

$$V\left(\text{min}^{-1}\right) = \frac{k_{cat} * [GTP]\ (\mu M)}{K_m + [GTP]\ (\mu M)}$$

V is the rate of dinucleotide synthesis from Equation 2. Parameters were estimated from three independent replicates.

EMSAs

EMSA reactions (20 µl) were carried out by incubating indicated concentrations of N4 RNAPII and gp2 with 1 nM 5' end-labeled ssDNA templates (see Table II.4) on ice in transcription buffer for 15 min. 5 µl 5x DNA loading buffer II was added and 5 µl of each product was loaded into native 6% (v/v) polyacrylamide gels (19:1 acrylamide:bisacrylamide, 0.25x TBE). Products were separated by electrophoresis for 2 hrs at 150 V. Gels were dried and analyzed by phosphorimaging. Free and shifted DNA complexes were quantified (TotalLab) and fraction DNA bound was calculated according to Equation 4:

$$\theta = 1 - \frac{DNA_{free}}{DNA_{free} + DNA_{bound} - DNA_{bound,0}}$$

θ is fraction DNA bound, DNA_{free} is the signal of unbound DNA, DNA_{bound} is the signal of all shifted DNA, and $DNA_{bound,0}$ is the background signal of all shifted DNA at in the absence of

protein. Dissociation constant (K_d) estimates were obtained by fitting data to the hill equation through non-linear least squares regression (nlstools, R statistical package).

Filter binding assays

0.1 μm nitrocellulose membranes were soaked in 0.4 M potassium hydroxide for 10 min with rotation followed by rinsing in Milli-Q H_2O thrice for 10 min each. Whatman chromatography paper and nitrocellulose membranes were then equilibrated with TME buffer for 1 hr with rotation. WT N4 RNAPII was diluted to the specified concentrations and incubated with 10 pM 5' end-labeled substituted Pm5 ssDNA templates (see Table II.4) in filter binding buffer in 100 μl total volume on ice for 15 min. Entire sample was then loaded into Minifold I microsample filtration manifold (Schleicher & Schuell; Keene, NH) and passed through nitrocellulose membranes by application of vacuum. Membranes were washed twice with 100 μl TME buffer before drying and visualization by phosphorimaging. Experiments were performed in quadruplicate. Spot intensity was quantified (TotalLab) and fraction DNA bound was calculated according to Equation 5:

$$\theta = \frac{V - V_0}{V_{max} - V_0}$$

θ is fraction DNA bound, V is the spot intensity, V_0 is the spot intensity in the absence of protein, and V_{max} is the spot intensity at the highest protein concentration. K_d estimates were obtained by fitting data to the hill equation through non-linear least squares regression (nlstools, R statistical package).

N4 RNAPII-gp2 immobilized metal affinity chromatography interaction assays

N4 RNAPII-gp2 interaction assays were performed as previously described with modifications (18). Reactions (100 μl) were carried out by incubating 0.5 μM N4 RNAPII with 1.5 μM gp2 in N4 RNAPII-gp2 interaction buffer at room temperature for 15 min. Samples were added to 25 μl

pre-equilibrated Ni-NTA resin in Illustra MicroSpin columns and resin-sample mix was incubated at room temperature for 30 min. Flow-through was collected by centrifugation (1 min, 250 rcf). Columns were washed with low salt and high salt buffer and flow-through was collected by centrifugation (1 min, 250 rcf). Samples were eluted with Ni-NTA elution buffer and flow-through was collected by centrifugation (1 min, 250 rcf). Samples were incubated for 5 min at 95°C and analyzed by 12% SDS-PAGE and Oriole staining. Protein molar concentration estimated through band quantification (TotalLab) and normalization to lysozyme standards.

N4 RNAPII-ssDNA template crosslinking

N4 RNAPII-ssDNA crosslinking reactions (20 µl) were carried out by incubating 1.0 µM *p*Bpa-containing N4 RNAPII alleles with 100 nM 5' end-labeled ssDNA templates in transcription buffer for 5 min at 37°C. Samples were irradiated with 365 nm UV light for 90 min in UV Stratalinker 2400 (Stratagene; La Jolla, CA). Samples were incubated for 5 min at 95°C and analyzed by 10% SDS-PAGE and Oriole staining. Gels were dried and analyzed by phosphorimaging.

N4 RNAPII-gp2 crosslinking

N4 RNAPII-gp2Bpa crosslinking reactions (60 µl) were carried out by incubating 2.5 µM N4 RNAPII with 500 nM F6Bpa gp2 in N4 RNAPII-gp2 interaction buffer at room temperature for 15 min. N4 RNAPIIBpa-gp2 crosslinking reactions (50 µl) were carried out by incubating 400 nM N4 RNAPIIBpa alleles with 2.0 µM WT or I3A gp2 in N4 RNAPII-gp2 interaction buffer II at room temperature for 15 min. Samples were irradiated with 365 nm UV light for 60 min in UV Stratalinker 2400. Samples were incubated for 5 min at 95°C and analyzed by 12% SDS-PAGE and either Oriole staining or western-blot (see below).

Mass spectroscopy and proteomics analysis

SDS-PAGE bands were excised, reduced with DTT (10 mM), cysteine thiols alkylated with iodoacetamide (40 mM, 60 min), washed, and digested overnight with mass spectrometry-grade proteases at either 37˚C (trypsin) or 25˚C (chymotrypsin). Peptides were extracted from gel pieces by 3 rounds of washing successively with 50 mM ammonium bicarbonate and AcN + 0.1% (v/v) formic acid, and dried by vacuum centrifugation. For liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis, peptide samples were reconstituted in 2% (v/v) AcN + 0.1% (v/v) formic acid and separated on a monolithic capillary C18 column (GL Sciences Monocap High Resolution Ultra, 100 µm I.D. × 200 cm length) using a water-AcN + 0.1% (v/v) formic acid gradient (2-50% (v/v) AcN over 180 min) at 360 nl min^{-1} using a Dionex Ultimate 3000 LC system with nanoelectrospray ionization (Proxeon Nanospray Flex source). Mass spectra were collected on an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) operating in a data-dependent acquisition mode, with one 120,000 $m/\Delta m$ MS1 parent ion full scan in the Orbitrap analyzer triggering ten 15,000 $m/\Delta m$ collision-induced dissociation MS2 fragment ion scans in the Orbitrap of intensity-selected precursors. Proteomic mass spectral data were analyzed using Stavrox (v. 3.6.6.6) (226). Precursor and product ion mass tolerances were set to 12 and 120 parts per million (ppm), respectively, *p*Bpa designated as the crosslinker, appropriate protease sites designated (KR for trypsin; FWY for chymotrypsin; blocked by C-terminal proline), and static cysteine carbamidomethylation and variable methionine oxidation were included as modifications.

Western blotting

0.45 µm PVDF membranes were rinsed with 40% (v/v) methanol for 30 sec. Membranes were rinsed with dH_2O and equilibrated with transfer buffer for 5 min with rotation. Proteins

were transferred to PVDF membranes with TransBlot-SD (Bio-Rad) for 30 min at 200 mA. PVDF membranes were washed with 20 ml TBST for 5 min with rotation. PVDF membranes were blocked in 20 ml blocking buffer for 1 hr at room temperature with rotation. PVDF membranes were incubated with 20 ml primary antibody solution for 1 hr at room temperature with rotation. PVDF membranes were washed with 20 ml TBST buffer three times for 10 min at room temperature with rotation. PVDF membranes were incubated with 20 ml secondary antibody solution for 1 hr at room temperature with rotation. PVDF membranes were washed with 20 ml TBST buffer three times for 10 min at room temperature with rotation. Membranes were imaged with Licor Odyssey CLx dual channel (LI-COR Biosciences) at 700 and 800 nm.

Secondary structure prediction of gp2

Gp2 secondary structure was predicted using PSIPRED (v. 3.3) with default settings (227).

Bacteriophage N4 bioinformatics annotations

N4 ORFs were compared with the proteins in the National Center for Biotechnology Information (NCBI) non-redundant protein database using domain enhanced lookup time accelerated basic local alignment search tool (DELTA-BLAST) (v. 2.8.0+) with 0.01 DELTA-BLAST and position-specific iterated BLAST (PSI-BLAST) E-value threshold values (228–230). Bioinformatics annotations were assigned to N4 ORFs based on homologous proteins with functionally validated annotations or sequence motifs identified by Interpro (v. 69.0), Conserved Domains Database (v. 3.16), and TMHMM (v. 2.0) searches (231–233). Predicted N4-encoded tRNAs were identified by ARAGORN (v. 1.2.36) using default settings (234).

Codon usage bias

The codon usage bias for bacteriophage N4 (Genbank accession number: NC_008720.1) and *E. coli* (Genbank accession number: U00096.3) ORFs was calculated as the normalized synonymous codon usage (NSCU) using DNA Master (v. 5.22.23).

Identification of N4-like phages

N4 vRNAP (gp50, Genbank accession number: YP_950528.1) was used as a query for DELTA-BLAST homology search against the NCBI non-redundant protein sequence database with 0.001 PSI-BLAST and DELTA-BLAST threshold values (228, 235). Phage species with homologous ORF(s) alignments spanning greater than 40% of the query length and scores below the E-value thresholds are considered N4-like and subjected to further bioinformatics analyses. N4 vRNAP homolog sequences were then aligned using Clustal Omega (v. 1.2.2) with default settings for detection of conservation at motifs required for vRNAP-like RNAP activity (82, 87, 236, 237).

N4-like phage genome-wide pairwise comparisons

Genbank and FASTA files for all N4-like phages identified were downloaded from NCBI databases in May 2017 and used for comparative genomics analyses. The pairwise fraction nucleotide alignment for all phage genomes was determined by NCBI nucleotide BLAST (BLASTn) (v. 2.8.0+) using a custom database of N4-like phages and default parameters. Calculation of the pairwise average nucleotide identity (ANI) was performed by DNA Master with default parameters. Pairwise measurements of fraction nucleotide alignment and ANI were summarized in pairwise distance matrices and visualized as heatmap with hierarchical clustering (gplots, R statistical package). N4-like phage genomes were clustered according to the methodologies for comparative genomics of mycobacteriophages (66). A 0.4 BLASTn alignment

score was used as the cutoff for N4-like phage cluster membership. Phages lacking pairwise nucleotide alignments with more than one other phage species above this threshold were not placed into an N4-like phage cluster and are considered "singleton" species.

N4-like phage genome alignment visualization

N4-like phage FASTA nucleotide sequences were concatenated by phage cluster assignment, and visualized by Gepard (v. 1.40) with default settings (238). N4-like phage Genbank files were loaded into EasyFig (v. 2.1), and whole genome BLASTn comparisons were made with default settings and visualized as linear genomes (239).

All-vs.-all BLAST of N4-like ORFs

All N4-like ORF sequences were concatenated into a single FASTA file. N4-like phage ORF custom BLAST database was created using BLASTDB (v. 2.5.0+). All N4-like ORFs were searched against custom database by protein BLAST (BLASTp) (v. 2.5.0+) using an E-value threshold of 0.001. Sequence alignments for homologous proteins were performed using Clustal Omega (v. 1.2.2) with default settings (236).

N4-like phage ORFam designation and network analysis

4,921 out of 4,922 ORFs in N4-like phage subfamily were sorted into 1,016 ORF families (ORFams) using the Markov Cluster (MCL) Algorithm (v. 12.135) (240). Briefly, pairwise similarities of all N4-like phage ORFs were calculated as the negative base ten logarithm conversions of the BLASTp E-values (cutoff of 0.001) and ORFams were derived by the MCL algorithm with an inflation factor of 1.2. One YH6 hypothetical protein (Genbank accession number: YP_009152534.1) was excluded from clustering analysis since no BLAST hit met the E-value threshold, while its small size (11 aa) suggests that this putative ORF may be falsely annotated. The ORFam membership table was converted into a presence/absence

incidence matrix; where rows represent bacteriophage, columns represent ORFams, and each cell contains a 1 when that phage encodes an ORF belonging to the family and a 0 if it does not. ORFam incidence matrix was converted to nexus format and reticulate network representation of N4-like phage subfamily relationships by shared ORFams was created by the NeighborNet method with default parameters in Splitstree (v. 4.14.4) (241).

The N4-like phage-ORFam incidence matrix was used to create a reticulate network of shared ORFams within the N4-like phage subfamily. Associations between N4-like phage species and ORFams were determined by findModules function with 50 iterations (lpbrim, R statistical package) (242). The probability of finding equally modular data by random chance was calculated through 99 randomly permuted matrices containing the same number of phage-ORFam associations.

N4-like phage cluster diversity and isolation

The ORFam incidence matrix was used to calculate the diversity and isolation of N4-like phage clusters. Cluster average shared ORFams (CLASO) was calculated as the fraction of ORFams shared between two genomes averaged across all pairs within the cluster. Cluster associated ORFams (CAO) was calculated as the number of ORFams conserved in all phages in the cluster divided by the average number of ORFams per genome. Cluster isolation index (CII) is calculated as the fraction of all ORFams in a cluster that are unique to that cluster.

Genbank accession numbers

N4, NC_008720.1; IME11, NC_019423.1; vBEcoPPhAPEC7, NC_024790.1; phiAxp-3, NC_028908.1; JWDelta, KF787094.1; JWAlpha, NC_023556.1; EC1-UPM, KC206276.2; ECBP1, NC_018854.1; Bp4, NC_024142.2; vBEcoPPhAPEC5, NC_024786.1; vBEcoPG7C, NC_015933.1; Frozen, NC_031062.1; Gutmeister, KX098391.1; Rexella, KX098390.1; Ea9-2,

NC_023579.1; RG-2014, NC_027348.1; pSb-1, NC_023589.1; phiCB2047-B, NC_020862.2; DSS3P2, NC_012697.1; EE36P1, NC_012696.1; vBDshPR2C, KJ803031.1; RD-1410Ws-07, KU885990.1; DFL12phi1, NC_024367.1; vBEamP-S6, NC_019514.1; DS-1410Ws-06, KU885988.1; FSL_SP-076, NC_021782.1; FSL_SP-058, NC_021772.1; Plymouth_1, FR719956.1; Pollock, NC_027381.1; Roseovarius_217, FR682616.1; RD-1410W1-01, KU885989.1; EcP1, NC_019485.1; LUZ7, NC_013691.1; VCO139, KC438283.1; vBPaePC2-10Ab09, NC_024140.1; DL64, NC_028885.1; KPP21, NC_029017.1; phi1, NC_028799.1; YH30, NC_029101.1; Pa2, NC_027345.1; phi176, KM411960.1; PEV2, NC_031063.1; RWG, KM411958.1; LIT1, NC_013692.1, vBPaePMAG4, NC_031104.1; Presley, NC_023581.1; PA26, JX194238.1; YH6, NC_027388.1; JA-1, NC_021540.1; VBP32, NC_020868.1; VBP47, NC_020848.1; pYD6-A, NC_020849.1.

Table II.1. Bacterial strains

Strain	Genotype	Use	Source
W3350	F^- *galK2 galT22* λ^- IN(*rrnD-rrnE*)1	Growth and study of N4 phage	Rothman-Denes lab stocks
W3350s*upF*	F^- *galK2 galT22* λ^- IN(*rrnD-rrnE*)1 *sup*F	Growth and study of N4 amber phage	Rothman-Denes lab stocks
BL21	F^- *ompT gal dcm lon hsdS*$_B$($r_B^- m_B^-$) [*malB*$^+$]$_{K\text{-}12}$(λ^S)	Protein expression	Rothman-Denes lab stocks
DH5α	F^- *endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG purB20* φ80d*lacZ*Δ M15 Δ(*lacZYA-argF*)U169, hsdR17($r_K^- m_K^+$), λ^-	Cloning and preparation of DNA	Rothman-Denes lab stocks

Table II.2. Buffers

Buffer	Composition	Purpose
TM	10 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$	N4 phage storage
TE	10 mM Tris-HCl pH 8.0, 1 mM EDTA	DNA resuspension
TA	0.7% (w/v) agar, 1x LB	Phage titer
5x SDS sample buffer	325 mM Tris-HCl pH 6.8, 5% (w/v) SDS, 62.5% (v/v) glycerol, 2.5% (v/v) β-mercaptoethanol, 0.05% (w/v) bromophenol blue	Protein sample loading buffer for SDS-PAGE
6x DNA loading buffer	10 mM Tris-HCl pH 8.0, 60% glycerol (v/v), 60 mM EDTA, 0.15% (w/v) orange G	DNA sample loading buffer for agarose gel electrophoresis
5x DNA loading buffer II	50% (v/v) glycerol, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol, 0.05% (w/v) orange G	DNA sample loading buffer for PAGE
TBE	89 mM Tris, 89 mM boric acid, 2 mM EDTA	Agarose and PAGE running buffer
Laemmli running buffer	25 mM Tris-HCl pH 6.8, 192 mM glycine, 1% (w/v) SDS	SDS-PAGE running buffer
Ni-NTA binding buffer	20 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole, 1 mM β-mercaptoethanol, EDTA-free protease inhibitor	N4 RNAPII IMAC purification
Ni-NTA wash buffer	20 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole, 1 mM β-mercaptoethanol	N4 RNAPII IMAC purification
Ni-NTA wash buffer II	20 mM Tris-HCl pH 8.0, 1 M NaCl, 20 mM imidazole, 1 mM β-mercaptoethanol	N4 RNAPII IMAC purification
Ni-NTA wash buffer III	20 mM Tris-HCl pH 8.0, 50 mM NaCl, 20 mM imidazole, 1 mM β-mercaptoethanol, 0.015% (v/v) Triton X-100	N4 RNAPII IMAC purification
Ni-NTA wash buffer IV	20 mM Tris-HCl pH 8.0, 50 mM NaCl, 20 mM imidazole, 1 mM β-mercaptoethanol	N4 RNAPII IMAC purification
Ni-NTA wash buffer V	20 mM Tris-HCl pH 8.0, 50 mM NaCl, 50 mM imidazole, 1 mM β-mercaptoethanol	N4 RNAPII IMAC purification
Ni-NTA elution buffer	20 mM Tris-HCl pH 8.0, 50 mM NaCl, 300 mM imidazole, 1 mM β-mercaptoethanol	N4 RNAPII IMAC purification
TMGE + 50 mM NaCl buffer	10 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 5% (v/v) glycerol, 0.1 mM EDTA, 1 mM β-mercaptoethanol, 50 mM NaCl	N4 RNAPII heparin sepharose chromatography purification
TMGE + 100 mM NaCl buffer	10 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 5% (v/v) glycerol, 0.1 mM EDTA, 1 mM β-mercaptoethanol, 100 mM NaCl	N4 RNAPII heparin sepharose chromatography purification
TMGE + 500 mM NaCl buffer	10 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 5% (v/v) glycerol, 0.1 mM EDTA, 1 mM β-mercaptoethanol, 500 mM NaCl	N4 RNAPII heparin sepharose chromatography purification
N4 RNAPII storage buffer	20 mM Tris-HCl pH 8.0, 50 mM NaCl, 50% (v/v) glycerol, 1 mM β-mercaptoethanol	N4 RNAPII protein storage buffer

Table II.2. Buffers (continued)

Buffer	Composition	Purpose
gp2 lysis buffer	20 mM Tris-HCl pH 8.0, 50 mM NaCl, 5 mM DTT, EDTA-free protease inhibitor	gp2 purification
gp2 1 M NaCl buffer	20 mM Tris-HCl pH 8.0, 1 M NaCl, 5 mM DTT	gp2 purification
gp2 detergent buffer	20 mM Tris-HCl pH 8.0, 50 mM NaCl, 5 mM DTT, 0.015% (v/v) Triton X-100	gp2 purification
gp2 50 mM NaCl buffer	20 mM Tris-HCl pH 8.0, 50 mM NaCl, 5 mM DTT	gp2 purification
gp2 denaturing buffer	50 mM Tris-HCl pH 8.0, 50 mM NaCl, 6 M guanidinium chloride	gp2 purification
gp2 denaturing buffer II	50 mM Tris-HCl pH 8.0, 50 mM NaCl, 3 M guanidinium chloride	gp2 purification
gp2 exchange buffer	20 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM EDTA, 30% (v/v) glycerol	gp2 purification
gp2 storage buffer	50 mM Tris-HCl pH 8.0, 50% (v/v) glycerol, 1 mM β-mercaptoethanol	gp2 protein storage buffer
Transcription buffer	10 mM Bis-Tris-Propane-HCl pH 7.0, 10 mM $MgCl_2$, 1 mM DTT, 5% (v/v) glycerol, 100 µg ml^{-1} BSA	*In vitro* runoff transcription reaction buffer
Transcription stop buffer	95% (v/v) formamide, 20 mM EDTA, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol	*In vitro* runoff transcription stop buffer
TME buffer	10 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 0.1 mM EDTA	Nitrocellulose membrane equilibration
Filter binding buffer	12 mM Tris-HCl pH 8.0, 5 mM NaCl, 10 mM $MgCl_2$, 0.1 mM EDTA, 5% (v/v) glycerol, 100 µg ml^{-1} BSA	Filter binding
N4 RNAPII-gp2 interaction buffer	32 mM Tris-HCl pH 8.0, 50 mM NaCl, 35% (v/v) glycerol	N4 RNAPII-gp2 crosslinking and IMAC interaction assays
N4 RNAPII-gp2 interaction buffer II	27 mM Tris-HCl pH 8.0, 50 mM NaCl, 30% (v/v) glycerol	N4 RNAPIIBpa-gp2 crosslinking assays
Low salt buffer	20 mM Tris-HCl pH 8.0, 50 mM NaCl	N4 RNAPII-gp2 IMAC interaction wash buffer
High salt buffer	20 mM Tris-HCl pH 8.0, 1 M NaCl	N4 RNAPII-gp2 IMAC interaction wash buffer
Transfer buffer	24 mM Tris-HCl pH 8.0, 192 mM glycine, 0.04% (w/v) SDS, 20% (v/v) methanol	Western blotting transfer buffer
TBST buffer	20 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.1% (v/v) Tween-20, pH 7.6	Western blotting wash buffer
Blocking buffer	20 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.1% (v/v) Tween-20, 1% (w/v) ECL prime blocking agent, pH 7.6	Western blotting antibody blocking buffer

Table II.2. Buffers (continued)

Buffer	Composition	Purpose
Primary antibody solution	1:2000 dilution rabbit α gp2, 1:3300 dilution mouse α HisG, 20 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.1% (v/v) Tween-20, 1% (w/v) ECL prime blocking agent, pH 7.6	Western blotting primary antibody solution for detection of gp2 and gp15
Secondary antibody solution	1:5000 dilution goat α rabbit Licor 680, 1:2500 dilution goat α mouse Licor 800, 20 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.1% (v/v) Tween-20, 1% (w/v) ECL prime blocking agent, pH 7.6	Western blotting secondary antibody solution for detection of gp2 and gp15

Table II.3. Oligonucleotides for cloning and mutagenesis

Primer	Sequence 5' to 3'	Purpose
BRL_13	GCGTCACACTTTGCTATGCC	Sequencing MCS region of pBAD vector
BRL_14	GCTTCTGCGTTCTGATTTAATCTG	Sequencing MCS region of pBAD vector
BRL_27	GTGCGTCCTAAGCTGTTAACTAAG	Sequencing N4 RNAPII gene constructs
BRL_28	GCCCTGTTCTATGCTGGTGTTA	Sequencing N4 RNAPII gene constructs
BRL_29	CGTCGTTGTGACTATGACAAGAACC	Sequencing N4 RNAPII gene constructs
BRL_30.1	CGTATATATCAAAGTAATGGTAAATTAGG TGGAGACTGTCCACTTCTTAGATAAGCC	E481am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_31.1	GGCTTATCTAAGAAGTGGACAGTCTCCAC CTAATTTACCATTACTTTGATATATACG	E481am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_32	GTATATATCAAAGTAATGGTAAATGAAGT GTAGACTGTCCACTTCTTAGATAAGCC	E483am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_33	GGCTTATCTAAGAAGTGGACAGTCTACAC TTCATTTACCATTACTTTGATATATAC	E483am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_34	GGAGACTGTCCACTTCTTAGATAAGCCAT AGGACTGTGTTCGTAAAGTACAGGG	Y492am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_35	CCCTGTACTTTACGAACACAGTCCTATGGC TTATCTAAGAAGTGGACAGTCTCC	Y492am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_36	GACTGTCCACTTCTTAGATAAGCCATATGA CTAGGTTCGTAAAGTACAGGGTACTG	C494am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_37	CAGTACCCTGTACTTTACGAACCTAGTCAT ATGGCTTATCTAAGAAGTGGACAGTC	C494am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_38	GATAAGCCATATGACTGTGTTTAGAAAGT ACAGGGTACTGAAGAGAAGACTCGTATGC	R496am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_39	GCATACGAGTCTTCTCTTCAGTACCCTGTA CTTTCTAAACACAGTCATATGGCTTATC	R496am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_45	GGAGACTGTCCACTTCTTAGATAAGCCAT ATTAGTGTGTTCGTAAAGTACAGGG	D493am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_46	CCCTGTACTTTACGAACACACTAATATGGC TTATCTAAGAAGTGGACAGTCTCC	D493am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_41	GACTGTCCACTTCTTAGATAAGCCATATGA CAGCGTTCGTAAAGTACAGGGTACTG	C494S N4 RNAPII Quikchange mutagenesis of pAD1
BRL_42	CAGTACCCTGTACTTTACGAACGCTGTCAT ATGGCTTATCTAAGAAGTGGACAGTC	C494S N4 RNAPII Quikchange mutagenesis of pAD1
BRL_43	GGAGACTGTCCACTTCTTAGATAAGCCATT TGACTGTGTTCGTAAAGTACAGGG	Y492F N4 RNAPII Quikchange mutagenesis of pAD1
BRL_44	CCCTGTACTTTACGAACACAGTCAAATGG CTTATCTAAGAAGTGGACAGTCTCC	Y492F N4 RNAPII Quikchange mutagenesis of pAD1

Table II.3. Oligonucleotides for cloning and mutagenesis (continued)

Primer	Sequence 5' to 3'	Purpose
BRL_166	GGCTAACAGGAGGAATTAACCATGGCTTA GACTACTTTTGCTAAA	I3am gp2 Quikchange mutagenesis of pOD9
BRL_167	CTGACCAAAGGTTTTAGCAAAAGTAGTCT AAGCCATGGTTAATTC	I3am gp2 Quikchange mutagenesis of pOD9
BRL_168	GGAGGAATTAACCATGGCTATCACTACTT AGGCTAAAACCTTTGGT	F6am gp2 Quikchange mutagenesis of pOD9
BRL_169	AGTAGAAGCCTGACCAAAGGTTTTAGCCT AAGTAGTGATAGCCAT	F6am gp2 Quikchange mutagenesis of pOD9
BRL_274	CGAACATCAGATGCACCTCGAAGCCCTGT ACAACAAGAACCAA	K12A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_275	ACAGTTGGTTCTTGTTGTACAGGGCTTCGA GGTGCATCTGATG	K12A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_276	AAAGCTGTACAACAAGAACCAAGCGTTAC CAAGAATGCGTCAG	L19A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_277	ACTCCTGACGCATTCTTGGTAACGCTTGGT TCTTGTTGTACAG	L19A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_280	GTACAACAAGAACCAACTGTTAGCGAGAA TGCGTCAGGAGTTT	P21A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_281	CTTCAAACTCCTGACGCATTCTCGCTAACA GTTGGTTCTTGTT	P21A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_282	GACTCATCGTCCGGATAAACGTGCGCGTA CCTATTCTCAGGGA	G239A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_283	GGTATCCCTGAGAATAGGTACGCGCACGT TTATCCGGACGATG	G239A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_284	TAAGAATGAGAACAACCTACTGGCGCTTG TAAGAGAAGCTGAA	N310A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_285	GTTCTTCAGCTTCTCTTACAAGCGCCAGTA GGTTGTTCTCATT	N310A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_286	GAATGAGAACAACCTACTGAACGCGGTAA GAGAAGCTGAAGAA	L311A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_287	CTGGTTCTTCAGCTTCTCTTACCGCGTTCA GTAGGTTGTTCTC	L311A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_288	TGAGAACAACCTACTGAACCTTGCGAGAG AAGCTGAAGAACCA	V312A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_289	TGGTTCTTCAGCTTCTCTCGCAAGGTTCAG TAGGTTGTTCTCA	V312A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_290	CGAACATCAGATGCACCTCGAATAGCTGT ACAACAAGAACCAA	K12am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_291	ACAGTTGGTTCTTGTTGTACAGCTATTCGA GGTGCATCTGATG	K12am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_292	AAAGCTGTACAACAAGAACCAATAGTTAC CAAGAATGCGTCAG	L19am N4 RNAPII Quikchange mutagenesis of pAD1

Table II.3. Oligonucleotides for cloning and mutagenesis (continued)

Primer	Sequence 5' to 3'	Purpose
BRL_293	ACTCCTGACGCATTCTTGGTAACTATTGGTTCTTGTTGTACAG	L19am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_294	GCTGTACAACAAGAACCAACTGTAGCCAAGAATGCGTCAGGAG	L20am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_295	CAAACTCCTGACGCATTCTTGGCTACAGTTGGTTCTTGTTGTA	L20am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_296	GTACAACAAGAACCAACTGTTATAGAGAATGCGTCAGGAGTTT	P21am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_297	CTTCAAACTCCTGACGCATTCTCTATAACAGTTGGTTCTTGTT	P21am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_298	GACTCATCGTCCGGATAAACGTTAGCGTACCTATTCTCAGGGA	G239am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_299	GGTATCCCTGAGAATAGGTACGCTAACGTTTATCCGGACGATG	G239am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_300	TAAGAATGAGAACAACCTACTGTAGCTTGTAAGAGAAGCTGAA	N310am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_301	GTTCTTCAGCTTCTCTTACAAGCTACAGTAGGTTGTTCTCATT	N310am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_302	GAATGAGAACAACCTACTGAACTAGGTAAGAGAAGCTGAAGAA	L311am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_303	CTGGTTCTTCAGCTTCTCTTACCTAGTTCAGTAGGTTGTTCTC	L311am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_304	TGAGAACAACCTACTGAACCTTTAGAGAGAAGCTGAAGAACCA	V312am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_305	TGGTTCTTCAGCTTCTCTCTAAAGGTTCAGTAGGTTGTTCTCA	V312am N4 RNAPII Quikchange mutagenesis of pAD1
BRL_308	GCTGTACAACAAGAACCAACTGGCACCAAGAATGCGTCAGGAG	L20A N4 RNAPII Quikchange mutagenesis of pAD1
BRL_309	CAAACTCCTGACGCATTCTTGGTGCCAGTTGGTTCTTGTTGTA	L20A N4 RNAPII Quikchange mutagenesis of pAD1

Table II.4. N4 RNAPII transcription template oligonucleotides

Template	Sequence 5' to 3'	Description
BL-4	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAACACTCACATTACTCTGTAGGG	-10C Pm5 template strand (-28 to +37)
BL-5	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAACAGTCACATTACTCTGTAGGG	-10G Pm5 template strand (-28 to +37)
BL-6	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAACATTCACATTACTCTGTAGGG	-10T Pm5 template strand (-28 to +37)
BL-7	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAACCATCACATTACTCTGTAGGG	-9C Pm5 template strand (-28 to +37)
BL-8	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAACGATCACATTACTCTGTAGGG	-9G Pm5 template strand (-28 to +37)
BL-9	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAACTATCACATTACTCTGTAGGG	-9T Pm5 template strand (-28 to +37)
BL-10	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAAAAATCACATTACTCTGTAGGG	-8A Pm5 template strand (-28 to +37)
BL-11	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAAGAATCACATTACTCTGTAGGG	-8G Pm5 template strand (-28 to +37)
BL-12	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAATAATCACATTACTCTGTAGGG	-8T Pm5 template strand (-28 to +37)
BL-13	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAACCAATCACATTACTCTGTAGGG	-7C Pm5 template strand (-28 to +37)
BL-14	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAGCAATCACATTACTCTGTAGGG	-7G Pm5 template strand (-28 to +37)
BL-15	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAATCAATCACATTACTCTGTAGGG	-7T Pm5 template strand (-28 to +37)
BL-16	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAACACAATCACATTACTCTGTAGGG	-6C Pm5 template strand (-28 to +37)
BL-17	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAGACAATCACATTACTCTGTAGGG	-6G Pm5 template strand (-28 to +37)
BL-18	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAATACAATCACATTACTCTGTAGGG	-6T Pm5 template strand (-28 to +37)
BL-19	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAACAACAATCACATTACTCTGTAGGG	-5C Pm5 template strand (-28 to +37)
BL-20	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAGAACAATCACATTACTCTGTAGGG	-5G Pm5 template strand (-28 to +37)
BL-21	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAATAACAATCACATTACTCTGTAGGG	-5T Pm5 template strand (-28 to +37)
BL-22	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATACAAACAATCACATTACTCTGTAGGG	-4C Pm5 template strand (-28 to +37)
BL-23	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAGAAACAATCACATTACTCTGTAGGG	-4G Pm5 template strand (-28 to +37)
BL-24	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATATAAACAATCACATTACTCTGTAGGG	-4T Pm5 template strand (-28 to +37)

Table II.4. N4 RNAPII transcription template oligonucleotides (continued)

Template	Sequence 5' to 3'	Description
BL-25	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATCAAAACAATCACATTACTCTGTAGGG	-3C Pm5 template strand (-28 to +37)
BL-26	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATGAAAACAATCACATTACTCTGTAGGG	-3G Pm5 template strand (-28 to +37)
BL-27	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATTAAAACAATCACATTACTCTGTAGGG	-3T Pm5 template strand (-28 to +37)
BL-28	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCAAAAAAACAATCACATTACTCTGTAGGG	-2A Pm5 template strand (-28 to +37)
BL-29	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCACAAAAACAATCACATTACTCTGTAGGG	-2C Pm5 template strand (-28 to +37)
BL-30	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCAGAAAAACAATCACATTACTCTGTAGGG	-2G Pm5 template strand (-28 to +37)
BL-31	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCCTAAAAACAATCACATTACTCTGTAGGG	-1C Pm5 template strand (-28 to +37)
BL-32	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCGTAAAAACAATCACATTACTCTGTAGGG	-1G Pm5 template strand (-28 to +37)
BL-33	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCTTAAAAACAATCACATTACTCTGTAGGG	-1T Pm5 template strand (-28 to +37)
BL-34	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTAATAAAAACAATCACATTACTCTGTAGGG	+1A Pm5 template strand (-28 to +37)
BL-35	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTGATAAAAACAATCACATTACTCTGTAGGG	+1G Pm5 template strand (-28 to +37)
BL-36	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTTATAAAAACAATCACATTACTCTGTAGGG	+1T Pm5 template strand (-28 to +37)
BL-37	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATACATAAAAACAATCACATTACTCTGTAGGG	+2A Pm5 template strand (-28 to +37)
BL-38	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATCCATAAAAACAATCACATTACTCTGTAGGG	+2C Pm5 template strand (-28 to +37)
BL-39	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATGCATAAAAACAATCACATTACTCTGTAGGG	+2G Pm5 template strand (-28 to +37)
BL-40	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAAACAATCACATTACTCTGTAGGG	WT Pm5 template strand (-28 to +37)
BL-41	CCCTACAGAGTAATGTGAACAACAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	WT Pm5 non-template strand (-10 to +4)
BL-44	CCCTACAGAGTAATGTGACCAACAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-10C Pm5 non-template bubble (-10 to +4)
BL-45	CCCTACAGAGTAATGTGAGCAACAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-10G Pm5 non-template bubble (-10 to +4)
BL-46	CCCTACAGAGTAATGTGATCAACAAACAATTCG TAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-10T Pm5 non-template bubble (-10 to +4)
BL-47	CCCTACAGAGTAATGTGAAAAACAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-9A Pm5 non-template bubble (-10 to +4)

Table II.4. N4 RNAPII transcription template oligonucleotides (continued)

Template	Sequence 5' to 3'	Description
BL-48	CCCTACAGAGTAATGTGAAGAACAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-9G Pm5 non-template bubble (-10 to +4)
BL-49	CCCTACAGAGTAATGTGAATAACAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-9T Pm5 non-template bubble (-10 to +4)
BL-50	CCCTACAGAGTAATGTGAACAACCAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-5C Pm5 non-template bubble (-10 to +4)
BL-51	CCCTACAGAGTAATGTGAACAACGAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-5G Pm5 non-template bubble (-10 to +4)
BL-52	CCCTACAGAGTAATGTGAACAACTAACAATTCG TAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-5T Pm5 non-template bubble (-10 to +4)
BL-53	CCCTACAGAGTAATGTGAACCACAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-8C Pm5 non-template bubble (-10 to +4)
BL-54	CCCTACAGAGTAATGTGAACGACAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-8G Pm5 non-template bubble (-10 to +4)
BL-55	CCCTACAGAGTAATGTGAACTACAAACAATTCG TAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-8T Pm5 non-template bubble (-10 to +4)
BL-56	CCCTACAGAGTAATGTGAACACCAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-7C Pm5 non-template bubble (-10 to +4)
BL-57	CCCTACAGAGTAATGTGAACAGCAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-7G Pm5 non-template bubble (-10 to +4)
BL-58	CCCTACAGAGTAATGTGAACATCAAACAATTCG TAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-7T Pm5 non-template bubble (-10 to +4)
BL-59	CCCTACAGAGTAATGTGAACAAAAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-6A Pm5 non-template bubble (-10 to +4)
BL-60	CCCTACAGAGTAATGTGAACAAGAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-6G Pm5 non-template bubble (-10 to +4)
BL-61	CCCTACAGAGTAATGTGAACAATAAACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-6T Pm5 non-template bubble (-10 to +4)
BL-62	CCCTACAGAGTAATGTGAACAACACACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-4C Pm5 non-template bubble (-10 to +4)
BL-63	CCCTACAGAGTAATGTGAACAACAGACAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-4G Pm5 non-template bubble (-10 to +4)
BL-64	CCCTACAGAGTAATGTGAACAACATACAATTCG TAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-4T Pm5 non-template bubble (-10 to +4)
BL-65	CCCTACAGAGTAATGTGAACAACAACCAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-3C Pm5 non-template bubble (-10 to +4)
BL-66	CCCTACAGAGTAATGTGAACAACAAGCAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-3G Pm5 non-template bubble (-10 to +4)
BL-67	CCCTACAGAGTAATGTGAACAACAATCAATTCG TAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-3T Pm5 non-template bubble (-10 to +4)
BL-68	CCCTACAGAGTAATGTGAACAACAAAAAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-2A Pm5 non-template bubble (-10 to +4)

Table II.4. N4 RNAPII transcription template oligonucleotides (continued)

Template	Sequence 5' to 3'	Description
BL-69	CCCTACAGAGTAATGTGAACAACAAAGAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-2G Pm5 non-template bubble (-10 to +4)
BL-70	CCCTACAGAGTAATGTGAACAACAAATAATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-2T Pm5 non-template bubble (-10 to +4)
BL-71	CCCTACAGAGTAATGTGAACAACAAACCATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-1C Pm5 non-template bubble (-10 to +4)
BL-72	CCCTACAGAGTAATGTGAACAACAAACGATTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-1G Pm5 non-template bubble (-10 to +4)
BL-73	CCCTACAGAGTAATGTGAACAACAAACTATTCG TAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	-1T Pm5 non-template bubble (-10 to +4)
BL-74	CCCTACAGAGTAATGTGAACAACAAACACTTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	+1C Pm5 non-template bubble (-10 to +4)
BL-75	CCCTACAGAGTAATGTGAACAACAAACAGTTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	+1G Pm5 non-template bubble (-10 to +4)
BL-76	CCCTACAGAGTAATGTGAACAACAAACATTTCG TAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	+1T Pm5 non-template bubble (-10 to +4)
BL-77	CCCTACAGAGTAATGTGAACAACAAACAAATC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	+2A Pm5 non-template bubble (-10 to +4)
BL-78	CCCTACAGAGTAATGTGAACAACAAACAACTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	+2C Pm5 non-template bubble (-10 to +4)
BL-79	CCCTACAGAGTAATGTGAACAACAAACAAGTC GTAGCCCTCCTTGTGAGCCGTTTTTAGTATCCC	+2G Pm5 non-template bubble (-10 to +4)
BL-84	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCAAGAAAACAATCACATTACTCTGTAGGG	-2A,-3G Pm5 template strand (-28 to +37)
BL-85	GGGATACTAAGGACGGCTCACAAGAGAAGCTA CATTCCTCAACGCAATCACATTACTCTGTAGGG	-1C,-3C,-6C,-7G Pm5 template strand (-28 to +37)
BL-86	GGGATACTAAGGACGGCTCACAAGAGAAGCTA CATTCCTCAAAACAATCACATTACTCTGTAGGG	-1C,-3C Pm5 template strand (-28 to +37)
BL-87	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCTTTAAAACAATCACATTACTCTGTAGGG	-1T,-3T Pm5 template strand (-28 to +37)
BL-88	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAAGGCAATCACATTACTCTGTAGGG	-6G,-7G Pm5 template strand (-28 to +37)
BL-89	GGGATACTAAGGACGGCTCACAAGAGAAGCTA CATTCCAAAAAACAATCACATTACTCTGTAGGG	-1C,-2A Pm5 template strand (-28 to +37)
BL-90	GGGATACTAAGGACGGCTCACAAGAGAAGCTA CATTCCTGAAAACAATCACATTACTCTGTAGGG	-1C,-3G Pm5 template strand (-28 to +37)
BL-91	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCAACAAAACAATCACATTACTCTGTAGGG	-2A,-3C Pm5 template strand (-28 to +37)
BL-92	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCAAAAACACAATCACATTACTCTGTAGGG	-6C,-2A Pm5 template strand (-28 to +37)
BL-93	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATGAACACAATCACATTACTCTGTAGGG	-6C,-3G Pm5 template strand (-28 to +37)

Table II.4. N4 RNAPII transcription template oligonucleotides (continued)

Template	Sequence 5' to 3'	Description
BL-94	GGGATACTAAAAACGGCTCACAAGGAGGGTCA CATTCAAAAAAGCAATCACATTACTCTGTAGGG	-7G,-2A Pm5 template strand (-28 to +37)
BL-95	GGGATACTAAAAACGGCTCACAAGGAGGGTCA CATTCATGAAAGCAATCACATTACTCTGTAGGG	-7G,-3G Pm5 template strand (-28 to +37)
BL-99	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCTTCAAAACAATCACATTACTCTGTAGGG	-1T,-3C Pm5 template strand (-28 to +37)
BL-100	GGGATACTAAGGACGGCTCACAAGAGAAGCTA CATTCCTTAAAACAATCACATTACTCTGTAGGG	-1C,-3T Pm5 template strand (-28 to +37)
BL-101	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCTTTAAGGCAATCACATTACTCTGTAGGG	-1T,-3T,-6G,-7G Pm5 template strand (-28 to +37)
BL-102	GGGATACTAAGGACGGCTCACAAGAGAAGCTA CATTCCTCAAGGCAATCACATTACTCTGTAGGG	-1C,-3C,-6G,-7G Pm5 template strand (-28 to +37)
BL-103	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCTTTAATTCAATCACATTACTCTGTAGGG	-1T,-3T,-6T,-7T Pm5 template strand (-28 to +37)
BL-104	GGGATACTAAGGACGGCTCACAAGAGAAGCTA CATTCCTCAACCCAATCACATTACTCTGTAGAG	-1C,-3C,-6C,-7C Pm5 template strand (-28 to +37)
BL-105	GGGATACTAAGGACGGCTCACAAGAGAAGCTA CATTCATAAACCCAATCACATTACTCTGTAGAG	-6C,-7C Pm5 template strand (-28 to +37)
BL-106	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAATTCAATCACATTACTCTGTAGGG	-6T,-7T Pm5 template strand (-28 to +37)
BL-107	GGGATACTAAAAACGGCTCACAAGGAGGGCTA CATTCATAAACGCAATCACATTACTCTGTAGGG	-6C,-7G Pm5 template strand (-28 to +37)
BL-108	GGGATACTAAGGACGCGTCACAAGAGAACGTA CATTCATAAAGCCAATCACATTACTCTGTAGGG	-6G,-7C Pm5 template strand (-28 to +37)
BL-111	GGGATACTAAAAACGGGTCACAACCATCACTA CATTCATAAAAACAATCACATTACTCTGTAGGG	Pm5 20 nt pause template strand (-28 to +37)
BL-121	ATCATAAAAACAATCACATTACTCTGTAGGG	+3A Pm5 template strand (-28 to +3)
BL-248	TAGCCAATTCACTTGGTTCTGAACCATCATGTA AGTCATAAATAAATGAAATTACATAAT	N4 Mc fragment template strand
CM-495	CATCGTCCGGATGCTCGTGGTCGTACC	27 nt ssDNA loading control

Table II.5. Plasmids

Plasmid	Description	Purpose	Source
pBAD-HisB	pBAD-HisB expression vector	protein expression	Invitrogen
pSupT/BpF	Expression of BPA synthetase/tRNA	incorporation of *p*Bpa at amber codon	Courtesy of Gourse lab
pOD9	pBAD-HisB[gp2]	Expression of untagged gp2	O. Demidenko
pAD1	pBAD-HisB[N4 RNAPII]	Expression of N-term His_6+leader-tagged N4 RNAPII	O. Demidenko
pCAM1	pAD1[K476A N4 RNAPII]	Expression of K476A N4 RNAPII	C. Markle
pCAM2	pAD1[N480A N4 RNAPII]	Expression of N480A N4 RNAPII	C. Markle
pCAM3	pAD1[E481A N4 RNAPII]	Expression of E481A N4 RNAPII	C. Markle
pCAM4	pAD1[E483A N4 RNAPII]	Expression of E483A N4 RNAPII	C. Markle
pCAM5	pAD1[T484A N4 RNAPII]	Expression of T484A N4 RNAPII	C. Markle
pCAM6	pAD1[H486A N4 RNAPII]	Expression of H486A N4 RNAPII	C. Markle
pCAM7	pAD1[D489A N4 RNAPII]	Expression of D489A N4 RNAPII	C. Markle
pCAM8	pAD1[K490A N4 RNAPII]	Expression of K490A N4 RNAPII	C. Markle
pCAM9	pAD1[Y492A N4 RNAPII]	Expression of Y492A N4 RNAPII	C. Markle
pCAM10	pAD1[D493A N4 RNAPII]	Expression of D493A N4 RNAPII	C. Markle
pCAM11	pAD1[C494A N4 RNAPII]	Expression of C494A N4 RNAPII	C. Markle
pCAM12	pAD1[R496A N4 RNAPII]	Expression of R496A N4 RNAPII	C. Markle
pCAM13	pAD1[K497A N4 RNAPII]	Expression of K497A N4 RNAPII	C. Markle
pCAM14	pAD1[Q499A N4 RNAPII]	Expression of Q499A N4 RNAPII	C. Markle
pCAM15	pAD1[G500A N4 RNAPII]	Expression of G500A N4 RNAPII	C. Markle
pCAM16	pAD1[T501A N4 RNAPII]	Expression of T501A N4 RNAPII	C. Markle
pCAM17	pAD1[E502A N4 RNAPII]	Expression of E502A N4 RNAPII	C. Markle
pCAM18	pOD9[I3A gp2]	Expression of I3A gp2	C. Markle
pCAM19	pOD9[W30A gp2]	Expression of W30A gp2	C. Markle
pBRL1	pAD1[E481am N4 RNAPII]	Expression of E481Bpa N4 RNAPII	This work
pBRL2	pAD1[E483am N4 RNAPII]	Expression of E483Bpa N4 RNAPII	This work
pBRL3	pAD1[Y492am N4 RNAPII]	Expression of Y492Bpa N4 RNAPII	This work
pBRL4	pAD1[D493am N4 RNAPII]	Expression of D493Bpa N4 RNAPII	This work
pBRL5	pAD1[C494am N4 RNAPII]	Expression of C494Bpa N4 RNAPII	This work
pBRL6	pAD1[R496am N4 RNAPII]	Expression of R496Bpa N4 RNAPII	This work
pBRL7	pAD1[C494S N4 RNAPII]	Expression of C494S N4 RNAPII	This work
pBRL8	pAD1[Y492F N4 RNAPII]	Expression of Y492F N4 RNAPII	This work
pBRL9	pAD1[K12A N4 RNAPII]	Expression of K12A N4 RNAPII	This work
pBRL10	pAD1[L19A N4 RNAPII]	Expression of L19A N4 RNAPII	This work
pBRL11	pAD1[L20A N4 RNAPII]	Expression of L20A N4 RNAPII	This work

Table II.5. Plasmids (continued)

Plasmid	Description	Purpose	Source
pBRL12	pAD1[P21A N4 RNAPII]	Expression of P21A N4 RNAPII	This work
pBRL13	pAD1[G239A N4 RNAPII]	Expression of G239A N4 RNAPII	This work
pBRL14	pAD1[N310A N4 RNAPII]	Expression of N310A N4 RNAPII	This work
pBRL15	pAD1[L311A N4 RNAPII]	Expression of L311A N4 RNAPII	This work
pBRL16	pAD1[V312A N4 RNAPII]	Expression of V312A N4 RNAPII	This work
pBRL17	pAD1[K12am N4 RNAPII]	Expression of K12Bpa N4 RNAPII	This work
pBRL18	pAD1[L19am N4 RNAPII]	Expression of L19Bpa N4 RNAPII	This work
pBRL19	pAD1[L20am N4 RNAPII]	Expression of L20Bpa N4 RNAPII	This work
pBRL20	pAD1[P21am N4 RNAPII]	Expression of P21Bpa N4 RNAPII	This work
pBRL21	pAD1[G239am N4 RNAPII]	Expression of G239Bpa N4 RNAPII	This work
pBRL22	pAD1[N310am N4 RNAPII]	Expression of N310Bpa N4 RNAPII	This work
pBRL23	pAD1[L311am N4 RNAPII]	Expression of L311Bpa N4 RNAPII	This work
pBRL24	pAD1[V312am N4 RNAPII]	Expression of V312Bpa N4 RNAPII	This work
pBRL25	pOD9[I3am gp2]	Expression of I3Bpa gp2	This work
pBRL26	pOD9[F6am gp2]	Expression of F6Bpa gp2	This work

CHAPTER III: ANNOTATION OF THE BACTERIOPHAGE N4 GENOME AND COMPARATIVE GENOMICS OF N4-LIKE PHAGES

INTRODUCTION

The bacteriophage N4 genome was previously sequenced using dideoxy dye terminator chemistry from a library of sheared DNA fragments and assembled in collaboration with members of the Pittsburgh Bacteriophage Institute. The N4 genome contains 70,153 bp of linear double-stranded DNA with a G+C content of 41.3%, non-permuted direct terminal repeats 390-440 bp in length, 73 predicted ORFs, and 4 tRNA genes (213, 243, 244). Excluding the last three ORFs (ORFs 70-72), the genome divides into two halves according to the direction of transcription. Early and middle transcripts map to the left end of the genome with rightward polarity, while late transcripts map to the right half of the genome with leftward polarity (25). Through extensive genetic, biochemical, structural, and bioinformatics studies, the Rothman-Denes lab had previously identified and characterized 20 gene products responsible for the hallmark properties of N4, while the functions of the remaining 53 predicted N4 ORFs have yet to be elucidated (Table III.1).

Bacteriophage N4 has three notable phenotypes: i) a reversed transcriptional program utilizing a virion-encapsidated RNAP for host-independent early transcription (vRNAP), a second N4-encoded heterodimeric RNAP for middle transcription (N4 RNAPII), and the host σ^{70}-RNAP directed to N4 late promoters by N4SSB; ii) delayed lysis and enlargement of host cells; iii) a large burst size of 3,000 pfu per infected cell (25, 31, 32). Until recently, these properties were completely unique to N4, which was considered a phenotypic and genomic orphan for over 40 years (245). Relatives of N4 were first discovered in 2008 with the isolation

of two marine Roseobacter phages, DSS3P2 and EE36P1, with similar morphology and encoding homologs to the well-characterized N4 RNAPs (246). Due to the advancement of next generation sequencing techniques and the increased appreciation for the role that phages play in the biosphere, the number of viral sequences in available databases has exploded over the past 10 years (39–43). Through these efforts, dozens more N4-like phages have since been discovered and described, although there is neither a clear definition for subfamily membership nor a disciplined methodology for the determination of phylogenetic relationships within this group of phages.

In this chapter, I aimed to leverage the influx of phage sequences in available databases to discover putative functions of unclassified N4 ORFs, provide a framework for the classification of newly sequenced or isolated phages as N4-like, and elucidate shared physiological strategies for host interaction and takeover. I have identified putative functions for an additional 12 N4 ORFs using bioinformatics approaches and utilized N4 vRNAP as a marker gene to assign 55 total phages to the N4-like phage subfamily. N4-like phages share similar genomic and morphological properties despite infecting a broad range of Proteobacteria. Comparative genomics of N4-like phages revealed that phages infecting closely related hosts share greater degrees of genetic similarity and identified the N4 transcriptional machinery as the hallmark of N4-like phages due to its conservation across the N4-like phage subfamily.

ANNOTATION OF THE BACTERIOPHAGE N4 GENOME

To define putative functions for the 53 uncharacterized N4 ORFs, I utilized a variety of bioinformatics tools to predict conserved domains and identify homologs of all N4 ORFs. Putative functions were assigned to 12 N4 ORFs based on the identification of a phage protein

homolog through BLAST search or the identification of a conserved domain of known function (E-value cutoff of 1.0E-5) through Interpro and the Conserved Domains Database analyses. The updated annotations for all N4 ORFs are summarized in Table III.1.

Table III.1. N4 ORF annotations

		Conserved Domains			
ORF	**Annotation**	**Name**	**Accession**	**Interval**	**E-value**
1	genome injection factor[a]				
2	N4 RNAPII activator[a]				
3	gp3				
4	gp4				
5	gp5				
6	cell division inhibitor[a]				
7	gp7				
8	host DNA replication inhibitor[a]				
9	gp9				
10	gp10				
11	gp11				
12	gp12	DUF4326	pfam14216	17-106	3.31E-33
13	gp13				
14	gp14				
15	N4 RNAPII subunit 1[a]	PHA00452	PHA00452	109-269	7.20E-07
16	N4 RNAPII subunit 2[a]	PHA00452	PHA00452	9-260	4.32E-18
17	capsid decorating[a]				
18	superinfection exclusion[a]				
19	gp19				
20	putative HflC protease modulator[c]	SPFH_prohibitin	cd03401	25-224	4.38E-20
21	gp21				
22	putative HNH homing endonuclease[c]	HNH_3	pfam13392	64-106	2.87E-16
23	gp23				
24	putative AAA+ ATPase[c]	AAA	pfam13392	13-164	3.48E-05
25	putative metallopeptidase[c]	DUF2201_N	pfam13203	9-245	1.03E-94
26	dCTP deaminase[b]	Dcd	COG0717	26-157	7.52E-25
27	gp27				
28	gp28				
29	gp29				
30	putative thymidylate synthase[c]	ThyX	COG1351	26-220	3.06E-06
31	gp31				
32	gp32				
33	putative rIIA-like protein[b]	HATPase_c_3	pfam13589	54-168	9.62E-06
34	putative rIIB-like protein[b]				
35	gp35				
36	putative NTP pyrophosphohydrolase[c]	PRA-PH	pfam01503	58-118	3.88E-10

Table III.1. N4 ORF annotations (continued)

		Conserved Domains			
ORF	**Annotation**	**Name**	**Accession**	**Interval**	**E-value**
37	putative DNA helicase[b]	RecD	COG0507	74-429	1.19E-19
39	DNA polymerase[a]	DNA_pol_A	pfam00476	421-809	3.47E-38
40	gp40				
41	gp41				
42	gp42				
43	putative DNA primase[c]	PriCT_1	pfam08708	651-705	4.51E-05
44	putative AAA ATPase[c]	AAA_24	pfam13479	15-162	8.66E-06
45	N4SSB[a]				
46	gp46				
47	gp47				
48	gp48	Phage_gp49_66	pfam13876	3-85	1.97E-37
49	gp49	DUF2829	pfam11195	82-159	6.42E-27
50	virion RNA polymerase[a]				
51	gp51				
52	gp52				
53	gp53				
54	gp54				
55	gp55				
56	major capsid[a]	capsid_maj_N4	TIGR04387	21-299	1.28E-87
57	gp57				
58	gp58				
59	portal[a]				
60	putative Rz-like spanin[c]				
60’	putative Rz1-like spanin[c]				
61	N-acetylmuramidase[a]	Glyco_hydro_108	pfam05838	26-110	1.70E-24
62	putative holin[c]				
63	gp63				
64	gp64				
65	tail sheath[a]				
66	tail appendage[a]				
67	gp67				
68	putative terminase[c]				
69	gp69				
70	gp70				
71	gp71				
72	gp72				

a, previous functional annotation; b, previous bioinformatics annotation; c, updated bioinformatics annotation

Analysis of the N4 genome by ARAGORN revealed that N4 encodes four putative tRNA genes located in the gene-poor region between ORFs 47 and 48: tRNAs TGG (Pro), TGT (Thr), TGA (Ser), and GTT (Asn) (Figure III.1). These four tRNA genes are predicted to be expressed from a single transcript originating from the Pm17 promoter located 3.7 kb upstream during middle transcription. To test whether phage-encoded tRNAs are under selection to ameliorate the differences in codon bias between phage and host ORFs, I calculated the normalized synonymous codon usage (NSCU) for each N4 and *E. coli* codon (Figure III.2). N4 exhibits a codon usage bias greater than that observed in its *E. coli* host for 23 out of 61 codons, including 10 codons for which the N4 bias is more than 2-fold higher. The N4 relative codon usage was consistent across early, middle, and late gene products including the highly expressed virion proteins (data not shown). These results suggest that N4 codon usage is not fully optimized for translation in *E. coli* K-12 host strains. If N4-encoded tRNAs compensate for differential codon usage between phage and host genes, codons for which N4 encodes a tRNA with a complementary anticodon are expected to be among the most differentially biased. Pro (CCA) codon usage is significantly more biased (greater than 2-fold) in N4 ORFs relative to *E. coli* hosts, while the remaining codons for which N4 encodes a complementary tRNA match the *E. coli* codon bias (Figure III.2). These data suggest that N4 tRNAs, with the possible exception of the tRNA-Pro (TGG), do not provide a selective advantage for N4 translation in *E. coli* hosts.

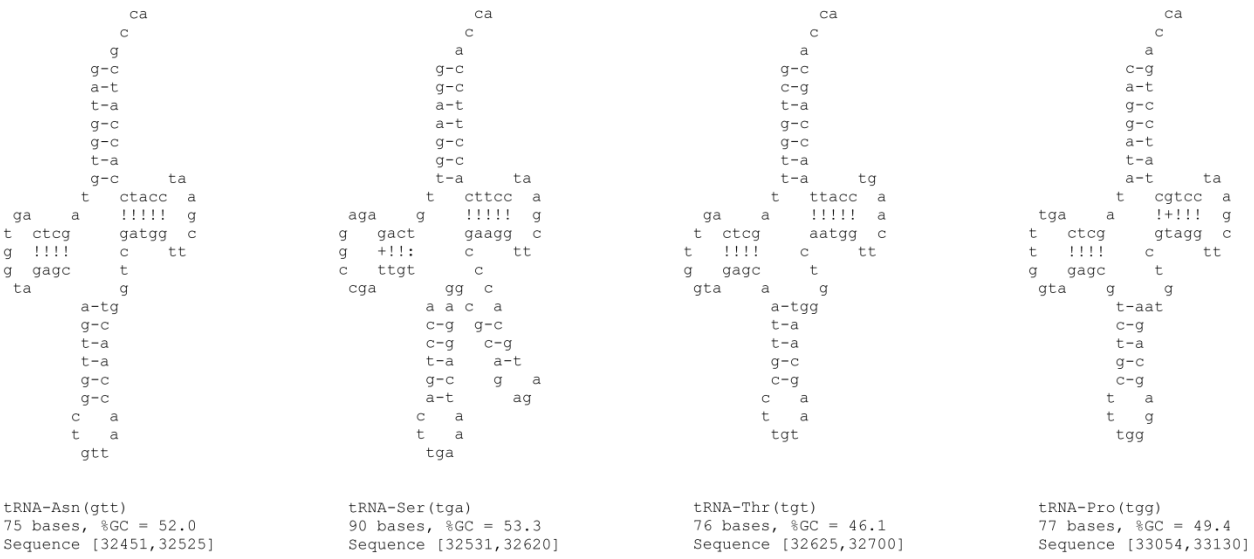

Figure III.1. Detection of four N4-encoded tRNAs. Analysis of the N4 genome by ARAGORN detected four putative tRNA genes: tRNAs TGG (Pro), TGT (Thr), TGA (Ser), and GTT (Asn).

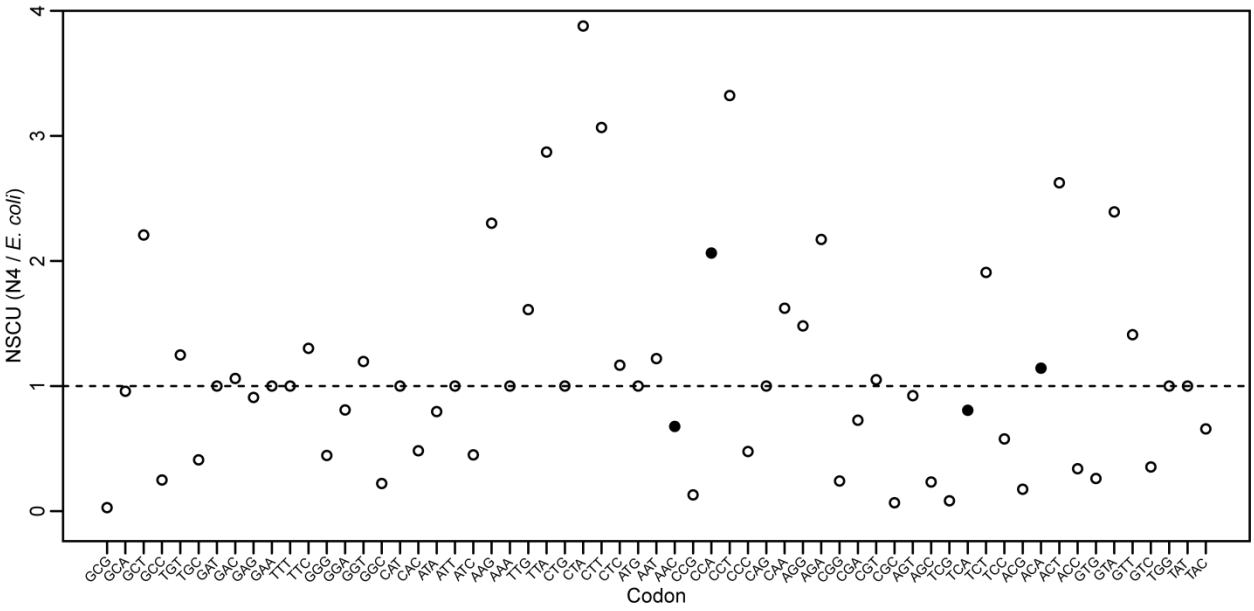

Figure III.2. N4-encoded tRNAs are not under selection to offset codon usage bias. The relative usage of codons in N4 vs. *E. coli* ORFs, measured by normalized synonymous codon usage (NSCU), is plotted for each codon. Codons for which N4 encodes a tRNA with a complementary anticodon are indicated as filled circles. Dashed line represents equal bias between N4 and *E. coli* codons.

IDENTIFICATION OF N4-LIKE PHAGES USING N4 vRNAP AS A MARKER GENE

Due to its activity across numerous N4 physiological processes (genome injection, early transcription, and virion assembly), vRNAP is considered the hallmark of bacteriophage N4 and may serve as a suitable marker gene for the classification of N4-like phages. To identify N4 relatives in available databases, I searched for N4 vRNAP homologs in the NCBI non-redundant protein sequence database by DELTA-BLAST homology search. 54 total phage species encode a homologous ORF with alignment spanning greater than 40% of the query sequence and scores below the E-value threshold of 0.001 (Table III.2). These DELTA-BLAST parameters are intentionally lenient to reduce the number of false negatives and maximize total discovery of N4-like phages. The vast majority of ORFs identified share very little sequence identity (20-30%), but align to over 70% of the 3,500 aa N4 vRNAP query sequence (Table III.2). For phage pSb-1, multiple ORFs (Genbank accession numbers YP_009008504.1, YP_009008503.1, and YP_009008502.1) were concatenated to reach the alignment threshold. Each of the three pSb-1 ORFs have significant homology to non-overlapping regions of N4 vRNAP and likely represent a single ORF mistakenly annotated as three separate ORFs due to many missense mutations introduced during genome assembly. All N4 vRNAP homologs identified are large polypeptides (> 3,000 aa) that align to the entire N4 mini-vRNAP domain and portions of the N- and C-terminal domains. Furthermore, sequence alignments show that all putative N4 vRNAP homologs contain the conserved sequence blocks required for catalysis in N4 vRNAP, including the A, B, C, and T/DxxGR motifs (87). Therefore, all 54 ORFs identified are likely N4 vRNAP homologs encoding functional polymerases with multiple roles throughout the phage life cycle.

Table III.2. Identification of N4-like phages by N4 vRNAP DELTA-BLAST homology search

Species	Subject accession	Subject coverage	Coverage length	Query cover	Identity
N4	YP_950528.1	1..3500	3500	100%	100%
IME11	YP_006990621.1	1..3450	3503	100%	66%
vBEcoPPhAPEC7	YP_009056186.1	1..3450	3503	100%	66%
EC1-UPM	AGC31565.1	1..997, 1140..3612	3490	99%	67%
ECBP1	YP_006908827.1	1..977, 1050..3570	3545	100%	66%
Bp4	YP_009113218.1	1..975, 1119..3570	3458	99%	67%
vBEcoPPhAPEC5	YP_009055564.1	1..978, 1073..3571	3518	100%	66%
vBEcoPG7C	YP_004782180.1	1..965, 1241..3748	3628	100%	64%
pSb-1	YP_009008504.1, YP_009008503.1, YP_009008502.1	1..892, 1..76, 172..1450, 1..1212	3491	67%	67%
FSL_SP-076	YP_008240191.1	229..3594	3626	96%	25%
FSL_SP-058	YP_008239463.1	158..3594	3714	99%	25%
Pollock	YP_009152160.1	5..3536	3792	99%	24%
VCO139	AGI61882.1	199..1630, 1946..3002	2727	72%	22%
phi1	YP_009198592.1	474..1630, 2164..3220	2460	66%	22%
JA-1	YP_008126816.1	199..1630, 1946..3002	2727	72%	21%
VBP32	YP_007676574.1	784..2025, 2078..3169	2469	65%	21%
VBP47	YP_007674140.1	618..2025, 2078..3169	2669	70%	21%
pYD6-A	YP_007674286.1	567..2115, 2126..3213	2838	75%	21%
phiCB2047-B	YP_007675808.1	28..3327	3582	99%	25%
DSS3P2	YP_002899070.1	9..925, 1374..3627	3480	96%	26%
EE36P1	YP_002898988.1	9..854, 1528..3780	3348	93%	26%
vBDshPR2C	AID16877.1	43..1003, 1314..3552	3475	95%	25%
RD-1410Ws-07	ANJ20865.1	2..960, 1413..3676	3471	95%	25%
DFL12phi1	YP_009043702.1	43..1003, 1314..3552	3475	95%	25%
DS-1410Ws-06	ANJ20714.1	2..960, 1413..3676	3471	95%	25%
Plymouth_1	CBX87992.1	70..957, 1392..3757	3567	98%	25%
Roseovarius_217	CBW47056.1	70..957, 1392..3757	3567	98%	25%
RD-1410W1-01	ANJ20798.1	5..3345	3672	99%	24%
phiAxp-3	YP_009208706.1	3..3426	3531	99%	55%
JWDelta	AHC56581.1	3..3423	3528	99%	54%
JWAlpha	YP_009004769.1	3..3423	3528	99%	54%
LUZ7	YP_003358355.1	382..3355	3231	84%	21%
vBPaePC2-10Ab09	YP_009031843.1	271..3346	3414	89%	20%
DL64	YP_009206224.1	272..3346	3388	89%	20%
KPP21	YP_009218950.1	396..3355	3200	83%	21%
YH30	YP_009226132.1	271..3346	3424	89%	20%
Pa2	YP_009148251.1	271..3346	3413	89%	20%
phi176	AIZ95005.1	71..3346	3663	96%	20%
PEV2	YP_009286295.1	71..3346	3648	96%	21%

Table III.2. Identification of N4-like phages by N4 vRNAP DELTA-BLAST homology search (continued)

Species	Subject accession	Subject coverage	Coverage length	Query cover	Identity
RWG	AIZ94822.1	71..3346	3648	96%	21%
LIT1	YP_003358468.1	271..3346	3423	89%	20%
vBPaePMAG4	YP_009290607.1	271..3346	3398	89%	20%
PA26	AFO70568.1	271..3346	3414	89%	20%
YH6	YP_009152574.1	271..3346	3384	89%	20%
Frozen	YP_009286200.1	49..3520	3592	99%	42%
Gutmeister	ANJ65375.1	49..3520	3592	99%	42%
Rexella	ANJ65299.1	49..3520	3592	99%	42%
Ea9-2	YP_009007447.1	49..3520	3594	99%	42%
RG-2014	YP_009148431.1	15..3411	3538	99%	40%
vBEamP-S6	YP_007005815.1	166..3535	3588	96%	25%
EcP1	YP_007003173.1	81..3489	3682	98%	24%
Presley	YP_009007647.1	372..959, 1517..3469	2781	74%	21%
ZC03	AMD43402.1	223..848, 1369..2146, 2249..3637	3024	82%	27%
ZC08	AMD43541.1	223..848, 1402..2179, 2282..3670	3027	82%	27%
pVa5	APC46019.1	1290..1799, 2269..3295	1603	43%	21%

COMPARATIVE GENOMICS OF N4-LIKE PHAGES

To validate the classification of putative N4-like phages through vRNAP homology, I searched for shared morphological, physiological, and genomic properties of these phages in the published literature (Table III.3). Where available, morphological and physiological data for the putative N4-like phages characterized suggest that they are all members of the *Podoviridae* family, are strictly lytic, and have similar plaque morphologies. Putative N4-like phages have capsids 61-85 nm in diameter and short, non-contractile tails 5-42 nm in length, which agree with the N4 capsid diameter and tail length (70 and 35 nm, respectively) (Table III.4). The genomes of putative N4-like phages share similarities with the N4 genome, ranging in size from 59-78 kbp, with G+C content 35-60% and encoding 0-15 tRNAs and 73-115 ORFs (Table III.3).

The presence of vRNAP homologs, classification as *Podoviridae*, and similar genome properties strongly suggest that these phages are, in fact, close relatives of N4.

Table III.3. N4-like phage subfamily genome characteristics

Species	Accession	Pubmed	Host	Genome size (bp)	%GC	ORFs	tRNAs	Cluster
N4	NC_008720.1	-	Escherichia coli	70153	41.3	73	4	1
IME11	NC_019423.1	23166261	Escherichia coli	72570	43.1	91	0	1
vBEcoPPhAPEC7	NC_024790.1	24269008	Escherichia coli	71778	43.3	83	1	1
EC1-UPM	KC206276.2	24134834	Escherichia coli	70912	42.9	80	0	1
ECBP1	NC_018854.1	23087106	Escherichia coli	69855	42.7	82	2	1
Bp4	NC_024142.2	-	Escherichia coli	72583	42.9	94	2	1
vBEcoPPhAPEC5	NC_024786.1	24269008	Escherichia coli	71248	43.5	83	1	1
vBEcoPG7C	NC_015933.1	22341309	Escherichia coli	72917	43.4	79	0	1
pSb-1	NC_023589.1	25283727	Shigella boydii	71629	42.7	103	0	1
FSL_SP-076	NC_021782.1	23865498	Salmonella enterica	72098	39.5	82	10	2
FSL_SP-058	NC_021772.1	23865498	Salmonella enterica	72394	39.6	86	10	2
Pollock	NC_027381.1	25635029	Escherichia coli	68365	36.0	85	4	2
VCO139	KC438283.1	23714204	Vibrio cholerae	68964	34.6	79	1	3
phi1	NC_028799.1	-	Vibrio cholerae	66708	34.5	110	0	3
JA-1	NC_021540.1	23714204	Vibrio cholerae	69278	34.6	79	1	3
VBP32	NC_020868.1	-	Vibrio parahaemolyticus	76718	42.5	115	2	4
VBP47	NC_020848.1	-	Vibrio parahaemolyticus	76705	42.5	115	2	4
pYD6-A	NC_020849.1	-	Pseudoalteromonas sp. YD6	76802	38.7	99	2	4
phiCB2047-B	NC_020862.2	24435853	Sulfitobacter sp. CB2047	74485	43.0	75	15	5
DSS3P2	NC_012697.1	19689706	Ruegeria pomeroyi	74611	47.9	81	0	5
EE36P1	NC_012696.1	19689706	Sulfitobacter sp. EE-36	73325	47.0	79	0	5
vBDshPR2C	KJ803031.1	25795023	Dinoroseobacter shibae	74806	49.2	85	0	5
RD-1410Ws-07	KU885990.1	27270945	Roseobacter denitrificans	76298	50.0	76	0	5
DFL12phi1	NC_024367.1	26380630	Dinoroseobacter shibae	75028	49.3	86	0	5
DS-1410Ws-06	KU885988.1	27270945	Dinoroseobacter shibae	76466	50.0	77	0	5
Plymouth_1	FR719956.1	25346726	Roseovarius nubinhibens	74704	49.1	91	3	5
Roseovarius_217	FR682616.1	25346726	Roseovarius sp. 217	74583	49.0	92	3	5
RD-1410W1-01	KU885989.1	27270945	Roseobacter denitrificans	72674	49.5	77	0	5
phiAxp-3	NC_028908.1	27094846	Achromobacter xylosoxidans	72409	55.2	80	0	6
JWDelta	KF787094.1	24468270	Achromobacter xylosoxidans	73659	54.3	89	0	6

Table III.3. N4-like phage subfamily genome characteristics (continued)

Species	Accession	Pubmed	Host	Genome size (bp)	%GC	ORFs	tRNAs	Cluster
JWAlpha	NC_023556.1	24468270	Achromobacter xylosoxidans	72329	54.4	91	0	6
LUZ7	NC_013691.1	20619867	Pseudomonas aeruginosa	74901	53.2	115	0	7
vBPaePC2-10Ab09	NC_024140.1	26115051	Pseudomonas aeruginosa	72028	54.9	83	0	7
DL64	NC_028885.1	-	Pseudomonas aeruginosa	72378	55.0	90	0	7
KPP21	NC_029017.1	26616567	Pseudomonas aeruginosa	73420	53.5	113	0	7
YH30	NC_029101.1	27283850	Pseudomonas aeruginosa	72192	54.9	86	0	7
Pa2	NC_027345.1	-	Pseudomonas aeruginosa	73008	54.9	91	0	7
phi176	KM411960.1	-	Pseudomonas aeruginosa	73048	54.9	92	0	7
PEV2	NC_031063.1	20619867	Pseudomonas aeruginosa	72697	54.9	93	0	7
RWG	KM411958.1	-	Pseudomonas aeruginosa	72646	54.9	92	0	7
LIT1	NC_013692.1	20619867	Pseudomonas aeruginosa	72544	55.0	90	0	7
vBPaePMAG4	NC_031104.1	-	Pseudomonas aeruginosa	72979	54.8	94	0	7
PA26	JX194238.1	22923802	Pseudomonas aeruginosa	72321	54.8	88	0	7
YH6	NC_027388.1	26254772	Pseudomonas aeruginosa	73050	54.9	90	0	7
Frozen	NC_031062.1	-	Erwinia amylovora	75147	46.9	92	8	8
Gutmeister	KX098391.1	-	Erwinia amylovora	71173	46.9	82	8	8
Rexella	KX098390.1	-	Erwinia amylovora	75448	46.9	92	7	8
Ea9-2	NC_023579.1	-	Erwinia amylovora	75568	47.0	94	7	8
RG-2014	NC_027348.2	25728819	Delftia tsuruhatensis	73990	59.9	88	0	S
vBEamP-S6	NC_019514.1	21764969	Erwinia amylovora	74669	52.1	115	0	S
EcP1	NC_019485.1	-	Enterobacter cloacae	59080	36.9	77	3	S
Presley	NC_023581.1	24309722	Acinetobacter baumannii	77792	37.8	95	0	S
ZC03	KU356690.1	28472930	Pseudomonas aeruginosa	69844	42.6	85	10	S
ZC08	KU356691.1	28472930	Pseudomonas aeruginosa	70774	43.1	83	9	S
pVa5	KX889068.1	-	Vibrio splendidus	78145	43.2	106	0	S

Table III.4. N4-like phage subfamily morphology and physiology

Species	Host	Latent period (min)	Burst size (pfu/cell)	Capsid diam. (nm)	Tail length (nm)	Plaque (mm)	Sample collection
N4	Escherichia coli	180	3,000	70	35	2	Sewer. Genoa, Italy
IME11	Escherichia coli	n/a	n/a	n/a	n/a	n/a	Hospital sewage. Beijing, China
vBEcoPPhAPEC7	Escherichia coli	n/a	n/a	70	n/a	n/a	River samples. Brussels, Belgium
EC1-UPM	Escherichia coli	n/a	n/a	n/a	n/a	n/a	Chicken feces
ECBP1	Escherichia coli	n/a	n/a	n/a	n/a	n/a	Chicken farms. Yesan, South Korea
Bp4	Escherichia coli	n/a	n/a	n/a	n/a	n/a	n/a
vBEcoPPhAPEC5	Escherichia coli	n/a	n/a	65	n/a	n/a	River samples. Brussels, Belgium
vBEcoPG7C	Escherichia coli	40-42	500-100	~70	~25	1-2	Horse feces
pSb-1	Shigella boydii	15	152.63	61±4	12±2	n/a	Dorimcheon stream. Seoul, South Korea
FSL_SP-076	Salmonella enterica	n/a	n/a	n/a	n/a	n/a	n/a
FSL_SP-058	Salmonella enterica	n/a	n/a	n/a	n/a	n/a	n/a
Pollock	Escherichia coli	n/a	n/a	n/a	n/a	n/a	Sewage sample. College Station, TX
VCO139	Vibrio cholerae	~60	~150	~64.8	n/a	n/a	Sewage effluent. Bangladesh, India
phi1	Vibrio cholerae	n/a	n/a	n/a	n/a	n/a	n/a
JA-1	Vibrio cholerae	~60	~150	~68.7	n/a	n/a	Sewage effluent. Bangladesh, India
VBP32	Vibrio parahaemolyticus	n/a	n/a	n/a	n/a	n/a	n/a
VBP47	Vibrio parahaemolyticus	n/a	n/a	n/a	n/a	n/a	n/a
pYD6-A	Pseudoalteromonas sp. YD6	n/a	n/a	n/a	n/a	n/a	n/a
phiCB2047-B	Sulfitobacter sp. CB2047	n/a	n/a	n/a	n/a	n/a	Raunefjorden, Norway
DSS3P2	Ruegeria pomeroyi	180	350	~70	~26	n/a	Seawater. Baltimore, MD
EE36P1	Sulfitobacter sp. EE-36	120	1500	~70	~26	n/a	Seawater. Baltimore, MD
vBDshPR2C	Dinoroseobacter shibae	n/a	n/a	72±5	28±2	n/a	Coastal seawater. Xiamen, China
RD-1410Ws-07	Roseobacter denitrificans	<60	341	69.6±3.8	41.4±2.8	1-2	Surface water. Sanya Bay, South China Bay
DFL12phi1	Dinoroseobacter shibae	n/a	n/a	~75	~35	n/a	Surface water. Xiamen, China

Table III.4. N4-like phage subfamily morphology and physiology (continued)

Species	Host	Latent period (min)	Burst size (pfu/cell)	Capsid diam. (nm)	Tail length (nm)	Plaque (mm)	Sample collection
DS-1410Ws-06	Dinoroseobacter shibae	120	298	70.8±1.9	41.6±1.8	1-2	Surface water, Sanya Bay, South China Bay
Plymouth_1	Roseovarius nubinhibens	160-360	~10	77.4±5	n/a	0.5-2	L4 sampling station. Plymouth, UK
Roseovarius_217	Roseovarius sp. 217	160-360	~100	72.4±2	n/a	0.5-2	Langstone Harbor. Hampshire, UK
RD-1410W1-01	Roseobacter denitrificans	60	27	63.2±1.6	40±3.7	1-2	Surface water. Sanya Bay, South China Bay
phiAxp-3	Achromobacter xylosoxidans	80	~9000	67	20	clear	Hospital sewage. China
JWDelta	Achromobacter xylosoxidans	90-120	~180	72	22	1-2	Waste water treatment. Braunschweig, Germany
JWAlpha	Achromobacter xylosoxidans	90-120	~180	69	22	1-2	Waste water treatment. Werl, Germany
LUZ7	Pseudomonas aeruginosa	n/a	n/a	76	30	1-2	Hospital sewage samples. Belgium
vBPaePC2-10Ab09	Pseudomonas aeruginosa	n/a	n/a	70	n/a	n/a	Sewer water. Abidjan, Ivory Coast
DL64	Pseudomonas aeruginosa	n/a	n/a	n/a	n/a	n/a	n/a
KPP21	Pseudomonas aeruginosa	n/a	n/a	67.0±2.4	5.3±1.4	n/a	Agricultural wastewater drain. Kochi City, Japan
YH30	Pseudomonas aeruginosa	20	n/a	65	40	n/a	Sewer system. China
Pa2	Pseudomonas aeruginosa	n/a	n/a	n/a	n/a	n/a	n/a
phi176	Pseudomonas aeruginosa	n/a	n/a	n/a	n/a	n/a	n/a
PEV2	Pseudomonas aeruginosa	25-30	100-150	~70	~30	1-2	Sewage treatment. Olympia, WA
RWG	Pseudomonas aeruginosa	n/a	n/a	n/a	n/a	n/a	n/a
LIT1	Pseudomonas aeruginosa	n/a	n/a	74	30	1-2	Local pond. Italy
vBPaePMAG4	Pseudomonas aeruginosa	n/a	n/a	n/a	n/a	n/a	n/a
PA26	Pseudomonas aeruginosa	n/a	n/a	n/a	n/a	n/a	Water reservoir. Naju City, South Korea
YH6	Pseudomonas aeruginosa	30	n/a	65	25	1-2	Sewer system. Changchun, China
Frozen	Erwinia amylovora	n/a	n/a	n/a	n/a	n/a	n/a

Table III.4. N4-like phage subfamily morphology and physiology (continued)

Species	Host	Latent period (min)	Burst size (pfu/cell)	Capsid diam. (nm)	Tail length (nm)	Plaque (mm)	Sample collection
Gutmeister	Erwinia amylovora	n/a	n/a	n/a	n/a	n/a	n/a
Rexella	Erwinia amylovora	n/a	n/a	n/a	n/a	n/a	n/a
Ea9-2	Erwinia amylovora strain	n/a	n/a	n/a	n/a	n/a	n/a
RG-2014	Delftia tsuruhatensis	10	150±9	85	n/a	1-2	Waste water treatment. Salt Lake City, UT
vBEamP-S6	Erwinia amylovora	n/a	n/a	66	n/a	n/a	Orchard samples. Switzerland
EcP1	Enterobacter cloacae	n/a	n/a	n/a	n/a	n/a	n/a
Presley	Acinetobacter baumannii	n/a	n/a	n/a	n/a	n/a	n/a
ZC03	Pseudomonas aeruginosa	50	10	72	21	0.5-1	Composting facility. Sau Paolo, Brazil
ZC08	Pseudomonas aeruginosa	50	10	72	21	0.5-1	Composting facility. Sau Paolo, Brazil
pVa5	Vibrio splendidus	n/a	n/a	n/a	n/a	n/a	Greece

To evaluate the diversity of N4-like phages and compare the evolutionary history of N4-like genomes, I sorted the N4-like phages into clusters based on full-genome nucleotide alignment according to the methodology utilized for the classification of mycobacteriophages (66). The pairwise fraction genome nucleotide alignment of N4-like phages was calculated using the NCBI nucleotide BLAST (BLASTn) algorithm, grouped into discrete clusters through hierarchical clustering, and visualized as a heatmap representation (Figure III.3). Hierarchical clustering separated the phages into eight well-defined clusters corresponding with a 0.4 fraction BLASTn alignment cutoff. Seven N4-like phage species (RG-2014, vBEamP-S6, EcP1, Presley, ZC03, ZC08 and pVa5) lacking more than one relative above the BLASTn alignment cutoff were not assigned to a cluster and were designated as singletons. Phages share significant nucleotide sequence similarity to other phages within the same cluster, (mean intra-cluster fraction BLASTn alignment 0.8±0.2), but largely lack sequence similarity with phages outside of their cluster (mean inter-cluster fraction BLASTn alignment 0.06±0.07).

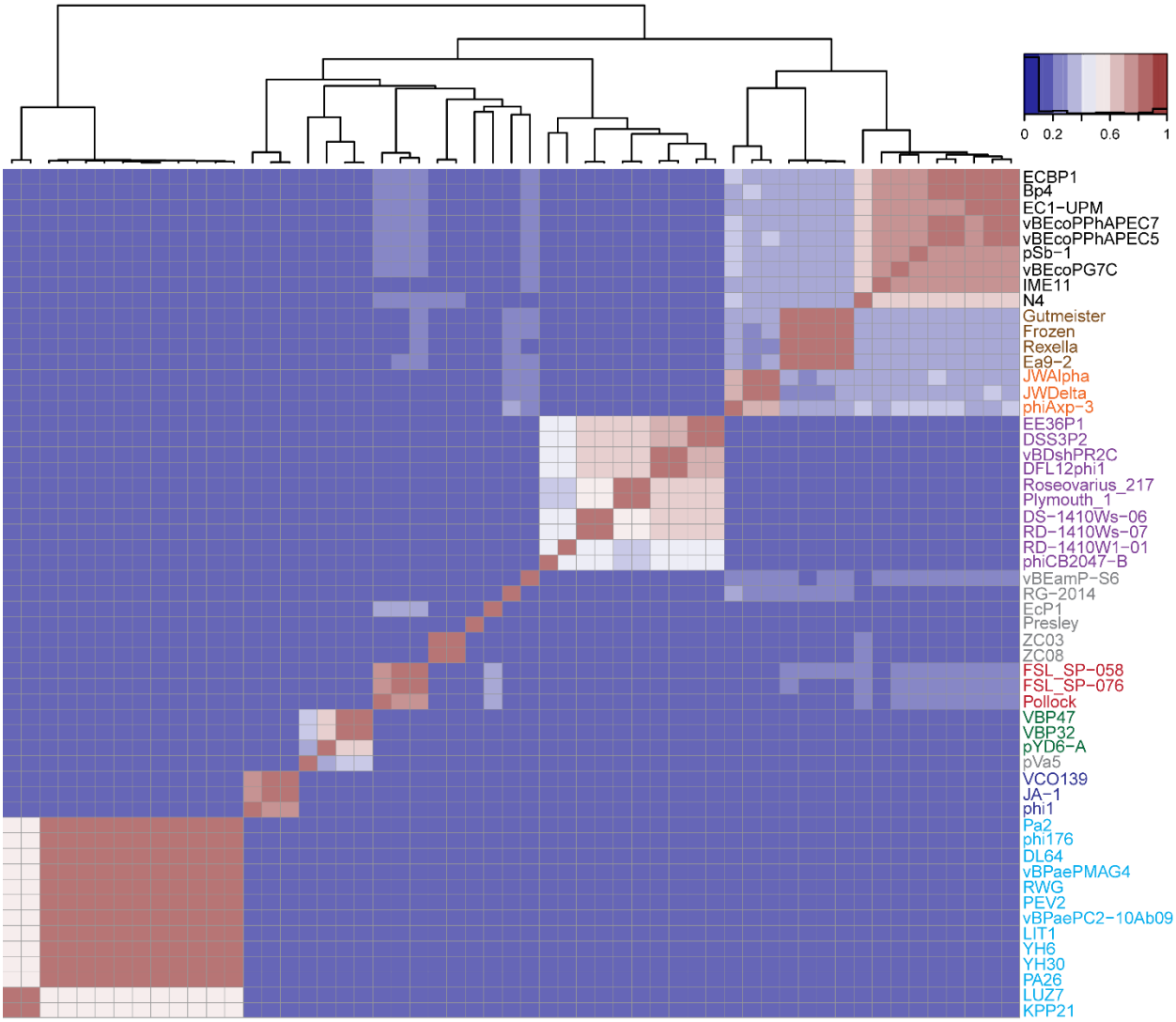

Figure III.3. Hierarchical clustering of N4-like phage species into eight clusters. Heatmap representation of fraction pairwise genome nucleotide alignment for all N4-like phages. Hierarchical clustering algorithm, displayed as dendrogram, groups N4-like phages into eight clusters, with seven phage species lacking more than one close relative remaining as singleton genomes. Phage names are indicated as row names and are color-coded according to cluster assignments, corresponding with a 0.4 BLASTn cutoff for intra-cluster pairwise comparisons. Cluster 1 phages, black; Cluster 2 phages, red; Cluster 3 phages, blue; Cluster 4 phages, green; Cluster 5 phages, purple; Cluster 6 phages, orange; Cluster 7 phages, cyan; Cluster 8 phages, brown; Singletons, grey. Cells are color-coded by fraction pairwise nucleotide alignment as indicated in legend, which also includes a histogram of dataset.

To independently validate genome clustering by BLASTn alignment, the pairwise average nucleotide identity (ANI) was calculated, summarized as a pairwise distance matrix, sorted into clusters through hierarchical clustering, and visualized as a heatmap representation

(Figure III.4). N4-like phage cluster assignment by BLASTn alignment was largely recapitulated by hierarchical clustering of N4-like phage ANI, with the exception of LUZ and KPP21 removal from the remaining Cluster 7 phages, suggesting that these two phages belong to a separate subcluster within Cluster 7.

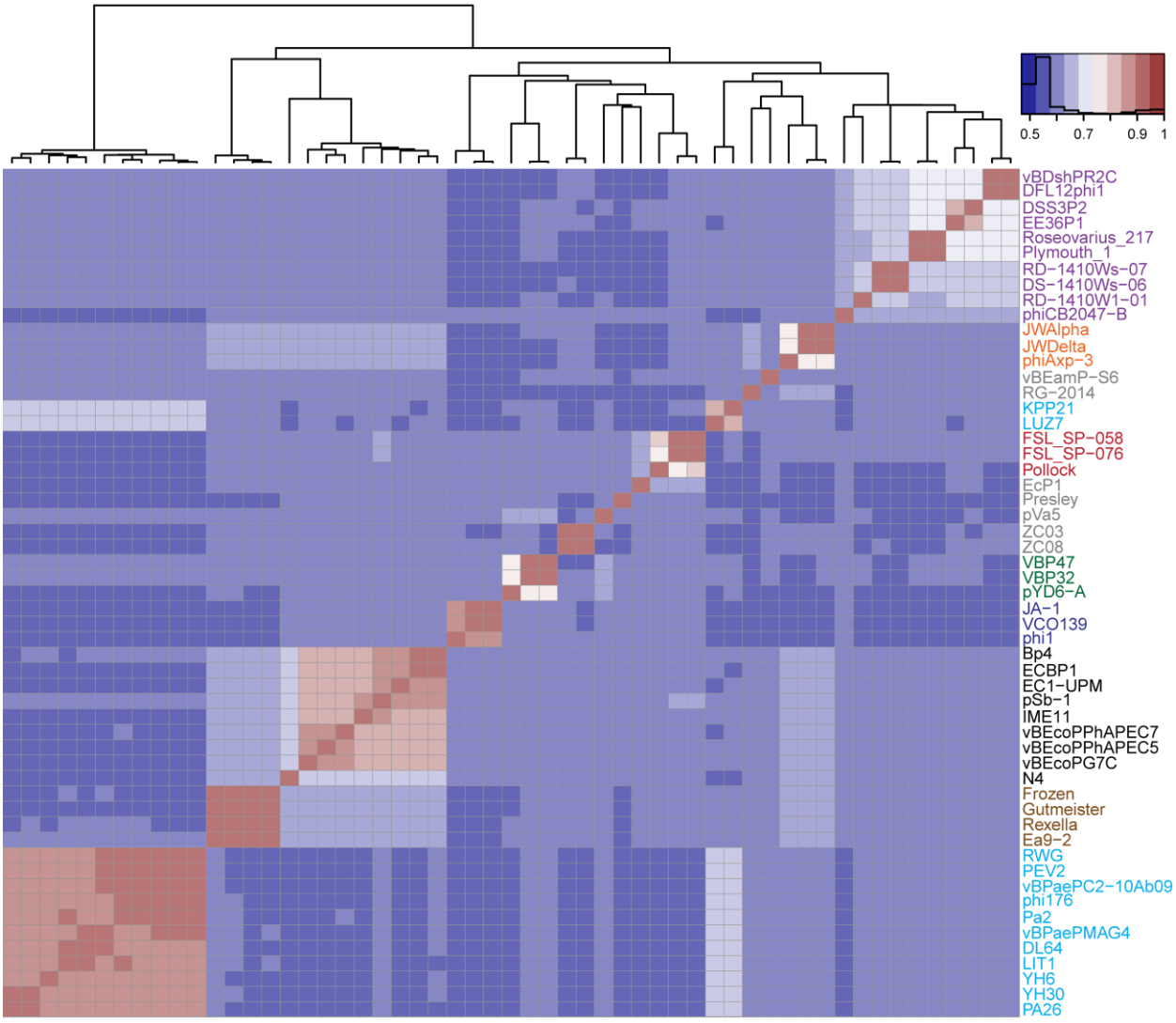

Figure III.4. N4-like phage clustering is supported by ANI. Heatmap representation of pairwise average nucleotide identity (ANI) shared between N4-like phages. ANI of N4-like phage genomes was calculated and plotted as heatmap representation. Phages are color-coded according to cluster assignment as in Figure III.3. Cells are color-coded by ANI as indicated in legend, which also includes a histogram of the dataset.

Dotplot visualization of N4-like phage genome nucleotide alignment (Figure III.5) shows that N4-like phages within the same cluster have strong sequence similarity throughout the length of the genome, suggesting that gene synteny and genomic organization is conserved across closely related species. In contrast, phages outside of the same cluster display weak conservation across the whole genome and only share sequence identity to limited regions, a hallmark of genomic mosaicism. The greatest variability in genome sequences occurs through the definition of genome ends and orientation. The published genomes of IME11, Bp4, VBP32, phiCB2047-B, DFL12phi1, DL64, YH30, and pVa5 have reversed orientations with respect to N4, while numerous untested end definitions are proposed throughout the subfamily without experimental validation. The genome ends play important roles in genome injection and early transcription at the onset of N4 infection, underscoring the necessity for genome end verification in other N4-like phages through biochemical techniques.

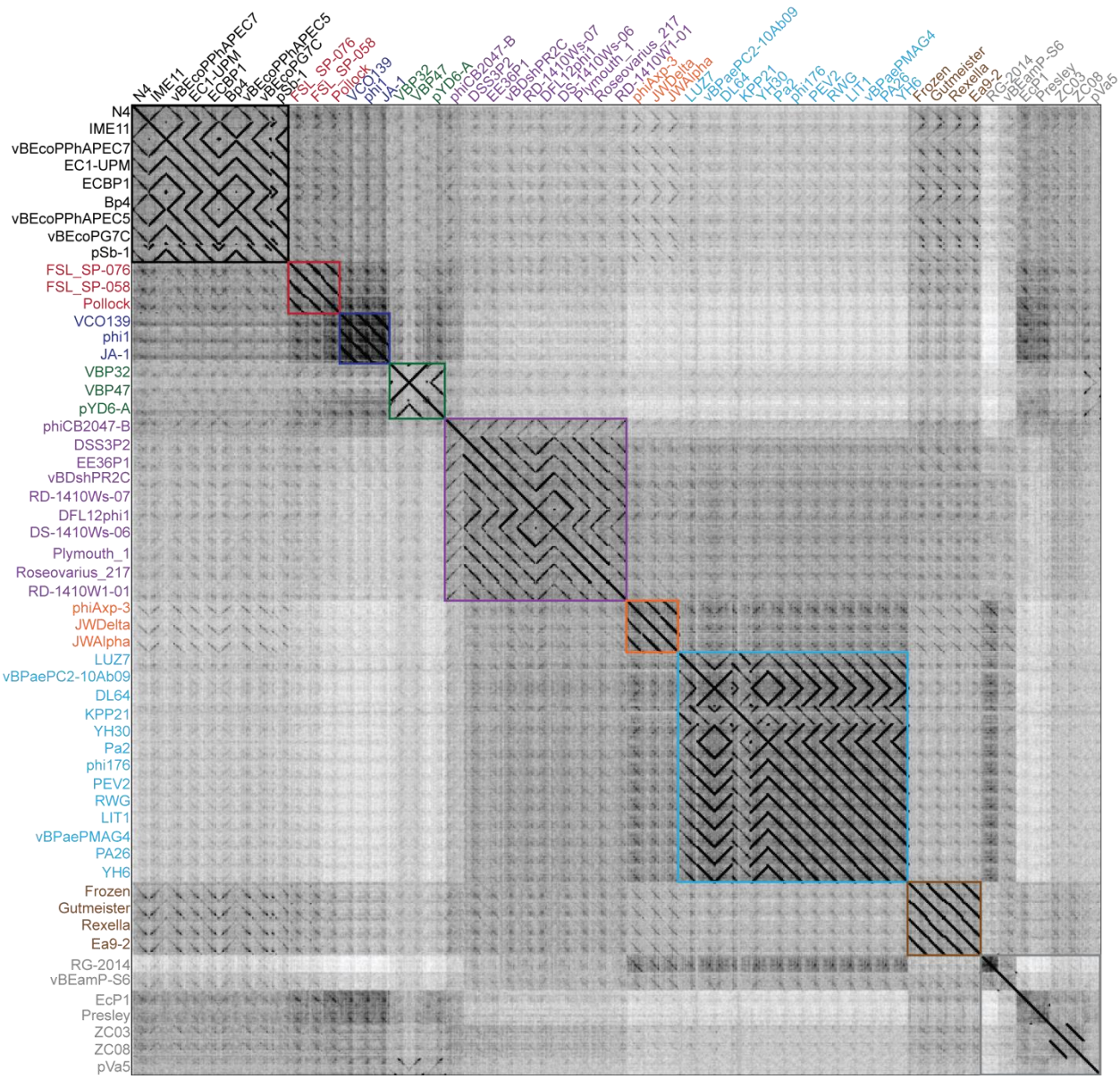

Figure III.5. Nucleotide sequence comparisons of all N4-like phage genomes. Individual genome sequences were concatenated by cluster affiliation into a single sequence and genome-wide nucleotide alignment were visualized as a dotplot using Gepard. Phage names are color-coded by cluster affiliation as in Figure III.3. Phage clusters are outlined for clarity.

Visualization of pairwise genome alignments by EasyFig representations allows for a fine-scale analysis of genome rearrangements among closely related phages (Figure III.6). These analyses show that the N4-like phages share a bipartite genome organization, with the vRNAP gene at the center of the genome near the transcription polarity transition. The left end of N4-like

phage genomes contain numerous ORFs transcribed with rightward polarity, while the right end of N4-like phage genomes contain larger morphogenetic ORFs transcribed with leftward polarity. These data suggest that all N4-like phages may have conserved strategies for the temporal expression of their genomes.

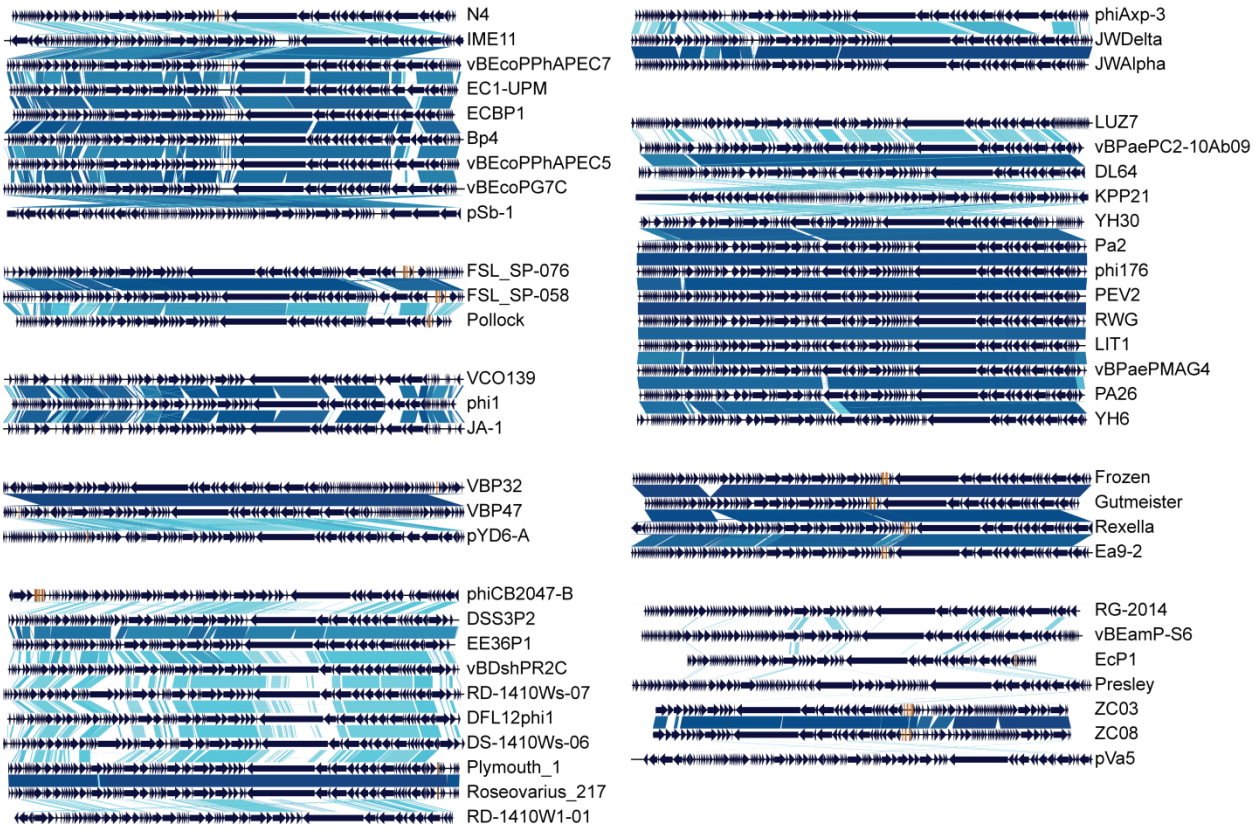

Figure III.6. N4-like phages share similar genome organization. N4-like phage genomes within each phage cluster were aligned through BLASTn in EasyFig. Nucleotide similarities between adjacent genomes displayed as color gradient from light to dark blue (60-100%). Genome representations of IME11, Bp4, VBP32, phiCB2047-B, DFL12phi1, DL64, YH30 and pVa5 were inverted to match orientation of N4 genome. ORFs, dark blue arrows; tRNAs, orange rectangles.

There are, however, a few notable exceptions to the bipartite genome organization where individual phage clusters have insertions containing ORFs of opposite orientation (Figure III.6). These ORF insertions are largely conserved within, but not between phage clusters. The change in orientation and lack of conservation outside of the phage cluster suggest that these insertions

result from non-homologous recombination with sequences outside of the N4-like phage subfamily. Cluster 2 phages have a 6-8 ORF cassette of rightward polarity inserted into the late region, encoding a putative lysis cassette as well as other morphogenetic proteins. Cluster 3, 4, 7, EcP1, and pVa5 phages have a similarly located 1-2 ORF insertion. The middle region also has insertions among several N4-like phages. Cluster 3 and 4 phages, as well as pVa5, have a single ORF (predicted to be an N-acetylmuramoyl-L-alanine amidase for Cluster 3 phages), with leftward polarity in the middle region. Cluster 7 and Presley phages have similarly located 4-8 ORF insertions encoding putative lysozymes. These results suggest that the evolutionary history and mechanisms of lysis are vastly different in Cluster 2, 3, and 7 phages.

CONSERVATION OF ORFs AMONG N4-LIKE PHAGES

To identify ORFs of related function across N4-like phage species, I performed an all-vs.-all BLAST homology search of ORFs in the N4-like phage subfamily (E-value cutoff of 0.001) and sorted the results into ORF families (ORFams) of related sequences using the Markov Clustering (MCL) Algorithm. Through this approach, I sorted 4,921 of the 4,922 N4-like ORFs into 1,016 ORFams (Table III.5, available as supplemental .xls file). An 11 aa YH6 ORF (Genbank accession numberYP_009152534.1) failed to meet the significance threshold (even for self-alignment) and was excluded from remaining analyses. A total of 20 (2.0%) ORFams were completely conserved across all N4-like phages, representing a “core genome” that acts to define this subfamily of bacteriophages. Conserved ORFams separate into three primary groups based on their annotation in the N4 genome (Figure III.7). The first group consists of proteins of unknown function (gp42, gp52, gp53, gp54, gp55, gp57, gp67, and gp69). Although proteins within this group are predicted to be morphogenetic based on their genome location and synteny,

no annotated functions have been assigned for any protein within these ORFams. The second group consists of generic phage proteins (putative AAA+ ATPase, putative metallopeptidase, DNAP, putative DNA primase, putative AAA ATPase, major capsid, portal, and putative terminase). Although putative functions have been assigned for all proteins within this group, their predicted functions apply to general phage functions that are shared across numerous phage families. In fact, ORFs within these ORFams share sequence conservation with numerous proteins outside of the N4-like phage subfamily. The final group of conserved ORFams consists of N4-specific proteins (N4 RNAPII subunit 1, N4 RNAPII subunit 2, N4SSB, and vRNAP). The N4-specific ORFams contain ORFs that are not only conserved in all N4-like phage genomes, but also lack sequence similarity to proteins encoded by phages outside of the N4-like phage subfamily defined in this work. Therefore, N4 RNAPII, N4SSB, and vRNAP and their homologs are the hallmark features of N4-like phages.

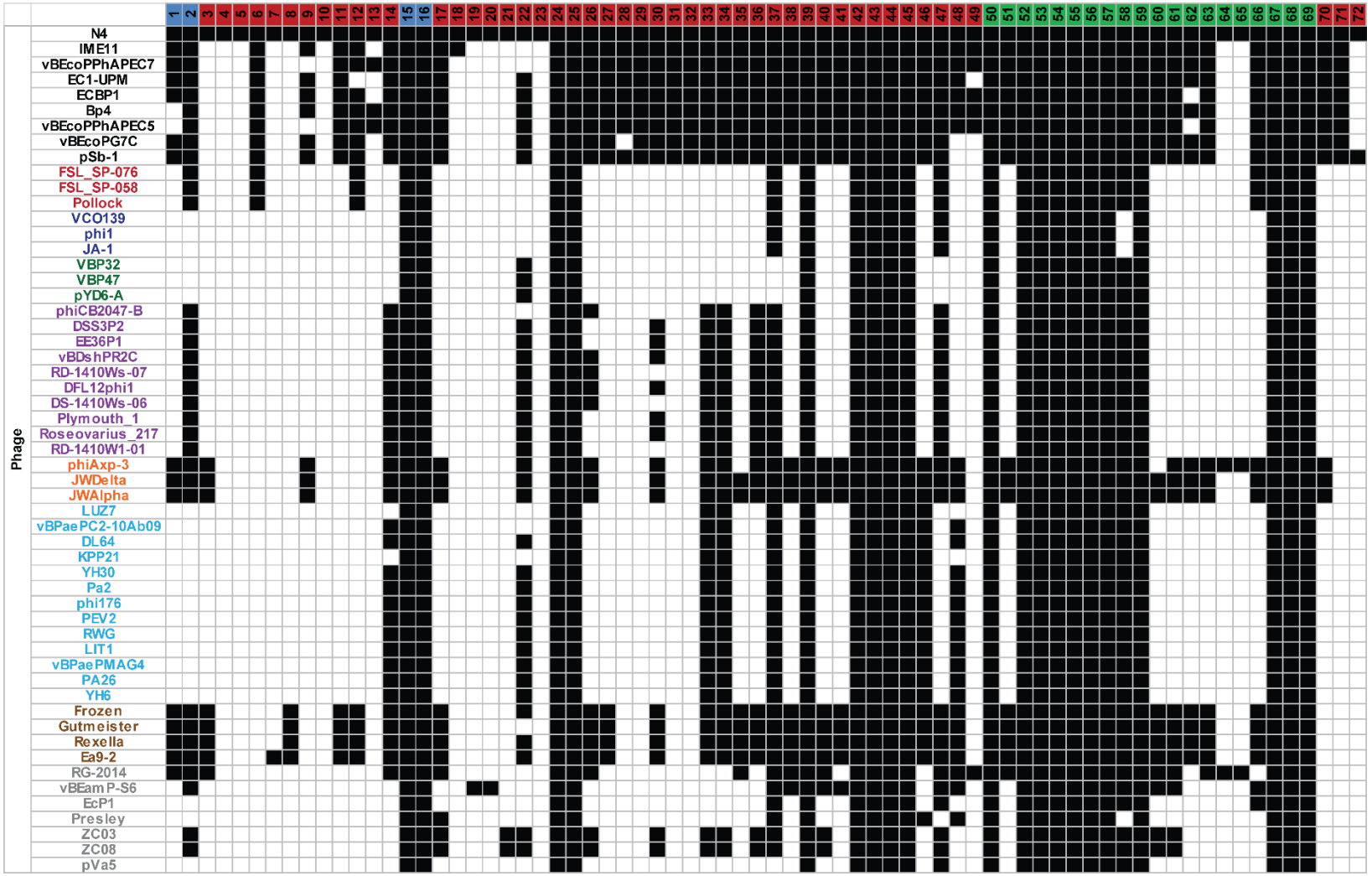

Figure III.7. N4 ORF conservation among N4-like phages. All-vs.-all BLAST results of N4-like phage ORFs were sorted into ORFams through the MCL clustering algorithm. ORFams that contain N4 ORFs are shown as presence/absence matrix, with phages encoding an ORF homologous to an N4 ORF denoted as filled cells. N4-like phages are color-coded as in Figure III.3. N4 ORFs are color-coded according to the timing of expression: blue, early genes; red, middle genes; green, late genes.

Due to their lack of conservation across all members of the N4-like phage subfamily, all other ORFs comprise the accessory genome. A large proportion of N4-like phage ORFs are not well conserved: 436 (42.9%) ORFams contain ORFs lacking homologs from other species (ORFans) and 185 (18.2%) ORFams contain only a pair of homologous ORFs. The vast majority of ORFs within ORFans encode short (less than 100 aa) proteins of unknown function, suggesting that these ORFs contain novel proteins and represent a trove of unexplored biology. Singleton phage genomes are highly enriched in ORFans and two-ORF ORFams, indicating their genetic isolation from the remaining N4-like phages and corroborating their exclusion from cluster assignment (Figure III.8).

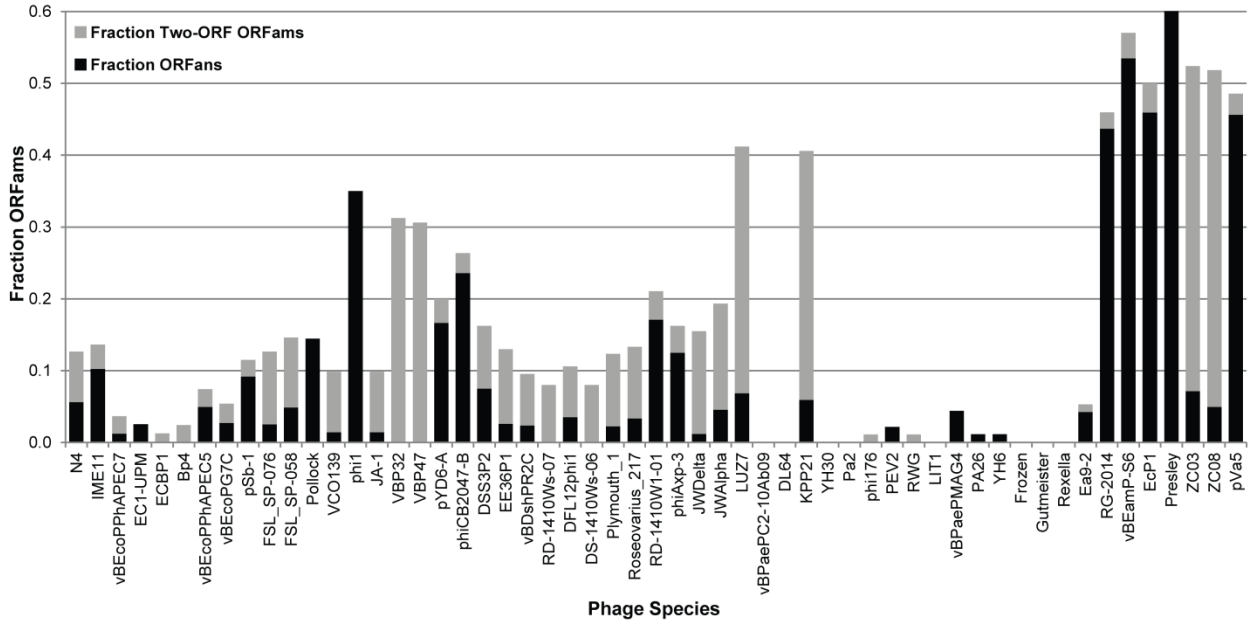

Figure III.8. N4-like phage ORFs are poorly conserved. The fraction of ORFams that comprise single-ORF ORFams (ORFans) (black) and double-ORF ORFams (grey) are plotted for each N4-like phage species.

While ORFs shared across the entire subfamily represent basic physiological strategies, ORFs associated with individual N4-like phage clusters likely represent accessory ORFs encoding proteins responsible for species-specific infection strategies. To identify ORFams

associated with each N4-like phage cluster, I searched for community structure within the N4-like phage-ORFam network using the Bipartite, Recursively Induced Modules algorithm. Through this method, the N4-like phage-ORFam network was separated into 13 distinct modules with a high modularity score (Q = 0.52) (Figure III.9). Randomly permuted matrices containing the same number of phage-ORFam associations produce 28±3 modules with Q = 0.267±0.007, suggesting that the modular distribution of ORFams in N4-like phage genomes is non-random. The phage-ORFam modules defined here recapitulate the phage-phage relationships established through nucleotide alignment, with the exception of the inclusion of phage pVa5 with the Cluster 4 phages due to the numerous ORFs shared between pVa5, VBP32, VBP47, and pYD6-A.

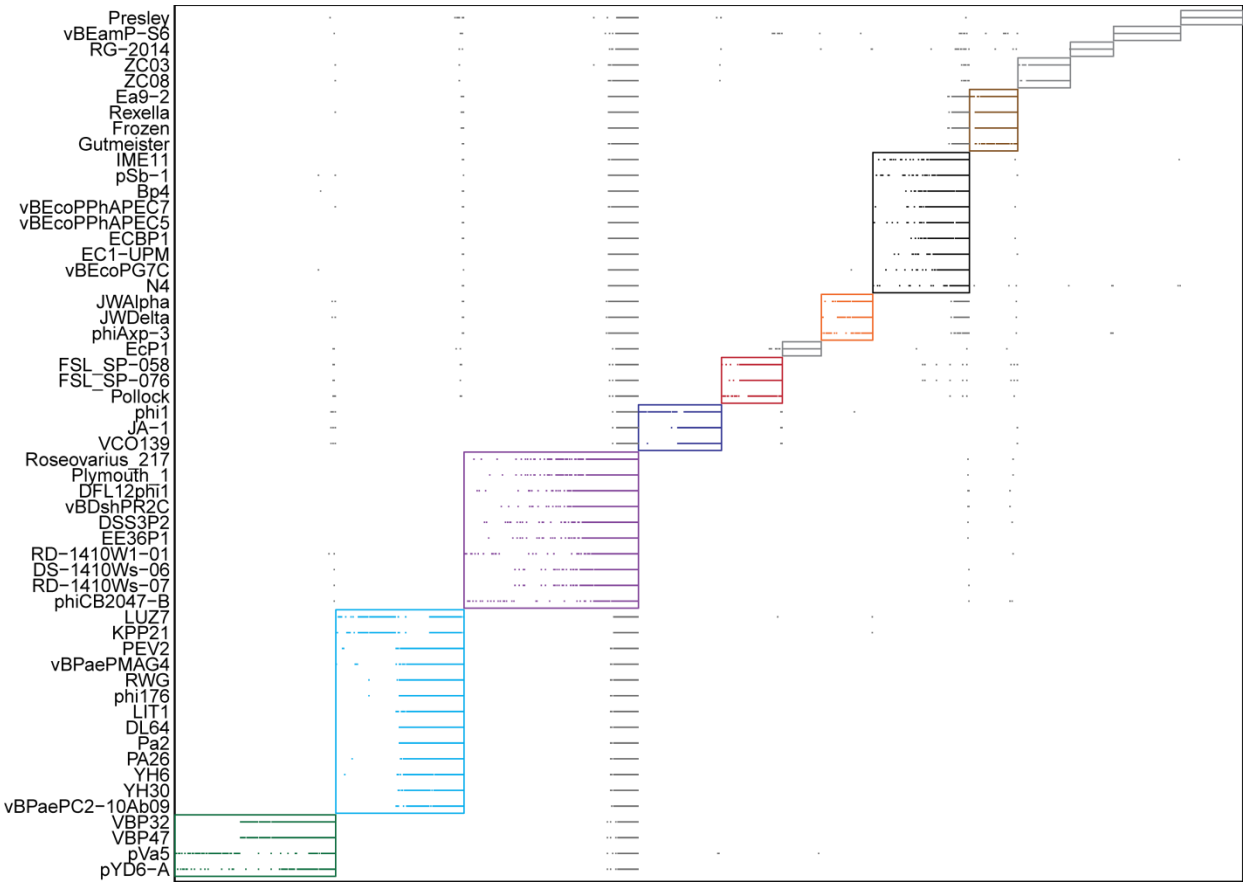

Figure III.9. Detection of N4-like phage cluster-associated ORFams. N4-like phage-ORFam incidence matrix was analyzed for modularity through the Bipartite, Recursively Induced Modules algorithm, which separates the network into 13 distinct clusters (modularity score; Q = 0.52) outlined with colored boxes. Presence of ORFams for each phage species marked by filled squares; intra-module relationships denoted as colored squares and inter-module relationships

Figure III.9 (continued)
denoted as light grey squares. Modules are color-coded to match the phage clusters in Figure III.3.

Since phages within each cluster infect the same or closely related host species, ORFams conserved and unique to these modules contain putative host specificity factors. Therefore, I have defined Cluster Identifier ORFams (ORFams with ORFs present in all genomes of a cluster but absent from all others) for each N4-like phage cluster. ORFs within these families are excellent candidates to encode host-specificity factors and warrant further characterization by biochemical and genetic methods (Table III.6).

Table III.6. N4-like phage Cluster Identifier ORFams

N4-like phage cluster	ORFams
Cluster 1	136-139
Cluster 2	186, 378-396
Cluster 3	170, 171, 285, 349-376
Cluster 4	203, 253, 255, 309-312, 314-317, 319-323, 325-327, 329-333
Cluster 5	49, 65 ,95, 118-121, 123-125
Cluster 6	289-292, 294-305
Cluster 7	50, 59, 66-76, 78-83
Cluster 8	217-219, 221-225, 227-237
Singletons	n/a

RETICULATE PHYLOGENY OF N4-LIKE PHAGES

Although the mosaic nature of bacteriophage genome evolution through the action of horizontal gene transfer makes the determination of a canonical hierarchical taxonomy challenging, relationships across phage populations are accurately represented as a reticulate network of shared ORFams. To that end, I constructed an incidence matrix of ORFam presence and absence for each N4-like phage and displayed the reticulate network using the NeighborNet

method in Splitstree (Figure III.10). In this representation, nodes represent phage species and edges represent shared ORFams, while multiple branches linking groups of phages are indicative of mosaicism across phage genomes. The reticulate phylogeny largely supports clustering of N4-like phages by nucleotide alignment. Phage clusters fall along distinct branches of the tree, while singleton phages do not strongly associate with any single N4-like phage cluster. Although not significant to assign these singletons to a single N4-like phage cluster, several N4-like singleton phages clearly share ORF homology with N4-like phage clusters. pVa5 shares a large number of ORFams with Cluster 4 phages, while EcP1 and vBEamP-S6 share numerous ORFams with Cluster 2 phages.

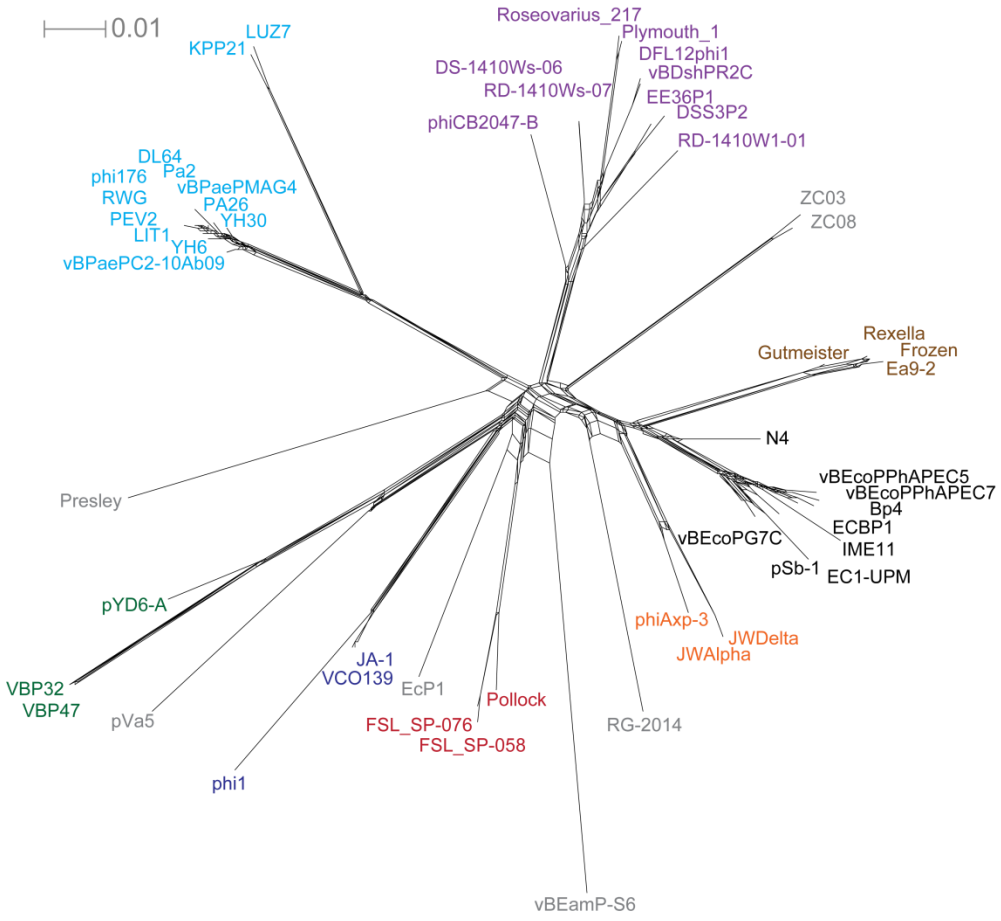

Figure III.10. Reticulate network of N4-like phage relationships by shared ORFams. N4-like phage-ORFam incidence matrix was converted to a reticulate network representation by SplitsTree 4. Scale bar represents number of substitutions per site. Phage clusters color-coded as in Figure III.3.

This reticulate phylogeny suggests that N4-like phages have variable degrees of diversity and isolation of gene flow within and between phage clusters. To measure the diversity within phage clusters, I calculated the pairwise fraction of shared ORFams averaged across all phages within each cluster (Cluster Average Shared ORFams; CLASO) and the fraction of ORFams in each phage genome that are conserved across the cluster (Cluster Associated ORFams; CAO). To measure gene flow between clusters, I calculated the Cluster Isolation Index (CII), which measures the fraction of ORFams in a cluster that are absent in other clusters (Figure III.11 and Table III.7). Clusters 5 and 7 contain the most diverse group of N4-like phages, but display significant isolation from the rest of the N4-like phage subfamily. The majority of diversity within the Cluster 7 phages is provided through the KPP21 and LUZ7 subcluster: phages within each Cluster 7 subcluster share over 90% of their ORFams, but share less than 70% of their ORFams with phages in the other subcluster. In contrast, Clusters 1, 6, and 8 contain few unique ORFams with apparent gene flow between phages in these clusters illustrated in the reticulate network (Figure III.10).

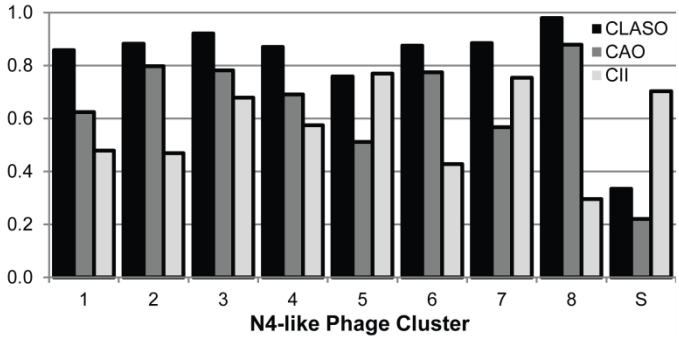

Figure III.11. N4-like phage cluster diversity and isolation. Cluster Average Shared ORFams (CLASO; black), Cluster Associated ORFams (CAO; dark grey), and Cluster Isolation Index (CII; light grey) values are plotted for each N4-like phage cluster. CLASO and CAO values are indicators of intra-cluster diversity, while CII values are indications of inter-cluster isolation.

Table III.7. N4-like phage cluster diversity and isolation

	# Phages	# ORFs	Avg # ORFs	# ORFams	Fraction ORFans	CLASO	CAO	CII
Cluster 1	9	767	85.2	138	0.04	0.86	0.62	0.48
Cluster 2	3	253	84.3	96	0.07	0.88	0.80	0.47
Cluster 3	3	268	89.3	109	0.13	0.92	0.78	0.68
Cluster 4	3	329	109.7	129	0.06	0.87	0.69	0.57
Cluster 5	10	819	81.9	173	0.06	0.76	0.51	0.77
Cluster 6	3	260	86.7	103	0.06	0.88	0.77	0.43
Cluster 7	13	1216	93.5	150	0.02	0.88	0.57	0.75
Cluster 8	4	360	90.0	98	0.01	0.98	0.88	0.30
Singletons	7	649	92.7	414	0.38	0.33	0.22	0.70
Subfamily	55	4921	89.5	1016	0.43	0.40	0.23	n/a

DISCUSSION AND FUTURE DIRECTIONS

Bacteriophage N4, an *E. coli* K-12 strain-specific bacteriophage isolated from the sewers of Genoa, Italy in the 1960s, is the founding member of the N4-like phage subfamily. Until recently, N4's novel transcriptional architecture, delayed lysis phenotype, and large burst size have been unique properties among well-characterized phages. Dozens of genome sequences for N4-like phages isolated from clinical and environmental samples have been added to available databases over the past 10 years. Here, I have utilized the increased phage sequence data available to identify the putative functions for 12 N4 ORFs and provide an updated annotation of the N4 genome. I have also provided a universal framework for the classification of N4-like phages using N4 vRNAP as a marker gene. By this definition, the N4-like phage subfamily currently consists of 55 species isolated over a wide geographic distribution that infect a broad range of Proteobacteria. All N4-like phages encode homologs of N4 vRNAP, N4 RNAPII, and

N4SSB, indicating that the N4 transcriptional architecture is the hallmark physiological feature of the N4-like phage subfamily.

Insights into N4 physiology through genome annotation

I identified putative functions for an additional 12 N4 ORFs using bioinformatics approaches. I have integrated these data with the functional characterization of other N4 elements from previous laboratory members to provide a comprehensive annotation of the location and sequences of all N4 transcriptional regulatory elements, tRNAs, and ORFs, which are summarized in Figure III.12.

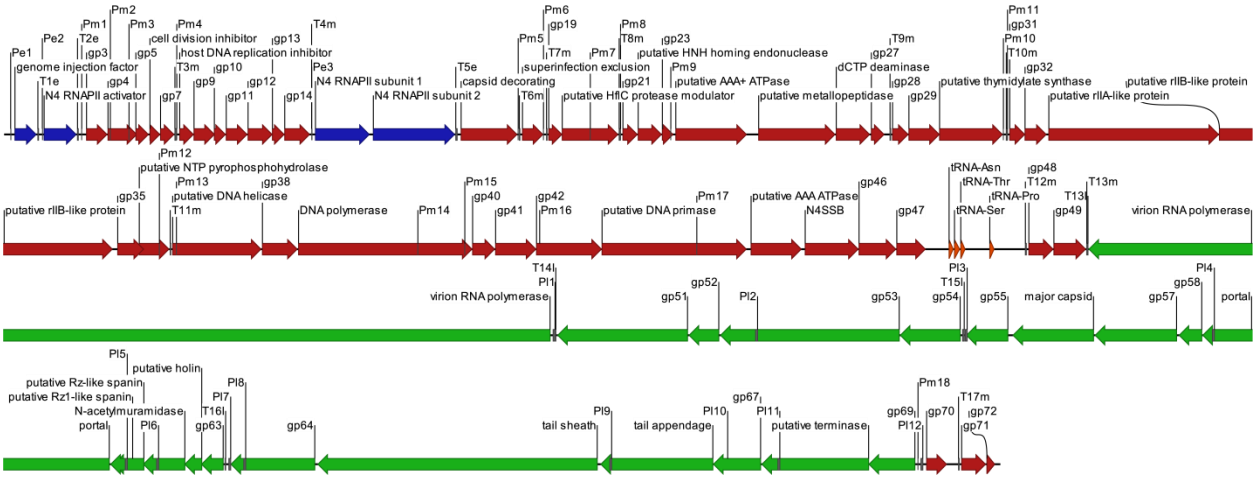

Figure III.12. N4 genome annotation. Schematic representation of the 70.2 kbp N4 genome, with all predicted ORFs, tRNAs, promoters, and termination sequences labeled with bioinformatics and functional annotations. Early genes (blue) and middle genes (red) are transcribed with rightward polarity from early promoters (Pe) and middle promoters (Pm), respectively. Late genes (green) are transcribed with leftward polarity from late promoters (Pl). Four tRNAs encoded by N4 (orange) lie in the gene-poor region of the genome between ORFs 47 and 48. Above each feature is the annotated function of the locus. ORFs are indicated by arrows and are named after functional or bioinformatics annotations. Hypothetical genes of unknown function are named for the gene product (gp).

In general, the bioinformatics approaches utilized to annotate N4 ORFs have provided limited insight into N4 biology. Despite the influx of bacteriophage sequences, the vast majority of N4 ORF BLAST searches failed to identify homologs of known function. In fact, the vast majority of N4 ORF BLAST hits belonged to N4-like phages that were merely annotated based

on homology to N4 ORFs. Other commonly used bioinformatics tools for homology detection such as HHpred utilize sequence databases for which bacteriophage sequences are underrepresented, resulting in the identification of proteins distantly related to those in N4 (247). While the ACLAME database of mobile genetic elements originally showed promise to identify protein families and assign functional annotations for large numbers of phage proteins, the database has not been updated for several years and has not been populated with recently sequenced phages for analysis (54). Therefore, the majority of N4 ORF annotations identified in this study were based largely upon the identification of a conserved domain signature in Interpro and the Conserved Domains Database analyses.

Through these methods, putative functions were assigned to 12 additional N4 ORFs; the vast majority of which correspond to middle genes (Table III.1 and Figure III.12). Although the putative functions assigned to N4 ORFs require validation through genetic and biochemical approaches, these studies may provide insight into N4 mechanisms of DNA replication, nucleotide biosynthesis, and host lysis.

Many bacteriophages degrade the host chromosome at the onset of infection and inhibit host transcription to reallocate host resources towards the synthesis of viral DNA and RNAs (29, 248, 249). N4, in contrast, does not degrade the host genome or fully inhibit host transcription, allowing the host cells to grow continuously throughout N4 infection (250, 251). N4 does, however, hijack the host σ^{70}-RNAP to initiate transcription from N4 late promoters and inhibits host DNA replication through the activity of gp8, which binds to the β clamp loader to shut down host DNA synthesis (22, 23, 34, 252). Although these proteins act to inhibit host macromolecular synthesis, N4 may require mechanisms to adjust the host nucleotide pool to increase the efficiency of N4 genome replication. Three middle ORFs putatively encode

enzymes responsible for nucleotide biosynthesis: i) ORF26 encodes a putative dCTP deaminase responsible for the deamination of dCTP to produce dUTP; ii) ORF36 encodes a putative NTP pyrophosphohydrolase responsible for the hydrolysis of NTPs to produce NMP plus pyrophosphate; iii) ORF30 encodes a putative thymidylate synthase responsible for the reductive methylation of dUMP to produce dTMP. These enzymes all belong to the dTTP biosynthetic pathway and may function to adjust the host nucleotide pool to produce dTTP for N4 genome replication to account for the differential G+C content between *E. coli* and N4 (50.6% vs. 41.3%, respectively).

One of the hallmark characteristics of bacteriophage N4 is the delayed lysis phenotype. N4 does not lyse the host cell until three hours after infection, releasing 3,000 pfu per infected cell (32). Gp61 was previously identified as an N-acetylmuramidase responsible for degradation of the host peptidoglycan (35). ORF61 is expressed as part of a polycistronic transcript containing ORFs 63-60, which together comprise the N4 lysis cassette (253). Bioinformatics analysis of the N4 lysis cassette uncovered putative functions for gp60 and gp62 (Table III.1). gp60 shared significant sequence homology with several Rz spanin proteins in BLAST searches. Analysis of ORF60 alternate reading frames revealed a secondary ORF (ORF60') located entirely within the codon of ORF60. N4 ORF60' is predicted to encode an Rz1 spanin protein. In canonical systems of gram-negative cell lysis, Rz/Rz1-spanin complexes mediate the fusion of the host inner and outer membranes to facilitate phage escape (254). Furthermore, ORF62 shared sequence homology with several putative phage holins, which are responsible for disruption of the host inner cell membrane. Structural predictions of gp62 suggested that this protein contains two transmembrane domains and a positively charged C-terminus, which are both characteristics of other well-defined phage holins (255).

The final component of the N4 lysis cassette, gp63, contains no conserved domains and no homologs of this protein were detected through BLAST search. The cytoplasmic gp63 protein was shown to be a positive regulator of N-acetylmuramidase activity: overexpression of gp63 leads to early cell lysis during N4 infection (253). N4-like phages within Clusters 1, 6, and 8 encode homologs of both N4 N-acetylmuramidase (gp61) and gp63, while the putative holin (gp62) and Rz/Rz1-like spanin (gp60/gp60') proteins are not universally conserved within N4-like phage lysis cassettes (Figure III.7). Interestingly, phages that encode an N4 gp63 homolog all contain putative N-acetylmuramidases with a positively charged N-terminal signal sequence, while N4 gp61 homologs in singleton phages that lack N4 gp63 homologs (vBEamP-S6, ZC03, and ZC08) have truncated N-termini (data not shown). These data suggest that homologs of N4 gp63 in Cluster 1, 6, and 8 phages may also positively regulate N-acetylmuramidase activity by promoting its release from the cytoplasmic membrane, as has been proposed for N4 (253).

Additionally, several gene products putatively involved in N4 DNA replication were also identified in N4 middle ORFs (Table III.1): i) ORF37 encodes a putative DNA helicase; ii) ORF43 encodes a putative DNA primase; iii) ORF44 encodes a putative AAA ATPase. Although AAA ATPases are involved in numerous cellular processes, the location of ORF44 between the putative DNA primase (ORF43) and N4SSB (ORF45) strongly suggests that this gene product is involved in N4 DNA replication (Figure III.12).

It has been postulated that many phages, specifically lytic phages, encode tRNA genes to compensate for the differences in codon bias between their own ORFs and those of the host to ensure efficient translation of phage mRNAs given the host tRNA pool (256). Analysis of the N4 genome by ARAGORN identified four N4-encoded tRNAs located in the gene poor region between ORFs 47 and 48, suggesting that this may be true for N4 (Figures III.1 & III.12).

However, N4-encoded tRNAs do not appear to ameliorate these codon usage differences to ensure N4 translational efficiency (Figure III.2). Therefore, the evolutionary advantage to maintain these N4-encoded tRNAs is unclear. These results suggest that N4 may have only recently acquired *E. coli* strain K-12 as a host species, which seems unlikely since N4 encodes numerous strategies for host takeover that require specific interactions between N4 and host proteins (e.g. gp6 interaction with host FtsA and FtsZ or gp8 interaction with host β clamp loader). Alternatively, the N4-encoded tRNAs may have been recently acquired in the evolutionary history of N4. This hypothesis is supported by the fact that phage-encoded tRNAs are not conserved throughout the N4-like phage subfamily (Figure III.6). The varying identity and location of tRNAs across N4-like phage clusters suggests that tRNAs are introduced through non-homologous recombination from sources outside the N4-like phage subfamily.

Identification and comparative genomics of N4-like phages

With the increased frequency of phage discovery and poor guidelines for general phylogenetic classification of viruses, I wished to provide a framework for the classification of newly sequenced or isolated phages to the N4-like phage subfamily. I propose using N4 vRNAP as a marker gene for the classification of phage species to the N4-like phage subfamily for two primary reasons. First of all, vRNAP is a 3,500 aa polypeptide that comprises ~15% of the N4 genome. Therefore, homologous proteins should be easy to identify in fully sequenced phage genomes and in isolated virions through SDS-PAGE analysis. Identification of vRNAP from purified virions of isolated phages would have the additional benefit of confirming RNAP encapsidation for N4 relatives. Secondly, vRNAP is involved in numerous processes throughout N4 infection and the presence of an N4 vRNAP homolog in other phage species would suggest

that they share similar mechanisms for genome injection, host-independent early transcription, and capsid maturation.

54 phages in current databases encode a homolog of N4 vRNAP containing all sequence motifs required for vRNAP activity (Table III.2). All phages encoding a homolog of N4 vRNAP share similar morphological and genomic properties, confirming that these are in fact close relatives of N4 and validating the marker gene approach for subfamily assignment (Tables III.3 & III.4). Despite having disparate geographic and host ranges, most N4-like phages were isolated from aqueous environments and hosts that have associations with the human microbiome (Table III.4). Correspondingly, a recent study of human gut metagenomics samples showed that N4-like phages were present in 99% of the 252 samples, while other phage families were present in only 20-60% of samples (257), suggesting that N4-like phages may play a role in shaping the human gut microbiome.

Hierarchical clustering of BLASTn alignment, ANI, and shared ORF comparisons sorted the 55 N4-like phages into eight clusters, while seven phages were left as singletons (Figures III.3, III.4, & III.9). In general, N4-like phage species cluster predominantly by host: i) Clusters 3, 6, 7, and 8 comprise phages that infect a single bacterial species; ii) Clusters 1, 2, and 5 comprise phages that infect a single bacterial family; iii) Cluster 4 phages, however, infect bacteria throughout the Gammaproteobacteria class. Despite infecting a wide range of host species, Cluster 4 phages do not contain a higher diversity of ORFams within their genomes or an enrichment of ORFans. Therefore, phages infecting closely related hosts share greater genetic similarity most likely due to shared evolutionary constraints and an increased opportunity for horizontal gene transfer.

The evolutionary history of N4-like phages is correlated with, although not completely dependent on the relationships with their bacterial hosts. Members of Cluster 8 all infect *Erwinia amylovora*, while members of Cluster 7 all infect *Pseudomonas aeruginosa*. Surprisingly, the Pseudomonas phages ZC03 and ZC08 exist as singletons and share very few ORFams with other Pseudomonas phages within Cluster 7, while the Erwinia phage vBEamP-S6 exists as a singleton and appears to share more ORFams with phages within Cluster 2 than the other Erwinia phages in Cluster 8.

I sorted all N4-like phage ORFs into ORFams to facilitate the discovery of ORFs of related function, create a reticulate phylogeny, and find shared ORFs associated with each phage cluster (Table III.5). The 20 ORFams completely conserved among N4-like phages comprise the "core genome" and suggest their importance in the N4-like lifestyle (Figure III.7). The majority are late genes that encode proteins of unknown function or generic phage proteins that are widely distributed outside of the N4-like phage subfamily. In contrast, homologs of the N4 transcriptional machinery (vRNAP, N4 RNAPII, and N4SSB) are exclusively found in N4-like phage genomes. Therefore, the N4 transcriptional strategy using a virion-encapsidated multi-domain RNAP for early transcription, a heterodimeric RNAPII for middle transcription, and SSB required for the coupling of late transcription and DNA replication is the hallmark feature used to define the N4-like phage subfamily.

ORFs responsible for other notable aspects of N4 physiology are less well conserved (Figure III.7). Homologs of proteins within the N4 lysis cassette were only found in 20 other phage species, while the N4 host cell division inhibitor (gp6) and host DNA replication inhibitor (gp8) are also poorly conserved among N4-like phages. These data suggest that many other N4-like phages utilize different methods of host takeover and display altered timing and mechanisms

of host lysis. These conclusions are well supported by the literature on N4-like phages, which describe N4-like phages with latent periods ranging from 10 min to 3 hrs and burst sizes ranging from 10 to 9,000 pfu per infected cell (Table III.4). Therefore, delayed host lysis and large burst sizes are not universal characteristics of N4-like phages.

The N4 tail sheath (gp65), responsible for interaction with the host cell receptor NfrA, has only two homologs among the N4-like phage subfamily (4) (Figure III.7). Surprisingly, N4 gp65 homologs were found in phiAxp-3 and RG-2014, which infect *Achromobacter xylosoxidans* and *Delftia tsuruhatensis*, respectively. Both of these bacterial species encode homologs of *E. coli* NfrA, however sequence alignment reveals limited sequence conservation among these receptor homologs. Therefore, the role of these N4 gp65 or *E. coli* NfrA homologs in host receptor recognition cannot be confirmed.

N4 RNAPII, responsible for bacteriophage N4 middle transcription, is incapable of initiating transcription from linear, dsDNA templates *in vitro* and requires the activity of gp2 to transcribe the N4 middle genes *in vivo* (18, 222). Although all N4-like phages encode homologs of N4 RNAPII, gp2 homologs were only detected within 32 other N4-like phage genomes (Figure III.7). These findings bring up numerous questions. Are these N4 RNAPII homologs capable of factor-independent transcription? If not, then do these phages encode an evolutionarily unrelated transcription factor? If so, do these N4 RNAPII homologs also recognize short, AT-rich promoters like other factor-dependent T7-like RNAPs? What sequence and structural changes exist between these RNAPs that determine each protein's dependence on transcription factors?

All N4 RNAPII homologs contain the conserved sequence blocks required for T7-like RNAP activity, suggesting that they are all functional RNAPs responsible for middle

transcription in N4-like phages (82). However, sequence alignment between N4 RNAPII homologs in gp2-encoding and gp2-deficient N4-like phages suggests that structural differences within the AT-rich recognition loop and β-IH may dictate the dependence of N4 RNAPII homologs on transcription factors (Figure III.13). All gp2-deficient phages encode gp15 homologs of greater length than N4 gp15. In gp2-deficient phages, the AT-rich recognition loop and β-IH lack sequence homology with the corresponding N4 RNAPII structures and contain insertions in these regions (Figure III.13). In contrast, gp1- and gp2-encoding phages contain gp15 homologs of equal length to N4 gp15 with strong sequence conservation at the AT-rich recognition loop and β-IH. Cluster 2 phages, which encode gp2 homologs but lack gp1 homologs, have extended AT-rich recognition loop and β-IH sequences similar to those found in phages lacking gp2. Cluster 5 phages, which also encode gp2 homologs but not gp1 homologs, lack sequence conservation with the N4 gp15 AT-rich recognition loop. The lack of sequence conservation and severe truncations in the AT-rich recognition loop and β-IH sequences of phages that contain N4 gp2 and N4 gp1 homologs suggest that these transcription factors may functionally complement the role of these structures during N4 RNAPII transcription initiation.

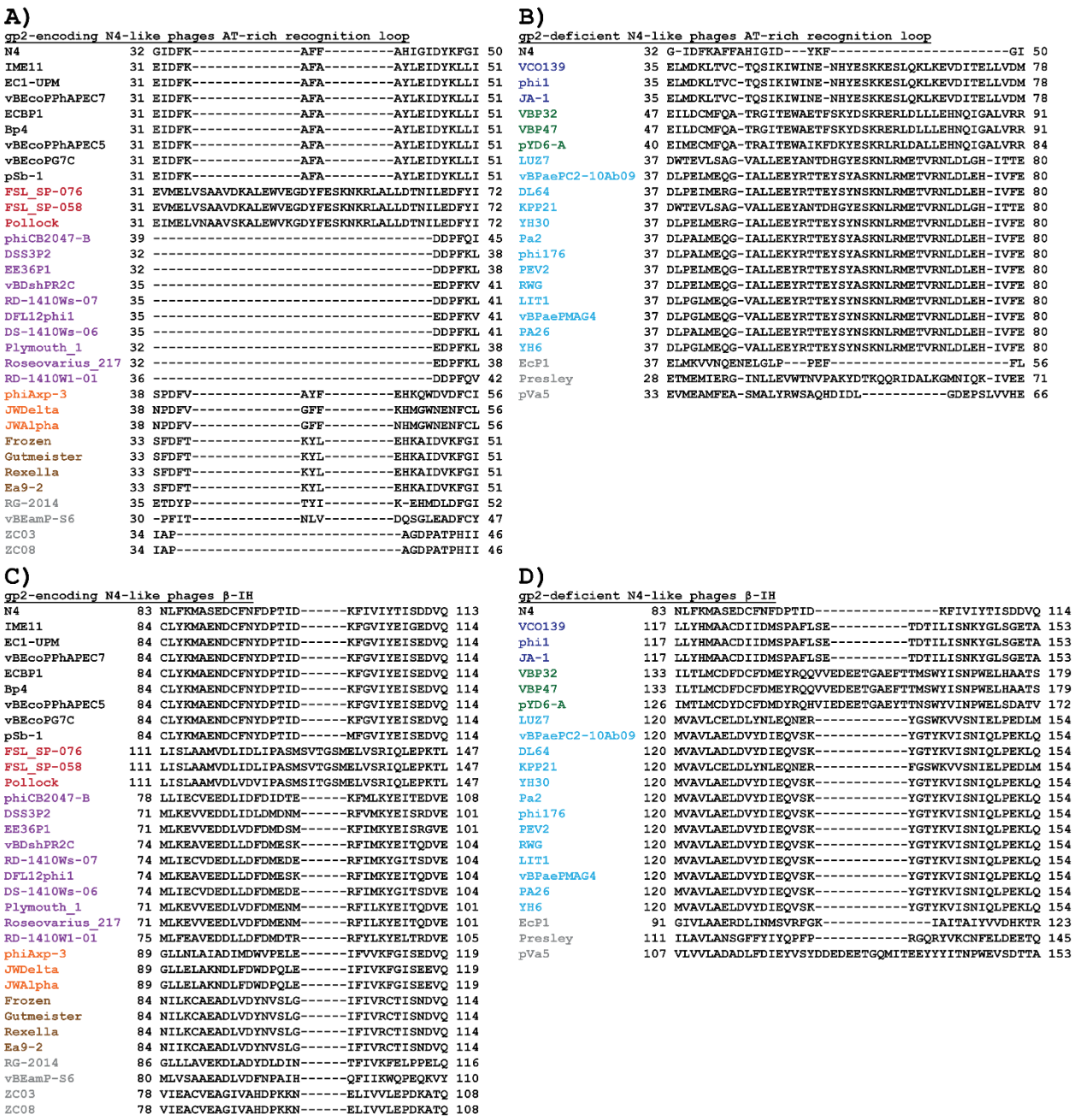

Figure III.13. N4 RNAPII AT-rich recognition loop and β-IH multiple sequence alignment. Clustal Omega multiple sequence alignment of the N4 RNAPII AT-rich recognition loop (A & B) and β-IH (C & D) between N4 RNAPII and its homologs in gp2-encoding N4-like phages (A & C) or gp2-deficient N4-like phages (B & D). Phage clusters color-coded as in Figure III.3.

Future directions

Although the expanded set of viral sequences in current databases have provided a valuable bioinformatics resource for the annotation of previously undefined N4 ORFs and

identification of ORFams in N4-like phages, the annotations presented here provide an incomplete picture of N4 and N4-like phage biology. Current bioinformatics tools and bacteriophage sequence databases are insufficient to identify close homologs of many N4-like phage ORFs, as 42.9% of the N4-like phage ORFams contain ORFans and only 32 of N4's 73 ORFs could be annotated. Additionally, the few annotations that could be assigned through bioinformatics methods are merely hypotheses for gene function and must be validated through additional genetic and biochemical approaches.

The analysis of the distribution of shared ORFams throughout the N4-like phage subfamily identified ORFams that were conserved among phages within a cluster, but absent from all other N4-like phages (Table III.6). These ORFams, termed Cluster Identifier ORFams, are components of the accessory genome and likely encode proteins that are non-essential, but might be positively selected to provide a fitness advantage in a given host species. These ORFs are largely unannotated and are excellent targets to discover host specificity factors or novel mechanisms for host takeover. Further characterization of these Cluster Identifier ORFams may lead to the development of new biotechnological tools or novel mechanisms for the inhibition of bacteria, which is increasingly important as the pipeline for novel antibiotics dwindles.

CHAPTER IV: MECHANISM OF N4 RNAPII PROMOTER RECOGNITION

INTRODUCTION

Transcription of bacteriophage N4 middle genes is performed by the phage-encoded heterodimeric N4 RNAPII along with its required transcription factor gp2 (15, 18). N4 RNAPII does not bind or initiate transcription from dsDNA templates *in vitro*, but can initiate transcription from fully or partially ssDNA templates such as supercoiled plasmids or bubbled DNA templates with low efficiency and limited sequence specificity (16, 17). These results suggest that N4 RNAPII contains the structural elements responsible for sequence-specific promoter recognition. However, the N4 RNAPII AT-rich recognition loop, β-IH, and specificity loop, which are required for promoter recognition in T7 RNAP, are poorly positioned to contact upstream DNA in the N4 RNAPII binary complex structure (Figure IV.1) (120, 223). In fact, no sequence-specific contacts were observed with upstream DNA, leaving the mechanism of N4 RNAPII promoter recognition unknown (Figure IV.1) (223). A previous graduate student (C. Markle) interrogated the role of the N4 RNAPII specificity loop in N4 RNAPII activity through alanine scanning mutagenesis of all polar and charged amino acids (residues 476-502) (all N4 RNAPII residues are numbered as a fusion of gp15-gp16 subunits), but failed to identify single residues required for promoter recognition *in vivo* or *in vitro* (224).

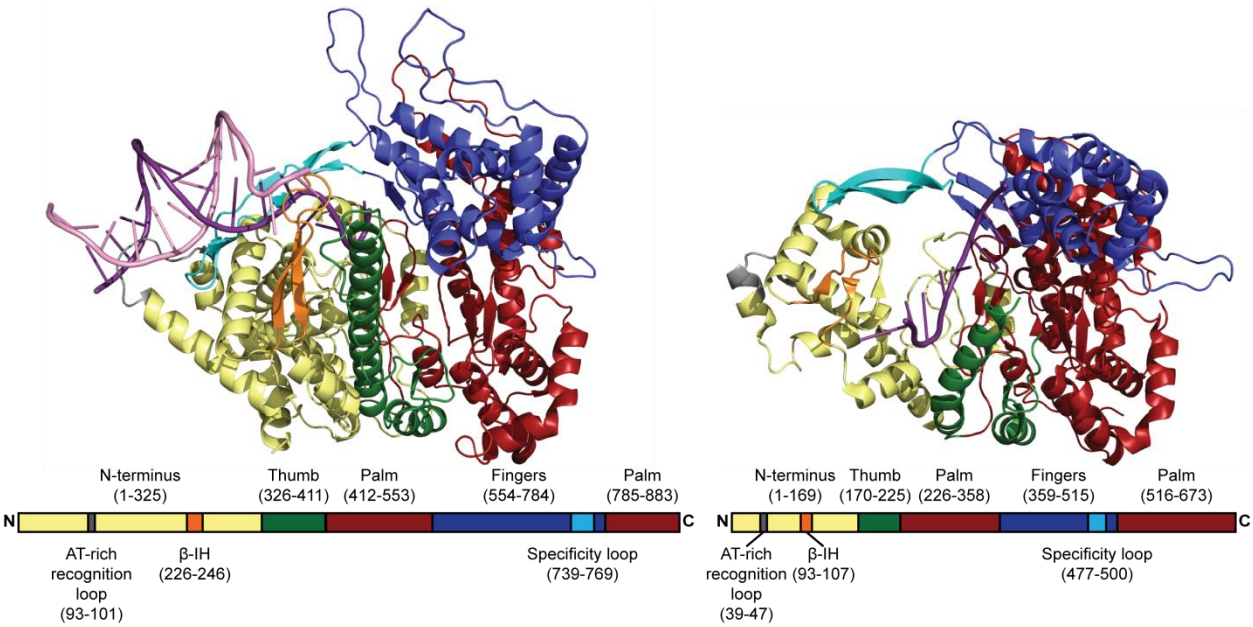

Figure IV.1. Structural comparison of T7 RNAP and N4 RNAPII. The cartoon representations of T7 RNAP (left, PDB: 1CEZ) and N4 RNAPII (right, PDB: 6DT7) binary complex structures are depicted. Bottom bars represent the primary sequences with amino acid numbering indicated in parentheses. N4 RNAPII residues are numbered as fusion of gp15-gp16 subunits. Domains and structural motifs are labeled and colored as in the crystal structures. N-terminus, yellow; thumb, green; palm, red; fingers, blue; AT-rich recognition loop, grey; β-IH, orange; specificity loop, cyan; template-strand DNA, purple; non-template strand DNA, pink. T7 RNAP structure adapted from Cheetham et al. (120); N4 RNAPII structure adapted from Molodtsov and Murakami (223).

The sequence requirements for N4 middle transcription initiation determined by Michael Hammer, a previous graduate student, are summarized below. He identified 18 N4 RNAPII sites of transcription initiation (designated Pm1-18), which he mapped to the N4 genome through RLM RT-PCR/TAP analysis followed by primer extension (19). The consensus N4 RNAPII promoter revealed a short, AT-rich sequence with no conservation upstream of -7 (Figure IV.2A) (19). Deletion mutagenesis confirmed the sequence conservation data and showed that N4 RNAPII promoter sequences are fully contained within the 10 bps upstream of transcription initiation sites. Saturation mutagenesis of a plasmid-resident Pm5 sequence showed that substitutions to any base at the -10, -6, -5, -1, and +1 position reduced transcription ($\leq$40% WT Pm5 activity) (Figure IV.2B) (19). However, plasmid-based transcriptional activity under these

conditions subsumes the processes of promoter unwinding, gp2 binding, N4 RNAPII recruitment, promoter recognition, transcription initiation, and promoter clearance.

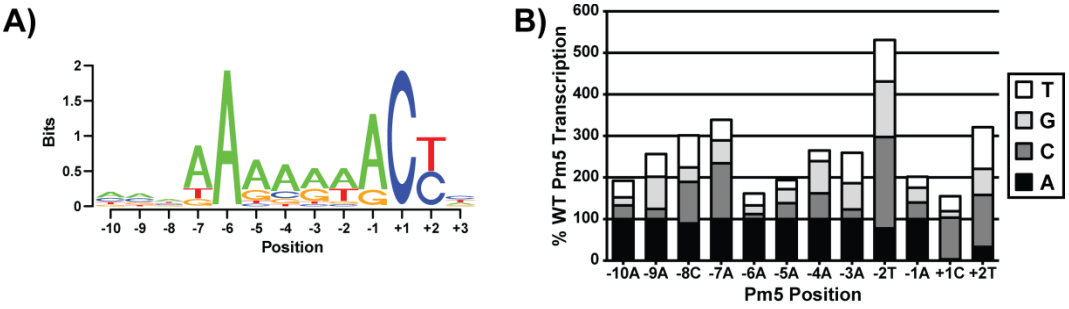

Figure IV.2. Definition of N4 RNAPII promoters. A) Consensus N4 RNAPII promoter sequence. Sequence logo of template-strand sequences from -10 to +3 (relative to transcription initiation site at +1). B) Relative transcriptional activity of Pm5 base substitutions *in vivo*. Primer extension reactions with 5' end-labeled primers detect transcripts from plasmid-resident Pm5 saturation mutagenesis templates analyzed by 8% PAGE. Relative expression plotted as percent WT Pm5 transcription for each base. WT Pm5 sequence template-strand sequence indicated on horizontal axes. All panels adapted from M. Hammer data (19).

In this chapter, I aimed to interrogate the sequence requirements for N4 RNAPII promoter binding and start site selection in isolation from other steps of transcription initiation and directly test the role of the N4 RNAPII specificity loop structure in sequence-specific promoter recognition. Results show that N4 RNAPII recognizes template-strand sequences at the -7, -6, -3, and -1 position of promoters through direct interactions with residues of the specificity loop.

SEQUENCE REQUIREMENTS FOR N4 RNAPII PROMOTER RECOGNITION

To isolate the nucleotides responsible for promoter recognition by N4 RNAPII in the absence of all other factors, I repeated the Pm5 saturation mutagenesis (-10 to +2) *in vitro* using ssDNA templates and purified recombinant WT N4 RNAPII in runoff transcription assays (Figure IV.3A). Single nucleotide substitutions at -10, -9, and -4 had little effect on N4 RNAPII

activity, while all substitutions at -7, -6, and +1 significantly reduced activity ($p \leq 0.05$, student's t-test). Template-strand purines were strongly favored at the -1 and -3 positions. -5A→T, -8C→G, and +2T→A substitutions significantly reduced transcription, while -2T→A significantly increased transcription from the Pm5 promoter.

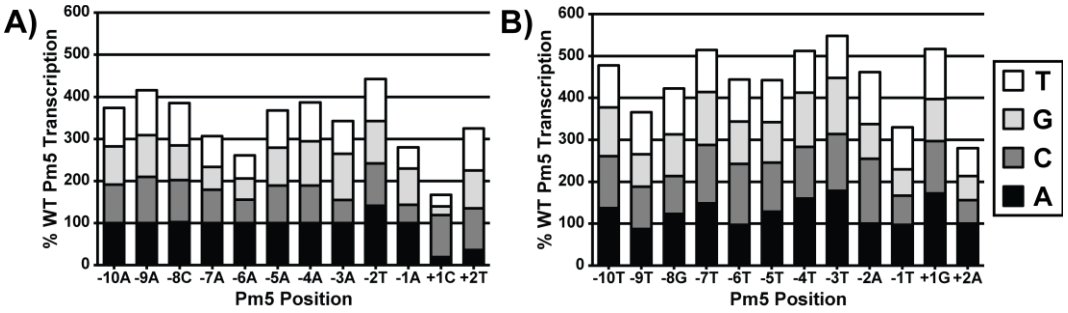

Figure IV.3. *In vitro* saturation mutagenesis of Pm5. A) Runoff transcription reactions of recombinant N4 RNAPII (50 nM) activity from Pm5 template-strand saturation mutagenesis (-10 to +2) ssDNA templates (50 nM) analyzed by 8% PAGE. B) Runoff transcription reactions of recombinant N4 RNAPII (50 nM) activity from Pm5 non-template strand saturation mutagenesis (-10 to +2) bubbled DNA templates (100 nM) analyzed by 8% PAGE. Bubbled templates created by annealing non-template strand oligonucleotides containing non-complementary bases from -10 to +4 to WT ssDNA Pm5 template-strand oligonucleotides. Relative expression from four independent replicates plotted as percent WT Pm5 transcription for each base. WT Pm5 template-strand (A) or non-template strand (B) sequences indicated on horizontal axes.

To determine whether N4 RNAPII exclusively interacts with nucleotides in the template strand, I tested N4 RNAPII transcriptional activity from bubbled templates *in vitro*. Pm5 non-template strand oligonucleotides containing mismatches spanning -10 to +4 were annealed to Pm5 template-strand oligonucleotides, creating runoff transcription templates with ssDNA from -10 to +4 (Figure IV.4A). N4 RNAPII transcription initiated from WT Pm5 bubbled templates was reduced nearly 2-fold relative to the transcriptional activity observed from fully ssDNA WT Pm5 templates, with transcription primarily initiating from the same site in both templates (Figure IV.4B). Filter binding experiments show that N4 RNAPII affinity for bubbled templates (K_d 112±5 nM) is reduced 1.5-fold relative to N4 RNAPII affinity for ssDNA templates (K_d 76±2 nM), partially explaining the reduction in transcriptional activity observed in runoff transcription

assays (Figure IV.4C). To determine whether the presence of the non-template strand influences the N4 RNAPII sequence preference for template-strand nucleotides, I tested N4 RNAPII transcriptional activity from bubbled templates containing single-nucleotide substitutions in the template strand. All template-strand substitutions at the -6 and -7 positions and pyrimidine substitutions at the -1 and -3 positions of bubbled templates significantly reduced N4 RNAPII activity in runoff transcription assays (Figure IV.4D). These base preferences closely match those observed from ssDNA Pm5 templates, confirming that the presence of the non-template strand in bubbled templates does not alter N4 RNAPII preference for template-strand sequences.

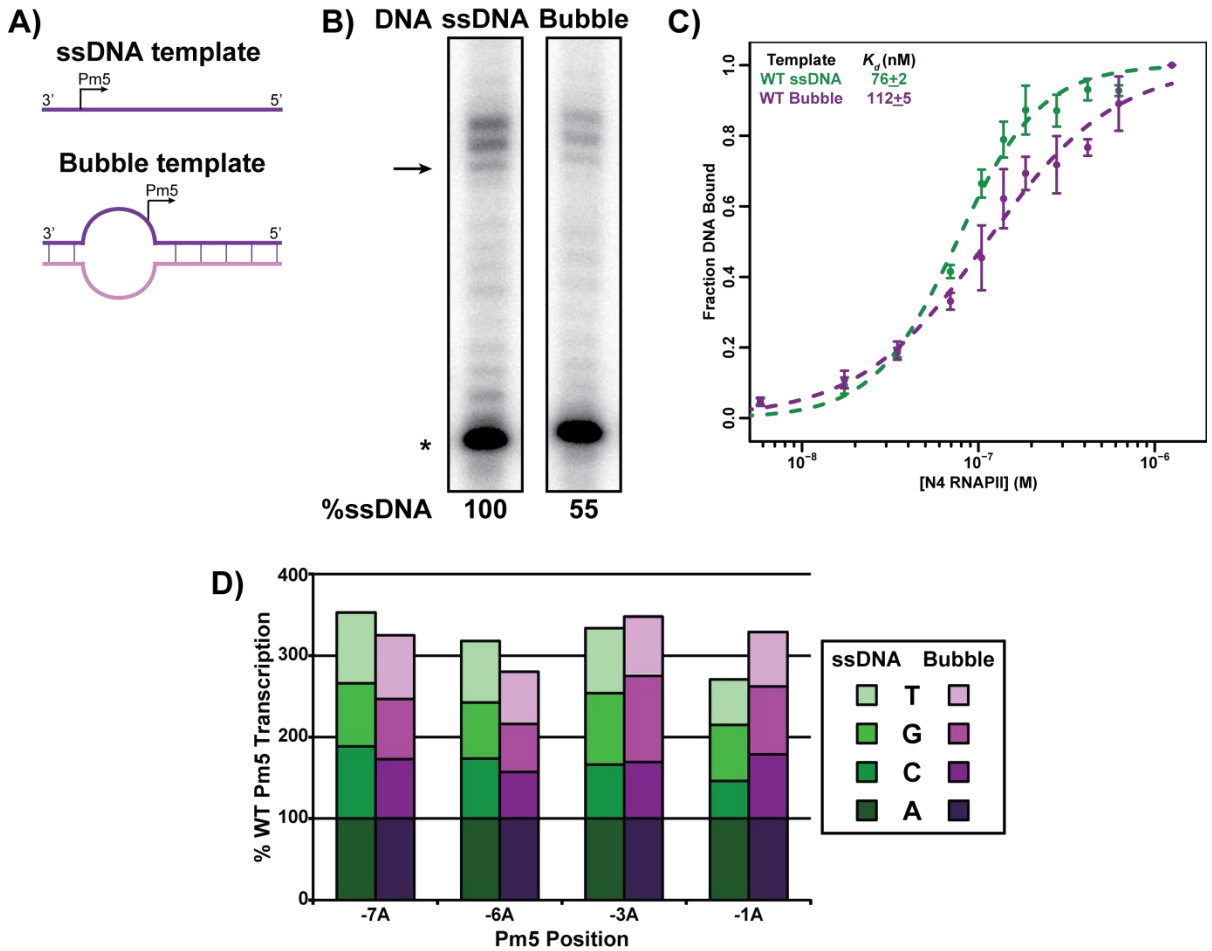

Figure IV.4. N4 RNAPII initiates transcription from bubbled templates with lower activity and unaltered sequence specificity. A) Cartoon representation of ssDNA and bubble templates used in this study. Template strand, purple; mismatched non-template strand, pink. B) N4 RNAPII initiates transcription from bubbled templates with reduced activity relative to ssDNA

Figure IV.4 (continued)
templates. Representative 8% PAGE autoradiogram of N4 RNAPII (50 nM) runoff transcription products from ssDNA or bubbled Pm5 DNA templates (100 nM). Relative activity from three independent replicates represented as percent ssDNA transcription below each lane. Arrow, expected 37 nucleotide runoff product; asterisk, loading control. C) N4 RNAPII binds to bubbled templates with reduced affinity relative to ssDNA templates. Binding of N4 RNAPII to ssDNA and bubbled Pm5 templates was measured by filter binding. 5' end-labeled ssDNA templates (10 pM) were incubated with increasing concentration of N4 RNAPII, applied to nitrocellulose filter, and analyzed by phosphorimaging. The normalized fraction bound (filled circles±SD) from four independent replicates was calculated according to Equation 5. Estimates of the dissociation constant (K_d±SEM) calculated through non-linear least squares regression to hill equation (dashed lines) are included for each template. D) The presence of the non-template strand does not alter N4 RNAPII template-strand sequence preference. Runoff transcription reactions of N4 RNAPII (50 nM) from ssDNA or bubbled Pm5 DNA templates (100 nM) with single-nucleotide substitutions in the template strand analyzed by 8% PAGE. Relative expression from three independent replicates plotted as percent WT Pm5 transcription for each base. WT Pm5 template-strand sequence indicated on horizontal axis.

To determine whether specific sequences within the non-template strand are recognized by N4 RNAPII, I performed saturation mutagenesis (-10 to +2) of Pm5 non-template strand sequences in bubbled templates and tested N4 RNAPII activity through runoff transcription assays (Figure IV.3B). Although, non-template strand substitutions did not reduce N4 RNAPII transcription as dramatically as substitutions to the template strand in aggregate, -1T→C, -1T→G, +2A→T, +2A→C, and +2A→G non-template strand substitutions did significantly reduce transcription from bubbled Pm5 templates. These data do not correlate with *in vivo* saturation mutagenesis and N4 RNAPII promoter sequence conservation data, suggesting that all N4 RNAPII sequence-specific contacts required for activity *in vivo* are limited to the template strand (Figures IV.2A, B, & IV.3B) (19).

The differences in promoter sequence preference observed between *in vivo* and *in vitro* transcription assays may be the result of gp2 altering the sequence preference of N4 RNAPII *in vivo*. To determine whether gp2 influences the N4 RNAPII sequence preference, I repeated the Pm5 ssDNA saturation mutagenesis (-7 to +1) *in vitro* runoff transcription assays in the presence

and absence of WT gp2 (Figure IV.5). Upon addition of WT gp2, N4 RNAPII template-strand sequence preference did not significantly change for any base except for one. In the absence of gp2, the -2T→C substitution had no effect on N4 RNAPII activity, but significantly reduced N4 RNAPII activity (<70% WT Pm5 activity) upon addition of gp2. These results show that gp2 does not function in promoter selection and support data indicating that gp2 binds to ssDNA non-specifically (Figure IV.5) (18).

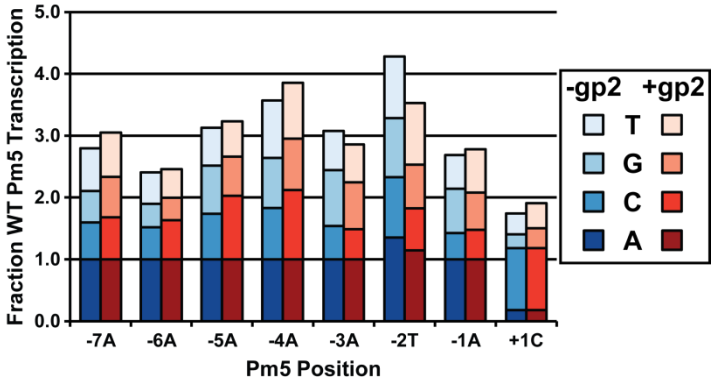

Figure IV.5. Gp2 does not alter N4 RNAPII sequence preference. Runoff transcription reactions of N4 RNAPII (10 nM) from Pm5 template strand saturation mutagenesis (-7 to +1) ssDNA templates (50 nM) activated by gp2 (0, 200 nM) analyzed by 8% PAGE. Relative transcription from four independent replicates plotted as fraction WT Pm5 transcription for each base. WT Pm5 template-strand sequence indicated on horizontal axis.

The lack of sequence conservation upstream of the -7 position suggests that N4 RNAPII promoters are not bipartite like T7 RNAP promoters, which have an upstream binding region and downstream initiation region. To confirm the absence of an upstream binding region in N4 RNAPII promoters, I measured the effect of single-nucleotide substitutions in Pm5 ssDNA templates on N4 RNAPII affinity through filter binding assays. I determined the fraction DNA bound across increasing concentrations of N4 RNAPII and determined the affinity of N4 RNAPII for each template sequence tested (Figure IV.6). N4 RNAPII binds to WT Pm5 ssDNA templates with a K_d of 76±2 nM and binds to ssDNA templates lacking a promoter sequence with a K_d of 116±2 nM: a 1.5-fold difference in binding preference (Figure IV.6, WT vs. NT). Pm5

substitutions that significantly reduced N4 RNAPII transcription *in vitro* also modestly reduce N4 RNAPII affinity. -1A→T, -3A→T, and -7A→G Pm5 substitutions reduced N4 RNAPII affinity as much as ssDNA templates lacking promoter sequences (K_d ≥116 nM), confirming the importance of these positions in N4 RNAPII promoter recognition (Figure IV.6). Although single-nucleotide substitutions to Pm5 ssDNA templates do reduce N4 RNAPII affinity, these substitutions are distributed throughout N4 RNAPII promoters and do not cluster together into a defined binding region. Furthermore, the small differences in affinity observed here do not fully explain the large disparities in promoter activity observed in transcription assays *in vitro* or *in vivo*, suggesting that N4 RNAPII transcription is only modestly regulated by preferential binding of N4 RNAPII to promoter sequences.

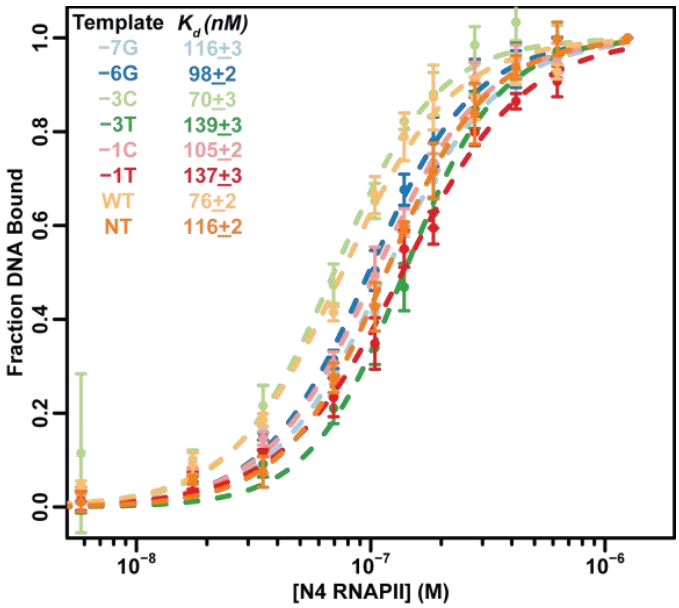

Figure IV.6. Promoter substitutions reduce N4 RNAPII promoter binding. Binding of N4 RNAPII to substituted Pm5 ssDNA templates was measured by filter binding. 5' end-labeled ssDNA templates (10 pM) were incubated with increasing concentration of N4 RNAPII, applied to nitrocellulose membrane, and analyzed by phosphorimaging. The normalized fraction bound (filled circles±SD) from four independent replicates was calculated according to Equation 5. Estimates of the dissociation constant (K_d±SEM) were calculated through non-linear least squares regression to the hill equation (dashed lines) are included for each template. NT, non-promoter template.

To determine whether deleterious Pm5 substitutions affect start site selection, I identified substitutions that alter the length of the RNA product from runoff transcription reactions. Runoff transcription from WT Pm5 ssDNA templates yields a 37 nt RNA runoff product initiated with GTP and extended up to two additional nucleotides at the 3' end through non-templated addition, a common characteristic among T7-like RNAPs (Figure IV.7A & B) (258, 259). Non-templated addition of one or two nucleotides at the 3' end of transcripts was confirmed through runoff transcription assays using [γ^{32}P] GTP to label the 5' end of transcripts (Figure IV.7B, lane 13). Runoff transcription initiating from -1C and -3C Pm5 ssDNA templates produces RNA two and three nucleotides longer, respectively (Figure IV.7A). Transcription from the -1C Pm5 ssDNA template initiates with [γ^{32}P] ATP two nucleotides upstream of the WT Pm5 start site, while -3C Pm5 ssDNA template initiates with [γ^{32}P] GTP three nucleotides upstream of WT Pm5 start site (Figure IV.7B, lanes 9 & 15). Interestingly, -1C and -3C Pm5 substitutions appear to affect start site selection rather than N4 RNAPII affinity, as cytidine substitutions at -1 and -3 positions in Pm5 ssDNA templates had less impact on N4 RNAPII binding than thymidine substitutions at the same positions (Figure IV.6). These results confirm that the change in transcript length is due to change in transcription start site and that residues within N4 middle promoters are responsible for start site determination and transcription initiation.

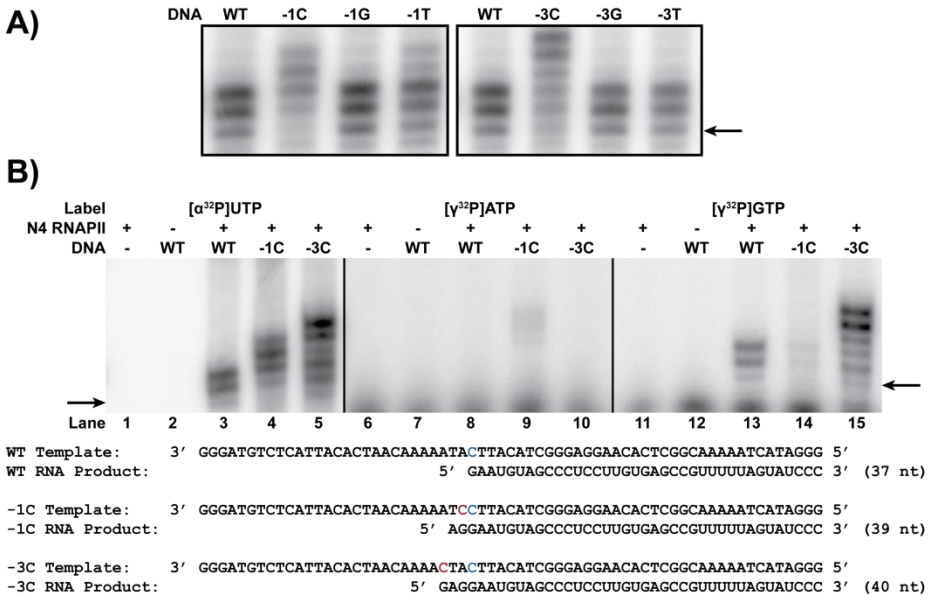

Figure IV.7. N4 RNAPII promoter substitutions shift start site selection. A) Base substitutions at -1 and -3 positions of Pm5 alter N4 RNAPII start site selection. Representative 8% PAGE autoradiograms of N4 RNAPII (50 nM) runoff transcription products from Pm5 ssDNA templates (50 nM) with the indicated nucleotide changes at positions -1 (left) and -3 (right). B) N4 RNAPII performs non-templated addition at the 3' end of transcripts and initiates transcription from alternate start sites from -1C and -3C Pm5 ssDNA templates *in vitro*. Representative 8% PAGE autoradiogram of N4 RNAPII (50 nM) runoff transcription products from WT, -1C, or -3C Pm5 ssDNA templates (100 nM) with the indicated radionucleotide as either the initiating ([γ^{32}P] ATP and [γ^{32}P] GTP, 20 μCi) or body labeling ([α^{32}P] UTP, 5 μCi) nucleotide. Lanes 6-15 are overexposed to compensate for decreased labeling efficiency for [γ^{32}P]NTPs. WT and Pm5 substitution template sequences, along with their expected RNA products, are listed (bottom). Arrow, expected 37 nucleotide RNA product; +1 nucleotides, blue; Pm5 substitutions, red.

The introduction of multiple mutations to the Pm5 promoter corroborates the single-base template strand saturation mutagenesis results (Figure IV.8). N4 middle promoter function is completely lost in runoff transcription assays initiating from either -1T,-3T or -6G,-7G double-substitution Pm5 ssDNA templates. Transcription from the -1C,-3C Pm5 ssDNA template initiates robustly, but from three nucleotides upstream of the WT Pm5 start site, matching the start site shift observed from the -3C single-substitution Pm5 ssDNA template. Combinations of these four mutations further exacerbate these characteristics (Figure IV.8, lanes 13 & 14). These

data support the importance of the -1, -3, -6, and -7 positions in N4 RNAPII promoter recognition *in vitro*.

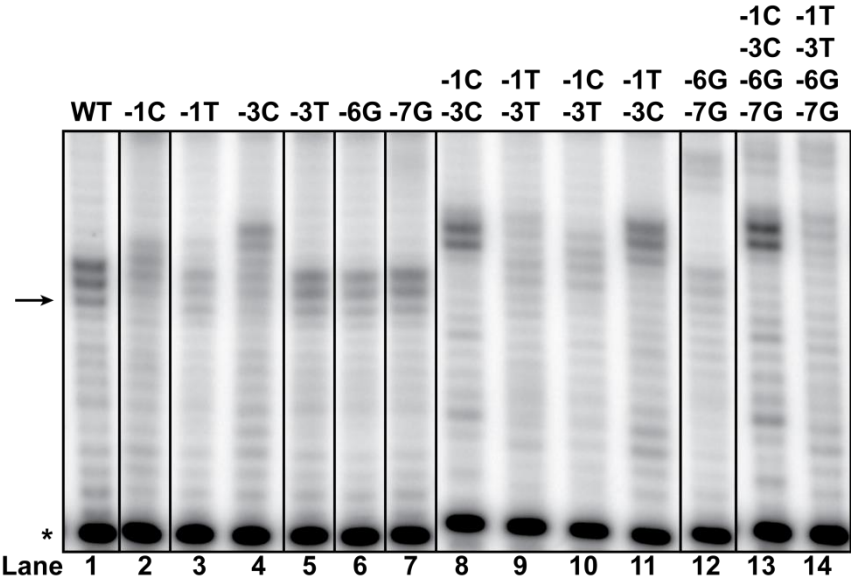

Figure IV.8. Introduction of multiple substitutions to Pm5 templates destroys promoter function. Representative 8% PAGE autoradiogram of N4 RNAPII (50 nM) runoff transcription products from Pm5 ssDNA templates (100 nM) with the indicated nucleotide substitutions. Arrow, expected 37 nucleotide runoff RNA product; asterisk, loading control.

THE N4 RNAPII SPECIFICITY LOOP IS RESPONSIBLE FOR PROMOTER RECOGNITION

To directly test whether the N4 RNAPII specificity loop is required for promoter recognition, I purified N4 RNAPII specificity loop alanine-substituted alleles (cloned by C. Markle) to homogeneity and tested each allele for deficiencies in transcriptional activity from WT Pm5 ssDNA templates *in vitro* (Figure IV.9). The E483A, Y492A, and C494A N4 RNAPII alleles showed significantly reduced transcription relative to WT N4 RNAPII ($p \leq 0.05$, student's t-test) in runoff transcription assays, while the E481A, K490A, and R496A N4 RNAPII alleles showed a moderate, although not statistically significant, reduction in transcription (Figure IV.9). These results show that specific residues within the N4 RNAPII specificity loop are required for

efficient transcription *in vitro* and may directly contribute to promoter recognition, although runoff transcription assays do not specifically determine which stage of transcription is affected.

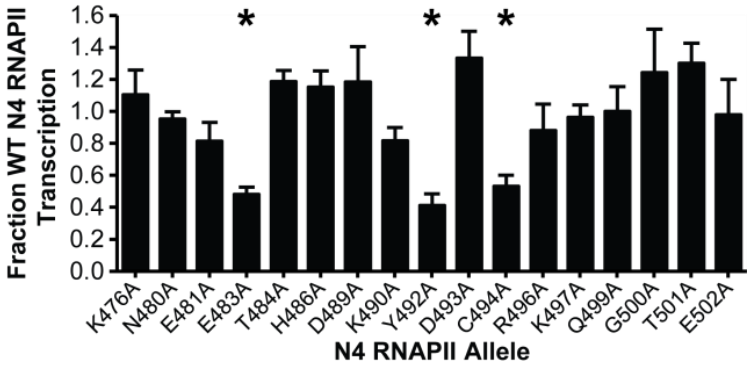

Figure IV.9. N4 RNAPII specificity loop substitutions reduce runoff transcription *in vitro*. Runoff transcription reactions of indicated N4 RNAPII specificity loop alanine-substituted alleles (50 nM) from Pm5 ssDNA templates (100 nM) analyzed by 8% PAGE. Relative transcription from three independent replicates plotted as fraction WT N4 RNAPII transcription for each N4 RNAPII allele. Error bars, SD; asterisks, statistically reduced runoff transcription relative to WT N4 RNAPII ($p \leq 0.05$, student's t-test). N4 RNAPII residues are numbered as fusion of gp15-gp16 subunits.

To determine whether the N4 RNAPII specificity loop alanine-substituted alleles are deficient in promoter recognition, I repeated the template-strand Pm5 saturation mutagenesis using ssDNA templates and N4 RNAPII specificity loop alleles in runoff transcription assays. I screened for N4 RNAPII alleles that did not discriminate against down promoter substitutions as efficiently as WT N4 RNAPII. Y492A N4 RNAPII showed increased specificity against pyrimidines at the -3 position, while C494A N4 RNAPII showed decreased discrimination against pyrimidines at the -1 position relative to WT N4 RNAPII (Figure IV.10). These results suggest that the C494 residue of N4 RNAPII makes specific contacts with the -1 position and the Y492 residue makes specific contacts with the -3 position of N4 RNAPII promoters.

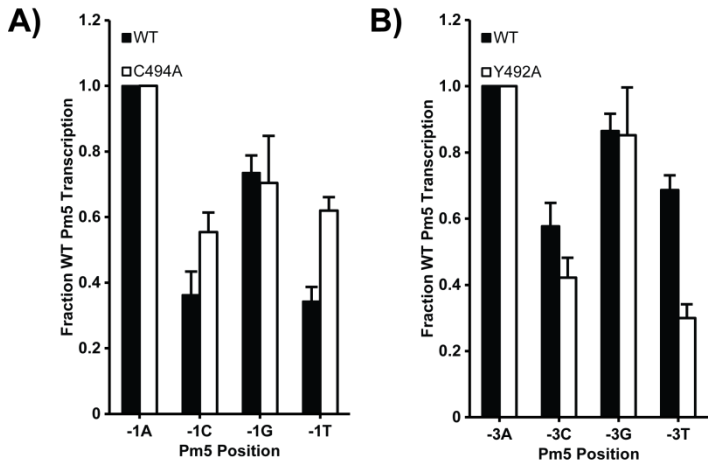

Figure IV.10. N4 RNAPII specificity loop alleles alter promoter specificity. Runoff transcription reactions of C494A (A) and Y492A (B) N4 RNAPII specificity loop alanine-substituted alleles (50 nM) from Pm5 ssDNA templates (100 nM) with the indicated nucleotide substitutions analyzed by 8% PAGE. Relative transcription from three independent replicates plotted as fraction WT Pm5 transcription for each base. Error bars, SD. N4 RNAPII residues are numbered as fusion of gp15-gp16 subunits.

To confirm direct interactions between the N4 RNAPII specificity loop and ssDNA templates, I site-specifically incorporated the photocrosslinking unnatural amino acid *p*-benzoyl-L-phenylalanine (*p*Bpa) into N4 RNAPII residues required for transcription *in vitro* (E481, E483, Y492, C494, and R496). *p*Bpa reacts with normally inert carbon-hydrogen bonds upon excitation with 350-360 nm ultraviolet (UV) light (260, 261). Upon UV excitation, E481Bpa and R496Bpa N4 RNAPII efficiently crosslinked to Pm5 ssDNA templates, observed as radiolabeled bands migrating with an apparent molecular weight of 100 kDa (Figure IV.11, lanes 5 & 13). The presence of two radiolabeled bands in the R496Bpa N4 RNAPII reactions corresponds to the crosslinking of one and two *p*Bpa-substituted gp16 subunits to a single ssDNA template, suggesting that the R496Bpa N4 RNAPII has reduced promoter specificity (Figure IV.11, lane 13).

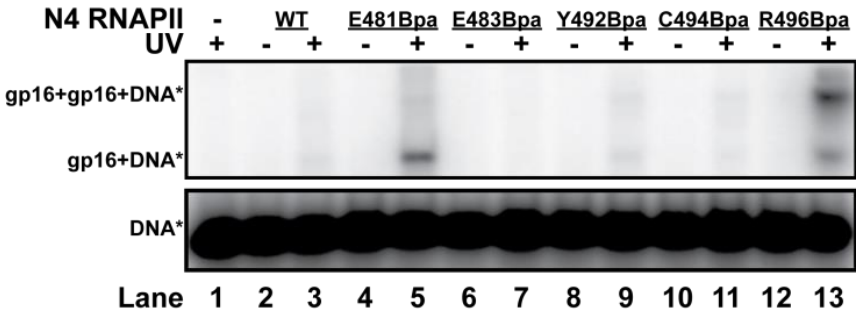

Figure IV.11. N4 RNAPII specificity loop crosslinks to ssDNA promoter templates. Indicated N4 RNAPII *p*Bpa-substituted alleles (0, 1 μM) were incubated with 5' end-labeled ssDNA Pm5 templates (100 nM) followed by UV crosslinking (365 nm, 90 min). Representative phosphorimage of products separated by 10% SDS-PAGE shown with bands of interest labeled (left). N4 RNAPII residues are numbered as fusion of gp15-gp16 subunits.

The failure of E483Bpa, Y492Bpa, and C494Bpa N4 RNAPII to efficiently crosslink to ssDNA templates under any conditions tested was surprising given the large impact the corresponding alanine substitutions had on N4 RNAPII transcriptional activity (Figures IV.9 & IV.11). These results could be obtained if these residues are poorly positioned to make contact with ssDNA or if the incorporation of the bulky benzophenone group reduces N4 RNAPII's ability to recognize and bind ssDNA templates. To differentiate between these possibilities, I tested each alanine- and *p*Bpa-substituted N4 RNAPII specificity loop allele for deficiencies in runoff transcription activity from WT Pm5 ssDNA templates (Figure IV.12). E481Bpa and R496Bpa N4 RNAPII alleles did not show reduced activity relative to WT N4 RNAPII, consistent with the ability of these enzymes to efficiently crosslink to ssDNA templates. *p*Bpa incorporation at E483 and C494 reduced N4 RNAPII activity to a greater degree than the corresponding alanine-substituted alleles, suggesting that the lack of crosslinking to promoter DNA is due to the reduction in ssDNA recognition and not because of improper amino acid positioning.

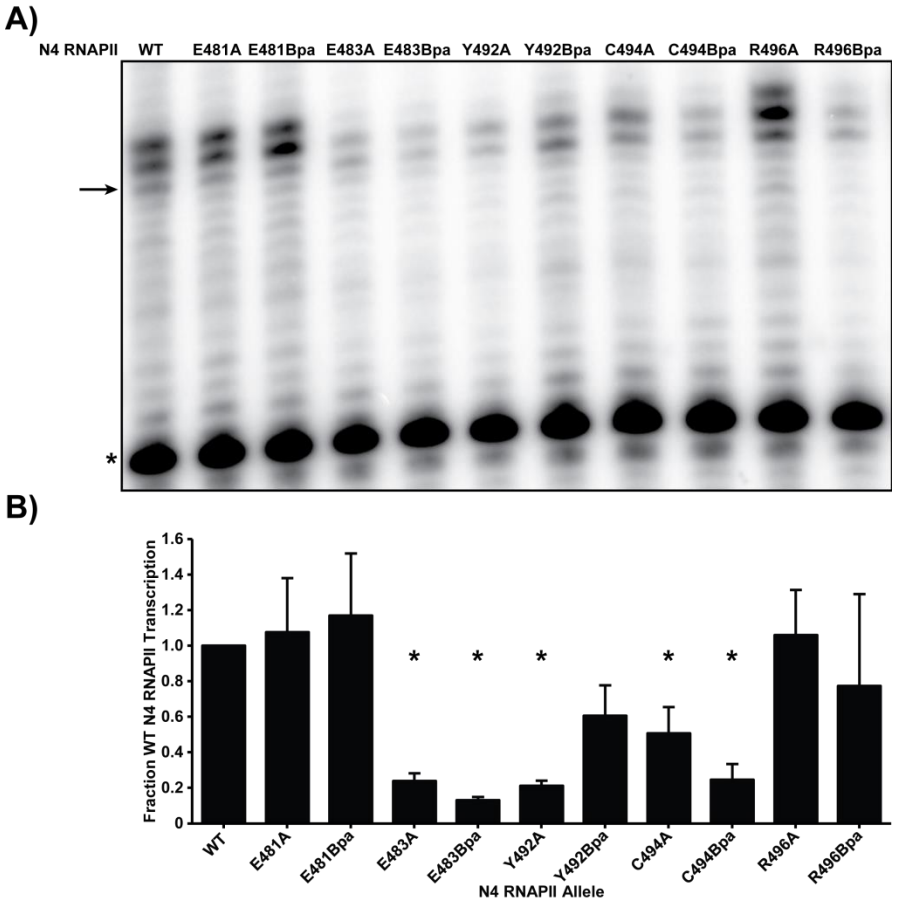

Figure IV.12. N4 RNAPII *p*Bpa specificity loop substitutions reduce runoff transcription *in vitro*. A) Representative 8% PAGE autoradiogram of N4 RNAPII specificity loop alanine- and *p*Bpa-substituted alleles (50 nM) runoff transcription products from WT Pm5 ssDNA templates (100 nM). Arrow, expected 37 nucleotide runoff RNA product; asterisk, loading control. B) Relative transcription from three independent replicates plotted as fraction WT N4 RNAPII transcription for each N4 RNAPII allele. Error bars, SD; asterisks, statistically reduced runoff transcription relative to WT N4 RNAPII ($p \leq 0.05$, student's t-test). N4 RNAPII residues are numbered as fusion of gp15-gp16 subunits.

Interestingly, *p*Bpa incorporation at Y492 did not significantly reduce N4 RNAPII transcription relative to WT N4 RNAPII. In fact, the presence of the benzophenone group at this position increased transcription approximately 3-fold over the Y492A allele (Figure IV.12). These results suggest that the presence of an aromatic residue at this position is required for promoter binding and recognition. Indeed, the Y492F N4 RNAPII allele also displayed

comparable transcriptional activity to WT N4 RNAPII in runoff transcription assays and increased transcription approximately 2-fold over the Y492A allele (Figure IV.13).

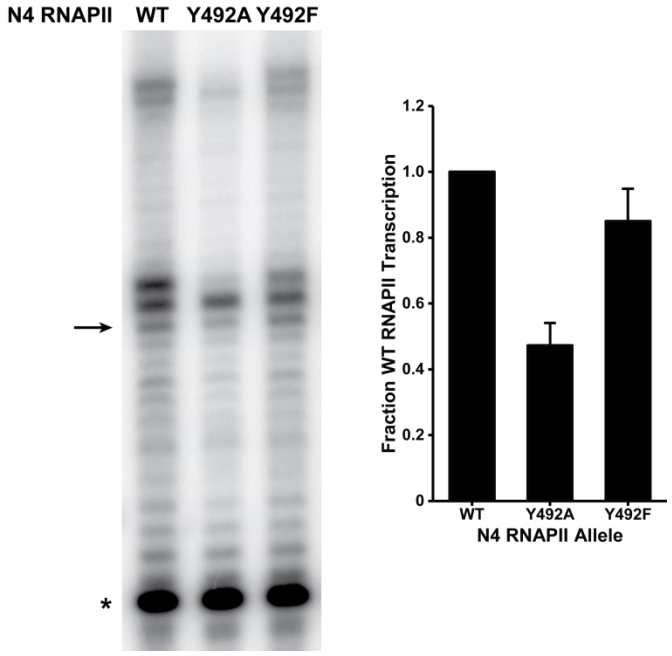

Figure IV.13. Aromatic amino acids are preferred at the Y492 position of N4 RNAPII. Representative 8% PAGE autoradiogram of N4 RNAPII Y492 alleles (50 nM) runoff transcription products from WT Pm5 ssDNA templates (100 nM) (left). Arrow, expected 37 nucleotide runoff RNA product; asterisk, loading control. Relative transcription from three independent replicates plotted as fraction WT N4 RNAPII transcription for each N4 RNAPII Y492 allele (right). Error bars, SD. N4 RNAPII residues are numbered as fusion of gp15-gp16 subunits.

In this chapter, I have shown that N4 RNAPII recognizes specific sequences in the template strand of AT-rich promoters through direct interactions with the specificity loop. These results and future directions are further examined in the discussion at the end of Chapter V.

CHAPTER V: MECHANISMS OF N4 RNAPII TRANSCRIPTION ACTIVATION BY GP2

INTRODUCTION

Gp2, which is tightly associated with the DNA-inner membrane complex during N4 infection, is required for N4 middle transcription *in vivo* (16, 222). *In vitro*, N4 RNAPII purified from the supernatant of N4-infected cells is inactive on native, double-stranded N4 DNA, but can initiate transcription from denatured N4 DNA or from the DNA-membrane complex of N4-infected cells (16). N4 ORF2 was cloned, purified to homogeneity, and characterized *in vitro* by a previous graduate student (R. Carter) to determine the role of gp2 in the localization and activation of N4 RNAPII transcription. Results indicate that gp2: i) is a non-sequence specific SSB; ii) stimulates N4 RNAPII transcription on ssDNA templates containing weak promoters; iii) binds cooperatively with N4 RNAPII to ssDNA; iv) directly interacts with N4 RNAPII with equimolar stoichiometry (18). Although required for N4 middle transcription *in vivo*, gp2 does not bind to dsDNA promoters and does not confer N4 RNAPII with the ability to initiate transcription from dsDNA promoters *in vitro* (18). Therefore, gp2 activates N4 RNAPII transcription primarily through recruitment of the polymerase to single-stranded templates, which are specifically recognized and melted by an unknown factor.

A previous graduate student (C. Markle) performed limited alanine scanning mutagenesis to identify gp2 residues required for interaction with N4 RNAPII and ssDNA-binding (summarized in Figure V.1) (224). Her results indicated that the N-terminus of gp2 is required for interaction with N4 RNAPII, while a central domain surrounding W30 is required for ssDNA-binding. These two classes of gp2 mutants are represented by the I3A and W30A gp2 alleles. The I3A gp2 allele does not interact with N4 RNAPII in solution, but binds to ssDNA

with higher affinity than WT gp2 in EMSA assays, while the W30A gp2 allele interacts with N4 RNAPII in solution, but does not bind to ssDNA in EMSA assays. Neither gp2 allele is capable of stimulating N4 RNAPII transcription *in vivo* or *in vitro*, suggesting that both N4 RNAPII interaction and ssDNA-binding are necessary for gp2-dependent activation of transcription (224). Furthermore, catalytic autolabeling experiments suggest that gp2 increases the rate of N4 RNAPII transcription initiation and shows that gp2 localizes within 12 Å of the initiating nucleotide at the N4 RNAPII active site; highlighting the importance in determining the exact site of N4 RNAPII required for gp2 binding (224).

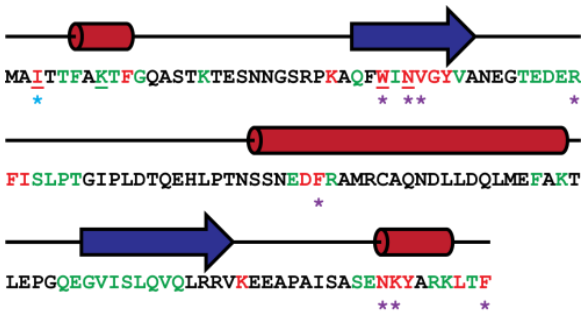

Figure V.1. Summary of gp2 alanine substitutions. The primary sequence of gp2 annotated with functional assignments based on *in vivo* and *in vitro* analyses of gp2 alanine scanning mutagenesis is depicted along with cartoon representation of secondary structures predicted by PSIPRED. Red letters, failure to complement N4 ORF2*am* phage infection; green letters, successfully complement N4 ORF2*am* phage infection; underlined letter, significantly reduced transcription activation *in vitro*; cyan asterisk, significantly reduced interaction with N4 RNAPII *in vitro*; purple asterisk, significantly reduced affinity for ssDNA templates *in vitro*; red cylinders, predicted α-helix; blue arrows, predicted β-sheets; black lines, loops. Adapted from C. Markle data (224).

In this chapter, I aimed to characterize the interaction between N4 RNAPII and its transcription factor gp2 to coordinate the N4 RNAPII site of catalysis and activate N4 middle transcription. Results show that gp2 localizes to the N4 RNAPII active site through interaction between the N-terminus of ssDNA-bound gp2 and the N4 RNAPII palm subdomain and DxxGR motif. Interaction between ssDNA-bound gp2 and N4 RNAPII activates transcription through

two mechanisms: recruiting N4 RNAPII to single-stranded middle promoters and increasing the catalytic efficiency of first phosphodiester bond formation.

WT GP2 RECRUITS N4 RNAPII TO ssDNA TEMPLATES

To confirm that gp2 recruits N4 RNAPII to ssDNA promoters, I evaluated the ability of gp2 to form stable ternary complexes with N4 RNAPII on ssDNA templates by EMSAs (Figure V.2). WT gp2 (14 kDa) bound to ssDNA with high affinity comparable to previously reported estimates (Figure V.2, top, lanes 2-7) (18, 224). N4 RNAPII (81 kDa) bound to ssDNA with lower affinity than gp2, as expected from the filter binding estimates of N4 RNAPII affinity for ssDNA (Figures IV.6 & V.2, lanes 8-13). The N4 RNAPII-ssDNA complex migrates with greater mobility and less stability than the gp2-ssDNA complexes, which suggests that gp2 binds to ssDNA as a multimer and supports previous observations of gp2 oligomerization and cooperative binding to ssDNA (18). Upon co-incubation of 50 nM WT gp2 and 25 nM N4 RNAPII with ssDNA, N4 RNAPII cooperatively bound to the gp2-ssDNA complex with high affinity observed as super-shifted ternary complexes (Figure V.2, top, lane 14). At increasing concentrations of N4 RNAPII, the amount of free ssDNA remained unaltered until N4 RNAPII was present in molar excess of gp2 (Figure V.2, top, lanes 14-19). These data suggest that N4 RNAPII binds to gp2-ssDNA complexes with significantly greater affinity (K_d < 50 nM) than free ssDNA (K_d 76±2 nM); this confirms the role of gp2 to recruit N4 RNAPII to single-stranded promoters (18, 224).

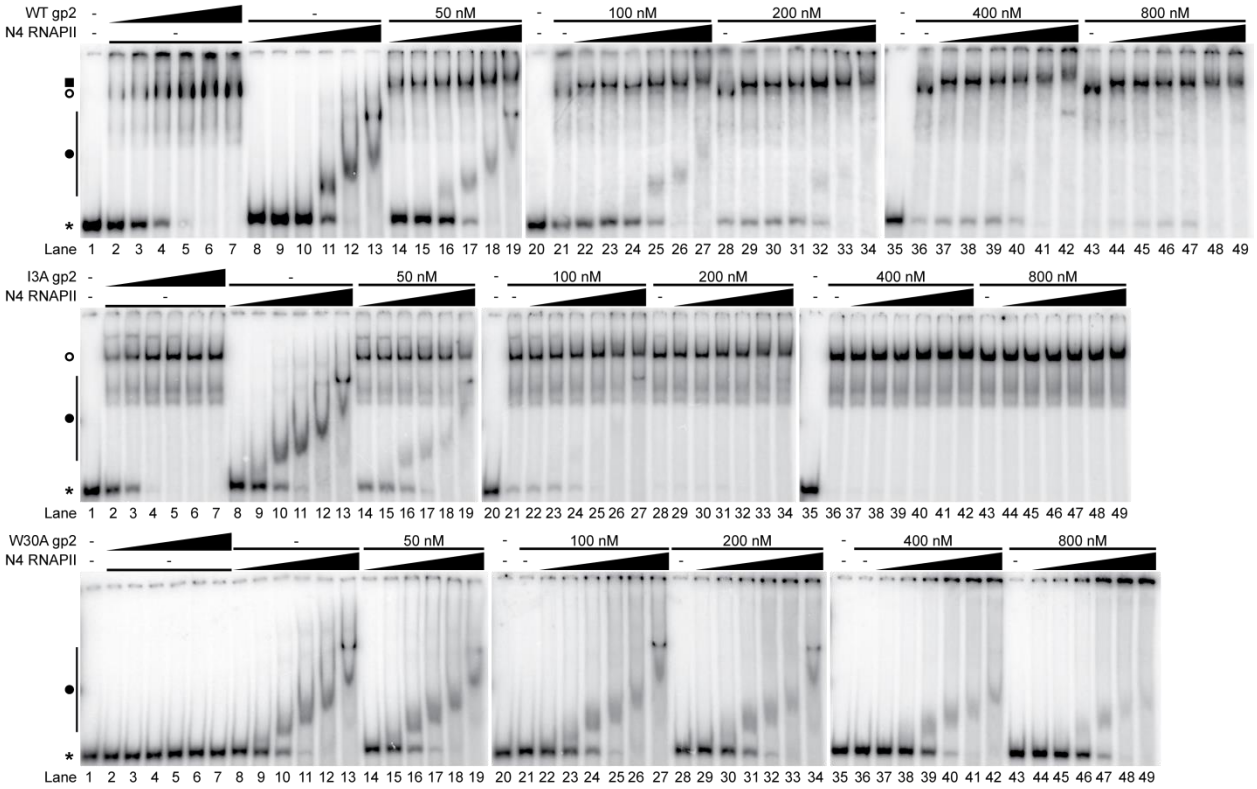

Figure V.2. Gp2 recruits N4 RNAPII to ssDNA templates. N4 RNAPII recruitment to ssDNA templates by WT (top), I3A (middle), and W30A (bottom) gp2 was determined through EMSAs. Indicated concentrations of gp2 alleles and N4 RNAPII were incubated with 5' end-labeled Pm5 ssDNA templates (1 nM) and DNA binding was analyzed by 6% PAGE and phosphorimaging. Representative phosphorimages shown with bands of interest labeled (left). Concentration gradients are as follows: 25, 50, 100, 200, 400, and 800 nM gp2 or N4 RNAPII. Asterisk, 5' end-labeled ssDNA; filled circle, N4 RNAPII+ssDNA; open circle, gp2+ssDNA; filled square, gp2+N4 RNAPII+ssDNA.

To determine whether gp2 recruitment of N4 RNAPII to ssDNA templates is dependent on direct interactions between the gp2 N-terminus and N4 RNAPII, I evaluated the ability of I3A gp2 to recruit N4 RNAPII to ssDNA templates (Figure V.2, middle). As expected, I3A gp2 bound to ssDNA with greater affinity than WT gp2 (Figure V.2, middle, lanes 2-7) (224). Co-incubation of I3A gp2, even at concentrations up to 800 nM, with N4 RNAPII and ssDNA templates failed to produce super-shifted complexes (Figure V.2, middle). When present at molar excess of I3A gp2, N4 RNAPII bound to free ssDNA to form N4 RNAPII-ssDNA complexes (Figure V.2, middle, lanes 14-27). These results indicate that N4 RNAPII recruitment to ssDNA

templates is dependent on interactions with the gp2 N-terminus. To determine whether gp2 binding to ssDNA is necessary for the recruitment of N4 RNAPII to ssDNA templates, I tested the ability of W30A gp2 to recruit N4 RNAPII to ssDNA templates (Figure V.2, bottom). On its own, W30A gp2 did not bind to ssDNA at any concentration tested, as previously determined (Figure V.2, bottom, lanes 2-7) (224). Co-incubation of W30A gp2 with N4 RNAPII and ssDNA did not produce super-shifted complexes at any concentration tested (Figure V.2, bottom). Additionally, N4 RNAPII affinity for free ssDNA was unchanged by the presence of high concentrations of W30A gp2. Together, these data suggest that gp2 activates middle transcription by recruitment of N4 RNAPII to ssDNA promoters and requires the interaction of N4 RNAPII with the N-terminus of ssDNA-bound gp2.

GP2 INCREASES THE CATALYTIC EFFICIENCY OF N4 RNAPII FIRST PHOSPHODIESTER BOND FORMATION

To determine whether gp2 increases the rate of N4 RNAPII catalysis, I determined the kinetic parameters for N4 RNAPII first phosphodiester bond formation in the absence and presence of gp2 (Figure V.3). I performed *in vitro* transcription assays with WT Pm5 ssDNA templates, measuring the rate of dinucleotide synthesis (GpA) across a gradient of initiating nucleotide (GTP) concentrations and determined the K_m and k_{cat} of transcription initiation. In the absence of gp2, N4 RNAPII initiates transcription with k_{cat} of 49±2 min^{-1} and K_m of 19±4 μM, resulting in an overall catalytic efficiency (k_{cat}/K_m) of 2.6 min^{-1} μM^{-1} (Figure V.3A). In general, the N4 RNAPII catalytic efficiency of first phosphodiester bond formation is comparable with the values reported for other T7-like RNAPs (158, 206, 218). The addition of equimolar WT gp2 (0.25 μM) did not significantly alter the affinity for the initiating nucleotide (K_m 19±4 vs. 23±7

μM), while the maximum velocity increased 1.7-fold (k_{cat} 49±2 vs. 82±5 min^{-1}), resulting in a 1.4-fold increase in catalytic efficiency (2.6 vs. 3.6 min^{-1} μM^{-1}) (Figure V.3A). Increasing gp2 concentration to 0.5 or 1.0 μM did not significantly increase the catalytic efficiency of first phosphodiester bond formation (3.9 and 3.5 min^{-1} μM^{-1}, respectively); increased maximum velocity was offset by reduced affinity for the initiating nucleotide (Figure V.3A). Therefore, gp2 increases the catalytic efficiency of first phosphodiester bond formation 1.4-fold at equimolar stoichiometry relative to basal N4 RNAPII independent from its role in recruiting N4 RNAPII to single-stranded promoters.

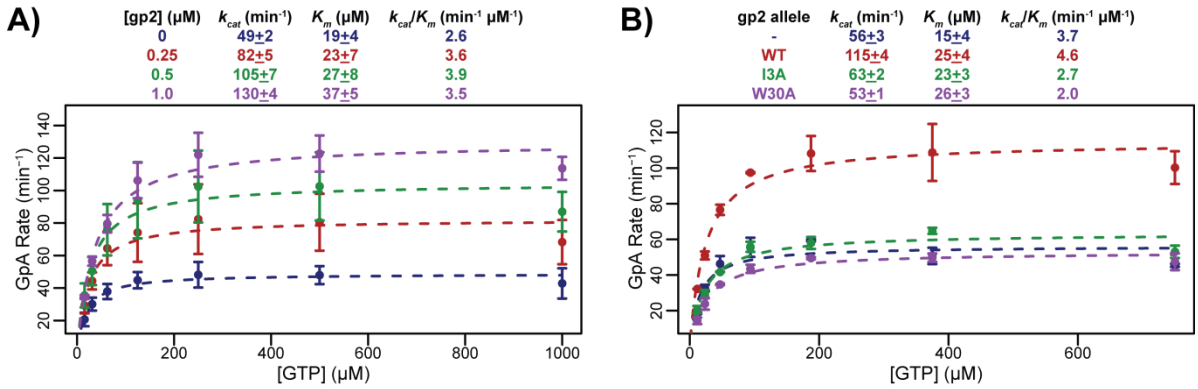

Figure V.3. Gp2 increases the catalytic efficiency of first phosphodiester bond formation. A) WT gp2 increases the maximum velocity of N4 RNAPII first phosphodiester bond formation. Rate of first phosphodiester bond formation by N4 RNAPII (250 nM) from Pm5 ssDNA templates (1 μM) with indicated concentrations of WT gp2 (0, 0.25, 0.5, 1.0 μM) and increasing concentrations of initiating nucleotide (15.6, 31.3, 62.5, 125, 250, 500, 1000 μM GTP) analyzed by 24% PAGE and phosphorimaging. Dinucleotide product (GpA) from three independent replicates was calculated according to Equation 1. Rate of GpA synthesis was calculated according to Equation 2 and plotted against GTP concentration (filled circles±SD). Kinetic parameters (K_m and k_{cat}) were calculated through non-linear least squares regression to Equation 3 (dashed lines) for each concentration of WT gp2. B) Interaction with N4 RNAPII and ssDNA-binding are both necessary for gp2-dependent increase in the maximum velocity of N4 RNAPII first phosphodiester bond formation. Rate of first phosphodiester bond formation of N4 RNAPII from Pm5 ssDNA templates with WT, I3A, or W30A gp2 (0, 0.5 μM) and increasing concentrations of initiating nucleotide (11.7, 23.4, 46.9, 93.8, 187.5, 375, 750 μM GTP) and kinetic parameters were determined as in (A).

To determine whether ssDNA-binding and N4 RNAPII interaction are necessary for the gp2-mediated increase in N4 RNAPII catalytic efficiency, I calculated the kinetic parameters for

N4 RNAPII first phosphodiester bond formation in the presence of 0.5 μM WT, I3A, or W30A gp2 (Figure V.3B). The addition of WT gp2 increased the maximum velocity 2-fold (k_{cat} 115±4 vs. 56±3 min^{-1}), matching the activity observed previously (Figure V.3A & B). The addition of neither I3A nor W30A gp2 increased the maximum velocity over basal N4 RNAPII activity (k_{cat} 63±2 and 53±1 vs. 56±3 min^{-1}, respectively), while WT, I3A, and W30A gp2 alleles reduced the affinity of N4 RNAPII for the initiating nucleotide relative to basal N4 RNAPII (K_m 23-26 μM vs. 15±4 μM) (Figure V.3B). WT gp2 increased catalytic efficiency 1.25-fold, while I3A and W30A gp2 decreased catalytic efficiency 1.4- and 1.9-fold, respectively (Figure V.3B). Therefore, ssDNA-binding and N4 RNAPII interaction mediated through the gp2 N-terminus are both necessary for the gp2-dependent increase in the rate of N4 RNAPII catalysis.

THE GP2 N-TERMINUS INTERACTS WITH THE N4 RNAPII PALM SUBDOMAIN AND DXXGR MOTIF

Gp2 alanine scanning mutagenesis and catalytic autolabeling experiments suggest that the gp2 N-terminus localizes near the N4 RNAPII active site (224). I mapped the site of N4 RNAPII-gp2 interaction to single amino acid resolution through a series of crosslinking experiments followed by mass spectrometry. I site-specifically incorporated the photocrosslinking unnatural amino acid *p*Bpa at I3 and F6 positions in the N-terminus of gp2 required for interaction with N4 RNAPII (224). As expected, purified I3Bpa gp2 had a similar activity profile *in vitro* as I3A gp2: significantly reduced affinity for N4 RNAPII in solution and failure to activate N4 RNAPII transcription in runoff transcription assays (Figure V.4). Purified F6Bpa gp2, however, displayed only a slight reduction in affinity for N4 RNAPII in solution and full capability to activate N4 RNAPII transcription in runoff transcription assays (Figure V.4).

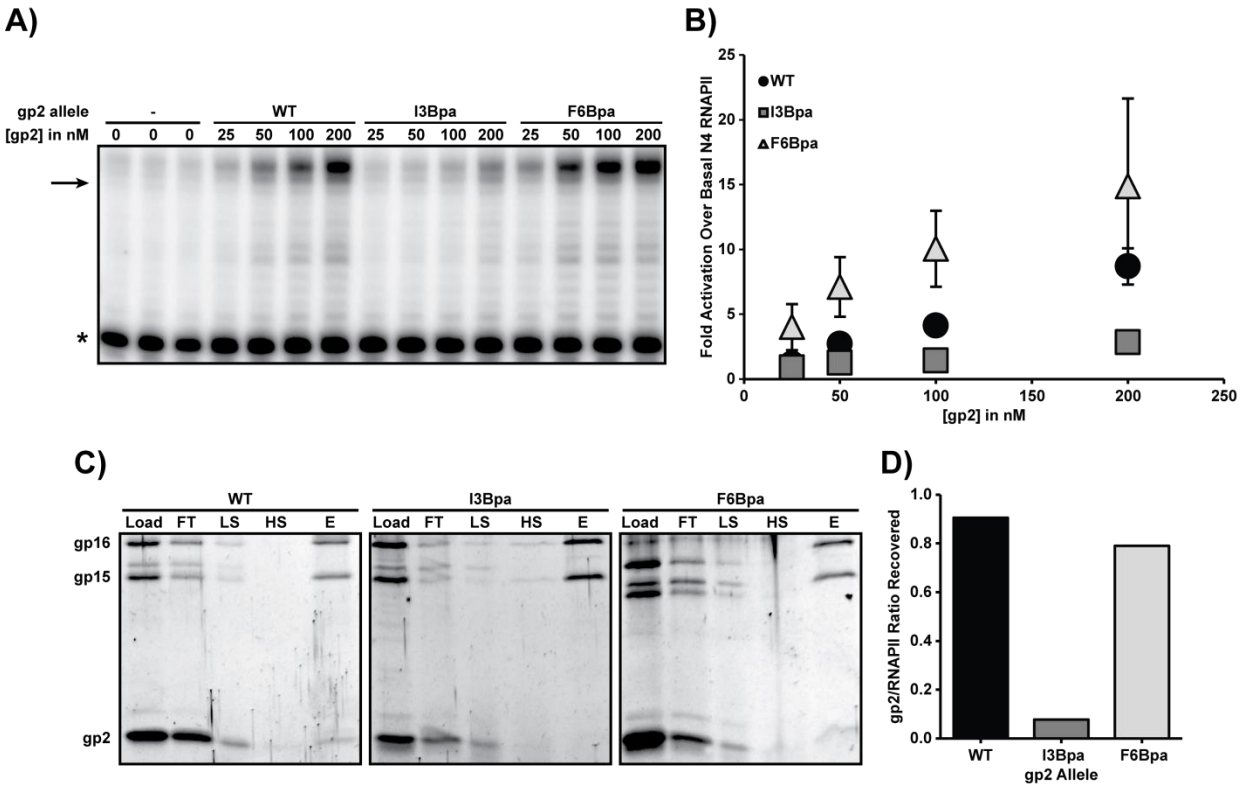

Figure V.4. Characterization of *p*Bpa-substituted gp2 alleles. A) Representative 8% PAGE autoradiogram of N4 RNAPII (25 nM) runoff transcription products from WT Pm5 ssDNA templates (100 nM) with increasing concentrations of WT, I3Bpa, and W30Bpa gp2 alleles (0, 25, 50, 100, 200 nM). Arrow, expected 37 nt runoff RNA product; asterisk, loading control. B) Relative transcription from three independent replicates plotted as fold activation over basal N4 RNAPII activity for each concentration of gp2 tested. Error bars; SD. C) Representative 12% SDS-PAGE gel of N4 RNAPII-gp2 IMAC interaction assays. Recombinant His_6-tagged WT N4 RNAPII (0.5 µM) and each gp2 allele (1.5 µM) were incubated together then applied to metal-affinity column. Bound complex was washed, eluted by imidazole, and all fractions were analyzed by 12% SDS-PAGE and Oriole staining. FT, flow through fraction; LS, low salt wash fraction; HS, high salt wash fraction; E, elution fraction. D) Ratio of gp2:N4 RNAPII in elution fraction quantified from a single replicate plotted.

No crosslinked species were observed by SDS-PAGE and Oriole staining of reactions containing N4 RNAPII with I3Bpa gp2 after UV excitation, confirming that the substitution of *p*Bpa at this position eliminates gp2 affinity for N4 RNAPII and prevents efficient crosslinking (data not shown). Upon co-incubation of N4 RNAPII with F6Bpa gp2 and UV excitation, species with migration corresponding to approximately 50, 75, and 120 kDa were observed by SDS-PAGE and Oriole staining (Figure V.5, left, lane 10). The presence of these bands were all

dependent on incubation with N4 RNAPII, exposure to UV light, and the F6Bpa gp2 allele; suggesting that these bands contain F6Bpa gp2-N4 RNAPII crosslinked species. The presence of F6Bpa gp2 in these species was confirmed by western blotting (Figure V.5, right). In the absence of N4 RNAPII, F6Bpa gp2 crosslinked to itself upon UV exposure observed as a ladder of multimeric products with ~15 kDa difference in apparent molecular weight (Figure V.5, right, lane 6). Upon addition of N4 RNAPII and UV excitation, unique higher molecular weight species containing F6Bpa gp2 were observed (Figure V.5, right, lane 10). F6Bpa gp2 (14 kDa) and gp15 (36 kDa) co-migrate in the 50 kDa band, matching the expected molecular weight of gp2-gp15 crosslinked products. Gp2 is also present in the N4 RNAPII-dependent bands migrating at approximately 75 and 120 kDa. These bands may contain F6Bpa gp2 crosslinked to the gp16 subunit (46 kDa) running at reduced mobility, although the lack of an antibody against gp16 prevents confirmation of this hypothesis.

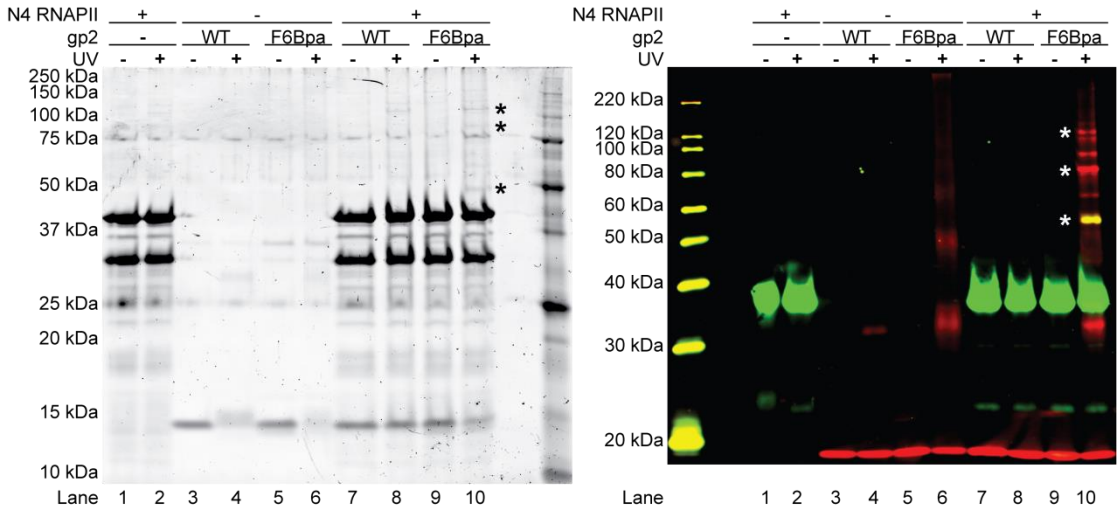

Figure V.5. The gp2 N-terminus crosslinks to N4 RNAPII. F6Bpa gp2 crosslinks to WT N4 RNAPII upon exposure to UV light. N4 RNAPII (0, 1.25 μM) was incubated with the indicated gp2 allele (0, 250 nM) followed by UV crosslinking (365 nm, 60 min). Products were analyzed by 12% SDS-PAGE and visualized by Oriole staining (left) and western blotting (right). Representative gels shown with bands of interest labeled with asterisks.

To validate the presence of F6Bpa gp2-N4 RNAPII crosslinks and map the site of interaction to single amino acid resolution, I analyzed the 50 kDa, 75 kDa, and 120 kDa gp2-containing bands by liquid chromatography-tandem mass spectrometry (LC-MS/MS) in collaboration with Dr. Jacob Waldbauer. Bands of interest were extracted from gel slices, digested with trypsin, subjected to LC-MS/MS, and mass spectral data were analyzed to identify peptides containing gp2 crosslinked to N4 RNAPII mediated through the F6Bpa residue. The top scoring putative gp2-N4 RNAPII crosslinked peptides in each band analyzed is summarized in Figure V.6 and Table V.1. F6Bpa gp2 crosslinked to L19-P21 residues within an unstructured loop in the N-terminus of the N4 RNAPII gp15 subunit was detected in both the 75 and 120 kDa bands (Figure V.6B & D, Table V.1). The 75 kDa band top scoring peptide, however, identified F6Bpa gp2 crosslinked to the N310-V312 residues within the palm subdomain of the N4 RNAPII gp16 subunit (Figure V.6C, Table V.1). Finally, the top scoring peptide from the 50 kDa band corresponded to F6Bpa gp2 crosslinked to the G239 residue within the DxxGR motif of the N4 RNAPII gp15 subunit (Figure V.6A, Table V.1).

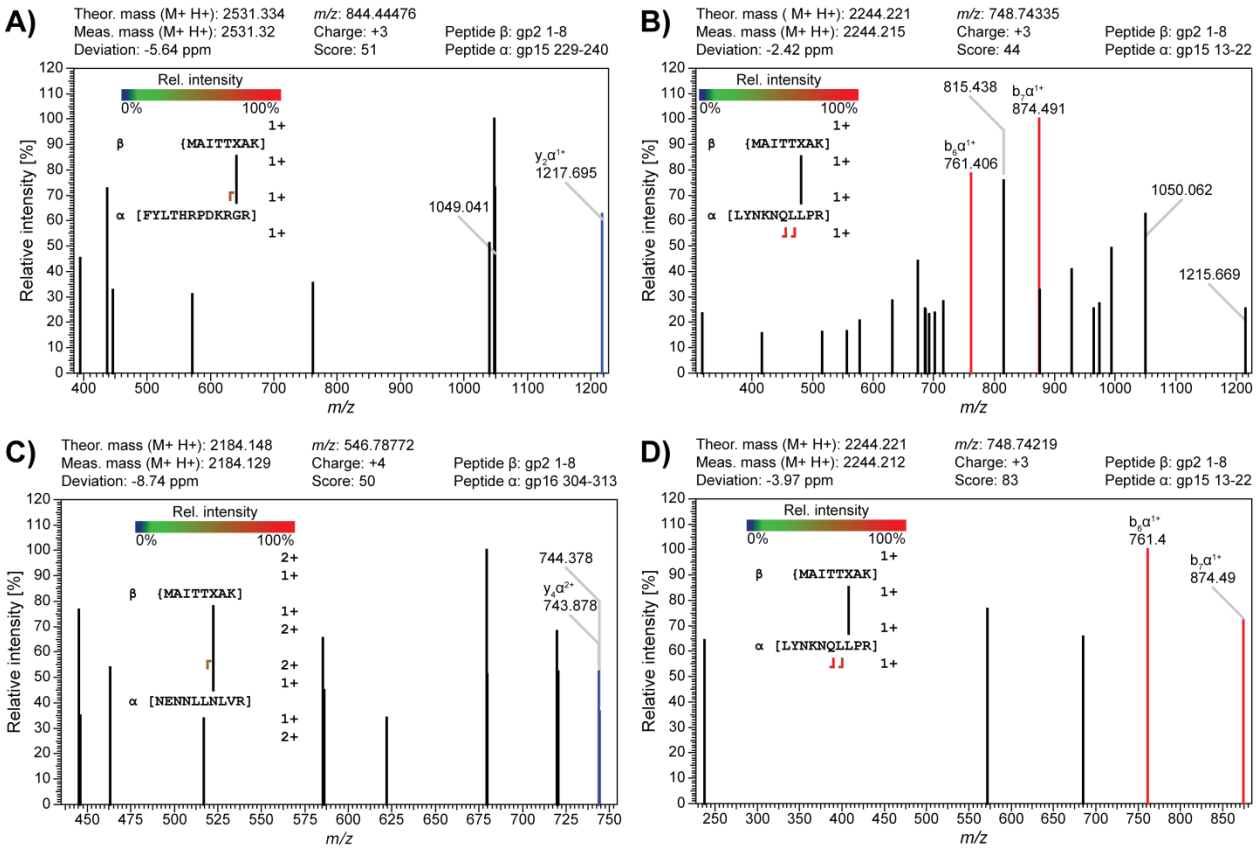

Figure V.6. Mapping sites of F6Bpa gp2 crosslinking to N4 RNAPII by LC-MS/MS. Results of Stavrox analysis of crosslinking mass spectral data for putative N4 RNAPII-F6Bpa gp2 interaction sites, showing MS/MS peptide fragmentation spectra with ions derived from crosslinked species highlighted. A) gp15 G239; B) gp15 L20-P21; C) gp16 N310-V312; D) gp15 L19-P21. N4 RNAPII residues are numbered as fusion of gp15-gp16 subunits.

Table V.1. Sites of N4 RNAPII interaction with F6Bpa gp2 identified by LC-MS/MS

Band	Protein 1	Peptide 1	Site 1	Protein 2	Peptide 2	Site 2[a]	Score
50 kDa	gp2	MAITTXAK	F6Bpa	N4 RNAPII, gp15 subunit	FYLTHRPDKRGR	G239	51
75 kDa	gp2	MAITTXAK	F6Bpa	N4 RNAPII, gp15 subunit	LYNKNQLLPR	L20, P21	44
75 kDa	gp2	MAITTXAK	F6Bpa	N4 RNAPII, gp16 subunit	NENNLLNLVR	N310, L311, V312	50
120 kDa	gp2	MAITTXAK	F6Bpa	N4 RNAPII, gp15 subunit	LYNKNQLLPR	L19, L20, P21	83

X, *p*Bpa; a, site 2 residues numbered as fusion of N4 RNAPII gp15-gp16 subunits.

Due to limitations in statistical analysis of mass spectrometry data, these are only putative sites of gp2-N4 RNAPII interaction and require additional biochemical validation. To validate the mass spectrometry data, I individually substituted each N4 RNAPII amino acid identified by mass spectrometry with alanine, purified the enzymes, and characterized each polymerase *in vitro*. G239A N4 RNAPII was inactive *in vitro*, as it failed to bind to ssDNA in EMSAs and failed to initiate transcription from Pm5 ssDNA templates in runoff transcription assays (data not shown). Therefore, G239A N4 RNAPII was omitted from further analysis. No other N4 RNAPII alanine allele had significantly reduced affinity for WT gp2 in N4 RNAPII-gp2 IMAC pulldown assays or significantly reduced gp2 activation of N4 RNAPII transcription *in vitro* (data not shown). Therefore, single alanine substitutions to N4 RNAPII residues were insufficient to validate putative sites of N4 RNAPII-gp2 interaction identified through mass spectrometry.

To validate the sites of N4 RNAPII-gp2 crosslinking identified by mass spectrometry, I used a reciprocal crosslinking approach. I site-specifically incorporated *p*Bpa into N4 RNAPII at all putative sites of gp2 interaction and interrogated each allele's ability to crosslink to gp2. As with the alanine-substituted enzymes, all *p*Bpa-substituted N4 RNAPII alleles were active in runoff transcription assays. Upon incubation of N310Bpa N4 RNAPII with 5-fold molar excess WT gp2 and UV excitation, a species with migration corresponding to approximately 60 kDa was observed by SDS-PAGE and Oriole staining (Figure V.7, top left, lane 10). This species was dependent on the N310Bpa N4 RNAPII allele, WT gp2, and UV excitation and migrated at the expected size (60 kDa) of gp2 (14 kDa) crosslinked to gp16 (46 kDa) (Figure V.7, left). Western blot analysis confirmed the presence of WT, but not I3A gp2 in the 60 kDa band, indicating that this interaction is mediated through the N-terminus of gp2 (Figure V.7, top right, lanes 10 & 12). Surprisingly, WT and I3A gp2 readily formed dimers (~30 kDa) and trimers (~45 kDa) in the

absence of N4 RNAPII and exposure to UV light, even in the absence of the photocrosslinking agent *p*Bpa (Figure V.7, bottom). No other *p*Bpa-substituted N4 RNAPII allele crosslinked with gp2 was detected by these methods, suggesting that all other putative sites of N4 RNAPII-gp2 interaction identified by mass spectrometry were false positives (data not shown). Together, these results suggest that the gp2 N-terminus is localized to the N4 RNAPII active site through direct interactions with the N310 residue of the palm subdomain.

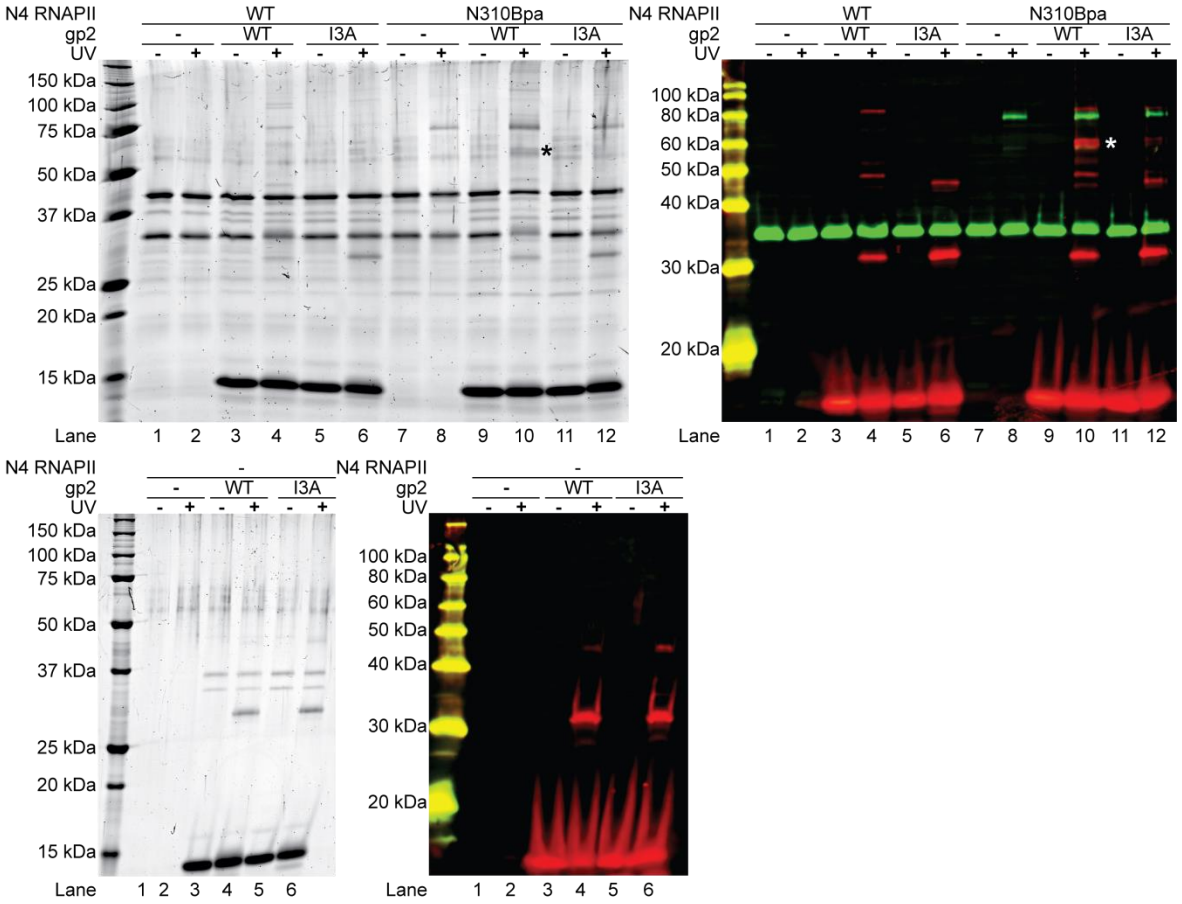

Figure V.7. N310Bpa N4 RNAPII crosslinks to the gp2 N-terminus. N310Bpa N4 RNAPII crosslinks to WT gp2 upon exposure to UV light. Indicated N4 RNAPII alleles (0, 0.4 µM) were incubated with indicated gp2 allele (0, 2 µM) followed by UV crosslinking (365 nm, 60 min). Products were analyzed by 12% SDS-PAGE and visualized by Oriole staining (left) and western blotting (right). Representative gels shown with bands of interest labeled with asterisks. Red bands, gp2; green bands, His_6-tagged gp15; yellow, co-migrating gp2 and His_6-tagged gp15. N4 RNAPII residues are numbered as fusion of gp15-gp16 subunits.

Since alanine and *p*Bpa substitutions to the G239 residue of N4 RNAPII completely inactivated N4 RNAPII activity *in vitro* (data not shown), I validated the putative F6Bpa gp2-G239 N4 RNAPII interaction through complementary mass spectrometry approaches in collaboration with Dr. Jacob Waldbauer. The 50 kDa band predicted to contain F6Bpa gp2-G239 N4 RNAPII crosslinked species was extracted from gel slices and split into two samples. One sample was digested with trypsin, while the other was digested with chymotrypsin, generating two distinct populations of peptide fragments for independent identification of F6Bpa gp2-G239 N4 RNAPII crosslinks by LC-MS/MS. The presence of peptides from gp15 (Y230-Y242) crosslinked to the N-terminal fragment of F6Bpa gp2 (M1-F10) in were detected at high abundance in MS1 data derived from chymotryptic cleavage of the crosslinked proteins (Figure V.8). These results validate the interaction between the gp2 N-terminus and the N4 RNAPII G239 residue of the conserved DxxGR motif located in the catalytic cleft and responsible for stabilization of the RNA:DNA hybrid during T7 RNAP transcription initiation (85).

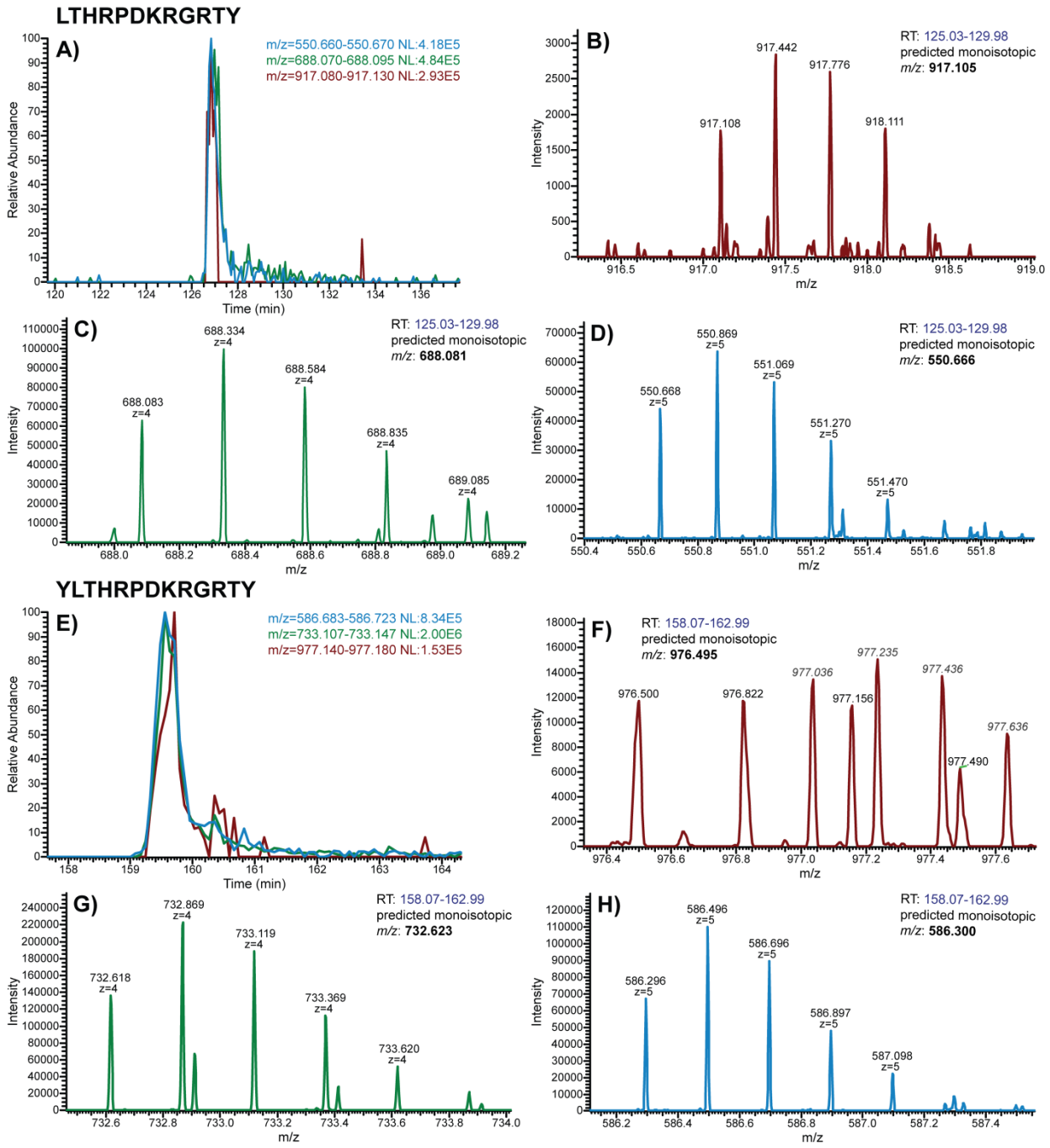

Figure V.8. High resolution MS1 data confirms crosslinking between F6Bpa gp2 and the N4 RNAPII DxxGR motif. High-resolution MS1 data indicating the presence of two peptides from gp15 (LTHRPDKRGRTY and YLTHRPDKRGRTY) crosslinked to mAITTxAKTF (where m = oxidized methionine and x = *p*Bpa) derived from chymotryptic cleavage of F6Bpa gp2. Predicted monoisotopic neutral masses for the crosslinked species are 2748.295 and 2927.472 Da, respectively. Extracted ion chromatograms for monoisotopic 3+, 4+ and 5+ ions are shown in (A) and (E), and mass spectra averaged over the elution window showing the isotopologue distributions are shown for 3+ ions in (B) and (F), 4+ ions in (C) and (G), and 5+ ions in (D) and (H). Note that peaks indicated in gray italics in (F) derive from an unrelated, co-eluting species in a 5+ charge state. RT, retention time; *m/z*, mass to charge ratio; z, charge state.

DISCUSSION AND FUTURE DIRECTIONS

Bacteriophage N4 middle transcription is performed by N4 RNAPII, one of the smallest members of the T7-like RNAP family, along with its required transcription factor gp2. N4 RNAPII has limited homology to N-terminal domains required for promoter recognition in other T7-like RNAPs, suggesting that N4 RNAPII recognizes promoters through factor-dependent mechanisms different from those observed in other T7-like RNAPs (15, 223). Therefore, I aimed to define the sequence and structural requirements for N4 RNAPII promoter recognition and elucidate the molecular mechanism of gp2 in transcription activation. The results presented in Chapter IV demonstrate that N4 RNAPII specifically recognizes short, AT-rich promoters at the -7, -6, -3, and -1 positions of the template strand through direct interactions with the specificity loop. Interaction between the N4 RNAPII specificity loop and sequences within the conserved core of promoters is required for promoter recognition, which is primarily regulated by template unwinding and transcription initiation in conjunction with additional protein factors. The results presented in Chapter V demonstrate that gp2 activates N4 RNAPII transcription by both recruiting N4 RNAPII to single-stranded promoters and increasing the maximum velocity of N4 RNAPII first phosphodiester bond formation. These activities are dependent on the localization of ssDNA-bound gp2 to the N4 RNAPII active site through direct interaction between the N-terminus of ssDNA-bound gp2 with the N4 RNAPII palm subdomain and DxxGR motif. Based on these data, I propose an updated model for gp2-dependent N4 RNAPII transcription activation unique among mechanisms of transcription initiation by T7-like RNAPs and their transcription factors.

N4 RNAPII recognizes AT-rich ssDNA promoters through its specificity loop

T7 RNAP recognizes a bipartite promoter sequence spanning -17 to +6, containing an upstream binding region (-17 to -5) and an initiation region (-4 to +6) (Figure V.9) (114). In contrast, upstream sequence conservation shows that N4 RNAPII recognizes short, AT-rich promoters with conservation (-10 to +2) limited to the initiation and conserved core regions, while the variable binding region is absent (Figures IV.2 & V.9) (19). Here, I have confirmed the role of these conserved sequences in N4 RNAPII promoter recognition through *in vitro* runoff transcription assays (Figure IV.3). Results indicate that N4 RNAPII promoter contacts are limited to template-strand bases (Figures IV.3 & IV.4), while N4 RNAPII displays an absolute requirement for initiation with GTP and incorporation of another purine at +2, a characteristic common to T7-like RNAPs where the initiating NTPs are required for stabilizing the open complex and unstacking the template-strand base at the -1 position (Figures IV.3 & IV.7) (93, 119, 158).

```
                -15  -10  -5   +1  +5
                 |    |    |    |   |
T7 RNAP:    ATTATGCTGAGTGATATCCTCT
Rpo41:      ---------TATATTCAT-----
POLRMT:     --------GGTGTNGGTTTCTN
N4 RNAPII:  ----------AAAAAACYC---
```

Figure V.9. Comparison of T7-like RNAP consensus promoter sequences. T7 RNAP, Rpo41, POLRMT, and N4 RNAPII consensus promoter template-strand sequences spanning -17 to +6. Light blue, transcription start site.

Surprisingly, runoff transcription assays with ssDNA templates did not reveal specific recognition of nucleotides at the -10, -9, or -4 position, which were previously shown to be important for activity on promoters within negatively supercoiled plasmid templates *in vivo* (Figures IV.2 & IV.3) (19). The preference for adenines at the -10, -9, and -4 position of

plasmid-resident Pm5 promoters *in vivo* and the enrichment of adenines from -10 to -3 in all N4 promoters suggests that a run of A-T base pairs may be an important factor in promoter melting *in vivo*. Therefore, DNA melting and topology play a crucial role in regulating N4 early and middle transcription by providing activated ssDNA promoters competent for polymerase binding.

Consistent with this, N4 RNAPII promoter substitutions caused modest differences (less than 2-fold) in N4 RNAPII affinity for N4 single-stranded promoter sequences (Figure IV.6), which do not fully explain the large differences in promoter activity observed in transcription assays *in vitro* or *in vivo* (Figures IV.2 & IV.3). These data confirm that N4 RNAPII promoters lack a defined upstream binding region and suggest that N4 RNAPII sequence specificity *in vivo* is primarily determined by melting of AT-rich promoters and only modestly regulated by N4 RNAPII affinity for ssDNA sequences.

Promoter selection by T7 RNAP occurs primarily through preferential binding to promoter sequences; T7 RNAP displays a 10^5-fold difference in affinity for promoter versus non-promoter dsDNA templates (123, 172). Other T7-like RNAPs, which also lack sequence conservation in the upstream binding region (Figure V.9), require the activity of additional transcription factors to differentiate promoter from non-promoter DNA. N4 vRNAP promoter binding is dependent on the extrusion and stabilization of a conserved promoter hairpin by the host-encoded architectural transcription factors DNA gyrase and *Eco*SSB (7, 12, 31). Rpo41, which also recognizes short (9 nt) AT-rich promoter, does not discriminate between promoter-containing and promoter-deficient DNA templates by differential affinity. Instead, Rpo41 requires the additional factor Mtf1 to facilitate sequence-dependent bending and melting of upstream DNA for promoter selection (143, 171). Thus, T7-like RNAPs outside of the

bacteriophage cluster rely on their transcription factors for both sequence-specific promoter recognition and assembly of the initiation complex. Therefore, the full understanding of the requirements for N4 RNAPII promoter recognition and transcription initiation is dependent on the characterization of its required transcription factors.

Although N4 RNAPII requires transcription factors for activity *in vivo*, N4 RNAPII is capable of initiating transcription from partially or fully ssDNA templates *in vitro*, suggesting that N4 RNAPII contains elements responsible for promoter recognition. T7 RNAP recognizes promoters through three structural elements: i) the AT-rich recognition loop, present at the N-terminal domain, inserts itself into the minor groove of upstream DNA; ii) the β-IH, present in the N-terminal domain, separates template and non-template strands; iii) the specificity loop, present in the fingers subdomain, makes sequence-specific contacts through the major groove and stabilizes the melted template strand through backbone contacts (Figures IV.1 & V.10) (109, 120, 262). Despite the lack of N-terminal sequence conservation (86), all well-studied T7-like RNAPs utilize structural homologs of the T7 RNAP specificity loop to make sequence-specific contacts within the major groove of promoter DNA (Figure V.10) (109, 145, 201, 207, 218). The specificity loops of these enzymes do not share sequence similarity, but all comprise a flexible antiparallel β-sheet located between conserved sequence blocks in the fingers subdomain.

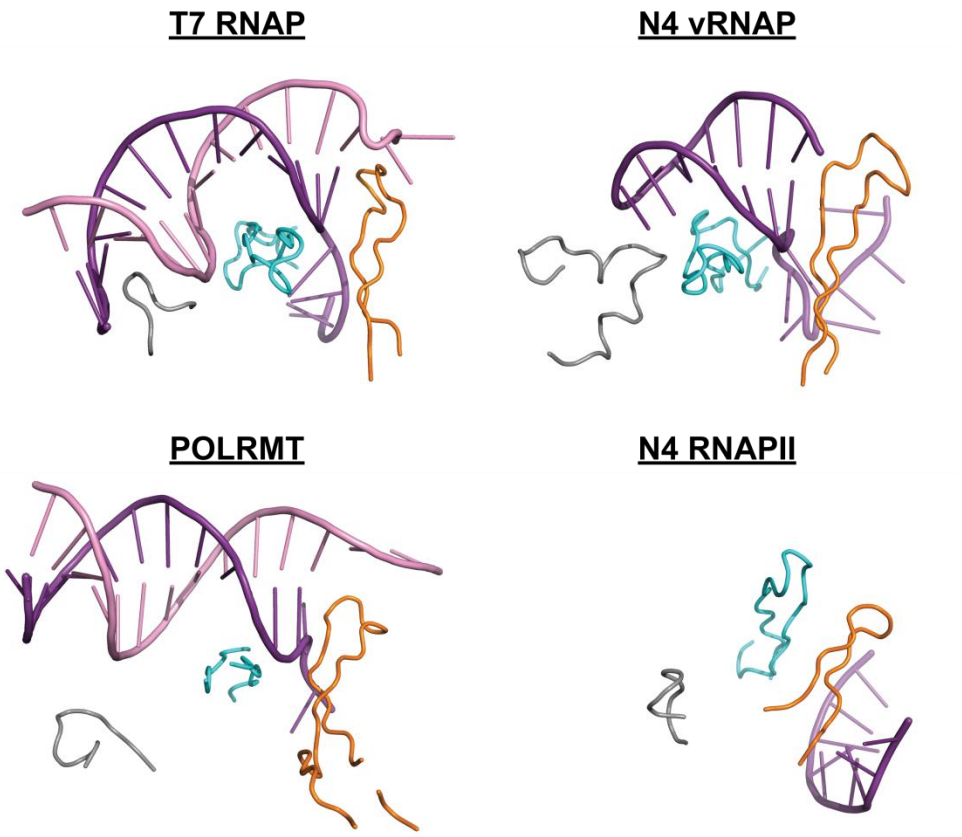

Figure V.10. Structural comparison of T7-like RNAP promoter recognition elements. Ribbon representation of the promoter recognition elements of T7 RNAP (PDB: 1CEZ), N4 vRNAP (PDB: 3C2P), POLRMT (PDB: 6ERP), and N4 RNAPII (PDB: 6DT7) binary complex structures are depicted along with cartoon representations of promoter DNA. AT-rich recognition loop, grey; β-IH, orange; specificity loop, cyan; template-strand DNA, purple; non-template strand DNA, pink. T7 RNAP structure adapted from Cheetham et al. (120); N4 vRNAP structure adapted from Gleghorn et al. (218); POLRMT structure adapted from Hillen et al. (207); N4 RNAPII adapted from Molodtsov and Murakami (223).

Results of mutagenesis and crosslinking approaches confirmed that residues within the N4 RNAPII specificity loop directly interact with the template strand of N4 RNAPII promoters and are required for activity on ssDNA templates (Figures IV.9-IV.13). Specificity loop residues required for promoter recognition (Figure IV.9) are located along the same face of the antiparallel β-hairpin facing towards the catalytic cleft and are well positioned to interact directly with template DNA in the active site. Although these residues are located a large distance from upstream DNA and the active site in the N4 RNAPII crystal structure, the specificity loops of

other T7-like RNAPs display considerable flexibility through the unstructured loops connecting them to the fingers domain (Figures IV.1 & V.10) (223). Correspondingly, direct interactions between the N4 RNAPII specificity loop and template-strand DNA were confirmed through crosslinking (Figure IV.11), suggesting that the N4 RNAPII specificity loop is also flexible or the upstream DNA is mislocalized in this structure (Figure IV.1) (223).

Furthermore, runoff transcription experiments demonstrated that the N4 RNAPII C494 recognizes purines at the -1 position, while Y492 recognizes purines at the -3 position, potentially through base stacking interactions (Figures IV.10 & IV.13). These results are consistent with Pm5 promoter sequence preferences established through *in vitro* runoff transcription assays and support the conservation of these sequences in N4 RNAPII promoters (Figures IV.2 and IV.3). Sequence-specific contacts between the N4 RNAPII specificity loop and template-strand nucleotides in ssDNA promoters would represent a new mechanism of promoter recognition for T7-like RNAPs. The T7 RNAP, vRNAP, and POLRMT specificity loops specifically recognize sequences within the major groove of dsDNA templates, but only make non-specific contacts with the melted template strand to stabilize the open complex (120, 207, 218).

These results are consistent with the role of the specificity loop in other T7-like RNAPs, but conflict with those observed in the N4 RNAPII binary complex structure (Figure V.10) (223). In this structure, the N4 RNAPII antiparallel β-hairpin specificity loop (residues 476-502) extends from the fingers subdomain and makes hydrophobic interactions with the N-terminal helix bundle. This interaction with the N4 RNAPII N-terminus leaves the specificity loop poorly positioned to recognize upstream DNA, which is localized to a basic patch between the thumb subdomain and the N-terminal domain (Figure IV.1). In these studies, N4 RNAPII was

crystallized with a DNA template containing dsDNA upstream of the -3 position and completely lacking a promoter consensus sequence (223). Since N4 RNAPII requires ssDNA templates from -10 to +2 for specific transcription initiation, the DNA in these structures are bound in a sequence-independent manner and more likely represent the N4 RNAPII elongation complex conformation. This would also explain the lack of conformational changes observed between the binary and elongation complexes (223). Furthermore, the N4 RNAPII specificity loop contacts with the N-terminal domain are likely artefacts of crystal packing, as the unstructured loops connecting this motif to the fingers domain should allow for flexibility and repositioning of the specificity loop as observed in other T7-like RNAPs. Therefore, the use of ssDNA promoter templates and flexibility of the N4 RNAPII specificity loop accounts for the different role of N4 RNAPII in promoter recognition observed in these biochemical assays.

Mechanism of gp2 activation of N4 RNAPII transcription

Since gp2 binds to ssDNA with significantly greater affinity than N4 RNAPII (K_d of ~45 nM and 76 nM, respectively) and since gp2 and N4 RNAPII directly interact in solution, it was hypothesized that gp2 may recruit N4 RNAPII to melted promoter sequences (18, 224). Although previous studies demonstrated cooperative binding to ssDNA and identify both N4 RNAPII and gp2 in the ternary complex (18, 224), estimates of the relative affinity of N4 RNAPII for gp2-bound ssDNA versus free ssDNA have not been shown until this study. Upon co-incubation with WT gp2 and ssDNA, N4 RNAPII displayed a clear preference for binding to the gp2-ssDNA complex to form fully super-shifted complexes over free ssDNA (Figure V.2), demonstrating that N4 RNAPII binds to gp2 or the gp2-ssDNA complex with a K_d much less than 50 nM. However, these assays merely provide an upper bound for the binding constants of N4 RNAPII for gp2 or gp2-ssDNA complexes. Direct measurements of the affinity of N4

RNAPII for gp2 in solution may be achieved through sensitive techniques such as surface plasmon resonance, while measurements of the binding constants for ternary complex formation may be achieved through steady-state binding assays such as isothermal calorimetry or fluorescence polarization.

To determine whether gp2 plays a role in transcription activation beyond its role in recruitment of N4 RNAPII to single-stranded promoters, a previous graduate student (C. Markle) performed catalytic autolabeling assays (224). In these experiments, both N4 RNAPII gp15 and gp16 subunits were labeled in a template-dependent fashion, while the addition of gp2 resulted in the strong labeling of gp2 and a 2-fold increase in the total amount of labeling observed (224). These results show that gp2 localizes to the N4 RNAPII active site and suggest that gp2 increases the efficiency of first phosphodiester bond formation. However, the overall increase in labeling observed does not confirm that gp2 directly increases the rate of N4 RNAPII catalysis since increased labeling could be due to increased efficiency of crosslinking to the initiating nucleotide or increased efficiency of first phosphodiester bond formation. The kinetics of transcription initiation experiments presented here directly measure the rate of first phosphodiester bond formation and confirm that gp2 increases the catalytic efficiency of first phosphodiester bond formation by increasing the maximum velocity of the reaction (Figure V.3).

The failure of both I3A and W30A gp2 to recruit N4 RNAPII to ssDNA in EMSAs (Figure V.2) and increase the rate of transcription initiation (Figure V.3) confirm that N4 RNAPII recruitment to ssDNA templates and activation of first phosphodiester bond formation are both dependent on the interaction between the ssDNA-bound gp2 and N4 RNAPII. Therefore, ssDNA-binding alters the positioning of gp2 at the N4 RNAPII active site and may be

required to induce structural rearrangements that facilitate substrate coordination and enable efficient transcription initiation.

The gp2 alanine scanning mutagenesis and catalytic autolabeling experiments performed by a previous graduate student (C. Markle) demonstrate that the N-terminus of gp2 is present at the N4 RNAPII active site (224). Further studies showed that gp2 interacts with the gp15 subunit of N4 RNAPII, however biochemical and genetic attempts to determine the N4 RNAPII residues required for interaction with gp2 were unsuccessful (224). Here, I confirmed that the gp2 N-terminus localizes to the N4 RNAPII active site through direct interactions between gp2 and the N4 RNAPII N310 and G239 residues through crosslinking and mass spectrometry approaches (Table V.1, Figures V.5-V.8). N310 is a surface-exposed residue in the α14 helix of the N4 RNAPII palm subdomain flanked by the thumb and fingers subdomains, while the G239 residue lies within the evolutionarily conserved DxxGR motif located within the N4 RNAPII catalytic cleft (Figure V.11). This motif is found in T7-like DNA-dependent RNAPs and has been shown to interact with the 3' end of the nascent transcript at the active site to stabilize the RNA:DNA hybrid during transcription initiation in T7 RNAP (84–86, 109). Therefore, gp2 interaction with N4 RNAPII DxxGR motif may help stabilize the nascent RNA product to increase the catalytic efficiency of transcription initiation (Figure V.3). Gp2 interaction with these sites provides access to the N4 RNAPII active site through large truncations within the N4 RNAPII thumb and fingers subdomain relative to other T7-like RNAPs and suggests that gp2 may functionally complement these structures (Figure V.11) (223). However, the interactions between the gp2 N-terminus and N4 RNAPII palm subdomain and DxxGR motif were determined through crosslinking reactions in the absence of ssDNA templates, which are representative of those

occurring in solution and therefore may not accurately reflect the gp2-N4 RNAPII interactions in the ternary complex required for transcription initiation.

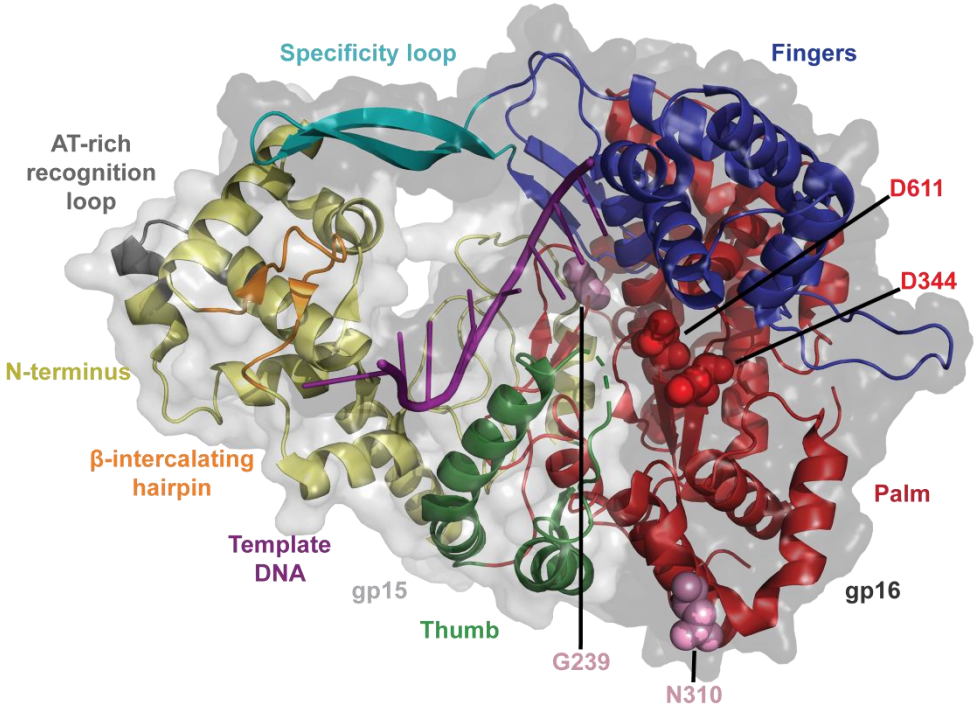

Figure V.11. Gp2 localizes to the N4 RNAPII active site through interactions with the N4 RNAPII palm subdomain and DxxGR motif. The cartoon representation of N4 RNAPII (PDB: 6DT7) indicating the sites of interaction with gp2 identified through crosslinking and mass spectroscopy is overlaid on surface models of the gp15 (light grey) and gp16 (dark grey) subunits. N4 RNAPII structural elements are color-coded as in Figure IV.1. The catalytic aspartates, D344 and D611, are represented as light red spheres. Residues that crosslink to F6Bpa gp2, G239 and N310, are represented as pink spheres. N4 RNAPII structure adapted from Molodtsov and Murakami (223). N4 RNAPII residues are numbered as fusion of gp15-gp16 subunits.

Gp2 may increase the rate of first phosphodiester bond formation by directly coordinating the rNTP substrates in the active site or through the induction of N4 RNAPII conformational changes. Although gp2 crosslinks to the initiating nucleotide during catalytic autolabeling reactions, crosslinking between lysine residues and the NTP derivative occurs across a distance of 12 Å and cannot confirm that gp2 residues are within hydrogen bonding distance of the initiating GTP (224, 263). Furthermore, gp2 increasing the rate of transcription initiation through

direct interaction with the initiating nucleotide is unlikely since N4 RNAPII affinity for the initiating nucleotide is substantially decreased upon addition of gp2 during transcription initiation assays (Figure V.3). In contrast, preliminary evidence suggests that gp2-binding induces a conformational change in N4 RNAPII. In the absence of gp2, N310Bpa and several other *p*Bpa-substituted N4 RNAPII alleles formed distinct crosslinks between gp15 and gp16 subunits upon UV exposure as evidenced by a gp15-containing band migrating at an apparent molecular weight of ~80 kDa, matching the predicted size of fused gp15-gp16 subunits (e.g. Figure V.7, top, lanes 7-12). Upon addition of WT gp2, *p*Bpa-substituted N4 RNAPII alleles display increased gp15-gp16 crosslinking (e.g. Figure V.7, top, lanes 8 & 10). These data suggest that N4 RNAPII interaction with gp2 in solution induces a conformational change in N4 RNAPII, potentially clenching the open cupped right hand architecture to bring the thumb and palm subdomains in closer proximity. Furthermore, addition of gp2 to N4 RNAPII catalytic autolabeling experiments performed by a previous graduate student (C. Markle) led to increased labeling of the N4 RNAPII gp15 subunit relative to the gp16 subunit (224). Gp2 localization to the N4 RNAPII active site therefore reorganizes the N4 RNAPII active site, potentially through the induction of N4 RNAPII conformational changes.

Model of N4 RNAPII transcription initiation

The data presented here have informed the working model for N4 RNAPII factor-dependent transcription (Figure V.12) (18). An unknown phage protein specifically recognizes dsDNA upstream of sites of N4 RNAPII transcription initiation to induce DNA bending and melting of AT-rich promoter sequences. Gp2 then binds to ssDNA with high affinity to stabilize the open complex and recruit N4 RNAPII to melted promoters through direct interactions with the N-terminus of ssDNA-bound gp2. Upon recruitment, N4 RNAPII recognizes nucleotides at

the -7, -6, -3, and -1 position of the melted template strand with the flexible specificity loop to coordinate the template strand for transcription initiation. The interaction between the N-terminus of ssDNA-bound gp2 reorganizes the N4 RNAPII active site through interactions with the DxxGR motif or palm subdomain to increase the catalytic efficiency of first phosphodiester bond formation.

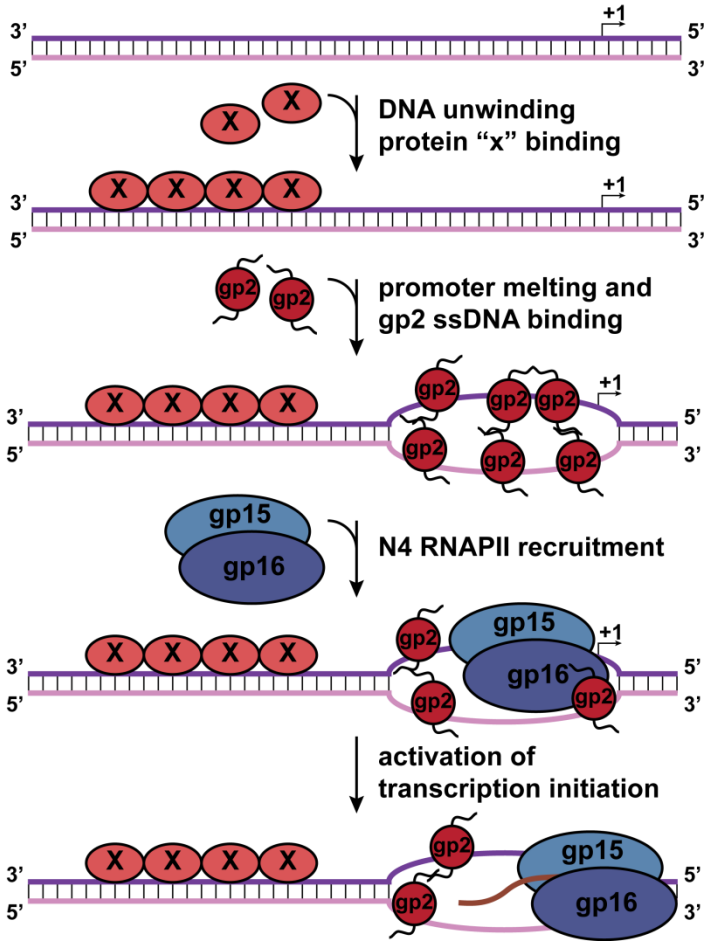

Figure V.12. Model of N4 RNAPII transcription initiation.

Similar to N4 RNAPII, the mtRNAPs require additional factors for transcription initiation *in vivo*, which activate transcription by: i) recruiting polymerases to promoter elements; ii) unwinding DNA and stabilizing the open promoter complex; and iii) increasing the rate of transcription initiation. The POLRMT transcription factor TFAM and gp2 both recruit the

polymerase to promoters through simultaneous protein-protein and protein-DNA interactions to increase the local concentration of the polymerase and increase the catalytic efficiency of transcription initiation (Figures V.2 & V.3) (18, 200, 206, 207). Like TFB2M and its homolog Mtf1, gp2 localizes to the polymerase active site and makes numerous contacts with nucleic acids and RNAP structural elements to activate the rate of transcription initiation (Figures V.3 & V.11) (170, 204, 207, 224). Since gp2 is an SSB and shows a modest preference for pyrimidines enriched in the non-template strand of N4 RNAPII promoters (18), gp2 may stabilize the open complex by sequestering the non-template strand as observed for Mtf1, TFB2M, and the bacterial sigma factors in multi-subunit RNAPs (160, 208).

Unlike the proteins above, gp2 does not confer N4 RNAPII the ability to melt dsDNA promoters. While TFB2M and Mtf1 interaction with their respective RNAPs is required for formation of a stable open promoter complex, gp2 interaction with N4 RNAPII does not induce polymerase conformational changes to facilitate promoter unwinding and does not activate N4 RNAPII transcription from linear, dsDNA templates *in vitro* (18, 143, 208). Furthermore, gp2 non-specifically recruits N4 RNAPII to previously melted promoters and does not contribute to N4 RNAPII promoter sequence recognition (Figure IV.5) (18). Thus, the determination of the complete mechanism of N4 RNAPII transcription initiation awaits the identification and characterization of the protein factor required to specifically melt AT-rich N4 RNAPII promoters.

In conclusion, these data describe a mechanism of N4 RNAPII factor-dependent transcription initiation distinct from mechanisms found in other T7-like RNAPs and underline the emerging diversity of transcription initiation strategies that exist within the T7-like RNAP family.

Future directions

Several outstanding questions remain to achieve a full understanding of the mechanism of N4 RNAPII transcription. First of all, promoter substitutions that shift start site selection or reduce N4 RNAPII transcription *in vitro* without reducing affinity for N4 RNAPII (-1C, -3C, -6G) presumably reduce the rate of transcription initiation. However, the data presented here do not directly address this assertion. This hypothesis can be tested more directly through calculating the N4 RNAPII kinetic parameters (K_m and k_{cat}) for transcription initiation using substituted ssDNA Pm5 templates. These assays can also be used to further interrogate enzyme kinetics underlying the requirement of purines as the initiating nucleotides. Furthermore, efforts to map the specific nucleotides to which E481Bpa and R496Bpa N4 RNAPII crosslink were unsuccessful (data not shown). Additional crosslinking experiments with templates containing the photocrosslinking nucleotide analog 5-IdU at the -7, -6, -3, and -1 position of ssDNA templates followed by proteolysis could be utilized to map the additional N4 RNAPII-ssDNA interactions and confirm the inferred specificity loop contacts with bases at -1 and -3.

Although the role of gp2 in transcription elongation is yet to be elucidated, preliminary evidence suggests that gp2 may function at steps beyond first phosphodiester bond formation. N4 RNAPII runoff transcription from WT Pm5 ssDNA templates produces a triplicate RNA product, comprised of a 37 nt RNA runoff product that is extended at the 3' end of transcription through non-templated addition. Upon addition of WT gp2, the longest RNA species, extended 2 nt at the 3' end, is more prominent than the other RNA species (Figure V.4). Since these products initiate from the same site (Figure IV.7B), any differences in product length must occur at the 3' end of transcripts. These results suggest that gp2 may function in N4 RNAPII transcription elongation or termination. To directly test whether gp2 activates N4 RNAPII

catalysis downstream of transcription initiation, kinetic assays of nucleotide addition from N4 RNAPII promoter-free elongation substrates extending a 9-12 nt RNA primer in the presence and absence of gp2 could be performed. Furthermore, if gp2 were to activate transcription beyond transcription initiation, it would be expected that gp2 would remain associated with N4 RNAPII upon promoter escape and transition into the elongation complex. This would be of considerable interest since mtRNAPs release their transcription initiation factors upon promoter clearance (166, 208).

Several lines of evidence suggest that the localization of ssDNA-bound gp2 to the N4 RNAPII catalytic cleft reorganizes the active site to active transcription. Additional N4 RNAPII structural studies utilizing ssDNA templates with the consensus promoter sequence and structures of the N4 RNAPII-gp2-DNA ternary complex could confirm these hypotheses and define all protein-protein and protein-DNA contacts required for N4 RNAPII transcription initiation. In the absence of structural data, techniques such H-D exchange or FRET may be utilized to characterize the conformational changes induced upon gp2-N4 RNAPII interaction and ternary complex formation upon promoter DNA binding.

Molecular genetics, biochemical, and structural studies are required to fully interrogate the role of the AT-rich recognition loop and β-IH on transcription factor dependence for N4 RNAPII homologs in other N4-like phages (Figure III.13). First of all, genetic or biochemical studies are required to define the required promoter sequences and determine whether each N4 RNAPII homolog requires additional factors to initiate transcription *in vivo*. Obtaining crystal structures of gp2-independent N4 RNAPII homologs in complex with promoter DNA would elucidate the differences in AT-rich recognition loop and β-IH structure and placement relative to those observed through corresponding N4 RNAPII structural analyses. It would also be of

considerable interest to swap the AT-rich recognition loop and β-IH sequences into N4 RNAPII to determine whether sequences at these positions are sufficient to confer transcription factor independence.

Finally, the determination of the complete mechanism of N4 RNAPII promoter recognition requires the identification and characterization of the protein factor required for promoter melting. Transcription from promoters located in N4 genomic DNA, but not plasmid-resident promoters, is dependent on the activity of gp1 during N4 infection of host cells expressing N4 RNAPII and gp2 (A. Demidenko, unpublished). This result implies that gp1 activates N4 RNAPII transcription in a template-conformation specific manner. Gp1 may act as an architectural transcription factor that binds dsDNA sequences on its own and induces DNA conformational changes that render the template competent for N4 RNAPII-gp2 binding. In this model, dsDNA binding and bending or unwinding could be directly measured with purified gp1 in DNA footprinting and 2-aminopurine assays *in vitro*. Conversely, gp1 may act as an allosteric effector similar to TFB2M or Mtf1 in mtRNAP transcription, reorienting the promoter recognition elements of the polymerase to render to enzyme competent for dsDNA-binding and melting. Under this mechanism, gp1 is expected to interact directly with N4 RNAPII, which could be directly tested through co-immunoprecipitation assays.

BIBLIOGRAPHY

1. Choi, K. H., McPartland, J., Kaganman, I., Bowman, V. D., Rothman-Denes, L. B., and Rossmann, M. G. (2008) Insight into DNA and protein transport in double-stranded DNA viruses: the structure of bacteriophage N4. *J. Mol. Biol.* **378**, 726–36

2. Kiino, D. R., and Rothman-Denes, L. B. (1989) Genetic analysis of bacteriophage N4 adsorption. *J. Bacteriol.* **171**, 4595–602

3. Kiino, D. R., Singer, M. S., and Rothman-Denes, L. B. (1993) Two overlapping genes encoding membrane proteins required for bacteriophage N4 adsorption. *J. Bacteriol.* **175**, 7081–5

4. McPartland, J., and Rothman-Denes, L. B. (2009) The tail sheath of bacteriophage N4 interacts with the *Escherichia coli* receptor. *J. Bacteriol.* **191**, 525–32

5. Falco, S. C., and Rothman-Denes, L. B. (1979) Bacteriophage N4-induced transcribing activities in *Escherichia coli*. I. Detection and characterization in cell extracts. *Virology*. **95**, 454–65

6. Falco, S. C., and Rothman-Denes, L. B. (1979) Bacteriophage N4-induced transcribing activities in *Escherichia coli*. II. Association of the N4 transcriptional apparatus with the cytoplasmic membrane. *Virology*. **95**, 466–75

7. Falco, S. C., Zivin, R., and Rothman-Denes, L. B. (1978) Novel template requirements of N4 virion RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **75**, 3220–4

8. Gellert, M., O'Dea, M. H., Itoh, T., and Tomizawa, J. (1976) Novobiocin and coumermycin inhibit DNA supercoiling catalyzed by DNA gyrase. *Proc. Natl. Acad. Sci. U. S. A.* **73**, 4474–8

9. Glucksmann, M. A., Markiewicz, P., Malone, C., and Rothman-Denes, L. B. (1992) Specific sequences and a hairpin structure in the template strand are required for N4 virion RNA polymerase promoter recognition. *Cell*. **70**, 491–500

10. Dai, X., Greizerstein, M. B., Nadas-Chinni, K., and Rothman-Denes, L. B. (1997) Supercoil-induced extrusion of a regulatory DNA hairpin. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 2174–9

11. Markiewicz, P., Malone, C., Chase, J. W., and Rothman-Denes, L. B. (1992) *Escherichia coli* single-stranded DNA-binding protein is a supercoiled template-dependent transcriptional activator of N4 virion RNA polymerase. *Genes Dev.* **6**, 2010–9

12. Glucksmann-Kuis, M. A., Dai, X., Markiewicz, P., and Rothman-Denes, L. B. (1996) *E. coli* SSB activates N4 virion RNA polymerase promoters by stabilizing a DNA hairpin required for promoter recognition. *Cell*. **84**, 147–54

13. Dai, X., and Rothman-Denes, L. B. (1998) Sequence and DNA structural determinants of N4 virion RNA polymerase-promoter recognition. *Genes Dev.* **12**, 2782–90

14. Davydova, E. K., Santangelo, T. J., and Rothman-Denes, L. B. (2007) Bacteriophage N4 virion RNA polymerase interaction with its promoter DNA hairpin. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7033–7038

15. Willis, S. H., Kazmierczak, K. M., Carter, R. H., and Rothman-Denes, L. B. (2002) N4 RNA polymerase II, a heterodimeric RNA polymerase with homology to the single-subunit family of RNA polymerases. *J. Bacteriol.* **184**, 4952–61

16. Zehring, W. A., and Rothman-Denes, L. B. (1983) Purification and characterization of coliphage N4 RNA polymerase II activity from infected cell extracts. *J. Biol. Chem.* **258**, 8074–80

17. Abravaya, K., and Rothman-Denes, L. B. (1989) *In vitro* requirements for N4 RNA polymerase II-specific initiation. *J. Biol. Chem.* **264**, 12695–9

18. Carter, R. H., Demidenko, A. A., Hattingh-Willis, S., and Rothman-Denes, L. B. (2003) Phage N4 RNA polymerase II recruitment to DNA by a single-stranded DNA-binding protein. *Genes Dev.* **17**, 2334–45

19. Hammer, M. D. (2004) *Nucleic acid sequence determinants of bacteriophage N4 middle and late transcription*. Ph.D. thesis, University of Chicago

20. Guinta, D., Stambouly, J., Falco, S. C., Rist, J. K., and Rothman-Denes, L. B. (1986) Host and phage-coded functions required for coliphage N4 DNA replication. *Virology*. **150**, 33–44

21. Lindberg, G., Kowalczykowski, S. C., Rist, J. K., Sugino, A., and Rothman-Denes, L. B. (1989) Purification and characterization of the coliphage N4-coded single-stranded DNA binding protein. *J. Biol. Chem.* **264**, 12700–8

22. Choi, M., Miller, A., Cho, N. Y., and Rothman-Denes, L. B. (1995) Identification, cloning, and characterization of the bacteriophage N4 gene encoding the single-stranded DNA-binding protein. A protein required for phage replication, recombination, and late transcription. *J. Biol. Chem.* **270**, 22541–7

23. Cho, N. Y., Choi, M., and Rothman-Denes, L. B. (1995) The bacteriophage N4-coded single-stranded DNA-binding protein (N4SSB) is the transcriptional activator of *Escherichia coli* RNA polymerase at N4 late promoters. *J. Mol. Biol.* **246**, 461–71

24. Falco, S. C., Zehring, W., and Rothman-Denes, L. B. (1980) DNA-dependent RNA polymerase from bacteriophage N4 virions. Purification and characterization. *J. Biol. Chem.* **255**, 4339–47

25. Zivin, R., Zehring, W., and Rothman-Denes, L. B. (1981) Transcriptional map of bacteriophage N4. Location and polarity of N4 RNAs. *J. Mol. Biol.* **152**, 335–56

26. Santangelo, T. J., and Artsimovitch, I. (2011) Termination and antitermination: RNA polymerase runs a stop sign. *Nat. Rev. Microbiol.* **9**, 319–29

27. Hinton, D. M. (2010) Transcriptional control in the prereplicative phase of T4 development. *Virol. J.* **7**, 289

28. Losick, R., and Pero, J. (1981) Cascades of sigma factors. *Cell*. **25**, 582–4

29. Studier, F. W. (1972) Bacteriophage T7. *Science*. **176**, 367–76

30. Semenova, E., Djordjevic, M., Shraiman, B., and Severinov, K. (2005) The tale of two RNA polymerases: transcription profiling and gene expression strategy of bacteriophage Xp10. *Mol. Microbiol.* **55**, 764–77

31. Lenneman, B. R., and Rothman-Denes, L. B. (2015) Structural and biochemical investigation of bacteriophage N4-encoded RNA polymerases. *Biomolecules*. **5**, 647–67

32. Schito, G. C. (1974) Development of coliphage N4: ultrastructural studies. *J. Virol.* **13**, 186–96

33. Khuong, N. (2014) *Coliphage N4 inhibits cell division by targeting FtsZ and FtsA*. Ph.D. thesis, University of Chicago

34. Yano, S. T., and Rothman-Denes, L. B. (2011) A phage-encoded inhibitor of *Escherichia coli* DNA replication targets the DNA polymerase clamp loader. *Mol. Microbiol.* **79**, 1325–38

35. Stojković, E. A., and Rothman-Denes, L. B. (2007) Coliphage N4 N-acetylmuramidase defines a new family of murein hydrolases. *J. Mol. Biol.* **366**, 406–19

36. Rohwer, F. (2003) Global phage diversity. *Cell*. **113**, 141

37. Suttle, C. A. (2007) Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–12

38. Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., and Hatfull, G. F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2192–7

39. Wommack, K. E., and Colwell, R. R. (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114

40. Wilhelm, S. W., Jeffrey, W. H., Suttle, C. a, and Mitchell, D. L. (2002) Estimation of biologically damaging UV levels in marine surface waters with DNA and viral dosimeters. *Photochem. Photobiol.* **76**, 268–73

41. Fuhrman, J. A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature*. **399**, 541–8

42. Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F., and Gordon, J. I. (2012) Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10**, 607–17

43. Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., Poulos, B. T., Solonenko, N., Lara, E., Poulain, J., Pesant, S., Kandels-Lewis, S., Dimier, C., Picheral, M., Searson, S., Cruaud, C., Alberti, A., Duarte, C. M., Gasol, J. M., Vaqué, D., Tara Oceans Coordinators, Bork, P., Acinas, S. G., Wincker, P., and Sullivan, M. B. (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*. **537**, 689–693

44. Hatfull, G. F. (2015) Dark matter of the biosphere: the amazing world of bacteriophage diversity. *J. Virol.* **89**, 8107–10

45. Hatfull, G. F., and Hendrix, R. W. (2011) Bacteriophages and their genomes. *Curr. Opin. Virol.* **1**, 298–303

46. Hayes, S., Mahony, J., Nauta, A., and van Sinderen, D. (2017) Metagenomic approaches to assess bacteriophages in various environmental niches. *Viruses*. **9**, 1–22

47. Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., and Ghai, R. (2013) Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**, e1003987

48. Roux, S., Hallam, S. J., Woyke, T., and Sullivan, M. B. (2015) Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife*. **4**, 1–20

49. Krishnamurthy, S. R., and Wang, D. (2017) Origins and challenges of viral dark matter. *Virus Res.* **239**, 136–142

50. Kleiner, M., Hooper, L. V., and Duerkop, B. A. (2015) Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics*. **16**, 7

51. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*. **5**, 69

52. Rosen, G. L., Reichenberger, E. R., and Rosenfeld, A. M. (2011) NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads.

Bioinformatics. **27**, 127–9

53. Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*. **3**, e985

54. Leplae, R., Hebrant, A., Wodak, S. J., and Toussaint, A. (2004) ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Res.* **32**, D45-9

55. Jurtz, V. I., Villarroel, J., Lund, O., Voldby Larsen, M., and Nielsen, M. (2016) MetaPhinder-identifying bacteriophage sequences in metagenomic data sets. *PLoS One*. **11**, e0163111

56. Westmoreland, B. C., Szybalski, W., and Ris, H. (1969) Mapping of deletions and substitutions in heteroduplex DNA molecules of bacteriophage lambda by electron microscopy. *Science*. **163**, 1343–8

57. Simon, M. N., Davis, R. W., and Davidson, N. (1971) Heteroduplexes of DNA molecules of lambdoid phages: physical mapping of their base sequence relationships by electron microscopy. in *The bacteriophage lambda* (Hershey, A. D. ed), pp. 313–328, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 10.1101/087969102.2.313

58. Susskind, M. M., and Botstein, D. (1978) Molecular genetics of bacteriophage P22. *Microbiol. Rev.* **42**, 385–413

59. Lawrence, J. G., Hatfull, G. F., and Hendrix, R. W. (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **184**, 4891–905

60. Hendrix, R. W. (2002) Bacteriophages: evolution of the majority. *Theor. Popul. Biol.* **61**, 471–80

61. Juhala, R. J., Ford, M. E., Duda, R. L., Youlton, A., Hatfull, G. F., and Hendrix, R. W. (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J. Mol. Biol.* **299**, 27–51

62. Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J. G., Jacobs, W. R., Hendrix, R. W., and Hatfull, G. F. (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell*. **113**, 171–82

63. Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25**, 762–77

64. Brüssow, H., and Hendrix, R. W. (2002) Phage genomics: small is beautiful. *Cell*. **108**, 13–16

65. Jang, H. Bin, Fagutao, F. F., Nho, S. W., Park, S. Bin, Cha, I. S., Yu, J. E., Lee, J. S., Im, S. P., Aoki, T., and Jung, T. S. (2013) Phylogenomic network and comparative genomics reveal a diverged member of the ΦKZ-related group, marine *Vibrio* phage ΦJM-2012. *J. Virol.* **87**, 12866–78

66. Hatfull, G. F., Jacobs-Sera, D., Lawrence, J. G., Pope, W. H., Russell, D. A., Ko, C.-C., Weber, R. J., Patel, M. C., Germane, K. L., Edgar, R. H., Hoyte, N. N., Bowman, C. A., Tantoco, A. T., Paladin, E. C., Myers, M. S., Smith, A. L., Grace, M. S., Pham, T. T., O'Brien, M. B., Vogelsberger, A. M., Hryckowian, A. J., Wynalek, J. L., Donis-Keller, H., Bogel, M. W., Peebles, C. L., Cresawn, S. G., and Hendrix, R. W. (2010) Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J. Mol. Biol.* **397**, 119–43

67. Pope, W. H., Bowman, C. A., Russell, D. A., Jacobs-Sera, D., Asai, D. J., Cresawn, S. G., Jacobs, W. R., Hendrix, R. W., Lawrence, J. G., Hatfull, G. F., Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science, Phage Hunters Integrating Research and Education, and Mycobacterial Genetics Course (2015) Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *Elife*. **4**, e06416

68. Brum, J. R., and Sullivan, M. B. (2015) Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**, 147–59

69. Youle, M., Haynes, M., and Rohwer, F. (2012) Scratching the surface of biology's dark matter. in *Viruses: Essential Agents of Life* (Witzany, G. ed), pp. 61–81, Springer, Dordrecht, Netherlands, 10.1007/978-94-007-4899-6_4

70. Comeau, A. M., Bertrand, C., Letarov, A., Tétart, F., and Krisch, H. M. (2007) Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology*. **362**, 384–96

71. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F., and Casjens, S. (2000) The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**, 504–8

72. Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., and Chisholm, S. W. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11013–8

73. Kelly, L., Ding, H., Huang, K. H., Osburne, M. S., and Chisholm, S. W. (2013) Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. *ISME J.* **7**, 1827–41

74. Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N. A. (2018) Phage puppet masters of the marine microbial realm. *Nat. Microbiol.* **3**, 754–766

75. Seed, K. D., Lazinski, D. W., Calderwood, S. B., and Camilli, A. (2013) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature*. **494**, 489–91

76. Bondy-Denomy, J., Pawluk, A., Maxwell, K. L., and Davidson, A. R. (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*. **493**, 429–32

77. Cramer, P. (2002) Common structural features of nucleic acid polymerases. *Bioessays*. **24**, 724–9

78. Joyce, C. M., and Steitz, T. A. (1995) Polymerase structures and function: variations on a theme? *J. Bacteriol.* **177**, 6321–6329

79. Joyce, C. M., and Steitz, T. A. (1994) Function and structure relationships in DNA polymerases. *Annu. Rev. Biochem.* **63**, 777–822

80. Sousa, R. (1996) Structural and mechanistic relationships between nucleic acid polymerases. *Trends Biochem. Sci.* **21**, 186–90

81. Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., and Steitz, T. A. (1992) Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science*. **256**, 1783–90

82. Delarue, M., Poch, O., Tordo, N., Moras, D., and Argos, P. (1990) An attempt to unify the structure of polymerases. *Protein Eng.* **3**, 461–7

83. Woody, A. Y., Eaton, S. S., Osumi-Davis, P. A., and Woody, R. W. (1996) Asp537 and Asp812 in bacteriophage T7 RNA polymerase as metal ion-binding sites studied by EPR, flow-dialysis, and transcription. *Biochemistry*. **35**, 144–52

84. Jeruzalmi, D., and Steitz, T. A. (1998) Structure of T7 RNA polymerase complexed to the transcriptional inhibitor T7 lysozyme. *EMBO J.* **17**, 4101–13

85. Imburgio, D., Anikin, M., and McAllister, W. T. (2002) Effects of substitutions in a conserved DX2GR sequence motif, found in many DNA-dependent nucleotide polymerases, on transcription by T7 RNA polymerase. *J. Mol. Biol.* **319**, 37–51

86. Cermakian, N., Ikeda, T. M., Miramontes, P., Lang, B. F., Gray, M. W., and Cedergren, R. (1997) On the evolution of the single-subunit RNA polymerases. *J. Mol. Evol.* **45**, 671–81

87. Kazmierczak, K. M., Davydova, E. K., Mustaev, A. A., and Rothman-Denes, L. B. (2002) The phage N4 virion RNA polymerase catalytic domain is related to single-subunit RNA polymerases. *EMBO J.* **21**, 5815–23

88. Cermakian, N., Ikeda, T. M., Cedergren, R., and Gray, M. W. (1996) Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. *Nucleic Acids Res.* **24**, 648–54

89. McAllister, W. T., and Raskin, C. A. (1993) The phage RNA polymerases are related to DNA polymerases and reverse transcriptases. *Mol. Microbiol.* **10**, 1–6

90. Chamberlin, M., and Ring, J. (1973) Characterization of T7-specific ribonucleic acid polymerase. 1. General properties of the enzymatic reaction and the template specificity of the enzyme. *J. Biol. Chem.* **248**, 2235–44

91. Studier, F. W., and Moffatt, B. A. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**, 113–30

92. Davanloo, P., Rosenberg, A. H., Dunn, J. J., and Studier, F. W. (1984) Cloning and expression of the gene for bacteriophage T7 RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 2035–9

93. Dunn, J. J., and Studier, F. W. (1983) Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.* **166**, 477–535

94. Bae, B., Davis, E., Brown, D., Campbell, E. A., Wigneshweraraj, S., and Darst, S. A. (2013) Phage T7 gp2 inhibition of *Escherichia coli* RNA polymerase involves misappropriation of σ70 domain 1.1. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19772–7

95. Nechaev, S., and Severinov, K. (1999) Inhibition of *Escherichia coli* RNA polymerase by bacteriophage T7 gene 2 protein. *J. Mol. Biol.* **289**, 815–26

96. Nechaev, S., and Severinov, K. (2003) Bacteriophage-induced modifications of host RNA polymerase. *Annu. Rev. Microbiol.* **57**, 301–22

97. McAllister, W. T., and Wu, H. L. (1978) Regulation of transcription of the late genes of bacteriophage T7. *Proc. Natl. Acad. Sci. U. S. A.* **75**, 804–8

98. Zhang, X., and Studier, F. W. (1997) Mechanism of inhibition of bacteriophage T7 RNA polymerase by T7 lysozyme. *J. Mol. Biol.* **269**, 10–27

99. Sousa, R., Chung, Y. J., Rose, J. P., and Wang, B. C. (1993) Crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution. *Nature*. **364**, 593–9

100. Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G., and Steitz, T. A. (1985) Structure of large fragment of *Escherichia coli* DNA polymerase I complexed with dTMP. *Nature.* **313**, 762–6

101. Muller, D. K., Martin, C. T., and Coleman, J. E. (1988) Processivity of proteolytically modified forms of T7 RNA polymerase. *Biochemistry*. **27**, 5763–71

102. He, B., Rong, M., Durbin, R. K., and McAllister, W. T. (1997) A mutant T7 RNA polymerase that is defective in RNA binding and blocked in the early stages of transcription. *J. Mol. Biol.* **265**, 275–88

103. Osumi-Davis, P. A., de Aguilera, M. C., Woody, R. W., and Woody, A. Y. (1992) Asp537, Asp812 are essential and Lys631, His811 are catalytically significant in bacteriophage T7 RNA polymerase activity. *J. Mol. Biol.* **226**, 37–45

104. Steitz, T. A., Smerdon, S. J., Jäger, J., and Joyce, C. M. (1994) A unified polymerase mechanism for nonhomologous DNA and RNA polymerases. *Science*. **266**, 2022–5

105. Bonner, G., Patra, D., Lafer, E. M., and Sousa, R. (1992) Mutations in T7 RNA polymerase that support the proposal for a common polymerase active site structure. *EMBO J.* **11**, 3767–75

106. Brieba, L. G., Gopal, V., and Sousa, R. (2001) Scanning mutagenesis reveals roles for helix n of the bacteriophage T7 RNA polymerase thumb subdomain in transcription complex stability, pausing, and termination. *J. Biol. Chem.* **276**, 10306–13

107. Bonner, G., Lafer, E. M., and Sousa, R. (1994) The thumb subdomain of T7 RNA polymerase functions to stabilize the ternary complex during processive transcription. *J. Biol. Chem.* **269**, 25129–36

108. Mentesana, P. E., Chin-Bow, S. T., Sousa, R., and McAllister, W. T. (2000) Characterization of halted T7 RNA polymerase elongation complexes reveals multiple factors that contribute to stability. *J. Mol. Biol.* **302**, 1049–62

109. Cheetham, G. M. T., and Steitz, T. A. (1999) Structure of a transcribing T7 RNA polymerase initiation complex. *Science*. **286**, 2305–9

110. Maksimova, T. G., Mustayev, A., Zaychikov, E. F., Lyakhov, D. L., Tunitskaya, V. L., Akbarov, A. K., Luchin, S. V, Rechinsky, V. O., Chernov, B. K., and Kochetkov, S. N. (1991) Lys631 residue in the active site of the bacteriophage T7 RNA polymerase. Affinity labeling and site-directed mutagenesis. *Eur. J. Biochem.* **195**, 841–7

111. Raskin, C. A., Diaz, G., Joho, K., and McAllister, W. T. (1992) Substitution of a single bacteriophage T3 residue in bacteriophage T7 RNA polymerase at position 748 results in a switch in promoter specificity. *J. Mol. Biol.* **228**, 506–15

112. Rong, M., He, B., McAllister, W. T., and Durbin, R. K. (1998) Promoter specificity determinants of T7 RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 515–9

113. Chapman, K. A., and Burgess, R. R. (1987) Construction of bacteriophage T7 late promoters with point mutations and characterization by *in vitro* transcription properties. *Nucleic Acids Res.* **15**, 5413–32

114. Chapman, K. A., Gunderson, S. I., Anello, M., Wells, R. D., and Burgess, R. R. (1988) Bacteriophage T7 late promoters with point mutations: quantitative footprinting and *in vivo* expression. *Nucleic Acids Res.* **16**, 4511–24

115. Ujvári, A., and Martin, C. T. (1997) Identification of a minimal binding element within the T7 RNA polymerase promoter. *J. Mol. Biol.* **273**, 775–81

116. Li, T., Ho, H. H., Maslak, M., Schick, C., and Martin, C. T. (1996) Major groove recognition elements in the middle of the T7 RNA polymerase promoter. *Biochemistry*. **35**, 3722–7

117. Lee, S. S., and Kang, C. (1993) Two base pairs at -9 and -8 distinguish between the bacteriophage T7 and SP6 promoters. *J. Biol. Chem.* **268**, 19299–19304

118. Klement, J. F., Moorefield, M. B., Jorgensen, E., Brown, J. E., Risman, S., and McAllister, W. T. (1990) Discrimination between bacteriophage T3 and T7 promoters by the T3 and T7 RNA polymerases depends primarily upon a three base-pair region located 10 to 12 base-pairs upstream from the start site. *J. Mol. Biol.* **215**, 21–9

119. Imburgio, D., Rong, M., Ma, K., and McAllister, W. T. (2000) Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. *Biochemistry*. **39**, 10419–30

120. Cheetham, G. M., Jeruzalmi, D., and Steitz, T. A. (1999) Structural basis for initiation of transcription from an RNA polymerase-promoter complex. *Nature*. **399**, 80–3

121. Schick, C., and Martin, C. T. (1995) Tests of a model of specific contacts in T7 RNA polymerase-promoter interactions. *Biochemistry*. **34**, 666–72

122. Muller, D. K., Martin, C. T., and Coleman, J. E. (1989) T7 RNA polymerase interacts with its promoter from one side of the DNA helix. *Biochemistry*. **28**, 3306–13

123. Bandwar, R. P., and Patel, S. S. (2002) The energetics of consensus promoter opening by T7 RNA polymerase. *J. Mol. Biol.* **324**, 63–72

124. Bandwar, R. P., and Patel, S. S. (2001) Peculiar 2-aminopurine fluorescence monitors the dynamics of open complex formation by bacteriophage T7 RNA polymerase. *J. Biol. Chem.* **276**, 14075–82

125. Ujvári, A., and Martin, C. T. (2000) Evidence for DNA bending at the T7 RNA polymerase promoter. *J. Mol. Biol.* **295**, 1173–84

126. Sousa, R., and Padilla, R. (1995) A mutant T7 RNA polymerase as a DNA polymerase. *EMBO J.* **14**, 4609–21

127. Temiakov, D., Patlan, V., Anikin, M., McAllister, W. T., Yokoyama, S., and Vassylyev, D. G. (2004) Structural basis for substrate selection by T7 RNA polymerase. *Cell*. **116**, 381–91

128. Martin, C. T., Muller, D. K., and Coleman, J. E. (1988) Processivity in early stages of transcription by T7 RNA polymerase. *Biochemistry*. **27**, 3966–74

129. Ikeda, R. A., and Richardson, C. C. (1986) Interactions of the RNA polymerase of bacteriophage T7 with its promoter during binding and initiation of transcription. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 3614–8

130. Tang, G.-Q., Roy, R., Ha, T., and Patel, S. S. (2008) Transcription initiation in a single-subunit RNA polymerase proceeds through DNA scrunching and rotation of the N-terminal subdomains. *Mol. Cell*. **30**, 567–77

131. Brieba, L. G., and Sousa, R. (2001) T7 promoter release mediated by DNA scrunching. *EMBO J.* **20**, 6826–35

132. Bandwar, R. P., Tang, G.-Q., and Patel, S. S. (2006) Sequential release of promoter contacts during transcription initiation to elongation transition. *J. Mol. Biol.* **360**, 466–83

133. Gong, P., and Martin, C. T. (2006) Mechanism of instability in abortive cycling by T7 RNA polymerase. *J. Biol. Chem.* **281**, 23533–44

134. Tahirov, T. H., Temiakov, D., Anikin, M., Patlan, V., McAllister, W. T., Vassylyev, D. G., and Yokoyama, S. (2002) Structure of a T7 RNA polymerase elongation complex at 2.9 Å resolution. *Nature*. **420**, 43–50

135. Bandwar, R. P., Ma, N., Emanuel, S. A., Anikin, M., Vassylyev, D. G., Patel, S. S., and McAllister, W. T. (2007) The transition to an elongation complex by T7 RNA polymerase is a multistep process. *J. Biol. Chem.* **282**, 22879–86

136. Yin, Y. W., and Steitz, T. A. (2004) The structural mechanism of translocation and helicase activity in T7 RNA polymerase. *Cell*. **116**, 393–404

137. Steitz, T. A. (2004) The structural basis of the transition from initiation to elongation phases of transcription, as well as translocation and strand separation, by T7 RNA polymerase. *Curr. Opin. Struct. Biol.* **14**, 4–9

138. Temiakov, D., Mentesana, P. E., Ma, K., Mustaev, A., Borukhov, S., and McAllister, W. T. (2000) The specificity loop of T7 RNA polymerase interacts first with the promoter and then with the elongating transcript, suggesting a mechanism for promoter clearance. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 14109–14

139. Yin, Y. W., and Steitz, T. A. (2002) Structural basis for the transition from initiation to elongation transcription in T7 RNA polymerase. *Science*. **298**, 1387–95

140. Greenleaf, A. L., Kelly, J. L., and Lehman, I. R. (1986) Yeast RPO41 gene product is required for transcription and maintenance of the mitochondrial genome. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 3391–4

141. Masters, B. S., Stohl, L. L., and Clayton, D. A. (1987) Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. *Cell*. **51**, 89–99

142. Tiranti, V., Savoia, A., Forti, F., D'Apolito, M. F., Centra, M., Rocchi, M., and Zeviani, M. (1997) Identification of the gene encoding the human mitochondrial RNA polymerase (h-mtRPOL) by cyberscreening of the Expressed Sequence Tags database. *Hum. Mol. Genet.* **6**, 615–25

143. Deshpande, A. P., and Patel, S. S. (2012) Mechanism of transcription initiation by the yeast mitochondrial RNA polymerase. *Biochim. Biophys. Acta*. **1819**, 930–8

144. Ringel, R., Sologub, M., Morozov, Y. I., Litonin, D., Cramer, P., and Temiakov, D. (2011) Structure of human mitochondrial RNA polymerase. *Nature*. **478**, 269–73

145. Nayak, D., Guo, Q., and Sousa, R. (2009) A promoter recognition mechanism common to yeast mitochondrial and phage T7 RNA polymerases. *J. Biol. Chem.* **284**, 13641–7

146. Paratkar, S., Deshpande, A. P., Tang, G.-Q., and Patel, S. S. (2011) The N-terminal domain of the yeast mitochondrial RNA polymerase regulates multiple steps of transcription. *J. Biol. Chem.* **286**, 16109–20

147. Shadel, G. S., and Clayton, D. A. (1995) A *Saccharomyces cerevisiae* mitochondrial transcription factor, sc-mtTFB, shares features with sigma factors but is functionally distinct. *Mol. Cell. Biol.* **15**, 2101–8

148. Falkenberg, M., Gaspari, M., Rantanen, A., Trifunovic, A., Larsson, N.-G., and Gustafsson, C. M. (2002) Mitochondrial transcription factors B1 and B2 activate transcription of human mtDNA. *Nat. Genet.* **31**, 289–94

149. McCulloch, V., Seidel-Rogol, B. L., and Shadel, G. S. (2002) A human mitochondrial transcription factor is related to RNA adenine methyltransferases and binds S-adenosylmethionine. *Mol. Cell. Biol.* **22**, 1116–25

150. Tzagoloff, A., and Myers, A. M. (1986) Genetics of mitochondrial biogenesis. *Annu. Rev. Biochem.* **55**, 249–85

151. Foury, F., Roganti, T., Lecrenier, N., and Purnelle, B. (1998) The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*. *FEBS Lett.* **440**, 325–31

152. Turk, E. M., Das, V., Seibert, R. D., and Andrulis, E. D. (2013) The mitochondrial RNA landscape of *Saccharomyces cerevisiae*. *PLoS One*. **8**, e78105

153. Biswas, T. K. (1998) Usage of non-canonical promoter sequence by the yeast mitochondrial RNA polymerase. *Gene*. **212**, 305–14

154. Biswas, T. K., Edwards, J. C., Rabinowitz, M., and Getz, G. S. (1985) Characterization of a yeast mitochondrial promoter by deletion mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 1954–8

155. Schinkel, A. H., Groot Koerkamp, M. J., Van der Horst, G. T., Touw, E. P., Osinga, K. A., Van der Bliek, A. M., Veeneman, G. H., Van Boom, J. H., and Tabak, H. F. (1986) Characterization of the promoter of the large ribosomal RNA gene in yeast mitochondria and separation of mitochondrial RNA polymerase into two different functional components. *EMBO J.* **5**, 1041–7

156. Wettstein-Edwards, J., Ticho, B. S., Martin, N. C., Najarian, D., and Getz, G. S. (1986) *In vitro* transcription and promoter strength analysis of five mitochondrial tRNA promoters in yeast. *J. Biol. Chem.* **261**, 2905–11

157. Biswas, T. K., and Getz, G. S. (1986) A critical base in the yeast mitochondrial nonanucleotide promoter. Abolition of promoter activity by mutation at the -2 position. *J. Biol. Chem.* **261**, 3927–30

158. Deshpande, A. P., and Patel, S. S. (2014) Interactions of the yeast mitochondrial RNA polymerase with the +1 and +2 promoter bases dictate transcription initiation efficiency. *Nucleic Acids Res.* **42**, 11721–32

159. Biswas, T. K., Ticho, B., and Getz, G. S. (1987) *In vitro* characterization of the yeast mitochondrial promoter using single-base substitution mutants. *J. Biol. Chem.* **262**, 13690–6

160. Paratkar, S., and Patel, S. S. (2010) Mitochondrial transcription factor Mtf1 traps the unwound non-template strand to facilitate open complex formation. *J. Biol. Chem.* **285**, 3949–56

161. Drakulic, S., Wang, L., Cuéllar, J., Guo, Q., Velázquez, G., Martín-Benito, J., Sousa, R., and Valpuesta, J. M. (2014) Yeast mitochondrial RNAP conformational changes are regulated by interactions with the mitochondrial transcription factor. *Nucleic Acids Res.* **42**, 11246–60

162. Matsunaga, M., and Jaehning, J. A. (2004) Intrinsic promoter recognition by a “core” RNA polymerase. *J. Biol. Chem.* **279**, 44239–42

163. Schinkel, A. H., Koerkamp, M. J., Touw, E. P., and Tabak, H. F. (1987) Specificity factor of yeast mitochondrial RNA polymerase. Purification and interaction with core RNA polymerase. *J. Biol. Chem.* **262**, 12785–91

164. Jang, S. H., and Jaehning, J. A. (1991) The yeast mitochondrial RNA polymerase specificity factor, MTF1, is similar to bacterial sigma factors. *J. Biol. Chem.* **266**, 22671–7

165. Tang, G.-Q., Paratkar, S., and Patel, S. S. (2009) Fluorescence mapping of the open complex of yeast mitochondrial RNA polymerase. *J. Biol. Chem.* **284**, 5514–22

166. Mangus, D. A., Jang, S. H., and Jaehning, J. A. (1994) Release of the yeast mitochondrial RNA polymerase specificity factor from transcription complexes. *J. Biol. Chem.* **269**, 26568–74

167. Schubot, F. D., Chen, C. J., Rose, J. P., Dailey, T. A., Dailey, H. A., and Wang, B. C. (2001) Crystal structure of the transcription factor sc-mtTFB offers insights into mitochondrial transcription. *Protein Sci.* **10**, 1980–8

168. Cliften, P. F., Park, J. Y., Davis, B. P., Jang, S. H., and Jaehning, J. A. (1997) Identification of three regions essential for interaction between a sigma-like factor and core RNA polymerase. *Genes Dev.* **11**, 2897–909

169. Cliften, P. F., Jang, S. H., and Jaehning, J. a (2000) Identifying a core RNA polymerase surface critical for interactions with a sigma-like specificity factor. *Mol. Cell. Biol.* **20**, 7013–23

170. Savkina, M., Temiakov, D., McAllister, W. T., and Anikin, M. (2010) Multiple functions of yeast mitochondrial transcription factor Mtf1p during initiation. *J. Biol. Chem.* **285**, 3957–64

171. Tang, G.-Q., Deshpande, A. P., and Patel, S. S. (2011) Transcription factor-dependent DNA bending governs promoter recognition by the mitochondrial RNA polymerase. *J. Biol. Chem.* **286**, 38805–13

172. Tang, G.-Q., and Patel, S. S. (2006) T7 RNA polymerase-induced bending of promoter DNA is coupled to DNA opening. *Biochemistry*. **45**, 4936–46

173. Kim, H., Tang, G.-Q., Patel, S. S., and Ha, T. (2012) Opening-closing dynamics of the mitochondrial transcription pre-initiation complex. *Nucleic Acids Res.* **40**, 371–80

174. Amiott, E. A., and Jaehning, J. A. (2006) Sensitivity of the yeast mitochondrial RNA polymerase to +1 and +2 initiating nucleotides. *J. Biol. Chem.* **281**, 34982–8

175. Karlok, M. A., Jang, S.-H., and Jaehning, J. A. (2002) Mutations in the yeast mitochondrial RNA polymerase specificity factor, Mtf1, verify an essential role in promoter utilization. *J. Biol. Chem.* **277**, 28143–9

176. Amiott, E. A., and Jaehning, J. A. (2006) Mitochondrial transcription is regulated via an ATP "sensing" mechanism that couples RNA abundance to respiration. *Mol. Cell*. **22**, 329–38

177. Arnold, J. J., Smidansky, E. D., Moustafa, I. M., and Cameron, C. E. (2012) Human mitochondrial RNA polymerase: structure-function, mechanism and inhibition. *Biochim. Biophys. Acta*. **1819**, 948–60

178. Falkenberg, M., Larsson, N.-G., and Gustafsson, C. M. (2007) DNA replication and transcription in mammalian mitochondria. *Annu. Rev. Biochem.* **76**, 679–99

179. Scarpulla, R. C. (2008) Transcriptional paradigms in mammalian mitochondrial biogenesis and function. *Physiol. Rev.* **88**, 611–38

180. Chang, D. D., and Clayton, D. A. (1984) Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. *Cell*. **36**, 635–43

181. Montoya, J., Christianson, T., Levens, D., Rabinowitz, M., and Attardi, G. (1982) Identification of initiation sites for heavy-strand and light-strand transcription in human mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 7195–9

182. Ojala, D., Montoya, J., and Attardi, G. (1981) tRNA punctuation model of RNA processing in human mitochondria. *Nature*. **290**, 470–4

183. Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., and Young, I. G. (1981) Sequence and organization of the human mitochondrial genome. *Nature*. **290**, 457–65

184. Zollo, O., Tiranti, V., and Sondheimer, N. (2012) Transcriptional requirements of the distal heavy-strand promoter of mtDNA. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 6508–12

185. Hixson, J. E., and Clayton, D. A. (1985) Initiation of transcription from each of the two human mitochondrial promoters requires unique nucleotides at the transcriptional start sites. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 2660–4

186. Dairaghi, D. J., Shadel, G. S., and Clayton, D. A. (1995) Human mitochondrial transcription factor A and promoter spacing integrity are required for transcription initiation. *Biochim. Biophys. Acta*. **1271**, 127–34

187. Gaspari, M., Falkenberg, M., Larsson, N.-G., and Gustafsson, C. M. (2004) The mitochondrial RNA polymerase contributes critically to promoter specificity in mammalian cells. *EMBO J.* **23**, 4606–14

188. Lodeiro, M. F., Uchida, A., Bestwick, M., Moustafa, I. M., Arnold, J. J., Shadel, G. S., and Cameron, C. E. (2012) Transcription from the second heavy-strand promoter of human mtDNA is repressed by transcription factor A *in vitro*. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 6513–8

189. Enríquez, J. A., Fernández-Silva, P., Pérez-Martos, A., López-Pérez, M. J., and Montoya, J. (1996) The synthesis of mRNA in isolated mitochondria can be maintained for several hours and is inhibited by high levels of ATP. *Eur. J. Biochem.* **237**, 601–10

190. Lodeiro, M. F., Uchida, A. U., Arnold, J. J., Reynolds, S. L., Moustafa, I. M., and Cameron, C. E. (2010) Identification of multiple rate-limiting steps during the human mitochondrial transcription cycle *in vitro*. *J. Biol. Chem.* **285**, 16387–402

191. Schwinghammer, K., Cheung, A. C. M., Morozov, Y. I., Agaronyan, K., Temiakov, D., and Cramer, P. (2013) Structure of human mitochondrial RNA polymerase elongation complex. *Nat. Struct. Mol. Biol.* **20**, 1298–303

192. Fisher, R. P., and Clayton, D. A. (1985) A transcription factor required for promoter recognition by human mitochondrial RNA polymerase. Accurate initiation at the heavy- and light-strand promoters dissected and reconstituted *in vitro*. *J. Biol. Chem.* **260**, 11330–8

193. Fisher, R. P., Lisowsky, T., Parisi, M. A., and Clayton, D. A. (1992) DNA wrapping and bending by a mitochondrial high mobility group-like transcriptional activator protein. *J. Biol. Chem.* **267**, 3358–67

194. Parisi, M. A., and Clayton, D. A. (1991) Similarity of human mitochondrial transcription factor 1 to high mobility group proteins. *Science.* **252**, 965–9

195. Malarkey, C. S., Bestwick, M., Kuhlwilm, J. E., Shadel, G. S., and Churchill, M. E. A. (2012) Transcriptional activation by mitochondrial transcription factor A involves preferential distortion of promoter DNA. *Nucleic Acids Res.* **40**, 614–24

196. Ngo, H. B., Lovely, G. A., Phillips, R., and Chan, D. C. (2014) Distinct structural features of TFAM drive mitochondrial DNA packaging versus transcriptional activation. *Nat. Commun.* **5**, 3077

197. Kaufman, B. A., Durisic, N., Mativetsky, J. M., Costantino, S., Hancock, M. A., Grutter, P., and Shoubridge, E. A. (2007) The mitochondrial transcription factor TFAM coordinates the assembly of multiple DNA molecules into nucleoid-like structures. *Mol. Biol. Cell*. **18**, 3225–36

198. Rubio-Cosials, A., Sidow, J. F., Jiménez-Menéndez, N., Fernández-Millán, P., Montoya, J., Jacobs, H. T., Coll, M., Bernadó, P., and Solà, M. (2011) Human mitochondrial transcription factor A induces a U-turn structure in the light strand promoter. *Nat. Struct. Mol. Biol.* **18**, 1281–9

199. Ngo, H. B., Kaiser, J. T., and Chan, D. C. (2011) The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. *Nat. Struct. Mol. Biol.* **18**, 1290–6

200. Morozov, Y. I., Agaronyan, K., Cheung, A. C. M., Anikin, M., Cramer, P., and Temiakov, D. (2014) A novel intermediate in transcription initiation by human mitochondrial RNA polymerase. *Nucleic Acids Res.* **42**, 3884–93

201. Morozov, Y. I., Parshin, A. V., Agaronyan, K., Cheung, A. C. M., Anikin, M., Cramer, P., and Temiakov, D. (2015) A model for transcription initiation in human mitochondria. *Nucleic Acids Res.* **43**, 3726–35

202. Posse, V., Hoberg, E., Dierckx, A., Shahzad, S., Koolmeister, C., Larsson, N.-G., Wilhelmsson, L. M., Hällberg, B. M., and Gustafsson, C. M. (2014) The amino terminal extension of mammalian mitochondrial RNA polymerase ensures promoter specific transcription initiation. *Nucleic Acids Res.* **42**, 3638–47

203. Posse, V., and Gustafsson, C. M. (2017) Human mitochondrial transcription factor B2 is required for promoter melting during initiation of transcription. *J. Biol. Chem.* **292**, 2637–2645

204. Sologub, M., Litonin, D., Anikin, M., Mustaev, A., and Temiakov, D. (2009) TFB2 is a transient component of the catalytic site of the human mitochondrial RNA polymerase. *Cell*. **139**, 934–44

205. Cotney, J., and Shadel, G. S. (2006) Evidence for an early gene duplication event in the evolution of the mitochondrial transcription factor B family and maintenance of rRNA methyltransferase activity in human mtTFB1 and mtTFB2. *J. Mol. Evol.* **63**, 707–17

206. Ramachandran, A., Basu, U., Sultana, S., Nandakumar, D., and Patel, S. S. (2017) Human mitochondrial transcription factors TFAM and TFB2M work synergistically in promoter melting during transcription initiation. *Nucleic Acids Res.* **45**, 861–874

207. Hillen, H. S., Morozov, Y. I., Sarfallah, A., Temiakov, D., and Cramer, P. (2017) Structural basis of mitochondrial transcription initiation. *Cell*. **171**, 1072-1081.e10

208. Hillen, H. S., Temiakov, D., and Cramer, P. (2018) Structural basis of mitochondrial transcription. *Nat. Struct. Mol. Biol.* **25**, 754–765

209. Rothman-Denes, L. B., and Schito, G. C. (1974) Novel transcribing activities in N4-infected *Escherichia coli*. *Virology*. **60**, 65–72

210. Falco, S. C., Laan, K. V, and Rothman-Denes, L. B. (1977) Virion-associated RNA polymerase required for bacteriophage N4 development. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 520–3

211. vander Laan, K., Falco, S. C., and Rothman-Denes, L. B. (1977) The program of RNA synthesis in N4-infected *Escherichia coli*. *Virology*. **76**, 596–601

212. Gellert, M., Mizuuchi, K., O'Dea, M. H., and Nash, H. A. (1976) DNA gyrase: an enzyme

that introduces superhelical turns into DNA. *Proc. Natl. Acad. Sci. U. S. A.* **73**, 3872–6

213. Zivin, R., Malone, C., and Rothman-Denes, L. B. (1980) Physical map of coliphage N4 DNA. *Virology*. **104**, 205–18

214. Haynes, L. L., and Rothman-Denes, L. B. (1985) N4 virion RNA polymerase sites of transcription initiation. *Cell*. **41**, 597–605

215. Dai, X., Kloster, M., and Rothman-Denes, L. B. (1998) Sequence-dependent extrusion of a small DNA hairpin at the N4 virion RNA polymerase promoters. *J. Mol. Biol.* **283**, 43–58

216. Davydova, E. K., and Rothman-Denes, L. B. (2003) *Escherichia coli* single-stranded DNA-binding protein mediates template recycling during transcription by bacteriophage N4 virion RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9250–5

217. Murakami, K. S., Davydova, E. K., and Rothman-Denes, L. B. (2008) X-ray crystal structure of the polymerase domain of the bacteriophage N4 virion RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 5046–51

218. Gleghorn, M. L., Davydova, E. K., Rothman-Denes, L. B., and Murakami, K. S. (2008) Structural basis for DNA-hairpin promoter recognition by the bacteriophage N4 virion RNA polymerase. *Mol. Cell*. **32**, 707–17

219. Gleghorn, M. L., Davydova, E. K., Basu, R., Rothman-Denes, L. B., and Murakami, K. S. (2011) X-ray crystal structures elucidate the nucleotidyl transfer reaction of transcript initiation using two nucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 3566–71

220. Basu, R. S., and Murakami, K. S. (2013) Watching the bacteriophage N4 RNA polymerase transcription by time-dependent soak-trigger-freeze X-ray crystallography. *J. Biol. Chem.* **288**, 3305–11

221. Abravaya, K., and Rothman-Denes, L. B. (1990) N4 RNA polymerase II sites of transcription initiation. *J. Mol. Biol.* **211**, 359–72

222. Zehring, W. A., Falco, S. C., Malone, C., and Rothman-Denes, L. B. (1983) Bacteriophage N4-induced transcribing activities in *E. coli*. III. A third cistron required for N4 RNA polymerase II activity. *Virology*. **126**, 678–87

223. Molodtsov, V., and Murakami, K. S. (2018) Minimalism and functionality: structural lessons from the heterodimeric N4 bacteriophage RNA polymerase II. *J. Biol. Chem.* **293**, 13616–13625

224. Markle, C. A. (2012) *Genetic and biochemical analysis of gp2 and its role in N4 RNAPII transcription initiation*. Ph.D. thesis, University of Chicago

225. R Core Team (2016) R: A language and environment for statistical computing

226. Götze, M., Pettelkau, J., Schaks, S., Bosse, K., Ihling, C. H., Krauth, F., Fritzsche, R., Kühn, U., and Sinz, A. (2012) StavroX--a software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom.* **23**, 76–87

227. Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K., and Jones, D. T. (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, W349-57

228. Boratyn, G. M., Schäffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J., and Madden, T. L. (2012) Domain Enhanced Lookup Time Accelerated BLAST. *Biol. Direct*. **7**, 12

229. Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., and Altschul, S. F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005

230. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–402

231. Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.-Y., Bateman, A., Punta, M., Attwood, T. K., Sigrist, C. J. A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D. A., Wu, C. H., Orengo, C., Sillitoe, I., Mi, H., Thomas, P. D., and Finn, R. D. (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**, D213-21

232. Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., and Bryant, S. H. (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222-6

233. Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–82

234. Laslett, D., and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–6

235. Ye, Y., Wei, B., Wen, L., and Rayner, S. (2013) BlastGraph: a comparative genomics tool based on BLAST and graph algorithms. *Bioinformatics*. **29**, 3222–4

236. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539

237. Méndez, J., Blanco, L., Lázaro, J. M., and Salas, M. (1994) Primer-terminus stabilization at the phi 29 DNA polymerase active site. Mutational analysis of conserved motif TX2GR. *J. Biol. Chem.* **269**, 30030–8

238. Krumsiek, J., Arnold, R., and Rattei, T. (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*. **23**, 1026–8

239. Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011) Easyfig: a genome comparison visualizer. *Bioinformatics*. **27**, 1009–10

240. Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–84

241. Huson, D. H., and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–67

242. Poisot, T., and B, S. D. (2015) lpbrim: LP-BRIM Bipartite Modularity.

243. Schito, G. C., Rialdi, G., and Pesce, A. (1966) The physical properties of the deoxyribonucleic acid from N4 coliphage. *Biochim. Biophys. Acta*. **129**, 491–501

244. Ohmori, H., Haynes, L. L., and Rothman-Denes, L. B. (1988) Structure of the ends of the coliphage N4 genome. *J. Mol. Biol.* **202**, 1–10

245. Wittmann, J., Klumpp, J., Moreno Switt, A. I., Yagubi, A., Ackermann, H.-W., Wiedmann, M., Svircev, A., Nash, J. H. E., and Kropinski, A. M. (2015) Taxonomic reassessment of N4-like viruses using comparative genomics and proteomics suggests a new subfamily - "*Enquartavirinae*." *Arch. Virol.* **160**, 3053–62

246. Zhao, Y., Wang, K., Jiao, N., and Chen, F. (2009) Genome sequences of two novel phages infecting marine roseobacters. *Environ. Microbiol.* **11**, 2055–64

247. Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951–60

248. Liu, B., Shadrin, A., Sheppard, C., Mekler, V., Xu, Y., Severinov, K., Matthews, S., and Wigneshweraraj, S. (2014) A bacteriophage transcription regulator inhibits bacterial transcription initiation by σ-factor displacement. *Nucleic Acids Res.* **42**, 4294–305

249. Center, M. S., Studier, F. W., and Richardson, C. C. (1970) The structural gene for a T7 endonuclease essential for phage DNA synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **65**, 242–8

250. Schito, G. C. (1973) The genetics and physiology of coliphage N4. *Virology*. **55**, 254–65

251. Rothman-Denes, L. B., Haselkorn, R., and Schito, G. C. (1972) Selective shutoff of catabolite-sensitive host syntheses by bacteriophage N4. *Virology*. **50**, 95–102

252. Miller, A., Wood, D., Ebright, R. H., and Rothman-Denes, L. B. (1997) RNA polymerase β' subunit: a target of DNA binding-independent activation. *Science*. **275**, 1655–7

253. Stojković, E. A. (2005) *Characterization of the coliphage N4-encoded N-acetylmuamidase, a member of a new family of peptidoglycan-hydrolyzing enzymes*. Ph.D. thesis, University of Chicago

254. Young, R. (2014) Phage lysis: three steps, three choices, one outcome. *J. Microbiol.* **52**, 243–58

255. Young, R., and Bläsi, U. (1995) Holins: form and function in bacteriophage lysis. *FEMS Microbiol. Rev.* **17**, 191–205

256. Bailly-Bechet, M., Vergassola, M., and Rocha, E. (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res.* **17**, 1486–95

257. Waller, A. S., Yamada, T., Kristensen, D. M., Kultima, J. R., Sunagawa, S., Koonin, E. V., and Bork, P. (2014) Classification and quantification of bacteriophage taxa in human gut metagenomes. *ISME J.* **8**, 1391–402

258. Milligan, J. F., Groebe, D. R., Witherell, G. W., and Uhlenbeck, O. C. (1987) Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic Acids Res.* **15**, 8783–98

259. Krupp, G. (1988) RNA synthesis: strategies for the use of bacteriophage RNA polymerases. *Gene*. **72**, 75–89

260. Chin, J. W., Martin, A. B., King, D. S., Wang, L., and Schultz, P. G. (2002) Addition of a photocrosslinking amino acid to the genetic code of *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11020–4

261. Lee, H. S., Dimla, R. D., and Schultz, P. G. (2009) Protein-DNA photo-crosslinking with a genetically encoded benzophenone-containing amino acid. *Bioorg. Med. Chem. Lett.* **19**, 5222–4

262. Oakley, J. L., and Coleman, J. E. (1977) Structure of a promoter for T7 RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 4266–70

263. Hartmann, G. R., Biebricher, C., Glaser, S. J., Grosse, F., Katzameyer, M. J., Lindner, A. J., Mosig, H., Nasheuer, H. P., Rothman-Denes, L. B., and Schäffner, A. R. (1988)

Initiation of transcription--a general tool for affinity labeling of RNA polymerases by autocatalysis. *Biol. Chem. Hoppe. Seyler.* **369**, 775–88