DISSERTATION

FACTORS AND MECHANISMS OF ARCHAEAL TRANSCRIPTION TERMINATION AND DNA REPAIR

Submitted by

Craig Marshall

Department of Biochemistry and Molecular Biology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2022

Doctoral Committee:

Advisor: Thomas J. Santangelo

Olve Peersen
Carol Wilusz
Tingting Yao

Copyright by Craig James Marshall 2022

All Rights Reserved

ABSTRACT

FACTORS AND MECHANISMS OF ARCHAEAL
TRANSCRIPTION TERMINATION AND DNA REPAIR

RNA synthesis by RNA polymerase (RNAP) is an essential process and must be properly regulated both temporally and spatially to ensure cellular health in dynamic environments. Regulation of RNA synthesis in response to internal and environmental stimuli is typically achieved through interactions with RNAP at all stages of the transcription cycle-initiation, elongation, and termination. While studies of transcription initiation and elongation have identified multiple regulatory transcription factors and defined mechanisms, only a handful of protein factors able to terminate transcription have yet been described, and the general mechanism of transcription termination is still highly debated. We previously identified the first two factors capable of terminating transcription elongation complexes (TECs) in Archaea from the genetically tractable *Thermococcus kodakarensis,* and use both factors as models to explore the molecular mechanisms involved in collapse of the TEC.

The Factor that terminates transcription in Archaea (FttA), a close homolog of the human CPSF subunit CPSF73, is completely conserved throughout Archaea, and appears to act analogously to the bacterial termination factor Rho, terminating transcription after the uncoupling of transcription and translation at the end of protein coding genes. We employed a novel genetic screen to verify the role of FttA in the polar repression of transcription, a phenomenon specific to regulation of genes contained within operons in prokaryotes.

Eta, a euryarchaeal-specific superfamily 2 (SF2) helicase, appears to terminate transcription in a more specialized context, potentially terminating transcription of TECs arrested at sites of DNA damage while concurrently recruiting appropriate DNA repair enzymes, akin to the bacterial termination factor Mfd. A structure-function study of Eta employing select mutations

derived from a crystallographic structure was conducted to elucidate the Eta-TEC contacts and various activities of Eta required for Eta-mediated termination. Further, many efforts were directed at establishing a role of Eta as an archaeal transcription-repair coupling factor (TRCF), and while this was not achieved, a state-of-the-art next-generation sequencing based approach to monitor nucleotide excision repair (NER) and the sub pathway transcription-coupled repair (TCR) genome-wide was developed and verified in *E.coli*.

The work in this dissertation adds valuable insight to multiple fields of research. First, exploration into the mechanism of Eta-mediated transcription termination reveals a potential shared susceptibility of core RNAP subunits to transcription termination while elucidating activities of SF2 helicases- enzymes which are ubiquitously distributed in multiple essential cellular pathways. Second, our genetic screen identifies FttA as the archaeal polarity factor, shedding light on functions of an ancestral factor indispensable in mammalian transcription termination pathways. Establishment of the novel RADAR-seq/RNA-seq measurement of NER genome-wide will likely prove instrumental in future studies of archaeal DNA repair, and potentially presents a new paradigm in research of eukaryotic-like NER by use of Archaea as a advantageous model organism.

ACKNOWLEDGEMENTS

Although only the 5th page, this acknowledgments section is the last to be completed, and maybe the most important to me. Getting to this point has involved countless experimental failures and some fulfilling successes, but mostly a lot of work. However, due to the relationships formed before and during my graduate career at Colorado State University, it never really felt like work- rather a rite of passage into a broader and exciting world of research.

I'd first like to thank my mentor, Tom Santangelo. Tom truly is an A-plus mentor who combines extreme honesty with a sincere care for and belief in his students. Without our candid scientific conversations throughout the years, I wouldn't yet be in this position. I always learned something from even our briefest conversations. Tom's advocacy for his students during the coronavirus pandemic allowed me to flourish rather than slump during this strange time, and I can't be more grateful.

To my committee- Olve, Carol, and Tingting- thank you for our great conversations over the years which fostered new ideas leading to better science. If I had to go back and do this all over, I would want these meetings to be more frequent.

My friends, both old and new, were a constant source of support and a constant reminder of how fortunate I have been throughout my life. Group chats with and visits from old friends (shout out Austin, Darren, Jimmy, JMac, Josh, Lilly, Luke, Max, Mike, and Marc) kept me going through the more frustrating times, as did the friends I have gained in Colorado. Travis, Bree, Brett, the Creek Crew (Dylan and Lindsay), and the winter gnar-shredding crews deserve a particular mention for preserving each other's sanity!

My family both overseas in the UK and back in Florida, Ohio & Pennsylvania have been invaluable to me. It has been challenging being thousands of miles away, but our constant communications and game-nights are some of my fondest memories from my time in graduate school. While my dad, Allan, is surely unhappy that I will soon be the most educated in the

family, his genuine interest in my work is truly appreciated. I can only hope to one day return the love and support I have received from him, my mum Debbie, my brother William, my sisters Tyler and Lucy, and of course, my new niece Emi!

Finally, I wouldn't even be half the person I am today without my own little family at home. Maybe I'm biased, but I'm quite sure Ambrose (AKA dogbreath, AKA stinker, AKA AmbroseDog, AKA fluffball) is the best dog in the world. Sage, Piper, and Mabel have also offered priceless furry-based support over the years.

To my fiancée Jenna- I can't quite explain how much you mean to me. Without you taking a chance and moving across the country with me, I'm not sure how things would have worked out. There really are too many ways to list that you have helped me become a better scientist, and better person. I can't wait for the next chapter of our lives.

-Craig

TABLE OF CONTENTS

ABSTRACT…….……ii
ACKNOWLEDGEMENTS…….iv

CHAPTER 1 – INTRODUCTION……...…….….1
1.1 Archaeal transcription..……..……..……1
1.1.1 Archaeal transcription initiation……..…….2
1.1.2 Archaeal transcription elongation……..4
1.1.3 Archaeal transcription termination…….……5
1.2 Archaeal transcription termination…….…….……...…6
1.3 Intrinsic transcription termination……...9
1.4 Factor dependent transcription termination..…….……10
1.4.1 Bacterial Rho-mediated termination...…….……..11
1.4.2 Mfd and transcription-coupled DNA repair in Bacteria……12
1.4.3 Factor-dependent Termination of Eukaryotic RNA Polymerase II...……12
1.4.4 FttA-mediated Archaeal Termination …..……...…….……..13
1.4.5 Eta-mediated Archaeal Termination……...15
1.5 Polar repression of transcription in prokaryotes……15
1.6 Transcription coupled nucleotide excision repair……..16
1.7 Thesis rationale…….……18
REFERENCES……..20

CHAPTER 2 – THE ARCHAEAL TERMINATION FACTOR ETA DETAIL VULNERABILITIES OF THE TRANSCRIPTION ELONGATION COMPLEX. …….……...…..…34
2.1 Introduction…...…….……..…34
2.1.1 Termination of transcription is subject to intricate control mechanisms...……...34
2.1.2 DNA vs. RNA dependent transcription termination……..……...……...33
2.1.3 Study rationale and summary……..37
2.2 Results……..…….……..37
2.2.1 Eta is a canonical superfamily 2 helicase/translocase……...37
2.2.2 Eta-mediated transcription termination requires the CTD of Eta. ……40
2.2.3 Conserved, solvent-exposed residues in the CTD of Eta are critical for motor, ATPase, and transcription termination activities.……40
2.3 Discussion…..…….……51
2.4 Methods……...…….……56
REFERENCES……..…….……..63

CHAPTER 3 – IDENTIFICATION OF THE FACTOR RESPONSIBLE FOR POLAR REPRESSION OF TRANSCRIPTION IN ARCHAEA……..……..70
3.1 Introduction…….……...…….70
3.2 Results……...……..74
3.2.1 *T. kodakarensis* strain CM003 harbors ‘polarity constructs’ upstream of essential metabolomic genes and exhibits poor growth on minimal His/Trp media…...……...74
3.2.2 A simple genetic screen performed on CM003 cells with basal mutations identified 10 mutant strains with observed Trp/His prototrophy..……..76
3.2.3 Whole genome sequencing revealed multiple mutant CM003 strains with mutations to FttA and/or RNA Polymerase ..……77

3.2.4 Mutations to both FttA and RNAP identified from the polarity screen map to regions of each enzyme critical for function and conserved between Archaea and humans……80
3.2.5 Strains harboring mutations to RNAP or FttA surviving the polarity genetic screen were able to transcribe *hisD/trpE* through polarity repressed constructs more efficiently than the CM003 mutant strain……...83
3.2.6 *T. kodakarensis* RNAP 'polarity mutant' performs comparably to the wild-type RNAP in *in vitro* transcription and transcription termination experiments……..87
3.2.7 Preliminary evidence suggests the FttA 'polarity mutant' can still efficiently couple to TECs through Spt4/5 and appears to out-perform wild-type FttA in *in vitro* transcription termination experiments…….89
3.3 Discussion……..……91
3.4 Methods…….…..……97
REFERENCES……..…….……...…..103

CHAPTER 4 – INVESTIGATIONS INTO ARCHAEAL TRANSCRIPTION COUPLED DNA REPAIR……..……..108
4.1 Introduction……..……..……..……108
4.2 Results……..……..……..……115
4.2.1 Identification of putative nucleotide excision repair enzymes from T. kodakarensis ……115
4.2.2 Putative NER enzymes do not form stably associated complexes in *T. kodakarensis*..……..……..117
4.2.3 Development of RADAR-seq to measure genome-wide transcription coupled DNA repair of UV-induced DNA damage in *E. coli*……..…122
4.3 Discussion and future directions……..……..127
4.4 Methods……...…..……..……..129
REFERENCES……..……..……132

APPENDIX A – ARCHAEAL DNA REPAIR MECHANSIMS……...139

APPENDIX B – THE ROLE OF ARCHAEAL CHROMATIN IN TRANSCRIPTION….……181

CHAPTER 1

INTRODUCTION[1]

1.1 Archaeal transcription

Proper regulation of RNA synthesis in response to internal and environmental stimuli is essential to ensure cellular health in dynamic environments, requiring interactions with RNA polymerase (RNAP) at all stages of the transcription cycle- initiation, elongation, and termination ((1–7)). RNAPs across all three Domains can be distilled down to three main domains; i) the core enzyme containing bridge helix (BH) and trigger loop (TL) elements around an active site for nucleotide addition (8–10), ii) a sliding mobile clamp domain which varies conformation across stages of the transcription cycle (11), iii) an Archaea/Eukarya-specific stalk domain making contacts with the DNA template during initiation, and nascent RNA during elongation and termination (12). The core RNAP enzyme is remarkably conserved across all extant life but there is substantially more homology between the single archaeal RNAP and eukaryotic RNAP II, adding to the body of evidence that the Archaea likely share a common ancestor with the Eukarya after the divergence of the Bacteria (13–16). The similarity in structure between the archaeal RNAP and eukaryotic RNAP II suggests regulation of the archaeal transcription cycle would resemble regulation of RNAP II more closely than bacterial RNAP, and this appears to be the case as transcription factors with eukaryotic homologs help to guide archaeal RNAP through the transcription cycle (Figure 1.1).

[1]Portions of this section are excerpts from the following publications: Sanders, T. J., Marshall, C.J., *et al.* (2021) 'Transcription | Transcription Termination', in *Encyclopedia of Biological Chemistry III*. Elsevier, pp. 435–442. and Marshall, C. J. and Santangelo, T. J. (2020) 'Archaeal DNA repair mechanisms', *Biomolecules*. MDPI AG, pp. 1–23.

1.1.1 Archaeal transcription initiation

Archaea utilize promoters which resemble eukaryotic promoters, placing a T/A rich TATA-like element ~25bp upstream of a transcription start site (TSS) and close to a purine-rich TFB recognition element (TFB)(17). However, compared to eukaryotes, Archaea have simpler enzymatic requirements for formation of pre-initiation complexes (PICs). Archaeal TATA binding protein (TBP) and TFB first interact with appropriate promoter elements to - bend DNA strands and establish the direction of transcription, respectively (18). Correct promoter TBP-TFB interactions are sufficient at many optimal promoter sequences for loading of RNAP and local melting of dsDNA required to form PICs (19, 20); the archaeal TFE facilitates efficient PIC formation in suboptimal conditions by binding initiating RNAPs and promoting an open-clamp configuration (21, 22) (Figure 1.1.i).

Transcription initiation is commonly regulated in archaeal cells as a means of controlling gene expression. Less conserved sequences within archaeal promoters can vary significantly, altering efficiency of PIC formation (and thus gene expression) through changed interactions between transcription initiation factors and the DNA template (23). Varying DNA sequences immediately downstream of the TSS can also regulate transcription initiation events by influencing the competition between promoter contacts of the PIC and active transcription elongation. Multiple isoforms of archaeal TBP, TFB, and TFE are commonly encoded in some archaeal clades (24–26), and it is plausible that varied combinations of different transcription initiation factor isoforms could allow for selective initiation at particular promoters- similar to a Bacterial sigma factor system (27). However, more investigation into transcription initiation factor isoforms is needed for any real conclusion on their effects on transcription initiation. Factor-mediated transcription initiation regulation outside of the basal transcription factors is less understood in Archaea, but is present. Many Archaea, including *T. kodakarensis*, encode histone proteins which chromatinize the genome and bare significant resemblances to the eukaryotic histone H3 (28). Archaeal histone proteins can repress transcription initiation, and be

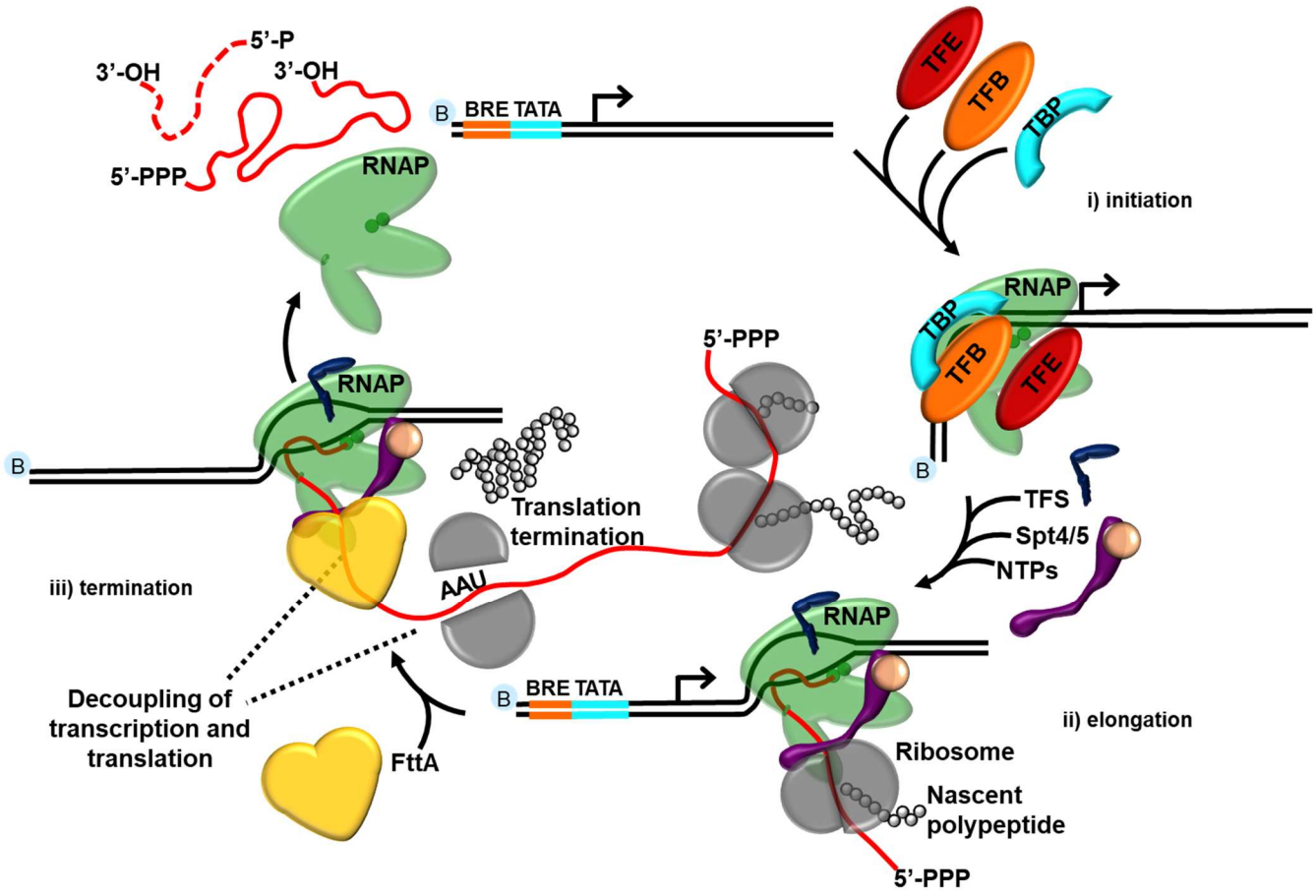


Figure 1.1. The archaeal transcription cycle. i) In initiation, RNA polymerase (green) makes interactions with TBP (cyan), TFB (orange) and TFE (red) at a BRE/TATA-based promoter, forming the pre-initiation complex. ii) During elongation, TFS aids stalled and backtracked transcription elongation complexes, and the universally conserved transcription factor Spt4/5 (purple) replaces TFE to promote a closed clamp configuration. Translation and transcription are coupled in Archaea, so during transcription of protein coding genes, ribosomes (grey) are trailing RNA polymerases during the elongation phase of transcription. iii) General transcription termination requires the decoupling of transcription and translation, where ribosome-free nascent RNA is recognized by FttA (yellow). FttA both cleaves the RNA releasing the full transcript, and releases the transcription elongation complex from the DNA template.

overcome by gene-specific transcription factors *in vitro* (29, 30), and directed changes to histone proteins which affect overall chromatin architecture alter global transcription events *in vivo* (31). Multiple transcription factors which have homologs in both Eukarya and Bacteria have been identified in Archaea which may act to regulate transcription initiation of specific individual or sets of genes although their mechanisms of action remain unresolved (18, 24, 25).

1.1.2 Archaeal transcription elongation

Promoter escape by initiation complexes and the transition into a transcription elongation complex (TEC) is typically signified by the exchange of the initiation factor TFE for the Spt4/5 complex composed of Spt4 and the universally conserved transcription factor Spt5 (NusG in Bacteria)(34, 35). Efficiency of transcription elongation is typically attributed to RNAP conformation, particularly the open/closed state of the RNAP clamp domain. Spt4/5 promotes active elongation by binding to both the stalk and clamp domains of RNAP, stimulating a more processive closed-clamp configuration of RNAP by enclosing the RNA/DNA hybrid and protecting the transcription bubble (Figure 1.1.ii)(22, 35–37)

Processivity of TECs can occasionally be interrupted during elongation by transcriptional roadblocks in the cell such as DNA damages and protein roadblocks. Archaeal histone-based chromatin takes the form of extended polymers on the genome with individual subunits that resemble the eukaryotic nucleosome (28). RNAPs stalled at transcriptional roadblocks (such as histone-based chromatin) will often ‘backtrack’ on the DNA template, causing the RNA 3’ end to move from the active site of RNAP into the secondary channel (35). Continued RNA synthesis is thus inhibited until the intrinsic cleavage activity of RNAP cleaves the RNA to produce a new 3’ end in the active site (22). In effect, the new RNA 3’ end gives the TEC another ‘chance’ to transcribe RNA through the transcriptional roadblock, which is often enough for continued elongation. TFS, an archaeal homolog of eukaryotic TFIIS, promotes RNAP-mediated cleavage of RNA transcripts by backtracked RNAPs, in effect aiding TECs through transcription

roadblocks and thus acting as a transcription elongation factor by temporarily shortening an RNA transcript (37).

Archaea are prokaryotic, and the processes of transcription and translation not only colocalize, but are functionally coupled (38). Actively elongating TECs transcribing between the start and stop codons of protein-coding genes are therefore trailed by a translating ribosome under most circumstances, which likely represents a significant portion of TECs given ~90% of the archaeal genome is protein-coding. Interplay between the translation and transcription apparatuses exists (42), but further investigation into any regulatory effects of such interplay during the elongation phase of transcription is warranted.

1.1.3 Archaeal transcription termination

While the mechanisms and factors surrounding archaeal transcription initiation and transcription elongation have long been studied, our understanding of the mechanisms and factors involved in transcription termination is less complete. The high stability of TECs and extremely small intergenic space in archaeal genomes necessitates efficient transcription termination to recycle RNAPs for new transcription events and prevent read-through transcription. Protein-mediated regulation of transcription termination has been known in Bacteria and Eukarya for some time, but is limited to only a handful of examples (43–47), and has only recently been described in Archaea through the identifications of transcription termination factors Eta and FttA (39, 40)(Figure 1.1.iii). This dissertation advances our understanding of the molecular mechanisms employed by Eta and FttA in protein-mediated archaeal transcription termination and investigates the cellular consequences of Eta and FttA-mediated transcription termination, and potential coupling of DNA repair pathways to transcription termination.

1.2 Archaeal transcription termination

RNA synthesis by RNAP is processive, requiring a single enzyme to transcribe the full length of a gene regardless of the length; some exceptionally long genes can take ~24 hours to transcribe at 20-40 nucleotides per second(41). The requirement for RNAP to remain resolutely associated with the DNA template and nascent transcript through multiple kilobases necessitates an extremely stable TEC that can transcribe through different sequences and protein-bound DNA templates. Despite this stability, cells must be able to halt RNA synthesis after transcription of a complete gene or operon and halt the activities of any RNAP that has initiated transcription aberrantly. Failure to terminate transcription of an upstream gene could allow regulation-independent expression of downstream genes or synthesis of untranslated or antisense transcripts with detrimental consequences; aberrant transcription is particularly problematic for the gene-dense chromosomes common to Bacteria and Archaea(42).

Multi-subunit RNAPs from each domain share a near identical core structure that envelopes an RNA:DNA hybrid within a tight-fitting pocket (8, 11, 13, 43, 44). High-resolution crystal and cryo-electron microscope structures and a wealth of biochemical data from many different RNAPs demonstrate that hydrogen bonding within the hybrid and contacts between the enclosed nucleic acids and RNAP provide stability to TECs. Despite similar TEC architecture, RNAPs from different domains, and each of the eukaryotic RNAPs, respond to different termination signals and factors, suggesting that several mechanisms of transcription termination are possible, or that a diverse set of factors and sequences use a common mechanism to disrupt the complex. Conserved elongation factors (i.e., NusG/Spt5 and NusA) modify RNAP activities and add an additional level of regulation to the elongation–termination decision(45–48). The mechanistic details of transcript release are best understood in Bacteria (49), although some features are shared in each domain.

Two general mechanisms of transcription termination have been characterized in both biochemical reactions with purified components and through genetic experiments (Figure 1.2).

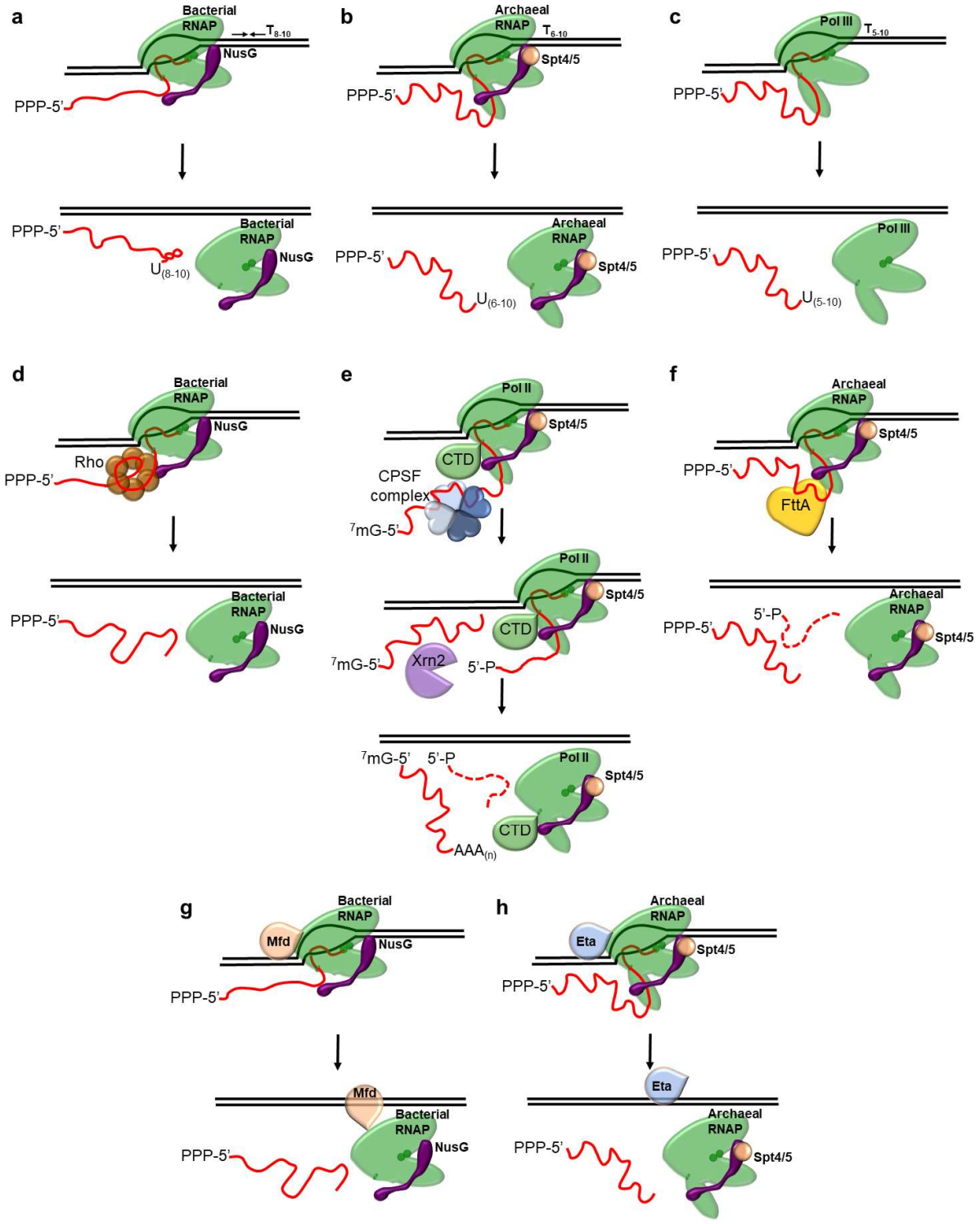

a
Bacterial RNAP
T8-10
NusG
PPP-5'
U(8-10)
b
Archaeal RNAP
T6-10
Spt4/5
U(6-10)
c
Pol III
T5-10
U(5-10)
d
Rho
e
Pol II
CTD
CPSF complex
7mG-5'
Xrn2
5'-P
AAA(n)
f
FttA
g
Mfd
h
Eta

Figure 1.2. Summary of transcription termination mechanisms commonly employed in Bacteria, Eukarya and Archaea. a. – c. Intrinsic transcription termination in Bacteria (a), Archaea (b), and for eukaryotic Pol III (c) results in release of the entire 5′-triphosphate-containing RNA transcript following transcription through a region of dyad-symmetry encoding an RNA hairpin immediately proceeded a T-rich non-template strand sequence (Bacteria) or T-rich non-template strand sequences (Archaea and eukaryotic Pol III). (d) Bacterial Rho proteins mediate transcription termination and release full-length RNA transcripts to solution. (e) CPSF- and Xrn2-mediated termination of eukaryotic Pol II complexes results first in cleavage of the 5′-methyl-G-capped 5′-transcript from the nascent RNA, and the resulting 3′-fragment is degraded by Xrn2 to mediate transcription termination by yet unknown mechanisms. (f) FttA can cleave the nascent transcript and terminate the archaeal transcription apparatus. (g) and (h) Both bacterial Mfd and archaeal Eta can disrupt stalled TECs and release full-length transcripts by rewinding the transcription bubble.

The first, termed ‘intrinsic termination’, requires only defined DNA sequences and the resultant RNA transcripts to disrupt the TEC. The second, factor-dependent termination, requires the activity of an additional protein(s) that interact with the TEC. Examples of each have been well defined in the three domains and point towards a set of universal principles for transcription termination.

1.3 Intrinsic Transcription Termination

Intrinsic terminators are utilized to control the activities of bacterial, archaeal and eukaryotic RNA polymerase III (Pol III). Bacterial intrinsic termination (Figure 1.2.a) sequences are perhaps best described and also the most complex, being composed of two essential elements: (1) a ~10–20 bp region of dyad symmetry in the DNA which is transcribed into an RNA sequence capable of forming an RNA–RNA duplex or RNA hairpin structure and (2) a T-rich sequence that directs the synthesis of a uridine-rich sequence in the RNA immediately following the RNA hairpin. Base pairing within the stem of the hairpin, the presence of the uridine-rich segment, and the spacing of these elements are critical to bacterial intrinsic termination efficiency; flanking sequences can also influence the efficacy of specific terminators (50). Intrinsic termination in bacterial species is typically quite precise, with transcript release occurring 8–10 bp downstream of the 3′ base of the hairpin, typically near the end of the uridine-rich sequence. Although the sequences directing intrinsic termination are well-described, the mechanics of RNA release at bacterial intrinsic-termination sequences likely differs for specific sequences (58, 60–64). TECs are weakened by sequence context alone; the hybrid at the normal position of termination is predominantly rU:dA, providing only the weakest hydrogen bonding of all possible combinations of nucleic acids. The uridine-rich sequence serves a second role, to pause RNAP at the position of termination, and any sufficiently uridine-rich sequence has been , shown to direct bacterial RNAP to pause. The paused, weakened TEC is now positioned immediately adjacent and downstream of the region capable of hairpin formation, and RNAP

interaction with the RNA hairpin results in transcript release. The role of hairpin formation and its effects on the TEC are debated, although most models predict that hairpin formation disrupts the upstream bases of the hybrid, triggering collapse of the TEC and RNA release. RNAP can be pushed forward by hairpin formation, and in the absence of continued synthesis, such movement unwinds the hybrid and eliminates RNAP-hybrid interactions necessary for complex stability (60, 63). Hairpin formation has also been shown to direct changes in RNAP structure 3,9. Allosteric models of transcription termination predict relatively large structural changes in RNAP that reduce the stability of the complex, hinder further synthesis, and allow dissociation of the complex (65).

Intrinsic transcription termination also occurs in Archaea and Eukarya (Figure 1.2.b & c), although the sequences that direct intrinsic termination differ from bacterial intrinsic termination. Archaeal RNAPs and eukaryotic RNA polymerase III (Pol III) do not require an RNA hairpin for intrinsic termination, although each enzyme does terminate transcription, *in vivo* and *in vitro*, in response to oligoT-rich template sequences (64, 66, 67). For these RNAPs the exact sequence of the oligo-T-rich region, its flanking sequences, and general sequence context dictate termination position and efficiency. Evidence suggests that oligo-T-rich sequences first evolved as termination sequences, and remain so for many RNAPs, with RNA hairpins adding specificity and precision to bacterial intrinsic termination. Eukaryotic RNAP polymerase II (Pol II) is notably resistant to bacterial intrinsic termination sequences and appears completely dependent on auxiliary proteins for transcription termination.

1.4 Factor-Dependent Termination

Genetic and biochemical evidence has demonstrated that termination of many transcription units relies on proteins that disrupt TECs. Many genes and operons lack easily identifiable intrinsic termination sequences and, in the case of genes transcribed by eukaryotic Pol II, rely entirely on termination by protein factors. The activity of these factors is also generally

employed for all transcription units, regardless of whether an intrinsic termination signal is also present; intrinsic termination sequences are rarely absolutely efficient, necessitating a backup mechanism to terminate TEC that escape even effective intrinsic termination sequences.

1.4.1 Bacterial Rho-mediated termination

Rho is an RNA-dependent ATPase encoded in most bacterial genomes, but orthologues are not obviously encoded in any archaeal or eukaryotic genomes. Rho is a surveillance factor that monitors the normal coupling of bacterial transcription and translation (58, 62, 68, 69). Rho accesses TECs by binding to ribosome-free RNA, wrapping ~60 nts of RNA around a ring-shaped hexameric structure (Figure 1.2.d). Rho provides a rapid response to the uncoupling of transcription and translation, typically halting transcription within seconds of uncoupling between the transcription and translation apparatuses. The mechanism(s) employed by Rho to terminate transcription are still debated, but the preponderance of evidence suggests that Rho translocates along the RNA and physically disrupts the TEC. Rho normally targets TECs that have reached the end of genes or operons, and thus are uncoupled from translation via ribosome release at the stop codon, but Rho can disrupt any TEC that is not immediately and consistently trailed by a ribosome, and is therefore responsible for polarity (section 1.5).

The major function of Rho may be to suppress futile transcription of host genes and limit expression of any introduced phage or foreign genes (55). Rho-dependent termination also is the natural mechanism of release at certain sites, likely serving as the sole mechanism to terminate transcription of many operons. Properties that favor Rho action are the absence of translation, a cytidine-rich transcript, and the absence of secondary structure in the nascent transcript (43). These characteristics are sufficiently common for Rho to engage any TEC that is not closely followed by a translating ribosome. Release of a nascent transcript is, in principle, possible at all promoter-distal locations, but release is normally only seen at sites that direct

RNAP to pause, likely allowing Rho to catch up to the TEC. Strong secondary structures in the RNA can impede Rho binding and activity.

The largely accepted mechanism of Rho activity invokes movement along the RNA until RNAP is reached, then pulling the RNA out of the TEC; alternatively, RNAP may be pushed forward, or allosterically modified to release the transcript. An alternative model suggests Rho may interact with RNAP throughout the transcription cycle *in vitro*, and thus be poised to halt transcription immediately in the absence of translation; stoichiometric details of this interaction are not known, nor has this proposed mechanism been evaluated *in vivo*.

1.4.2 Mfd and transcription-coupled DNA repair in Bacteria

Repair of highly transcribed DNA is more rapid than noncoding or poorly transcribed DNA and the template strand is repaired faster than the non-template strand of the same gene. DNA damage on the template strand stalls RNAP, and so a single lesion in an essential gene could have deleterious effects. RNAP stalled on a DNA lesion serves as a marker for DNA damage and recruits a factor (Mfd) that initiates RNAP removal and DNA repair. Mfd is a DNA-dependent translocase that consistently scans the genome, binding directly to stalled TECs and rewinds the DNA duplex at the upstream edge of the TEC, injecting torque that collapses the transcription bubble, moves RNAP forward, and ultimately disrupts the TEC (Figure 1.2.g) (70–75). If RNAP is stalled in the absence of damage or backtracked, Mfd activity can rescue the TEC to an active configuration. When continued elongation is blocked by a DNA lesion, the TEC is removed, and Mfd then recruits the UvrABC complex to excise the lesion and initiate DNA repair. Like Rho, Mfd can terminate transcription of almost any TEC.

1.4.3 Factor-dependent Termination of Eukaryotic RNA Polymerase II

Models of RNA Polymerase II (Pol II) transcription termination have been the subject of contentious debate for several decades(44, 76–82). At least two general mechanisms are

employed, but only polyA-dependent mechanisms of transcription termination are discussed here. Although the exact mechanics of transcription termination coupled to 3'-polyadenylation remain obscure, the factors promoting termination of Pol II are generally agreed upon. Broadly, the polyA-dependent eukaryotic transcription termination machinery is made up of a cleavage and polyadenylation specificity factor complex (CPSF) that first severs the nascent RNA, thereby generating a new 5'-monophosphate (uncapped) terminus that is recognized by the exonuclease Rat1/Xrn2 (yeast/human) which "torpedoes" or allosterically disrupts Pol II, terminating transcription (Figure 1.2.e). The polyadenylation signal (PAS; most commonly 5'-AAUAAA) is a uniquely eukaryotic termination signal generally located downstream of the most 3' exon (83, 84). After being transcribed, CPSF recognizes and binds the nascent polyadenylation signal and the metallo-beta lactamase CPSF-73 subunit of CPSF makes an endonucleolytic cleavage in the RNA downstream of the PAS (83–86). Cleavage by CPSF-73 results in two key biochemical contributions to termination. (1) Eukaryotic mRNAs are typically capped by a 7-methylguanosine to protect them from endogenous RNA degrading nucleases. CPSF-73 cleavage results in an uncapped nascent RNA – a suitable substrate for the 5'-3' exonuclease Xrn2 (Rat1 in yeast) (76, 80, 87). (2) CPSF provides a scaffold for additional molecular machinery that dephosphorylates the still actively elongating transcription complex, slowing transcription elongation rate. Ultimately, Xrn2 degradation of the nascent transcript reaches Pol II itself and interactions with a slowed yet still elongating PolI II all contribute to the termination of Pol II (88–90). Whole transcriptome, deep-sequencing analyses of CPSF- and Xrn2-depleted cells and biochemically reconstituted transcription systems support a shared role of these factors in transcription termination (76, 80, 87–92).

1.4.4 FttA-mediated Archaeal Termination

The absence of Rho or a Rho-like activity in eukaryotes is not surprising given the physical separation of transcription and translation. As such, eukaryotic transcription termination

employs different mechanisms (see below). Archaea, however, are prokaryotes wherein transcription and translation are coupled (93), suggesting a Rho-like mechanism could also be employed in Archaea. While decoupling of the transcription and translation apparatuses in archaea does lead to polar suppression of downstream gene expression(94), Archaea lack obvious Rho-homologues. Instead, all Archaea encode a eukaryotic-like termination factor that directs transcription termination and may also be responsible for polarity (section 1.4) (95).

FttA (Factor that terminates transcription in Archaea) is an essential, energy-independent transcription termination factor conserved throughout the archaeal domain. FttA is an ortholog of the CPSF-73 subunit of the cleavage and polyadenylation factor complex (CPSF). In vitro experiments demonstrated both FttA-mediated nascent RNA cleavage and transcription termination. FttA can independently mediate termination of stalled or slowly elongating RNAP and FttA-mediated termination becomes competitive with normal elongation rates when FttA is kinetically coupled to both the stalk domain of RNAP and the conserved elongation factor, Spt5 (see below for more details). FttA activity, unlike the CPSF complex (86, 96), does not require an upstream signaling sequence and *in vitro* experimentation suggests that FttA-mediated termination is likely analogous to Rho-mediated termination, which recognizes long stretches of unstructured, C-rich nascent RNA transcripts due to uncoupling of transcription and translation (43, 97, 98). Combined endonucleolytic cleavage and subsequent termination activity in a single protein factor demonstrates the molecular similarities of archaeal transcription termination to eukaryotic transcription termination supporting their shared evolutionary lineage.

In vivo analyses of FttA suggest it is responsible for proper 3’ end formation of transcripts and limits aberrant run-on transcription much like Rho and the Pol II termination machinery. FttA inhibition in archaeal cells results in significantly longer RNA 3’ termini, a phenotype that matches both Xrn2 and CPSF-depleted eukaryotic cells (76, 89, 90).

1.4.5 Eta-mediated Archaeal Termination

TECs do not always reach the end of a gene, and may encounter roadblocks to transcription which can have consequences on proper gene expression and DNA replication. Thus, cells must encode some factor capable of removing arrested TECs from the genome akin to Mfd in Bacteria. No conserved factor has been identified across the various taxonomical clades of Archaea (i.e. crenarchaea, euryarchaea, ASGARD, etc.) that terminates arrested TECs. Thus, it is likely that clade-specific transcription termination factors perform this important role. Eta, a superfamily 2 (SF2) helicase conserved with in the euryarchaeal clade of Archaea, was identified from archaeal lysates for its ability to disrupt TECs and evidence suggests it acts analogously to bacterial Mfd (99). Both factors require DNA upstream of TECs, are ATP dependent and are non-competitive with an actively elongating TEC indicating they are likely not responsible for global termination events. The established role of Mfd in bacterial transcription coupled DNA repair (TCR) (58, 71, 100), coupled with evidence for TCR in Archaea (101–103), implies a similar role for Eta in Archaea. A deletion of Eta results in a UV irradiation sensitivity phenotype consistent with impaired DNA damage responses (99).

1.5 Polar repression of transcription in prokaryotes

In most prokaryotes, the lack of a nucleus has resulted in the functional coupling of transcription and translation (93, 104). In Bacteria, the transcription and translation apparatuses normally become uncoupled following the termination of translation, leaving the isolated TEC vulnerable to the activities of the termination factor Rho. Rho targets uncoupled TECs by recognizing ribosome-free nascent transcripts and ultimately terminates transcription by disrupting the TEC. Rho-mediated termination, however, is not limited to instances where ribosomes were halted due to natural stop codons(3, 105). Prokaryotic genomes commonly contain operons, sets of multiple distinct genes driven by a single upstream promoter. In many operons, unique folding patterns within the RNA transcript can drive the uncoupling of transcription and translation,

selectively permitting Rho-mediated transcription termination to regulate gene expression. Genes downstream of the 'attenuation' event are therefore not expressed under certain conditions, due to a genetic element in an upstream gene; the influence of upstream sequences on the transcription of downstream sequences results in a biological concept known as polarity(94, 106).

1.6 Transcription coupled nucleotide excision repair

Many archaea thrive within niche and extreme environments which can increase rates of DNA damage. Many halophilic archaea, for example, thrive in shallow salt plains and endure extreme levels of UV radiation (107), while some hyperthermophilic species persist at temperatures that would easily denature purified DNA (108, 109), and yet, the presumed increased rates of deamination, depurination, and oxidation are somehow tolerated (110–112). In addition to growth in the extremes, many archaeal species maintain genomic stability levels to display similar rates of spontaneous mutation to mesophilic prokaryotes such as Escherichia coli (113–115). Perhaps surprisingly, no unique DNA repair pathways have been described in Archaea, nor extremophilic Bacteria.

Some DNA damages, i.e., UV-induced photoproducts, result in a distortion of the dsDNA helix which has stalling effects on critical processes such as replication and transcription. DNA repair mechanisms have evolved to detect the general distortions of the DNA backbone rather than the actual modification, which allows detection at a broad range of DNA lesions. Global genomic nucleotide excision repair (GG-NER) in Bacteria and Eukarya relies on enzymes to recognize the "bulky lesion" and direct strand-specific cuts on the damaged DNA strand (116, 117). The DNA damage, now between two nicks, is thus primed for "excision" from the DNA allowing resynthesis from the undamaged strand, and nick ligation to complete repair.

In Bacteria and Eukarya, NER can be initiated by recognition of TECs which stall upon DNA lesions entering the active site of RNA polymerase (RNAP) during transcription, a process termed transcription coupled DNA repair (TCR) (Figure 1.3). Utilizing actively transcribing RNAPs to sense DNA damage offers an evolutionary advantage as actively transcribed regions of the genome are actively monitored for lesions. Akin to global NER, TCR has yet to be described in Archaea but current evidence suggests it is an active pathway in some clades. While studies in crenarchaea have revealed no significant change in DNA repair of transcribed versus non-transcribed strands (118), euryarchaeal species have displayed preferential repair of transcribed DNA strands—a hallmark of TCR (101). Additionally, the archaeal RNAP—which closely resembles eukaryotic RNAPII—has been shown to stall specifically at template strand DNA damage [131]. In Bacteria, the transcription termination factor Mfd acts as the TRCF, simultaneously recruiting the Uvr family of NER enzymes and terminating transcription to prevent the formation of mutant transcripts (74, 118, 119). Euryarchaeal termination activity (Eta), an archaeal transcription termination factor, appears to act analogously to Mfd, and is intimately linked with other nucleic acid metabolic pathways. However, it is unknown whether Eta acts as an archaeal TRCF. Eta requires DNA sequences upstream of RNAP, aids backtracked RNAPs, is ATP-dependent, and is non-competitive with an actively elongating TEC. Deletion of both Mfd in Bacteria and Eta in Archaea produce a UV sensitivity phenotype, further suggesting Eta also has a role in DNA repair (98, 120). Species which encode Eta also encode eukaryotic XP NER enzymes which have yet to be implicated in an NER pathway. Without an obvious damage recognition NER enzyme encoded, it is attractive to think of damage stalled RNAP fulfilling this role. If Eta acts as an archaeal TRCF analogous to Mfd, but recruits eukaryotic-like NER enzymes, another intriguing example of an archaeal physiological pathway with both bacterial and eukaryotic-like elements would be presented and explicitly evidence TCR as a universally conserved DNA repair pathway for the first time.

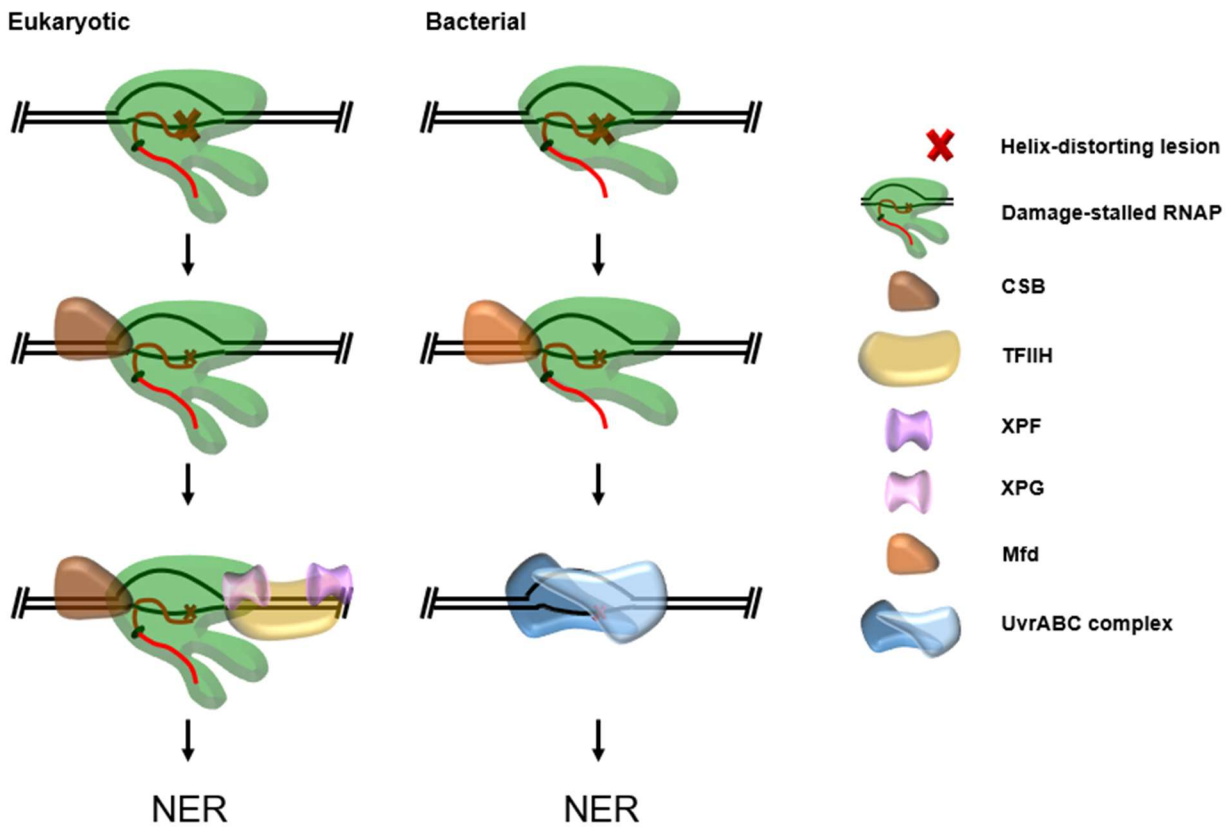


Figure 1.3. Transcription-coupled nucleotide excision repair in Eukarya and Bacteria. – c. Intrinsic transcription termination in Bacteria (a), Archaea (b), and for eukaryotic Pol III (c) results in release of the entire 5′-triphosphate-containing RNA transcript following transcription through a region of dyad-symmetry encoding an RNA hairpin immediately proceeded a T-rich non-template strand sequence (Bacteria) or T-rich non-template strand sequences (Archaea and eukaryotic Pol III). (d) Bacterial Rho proteins mediate transcription termination and release full-length RNA transcripts to solution. (e) CPSF- and Xrn2-mediated termination of eukaryotic Pol II complexes results first in cleavage of the 5′-methyl-G-capped 5′-transcript from the nascent RNA, and the resulting 3′-fragment is degraded by Xrn2 to mediate transcription termination by yet unknown mechanisms. (f) FttA can cleave the nascent transcript and terminate the archaeal transcription apparatus. (g) and (h) Both bacterial Mfd and archaeal Eta can disrupt stalled TECs and release full-length transcripts by rewinding the transcription bubble.

1.7 Thesis Rationale

Investigations thus far into transcription termination have revealed a handful of protein factors capable of disrupting the extremely stable TEC, and in many cases, their situational and substrate requirements. However, knowledge of the exact molecular mechanisms involved in factor-mediated transcription termination is still incomplete, and the factor responsible for polar repression of transcription in Archaea has yet to be identified. The work presented here investigates molecular mechanisms of two transcription termination factors recently identified in hyperthermophilic Archaea and uses novel experimentation to identify amino acid residues critical for the overall function of both enzymes. Investigations into the euryarchaeal-specific SF2 helicase Eta provide additional understanding of ubiquitously distributed SF2 helicases and the contested mechanism of transcription termination. The results of a novel genetic screen reveal conserved regions of FttA, an ancestral form of factors involved in mammalian transcription termination, as critical for proper transcription termination in Archaea. The results highlight evolutionarily conserved functions between the archaea and eukaryotes, as well as a shared susceptibility of the archaeal and bacterial RNA polymerases to transcription termination factors. Further, innovative strategies were developed to explore the potential coupling of DNA repair pathways to transcription termination. Evidencing transcription coupled DNA repair in the third Domain for the first time potentially opens up the Archaea as a unique model organism to study eukaryotic-like DNA repair.

REFERENCES

1. D. F. Browning, S. J. W. Busby, Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology* **14**, 638–650 (2016).

2. N. Sharma, Regulation of RNA polymerase II-mediated transcriptional elongation: Implications in human disease. *IUBMB Life* **68**, 709–716 (2016).

3. C. L. Turnbough, Regulation of Bacterial Gene Expression by Transcription Attenuation. *Microbiology and Molecular Biology Reviews* **83** (2019).

4. J. Guo, M. E. Turek, D. H. Price, Regulation of RNA polymerase II termination by phosphorylation of GDOWN1. *Journal of Biological Chemistry* **289**, 12657–12665 (2014).

5. K. Izumi, Disorders of Transcriptional Regulation: An Emerging Category of Multiple Malformation Syndromes. *Molecular Syndromology* **7**, 262–273 (2016).

6. J. Ma, *et al.*, Transcription factor regulation of RNA polymerase's torque generation capacity. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 2583–2588 (2019).

7. A. Agapov, D. Esyunina, D. Pupov, A. Kulbachinskiy, Regulation of transcription initiation by Gfh factors from Deinococcus radiodurans. *Biochemical Journal* **473**, 4493–4505 (2016).

8. J. Griesenbeck, H. Tschochner, D. Grohmann, Structure and function of RNA polymerases and the transcription machineries. *Sub-Cellular Biochemistry* **83**, 225–270 (2017).

9. P. P. Hein, R. Landick, The bridge helix coordinates movements of modules in RNA polymerawse. *BMC Biology* **8** (2010).

10. A. Mazumder, M. Lin, A. N. Kapanidis, R. H. Ebright, Closing and opening of the RNA polymerase trigger loop. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 15642–15649 (2020).

11. A. Hirata, B. J. Klein, K. S. Murakami, The X-ray crystal structure of RNA polymerase from Archaea. *Nature* **451**, 851–854 (2008).

12. D. Grohmann, F. Werner, Cycling through transcription with the RNA polymerase F/E (RPB4/7) complex: Structure, function and evolution of archaeal RNA polymerase. *Research in Microbiology* **162**, 10–18 (2011).

13. F. Werner, D. Grohmann, Evolution of multisubunit RNA polymerases in the three domains of life. *Nature Reviews Microbiology* **9**, 85–98 (2011).

14. B. J. Baker, *et al.*, Diversity, ecology and evolution of Archaea. *Nature Microbiology* **5**, 887–900 (2020).

15. K. Zaremba-Niedzwiedzka, *et al.*, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

16. L. Eme, A. Spang, J. Lombard, C. W. Stairs, T. J. G. Ettema, Archaea and the origin of eukaryotes. *Nature Reviews Microbiology* **15**, 711–723 (2017).

17. E. Peeters, N. Peixeiro, G. Sezonov, Cis-regulatory logic in archaeal transcription in *Biochemical Society Transactions*, (Portland Press, 2013), pp. 326–331.

18. O. Littlefield, Y. Korkhin, P. B. Sigler, The structural basis for the oriented assembly of a TBP/TFB/promoter complex. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 13668–13673 (1999).

19. B. R. Wenck, T. J. Santangelo, Archaeal transcription. *Transcription* **11**, 199–210 (2020).

20. S. D. Bell, P. L. Kosa, P. B. Sigler, S. P. Jackson, Orientation of the transcription preinitiation complex in Archaea. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 13662–13667 (1999).

21. S. H. Jun, *et al.*, Direct binding of TFEα opens DNA binding cleft of RNA polymerase. *Nature Communications* **11** (2020).

22. S. Schulz, *et al.*, TFE and Spt4/5 open and close the RNA polymerase clamp during the transcription cycle. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E1816–E1825 (2016).

23. N. Peng, Q. Xia, Z. Chen, Y. X. Liang, Q. She, An upstream activation element exerting differential transcriptional activation on an archaeal promoter. *Molecular Microbiology* **74**, 928–939 (2009).

24. A. Bleiholder, R. Frommherz, K. Teufel, F. Pfeifer, Expression of multiple tfb genes in diVerent Halobacterium salinarum strains and interaction of TFB with transcriptional activator GvpE. *Archives of Microbiology* **194**, 269–279 (2012).

25. F. Blombach, *et al.*, Archaeal TFEα/β is a hybrid of TFIIE and the RNA polymerase III subcomplex hRPC62/39. *eLife* **4** (2015).

26. T. J. Santangelo, L. Čuboňová, C. L. James, J. N. Reeve, TFB1 or TFB2 Is Sufficient for Thermococcus kodakaraensis Viability and for Basal Transcription in Vitro. *Journal of Molecular Biology* **367**, 344–357 (2007).

27. A. Feklístov, B. D. Sharon, S. A. Darst, C. A. Gross, Bacterial sigma factors: A historical, structural, and genomic perspective. *Annual Review of Microbiology* **68**, 357–376 (2014).

28. F. Mattiroli, *et al.*, Structure of histone-based chromatin in Archaea. *Science* **357**, 609–612 (2017).

29. S. P. Wilkinson, M. Ouhammouch, E. P. Geiduschek, Transcriptional activation in the context of repression mediated by archaeal histones. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 6777–6781 (2010).

30. D. Soares, *et al.*, Archaeal histone stability, DNA binding, and transcription inhibition above 90°C. *Extremophiles* **2**, 75–81 (1998).

31. T. J. Sanders, *et al.*, Extended Archaeal Histone-Based Chromatin Structure Regulates Global Gene Expression in Thermococcus kodakarensis. *Frontiers in Microbiology* **12**, 1071 (2021).

32. M. A. Pritchett, S. P. Wilkinson, E. P. Geiduschek, M. Ouhammouch, Hybrid Ptr2-like activators of archaeal transcription. *Molecular Microbiology* **74**, 582–593 (2009).

33. L. Lemmens, H. R. Maklad, I. Bervoets, E. Peeters, Transcription Regulators in Archaea: Homologies and Differences with Bacterial Regulators. *Journal of Molecular Biology* **431**, 4132–4146 (2019).

34. D. Grohmann, *et al.*, The Initiation Factor TFE and the Elongation Factor Spt4/5 Compete for the RNAP Clamp during Transcription Initiation and Elongation. *Molecular Cell* **43**, 263–274 (2011).

35. F. Werner, A nexus for gene expression-molecular mechanisms of Spt5 and NusG in the three domains of life. *Journal of Molecular Biology* **417**, 13–27 (2012).

36. F. W. Martinez-Rucobo, S. Sainsbury, A. C. M. Cheung, P. Cramer, Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. *EMBO Journal* **30**, 1302–1310 (2011).

37. G. A. Hartzog, T. Wada, H. Handa, F. Winston, Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in Saccharomyces cerevisiae. *Genes and Development* **12**, 357–369 (1998).

38. J. W. Shaevitz, E. A. Abbondanzieri, R. Landick, S. M. Block, Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. *Nature* **426**, 684–687 (2003).

39. P. Čabart, H. Jin, L. Li, C. D. Kaplan, Activation and reactivation of the RNA polymerase II trigger loop for intrinsic RNA cleavage and catalysis. *Transcription* **5** (2014).

40. T. J. Sanders, *et al.*, TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Molecular Microbiology* **111**, 784–797 (2019).

41. S. L. French, T. J. Santangelo, A. L. Beyer, J. N. Reeve, Transcription and translation are coupled in Archaea. *Molecular Biology and Evolution* **24**, 893–895 (2007).

42. M. Brenneis, O. Hering, C. Lange, J. Soppa, Experimental characterization of cis-acting elements important for translation and transcription in halophilic archaea. *PLoS Genetics* **3**, 2450–2467 (2007).

43. C. M. Hart, J. W. Roberts, Rho-dependent transcription termination. Characterization of the requirement for cytidine in the nascent transcript. *Journal of Biological Chemistry* **266**, 24140–24148 (1991).

44. W. Luo, A. W. Johnson, D. L. Bentley, The role of Rat1 in coupling mRNA 3′-end processing to transcription termination: Implications for a unified allosteric-torpedo model. *Genes and Development* **20**, 954–965 (2006).

45. J. Shi, *et al.*, Structural basis of Mfd-dependent transcription termination. *Nucleic Acids Research* **48**, 11762–11772 (2020).

46. Y. Jiang, M. Liu, C. A. Spencer, D. H. Price, Involvement of transcription termination factor 2 in mitotic repression of transcription elongation. *Molecular Cell* **14**, 375–386 (2004).

47. T. P. Newing, *et al.*, Molecular basis for RNA polymerase-dependent transcription complex recycling by the helicase-like motor protein HelD. *Nature Communications* **11**, 1–11 (2020).

48. J. E. Walker, O. Luyties, T. J. Santangelo, Factor-dependent archaeal transcription termination. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E6767–E6773 (2017).

49. T. J. Sanders, *et al.*, FttA is a CPSF73 homologue that terminates transcription in Archaea. *Nature Microbiology* **5**, 545–553 (2020).

50. Y. Xie, J. N. Reeve, Transcription by an archaeal RNA polymerase is slowed but not blocked by an archaeal nucleosome. *Journal of Bacteriology* **186**, 3492–3498 (2004).

51. M. Krzyszton, *et al.*, Defective XRN3-mediated transcription termination in Arabidopsis affects the expression of protein-coding genes. *Plant Journal* **93**, 1017–1031 (2018).

52. J. Y. Kang, *et al.*, Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell* **173**, 1650–1662 (2018).

53. Z. Hao, *et al.*, Pre-termination Transcription Complex: Structure and Function. *Molecular Cell* **81**, 281-292.e8 (2021).

54. J. Y. Kang, *et al.*, Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell* **173**, 1650–1662 (2018).

55. C. J. Cardinale, *et al.*, Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. coli. *Science* **320**, 935–938 (2008).

56. M. Strauß, *et al.*, Transcription is regulated by NusA:NusG interaction. *Nucleic Acids Research* **44**, 5971–5982 (2016).

57. C. M. Burns, L. v. Richardson, J. P. Richardson, Combinatorial effects of NusA and NusG on transcription elongation and Rho-dependent termination in Escherichia coli. *Journal of Molecular Biology* **278**, 307–316 (1998).

58. J. W. Roberts, Mechanisms of Bacterial Transcription Termination. *Journal of Molecular Biology* **431**, 4030–4039 (2019).

59. J. C. Mcdowell, J. W. Roberts, D. J. Jin, C. Gross, Determination of intrinsic transcription termination efficiency by RNA polymerase elongation rate. *Science* **266**, 822–825 (1994).

60. T. J. Santangelo, J. W. Roberts, Forward translocation is the natural pathway of RNA release at an intrinsic terminator. *Molecular Cell* **14**, 117–126 (2004).

61. I. Gusarov, E. Nudler, The mechanism of intrinsic transcription termination. *Molecular Cell* **3**, 495–504 (1999).

62. V. Epshtein, D. Dutta, J. Wade, E. Nudler, An allosteric mechanism of Rho-dependent transcription termination. *Nature* **463**, 245–249 (2010).

63. A. Ray-Soni, M. J. Bellecourt, R. Landick, Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annual Review of Biochemistry* **85**, 319–347 (2016).

64. T. J. Santangelo, L. Cubonová, K. M. Skinner, J. N. Reeve, Archaeal intrinsic transcription termination in vivo. *Journal of Bacteriology* **191**, 7102–7108 (2009).

65. V. Epshtein, C. J. Cardinale, A. E. Ruckenstein, S. Borukhov, E. Nudler, An Allosteric Path to Transcription Termination. *Molecular Cell* **28**, 991–1001 (2007).

66. S. Mishra, R. J. Maraia, RNA polymerase III subunits C37/53 modulate rU:dA hybrid 3′ end dynamics during transcription termination. *Nucleic Acids Research* **47**, 310–327 (2019).

67. T. J. Santangelo, J. N. Reeve, Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *Journal of Molecular Biology* **355**, 196–210 (2006).

68. M. S. Ciampi, Rho-dependent terminators and transcription termination. *Microbiology* **152**, 2515–2528 (2006).

69. S. Jain, R. Gupta, R. Sen, Rho-dependent transcription termination in bacteria recycles RNA polymerases stalled at DNA lesions. *Nature Communications* **10** (2019).

70. T. R. Strick, J. R. Portman, Transcription-Coupled Repair: From Cells to Single Molecules and Back Again. *Journal of Molecular Biology* **431**, 4093–4102 (2019).

71. C. P. Selby, Mfd Protein and Transcription–Repair Coupling in Escherichia coli. *Photochemistry and Photobiology* **93**, 280–295 (2017).

72. J. Y. Kang, *et al.*, Structural basis for transcription complex 1 disruption by the mfd translocase. *eLife* **10**, 1–86 (2021).

73. T. T. Le, *et al.*, Mfd Dynamically Regulates Transcription via a Release and Catch-Up Mechanism. *Cell* **172**, 344-357.e15 (2018).

74. O. Adebali, A. Sancar, C. P. Selby, Mfd translocase is necessary and sufficient for Transcription-coupled repair in Escherichia coli. *Journal of Biological Chemistry* **292**, 18386–18391 (2017).

75. O. Adebali, Y. Y. Chiou, J. Hu, A. Sancar, C. P. Selby, Genome-wide transcription-coupled repair in Escherichia coli is mediated by the Mfd translocase. *Proceedings of the*

National Academy of Sciences of the United States of America **114**, E2116–E2125 (2017).

76. J. D. Eaton, *et al.*, Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Genes and Development* **32**, 127–139 (2018).

77. J. D. Eaton, L. Francis, L. Davidson, S. West, A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. *Genes & development* **34**, 132–145 (2020).

78. S. McCracken, *et al.*, The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385**, 357–360 (1997).

79. S. Dengl, P. Cramer, Torpedo nuclease Rat1 is insufficient to terminate RNA polymerase II in Vitro. *Journal of Biological Chemistry* **284**, 21270–21279 (2009).

80. W. Luo, D. Bentley, A ribonucleolytic rat torpedoes RNA polymerase II. *Cell* **119**, 911–914 (2004).

81. T. Nojima, *et al.*, Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).

82. N. J. Proudfoot, Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352** (2016).

83. M. Clerici, M. Faini, L. M. Muckenfuss, R. Aebersold, M. Jinek, Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex. *Nature Structural and Molecular Biology* **25**, 135–138 (2018).

84. M. Clerici, M. Faini, R. Aebersold, M. Jinek, Structural insights into the assembly and polya signal recognition mechanism of the human CPSF complex. *eLife* **6** (2017).

85. N. G. Kolev, T. A. Yario, E. Benson, J. A. Steitz, Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3′-end maturation. *EMBO Reports* **9**, 1013–1018 (2008).

86. C. R. Mandel, *et al.*, Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature* **444**, 953–956 (2006).

87. S. West, N. Gromak, N. J. Proudfoot, Human 5′ → 3′ exonuclease Xm2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* **432**, 522–525 (2004).

88. N. Fong, *et al.*, Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Molecular Cell* **60**, 256–267 (2015).

89. J. D. Eaton, L. Francis, L. Davidson, S. West, A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. *Genes & development* **34**, 132–145 (2020).

90. M. A. Cortazar, *et al.*, Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a “Sitting Duck Torpedo” Mechanism. *Molecular Cell* **76**, 896-908.e4 (2019).

91. C. Baejen, *et al.*, Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes. *Molecular Cell* **66**, 38-49.e6 (2017).

92. J. Park, M. Kang, M. Kim, Unraveling the mechanistic features of RNA polymerase II termination by the 5’-3’ exoribonuclease Rat1. *Nucleic Acids Research* **43**, 2625–2637 (2015).

93. S. L. French, T. J. Santangelo, A. L. Beyer, J. N. Reeve, Transcription and translation are coupled in Archaea. *Molecular Biology and Evolution* **24**, 893–895 (2007).

94. T. J. Santangelo, *et al.*, Polarity in archaeal operon transcription in Thermococcus kodakaraensis. *Journal of Bacteriology* **190**, 2244–2248 (2008).

95. T. J. Sanders, *et al.*, FttA is a CPSF73 homologue that terminates transcription in Archaea. *Nature Microbiology* **5**, 545–553 (2020).

96. L. Schönemann, *et al.*, Reconstitution of CPSF active in polyadenylation: Recognition of the polyadenylation signal by WDR33. *Genes and Development* **28**, 2381–2393 (2014).

97. C. M. Hart, J. W. Roberts, Deletion analysis of the lambda tR1 termination region: Effect of sequences near the transcript release sites, and the minimum length of Rho-dependent transcripts. *Journal of Molecular Biology* **237**, 255–265 (1994).

98. M. Boudvillain, N. Figueroa-Bossi, L. Bossi, Terminator still moving forward: Expanding roles for Rho factor. *Current Opinion in Microbiology* **16**, 118–124 (2013).

99. J. E. Walker, O. Luyties, T. J. Santangelo, Factor-dependent archaeal transcription termination. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E6767–E6773 (2017).

100. J. S. Park, M. T. Marr, J. W. Roberts, E. coli transcription repair coupling factor (Mfd protein) rescues arrested complexes by promoting forward translocation. *Cell* **109**, 757–767 (2002).

101. N. Stantial, J. Dumpe, K. Pietrosimone, F. Baltazar, D. J. Crowley, Transcription-coupled repair of UV damage in the halophilic archaea. *DNA Repair* **41**, 63–68 (2016).

102. T. J. Sanders, C. J. Marshall, T. J. Santangelo, The Role of Archaeal Chromatin in Transcription. *Journal of Molecular Biology* **431**, 4103–4115 (2019).

103. A. M. Gehring, T. J. Santangelo, Archaeal RNA polymerase arrests transcription at DNA lesions. *Transcription* **8**, 288–296 (2017).

104. C. Wang, *et al.*, Structural basis of transcription-translation coupling. *Science* **369**, 1359–1365 (2020).

105. C. Yanofsky, Attenuation in the control of expression of bacterial operons. *Nature* **289**, 751–758 (1981).

106. Y. Li, S. Altman, Polarity effects in the lactose operon of Escherichia coli. *Journal of Molecular Biology* **339**, 31–39 (2004).

107. D. L. Jones, B. K. Baxter, DNA repair and photoprotection: Mechanisms of overcoming environmental ultraviolet radiation exposure in halophilic archaea. *Frontiers in Microbiology* **8** (2017).

108. H. Atomi, J. Reeve, Microbe profile: Thermococcus kodakarensis: The modehyperthermophilic archaeon. *Microbiology (United Kingdom)* **165**, 1166–1168 (2019).

109. C. Vieille, G. J. Zeikus, Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. *Microbiology and Molecular Biology Reviews* **65**, 1–43 (2001).

110. T. Lindahl, Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).

111. C. A. Lewis, J. Crayle, S. Zhou, R. Swanstrom, R. Wolfenden, Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 8194–8199 (2016).

112. M. Ehrlich, K. F. Norris, R. Y. Wang, K. C. Kuo, C. W. Gehrke, DNA cytosine methylation and heat-induced deamination. *Bioscience Reports* **6**, 387–393 (1986).

113. Y. Ishino, I. Narumi, DNA repair in hyperthermophilic and hyperradioresistant microorganisms. *Current Opinion in Microbiology* **25**, 103–112 (2015).

114. D. W. Grogan, G. T. Carver, J. W. Drake, Genetic fidelity under harsh conditions: Analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 7928–7933 (2001).

115. A. Palud, *et al.*, Intrinsic properties of the two replicative DNA polymerases of pyrococcus abyssi in replicating abasic sites: Possible role in DNA damage tolerance? *Molecular Microbiology* **70**, 746–761 (2008).

116. J. J. Truglio, D. L. Croteau, B. van Houten, C. Kisker, Prokaryotic nucleotide excision repair: The UvrABC system. *Chemical Reviews* **106**, 233–252 (2006).

117. M. G. Kemp, J. T. Reardon, L. A. Lindsey-Boltz, A. Sancar, Mechanism of release and fate of excised oligonucleotides during nucleotide excision repair. *Journal of Biological Chemistry* **287**, 22889–22899 (2012).

118. R. Dorazi, D. Götz, S. Munro, R. Bernander, M. F. White, Equal rates of repair of DNA photoproducts in transcribed and non-transcribed strands in Sulfolobus solfataricus. *Molecular Microbiology* **63**, 521–529 (2007).

119. A. M. Deaconescu, M. M. Suhanovsky, From Mfd to TRCF and Back Again—A Perspective on Bacterial Transcription-coupled Nucleotide Excision Repair. *Photochemistry and Photobiology* **93**, 268–279 (2017).

120. A. M. Deaconescu, A. Sevostyanova, I. Artsimovitch, N. Grigorieff, Nucleotide excision repair (NER) machinery recruitment by the transcription-repair coupling factor involves unmasking of a conserved intramolecular interface. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3353–3358 (2012).

121. B. J. Schalow, C. T. Courcelle, J. Courcelle, Mfd is required for rapid recovery of transcription following UV-induced DNA damage but not oxidative DNA damage in Escherichia coli. *Journal of Bacteriology* **194**, 2637–2645 (2012).

CHAPTER 2

THE STRUCTURE AND ACTIVITIES OF THE ARCHAEAL TRANSCRIPTION TERMINATION FACTOR ETA DETAIL VULNERABILITIES OF THE TRANSCRIPTION ELONGATION COMPLEX.

2.1 Introduction

2.1.1 Termination of transcription is subject to intricate control mechanisms

Factors that promote or inhibit efficient transcription initiation are abundantly encoded in most genomes, with many species harboring dozens to hundreds of site-specific DNA binding proteins that can influence the assembly and activities of basal transcription factors and RNA polymerase (RNAP) at promoter sequences (1–6). The abundance of transcription factors that control initiation does not preclude regulation throughout the remainder of the transcription cycle and an increasing number of factors have been demonstrated to influence post-initiation transcription that are often rate-limiting for gene expression (7–10). DNA bound proteins and nucleoid or chromatin structures typically hinder the progression of RNAP along the DNA template, slowing elongation and providing regulatory pauses that can be exploited to control the rate of RNA production (11, 12). Many eukaryotic factors can post-translationally modify RNAP or intimately associate with RNAP to control the rate of elongation and translocation (13, 14), but only a few factors can reduce the elongation rate to zero and disrupt the normally extremely-stable transcription elongation complex (TEC) (15–21).

The termination phase of the transcription cycle is subject to intricate control mechanisms that take advantage of vulnerabilities to the stability of the TEC. The bacterial and archaeal RNAP, as well as eukaryotic RNAP III (Pol III) and bacteriophage RNAP, are responsive to intrinsic termination sequences wherein DNA sequences encoding weak, rU:dA RNA:DNA hybrids, often in conjunction with hairpin RNA structures, can stall and disrupt TECs (22–25). While intrinsic termination sequences are often encoded downstream of genes and

operons, intrinsic termination signals embedded within the 5'-UTR and coding sequences of transcripts form the foundation of many riboswitches and attenuation mechanisms (26). While the absence or non-essential nature of known termination factors implies intrinsic termination mechanisms alone may suffice for some bacterial species, most Bacteria, all Archaea, and all Eukarya are dependent on protein factors that can stimulate transcription termination to ensure proper expression of the genome (27). Only a few domain-specific transcription factors have been identified that can disrupt TECs to release both RNAP and the RNA transcript from DNA, and with just one exception (e.g., FttA, also termed aCPSF1, and the eukaryotic CPSF73 (28)), there is no cross-domain conservation of any known transcription termination factor. This suggests that in each domain, unique proteins have evolved that likely target vulnerabilities of the TECs to tip the energetic balance in favor of TEC disassembly versus continued elongation. Given the known structures and conserved nature of the contacts that stabilize TECs in each domain, it is possible, if not likely, that the known termination factors may be reliant on similar mechanisms to disrupt TECs and control gene expression.

2.1.2 DNA vs. RNA dependent transcription termination

Debate remains regarding the exact mechanism(s) employed by transcription factors to disrupt TECs, but the essentiality of many transcription termination factors underlies the importance of termination factors to control gene expression. Transcription termination factors can be broadly split into two clades: 'RNA-dependent' versus 'DNA-dependent'. RNA-dependent termination factors (e.g., Rho in bacteria (29) and FttA in archaea (28) are typically associated with general governance of TEC activity, terminating rogue TECs as well as TECs that have transcribed to the end of a gene or operon. In many prokaryotic species, the uncoupling of transcription and translation provides an RNA binding site and ultimate access to the TEC (28, 30–32). In eukaryotes, RNA cleavage associated with polyadenylation signals provides access to an uncapped 5' RNA terminus that permits RNA degradation and TEC access (33).

Conversely, DNA-dependent termination factors (e.g., Mfd in bacteria (34), TTF2 in eukaryotes (35)), and Eta in archaea (16)) tend to have more specialized functions, typically acting to disrupt TECs regardless of position or coupling to the translation apparatus to recycle RNAPs irreversibly stalled due to DNA damage or to clear chromosomes prior to condensation or replication(21, 36).

Archaea encode a single multi-subunit RNAP that shares substantial structural similarities to eukaryotic RNAP II (Pol II) (37, 38). The archaeal RNAP is sensitive to intrinsic termination signals, and at least *in vitro*, the archaeal RNAP is also sensitive to bacterial rho-mediated termination (39); despite this susceptibility, rho is not encoded in any archaeal genome. Instead, archaeal species universally encode an essential transcription termination factor termed FttA (Factor that terminates transcription in Archaea; also termed aCPSF1 (28)) and most euryarchaeal species encode a second termination factor, Eta (Euryarchaeal termination activity (16)). FttA is an β-CASP, metallo-β-lactamase RNA-dependent termination factor likely responsible for global transcription termination events that provides the regulation normally afforded by rho in Bacteria and the poly-A dependent transcription termination common in eukaryotes. Eta, in contrast, is a DNA-dependent, superfamily 2 (SF2) helicase/translocase transcription termination factor. SF2 helicases are ubiquitously distributed across the domains with varying yet often indispensable roles in nucleic acid metabolism (40–42). Eta-mediated termination shares some similarities with bacterial Mfd-mediated termination, with both termination mechanisms being relatively slow and thus poorly suited to terminate rapidly transcribing TECs. Both Eta and Mfd are reliant on access to the TEC through DNA upstream to restart stalled or backtracked TECs or terminate TECs that cannot continue elongation due to damage in the template strand of DNA or nucleotide deprivation.

2.1.3 Study rationale and summary

To detail the mechanisms of SF2 helicase/translocase function and DNA-dependent transcription termination, we report the X-ray crystal structure of Eta (lacking the N-terminal domain (NTD)) and a structure-guided functional analysis of Eta-variants. The structure of Eta defines two highly conserved SF2 helicase/translocase domains flanked by a N-terminal domain that is dispensable for transcription termination, and a globular C-terminus harboring conserved residues that are critical to Eta function. Single amino acid changes in the C-terminus of Eta suffice to separately disrupt factor-dependent transcription termination, ATPase, and motor activities. The combined results provide the basis for a structural model of Eta-TEC interactions and a model of Eta-mediated transcription termination. Disrupting the TEC from the upstream edge of the transcription bubble in both intrinsic and factor-dependent termination suggest that this point of access takes advantage of a vulnerability to TEC stability by promoting forward translocation in the absence of continued RNA synthesis to undermine the strength of the TEC through successive disruptions to the RNA-DNA hybrid and collapse of the transcription bubble.

2.2 Results

2.2.1 Eta is a canonical superfamily 2 helicase/translocase

Eta from *Thermococcus kodakarensis* (Figure 2.1) is an 832 aa, ~96 kDa monomer with eleven readily identified motifs (Q, I, Ia, Ic, II, III, IV, V, Va, Vc) that define Eta as a member of the DEAD-box family of SF2 helicases/translocases (43). Walker A and Walker B motifs that promote ATP and Mg^{2+} binding are also defined, and primary sequence alignments of > 100 Eta-homologues, all from euryarchaeal species, reveal a weakly conserved N-terminal domain (NTD) harboring four almost universally conserved cysteines that likely coordinate a metal ion(s). Modeling of the presumptive metal-binding N-terminus (aa 1 - 192) predicts an extended alpha-helical structure that is dispensable for Eta-mediated transcription termination (16). Full-length Eta is insoluble at high concentrations; deletion of the NTD permitted sufficient

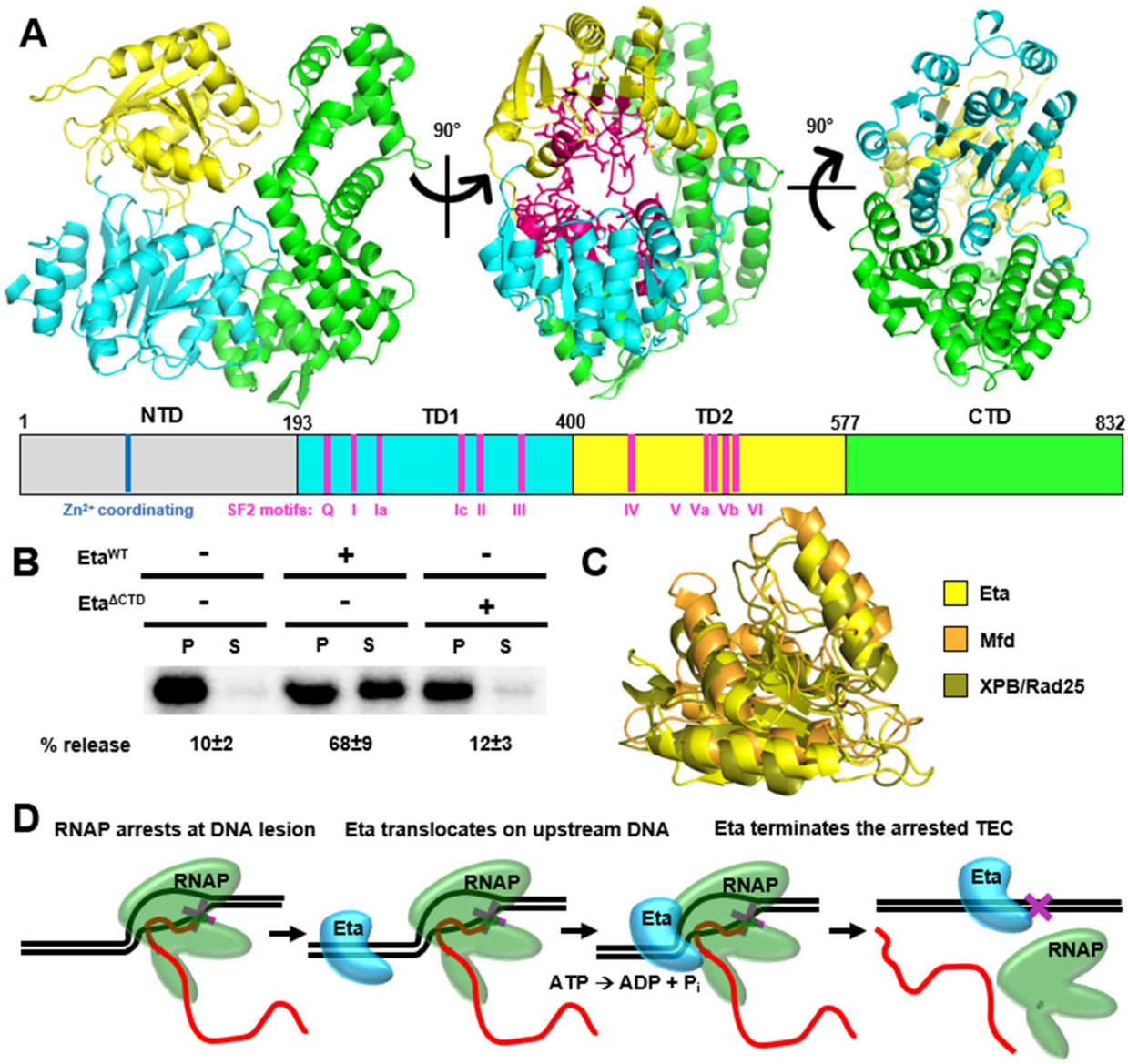


Figure 2.1: The crystal structure of Eta$^{\Delta NTD}$ and proposed mechanism of Eta-mediated transcription termination.
A) The X-ray crystal structure of Eta$^{\Delta NTD}$ (aa 194-832) reveals a three-domain architecture; two RecA-like translocase domains (aa 193-577, cyan & yellow), and a helically structured unique C-terminal domain (aa 578-832, green). 11 conserved SF2 helicase motifs identified in the primary sequence of Eta are highlighted in magenta. B) The unique C-terminal domain of Eta is indispensable for Eta-mediated transcription termination. P = pellet, S = supernatant. Percent transcript release is reported as mean +/- standard deviation of three replicates. C) Alignment of the TD2 translocase domain of Eta (yellow), the bacterial transcription termination factor Mfd (orange, PDB: 6XEO), and XPB/Rad25 (olive, PDB: 2FWR) reveal near identical structures. D) Eta-mediated termination necessitates ATP hydrolysis, efficient motor activity, and correct contacts with RNAP to destabilize stalled TECs. The magenta X represents a transcription roadblock, i.e., template-strand DNA damage.

concentration of preparations of the remainder of Eta (aa 193-804, termed Eta$^{\Delta NTD}$) to establish conditions suitable for crystallization. The X-ray crystal structure was solved by single wavelength anomalous dispersion (SAD) at 4.5 Å and then the resolution of the structure was improved to 4.1 Å with a native crystal. Four almost identical (0.179~0.219-Å root mean square deviations over 597 residues) Eta$^{\Delta NTD}$ molecules were resolved per asymmetric unit, and residues after 635 were modeled as poly-alanine due to electron density limitations. The Eta$^{\Delta NTD}$ structure reveals a mostly alpha-helical flattened-disc structure ~75 Å in diameter by ~40 Å deep defining three approximately equally sized primary domains: aa 193-400 comprise the first helicase/translocase domain (TD1, Figure 2.1.A, cyan), aa 401-577 define a second helicase/translocase domain (TD2, Figure 2.1.A, yellow), and aa 578 to the C-terminus form a completely alpha-helical C-terminal domain (CTD, Figure 2.1.A, green). The two helicase/translocase domains share a near identical fold with each other and the SF2 conserved sequences (Figure 1A, purple) form the bulk of the interface between TD1 and TD2. Dali-based structural alignments reveal that the twin translocase/helicase domains of Eta overlay nicely with the translocase/helicase domains of many other SF2 helicase/translocases, including the helicase domain of eukaryotic DNA polymerase theta(44), Rad25 (XPB)(45, 46), bacterial Mfd proteins (Figure 1C), and many RNA and DNA helicases. Dali alignments (47) of the Eta CTD reveal some structural conservation with the archaeal Hel308 (48)and the Ski2 RNA-helicase Brr2 (49). Contacts between TD1/TD2 and the CTD are minor but include a globular region of the CTD (aa 577-630) that packs against TD1, an extended loop (residues 295-310) from TD1 that contacts the CTD, and the C-terminus of an α-helix (aa 747-770) within the CTD that contacts TD2 around aa 473. Minimal contact points between TD1/TD2 and the CTD suggests potential rearrangements of the domains upon RNAP/TEC and/or DNA binding during the RNAP displacement process, as observed in other RNAP binding helicases such as HelD and Mfd (15, 50). The primary sequences alignments of Eta to other SF2 members predicts a Winged-Helix Domain (WHD) within the CTD (starting at aa 626) that was not well resolved in

the crystal structure, but whose characterization in other SF2 members was predicted to be important for proper activity (40).

2.2.2 Eta-mediated transcription termination requires the CTD of Eta.

Eta-mediated transcription termination is energy-dependent (dATP and ATP suffice equivalently) and requires access to DNA upstream of the TEC (16). These activities, and retention of other hallmarks of Mfd-mediated termination in Bacteria, including rescue of backtracked complexes and the ability to terminate slowly elongating or static TECs, suggested a model of Eta-mediated termination wherein Eta binds DNA upstream of the TEC, then translocates along dsDNA, collapsing the transcription bubble and forcing RNAP to translocate forward in the absence of continued RNA synthesis (Figure 2.1.E). Retention or deletion of the NTD did not impact the efficiency of Eta-mediated transcription termination, suggesting that the NTD may play a role in coordinating Eta-mediated transcription termination with other proteins and processes, perhaps linked to DNA repair (16, 51). The Eta-CTD, in contrast, is essential for Eta-mediated termination (Figure 2.1.D), hinting that the CTD may be the interaction surface with the archaeal RNAP during factor-dependent transcription termination. Deletion of the Eta CTD did not impact ATPase activities (Figure 2.S.2), further defining the Walker motifs and translocase/helicase activities to the twin TDs.

2.2.3 Conserved, solvent-exposed residues in the CTD of Eta are critical for motor, ATPase, and transcription termination activities.

The function of CTD was further investigated for its Eta-mediated transcription termination and helicase/translocase activities. The ~100 closest homologs of T. kodakarensis Eta were aligned to identify residues that were highly (> 90%) conserved within the CTD that were likely solvent exposed based on the $Eta^{\Delta NTD}$ crystal structure determined this study (Figure 2.S.1). 37 highly conserved residues were identified within the CTD, with one-third (Q588,

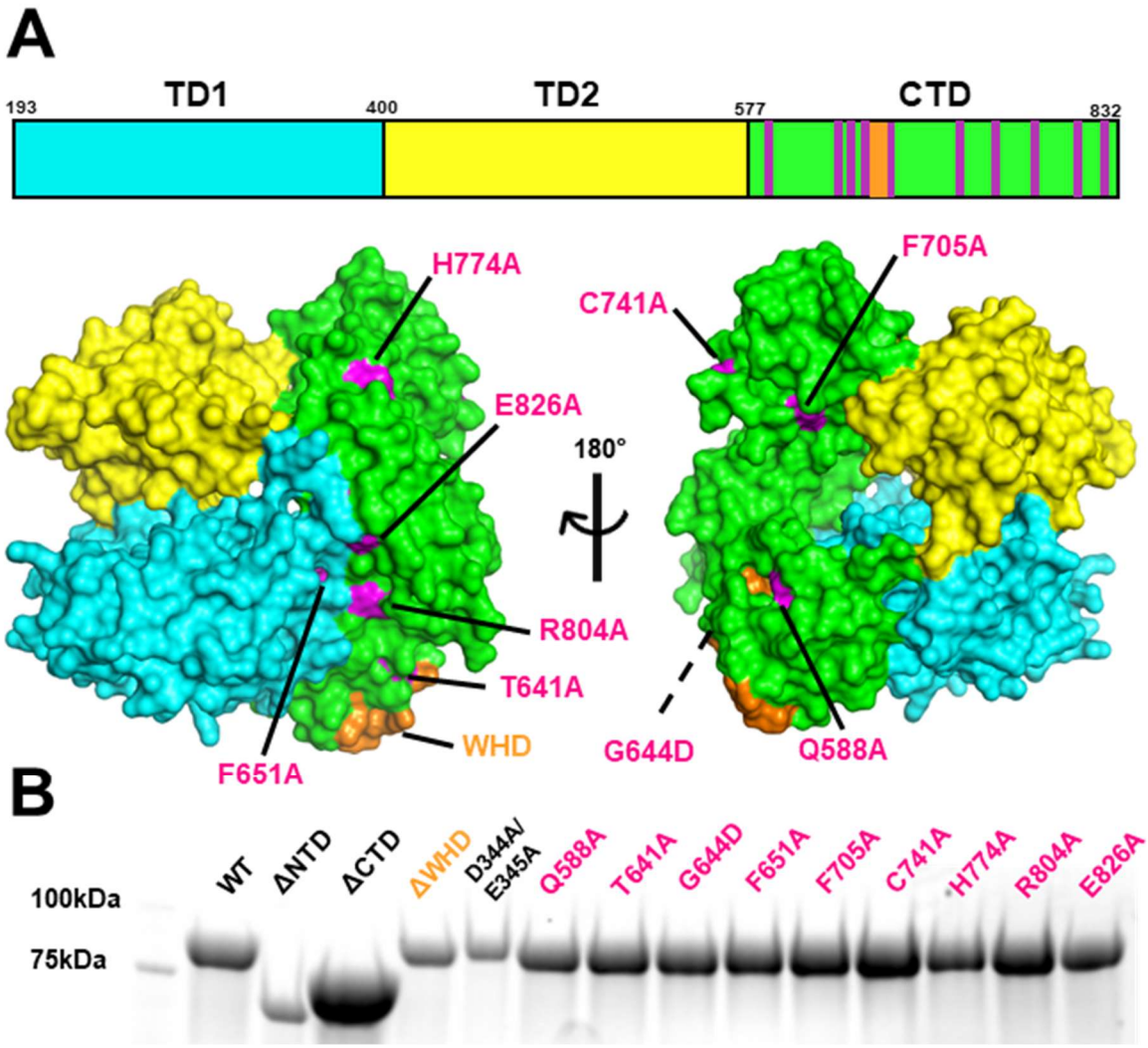


Figure 2.2: Conserved and likely solvent-exposed residues and sub-domains of Eta targeted for mutagenesis in this study.
A) Surface modeling of the EtaΔNTD crystal structure allows identification of conserved and likely solvent-exposed residues in the CTD of Eta (green). Identified residues (magenta), in addition to a putative Winged-Helix domain (orange), and the Walker B domain (D344, E345) were perturbed via site-directed mutagenesis, recombinantly expressed, and purified (B). Molecular weight standards are indicated to the left.

G644, T641, F651, F705, C741, H774, R804, and E826) likely solvent exposed (Figure 2.2.A). To determine the functional impacts of these conserved residues on Eta activity, we generated single amino acid variants of Eta wherein non-glycine residues were substituted with alanine, and G644 was substituted for an aspartic acid (Figure 2.2.B). We also generated a variant wherein the predicted WHD was deleted (ΔWHD) and purified each variant to homogeneity for use in *in vitro* helicase, ATPase and transcription termination assays. Each variant was generated in full-length Eta, remained stably folded at 85˚, and was recombinantly purified under identical procedures to Eta^{WT}.

Eta-mediated termination relies on DNA translocase activity to terminate stalled TECs (16). While the structure of Eta predicts translocase/helicase activities to be contained within the TDs, it remained plausible that specific changes within the CTD might impact helicase activities through domain crosstalk or local misfolding. To ensure our purified Eta variants retained translocase activity, the ability of Eta and Eta variants to displace streptavidin bound to an internally biotinylated and radiolabeled DNA template was investigated in both a single-stranded and double-stranded DNA context (Figure 2.3.A). In our ssDNA assay, Eta^{WT} was able to displace ~90% of streptavidin bound to a biotinylated ssDNA substrate across the 8-minute time course (Figure 2.3.B). To ensure that once streptavidin was displaced it would not simply re-bind the biotinylated DNA substrate, reactions were carried out in a vast excess of free biotin. Biotinylated ssDNA bound streptavidin displacement curves generated using Eta^{WT} and Eta variants were fit to a Michaelis-Menten-like curve (average $R^2 = 0.986$), and maximum reaction rates were determined for each Eta variant (Figure 2.3.D, grey bars). Most Eta variants (Eta^{Q588A}, Eta^{T641A}, Eta^{F705A}, Eta^{C741A}, Eta^{R804A}, Eta^{E826A}) retained wild-type like activity or had extremely minimal perturbations with normalized (WT=1.0) maximum reaction rates >0.8. Eta^{F651A} and Eta^{H774A} displayed modest disruptions to ssDNA translocase activity with normalized maximum reaction rates of 0.63 and 0.73, respectively. $Eta^{\Delta WHD}$, $Eta^{\Delta CTD}$ and Eta^{G644D} had severely hindered ssDNA translocase activity displaying maximum rates of

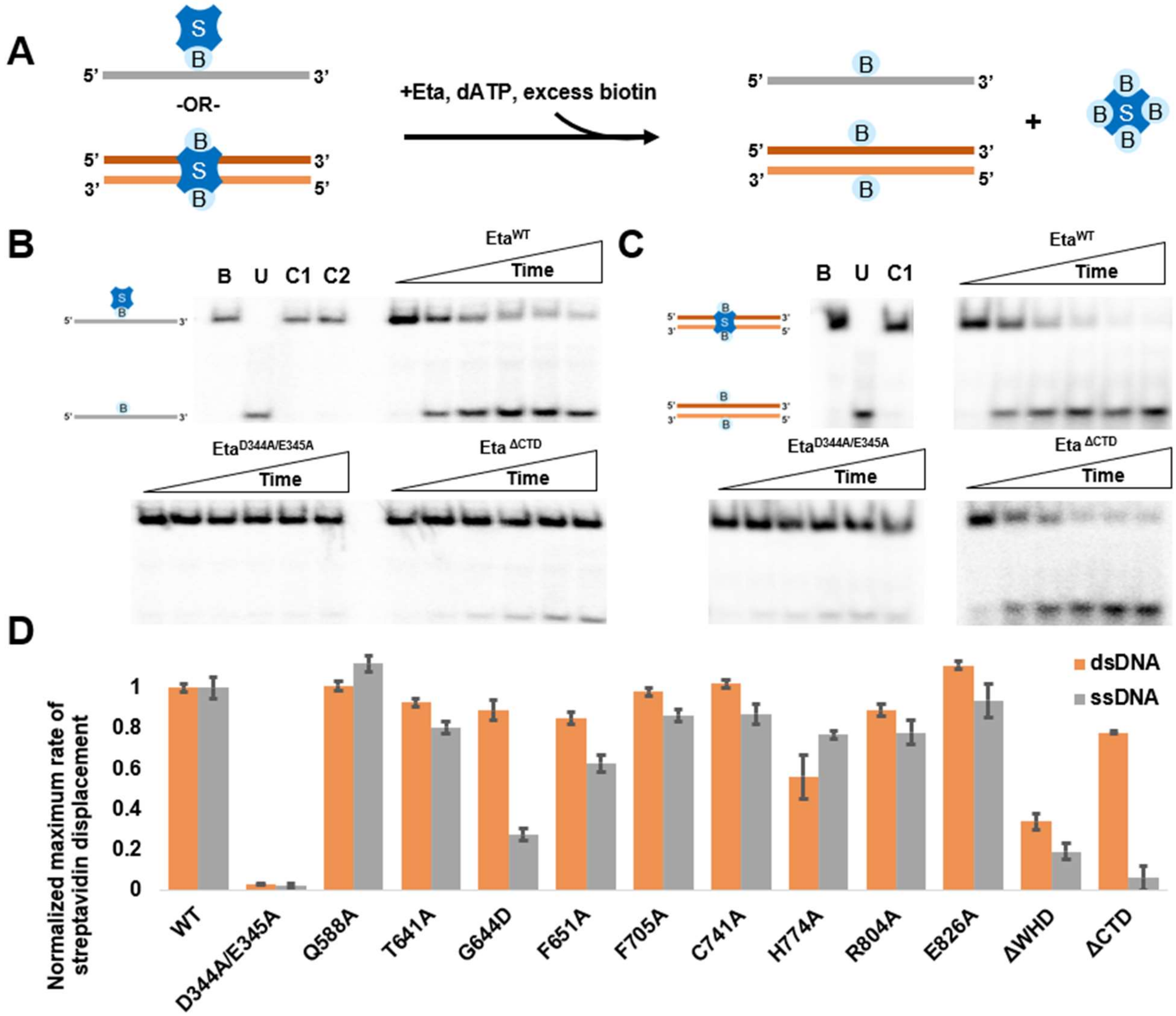


Figure 2.3: Varied effects of amino acid substitutions to the CTD of Eta on ssDNA/dsDNA translocase activity.
A) Single-stranded (grey) or double-stranded (orange initially) DNA is internally biotinylated and initially bound by streptavidin to produce the translocase substrate. ATP-dependent translocase activity displaces streptavidin from the DNA substrate where free biotin traps the streptavidin from rebinding the DNA. B+C) Populations of radiolabeled streptavidin bound (B) and unbound (U) ssDNA or dsDNA can be resolved by native PAGE. Reactions lacking dATP (C1) or Eta (C2) display no intrinsic release of streptavidin, but enzymatic release of streptavidin by Eta^{WT} and Eta variants can be tracked over time. D) Quantification of maximum reaction rates of Eta and Eta variant translocase activity on both ssDNA and dsDNA. Error bars represent one standard deviation from the mean (n=3).

streptavidin displacement less than a quarter of Eta^{WT}. The broad range of disruptions to ssDNA translocase activity of Eta CTD variants suggest currently unknown mechanisms of crosstalk between TD1 and TD2 of Eta with the CTD. However, it is likely that translocation on a dsDNA substrate is required for Eta-mediated transcription termination, and it is plausible dsDNA translocation is achieved through an alternative mechanism. Thus, we performed streptavidin displacement assays for each Eta variant on dsDNA substrates with one radiolabeled strand to examine streptavidin release via native PAGE (Figure 2.3.C). Maximum reaction rates were also generated after fitting streptavidin-displacement data to a Michaelis-Menten like curve (average R^2=0.988) and normalized to Eta^{WT} (Figure 2.3.D, orange bars). Most variants (Eta^{Q588A}, Eta^{T641A}, Eta^{F705A}, Eta^{C741A}, Eta^{H774A}, Eta^{R804A}, Eta^{E826A}, and $Eta^{\Delta WHD}$) had comparable translocase activity rates on dsDNA and ssDNA. Interestingly, Eta^{G644D} and $Eta^{\Delta CTD}$ displayed significantly improved translocase activity on the dsDNA substrate, indicating some regions of the CTD of Eta are instrumental in establishing the correct substrate selection for correct enzyme function. The observed dsDNA translocase activity of $Eta^{\Delta CTD}$, which in unable to collapse stalled TECs, reinforces the idea that translocase activity is required but not sufficient for Eta-mediated transcription termination. Overall, the ssDNA and dsDNA translocase assays together suggest it is likely that Eta CTD variants, with the exception of Eta^{H774A} and $Eta^{\Delta WHD}$, retain motor activities required during Eta-mediated termination.

Eta retains 3’ to 5’ helicase *in vitro* and it is plausible that helicase activity is required for Eta-mediated termination(52). Translocase assays alone do not assess strand-separation activity and thus, to ensure our purified Eta variants retained helicase activity, we tested the ability of each variant to unwind double stranded DNAs with a 3’ overhang (Figure 2.4). Eta^{WT} directed the complete, ATP-dependent (dATP was used to fall in line with transcription termination assays which use dATP to ensure no use of ATP by our stalled TECs) release of a small, radiolabeled oligonucleotide from dsDNA substrates (Figure 2.4.B) within just a few minutes. A DNA oligo completely complementary to the radiolabeled short oligonucleotide was

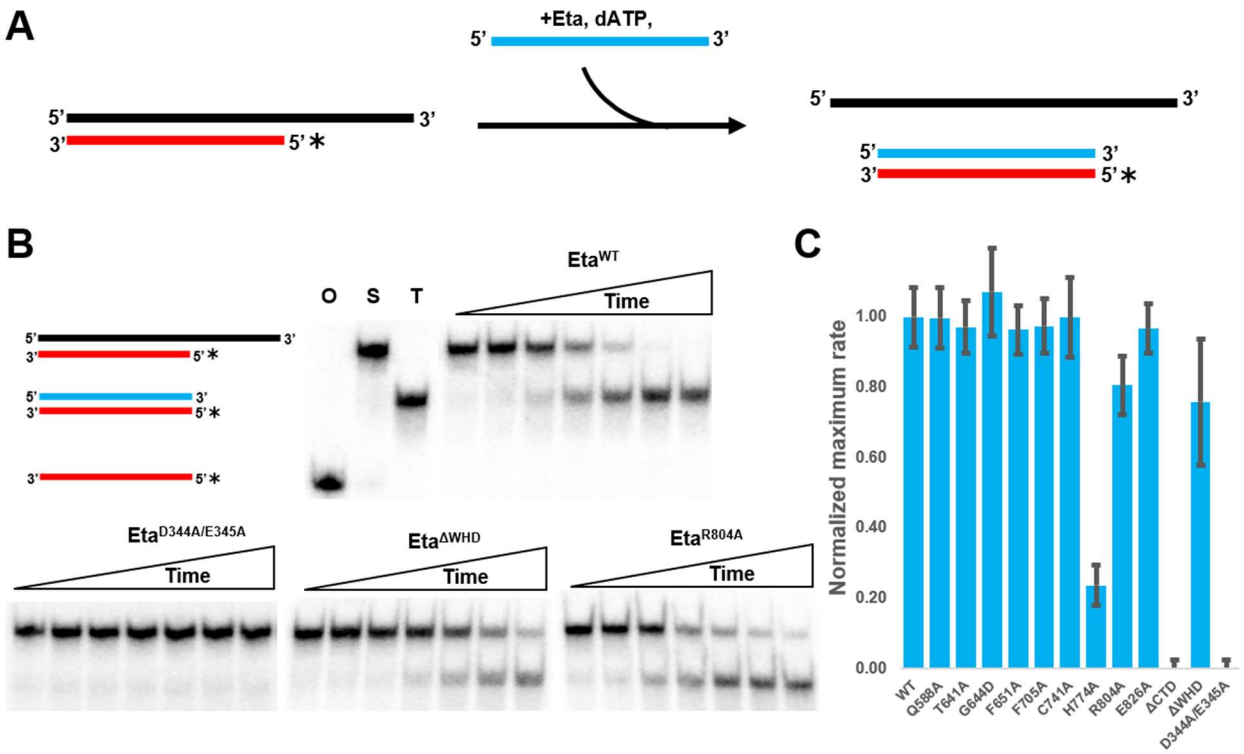


Figure 2.4: Most Eta variants retain wild-type like levels of motor activity.
A) A short radiolabeled DNA strand (red), initially paired to a longer complementary strand (black) generating a 3'-overhung dsDNA substrate, is separated by Eta in an ATP-dependent manner and anneals to the added complementary trap DNA (blue). B) The radiolabeled short oligo (O) can be tracked as it progresses from the substrate DNA (S) to the trap DNA (T) upon strand separation by Eta or Eta variants to monitor helicase activity efficiency. C) Quantification of maximum reaction rates of Eta and Eta variant helicase activity. Error bars represent one standard deviation from the mean (n=3).

added to ensure the radiolabeled strand did not simply re-anneal to reform the original substrate; Eta cannot unwind blunt ended dsDNA substrates (52). With few exceptions (Eta^{H774A}, Eta^{R804A}, and $Eta^{\Delta WHD}$), the purified Eta CTD variants (Figure 2.4.B,C) showed no deficiencies in helicase activities. As anticipated, disruption of the Walker B motif completely abrogates motor activity (Figure 2.4.C). Eta^{R804A} and $Eta^{\Delta WHD}$ display slower helicase activities, only unwinding ~ 60-80% of dsDNA substrates in the same time required for Eta^{WT} to completely unwind the substrates; the kinetic trajectory of Eta^{R804A} and $Eta^{\Delta WHD}$ helicase activities suggest these enzymes are simply slower. In contrast, the activity of Eta^{H774A} suggests this amino acid substitution either perturbs local folding dynamics congruent with proper helicase activity, or lacks the proper prerequisite ATPase activity. Importantly, most Eta mutants retained motor activities suggesting any defects in transcription termination activities are not due to immobility of the enzyme.

Retained helicase and translocase activities in the bulk of the Eta variants suggested that each variant also retained robust ATPase activities, however, we nonetheless wanted to test if amino acid substitutions introduced into our Eta variants altered ATPase activity. ATPase activity of Eta and Eta variants is dependent on a nucleic acid substrate, and disrupting the Walker B motif (D344A/E345A) abrogates ATPase activity, as expected (Figure 2.S.3). Most Eta CTD substitutions have no significant impact on ATPase activity. Despite slower helicase activity, Eta^{R804A} consistently displays better than Eta^{WT} ATPase activities, whereas the $Eta^{\Delta WHD}$ variant displays just ~60% of Eta^{WT} activity. Congruent with the weak helicase activity, Eta^{H774A} displays just ~25% of the Eta^{WT} ATPase activities. Further, while variations to some individual amino acids within the CTD of Eta had severe consequences for ATPase activity, $Eta^{\Delta CTD}$ displayed only a very slight perturbation to ATPase activity (~90% of Eta^{WT}). This is suggestive of a role of the CTD in coordinating TD1 and TD2 into correct conformations for DNA-dependent ATP hydrolysis.

Stalled TECs are easily formed via promoter-directed initiation and selective nucleotide deprivation on solid supports to provide an ideal substrate to monitor the rate and efficiency of Eta-mediated transcription termination (Figure 2.5.A)(16). The archaeal basal transcription factors TFB and TBP suffice to permit the archaeal RNAP to recognize BRE and TATA promoter elements, respectively, and initiate transcription with just three of the four NTPs (ATP, UTP and CTP); elongation on templates with a G-less cassette allows formation of promoter proximal stalled TECs. Magnetic separations collect intact TECs with 58 nucleotide radiolabeled transcripts (TECs+58) within a pellet (P) fraction, and in the absence of any protein additions, the extreme stability of TECs results in only minor amounts of nascent transcript release to solution (S) even after extended incubation at 85˚C. While addition of Eta^{WT} to stalled TECs results in ~80% efficient transcription termination (Figure 2.5.B), many Eta CTD variants completely lose the ability to terminate transcription or display reduced termination efficiencies (Figure 2.5.C,D). As anticipated, Eta variants with poor motor and ATPase activities (Eta^{H774A}, $Eta^{\Delta WHD}$) also display poor transcription termination activity. Substitutions at the far ends of the CTD (Eta^{Q588A}, Eta^{E826A}) did not affect transcription termination efficiency; in the case of Eta^{E826A} a modest increase in transcription termination activity is observed. Importantly, many substitutions that have no impact on motor or ATPase activities were identified that detrimentally affect transcription termination activity. Two variants, Eta^{G644D} and Eta^{R804A}, are of particular interest as both induce almost no transcription termination while retaining high levels of motor/ATPase activity. Eta^{T641A}, Eta^{F651A}, Eta^{F705A}, and Eta^{C741A} are also of interest as termination efficiencies decreased to just ~50-70% of Eta^{WT}.

Direct interactions between Eta and the TEC are anticipated based on our modeling and the ability of Eta to disrupt TECs, however, quantifying and comparing the stability of interactions between Eta or Eta-variants with TECs has not been experimentally possible

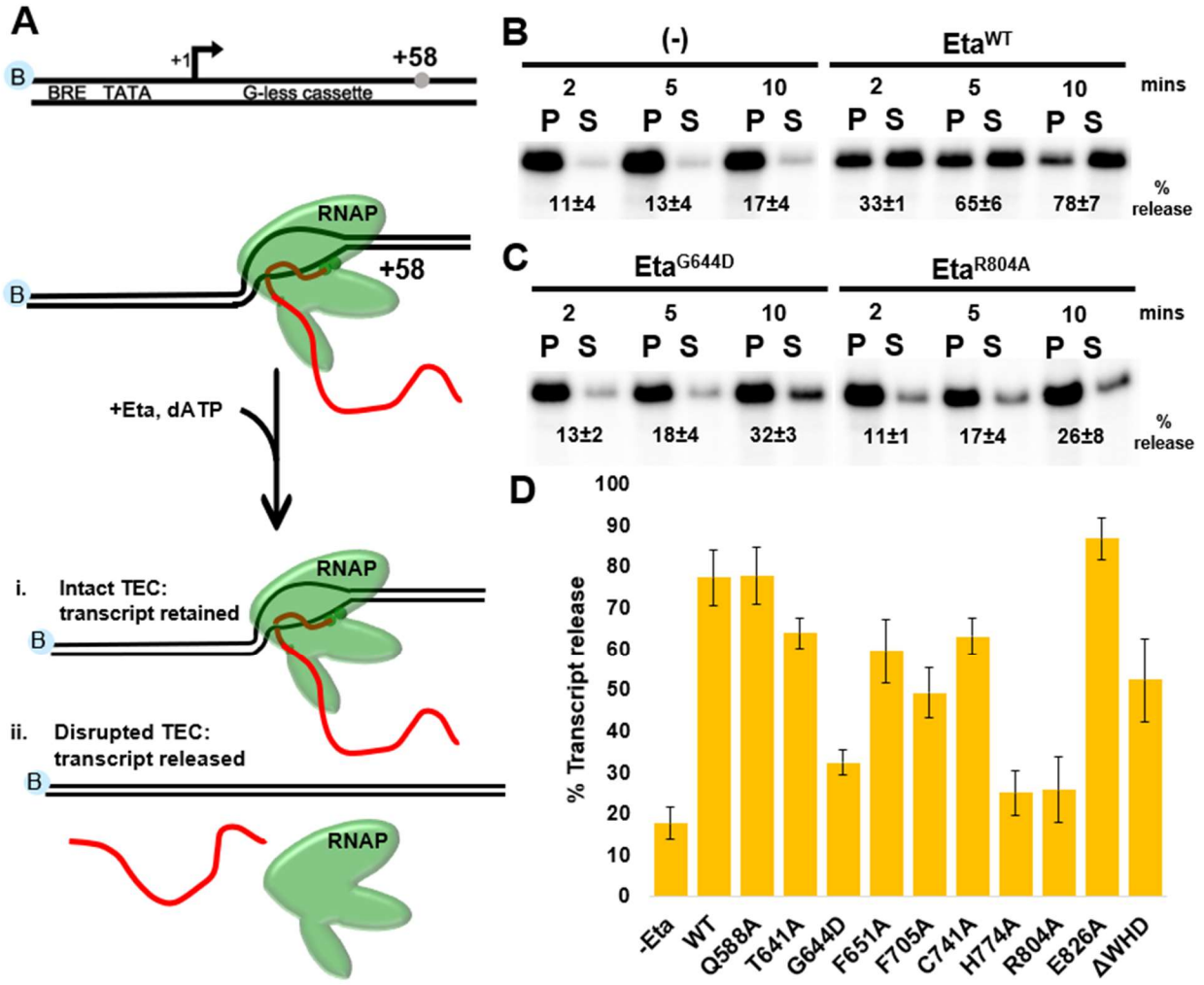


Figure 2.5: Multiple conserved C-terminal residues are involved in efficient Eta-mediated transcription termination

A) The *in vitro* transcription termination assay relies on nucleotide deprivation-based stalling of promoter-initiated TECs at +58 on a biotinylated dsDNA template. Release or retention of radiolabeled transcripts from TECs+58 allows quantification of Eta-mediated termination.
B) Representative transcription termination assays in the absence or presence of EtaWT. Percent transcript release is reported as mean +/- standard deviation of three replicates.
C) Representative transcription termination assays in the presence of EtaG644D and EtaR804A. Percent transcript release is reported as mean +/- standard deviation of three replicates. D) Quantification and comparison of transcription termination activities of Eta and Eta variants.

despite exhaustive attempts at pull-down or direct binding assays. Given these limitations, we used the more-readily detected movements and activities of TECs to monitor the impacts of select amino acid substitutions within Eta on putative Eta-TEC interactions (Figure 2.6). TECs prepared by nucleotide deprivation cannot accurately transcribe further along the DNA template than substrate availability permits, but TECs are free to undergo retrograde movement, often coincident with endonucleolytic cleavage of RNA phosphodiester linkages that shorten nascent transcripts. Retrograde movement, termed backtracking, is stimulated by extended incubations at physiological temperatures and is often coupled to cleavage of the transcript. When limited NTP subsets are available, a population of $TECs_{+58}$ is not static, but is most accurately described as a dynamic and interchanging population of forward translocated, catalytically-proficient $TECs_{+58}$, backtracked $TECs_{+58}$ with internal phoshodiester linkages occupying the active center, and $TECs_{+\sim 50\text{-}57}$ that could either be backtracked or catalytically-proficient but with shortened transcripts. Eta, like Mfd, is known to not only terminate TECs, but also influence the propensity of TECs to backtrack, with Eta and Mfd both capable of rescuing backtracked complexes to catalytically competent conformations.

We tested whether Eta-variants that retained ATPase, helicase, and translocase activities, and thus appeared to have no deficiencies in DNA interactions, but that lacked robust termination functions could influence TEC activities beyond termination. The capacity to alter the dynamics of a population of TECs is suggestive of retained engagements with TECs when DNA bound, whereas minimal changes to the dynamics of a population of stalled TECs is suggestive of impaired TEC interactions when Eta-variants are known to retain all tested ATP and DNA interactions like Eta^{WT}. When $TECs_{+58}$ are first generated, captured to a solid support, and washed to remove NTPs at reduced temperatures, nearly all (≥ 98%) of transcripts are retained within TECs and no obvious reduction of nascent RNA length occurs (Figure 2.6, lanes 1-4). When $TECs_{+58}$ are shifted to and incubated at 85˚C in the absence of any NTPs, most

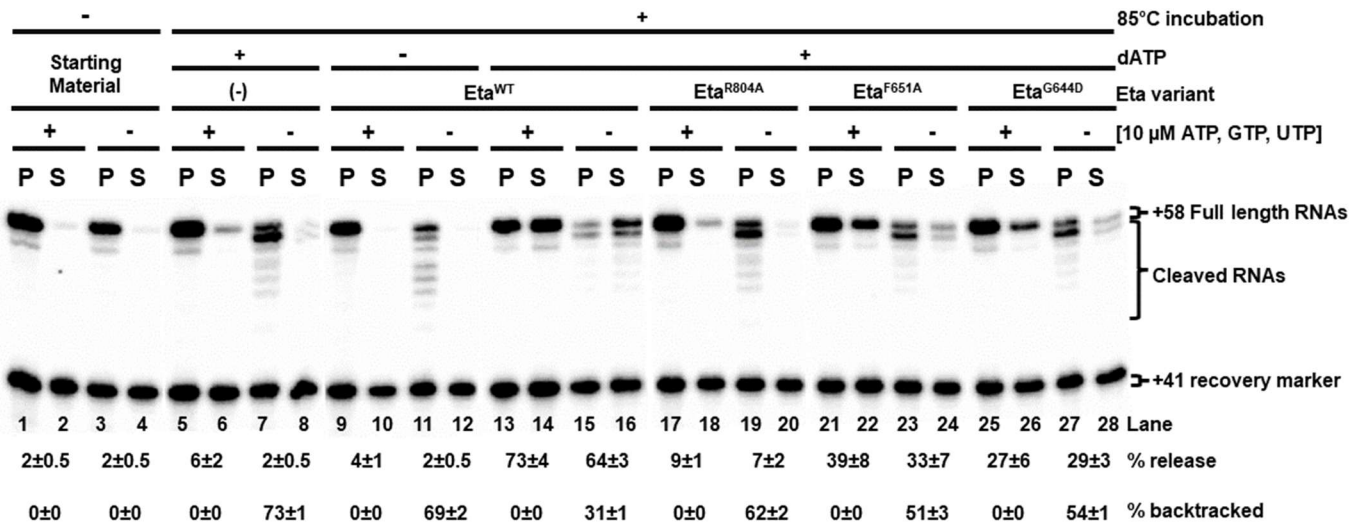


Figure 2.6: Eta variants with wild-type like motor activities but disruptions to transcription termination activity lack the correct contacts to rescue backtracked TECs. Stable $TECs_{+58}$ were assembled and resuspended in buffer lacking or including 10 μM ATP, CTP UTP (Lanes 1-4). Upon incubation at 85C, $TECs_{+58}$ in the presence of NTPs backtrack but can immediately resynthesize RNA, producing RNA products that resolve at +58 (lanes 5,6). $TECs_{+58}$ lacking NTPs backtrack but cannot resynthesized shortened RNA products, resulting in multiple RNA species less than 58 nt in length (lanes 7, 8). Incubation of Eta^{WT} without an energy source has no effect on the stability or backtracking ability of $TECs_{+58}$ (lanes 9-12). Addition of Eta^{WT} and dATP terminates TECs if arrested at +58 with no potential extension of RNA due to nucleotide deprivation. However, Eta^{WT} preferentially rescues and promotes nucleotide addition by backtracked TECs, allowing them to extend back to +58 before being terminated by Eta^{WT} (lanes 13-16). Rescue of backtracked TECs is also abrogated in Eta variants which retain motor activities but displayed abrogated transcription termination activity in Figure 5 (G644D, F651A, and R804A, lanes 17-28).

backtrack, and ~75% show evidence of transcript cleavage (Figure 2.6, lanes 7&8); the addition of even low concentrations of ATP, UTP, and CTP allow cycles of backtracking and resynthesis that maintain a much more uniform TEC_{+58} population (Figure 2.6, lanes 5&6). The addition of Eta^{WT} alone (Figure 2.6, lanes 11&12) does not influence population dynamics, whereas the addition of Eta^{WT} with an energy source exclusive to Eta (e.g. dATP) results in release of most transcripts to solution, and importantly, the number of TECs that backtrack is reduced from ~70% to just ~30% (Figure 2.6, lanes 13-16). Loss of $TECs_{\sim+50-57}$ and coincident gains in $TECs_{+58}$ implies that proper engagement of Eta^{WT} with TECs both stimulated catalytically-competent conformations of many TECs that permit extension of previously shortened transcripts and an overall reduction in backtracking that limits the potential to shorten transcripts. While Eta^{WT} clearly influences TEC activities beyond termination, Eta^{R804V}, Eta^{F651A}, and Eta^{G644D}, each of which retains strong ATPase, helicase, and dsDNA translocase activities, all fail not only to elicit efficient termination of TECs but also fail to significantly rescue backtracked complexes or influence TEC population dynamics (Figure 2.6, lanes 17-28). While Eta^{WT} rescues ~40% of backtracked complexes, Eta^{R804V}, Eta^{F651A}, and Eta^{G644D} rescue just 8%, 18%, and 15% of backtracked TECs. The most parsimonious explanation for the combined reductions in termination activity and loss of forward translocation of TECs is that although these Eta variants are likely fully capable of engaging DNA, these variants cannot correctly engage TECs, either by blocking direct Eta-RNAP interactions or altering Eta-dynamics such that engagement of TECs is impaired.

2.3 Conclusions and discussion

- **The overall function of superfamily II helicases is dependent on coordination of motor domains with other functional domains of the enzyme**
- **Single amino acid residues in the CTD of Eta are critical for correct coordination with the motor domains required for Eta-mediated transcription termination.**

- **Arginine at position 804 of Eta is required for transcription termination and likely contacts a susceptible region of arrested TECs that is shared between Archaea and Bacteria.**

Transcription elongation complexes are necessarily stable; many functional transcriptional units are thousands to millions of base pairs in length and must be processively transcribed. The requirement for regulated transcription termination ensures that TECs retain structural and energetic vulnerabilities that can be exploited by intrinsic- and factor-dependent termination mechanisms. Eta-mediated termination provides an ideal model for understanding TEC susceptibility and the activities of SF2 helicases. Determination of the X-ray crystal structure of $Eta^{\Delta NTD}$, combined with construction and analyses of more than a dozen Eta variants, ranging from large deletions to individual amino acid substitutions, defines the essentiality of the CTD for Eta-mediated transcription termination and highlights specific regions of Eta that play roles in ATPase, motor, and termination activities (Table 2.1).

Eta has been implicated in protein networks likely involved in DNA repair (16, 48), and translation and RNA metabolism (50, 51), thus it was anticipated that not all conserved residues in the CTD of Eta would be directly related to Eta-mediated transcription termination. The NTD of Eta is dispensable for Eta-mediated termination (16) and may also have roles in stimulating additional activities. Substitutions to the near-universally conserved residues Q588 and E826, located at the extremities of the Eta CTD, resulted in no significant changes to ATPase, motor, or transcription termination activities. In contrast, Eta-variants H774A and ΔWHD were severely compromised in all tested activities; given the thermostability of these variants, it is likely that they retained a folded structure, but that substitutions to the CTD impacted the coordination between the TDs and the CTD. H774 is close to the twin-translocase domains of Eta and could potentially be crucial in correct domain packing and maintaining an appropriate and functionally relevant tertiary structure. The Winged-Helix domain of Hel308, another well-studied archaeal

Table 2.1: Termination, helicase, ATPase, and translocase activities of Eta and Eta variants normalized to WT = 1.0.

Variant	Termination activity	Helicase activity	ATPase activity	Translocase activity (ssDNA)	Translocase activity (dsDNA)
WT	1.00X	1.00X	1.00X	1.00X	1.00X
Q588A	1.01X	1.00X	1.00X	1.12X	1.12X
T641A	0.78X	0.97X	1.15X	0.81X	0.93X
G644D	0.25X	1.07X	1.12X	0.27X	0.89X
F651A	0.70X	0.96X	1.01X	0.63X	0.85X
F705A	0.53X	0.97X	1.08X	0.87X	0.98X
C741A	0.76X	0.98X	1.04X	0.87X	1.02X
H774A	0.12X	0.24X	0.31X	0.77X	0.56X
R804A	0.14X	0.82X	1.53X	0.78X	0.90X
E826A	1.16X	0.97X	1.02X	0.94X	1.11X
ΔWHD	0.58X	0.76X	0.67X	0.19X	0.34X
ΔCTD	0.04X	0.01X	0.85X	0.06X	0.78X
	1.0 0.0	1.0 0.0	1.0 0.0	1.0 0.0	1.0 0.0

SF2 helicase, is required for proper DNA interactions (40). Given that DNA binding is essential for Eta-ATPase activity, the ~30% reduction in ATPase activity and ~80% reduction in translocase activities in the EtaΔWHD variant suggests the WHD contributes to coordinating ATP-dependent translocase activities and that energy release is poorly coupled to motor activity on DNA in the absence of the WHD. Several highly conserved residues in the CTD of Eta are specifically critical for Eta-mediated transcription termination, with little to no impact on ATPase, translocase, and helicase activity: substitution of T641, F651, and C741 significantly reduced termination activity, while substitution of G644, F651 and R804 nearly or completely abolished transcription termination activity. These single substitution variants behaved, in termination assays, similarly to a complete deletion of the CTD, once again highlighting the necessity of the CTD for Eta-mediated termination and adumbrating direct contact of the TEC and Eta-CTD.

The conservation of the TDs permits structural modeling based on other SF2 helicases with respect to the TEC. Modeling of $Eta^{\Delta NTD}$-TEC interactions using the cryo-EM structures of *T.kodakarensis* RNAP (57) and bacterial Mfd-TEC complexes (58) provides a plausible arrangement of Eta and the archaeal TEC (Figure 2.7). When the twin TDs of Eta and Mfd are rigidly superimposed, the C-terminus of Eta conflicts with the path of the upstream DNA resolved in the Mfd-TEC structure, implying either that the essential C-terminus of Eta is likely to rearrange upon TEC engagement, or the Eta-DNA interactions generate a new path for the DNA that does not require CTD rearrangements. When bound to upstream DNA in the same manner as Mfd, Eta is positioned such that the highly conserved, functionally critical CTD could make contact with the protrusion domain of RNAP, in many ways analogous to the primary point of contact between Mfd and the bacterial TEC (59). The weak interactions supporting Eta-TEC engagements have beguiled imaging a co-complex, but the Eta-variants that retain all activities except termination activity are also limited in altering backtracking (Figure 6), implying a defect in Eta-TEC interactions.

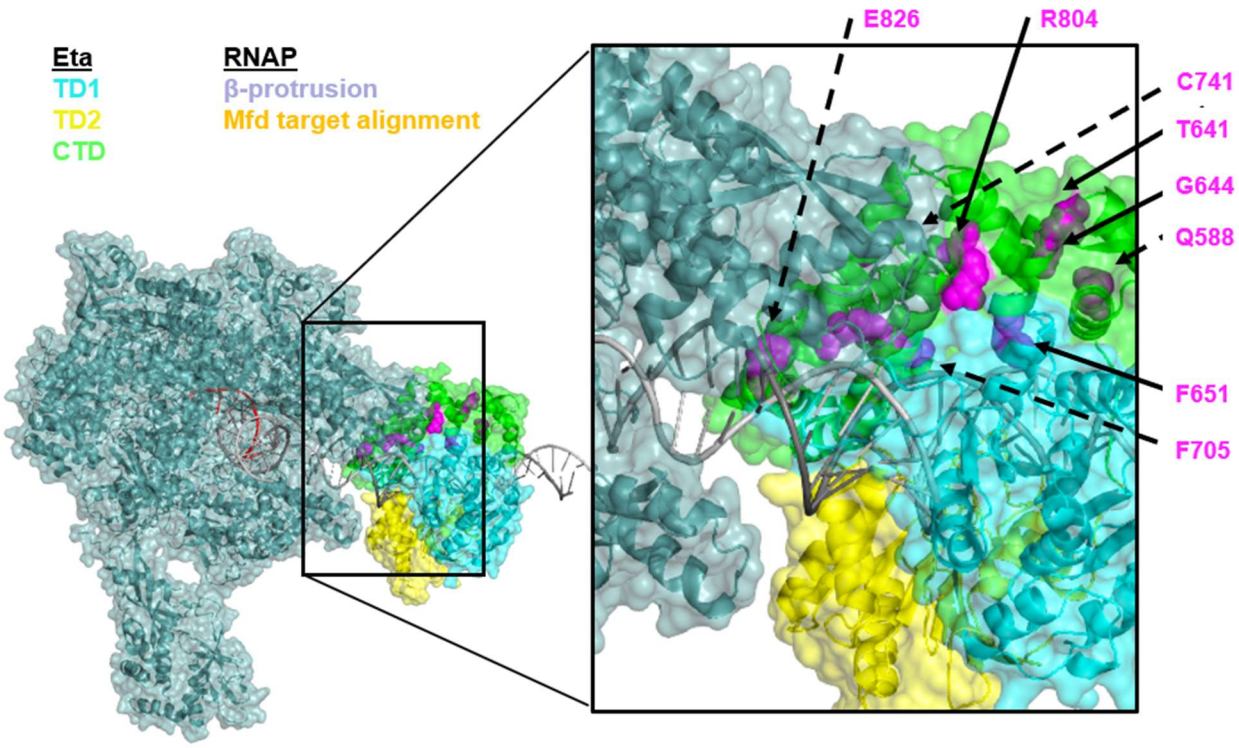


Figure 2.7: Modeling of Eta$^{\Delta NTD}$-TEC interactions using the cryo-EM structures of *T.kodakarensis* RNAP and bacterial Mfd-TEC complexes.
The high conservation of the core of bacterial and archaeal RNAPs, and that of the SF2 translocase domains of Mfd and Eta, allowed modeling of the Eta-TEC complex using the previously reported Mfd-TEC structure. Several termination-deficient CTD mutants (magenta) target amino acid residues in the plane of Eta closest to the beta-protrusion of RNAP (blue). Residues which when mutated in the bacterial RNAP lead to Mfd-mediated termination resistance are highlighted in orange and are plausible targets of Eta-mediated termination.

While termination-deficient specific variants (T641A, G644D, F651A, F705A, C741A, R804A), are dispersed in linear sequence, folding of the CTD positions these residues on a single surface where they could plausibly interact with the TEC (Figure 2.7). Unfortunately, R804 is not resolved in the Eta$^{\Delta NTD}$ crystal structure. The mode of TEC engagement with Mfd, and that modeled for Eta, suggest that both termination factors drive RNAP forward using their motor activities to initiate transcription termination. Coordinating Eta-mediated ATP or dATP hydrolysis to movement and ultimately dissociation of the TEC likely involves not only direct contacts between Eta and RNAP, but also intramolecular movements within Eta that are potentially hindered by substitutions that reduce, but do not eliminate Eta-mediated transcription termination.

In Bacteria, Mfd acts as the transcription repair coupling factor (TRCF) and terminates TECs stalled at bulky DNA lesions while concurrently recruiting bacterial nucleotide excision repair (NER) enzymes to the site of damage. NER or the transcription-coupled repair (TCR) sub-pathway have not been explicitly demonstrated in Archaea, although tantalizing evidence implies NER may be active in diverse archaeal clades (51, 60–63). Eta-mediated transcription termination resembles Mfd-mediated termination, and deletion of both factors from their respective organisms results in sensitivity of cells to UV-induced DNA damage (16, 64) . However, continued efforts will be necessary to detail whether Eta-mediated transcription termination is linked to DNA repair and to identify the factors that may direct archaeal NER and TCR.

2.4 Methods

X-ray crystal structure determination and refinement of Eta$^{\Delta NTD}$

Eta$^{\Delta NTD}$ crystals were prepared by hanging drop vapor diffusion at 22°C by mixing equal volumes of Eta$^{\Delta NTD}$ (~8 mg/ml) and crystallization solution (0.1 M Tris-HCl (pH 8.0), 4.5 M NaCl,

and ~3% ethylene glycol). Crystals were plunge-frozen in liquid nitrogen. To obtain experimental phase information, native crystals were soaked in crystallization solution containing 0.1 mM hexatantalum bromide cluster (Ta_6Br_{12}) for 4 hours before flash freezing in liquid nitrogen. Diffraction data were collected at the F1 beamline of the Cornell University High Energy Synchrotron Source (CHESS, Ithaca, NY) at 100 K and data were processed by HKL2000 (59). Both native and Ta_6Br_{12} containing crystals belonged to space group P32 and contained four molecules per asymmetric unit. With both isomorphous and anomalous signals from native and Ta_6Br_{12} datasets, 11 Ta_6Br_{12} sites in the asymmetric unit were located and the experimental phase (figure of merit: 0.480) was calculated using Automated structure solution (AutoSol) in PHENIX (60). Density modification by Automated model building (AutoBuild) in PHENIX yielded a map, which could be used for manual model building using Coot (61), and the structure was refined using PHENIX. The final model contains residues from 194 to the C-terminus; residues after aa 635 were modeled as poly-alanine due to relatively poor electron density of this region. Final coordinates and structure factors have been deposited to the Protein Data Bank (PDB) with the accession codes listed in the supplementary data.

Modeling the Tko RNAP elongation complex with Eta$^{\Delta NTD}$

We modeled the T. kodakarensis RNAP elongation complex with Eta$^{\Delta NTD}$ using the cryo-EM structure of T. kodakarensis RNAP (PDB: 6KF3) (52), the cryo-EM structure of the E. coli RNAP elongation complex (EC) with Mfd (PDB: 6X2N) (15) and the crystal structure of Eta$^{\Delta NTD}$ determined in this study. T. kodakarensis and E. coli RNAPs were superimposed using their catalytic domains including double-psi beta barrel (DPBB) domains in their largest (RpoA1/β') and second largest subunits (RpoB/β). Eta$^{\Delta NTD}$ was aligned with Mfd using the translocase domains (TD1 and TD2).

Multiple sequence alignment and identification of conserved amino acids within the C-terminal domain of Eta

The amino acid sequence for TK0566, encoding Eta in T. kodakarensis, was queried using the blastp suite and aligned using COBALT (62) versus the 100 top matches from the blastp query (https://www.ncbi.nlm.nih.gov/). Individual residues in the CTD (aa 577-832) were scored for conservation based on retention frequency. Residues conserved in >90% of sequences were deemed conserved. Surface modeling of the Eta CTD in the crystal structure identified conserved residues likely to be solvent exposed for mutagenesis.

Site-directed mutagenesis

Site-directed mutagenesis was performed on plasmid pTS481(16) with the QuickChange II XL kit (Agilent Technologies). Conserved and like solvent exposed residues were replaced with an amino acid predicted to disrupt the normal function of the residue (Q588A, T641A, G644D, F651A, F705A, C741A, H774A, R804A, E826A). A small deletion to the putative winged-helix domain (aa 630-645) was generated in the same manner.

Protein purifications

RNAP (RpoL-HA-6xHis) was purified from T. kodakarensis strain TS413 as previously described (63). Eta^{WT}, $Eta^{\Delta NTD}$, $Eta^{D344A/E345A}$, $Eta^{\Delta CTD}$, $Eta^{\Delta WHD}$, Eta^{Q588A}, Eta^{T641A}, Eta^{G644D}, Eta^{F651A}, Eta^{F705A}, Eta^{C741A}, Eta^{H774A}, Eta^{R804A}, and Eta^{E826A} were expressed and purified from Rosetta 2 (DE3) cells cultured in LB supplemented with 3% D-sorbitol, 100 µg/mL ampicillin, and 34 µg/mL chloramphenicol. Expression was induced by addition of 0.2 mM isopropyl β-D-1-thiogalactopyranoside, and cultures were grown overnight at 22°C. Biomass was pelleted and initially lysed via sonication in 15 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 100 mM NaCl, 2 mM β-

mercaptoethanol. Material was centrifuged at 22,000 x g for 15 mins at 4°C to produce a clarified lysate and debris pellet. The 'debris' pellet was lysed again in 15 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 500 mM NaCl, 2 mM β-mercaptoethanol and re-clarified as described above. The supernatant from this re-clarification in higher salt contained Eta, and was heat treated at 85°C for 30 mins. The high-salt heat-treated lysate was again clarified by centrifugation at 22,000 x g for 15 mins at 4°C, and the resulting supernatant was step-dialyzed into 15 mM Tris-HCl pH 8.0, 10mM $MgCl_2$, 300 mM NaCl, 2 mM β-mercaptoethanol and loaded onto a 5mL HiTrap Heparin column (Cytiva) using an AKTA Pure FPLC system (GE Healthcare). Proteins were eluted over a 60mL gradient to 15 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 1 M NaCl, 2 mM β-mercaptoethanol. Fractions containing Eta were identified by SDS-PAGE, pooled, and dialyzed into storage buffer, 15 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 250 mM KCl, 2 mM β-mercaptoethanol, 50% glycerol and concentrated by column centrifugation (Vivaspin 50kDa MWCO). Resulting enzymes was quantified using a Qubit™ Protein Assay Kit (Invitrogen).

Translocase (streptavidin displacement) assays

CTGGCTGTGGCGTGTTTCTGGTGGTTCCTAGGTCTTAGCCGTCTACGCCTCACT (CM0082-IB; 5'- CTGGCTGTGGCGTGTTTCTGGTGG[bio-dT]ACCTAGGTCTTAGCCGTCTACGC CTCACT -3') was radiolabeled with T4 Polynucleotide Kinase (New England Biolabs) in the presence of [γP32]-ATP (Perkin Elmer). For dsDNA assays, radiolabeled CM0082-IB was incubated in 5-fold molar excess CM0082-RC-IB (5'-AGTGAGGCGTAGACGGCTAAGACCTA GG[bio-dT]ACCACCAGAAACACGCCACAGCCAG-3', heated to 95 °C for 2 mins, and slowly cooled to room temperature to anneal strands. Single stranded radiolabeled CM0082-IB or radiolabeled double stranded CM0082-IB/CM0082-RC-IB was bound to 40-fold molar excess streptavidin. Translocase reactions were assembled with final concentrations of 5 nM ssDNA or

dsDNA substrate and 15 nM Eta in 20 mM Tris HCl pH 8.0, 10 mM MgCl2, 10 µg/mL BSA, 1 mM DTT. Reactions were heated to 50°C and started by addition of dATP and excess biotin (400 µM) to trap streptavidin displacement events. At the appropriate time points, 10 µL aliquots were removed and stopped on ice with the addition of 3 µL 1.5% SDS, 50% glycerol. Terminated reactions were separated on a 15% native polyacrylamide gel, and radiolabeled products were detected by exposure of the gel to a phosphorimaging screen (GE Healthcare) and analyzed using GE Imagequant 5.2.

Helicase assays

The short DNA strand of the 3'-overhung dsDNA substrate (CM0080; 5'-GACCTAGGAACCACCAGAAACACGCCACAGCCAG-3') was radiolabeled with T4 Polynucleotide Kinase (New England Biolabs) in the presence of [γP^{32}]-ATP (Perkin Elmer). Equal molar amounts of CM0080 and the long DNA strand of the 3'-overhung dsDNA substrate (CM0082; 5'-CTGGCTGTGGCGTGTTTCTGGTGGTTCCTAGGTCTTAGCCGTCTACGC CTCACT-3') were mixed, heated to 95°C for 2 mins, and slowly cooled to room temperature to anneal strands, forming the substrate. Helicase reactions were assembled with final concentrations of 5 nM substrate and 15 nM Eta in 20 mM Tris HCl pH 8.0, 10 mM $MgCl_2$, 10 µg/mL BSA, 1 mM DTT. Reactions were heated to 50°C and started by addition of dATP and trap DNA (CM0079; 5'-CTGGCTGTGGCGTGTTTCTGGTGGTTCCTAGGTC-3') to 3.5 mM and 25 nM, respectively. At the appropriate time points, 10 µL aliquots were removed and stopped on ice with the addition of 3 µL 1.5% SDS, 50% glycerol. Terminated reactions were separated on a 15% native polyacrylamide gel, and radiolabeled products were detected by exposure of the gel to a phosphorimaging screen (GE Healthcare) and analyzed using GE Imagequant 5.2.

Malachite Green ATPase assays

100 µL reactions were assembled with 500 nM substrate (prepared from CM0080 and CM0082 DNA oligonucleotides, as above) and 250 nM Eta in 20mM Tris HCl pH8.0, 10mM $MgCl_2$, 10 µg/mL BSA, 1 mM DTT with 200 µM ATP. Reactions were heated to 50°C for 90 secs, stopped by cooling on ice and the addition of 20 µL Malachite Green reagent (MAK-307, Sigma-Aldrich), followed by 15 mins incubation at room temperature in a 96-well-plate to develop. Absorbance at 620 nm was determined for each reaction in the 96-well plate using a BioTek Synergy 2.

In vitro transcription termination assays

Assembly of pre-initiation complexes and elongation to generate TECs+58 were performed as described (63). TECs+58 were captured using Nickel-coated magnetic beads (ThermoFisher), washed 3x in 20 mM Tris·HCl pH 8.0, 0.1 mM EDTA, 250 mM KCl, 4 mM $MgCl_2$, 20 µg/mL BSA, and resuspended 250 mM KCl, 20 mM Tris·HCl pH 8.0, 10 mM $MgCl_2$, and 2 mM DTT. TECs+58 were incubated in the absence or presence of 500 nM Eta (or Eta-variant) and 5 mM dATP at 85°C for 2, 5, or 10 minutes. Reaction aliquots (20 µL) were removed to ice-cold tubes containing streptavidin paramagnetic particles (Promega) that permit separation of intact and disrupted TECs. Supernatants were added to 100 µL 0.6 M Tris·HCl pH 8.0 and 12 mM EDTA and streptavidin bead pellets were resuspended in 120 µL 0.5 M Tris·HCl pH 8.0 and 10 mM EDTA. Pellets and supernatants of each reaction were then subjected to an equal volume P/C/I (25:24:1 (v/v/v) phenol/chloroform/isoamyl alcohol) extraction. Nucleic acids were precipitated with ethanol and resuspended in 4 µL formamide loading dye before separation by electrophoresis through a 15% polyacrylamide/8 M urea denaturing gel. Radiolabeled RNAs were detected by exposure to a phosphorimaging screen (GE Healthcare) and analyzed using GE ImageQuant 5.2. Release of RNAs was calculated by quantifying transcripts in the

supernatant samples divided by transcripts quantified in both the pellet and supernatant samples. *In vitro* transcription termination experiments presented in figure 6 required assembly of pre-initiation complexes and elongation to generate $TECs_{+58}$ as previous described (69). After nickel capture and washing of TECs as described above, TECs were either resuspended in Buffer A (250 mM KCl, 20 mM Tris·HCl pH 8.0, 10 mM $MgCl_2$, 2 mM DTT, 10 μM ATP, 10 μM GTP, 10 μM UTP) to prevent backtracking of TECs, or Buffer B (250 mM KCl, 20 mM Tris·HCl pH 8.0, 10 mM $MgCl_2$, 2 mM DTT) to allow backtracking of TECs. Buffer A and buffer B resuspended TECs were incubated in the absence or presence of 500 nM Eta (or Eta-variant) and 5 mM dATP at 85°C for 5 minutes. Reactions (20 μL) were removed to ice-cold tubes containing streptavidin paramagnetic particles (Promega) that permit separation of intact and disrupted TECs through binding to biotin ligated to the 5' end of the non-template strand of DNA substrates. Supernatants were added to 100 μL 0.6 M Tris·HCl pH 8.0 and 12 mM EDTA and streptavidin bead pellets were resuspended in 120 μL 0.5 M Tris·HCl pH 8.0 and 10 mM EDTA. Pellets and supernatants of each reaction were then subjected to an equal volume P/C/I (25:24:1 (v/v/v) phenol/chloroform/isoamyl alcohol) extraction. Nucleic acids were precipitated with ethanol and resuspended in 4 μL formamide loading dye before separation by electrophoresis through a 15% polyacrylamide/8 M urea denaturing gel. Radiolabeled RNAs were detected by exposure to a phosphorimaging screen (GE Healthcare) and analyzed using GE ImageQuant 5.2. Release of RNAs was calculated by quantifying transcripts in the supernatant samples divided by transcripts quantified in both the pellet and supernatant samples.

REFERENCES

1. W. Lin, *et al.*, Structural basis of ECF-σ-factor-dependent transcription initiation. *Nature Communications* **10** (2019).
2. D. Grohmann, *et al.*, The Initiation Factor TFE and the Elongation Factor Spt4/5 Compete for the RNAP Clamp during Transcription Initiation and Elongation. *Molecular Cell* **43**, 263–274 (2011).
3. A. Agapov, D. Esyunina, D. Pupov, A. Kulbachinskiy, Regulation of transcription initiation by Gfh factors from Deinococcus radiodurans. *Biochemical Journal* **473**, 4493–4505 (2016).
4. M. Lilic, S. A. Darst, E. A. Campbell, Structural basis of transcriptional activation by the Mycobacterium tuberculosis intrinsic antibiotic-resistance transcription factor WhiB7. *Molecular Cell* (2021) https:/doi.org/10.1016/j.molcel.2021.05.017 (July 5, 2021).
5. A. Wu, *et al.*, DOT1L complex regulates transcriptional initiation in human erythroleukemic cells. *Proceedings of the National Academy of Sciences* **118**, e2106148118 (2021).
6. D. F. Browning, S. J. W. Busby, Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology* **14**, 638–650 (2016).
7. J. Y. Kang, *et al.*, Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell* **173**, 1650–1662 (2018).
8. J. Ma, *et al.*, Transcription factor regulation of RNA polymerase's torque generation capacity. *Proc Natl Acad Sci U S A* **116**, 2583–2588 (2019).
9. Z. Hao, *et al.*, Pre-termination Transcription Complex: Structure and Function. *Molecular Cell* **81**, 281-292.e8 (2021).
10. V. Svetlov, E. Nudler, Reading of the non-template DNA by transcription elongation factors. *Molecular Microbiology* **109**, 417–421 (2018).

11. T. J. Sanders, *et al.*, TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Molecular Microbiology* **111**, 784–797 (2019).

12. W. Kobayashi, H. Kurumizaka, Structural transition of the nucleosome during chromatin remodeling and transcription. *Current Opinion in Structural Biology* **59**, 107–114 (2019).

13. S. M. Fuchs, R. N. Laribee, B. D. Strahl, Protein modifications in transcription elongation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1789**, 26–36 (2009).

14. J. P. Hsin, J. L. Manley, The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes and Development* **26**, 2119–2137 (2012).

15. J. Y. Kang, *et al.*, Structural basis for transcription complex 1 disruption by the mfd translocase. *Elife* **10**, 1–86 (2021).

16. J. E. Walker, O. Luyties, T. J. Santangelo, Factor-dependent archaeal transcription termination. *Proc Natl Acad Sci U S A* **114**, E6767–E6773 (2017).

17. N. Said, *et al.*, Steps toward translocation-independent RNA polymerase inactivation by terminator ATPase p. *Science (1979)* **371** (2021).

18. V. Epshtein, D. Dutta, J. Wade, E. Nudler, An allosteric mechanism of Rho-dependent transcription termination. *Nature* **463**, 245–249 (2010).

19. J. D. Eaton, *et al.*, Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Genes and Development* **32**, 127–139 (2018).

20. H. E. Mischo, *et al.*, Cell-Cycle Modulation of Transcription Termination Factor Sen1. *Molecular Cell* **70**, 312-326.e7 (2018).

21. Y. Jiang, M. Liu, C. A. Spencer, D. H. Price, Involvement of transcription termination factor 2 in mitotic repression of transcription elongation. *Molecular Cell* **14**, 375–386 (2004).

22. A. G. Arimbasseri, K. Rijal, R. J. Maraia, Transcription termination by the eukaryotic RNA polymerase III. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1829**, 318–330 (2013).

23. T. J. Santangelo, L. Cubonová, K. M. Skinner, J. N. Reeve, Archaeal intrinsic transcription termination in vivo. *Journal of Bacteriology* **191**, 7102–7108 (2009).

24. J. W. Roberts, Mechanisms of Bacterial Transcription Termination. *Journal of Molecular Biology* **431**, 4030–4039 (2019).

25. K. M. Kazmierczak, E. K. Davydova, A. A. Mustaev, L. B. Rothman-Denes, The phage N4 virion RNA polymerase catalytic domain is related to single-subunit RNA polymerases. *EMBO Journal* **21**, 5815–5823 (2002).

26. C. L. Turnbough, Regulation of Bacterial Gene Expression by Transcription Attenuation. *Microbiology and Molecular Biology Reviews* **83** (2019).

27. T. J. Sanders, C. J. Marshall, B. R. Wenck, J. N. Selan, T. J. Santangelo, "Transcription Termination" in *Reference Module in Life Sciences*, 3rd Ed., (Elsevier, 2021) https:/doi.org/10.1016/B978-0-12-819460-7.00129-8.

28. T. J. Sanders, *et al.*, FttA is a CPSF73 homologue that terminates transcription in Archaea. *Nature Microbiology* **5**, 545–553 (2020).

29. P. Mitra, G. Ghosh, M. Hafeezunnisa, R. Sen, Rho Protein: Roles and Mechanisms. *Annual Review of Microbiology* **71**, 687–709 (2017).

30. M. S. Ciampi, Rho-dependent terminators and transcription termination. *Microbiology (N Y)* **152**, 2515–2528 (2006).

31. A. Weixlbaumer, F. Grünberger, F. Werner, D. Grohmann, Coupling of Transcription and Translation in Archaea: Cues From the Bacterial World. *Frontiers in Microbiology* **12** (2021).

32. S. L. French, T. J. Santangelo, A. L. Beyer, J. N. Reeve, Transcription and translation are coupled in Archaea. *Molecular Biology and Evolution* **24**, 893–895 (2007).

33. N. J. Proudfoot, Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science (1979)* **352** (2016).

34. J. Shi, *et al.*, Structural basis of Mfd-dependent transcription termination. *Nucleic Acids Research* **48**, 11762–11772 (2020).

35. Z. Xie, D. Price, Drosophila factor 2, an RNA polymerase II transcript release factor, has DNA-dependent ATPase activity. *Journal of Biological Chemistry* **272**, 31902–31907 (1997).

36. O. Adebali, Y. Y. Chiou, J. Hu, A. Sancar, C. P. Selby, Genome-wide transcription-coupled repair in Escherichia coli is mediated by the Mfd translocase. *Proc Natl Acad Sci U S A* **114**, E2116–E2125 (2017).

37. J. Griesenbeck, H. Tschochner, D. Grohmann, Structure and function of RNA polymerases and the transcription machineries. *Sub-Cellular Biochemistry* **83**, 225–270 (2017).

38. A. Hirata, B. J. Klein, K. S. Murakami, The X-ray crystal structure of RNA polymerase from Archaea. *Nature* **451**, 851–854 (2008).

39. T. J. Santangelo, J. N. Reeve, Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *Journal of Molecular Biology* **355**, 196–210 (2006).

40. S. J. Northall, *et al.*, DNA binding and unwinding by Hel308 helicase requires dual functions of a winged helix domain. *DNA Repair* **57**, 125–132 (2017).

41. B. G. Wilson, C. W. M. Roberts, SWI/SNF nucleosome remodellers and cancer. *Nature Reviews Cancer* **11**, 481–492 (2011).

42. L. He, M. S. J. James, M. Radovcic, I. Ivancic-Bace, E. L. Bolt, Cas3 protein—a review of a multi-tasking machine. *Genes (Basel)* **11** (2020).

43. A. K. Byrd, K. D. Raney, Superfamily 2 helicases. *Frontiers in Bioscience* **17**, 2070–2088 (2012).

44. S. J. Black, *et al.*, Molecular basis of microhomology-mediated end-joining by purified full-length Polθ. *Nature Communications* **10** (2019).

45. F. He, K. Duprez, E. Hilario, Z. Chen, L. Fan, Structural basis of the XPB helicase-Bax1 nuclease complex interacting with the repair bubble DNA. *Nucleic Acids Research* **48**, 11695–11705 (2020).

46. T. van Eeuwen, *et al.*, Cryo-EM structure of TFIIH/Rad4–Rad23–Rad33 in damaged DNA opening in nucleotide excision repair. *Nature Communications* **12** (2021).

47. L. Holm, “Using Dali for Protein Structure Comparison” in *Methods in Molecular Biology*, (Humana Press Inc., 2020), pp. 29–42.

48. J. D. Richards, *et al.*, Structure of the DNA repair helicase Hel308 reveals DNA binding and autoinhibitory domains. *Journal of Biological Chemistry* **283**, 5118–5126 (2008).

49. E. Absmeier, K. F. Santos, M. C. Wahl, Molecular Mechanism Underlying Inhibition of Intrinsic ATPase Activity in a Ski2-like RNA Helicase. *Structure* **28**, 236-243.e3 (2020).

50. T. Kouba, *et al.*, Mycobacterial HelD is a nucleic acids-clearing factor for RNA polymerase. *Nature Communications* **11** (2020).

51. C. J. Marshall, T. J. Santangelo, Archaeal DNA Repair Mechanisms. *Biomolecules* **10**, 1472 (2020).

52. A. Fujiwara, *et al.*, Application of a Euryarchaeota-specific helicase from Thermococcus kodakarensis for noise reduction in PCR. *Applied and Environmental Microbiology* **82**, 3022–3031 (2016).

53. B. Clouet-D’Orval, *et al.*, Insights into RNA-processing pathways and associated RNA-degrading enzymes in Archaea. *FEMS Microbiology Reviews* **42**, 579–613 (2018).

54. D. K. Phung, *et al.*, RNA processing machineries in Archaea: The 5’-3’ exoribonuclease aRNase J of the β-CASP family is engaged specifically with the helicase ASH-Ski2 and the 3’-5’ exoribonucleolytic RNA exosome machinery. *Nucleic Acids Research* **48**, 3832–3847 (2020).

55. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**, 845–858 (2015).

56. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

57. S. H. Jun, *et al.*, Direct binding of TFEα opens DNA binding cleft of RNA polymerase. *Nature Communications* **11** (2020).

58. J. Y. Kang, *et al.*, Structural basis for transcription complex disruption by the Mfd translocase. *Elife* **10** (2021).

59. A. J. Smith, N. J. Savery, RNA polymerase mutants defective in the initiation of transcription-coupled DNA repair. *Nucleic Acids Research* **33**, 755–764 (2005).

60. N. Stantial, J. Dumpe, K. Pietrosimone, F. Baltazar, D. J. Crowley, Transcription-coupled repair of UV damage in the halophilic archaea. *DNA Repair* **41**, 63–68 (2016).

61. A. M. Gehring, T. J. Santangelo, Archaeal RNA polymerase arrests transcription at DNA lesions. *Transcription* **8**, 288–296 (2017).

62. J. E. Walker, O. Luyties, T. J. Santangelo, Factor-dependent archaeal transcription termination. *Proc Natl Acad Sci U S A* **114**, E6767–E6773 (2017).

63. M. Ogrunc, D. F. Becker, S. W. Ragsdale, A. Sancar, Nucleotide excision repair in the third kingdom. *Journal of Bacteriology* **180**, 5796–5798 (1998).

64. C. P. Selby, Mfd Protein and Transcription–Repair Coupling in Escherichia coli. *Photochemistry and Photobiology* **93**, 280–295 (2017).

65. Z. Otwinowski, W. Minor, Processing of X-ray diffraction data collected in oscillation mode. *Methods in Enzymology* **276**, 307–326 (1997).

66. P. D. Adams, *et al.*, PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography* **66**, 213–221 (2010).

67. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography* **60**, 2126–2132 (2004).

68. J. S. Papadopoulos, R. Agarwala, COBALT: Constraint-based alignment tool for multiple protein sequences. *Bioinformatics* **23**, 1073–1079 (2007).

69. A. M. Gehring, T. J. Santangelo, Manipulating archaeal systems to permit analyses of transcription elongation-termination decisions in vitro. *Methods Mol Biol* **1276**, 263–279 (2015).

CHAPTER 3

A NOVEL GENETIC SCREEN INFERS FTTA IS RESPONSIBLE FOR POLAR REPRESSION OF TRANSCRIPTION IN ARCHAEA.

3.1 Introduction

Transcription must be appropriately and dynamically controlled at all stages of the transcription cycle to ensure proper gene expression and total cellular health. Transcription elongation complexes (TECs) must sometimes transcribe outstandingly (>50kbp) long genes and are thus necessarily stable, consisting of RNA polymerase (RNAP), associated factors, nascent transcript, and the dsDNA template(1). The stability of the TEC 'juggernaut' makes TECs difficult to disrupt, and thus how cells terminate transcription is of research interest. Many prokaryotic transcripts contain RNA hairpins and/or U-rich RNAs which have the potential to terminate TECs based on nucleic acid sequence alone, a concept known as intrinsic termination, but these intrinsic terminators do not allow more nuanced regulation and are not universally distributed at the end of coding regions (2–5). Factor-mediated termination allows for more control of transcription termination in both a general (at the 3' end of coding genes) and specialized (sites of RNAP arrest) sense. However, only a handful of protein factors across life have been identified which can dissociate the extremely stable TEC(6–12).

Prokaryotes lack nuclei, and the processes of transcription and termination are functionally coupled in most prokaryotes (13, 14). In Bacteria, the transcription and translation apparatuses are coupled by the NusG protein and normally become uncoupled following the stop codon-mediated termination of translation. The nascent RNA protruding from the TEC is then vulnerable to the activities of the termination factor Rho. Rho is a hexamer which targets ribosome-free nascent transcripts, wrapping ~60nt of RNA around the core subunits. Rho uses robust ATP hydrolysis to translocate the RNA and catch-up with the vulnerable TEC, and

ultimately terminates transcription (15–19). The exact mechanism by which Rho terminates the TEC is still disputed, but evidence suggests important allosteric changes and/or kinetic competition between Rho and TECs are necessary for termination (20, 21). Rho-mediated termination occurs whenever TECs are not functionally coupled to a ribosome, and is not limited to instances where ribosomes were halted due to natural stop codons.

We recently identified the first two protein factors capable of terminating transcription in the prokaryotic Archaea. One of which, the Factor that terminates transcription in Archaea (FttA) appears to act analogously to Rho, terminating transcription after the uncoupling of transcription and translation(10). FttA is a homolog of human CPSF73, a subunit of the Cleavage and Polyadenylation Specificity Factor complex responsible for identification, cleavage and subsequent polyadenylation at AAUAAA tracts at the end of almost all eukaryotic mRNA transcripts (22–24). The conservation of FttA with a member of a eukaryotic termination pathway is not surprising, due to the substantial similarities between the lone multi-subunit archaeal RNAP and eukaryotic RNAP II (25). While the human CPSF complex works to cleave mRNA transcripts and direct 3' end transcript maturation, the complex itself does not terminate transcription, but produces an uncapped 5' end vulnerable to the transcription termination activities of Xrn2 (26, 27). FttA, however, both cleaves the RNA *and* terminates TECs. It is likely that the four similar-yet-diverged subunits of the CPSF complex arose from the single archaeal form, and lost termination activity during speciation events allowing eukaryotic organisms more control over 3' end formation (23, 28). Recent biochemical analyses of FttA reveal it is responsible for general termination at the end of 3' coding genes; reduced expression of FttA leads to RNA transcripts with longer 3'UTR regions (10).

Prokaryotic genomes commonly have reduced intergenic space and contain operons, sets of multiple distinct (but usually related) genes driven by a single upstream promoter, allowing regulation of multiple genes by one regulatory event. Attenuation occurs in many

operons, i.e. the well-studied bacterial Trp operon, where unique folding patterns within the RNA transcript can drive the uncoupling of transcription and translation, selectively permitting Rho-mediated transcription termination to regulate gene expression (29, 30). Genes downstream of the 'attenuation' event are therefore not expressed under certain conditions, due to a genetic element in an upstream gene; the influence of upstream sequences on the transcription of downstream sequences results in a biological concept known as polarity (Figure 3.1) (31, 32). The mechanistic nature of polarity implies the termination event can also likely occur in undesirable instances such as a stop codon mutation in a gene body upstream within an operon, or amino acid starvation leading to ribosome arrest.

The polar repression of transcription is also present in Archaea (31) and the mode of activity of FttA (recognizing the uncoupling of transcription and translation to terminate transcription) suggests that it is responsible for the polarity phenomenon. While it is well documented that the transcription termination factor Rho is responsible for polarity in Bacteria (33, 34), there is no concrete evidence that FttA is responsible for the polar repression of transcription in Archaea. The importance of properly regulated transcription termination is highlighted by the strict essentiality of FttA; it has been challenging designing rational mutant forms of FttA for *in vivo* investigations of the potential role of FttA in the polar repression of transcription as significant perturbations to the activity of FttA are also lethal. Thus, to determine the transcription termination factor responsible for polar repression of archaeal transcription *in vivo,* we designed a novel genetic screen in the euryarchaeon *Thermococcus kodakarensis* reliant of genetically encoded polarity-regulated constructs that, in wild-type cells, limited expression of essential genes and blocked rapid growth due to the polarity phenomenon. Only cells which obtained a functional mutation during our screen which allowed readthrough of our polarity-regulated constructs (presumably due to perturbations to factors involved the polar repression of transcription) could survive. As our screen simply required cells to read through our polarity-regulated constructs, we expected a significant background of cells harboring

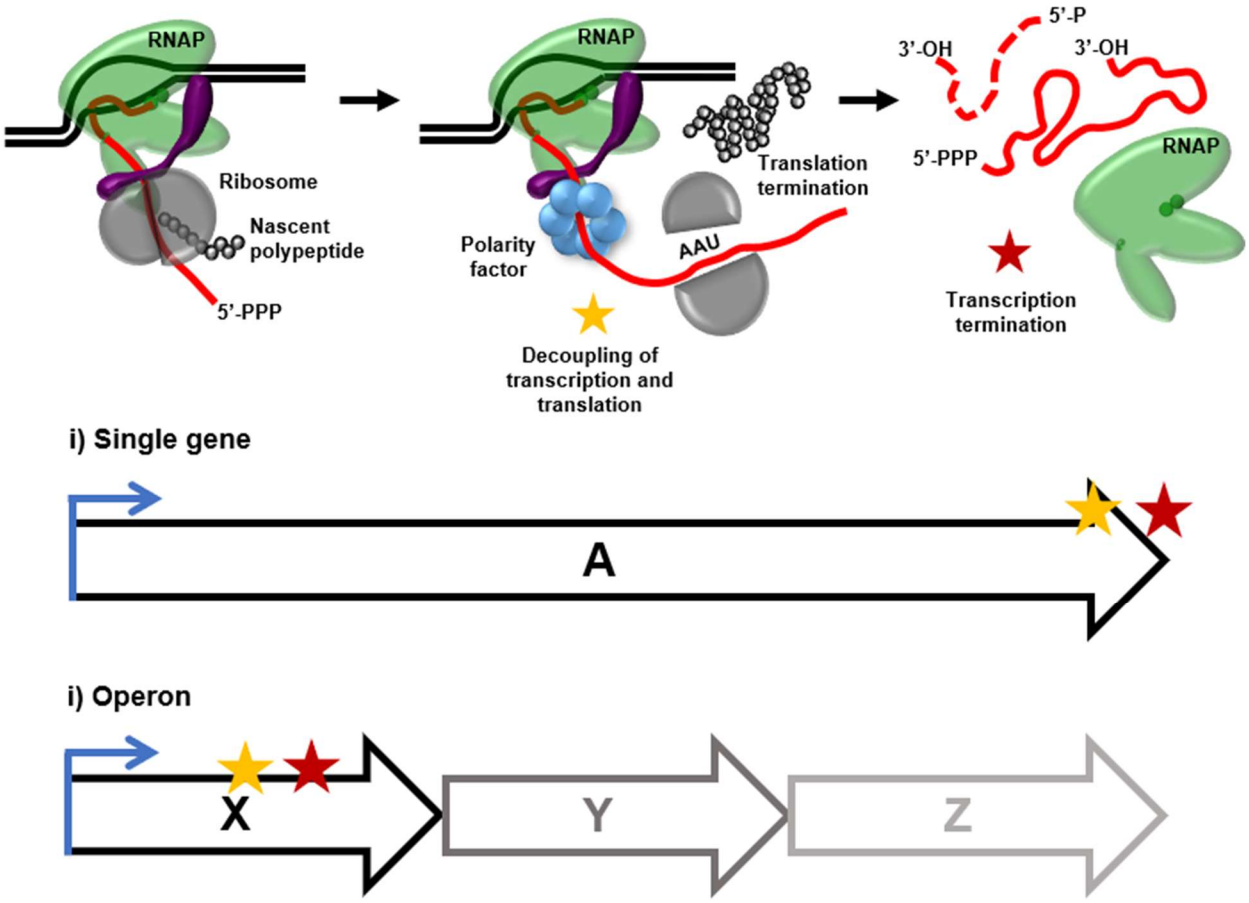


Figure 3.1: General termination and polar repression of transcription in prokaryotes. i) General transcription termination at the end of most coding genes (A) in prokaryotes occurs when transcription and translation are uncoupled after stop codon-mediated translation termination (yellow star). Termination occurs at varied distances after the stop codon (red star), producing varied 3'UTRs and is dependent on recognition of the free RNA by Rho in Bacteria and FttA in Archaea. ii) The polar repression of transcription occurs when transcription and translation become uncoupled in an upstream gene (X) of a multi-cistronic operon. Transcription termination will occur slightly downstream of this uncoupling event in a similar mechanism as stop-codon mediated general termination. Thus, genes downstream (Y & Z) are repressed in a 'polar' manner. Rho is known to also be responsible for the polar repression of transcription in Bacteria, and we hypothesize FttA is the archaeal 'polarity factor'.

perturbations to factors not involved in polarity, i.e. disruptions to translation termination, increased RNA binding of an RNA binding factor to shield TECs from transcription termination, etc. While we did observe the expected background, 3/10 candidate surviving colonies contained mutations to FttA or RNAP, two factors hypothesized to be involved in the polar repression of transcription before the genetic screen was performed. The genetic screen strongly suggests that FttA is responsible for polar repression of transcription in Archaea and mutations impacting both RNAP and FttA can impact the efficiency of FttA-mediated transcription termination. Our results provide valuable insights into susceptibility of prokaryotic TECs to transcription termination and, given the homology of FttA to conserved eukaryotic transcription termination factors, evolutionary insight into aspects of eukaryotic transcription termination.

3.2 Results

3.2.1 *T. kodakarensis* strain CM003 harbors 'polarity constructs' upstream of essential metabolomic genes and exhibits poor growth on minimal His/Trp media.

Our genetic screen was dependent on 'polarity constructs' placed upstream of genes responsible for histidine (TK0244, *hisD*) and tryptophan (TK0254, *trpE*) synthesis in a single laboratory strain of the genetically tractable *T. kodakarensis*. The design of the polarity constructs ensured that *hisD* and *trpE* genes cannot be expressed highly enough in cells with a healthy system for the polar repression of transcription to prevent rendering the cells His/Trp auxotrophs. Polarity-repressed constructs (Figure 3.2.A) were designed to emulate a bicistronic operon consisting of a non-essential and non-interfering gene upstream of the essential *hisD* and *trpE* genes. In the non-essential/non-interfering gene, a constitutive promoter ensured transcription was efficiently initiated, and two stop codon mutations were placed in the open

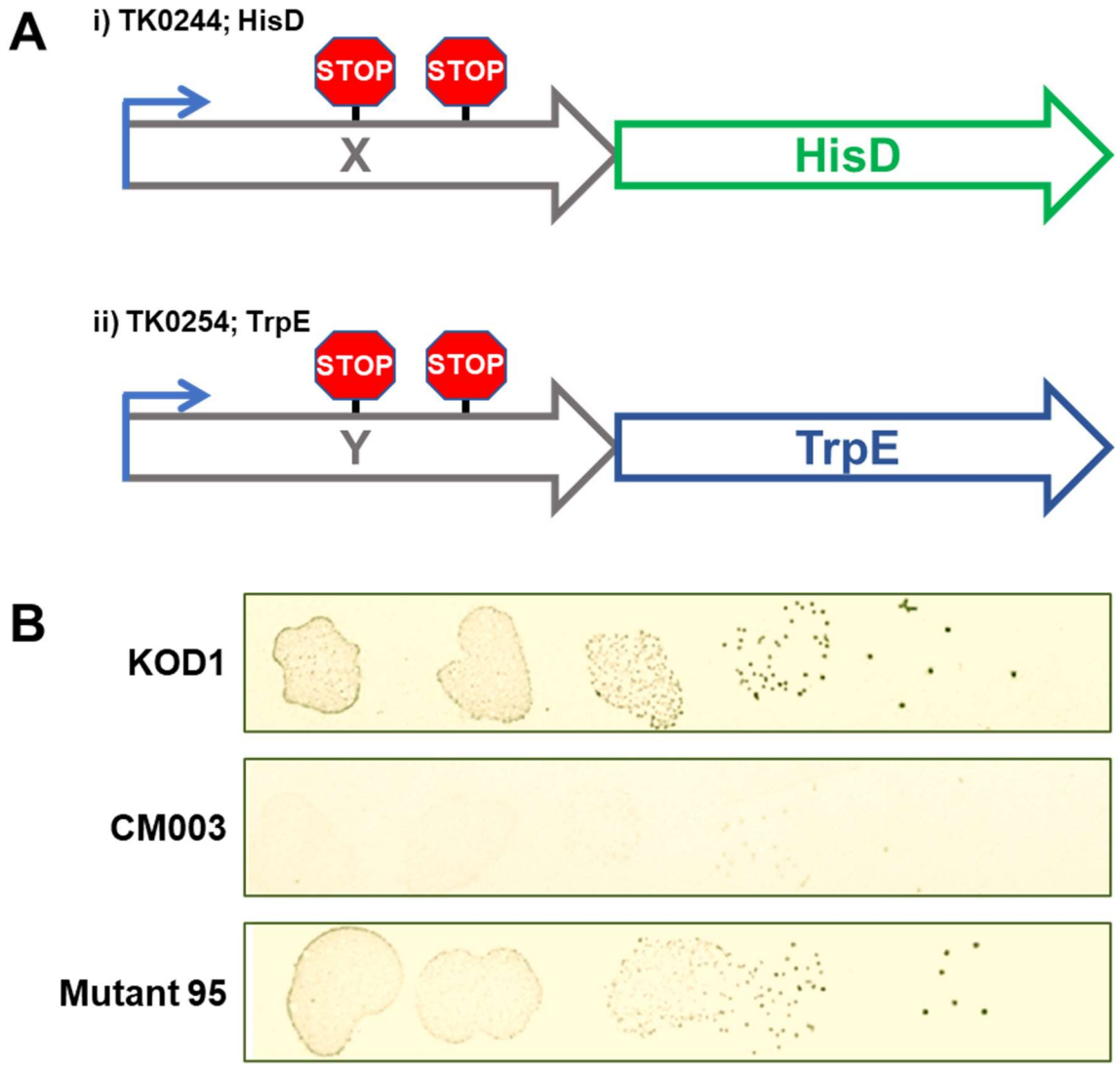


Figure 3.2: "Polarity constructs" encoded in *T.kodakarensis* strain CM003 render cells Trp/His auxotrophic.
A) "Polarity constructs" were genetically engineered into strain TS559 upstream of gene TK0254 (*trpE*) and TK0244 (*hisD*). Constructs placed a strong promoter (blue arrow) to ensure transcription of a synthetic bicistronic operon consisting of i) an upstream non-essential & non-interfering gene and ii) the target gene, *trpE* or *hisD*. Two in-frame stop codons were encoded in the upstream gene, preventing its proper expression but also inducing the polar repression of transcription of the downstream *trpE*/*hisD* gene. B) Strain CM003 is auxotrophic when plated with minimal amounts of tryptophan and histidine compared to the prototrophic wild-type strain, KOD1. In this study, mutants were generated (i.e. Mutant 95), that reduced the polar repression of *hisD*/*trpE* transcription and rescued tryptophan/histidine prototrophy of cells.

reading frame. The stop codon mutations not only prevent expression of the upstream gene, but also cause translation termination and the subsequent uncoupling of translation from transcription. The transcription termination event prevents proper expression of the downstream *hisD*/*trpE* genes. Placing *two* stop codons in upstream genes of polarity-repressed constructs in front of both *hisD* and *trpE* simultaneously reduced the likelihood of random basal mutations changing intrinsic nucleic acid sequences in the polarity construct to allow readthrough. Separate plasmids containing 'polarity constructs' targeting both the *hisD* and *trpE* genes were generated for homologous-recombination based strain construction. Successful and successive incorporations of each construct yielded strain CM003. Depending on the levels of basal readthrough of our polarity constructs, CM003 would likely be His/Trp auxotrophic at sufficiently low levels of His/Trp. The maximum levels of His/Trp allowed in media were determined before strain CM003 exhibited abrogated growth compared to the wild-type His/Trp prototrophic strain KOD1. Plated on media containing 1.6 μM histidine and 27.5 μM tryptophan, the auxotrophic CM003 exhibits severely compromised growth compared to the prototrophic wild-type strain KOD1, indicating the polar repression of transcription in CM003 cells is preventing expression of *hisD* and *trpE* in the context of our polarity constructs. Our genetic screen consequently aimed to identify mutations to CM003 cells which eliminated this growth speed defect in minimal His/Trp conditions, presumably due to increased expression of *hisD* and *trpE* resulting from mutations to genes encoding factors involved in the polar repression of transcription.

3.2.2 A simple selection of CM003 cells with mutations allowing transcription of essential genes through the polarity constructs yielded 10 mutant strains.

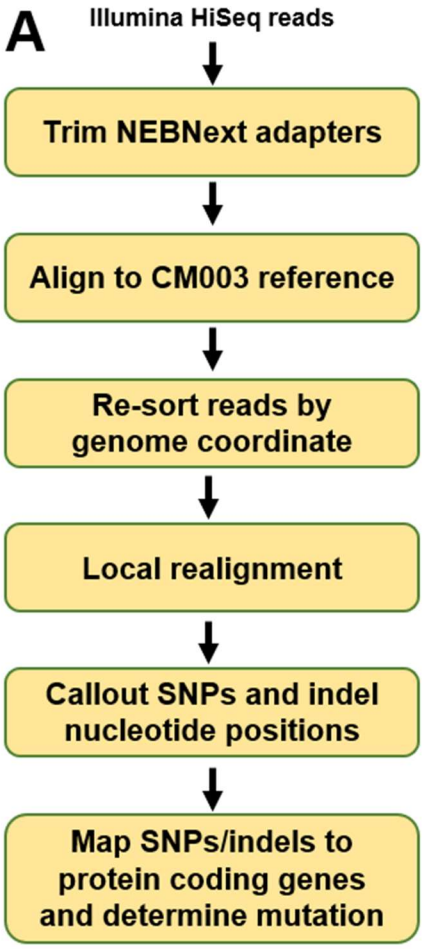
Sheer numbers (~10^9 cells per 5mL culture) and a basal archaeal mutation rate likely comparable to mesophilic prokaryotes (35) permitted us to simply passage CM003 cultures twice overnight in rich media supplemented with histidine and tryptophan. Mutations randomly

distributed at a low level throughout the trillions of genomes in the final culture due to rare errors produced by healthy DNA replication and repair pathway components. Under rich media conditions, mutations to genes encoding factors involved in polarity may prove inconsequential allowing their detection before subsequent challenge steps. Cells were then immediately plated on minimal His/Trp media plates (containing levels of His/Trp determined above) allowing us to screen for individual cells harboring mutations to genes which conferred Trp/His prototrophy. Without supplementation of His/Trp, and because of our genetically encoded polarity constructs, cells would die unless one of the random basal mutations harbored in an individual cell changed the function of some factor(s) to allow expression of the *hisD* or *trpE* genes. Such a mutation presumably perturbs proper polar repression of transcription allowing the mutant strain to transcribe the histidine and tryptophan synthesis genes through the polarity construct. An overwhelming majority of cells were unable to survive due to the tremendous unlikelihood that a randomly distributed mutation would disrupt polar repression of transcription and allow Trp/His prototrophy of CM003. Further, many mutations that perturb the polar repression of transcription could also lethally perturb transcription termination in general – killing the cells due to a an unsurmountable disruption of the transcription apparatus. Nonetheless, we were successfully able to isolate colonies from 10 mutant cells of CM003 which were able to grow in the absence of histidine or tryptophan and exhibited growth similar to the parent strain when plated on minimal media (Fig 3.2.B).

3.2.3 Whole genome sequencing revealed multiple mutant CM003 strains with mutations to FttA and/or RNA Polymerase.

To identify mutations harbored within the 10 mutant strains (mutants 60, 66, 71, 77, 80, 95, 99, 100, 105, & 109) which were able to survive in the absence of histidine or tryptophan, individual colonies were cultured and genomic DNA was extracted. Illlumina NextSeq.libraries were then

prepared from the mutant strain gDNA samples (NEBNext) and individual fragments were sequenced. A bioinformatic pipeline was developed (Figure 3.3.A) to trim adapters from raw reads, align sequenced DNA to a reference genome of CM003, and call out mutations as single-nucleotide polymorphisms (SNPs), insertions, or deletions, and determine non-silent mutations to gene-coding and non-coding areas of the genome. Across 10 candidate mutant strains sequenced, a total of 271 individual point mutations were identified, ranging from 6 to 106 per individual strain. This list was refined down to 127 by mapping each point mutation to the annotated gene / genetic element at that position in the genome, discarding mutations to non-protein/non-regulatory regions of the genome, and removing those also identified as point mutations in the parent strain (i.e. errors in the reference genome) (Figure 3.3.B). Further, the depth of sequencing (100-fold genome coverage) suggests that while *T. kodakarensis* may contain up to ~12 genomes, each mutation identified was a total conversion. Seven out of the ten mutant strains contained many mutations to genes encoding factors which do not appear to function in transcription or transcription termination (tRNAs, ribosomal subunits, 'hypothetical' proteins, etc.) and while we could hypothesize how these strains gained Trp/His prototrophy (see discussion), they did not help us obviously identify an archaeal polarity factor. Further, no surviving strains contained any mutagenesis to any elements of either polarity construct, suggesting they still operate as intended. Excitingly, three out of the ten mutant strains harbored mutations to either genes encoding FttA (TK1428) or the RNAP beta subunit (TK1083), two factors hypothesized to be central to the polar repression of transcription (Figure 3.3.C). One mutant strain in particular (mutant 109) was able to grow in the absence of histidine and tryptophan with only a mutation to the protein coding region of the beta-subunit of RNAP and to two hypothetical proteins.



B

mutant #	SNPs/indels in coding regions	notable mutations
60	6	
66	32	methionine tRNA
71	8	
77	9	
80	12	ribosomal subunit
95	38	RNAP beta subunit
99	4	
100	10	FttA
105	5	
109	3	RNAP beta subunit
Total	127	

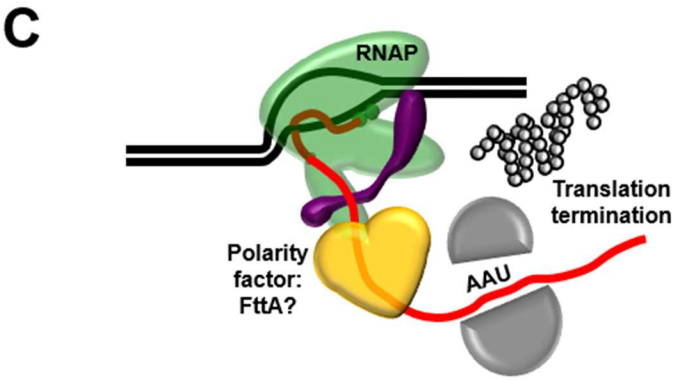
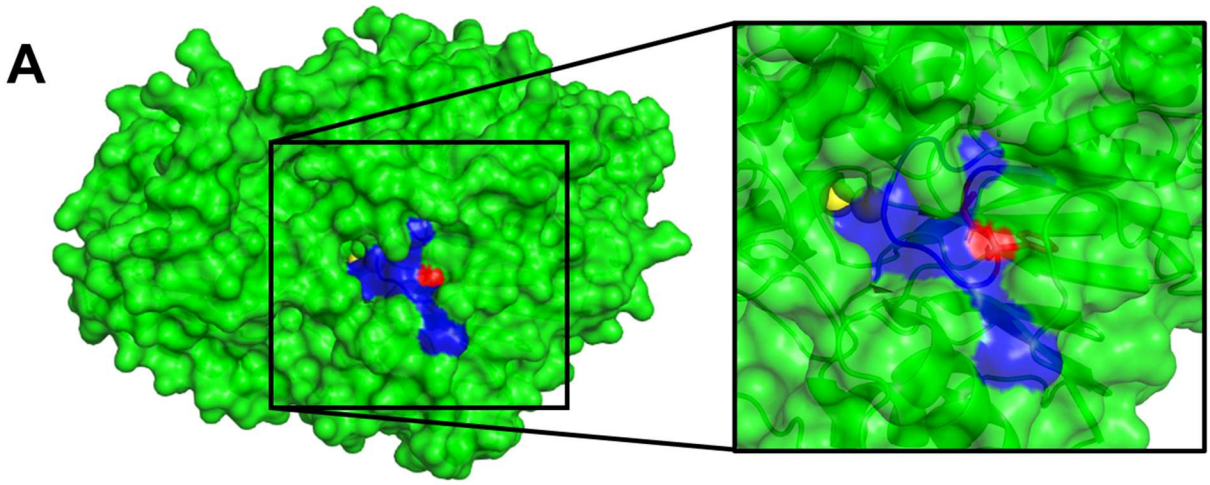


Figure 3.3 Next-generation sequencing of mutant strains passing the 'polarity' genetic screen allowed determination of open reading frame SNPs, revealing three mutant strains containing mutations to RNAP or FttA.
A) A biofnformatic pipeline was developed on Galaxy tools (usegalaxy.org) to prepare raw DNA sequencing reads from mutant strains and align them to a CM003 reference genome while allowing for determination of SNPs, insertions, and deletions in the mutant strain genome. B) Ten mutant strains were sequenced and mutant genomes contained a range of 3-37 total SNPs/insertions/deletions. Mutants 95, 100, and 109 each contained a non-silent mutation to FttA or RNAP. C) Mutations to RNAP (green) and FttA (yellow) could plausibly render RNAP resistant to FttA, impact proper transcription termination by FttA, or prevent proper coupling of FttA to RNAP by Spt4/5 (purple).

3.2.4 Mutations to both FttA and RNAP identified from the polarity screen map to regions of each enzyme critical for function and conserved between Archaea and humans.

The FttA mutation identified from our polarity screen resulted in a glycine to serine mutation at position 193 and a phenylalanine to leucine mutation at position 194- substitutions to two adjacent amino acids likely caused by one mutational event. The X-ray crystal structure of the close relative of FttA from *Pyrococcus horikoshii* (36)(PDB: 3AF5) shares ~90% sequence identity with FttA from *T. kodakarensis* and was used to map the positions of G193 and F194 relative to known regions of the enzyme (Figure 3.4). The structure suggests F194 is positioned pointing inwards towards the core of the enzyme close to the active site-coordinated Zn^{2+}, where RNA cleavage likely occurs.. G193 is completely conserved across euryarchaea, and appears as a surface exposed residue between two highly conserved regions of FttA which are retained between FttA and human CPSF73 .These conserved regions appear to form the entry channel into the active site of the enzyme and could be potentially involved in mediating correct nucleic acid contacts or proper coupling events for FttA-mediated termination. It is very plausible that substitutions to amino acids in this location of FttA could alter the function of the enzyme enough to allow cells harboring the mutation to survive the polarity-based genetic screen.

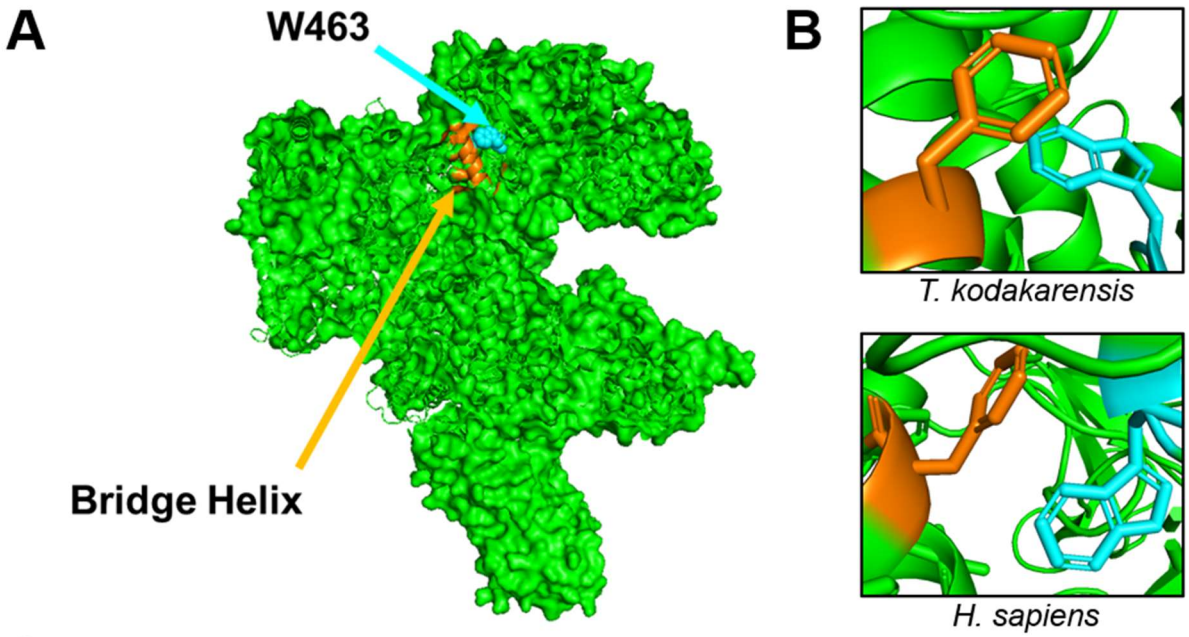
The same RNAP mutation was identified from both mutant strains 95 and 109 in our genetic screen results in a substitution of tryptophan to arginine at position 463 of the RNAP beta subunit, encoded by TK1083. A recent cryo-EM structure of the single RNAP encoded in *T.kodakarensis* (37)(PDB: 6KF4) was used to map the position of the RNAP beta subunit W463. W463 is buried in the catalytic core of RNAP, at the interface between the beta and alpha prime subunits, where direct coordination is observed with W463 and a phenylalanine of the RNAP 'bridge helix' (Figure 3.4.A). The RNAP bridge helix is a highly conserved structural element in



B

```
Human CPSF73    EESDQLLIRPLGAGQEVGRSCIILEFKGRKIMLDCGIHPG 58
T.k FttA        YKSRWIRITGLGGFREVGRSALLVQTDESYVLVDFGVNIA 240
                 :*  : *  **. :*****.:::: .   :::* *:: .
```

Figure 3.4: Mutations to FttA identified in the polarity screen border a highly conserved region leading to the active site.
A) Residues G193 and F194 (red) were mutated in the polarity-based genetic screen and are surface accessible along a channel completely conserved between *T. kodakarensis* FttA and *H. sapiens* CPSF73 (blue). The active site is evidenced by the coordinating Zn^{2+} (yellow).
B) Alignment of *T. kodakarensis* FttA and *H. sapiens* CPSF73 centered around G193/F194 of FttA reveals conservation of residues after billions of years of divergent evolution.



C

```
human Rpb2   MSSRAGVSQVLNRYTYSSTLSHLRRTNTPIGRDGKLAKPRQLHNTHWGLVCPAETPEG  532
T.K RpoB     --GRTGVSQLLDRTNYMSTLSHLRRVTSPLSRDQPHFEARDLHGTHWGRICPTETPEG  476
               .*:****:*:* .* ********..:*:.**    : *:**.**** :**:*****

human Rpb1   RGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDA  879
T.k RpoA'    RGFVTNSYKSGLTPQEYFFHAMGGREGLVDTAVRTAQSGYMQRRLINALQDLKVDYDG  860
             **** *** :**** *:***********:****:**::**:*****:::::.: *.**.
```

Figure 3.5 Mutation to RNAP beta-subunit identified in the polarity screen effects a highly conserved residue in the core of RNAP which is conserved across life.
A) W463 (blue) of the *T. kodakarensis* RNAP beta-subunit was mutated in the polarity-based genetic screen and makes critical contacts with the bridge helix (orange) in the alpha prime subunit of RNAP. B) The beta-subunit tryptophan coordinates with a phenylalanine of the bridge-helix. Both residues and their interaction is conserved between *T. kodakarensis* and *H. sapiens*. C) Alignment of *T. kodakarensis* and *H. sapiens* RNAP beta and alpha-prime subunits centered around regions containing the beta-subunit tryptophan identified in this study and coordinating bridge-helix phenylalanine indicates extremely high levels of conservation of the RNAP core between humans and Archaea.

the alpha prime subunit of RNAP critical for facilitating the correct catalytic contacts required for nucleotide addition(38). W463 and,unsurprisingly, the coordinating bridge-helix phenylalanine are both conserved between *T. kodakarensis* RNAP and human RNAP II- almost a billion years of evolutionary divergence- further illustrating the importance of W463, and the coordination of the residue with the RNAP bridge helix (Figure 3.4.B). It is interesting the same genetic mutation was identified in two separate mutant strains of *T.kodakarensis,* but the nature of the mutation and almost total conservation of this residue suggest this is not a functionally relevant polymorphism and hints at a functional basis for the apparent lack of polarity in encoding mutant strains.

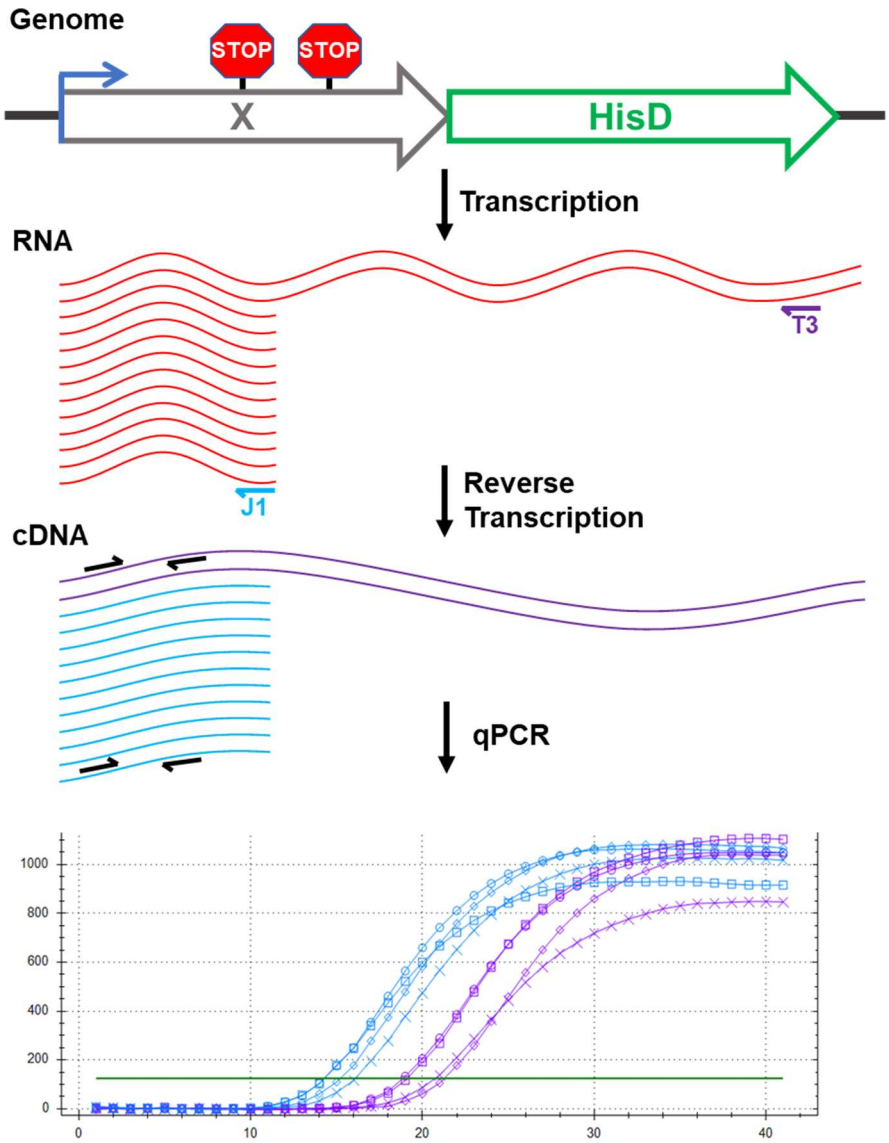
Several mutant strains recovered from our polarity construct-based genetic screen contained mutations resulting in changes to functionally relevant residues of two factors which were predicted to be involved in polarity beforehand (FttA and RNAP). Strains without mutations resulting in changes to predicted factors were evaluated, but no obvious or potential transcription termination factors playing a role in the polar repression of transcription were recognized. However, plausible explanations for their passing of our genetic screen could be made (see discussion). Taken together, our results suggest the genetic screen was successful in identifying mutant strains which allowed readthrough of our polarity constructs and that FttA may be responsible for the polar repression of transcription in Archaea.

3.2.5 Strains harboring mutations to RNAP or FttA surviving the polarity genetic screen were able to transcribe *hisD*/*trpE* through polarity repressed constructs more efficiently than the CM003 mutant strain.

It is plausible that some other metabolic process has been altered which allowed our mutant strains to thrive in minimal amounts of histidine and tryptophan. To ensure mutant strains were

able to pass our polarity-based genetic screen specifically because of transcriptional readthrough of the genetically encoded polarity-repressive constructs, we performed RT-qPCR on the polarity construct loci at *trpE* and *hisD* (Figure 3.6.A). Primers were designed for reverse transcription which annealed to two strategic regions of each polarity construct; i) RNA immediately downstream of the first genetically encoded stop codon in the upstream non-essential gene (Figure 3.6.B, J1) and ii) RNA close to the stop codon of *hisD* or *trpE* (Figure 3.6.B, T3). Reverse transcription of RNA using the J1 primer yields all RNAs that made it to the first upstream stop codon, whereas reverse transcription of RNA using the T3 primer yields only cDNA from full-length RNAs that were able to read-through our polarity constructs. qPCR quantification of the resultant cDNAs obtained from each reverse transcription primer using a primer pair upstream of the J1 reverse transcription primer (Figure 3.6.C) allows determination of the relative readthrough of the polarity constructs between the parent strain CM003 and mutant strains. The difference between the Ct values (ΔCt) obtained from qPCR amplification of J1 and T3 derived cDNAs displays the proportion of TECs which were able to read through the polarity constructs. A smaller ΔCt indicates a higher percentage of transcription events successfully reading through the polarity construct. A decrease in this ΔCt value in a mutant strain, or a more positive ΔΔCt between the mutant strain ΔCt and CM003 ΔCt, indicates more full length RNA indicative of readthrough of the polarity constructs.

The RT-qPCR experiment was performed on the *hisD* and *trpE* polarity construct loci of the CM003 parent strain, mutant strain 95 (containing ~20 point mutations including $FttA^{G193S/F194L}$), CM005 ($FttA^{G193S/F194L}$ in a CM003 background), and CM007 ($RNAP^{\beta\text{-}W463A}$ in a CM003 background) (Figure 3.6.D). RT-qPCR of the *hisD* polarity locus revealed a subtle but reproducible ~2 fold increase in the proportion of full-length *hisD* polarity construct transcripts gene in mutant 95 and CM005, and a ~4 fold increase of this proportion in CM007. An identically designed experiment of the *trpE* polarity locus reveals very similar numbers. While



TK0244 polarity construct					
Strain	**J1 Ct**	**T3 Ct**	**ΔCt**	**ΔΔCt CM003**	**Fold Change Full length RNA**
CM003	14.59	19.78	5.19	0	**1.00**
CM005	13.25	17.28	4.03	1.16	**2.23**
CM007	13.78	17.01	3.23	1.96	**3.89**
Mut 95	13.86	17.85	3.99	1.2	**2.30**

TK0254 polarity construct					
Strain	**J1 Ct**	**T3 Ct**	**ΔCt**	**ΔΔCt CM003**	**Fold Change Full length RNA**
CM003	21.28	27.59	6.31	0	**1.00**
CM005	20.67	25.02	4.35	0.84	**1.79**
CM007	21.01	24.28	3.27	1.92	**3.78**
Mut 95	21.52	25.36	3.84	1.35	**2.55**

Figure 3.6: RT-qPCR provides evidence of increased readthrough transcription of polarity constructs in strains recovered from polarity screen and strains harboring mutations to RNAP and FttA identified in the polarity screen.
A) Polarity loci encoded in CM003 have the same basic structure and should act to repress transcription of *hisD* or *trpE* by the polarity mechanism. B) RNAs produced by transcription of the polarity loci in CM003 should mostly be truncated before reaching *hisD* or *trpE* due to the polar repression of transcription. Primer J1 allows cDNA synthesis of all RNAs that were initiated from the polarity locus promoter. Primer T3 only allows cDNA synthesis of full length RNAs. C) Primers upstream of the target sequence of the J1 reverse transcription primer allow relative quantification of total vs. full length cDNAs across parent and mutant strains by qPCR. D) Quantified qPCR results of the *hisD* (TK0244) and *trpE* (TK0254) polarity construct loci. The ΔΔCt value represents the change in proportion of full-length vs. total cDNA between the CM003 parent strain and polarity mutant strains. All mutant strains tested indicated a 2-4X increase in proportion of full-length RNA.

the numbers are subtle, it is plausible that CM003 only needed a small increase in *hisD*/*trpE* transcription to produce enough active enzyme to relay prototrophy under our screen conditions. However, the RT-qPCR experiment evidences the screen worked as expected, and the reduced polar repression of transcription in strain CM005 (FttA$^{G193S/F194L}$ in CM003) strongly suggests FttA is responsible for the polar repression of transcription in Archaea.

3.2.6 T. kodakarensis RNAP 'polarity mutant' performs comparably to the wild-type RNAP in *in vitro* transcription and transcription termination experiments.

To determine a biochemical explanation for a mutation to the RNAP beta subunit allowing cells to read through our polarity constructs, a *T. kodakarensis* strain harboring the mutant form of RNAP (CM007) was constructed and cultured to isolate the mutant form of RNAP (RNAP beta subunit (TK1083) W463A; RNAP^{PM}) for biochemical experimentation. It could be rationalized that RNAP^{PM} contains a substitution that affects the nucleotide addition cycle and that TECs formed with the mutant polymerase are simply faster or pause resistant, reaching the end of our polarity construct before FttA can terminate transcription. Conversely, the mutant RNAP could also simply be termination resistant, preventing FttA (and potentially other transcription termination factors) from performing transcription termination efficiently before it is able to read through the polarity construct.

Differences between overall RNA synthesis speed and pausing patterns in RNAP^{WT} and RNAP^{PM} were first investigated using our *in vitro* transcription system (Figure 3.7.A), and it was quickly apparent that under our *in vitro* transcription system conditions, there were no differences between the transcription speed and pausing frequency of the wild-type and polarity mutant RNAPs (Figure 3.7.B). The ability of known archaeal transcription termination factors Eta and FttA to terminate stalled TECs assembled with RNAP^{WT} or RNAP^{PM} was then tested to

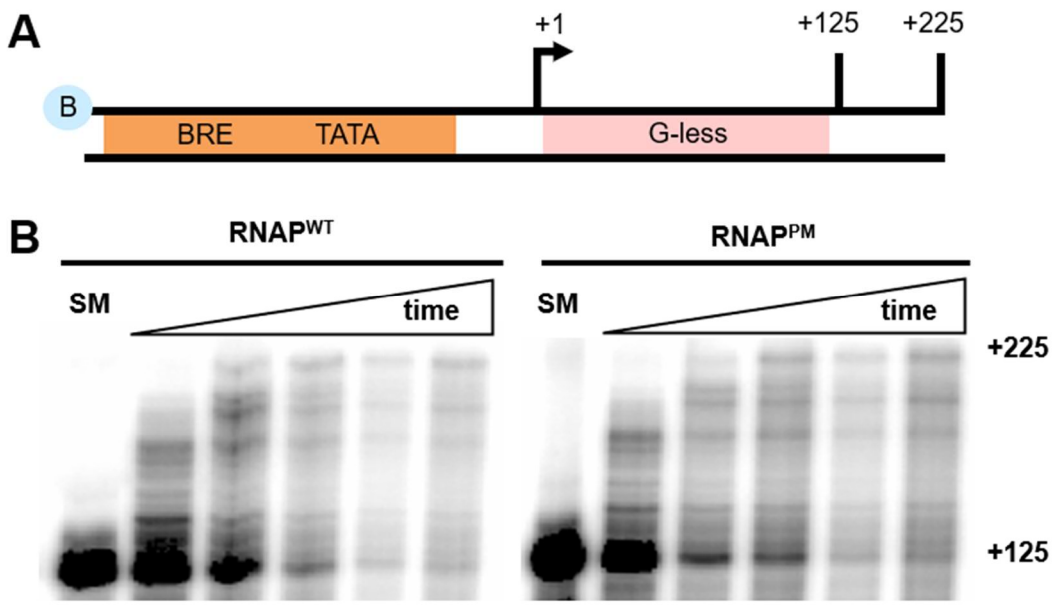


Figure 3.7: RNAP^{PM} transcribes transcription templates similarly to RNAP^{WT} *in vitro*. A) 5' biotinylated transcription templates consist of a strong BRE/TATA based promoter (orange) mapped to a known transcription start site (+1), a G-less cassette to the 125th nucleotide position relative to the known transcription start site (+125), and an unrestrained sequence after the G-less cassette allowing full-length extension to +225. B) RNAP^{WT} and RNAP^{PM} were both initiated onto transcription templates and stalled at +125 by withholding of GTP (SM). GTP was then added, and samples stopped over a short time course to examine the extension of +125 RNAs by both RNAPs. RNAP^{WT} and RNAP^{PM} show minimal differences in extension from +125 to +225 with comparable speeds and pause positions.

determine if the polarity mutant RNAP is termination resistant. Similarly, no differences were observed in the susceptibility of $RNAP^{WT}$ or $RNAP^{PM}$ to termination by either Eta of FttA. Our *in vitro* evidence suggests that the mutant RNAP identified in our genetic screen is not simply a faster or pause-resistant RNAP and is not resistant to transcription termination. More investigation into the biochemical activities of this mutant RNAP will be required to fully determine reasons for phenotypic differences in strains harboring the mutant enzyme. Tracking the location of TECs genome wide through new techniques such as NET-seq may illuminate more subtle transcriptional differences of $RNAP^{PM}$.

3.2.7 Preliminary evidence suggests the FttA 'polarity mutant' can still efficiently couple to TECs through Spt4/5 and appears to out-perform wild-type FttA in *in vitro* transcription termination experiments.

Current models of FttA-mediated transcription suggest FttA recognizes ribosome-free nascent RNA protruding from TECs not coupled to translation, where it cleaves the RNA ~25nt from the 3' end and concurrently disrupts the TEC (10). FttA activity is non-competitive with actively elongating TECs, and is perturbed by G-rich templates, unless coupled to TECs by the universally conserved Spt4/5 complex (NusG in Bacteria; DSIF in Eukarya). It was plausible that many activities of FttA could be perturbed in the mutant form of FttA (G193S, F194L; $FttA^{PM}$) identified in our genetic screen that permitted strains harboring the mutation to survive. The first rationale for a variant FttA enzyme allowing readthrough of the polarity constructs was an abrogation to the proper coupling to the Spt4/5 complex required for efficient termination of actively elongating TECs without nucleic acid sequence requirements. This would potentially allow TECs uncoupled from ribosomes after our mutated stop codon in the upstream non-essential gene to reach the start codon of the *hisD*/*trpE* genes before being efficiently terminated by FttA. Once TECs have reached the start codon of the *hisD* or *trpE* genes,

ribosomes can re-initiate translation and once again protect TECs from FttA-mediated transcription termination. Secondly, disruption of the TEC-collapsing ability of the FttA mutant could potentially allow readthrough of the polarity constructs. The constructs contain relatively strong promoters, and thus many TECs are likely attempting to transcribe through the non-essential at a given moment. If the FttA variant cannot efficiently disrupt TECs before they reach the *hisD*/*trpE* start codon, then cells would become His/Trp prototrophic. If RNA cleavage is a prerequisite to transcription termination, disruption to RNA cleavage activity could also potentially allow readthrough of the polarity constructs. Both $FttA^{WT}$ and $FttA^{PM}$ were recombinantly expressed in *E.coli* and isolated for biochemical experimentation.
o characterize altered biochemical properties of the $FttA^{G193S/F194L}$. The overall efficiency of FttA-mediated termination between $FttA^{WT}$ and $FttA^{PM}$ has been examined on stalled TECs where surprisingly, $FttA^{PM}$ consistently outperformed $FttA^{WT}$. To test whether $FttA^{PM}$ is unable to properly couple to TECs via interactions with Spt4/5, the ability of $FttA^{PM}$ to terminate actively elongating TECs in the presence and absence of Spt4/5 was compared to $FttA^{WT}$. Stalled $TECs_{+125}$ were incubated in the presence of NTPs, FttA, and if necessary, Spt4/5 allowing kinetic competition between further transcription to +225 and an FttA-mediated termination event. The same pattern of activity and requirement of Spt4/5 coupling for FttA-mediated termination of actively elongating TECs is observed in reactions containing $FttA^{WT}$ $FttA^{PM}$. In fact, under current assay conditions, $FttA^{PM}$ has consistently displayed modestly but significantly higher transcription termination efficiencies overall. These results indicate that the mutant strains harboring the G193S/F194L mutation to FttA were not able to overcome our polarity-repressed constructs due to abnormal coupling between FttA and Spt4/5.

3.3 Conclusions and discussion

- **Mutations to genes of the normally Trp/His auxotrophic CM003 relay Trp/His prototrophy through polarity-repressive constructs.**
- **CM003 harboring only mutations to FttA identified from the genetic screen display significant readthrough of polarity-repressive constructs indicating FttA is responsible for the polar repression of transcription in Archaea.**
- **Preliminary evidence suggests the mutation to FttA is adjacent to a archaeal-eukaryotic conserved dimerization interface which could potentially reduce the efficiency of the polar repression of transcription by altered dimerization kinetics.**

Timely control of the transcription elongation/termination decision is necessary for cells to produce RNA transcripts of appropriate length (and thus function), but in operon-encoding prokaryotes transcription termination regulation can help cells orchestrate appropriate transcriptional responses to internal and environmental stimuli. In Bacteria, the transcription termination factor Rho recognizes the uncoupling of transcription and translation to terminate transcription genome wide. In Archaea, this role is fulfilled by FttA. The prerequisites for Rho- and FttA-mediated transcription termination suggest that both act any time translation stalls and the nascent RNA is susceptible- not just at a stop codon. In operons, if such an event occurs in upstream genes, transcription can be terminated before expression of downstream genes, i.e. polarity. The polar repression of transcription (and global non-canonical transcription termination) can thus be induced by multiple events: i) amino acid starvation can stall ribosomes leading to transcription termination when protein synthesis is unfavorable, ii) attenuation events (i.e. the well-studied *E.coli* Trp operon) allow transcription termination of specific gene-sets surplus to current cellular demands, and iii) transcription termination of TECs transcribing mutated genes containing premature stop codons (or transcriptional infidelity leading to a stop

codon in the RNA specifically) that uncouple transcription and translation early. It is well documented that Rho is the factor responsible for the polar repression of transcription in Bacteria, but no concrete evidence was available implicating FttA as the archaeal polarity factor. The novel genetic screen here provides substantial evidence that FttA is responsible for the polar repression of transcription in Archaea and identifies previously unrevealed amino acid residues within FttA which, when altered, disrupt polarity in archaeal cells.

From trillions of cells originating from 100 individual CM003 cultures, our screen identified only ten CM003 starter cultures which produced colonies when challenged, indicative of the high unlikelihood of a functional mutation allowing transcriptional readthrough of our genetically encoded polarity constructs. One colony (to avoid sister-colonies) from each colony-forming plate were thereafter were proliferated as 'mutant strains'. Subsequent genetic analysis of surviving strains revealed on average 20 identified point mutations per strain, ranging from 4-37 per individual strain. None of the surviving mutant strains appeared to have any changes to the nucleotide sequences at any region of any of the polarity constructs, suggesting they still operate as intended in these strains. Mutant strains contained mutations in a broad range of coding genes, and seven out of ten had no perturbations to an obvious primary player in the polar repression of transcription. These strains contained some mutations to coding regions of ribosomal subunits, which could affect uncoupling activities and thus the competition between ribosomes and FttA, or have ribosomes readthrough stop-codons at an increased rate. Some strains contained mutations to tRNA coding regions, which may have an effect on codon usage at a stop codon, i.e. mutant 66 contained two mutations impacting the methionine initiator tRNA, which could feasibly adjust ribosomal initiation events to prevent FttA activity before re-initiating after a stop-codon at our polarity loci. Many mutant strains contained mutations to coding regions of 'hypothetical proteins', archaeal-specific proteins which are poorly annotated and as of yet have an unknown function. While it is possible some of these proteins play a role in

polarity, none at first glance appeared to have the attributes of transcription factors, but this should be confirmed. At this time, RT-qPCR experiments have not confirmed that all mutant strains have increased readthrough of the polarity constructs- only strains with mutations resulting in functional changes to FttA and RNAP, hypothesized primary players in the polar repression of transcription, have been tested so far. It will be interesting to gather RT-qPCR results from all ten mutant strains to investigate if these mutant cells can pass our genetic screen independent of transcription through the polarity loci.

Excitingly, three of our mutant strains contained mutations impacting the RNAP beta-subunit of FttA- two protein factors hypothesized to be involved in the polar repression of transcription before the genetic screen was conducted. Analysis of the specific changes these mutations relayed at the protein level ($FttA^{G193S/F194L}$ and $RNAP^{\beta\text{-}W463A}$) reveals mutations to both factors are in extremely conserved regions which are likely critical for proper function. It is likely that the mutations do not severely abrogate the function of FttA or RNAP at the protein level, as both enzymes activities are critical and essential for cellular health. Although preliminary evidence suggests that the mutations impacting RNAP do not impact the speed, pause preferences, or termination susceptibility of TECs *in vitro*, Initial work thus far has focused on biochemically characterizing the effects of the mutations to FttA identified from the genetic screen. Ongoing experiments currently suggest no abrogation to the ability of $FttA^{G193S/F194:}$ to cleave RNA, terminate stalled complexes, or couple to and terminate actively elongating TECs via Spt4/5. Our lab recently entered into collaboration with Dr. Richard Ebright at Rutgers University, in an attempt to visualize an FttA-Spt4/5-TEC structure by cryo-EM. Initial images from that project have recently been obtained, and while the data is not finalized and a more efficient approach is currently underway to increase the resolution of the final image, the current structure is extremely valuable in determining the effect of the G193S/F194L on the activity of FttA *in vivo* (Figure 3.8.A). The structure, assembled with a catalytically inactive mutant of FttA,

reveals FttA acts as a dimer during FttA-mediated transcription. One homodimer makes predominant contacts with Spt4/5, while the second homodimer makes predominant interactions with the stalk domain of RNAP, composed of RNAP subunits E and F. The stalk domain of RNAP and Spt4/5 thus direct the correct conformation of the FttA dimer to both cleave RNA and terminate the TEC. From the initial cryo-EM images, the dimerization interface of FttA was determined to likely be at the very C-terminus of the protein where the last ~7 amino acids are highly conserved within the archaeal Domain. A model of the dimerization of FttA was made using the cryo-EM structure as a guide, and reveals that the mutations discovered in our polarity genetic screen affect residues directly adjacent to the dimerization interface (Figure 3.8.B). It is plausible that our mutations result in small perturbations to local protein folding, thus effecting the proper dimerization between two FttA subunits. Preliminary data was generated by incubating either $FttA^{WT}$ or $FttA^{PM}$ in increasing salt concentrations followed by native PAGE to assess the efficiency of dimerization of each FttA variant under different salt conditions (Figure 3.8.C). Surprisingly, initial experiments suggest that $FttA^{PM}$ dimerizes *more* efficiently than $FttA^{WT}$, as evidenced by the disappearance of bands corresponding to monomeric FttA when the enzyme contains the G193S/F194L mutation. This could explain the slight ~3-4X increase in termination activity of the FttA mutant in *in vitro* transcription experiments, but does not help rationalize how strain CM005, containing the FttA polarity mutant in a CM003 background, produces more full-length transcripts as a result of readthrough of our polarity constructs. FttA, if terminating transcription more efficiently, could potentially autoregulate its own transcription, leading to a lower FttA copy number in the cell and thus more transcriptional readthrough of our polarity constructs. The increased dimerization could effect FttA turnover after a transcription termination event, remaining engaged with a collapsed TEC and/or Spt4/5. Alternatively, there could be one or more unknown factors which regulate the interplay between FttA, Spt4/5, and TECs during the transcription cycle and interactions with FttA are changed in the polarity mutant variant. It is equally plausible that the dimerization is not effecting polarity, and the reduced

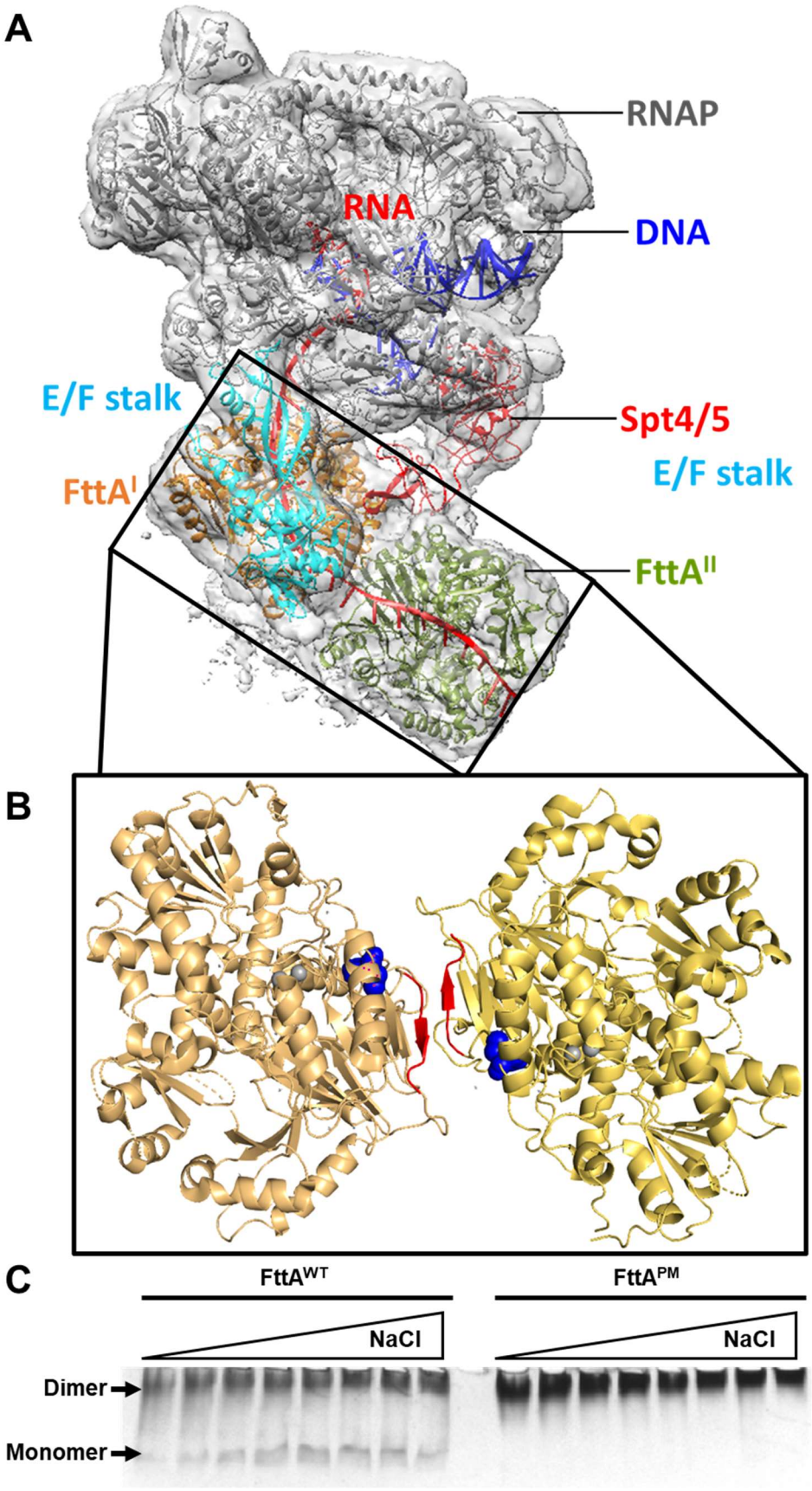

A
RNAP
RNA
DNA
E/F stalk
Spt4/5
E/F stalk
FttA^{I}
FttA^{II}
B
C
FttA^{WT}
FttA^{PM}
NaCl
NaCl
Dimer
Monomer

Figure 3.8: CryoEM structure of $(FttA)_2$-(Spt4/5)-TEC reveals a dimerization interface which may be perturbed by G193S/F194L mutations
A) Initial cyro-EM structure of $(FttA)_2$-(Spt4/5)-TEC obtained in collaboration with the Ebright lab. Two FttA molecules (orange, gold) are required for FttA-mediated termination, each directed by interactions with the RNAP stalk domain (cyan) and Spt4/5 (red). B) Model of the FttA dimer interface based on the cryo-EM structure. A predicted archaeal-conserved C-terminal dimerization interface (red) is located adjacent to the polarity mutations identified in this study (blue). C) Native PAGE of $FttA^{WT}$ and $FttA^{PM}$ suggests $FttA^{PM}$ dimerizes more efficiently *in vitro* evidenced by the lack of monomeric FttA in these conditions.

efficiency of the polar repression of transcription exhibited by the mutant FttA enzyme is a result of a yet to be elucidated mechanism. Alternate scenarios, such as reduced expression or higher turnover of FttA at the RNA or protein level could explain the reduction in the polar repression of transcription in strains encoding the mutant FttA. Additional experiments, such as Western blotting and RT-qPCR, will be critical in determining any changes in proper FttA expression due to the observed mutations. Further, it is important to determine whether the reduction in polar repression of transcription is simply a result of perturbed transcription termination globally at 3' end of genes by the mutant FttA. Performing experiments to assess length of RNA 3' ends globally will be instrumental in establishing whether or not the activity of the mutant FttA is perturbed specifically in the context of polarity.

Overall, our genetic screen successfully identified functionally relevant mutations to both FttA and RNAP, which when reconstituted in a clean parent strain background, lead to significantly increased transcriptional readthrough of our genetically encoded polarity constructs. The increased readthrough of the polarity constructs is presumably due to impacted polar repression of transcription. While the biochemical explanations underlying the passage of mutant strains through the genetic screen and strains CM005/CM007 producing more full-length transcripts through the polarity constructs remain incompletely resolved, the current work provides strong evidence FttA is the factor responsible for the polar repression of transcription in Archaea.

3.4 Methods

Strain Construction

Strain CM003 was constructed using two successions of the same transformation, recombination selection, and confirmation techniques as previously described (39) First, plasmid pCM0020 was transformed into strain TS559 yielding strain CM002. CM002 was then

transformed with pCM0019 yielding CM003. Strain CM007, harboring the RNAP beta subunit (TK1083) W463A mutation identified in this study, was constructed from plasmid CM0025 for purification of the mutant RNAP. Sequencing of PCR products obtained from the appropriate loci, and whole-genome sequencing (details below) confirmed correct strain construction.

Passage of strain CM003 through genetic screen

Individual colonies of CM003 were anaerobically isolated by spotting on ASW-YT-Gelzan plates at 85°C as previously described (40). Individual colonies were anaerobically inoculated into 5mL of ASW-YT rich media and grown 16 hours at 85°C. 50uL of the resultant culture was then passaged into 5mL minimal liquid media composed of ASW and 20 amino acid mix and grown an additional 16 hours anaerobically at 85°C. The passaged 5mL culture was then centrifuged anaerobically (5000xg, 10mins) and cells were resuspended in 200uL 0.8X ASW. Resuspended cells were plated on minimal plates composed of ASW and 18 amino acid mix (lacking histidine and tryptophan), and incubated anaerobically at 85°C for 96 hours. One colony from each plate exhibiting colony development was individually inoculated into 5mL ASW-YT media and assigned a 'mutant number'. Cultures were grown 16 hours anaerobically at 85°C and genomic DNA was isolated for genomic variant analysis.

Challenge plates

T. kodakarensis strain KOD1 (wild-type), CM003, and mutant strains were individually plated on minimal media containing ASW and 20 amino acid mix and grown anaerobically 24 hours at 85°C. Individual colonies of each strain were resuspended in 50uL 0.8XASW and subjected to a 10X dilution series with 0.8X ASW in a 96-well plate. Dilution series of each strain were then plated on the same plate of minimal media containing ASW and 18 amino acid mix (lacking histidine and tryptophan) and grown anaerobically 85°C for 24 hours. The resulting plate growth

was transferred to a methanol soaked-PVDF membrane at room temperature for 15 minutes and stained with Coomassie dye. Membranes were visualized with a BioRad GelDoc+.

Bioinformatic pipeline development and mutational analysis of surviving mutant strains

Genomic DNA samples of CM003 and mutant strains were isolated, and prepared for Illumina whole genome sequencing using the NEBNext Multiplex Oligos for Illumina kit (New England Biolabs, E7335S). Samples were read on an Illumina NextSeq. DNA-seq reads were first quality-checked (FastQC, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and then sequences representing indexing barcodes and adapters were removed from the reads (Trimmomatic (41)). Trimmed reads were aligned to a CM003 reference genome (Bowtie 2 (42)) to obtain genome coordinates before subsequent local realignment and identification of insertions, deletions, and point mutations (The Genome Analysis Toolkit [GATK] (43)). Tabulated calls of variants were manually mapped to annotated genome features and genes of *T. kodakarensis*. Silent mutations were discounted, and amino acid substitutions caused by each insertion, deletion, or point mutation in protein coding genes was determined and tabulated.

RNA extraction from CM003 and 'polarity' mutant strains for RT-qPCR

Cultures of CM003, CM005, CM007, and Mutant 95 were anaerobically grown to an OD_{600} of 0.5 overnight at 85°C in ASW-YT-Pyruvate. 10 mL of cells were pelleted and resuspended in 1 mL of TRIzol reagent (Invitrogen, Inc.) and 200 µL chloroform to extract ribonucleic acids. 400 µL aqueous layer from the TRIzol extraction was added to 500 µL 100% isopropanol to precipitate ribonucleic acids and pelleted by centrifugation 12000 x *g,* 10mins, 4°C. The RNA pellet was washed in 75% EtOH and re-pelleted by centrifugation. The EtOH was removed and the pellet was air dried for 30 minutes at room temperature, followed by resuspending in 50 µL of 1X DNase I Buffer (NEB). 2 µL of 1U/µL DNase I (NEB) was added to RNA samples and

incubated 90 mins @ 37°C. 2 µL 0.5 M EDTA was then added and the reaction heated at 80°C for 10 minutes. 1 mL of TRIzol reagent was then added to the reaction along with 200 µL chloroform to re-extract ribonucleic acids. 400 µL aqueous layer from the TRIzol extraction was added to 500 µL 100% isopropanol to precipitate ribonucleic acids and pelleted by centrifugation 12000 x *g,* 10mins, 4°C. The RNA pellet was washed in 75% EtOH and re-pelleted by centrifugation. The EtOH was removed and the pellet was air dried for 30 minutes at room temperature, followed by resuspension in 50 µL RNase-free water. DNase treated RNA samples were quantified using the Qubit BR RNA Assay kit (Thermo Fisher).

Reverse transcription of RNA isolated from CM003 and 'polarity' strains

500ng of each RNA sample was added to 1.2 µL 10 mM dNTPs, 1.2 µL 1 µM primer (J1 or T3 primers specific to the *hisD* polarity construct or the *trpE* polarity construct; see Figure 3.6) and brought to 15.6 µL with RNase-free water. Primers were annealed to the RNA template by heating to 65°C for 10 minutes and slowly cooling to room temperature. Reactions were then brought to 24 µL containing 5mM DTT, 1X SuperScript IV Buffer (Invitrogen, Inc.), and 5U SuperScript IV Reverse Transcriptase (Invitrogen, Inc.). Reactions were incubated at 55°C for 10 minutes to synthesize cDNA and 85°C for 10 minutes to denature the reverse transcriptase enzyme. Samples were used directly in qPCR reactions

qPCR reactions on cDNA templates

1.5 µL of cDNA template was used in 10 µL qPCR reactions containing 0.25 µM of primers J1F and J1R and 1X SYBR-based QuantiNova qPCR master-mix (Qiagen). Reactions were assembled in 96-well clear PCR plates (BioRAD) and sealed with Microseal 'B' sealing film (BioRAD). Reactions were loaded for qPCR on a C1000 Thermal Cycler (BioRAD) fitted with the CFX96 Real-Time System (BioRAD) and incubated at 95°C for 3 minutes, followed by 40 cycles

of 95°C for 15 seconds and 60 °C for 15 seconds. After each cycle, tube fluorescence was read at 480nm, indicative of SYBR green fluorescence due to presence of double-stranded DNA. Curves were analyzed and Ct values determined for each reaction using the CFX96 software (BioRAD) and tabulated in Microsoft Excel.

Protein purifications

RNAP^{PM} (RpoL-HA-6xHis) was purified in native form from *T. kodakarensis* strain CM007 as previously described. FttA^{WT} and FttA^{PM} were expressed and purified from Rosetta 2 (DE3) cells harboring plasmid pCM025 (pQE-80L-FttA^{PM}) cultured in LB supplemented with 3% D-sorbitol, 100 µg/mL ampicillin, and 34 µg/mL chloramphenicol. 0.2 mM isopropyl β-D-1-thiogalactopyranoside was used to induce expression from the plasmid at an OD_{600} of 0.5, and induced cultures were grown overnight at 22°C. Biomass was pelleted and sonically lysed in 15 mM Tris-HCl pH 8.0, 10 mM MgCl2, 1 M NaCl, 2 mM β-mercaptoethanol. Material was centrifuged at 22,000 x g for 15 mins at 4°C to produce a clarified lysate and debris pellet. The clarified cell lysate was heat treated at 85°C to denature residual *E.coli* proteins and again clarified by centrifugation at 22,000 x g for 15 mins at 4°C. The resulting supernatant was step-dialyzed into 15 mM Tris-HCl pH 8.0, 10mM MgCl2, 300 mM NaCl, 2 mM β-mercaptoethanol and loaded onto a 5mL HiTrap Heparin column (Cytiva) using an AKTA Pure FPLC system (GE Healthcare). Proteins were eluted over a 40mL gradient to 15 mM Tris-HCl pH 8.0, 10 mM MgCl2, 600 mM NaCl, 2 mM β-mercaptoethanol. Fractions containing FttA^{WT} or FttA^{PM} were identified by SDS-PAGE, pooled, and dialyzed into storage buffer, 15 mM Tris-HCl pH 8.0, 10 mM MgCl2, 250 mM KCl, 2 mM β-mercaptoethanol, 50% glycerol. Protein samples were concentrated by size-exclusion centrifugation (Vivaspin 50kDa MWCO). Resulting enzymes were quantified using a Qubit™ Protein Assay Kit (Invitrogen).

In vitro transcription assays

Assembly of pre-initiation complexes and elongation to generate $TECs_{+125}$ were performed as described previously (63). $TECs_{+125}$ were captured using Nickel-coated magnetic beads (ThermoFisher), washed 3x in 20 mM Tris·HCl pH 8.0, 0.1 mM EDTA, 250 mM KCl, 4 mM MgCl2, 20 µg/mL BSA, and resuspended 250 mM KCl, 20 mM Tris·HCl pH 8.0, 10 mM MgCl2, and 2 mM DTT. 20 µL$TECs_{+125}$ were incubated in the presence of 1 mM NTPs at 85°C for 1, 2, 3, 4, or 5 minutes and stopped by addition of 100 µL 1.2 X STOP buffer (600 mM Tris-HCl pH 8.0, 12 mM EDTA) on ice. Radiolabeled RNAs were extracted by addition of equal volume P/C/I. Nucleic acids were precipitated with ethanol and resuspended in 4 µL formamide loading dye before separation by electrophoresis through a 15% polyacrylamide/8 M urea denaturing gel. Radiolabeled RNAs were detected by exposure to a phosphorimaging screen (GE Healthcare) and analyzed using GE ImageQuant 5.2.

REFERENCES

1. N. J. Proudfoot, Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352**, aad9926 (2016).

2. I. Gusarov, E. Nudler, The mechanism of intrinsic transcription termination. *Molecular Cell* **3**, 495–504 (1999).

3. T. J. Santangelo, L. Cubonová, K. M. Skinner, J. N. Reeve, Archaeal intrinsic transcription termination in vivo. *Journal of Bacteriology* **191**, 7102–7108 (2009).

4. J. C. Mcdowell, J. W. Roberts, D. J. Jin, C. Gross, Determination of intrinsic transcription termination efficiency by RNA polymerase elongation rate. *Science* **266**, 822–825 (1994).

5. T. J. Santangelo, J. N. Reeve, Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *Journal of Molecular Biology* **355**, 196–210 (2006).

6. J. E. Walker, O. Luyties, T. J. Santangelo, Factor-dependent archaeal transcription termination. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E6767–E6773 (2017).

7. J. W. Roberts, Termination factor for RNA synthesis. *Nature* **224**, 1168–1174 (1969).

8. H. E. Mischo, *et al.*, Cell-Cycle Modulation of Transcription Termination Factor Sen1. *Molecular Cell* **70**, 312-326.e7 (2018).

9. Y. Jiang, M. Liu, C. A. Spencer, D. H. Price, Involvement of transcription termination factor 2 in mitotic repression of transcription elongation. *Molecular Cell* **14**, 375–386 (2004).

10. T. J. Sanders, *et al.*, FttA is a CPSF73 homologue that terminates transcription in Archaea. *Nature Microbiology* **5**, 545–553 (2020).

11. J. Shi, *et al.*, Structural basis of Mfd-dependent transcription termination. *Nucleic Acids Research* **48**, 11762–11772 (2020).

12. J. D. Eaton, S. West, An end in sight? Xrn2 and transcriptional termination by RNA polymerase II. *Transcription* **9**, 321–326 (2018).

13. S. L. French, T. J. Santangelo, A. L. Beyer, J. N. Reeve, Transcription and translation are coupled in Archaea. *Molecular Biology and Evolution* **24**, 893–895 (2007).

14. C. Wang, *et al.*, Structural basis of transcription-translation coupling. *Science* **369**, 1359–1365 (2020).

15. M. S. Ciampi, Rho-dependent terminators and transcription termination. *Microbiology* **152**, 2515–2528 (2006).

16. V. Valabhoju, S. Agrawal, R. Sen, Molecular basis of NusG-mediated regulation of Rho-dependent transcription termination in bacteria. *Journal of Biological Chemistry* **291**, 22386–22403 (2016).

17. C. M. Hart, J. W. Roberts, Rho-dependent transcription termination. Characterization of the requirement for cytidine in the nascent transcript. *Journal of Biological Chemistry* **266**, 24140–24148 (1991).

18. C. M. Burns, L. v. Richardson, J. P. Richardson, Combinatorial effects of NusA and NusG on transcription elongation and Rho-dependent termination in Escherichia coli. *Journal of Molecular Biology* **278**, 307–316 (1998).

19. P. Mitra, G. Ghosh, M. Hafeezunnisa, R. Sen, Rho Protein: Roles and Mechanisms. *Annual Review of Microbiology* **71**, 687–709 (2017).

20. V. Epshtein, D. Dutta, J. Wade, E. Nudler, An allosteric mechanism of Rho-dependent transcription termination. *Nature* **463**, 245–249 (2010).

21. A. Q. Zhu, P. H. von Hippel, Rho-dependent termination within the trp t' Terminator. II. Effects of kinetic competition and Rho Processivity. *Biochemistry* **37**, 11215–11222 (1998).

22. C. R. Mandel, *et al.*, Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature* **444**, 953–956 (2006).

23. N. G. Kolev, T. A. Yario, E. Benson, J. A. Steitz, Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3′-end maturation. *EMBO Reports* **9**, 1013–1018 (2008).

24. M. Clerici, M. Faini, L. M. Muckenfuss, R. Aebersold, M. Jinek, Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex. *Nature Structural and Molecular Biology* **25**, 135–138 (2018).

25. F. Werner, Structural evolution of multisubunit RNA polymerases. *Trends in Microbiology* **16**, 247–250 (2008).

26. J. D. Eaton, *et al.*, Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Genes and Development* **32**, 127–139 (2018).

27. M. Clerici, M. Faini, R. Aebersold, M. Jinek, Structural insights into the assembly and polya signal recognition mechanism of the human CPSF complex. *eLife* **6** (2017).

28. C. R. Mandel, *et al.*, Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature* **444**, 953–956 (2006).

29. C. L. Turnbough, Regulation of Bacterial Gene Expression by Transcription Attenuation. *Microbiology and Molecular Biology Reviews* **83** (2019).

30. C. Yanofsky, Attenuation in the control of expression of bacterial operons. *Nature* **289**, 751–758 (1981).

31. T. J. Santangelo, *et al.*, Polarity in archaeal operon transcription in Thermococcus kodakaraensis. *Journal of Bacteriology* **190**, 2244–2248 (2008).

32. Y. Li, S. Altman, Polarity effects in the lactose operon of Escherichia coli. *Journal of Molecular Biology* **339**, 31–39 (2004).

33. S. Saxena, J. Gowrishankar, Modulation of Rho-dependent transcription termination in Escherichia coli by the H-NS family of proteins. *Journal of Bacteriology* **193**, 3832–3841 (2011).

34. M. H. de Smit, P. W. G. Verlaan, J. van Duin, C. W. A. Pleij, Intracistronic transcriptional polarity enhances translational repression: A new role for Rho. *Molecular Microbiology* **69**, 1278–1289 (2008).

35. D. W. Grogan, G. T. Carver, J. W. Drake, Genetic fidelity under harsh conditions: Analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 7928–7933 (2001).

36. B. Mir-Montazeri, M. Ammelburg, D. Forouzan, A. N. Lupas, M. D. Hartmann, Crystal structure of a dimeric archaeal Cleavage and Polyadenylation Specificity Factor. *Journal of Structural Biology* **173**, 191–195 (2011).

37. S. H. Jun, *et al.*, Direct binding of TFEα opens DNA binding cleft of RNA polymerase. *Nature Communications* **11** (2020).

38. P. P. Hein, R. Landick, The bridge helix coordinates movements of modules in RNA polymerase. *BMC Biology* **8** (2010).

39. T. H. Hileman, T. J. Santangelo, Genetics techniques for Thermococcus kodakarensis. *Frontiers in Microbiology* **3** (2012).

40. A. Gehring, T. Sanders, T. J. Santangelo, Markerless Gene Editing in the Hyperthermophilic Archaeon Thermococcus kodakarensis. *BIO-PROTOCOL* **7**, e2604 (2017).

41. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

42. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).

43. A. McKenna, *et al.*, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).

CHAPTER 4

INVESTIGATIONS INTO ARCHAEAL TRANSCRIPTION COUPLED DNA REPAIR[2]

4.1 Introduction

Aside from the intrinsic instability of DNA, genomes are constantly threatened by a plethora of endogenous and exogenous insults. Left unrepaired, DNA damage increases mutation rates, causing adverse effects on cellular health, with often drastic consequences to cellular and organismal fitness (1, 2). Endogenous damage has many sources: genomic material can spontaneously undergo base hydrolysis or deaminate, and torsional stresses brought about by information processing systems can bring about genomic instability (2, 3). Cellular machineries will occasionally incorporate mismatch errors or ribonucleotides (rNMPs) into newly synthesized DNA (4), and many metabolic enzymes produce reactive oxygen species (ROS) which may oxidize DNA bases (5). Cells must also tolerate exogenous sources of DNA damage which vary depending on the external environment. Chemical crosslinkers, environmentally-generated ROS, ultraviolet light and ionizing radiations from within or which penetrate the atmosphere all have mutagenic effects on DNA (6–8).

Many archaea thrive within niche and extreme environments which can increase rates of DNA damage. Many halophilic archaea, for example, thrive in shallow salt plains and endure extreme levels of UV radiation (9), while some hyperthermophilic species persist at temperatures that would easily denature purified DNA (10, 11), and yet the presumed increased rates of deamination, depurination and oxidation are somehow tolerated (12, 13). In

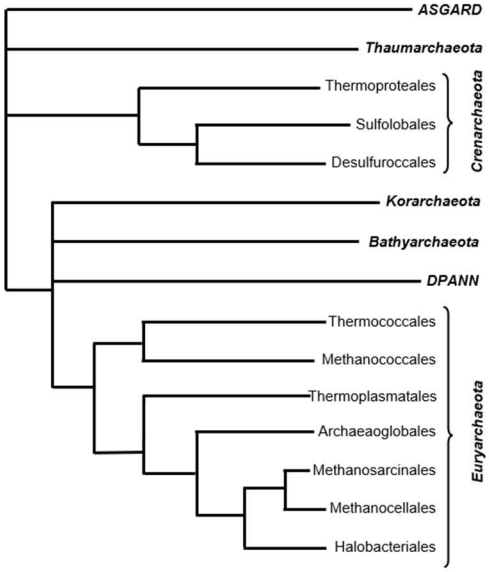
[2]Portions of the introduction to this section are excerpts from the following publication: Marshall, C. J. and Santangelo, T. J. (2020) 'Archaeal DNA repair mechanisms', *Biomolecules*. MDPI AG, pp. 1–23.

addition to growth in the extremes, many archaeal species maintain genomic stability levels to display similar rates of spontaneous mutation to mesophilic prokaryotes such as Escherichia coli (14–16). It is perhaps surprising then that only DNA repair strategies conserved in Bacteria or Eukarya have yet been recognized in Archaea. While DNA repair strategies vary by clade, no new uniquely archaeal DNA repair pathway has been described (Table 4.1). Insight into how Archaea detect and convert damaged DNA bases into repairable substrates has begun to reveal how genomic integrity is preserved *in extremis*.

A multitude of strategies to identify modified nucleotides or damaged DNA structures and initiating repair are encoded in most genomes, with processes for recognition and repair perhaps best studied within mesophilic bacteria and eukarya. Some DNA damage repair can be directly reversed, i.e. photoreactivation of thymine-thymine dimers by photolyases and repair of methylation adducts by alkyltransferases such as AlkB (17, 18). However, DNA repair more commonly involves pathways which require several specialized enzymes through steps of damage recognition, initiation of repair and final polymerization/ligation of resynthesized DNA. Collectively, the cycle of recognition-, initiation- and ligation-based DNA repair dominates the conserved DNA repair pathways that account for the majority of DNA repair, be it double-strand break (DSB) repair, mismatch repair (MMR), ribonucleotide excision repair (RER), or base excision repair (BER)- pathways all conserved across the three Domains (19–26).

Some DNA damages, i.e. UV induced photoproducts, result in a distortion of the dsDNA helix which has stalling effects on critical processes such as replication and transcription(27). It is unclear how such damages are resolved in Archaea, but in Bacteria and Eukarya DNA repair mechanisms have evolved to detect the general distortions of the DNA backbone rather than the actual modification, which allows detection a broad range of DNA lesions. Global genomic nucleotide excision repair (GG-NER) in Bacteria and Eukarya relies on enzymes to recognize the ‘bulky lesion’ and direct strand specific cuts on the damaged DNA strand. The DNA damage, now between two nicks, is thus primed for ‘excision’ from the DNA allowing

Table 4.1: Predicted distribution of pathway-specific archaeal DNA repair proteins by clade, according to KEGG (Kyoto Encyclopedia of Genes and Genomes) orthologies: Many pathways appear conserved, with most variation found in distribution of mismatch repair (MMR) and nucleotide excision repair (NER) proteins.



Mismatch repair	DSB Repair		Ribonucleotide Excision Repair	Base Excision Repair	Nucleotide Excision Repair*
	NHEJ/MMEJ	HR			
MutS, MutL, EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
EndoMS	Ku	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
MutL-MutS	?	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
MutL-MutS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
MutS, MutH, EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrB, UvrD, XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABC, XPB, XPD, XPF
MutS, MutH	Ku	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD
?	Ku	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
MutS, EndoMS	Ku	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
MutS, EndoMS	Ku	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPD, XPF
MutL/S/H, EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPF

resynthesis from the undamaged strand, and nick ligation to complete repair. In Bacteria, NER is mediated by the UvrA2/B/C/D enzymes. Helix distortions are first recognized by the UvrA dimer and damage is subsequently verified by UvrB (28). The activity of UvrB converts general strand distortion detection by UvrA into damage and strand specific detection, which directs the nuclease activity of UvrC either side of the DNA damage (29, 30). The UvrD helicase can then excise the damage containing strand from the genome allowing for strand resynthesis by DNA polymerase I and nick sealing by DNA ligase (31). The core steps of eukaryotic NER resemble a slightly more sophisticated bacterial NER (Figure 4.1). DNA helix distortions are first recognized by the XPC repair protein, and then damage is verified by the XPA protein to form a pre-incision complex. Helicases XPB and XPD then separate DNA strands at the site of damage- and the orientation of the resulting complex allows strand-specific cuts by XPG and XPF on either side of the site of DNA damage (32). The damage containing strand is excised in complex with TFIIH (33–35), allowing strand resynthesis by DNA polymerase δ or ε, and nick sealing by DNA ligase I (27).

In Bacteria and Eukarya, NER can alternatively be initiated by recognition of transcription elongation complexes (TECs) which stall upon DNA lesions entering the active site of RNA polymerase (RNAP) during transcription- a process termed transcription coupled DNA repair (TCR) (36). Utilizing actively transcribing RNAPs to sense DNA damage offers an evolutionary advantage as actively transcribed regions of the genome are actively monitored for lesions. Akin to global NER, TCR has yet to be described in Archaea but current evidence suggests it is an active pathway in some clades. While studies in crenarchaea have revealed no significant change in DNA repair of transcribed versus non-transcribed strands (37), euryarchaeal species have displayed preferential repair of transcribed DNA strands- a hallmark of TCR (38). Additionally, the archaeal RNAP- which closely resembles eukaryotic RNAPII- has been shown to stall specifically at template strand DNA damage (39).

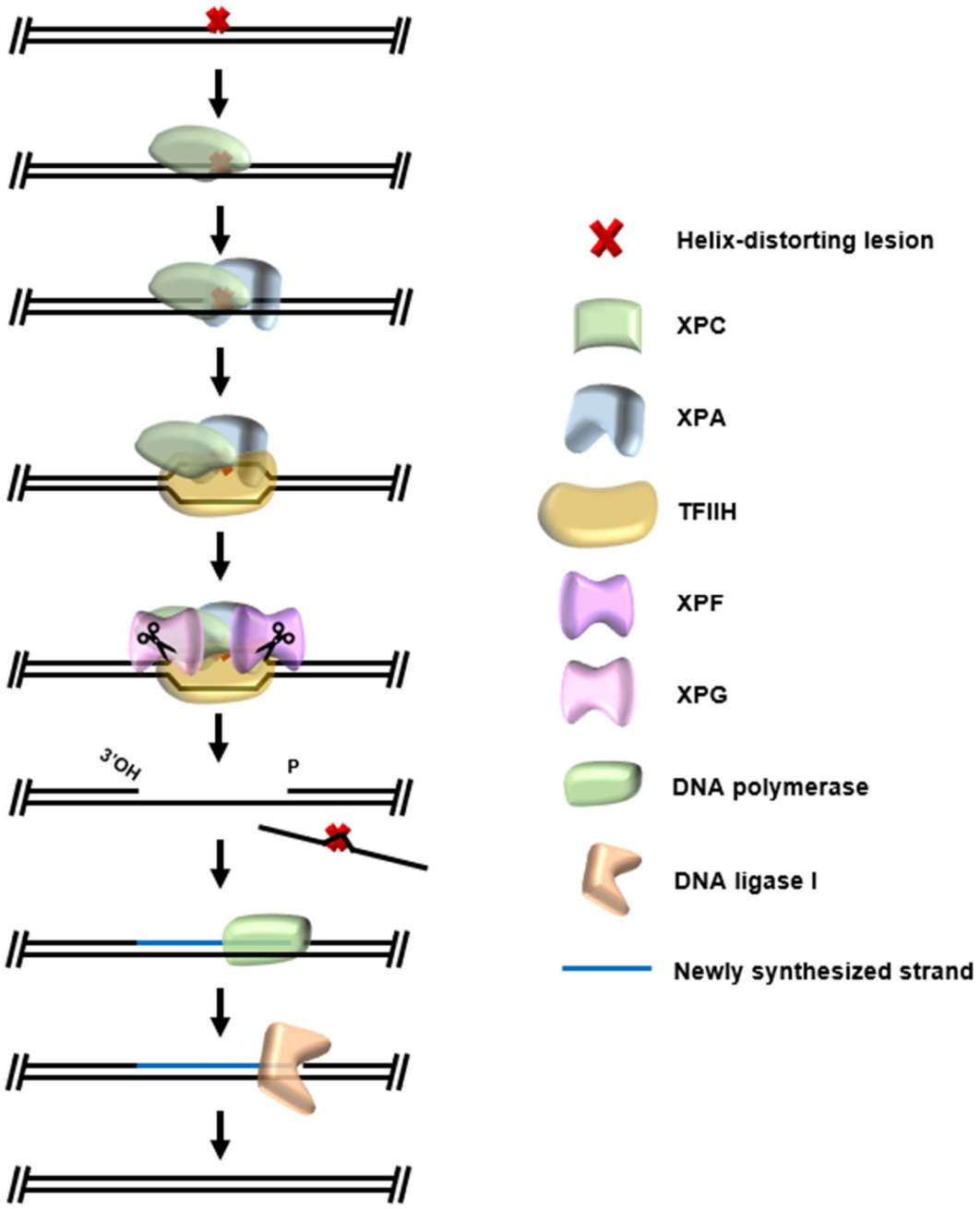


Figure 4.1: Eukaryotic global genomic nucleotide excision repair. DNA damages which distort the DNA double helix are recognized by XPC, which recruits the damage recognition XPA and TFIIH complex. Components of the TFIIH complex melt strands of DNA around a verified DNA lesion, allowing cuts of the damaged strand by XPG and XPF. The TFIIH complex uses helicase activity to “excise” the damaged strand, allowing conclusion of repair by DNA polymerase and DNA ligase I.

In eukaryotes, the CSB protein acts as the transcription repair coupling factor (TRCF)- initially recognizing stalled TECs and allowing localization of TFIIH and other NER enzymes directly to the site of damage (40–42). In Bacteria, the transcription termination factor Mfd acts as the TRCF- simultaneously recruiting the Uvr family of NER enzymes and terminating transcription to prevent formation of mutant transcripts (42–44). There are no homologs of either CSB of Mfd found in the archaeal Domain, suggesting a potential archaeal TCR pathway- and TRCF- evolved separately. Recently, the first enzyme with transcription termination activity was reported in Archaea- euryarchaeal termination activity (Eta)- and is intimately linked with other nucleic acid metabolic pathways and is a candidate for acting as the archaeal TRCF (45). Euryarchaeal termination activity (Eta) requires DNA sequences upstream of RNAP, aids backtracked RNAPs, is ATP-dependent and non-competitive with elongation- all attributes shared with the bacterial TRCF Mfd (46). The eukaryotic TRCF, CSB, also requires DNA sequences upstream of a stalled RNAP. Mfd catches up to backtracked or stalled polymerases by 'autonomously' patrolling DNA upstream of TECs (47).

Some Archaea encode homologs of bacterial Uvr proteins, but the majority encode homologs of critical eukaryotic NER proteins- in particular helicases XPB/XPD and endonuclease XPF (Table 4.1). No NER pathway (GG- or TC-), however, has yet been explicitly defined in Archaea, with research focusing on drawing parallels from individual enzymes conserved between Eukarya and Archaea. Such enzymes tend to exist outside the context of a multi-protein complex, allowing for ease of purification and crystallization. For example, independent structures of archaeal XPD, normally a component of the multienzyme TFIIH complex in eukaryotes, from *Thermoplasma acidophilum* and *Sulfolobus acidocaldarius* revealed a distinct 4 domain structure. Amino acids which when mutated disrupt the role of human XPD in NER but not transcription initiation (and cause the disease Xeroderma Pigmentosum) could be directly mapped to functionally critical sites of the archaeal structures (35).

While research into archaeal XP homologs has been structurally fruitful, establishing the NER pathway in Archaea has remained challenging and elusive. Perturbations in the UvrA, UvrB, and UvrC homologs found in *Halobacterium* resulted in almost total loss of resistance to UV exposure- but it remains unseen if these homologs function in a recognized NER pathway, and the Uvr proteins are only found in a minority of Archaea (48). Conversely, deletions of XPB, XPD, and XPF from *Thermococcus kodakarensis* resulted in only slight sensitivity to moderate doses of UV irradiation (49), suggesting these enzymes are involved in- but not required for- the UV induced damage response. Additional factors could potentially play a role in archaeal NER, and in some cases the eukaryotic-like NER enzymes are paired with auxiliary nucleases. XPB helicase is sometimes found encoded in an operon with a nuclease named Bax1 and these enzyme act in concert to open a DNA bubble and make cuts (50). In many XPF encoding species, the 3'-5' exonuclease HAN is often encoded, potentially recapitulating *in vitro* experiments where XPF and HAN form a functional nuclease complex (51). Recent biochemical examinations of archaeal XPF have investigated the enzyme in the context of replication restart and Holliday junction formation (52), but it is possible that XPF performs multiple functions within the cell- including one in an NER pathway.

NER in Archaea has so far proven elusive. The diverse patchwork of NER enzymes encoded in the many clades of Archaea suggests that no one NER pathway dominates across the Domain, or that an alternative mechanism is used for repair of bulky DNA lesions. Differing suites of NER enzymes across clades has led to conflicting limited studies into archaeal TC-NER, with some suggesting no TC-NER pathway exists in crenarchaea (37), and some suggesting preferential repair of actively transcribed genes exists in euryarchaea (38). To date, no genome-wide approaches into the existence of GG-NER or TC-NER has been conducted in Archaea. After unfruitful molecular biology approaches, involving placing a CPD substrate in a closed-end dsDNA for measurements of CPD repair with purified enzymes and cellular lysates, we developed novel *in vivo* and next generation sequencing approaches to investigate a

potential TC-NER in the euryarchaeon *Thermococcus kodakarensis*. Exciting recent progress has been made and has laid the groundwork for further research into archaeal NER. Investigations into complex(es) formed by putative archaeal NER enzymes *in vivo* were performed, and concluded that our potential suite of archaeal NER enzymes do not form stably-associated complexes. Further, a experiment was also designed to investigate the rate of repair of UV-induced DNA damages genome-wide in relation to transcriptional activity *in vivo* through combinatorial data generated from RADAR-seq (53) and RNA-seq. Our time-resolved RADAR/RNA-seq experiment has been optimized for *E. coli* control strains and is primed for investigations into GG-NER/TC-NER in *T. kodakarensis.* While the results discussed here failed to completely resolve an archaeal NER pathway, insight has been gained and progress made- and experimentation still actively pursued.

4.2 Results

4.2.1 Identification of putative nucleotide excision repair enzymes from *T. kodakarensis.*

Archaea are the likely progenitor of Eukarya, but hold significant similarities with Bacteria; homologs of both eukaryotic and bacterial proteins are distributed throughout the archaeal Domain. Thus, it is plausible that our model euryarchaeon *Thermococcus kodakarensis* could have an NER pathway which resembles the bacterial UvrA/B/C mediated pathway, the eukaryotic XP-family mediated pathway, or a previously unreported archaeal NER mechanism involving a unique family of enzymes. Genetic analysis of the entire archaeal Domain suggests that while some methanogenic and halophilic Archaea encode homologs of the bacterial UvrABC system, most encode homologs of the XP-family of NER enzymes- in particular helicases XPB and XPD, and the endonucleases XPF and XPG. Genetic analysis of *T. kodakarensis* specifically reveals homologs of eukaryotic NER enzymes XPB, XPD, and XPF are encoded (Figure 4.3). Further, there appear to be no encoded photolyases, suggesting a

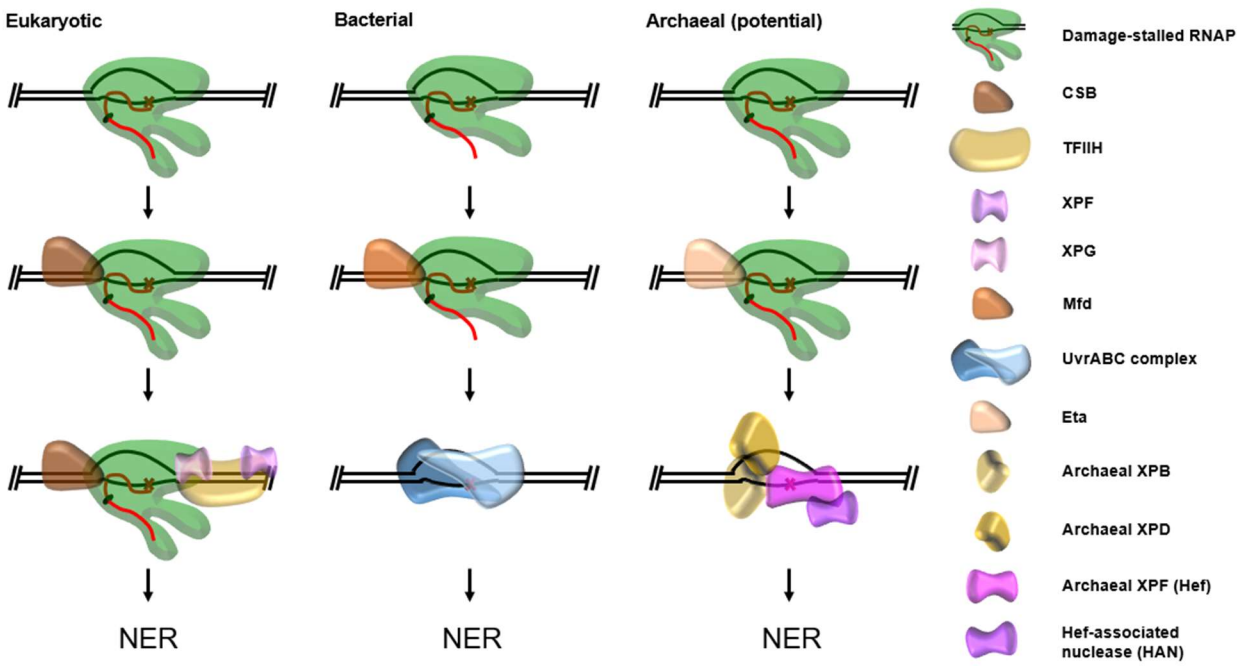


Figure 4.2: Current eukaryotic and bacterial models of transcription coupled nucleotide excision repair (TC-NER) and a hypothetical archaeal model. In all cases, RNA polymerase (RNAP) is arrested at template-strand DNA damage and recognized by the TRCF-CSB in Eukarya, Mfd in Bacteria, and potentially Eta in Archaea. The TRCF either backtracks RNAP or terminates transcription while recruiting NER enzymes directly to the site of damage. Homologs of the eukaryotic XP proteins found in many Archaea act in our archaeal model.

NER pathway is likely required to repair bulky helix-distorting DNA lesions versus a direct reversal pathway. The general pathway of NER suggests that the XP proteins found encoded in *T. kodakarensis* could be sufficient for an NER pathway. XPF is a nuclease which could plausibly be responsible for dual incisions required on each side of the bulky DNA lesion. XPB and XPD are DNA helicases with opposite polarities, and could thus be involved in damaged strand excision after dual incisions are made, or for damage recognition and/or validation. In fact, amino acid residues within the human XPD which when mutated specifically effect the function of XPD in NER are conserved between humans and Archaea. Recently, the XP-homologs encoded in *T. kodakarensis* have been mostly investigated in the context of DNA replication where XPF (Hef) is suggested to aid replication restart of stalled replisomes, but strains with single genetic deletions of XPB, XPD, and XPF render cells UV-sensitive, suggesting a role in UV-induced DNA damage repair. Thus, we hypothesized that a potential NER pathway in *T. kodakarensis* utilizes a eukaryotic-like suite of enzymes including XPB (TK0928), XPD (TK0784), and XPF (TK1021).

Additional enzymes were also selected for investigations into archaeal NER (Figure 4.3). Eta (TK0566) (see: chapter 2) is a euryarchaeal-specific transcription termination factor which appears to act analogously to bacterial Mfd, and thus may serve as an archaeal transcription-repair coupling factor (TRCF). Deletion of Eta also renders cells sensitive to UV irradiation. Hef associated nuclease (HAN – TK0155) is a 3' to 5' exonuclease which has been shown to form a stable complex with XPF and could potentially play a role in an archaeal NER pathway.

4.2.2 Putative NER enzymes do not form stably associated complexes in *T.kodakarensis*.

In eukaryotic NER, XPB, XPD, and XPF are part of the stable 10-protein complex TFIIH. To investigate if the putative NER proteins encoded in *T. kodakarensis* also form a stable complex,

	Protein	Encoding Gene	Annotated Function	Potential NER Role
Homologs in Eukarya	aXPB	TK0928	3' to 5' helicase	Helicase-mediated removal of damaged strand
Homologs in Eukarya	aXPD	TK0784	5' to 3' helicase	Helicase-mediated removal of damaged strand, damage recognition and/or verification?
Homologs in Eukarya	aXPF (Hef)	TK1021	3' to 5' helicase/endonuclease	Strand incisions pre-strand excision?
Archaeal-specific	Eta	TK0566	3' to 5' helicase, transcription termination factor	Transcription-repair coupling factor (Stalled RNAP initiated NER)
Archaeal-specific	HAN	TK0155	3' to 5' deoxyriboexonuclease XPF binding partner	Link between Eta and NER enzymes? Degrades damaged strand post-incision?

Figure 4.3: Putative NER enzymes identified from annotated *T. kodakarensis* genome (KEGG) and their potential roles in archaeal NER.

Multidimensional Protein Identification Technology (MudPIT) was performed on *T. kodakarensis* strains genetically manipulated to place N-terminal HA-6xHistidine tags in front of one of the potential NER proteins of interest (XPB, XPD, XPF, Eta, HAN). Strains harboring a tagged putative NER enzyme were cultured, lysed, and subjected to a Nickel chelating column, where the 6xHistidine tagged proteins and any stably interacting partners were selectively bound and eluted over a shallow imidazole gradient. For each tagged strain, Western blots were first performed using an anti-HA primary antibody to confirm both i) expression of the enzyme and ii) accessibility of the 6xHistidine tag to the Nickel charged column. Strain TS559, the parent strain of all tagged strains, was processed identically and sent as a negative control, and no HA-6xHis tagged proteins were detected in the resulting imidazole gradient (Figure 4.4.A, TS559). The presence of tagged proteins in imidazole elution fractions was readily apparent in Western blots (Figure 4.4.A, TK0155D). Tagged NER proteins from different strains bound and eluted identically from the Nickel column, allowing the same fractions from each strain to be sent alongside identical fractions of TS559, allowing determination of proteins which randomly co-elute in the same fractions as tagged proteins during the imidazole gradient elution. Fractions containing tagged protein of interest determined by Western blotting were pooled and quantified, and 10 µg of material was sent for LC-MS/MS identification of proteins co-eluting with the tagged NER protein of interest. Strain TS413, containing an identical tag on the C-terminus of the RpoL subunit of RNAP was processed identically as a positive control. RNAP is an extremely stable and long-lived complex, and thus other subunits of RNAP should co-elute and be present in significant amounts in fractions containing HA-6xHistidine tagged RpoL Protein levels detected by LC-MS/MS in each tagged strain were compared to those found in the TS559 elution using a Fisher's exact test; proteins significantly enriched in the tagged strain sample likely co-eluted with the tagged protein of interest. The TS413 positive control worked as expected, as untagged RNAP subunits were significantly enriched in the TS413 elution fractions vs. the TS559 elution fractions, indicating they co-eluted in large quantities with the tagged

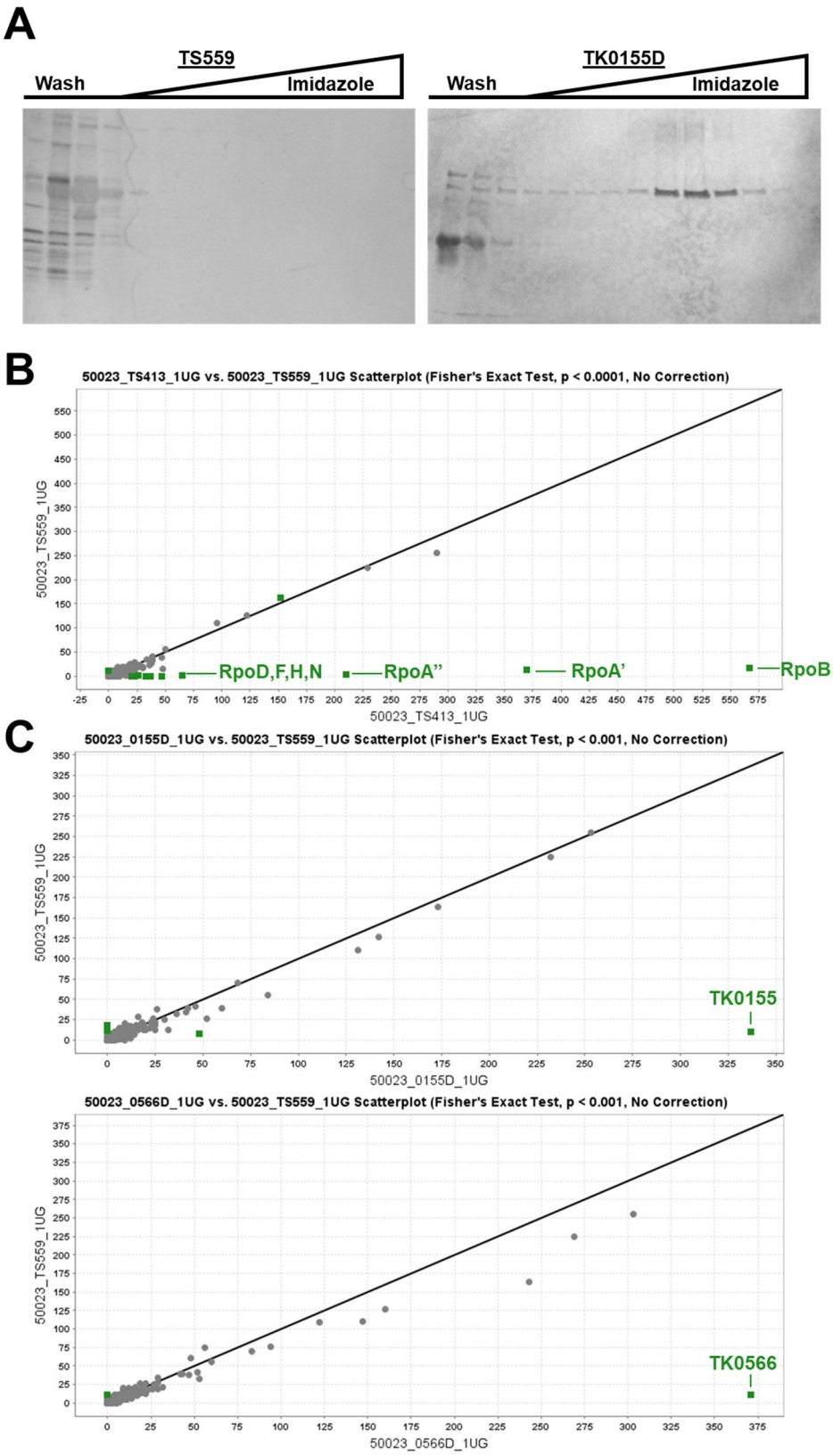

A
TS559
Wash
Imidazole
TK0155D
Wash
Imidazole
B
50023_TS413_1UG vs. 50023_TS559_1UG Scatterplot (Fisher's Exact Test, p < 0.0001, No Correction)
50023_TS559_1UG
50023_TS413_1UG
RpoD,F,H,N
RpoA"
RpoA'
RpoB
C
50023_0155D_1UG vs. 50023_TS559_1UG Scatterplot (Fisher's Exact Test, p < 0.001, No Correction)
50023_TS559_1UG
50023_0155D_1UG
TK0155
50023_0566D_1UG vs. 50023_TS559_1UG Scatterplot (Fisher's Exact Test, p < 0.001, No Correction)
50023_TS559_1UG
50023_0566D_1UG
TK0566

Figure 4.4: MudPIT analysis reveals no stable complex formed between putative NER enzymes identified in *T.kodakarensis*. A) Western blots of imidazole elution fractions of strain lysates loaded onto Nickel chelating columns. Strain TS559 contains no tagged proteins and thus no tagged proteins are detected in the elution fractions, whereas a single, specific population is detected corresponding to tagged HAN in the TK0155D strain. B) Proteins detected in strain TS413 (RpoL-HA-6xHis) imidazole elution fractions were compared to identical fractions of TS559 using Fisher's exact test. Untagged subunits of RNAP (RpoD, RpoF, RpoH, RpoN, RpoA', RpoA'', RpoB) co-eluted with the tagged RpoL subunit and were significantly more abundant in the TS413 sample than the TS559 sample, as expected. C) Proteins detected in strains TK0155D (HAN-HA-6xHis) and TK0566 (Eta-HA-6xHis) imidazole elution fractions were compared to identical fractions of TS559 using Fisher's exact test. Only the tagged proteins of interest (TK0155, TK0566) were found in significantly higher abundance in the tagged strain imidazole elution fractions, indicative of no co-eluting partners, and likely no *in vivo* NER complex.

RpoL subunit (Figure 4.4.B). Surprisingly, however, MudPIT data obtained from tagged putative NER enzymes provided no evidence of any stable NER complex being formed *in vivo*. When comparing protein content of imidazole elution fractions from tagged strains to TS559 fractions, only the tagged protein of interest was enriched in each sample. For example, only TK0155 (HAN) was enriched in TK0155D encoding the N-terminally tagged HAN protein product, and only TK0566 (HAN) was enriched in TK0566D encoding the N-terminally tagged Eta protein product (Figure 4.4.C). There were no obvious co-eluting factors with N-terminally tagged XPB, XPD, XPF, Eta, or HAN, suggesting that if these enzymes perform NER in *T. kodakarensis,* they do so without forming a long-lived stable complex. It remains plausible, however, that such a complex is not necessary for repair of bulky DNA lesions in Archaea.

4.2.3 Development of RADAR-seq to measure genome-wide transcription coupled DNA repair of UV-induced DNA damage in E. coli.

Repeated attempts to define either GG-NER or TC-NER in Archaea and identify factors which participate in either pathway either *in vitro* or *in vivo* were previously unsuccessful, necessitating an alternative experimental approach. The development of RADAR-seq (53) by collaborators at New England Biolabs permitted a genome wide snapshot of locations of a specific DNA-damage, including UV-induced CPDs which are repaired by the NER pathway in Bacteria and Eukarya. We devised an experiment using RADAR-seq in tandem with strand-specific RNA-seq to investigate the existence of GG-NER and TC-NER in Archaea for the first time, and this technique was first verified in *E.coli* where both GG-NER and TC-NER have been well defined.

Experimental design of RADAR-seq experiment in E.coli

A 10 mL culture of *E.coli* strain MGP (deficient in photolyase-mediated direct repair of CPDs) was irradiated at 254 nm to a final UV-exposure of 1.2 J/m^2 to induce CPDs in the genome

(Figure 4.5.A). After irradiation, the culture was inoculated into 40 mL 1.25X LB media and allowed to 'recover' from irradiation at 37°C for 60 minutes, allowing CPD repair by the bacterial Mfd and UvrABCD mediated NER-pathways. Aliquots of culture were taken before and after UV irradiation, and over a time course of the 60-minute UV-recovery period. Genomic DNA and RNA were both isolated from each aliquot of culture. CPDs in each sample can be mapped genome wide in a strand specific manner by RADAR-seq, and strand-specific RNA-seq reveals which strands of the genome are highly transcribed in each sample. Since TECs stall specifically at CPDs on template strand DNA before TC-NER is initiated, TC-NER can be evidenced by observing a faster rate of CPD loss on genomic DNA strands that correspond to the template strand of transcripts detected from strand-specific RNA-seq data. By performing the experiment on cells genetically deleted for enzymes involved in TC-NER (i.e. the bacterial TRCF Mfd) the faster repair rate of highly transcribed areas of the genome would disappear, as cells could only rely on GG-NER.

Induction and repair of CPDs can be detected in irradiated E. coli cells

Levels of CPDs were successfully detected in *E.coli* strain MGP before UV irradiation, immediately after UV irradiation, and over a 1-hour recovery time course. Before UV irradiation, cells contained a baseline level of ~4 CPDs per million base pairs sequenced- likely a result of ambient UV light. After cells were irradiated to 1.2 J/m^2, the number of CPDs per million base pairs jumped to ~45, indicating UV irradiation was successful in inducing genomic CPDs which could be detected by RADAR-seq (Figure 4.5.B, red curve). Further, we were able to observe the repair of these CPDs over the 1 hour recovery time-course. While repair is initially slow until ~20minutes recovery time, CPDs appear to return to baseline levels after ~30 minutes which is congruent with the literature estimate of bacterial NER taking 20 minutes to complete (54) and suggest successful detection of genome-wide CPD repair after UV irradiation.

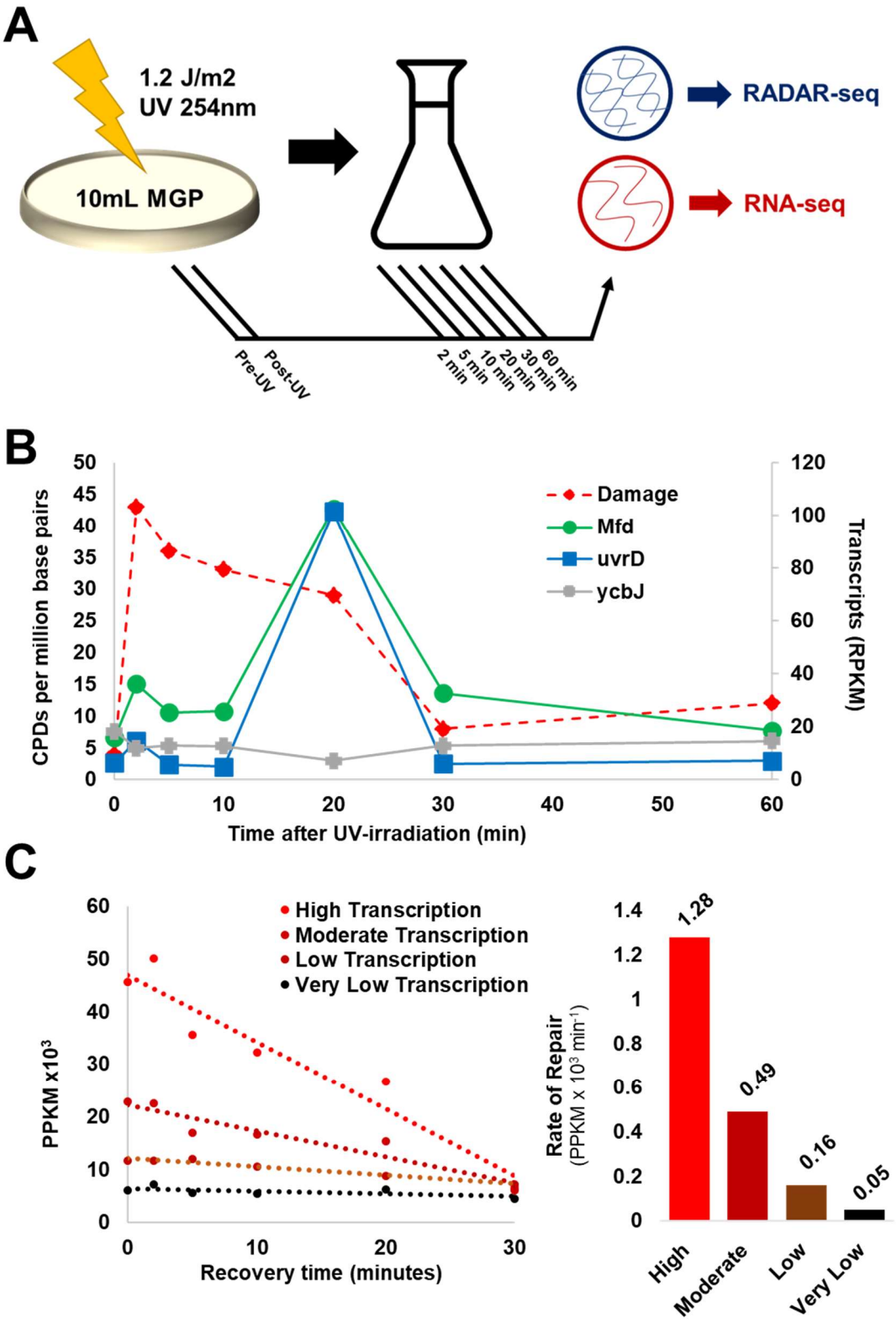

A
1.2 J/m2
UV 254nm
10mL MGP
RADAR-seq
RNA-seq
Pre-UV
Post-UV
2 min
5 min
10 min
20 min
30 min
60 min
B
CPDs per million base pairs
Transcripts (RPKM)
Time after UV-irradiation (min)
Damage
Mfd
uvrD
ycbJ
C
PPKM x10³
Recovery time (minutes)
High Transcription
Moderate Transcription
Low Transcription
Very Low Transcription
Rate of Repair
(PPKM x 10³ min⁻¹)
1.28
0.49
0.16
0.05
High
Moderate
Low
Very Low

Figure 4.5: RADAR-seq, in tandem with RNA-seq, can evidence transcription-coupled NER in *E.coli* cells. A) The time-resolved RADAR-seq/RNA-seq experiment involves initial irradiation of cell culture and recovery period. Samples of culture are taken pre-exposure, post-exposure, and during the recovery time course to isolate samples of DNA for RADAR-seq (allowing genome-wide detection of CPDs) and RNA for RNA-seq. B) CPD levels (red dotted line) significantly increase after UV-irradiation, and the repair of CPDs correlates with the increased expression of known NER enzymes Mfd (green line) and UvrD (blue line). C) When *E.coli* genes are binned into quartiles based on expression levels from RNA-seq data, it is observed that repair rates of CPDs (bar chart) increase with increased transcription, indicative of TC-NER.

Transcription of known bacterial nucleotide excision repair enzymes increases in E.coli after UV-irradiation

RNA-samples, isolated from the same aliquots as genomic DNA for RADAR-seq, were subjected to a library preparation and sequencing that allowed strand-specific mapping of reads, allowing observation of transcriptome changes after UV-irradiation and during the 1 hour recovery period. Reads Per Kilobase of transcript, per Million mapped reads (RPKM) levels were determined for Mfd, the bacterial TRCF required for TC-NER, and the UvrD helicase, required for 'excision' of the damaged strand during NER. While there appeared to be no transcriptional changes of NER genes ~10 minutes post-recovery, transcript levels of Mfd (Figure 4.5.B, green curve) and UvrD (Figure 4.5.B, blue curve) increased 5-fold and 10-fold, respectively, after 20 minutes. The increased expression of Mfd and UvrD appears to be induced specifically by UV irradiation, as non NER genes such as the protein kinase encoded by *ycbJ* (Figure 4.5.B, grey curve) had no changes to expression levels during the recovery time course. Further, the timing of increased expression of these genes is again congruent with the literature estimate of the NER timeline. Taken together, the increased expression of Mfd and UvrD in response to UV-irradiation within the predicted bacterial NER timeline suggests our time-resolved RADAR-seq experiment is successfully detecting loss of CPDs overtime specifically due to the bacterial NER pathways.

Time-resolved RADAR-seq/RNA-seq indicates CPD repair rate correlates to transcriptional activity, indicative of TC-NER in E.coli.

RPKM levels were determined from RNA-seq data to determine relative levels of transcription of each gene in the genome- and genes were then assigned to four quartiles based on the levels of transcription, allowing binning of CPDs into areas of high transcription, moderate

transcription, low transcription, and very low transcription, reported as Patches Per Kilobase of gene per Million reads (PPKM) to normalize CPDs to genes of varying lengths- where patches correspond to patches of methylated dNTPs incorporated at sites of CPDs during the RADAR-seq library preparation. Plotting CPD levels over time for each quartile (Figure 4.5.C) revealed a linear decrease in CPD levels, allowing a linear fit to estimate the rate of repair for each level of transcription. While low and very low transcribed areas of the genome exhibited relatively slow rates of CPD repair (0.16 and 0.05 PPKMx10^3 per minute), rates of repair increased significantly with transcriptional activity. Moderately expressed areas of the genome displayed higher levels of CPD repair (0.49 PPKMx10^3 per minute), and highly expressed areas of genome exhibited the highest repair rate (1.28 PPKMx10^3 per minute), indicating that the time-resolved RADAR-seq experiment successfully demonstrates CPD repair is coupled to transcription in *E. coli*. Interestingly, preliminary data suggests when separating CPD repair in highly transcribed areas of the genome between template and non-template strand indicated no difference in repair rate between the strands. This seems unexpected, and TECs pause specifically at template strand CPDs allowing initiation of NER. However, some research suggests that while the TEC stalls at a template strand DNA lesion, downstream NER activities are not strand specific (55), and may explain this observed phenomenon. Nonetheless, our time resolved RADAR-seq/RNA-seq appears a legitimate tool to investigate transcription-coupled repair of UV-induced DNA damages.

4.3 Conclusions, discussion and future directions

- ***T. kodakarensis* encodes homologs of enzymes involved in eukaryotic NER.**
- **Putative NER proteins in *T. kodakarensis* do not form stable long-lived complexes.**
- **Time-resolved RADAR-seq/RNA-seq can be used to monitor both transcription coupled DNA repair and UV-induced expression of NER proteins in *E.coli*.**

Archaeal DNA repair research has offered insight into the strategies of preserving DNA stability in extremes once thought inhospitable to life. Surprisingly, although some Archaea inhabit extreme environments, archaeal DNA repair pathways do not appear unique, but directly mirror pathways found in the bacterial and/or eukaryotic Domains. Unresolved is how Archaea deal with 'bulky' DNA lesions which cause perturbations to the DNA backbone. In Bacteria and Eukarya, NER is the predominant mechanism for repairing bulky DNA lesions, aside from direct reversal of the damage by photolyase enzymes. NER is yet to be resolved in Archaea, and research performed as part of this dissertation has taken steps towards elucidating its existence as a universally conserved mechanism of repairing bulky DNA lesions. *In vitro* investigations have thus far been unsuccessful in identifying desired NER-like activities from recombinantly expressed putative archaeal NER enzymes, but MudPIT LC/MS-MS offers insight into the complexes (or lack thereof) formed by putative NER enzymes and RADAR-seq/RNA-seq has provided a valuable tool for investigating genome-wide transcription coupled DNA repair. NER in Bacteria and Eukarya generally involves steps of damage recognition, incision each side of the DNA lesion on the same strand, excision of the damaged strand, and finally re-synthesis and ligation of the resultant gap by DNA polymerase and DNA ligase. It is plausible that the helicase activities of XPB/XPD and nuclease activity of XPF are sufficient for archaeal NER, or additional factors are required. For example, no homologs of enzymes required for recognition of bulky DNA lesions (XPA in Eukarya; UvrA in Bacteria) in GG-NER are found in the genome of *T. kodakarensis* and thus this activity may be performed by a currently unknown factor. Further, known NER mechanisms involve protein complexes such as the bacterial UvrABC complex and eukaryotic TFIIH. If the enzymes described here *do* perform NER in Archaea, they likely do so without forming a large protein complex, suggesting that an archaeal NER mechanism may follow similar steps, but these steps are achieved in a much different context. It is also possible that an archaeal NER complex does not form until UV-induced damage is detected by the cell,

and thus current efforts are directed at re-performing the MudPIT-LC-MS/MS experiment with cell populations that have been recovering for 20 minutes post-UV irradiation.

Adapting the RADAR-seq experiment for use with *T. kodakarensis* will be instrumental in determining any transcription-coupled repair of UV-induced DNA damage in Archaea for the first time. It is likely that transcription coupled repair of such damages involves an NER pathway. Strains of *T. kodakarensis* have already been generated with deletions for each putative NER enzyme- XPB, XPD, XPF, Eta, and HAN, and such strains will be valuable in determining the prerequisites for TC-NER through the time-resolved RADAR-seq/RNA-seq experiment. Once the enzymatic requirements for NER have been determined, development of *in vitro* techniques for determination of the exact role of each enzyme in NER will likely be required. If *T. kodakarensis* does utilize the XP family proteins to perform NER, the homologies to eukaryotic XP protein counterparts could open up a new paradigm for research of eukaryotic-like NER outside the context of a multi-protein subunit such as TFIIH. Although unlikely, it remains possible that NER does not exist in Archaea, and that currently unknown enzymes mediate photo-independent direct reversal or an alternative pathway for repair of bulky DNA lesions.

4.4 Methods

MudPIT analysis of putative NER proteins

N-Terminally HA-6xHistidine tagged strains ('D' strains) for genes TK0155, TK0566, TK0784, TK0928, and TK1021 were constructed as previously described (56). Strains were anaerobically cultured to an OD_{600} of 0.5 overnight at 85°C and immediately plunged on ice. Cells were pelleted by centrifugation 10k x g at 4°C for 20 minutes, and lysed by sonication in buffer A (10 mM Tris•HCl pH 8.0, 300 mM NaCl, 5 mM MgCl2, 1 mM βME). Cell lysates were clarified by centrifugation 10k x g at 4C for 20 minutes, and soluble material was immediately loaded onto a

1mL Nickel chelating column in buffer A at 1.0 mL/min. Bound material was eluted over a 0% to 30% gradient of Buffer B (10 mM Tris pH 8.0, 300 mM NaCl, 5 mM MgCl2, 1 mM βME, 200 mM imidazole) over 20CV, collecting 1mL fractions. 10 µL samples of obtained fractions were resolved by SDS-PAGE (4-20% criterion TGX (BioRAD). The presence of HA-tagged proteins in the gel was determined by Western blotting. Gel-resolved proteins were transferred to a PVDF membrane, and blocked for 30 minutes in 5% BSA and washed in TBST. Membranes were incubated in 1:1000 anti-HA mouse IgG in TBST 4°C overnight, washed with TBST, and incubated in 1:5000 anti-mouse rat IgG linked to HRP at room temperature for one hour. HRP was activated and imagined via colorimetric detector. Fractions containing tagged forms of TK0155, TK0566, TK0784, TK0928, or TK1021 were determined by presence of the appropriate band. Fractions corresponding to bands in the center 3 mL of the elution distribution were pooled and quantified via Qubit (Thermo Fisher). 10 µg of protein in solution was precipitated with equal volume 30% TCA, and shipped to the Mass Spectrometry and Proteomics Facility at The Ohio State University for LC/MS-MS analysis. All tagged proteins eluted almost identically, so corresponding fractions from a TS559 control were also treated identically and shipped for LC/MS-MS analysis. LC/MS-MS data was analyzed in Scaffold 5. Briefly, protein polypeptides from trypsin digest were detected by MS-MS and mapped to a TS559 reference proteome. Hits from each strain were assembled into digital files for analysis in Scaffold 5. Protein hits from TS559 were tested pairwise using Fisher's Exact Test to determine significantly different populations of enzymes in protein samples.

UV irradiation and recovery culture of E. coli for time resolved RADAR-seq and RNA-seq

100mL of *E. coli* strain MGP was grown to an OD_{600} of 1.25 in LB/Kanamycin at 37°C. Cells were pelleted by centrifugation at 4000 x *g* for 5 minutes and resuspended in 10 mL 1.0X ice cold PBS. 1 mL was taken for 'pre irradiation' nucleic acid samples. The remaining 9mL of cells

were spread on a 10 cm diameter glass petri dish and irradiated with 254 nm UV light to a total of 1.2 J/m^2 exposure. 1mL was taken for 'post irradiation' nucleic acid samples. The remaining 8mL was inoculated into 42 mL 1.2X LB/Kanamycin pre-heated to 37°C. The culture was incubated at 37°C and 5 mL of was removed at 2, 5, 10, 20, 30, and 60 minute time points for immediate nucleic acid extraction. Genomic DNA was isolated from samples (NEB Monarch), quantified, and 2 µg of material was prepared for Pacific-Bioscience based RADAR-seq analysis of CPDs as previously described (53). To obtain RNA, cells were resuspended and disrupted in 1 mL TRIzol reagent (Invitrogen) and 200 µL chloroform. After centrifugation at 8000 x *g*, the top aqueous layer was precipitated in isopropanol. RNA pellets were washed in 75% ethanol, and resuspended in RNase-free water. Extracted RNA was treated with 2 U DNaseI (NEB) for 90 mins at 37 °C, and subjected to an additional TRIzol extraction to isolate RNA. RNA was prepared for strand-specific RNA-seq as previously described (57). Libraries of DNA were sequenced on a Pacific Biosciences Sequel 1 and RNA libraries were sequenced by collaborators at New England Biolabs.

REFERENCES

1. S. Uphoff, *et al.*, Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation. *Science* **351**, 1094–1097 (2016).

2. I. Vitale, G. Kroemer, Spontaneous DNA damage propels tumorigenicity. *Cell Research* **27**, 720–721 (2017).

3. Z. Li, A. H. Pearlman, P. Hsieh, DNA mismatch repair and the DNA damage response. *DNA Repair* **38**, 94–101 (2016).

4. E. Crespan, *et al.*, Impact of ribonucleotide incorporation by DNA polymerases β and δ on oxidative base excision repair. *Nature Communications* **7** (2016).

5. L. He, *et al.*, Antioxidants Maintain Cellular Redox Homeostasis by Elimination of Reactive Oxygen Species. *Cellular Physiology and Biochemistry* **44**, 532–553 (2017).

6. W. L. Santivasi, F. Xia, Ionizing radiation-induced DNA damage, response, and repair. *Antioxidants and Redox Signaling* **21**, 251–259 (2014).

7. A. Carusillo, C. Mussolino, DNA Damage: From Threat to Treatment. *Cells* **9** (2020).

8. J. Cadet, J. Richard Wagner, DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation. *Cold Spring Harbor Perspectives in Biology* **5** (2013).

9. D. L. Jones, B. K. Baxter, DNA repair and photoprotection: Mechanisms of overcoming environmental ultraviolet radiation exposure in halophilic archaea. *Frontiers in Microbiology* **8** (2017).

10. H. Atomi, J. Reeve, Microbe profile: Thermococcus kodakarensis: The modehyperthermophilic archaeon. *Microbiology (United Kingdom)* **165**, 1166–1168 (2019).

11. M. Ehrlich, K. F. Norris, R. Y. Wang, K. C. Kuo, C. W. Gehrke, DNA cytosine methylation and heat-induced deamination. *Bioscience Reports* **6**, 387–393 (1986).

12. C. A. Lewis, J. Crayle, S. Zhou, R. Swanstrom, R. Wolfenden, Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 8194–8199 (2016).

13. T. Lindahl, Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).

14. Y. Ishino, I. Narumi, DNA repair in hyperthermophilic and hyperradioresistant microorganisms. *Current Opinion in Microbiology* **25**, 103–112 (2015).

15. D. W. Grogan, G. T. Carver, J. W. Drake, Genetic fidelity under harsh conditions: Analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 7928–7933 (2001).

16. A. Palud, *et al.*, Intrinsic properties of the two replicative DNA polymerases of pyrococcus abyssi in replicating abasic sites: Possible role in DNA damage tolerance? *Molecular Microbiology* **70**, 746–761 (2008).

17. C. Yi, C. He, DNA repair by reversal of DNA damage. *Cold Spring Harbor Perspectives in Biology* **5** (2013).

18. A. Sancar, Mechanisms of DNA Repair by Photolyase and Excision Nuclease (Nobel Lecture). *Angewandte Chemie - International Edition* **55**, 8502–8527 (2016).

19. A. Minobe, *et al.*, Biochemical characterization of mismatch-binding protein MutS1 and nicking endonuclease MutL from a euryarchaeon Methanosaeta thermophila. *DNA Repair* **75**, 29–38 (2019).

20. J. S. Lenhart, M. C. Pillon, A. Guarné, J. S. Biteen, L. A. Simmons, Mismatch repair in Gram-positive bacteria. *Research in Microbiology* **167**, 4–12 (2016).

21. A. Vaisman, *et al.*, Investigating the mechanisms of ribonucleotide excision repair in Escherichia coli. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **761**, 21–33 (2014).

22. M. R. Heider, B. W. Burkhart, T. J. Santangelo, A. F. Gardner, Defining the RNaseH2 enzyme-initiated ribonucleotide excision repair pathway in Archaea. *Journal of Biological Chemistry* **292**, 8835–8845 (2017).

23. S. S. Wallace, Base excision repair: A critical player in many games. *DNA Repair* **19**, 14–26 (2014).

24. A. M. Gehring, *et al.*, Biochemical reconstitution and genetic characterization of the major oxidative damage base excision DNA repair pathway in Thermococcus kodakarensis. *DNA Repair* **86** (2020).

25. J. Wiktor, M. van der Does, L. Büller, D. J. Sherratt, C. Dekker, Direct observation of end resection by RecBCD during double-stranded DNA break repair in vivo. *Nucleic Acids Research* **46**, 1821–1833 (2018).

26. B. B. Hopkins, T. T. Paull, The P. furiosus Mre11/Rad50 Complex Promotes 5′ Strand Resection at a DNA Double-Strand Break. *Cell* **135**, 250–260 (2008).

27. S. C. Shuck, E. A. Short, J. J. Turchi, Eukaryotic nucleotide excision repair: From understanding mechanisms to influencing biology. *Cell Research* **18**, 64–72 (2008).

28. M. Jaciuk, *et al.*, A combined structural and biochemical approach reveals translocation and stalling of UvrB on the DNA lesion as a mechanism of damage verification in bacterial nucleotide excision repair. *DNA Repair* **85** (2020).

29. A. Sancar, J. T. Reardon, Nucleotide excision repair in E. coli and man. *Advances in Protein Chemistry* **69**, 43–71 (2004).

30. G. Spivak, Nucleotide excision repair in humans. *DNA Repair* **36**, 13–18 (2015).

31. R. Tuteja, N. Tuteja, Analysis of DNA repair helicase UvrD from Arabidopsis thaliana andOryza sativa. *Plant Physiology and Biochemistry* **71**, 254–260 (2013).

32. M. G. Kemp, J. T. Reardon, L. A. Lindsey-Boltz, A. Sancar, Mechanism of release and fate of excised oligonucleotides during nucleotide excision repair. *Journal of Biological Chemistry* **287**, 22889–22899 (2012).

33. L. Fan, *et al.*, XPD Helicase Structures and Activities: Insights into the Cancer and Aging Phenotypes from XPD Mutations. *Cell* **133**, 789–800 (2008).

34. J. Kuper, S. C. Wolski, G. Michels, C. Kisker, Functional and structural studies of the nucleotide excision repair helicase XPD suggest a polarity for DNA translocation. *EMBO Journal* **31**, 494–502 (2012).

35. S. C. Wolski, *et al.*, Crystal structure of the FeS cluster-containing nucleotide excision repair helicase XPD. *PLoS Biology* **6**, 1332–1342 (2008).

36. N. J. Savery, The molecular mechanism of transcription-coupled DNA repair. *Trends in Microbiology* **15**, 326–333 (2007).

37. R. Dorazi, D. Götz, S. Munro, R. Bernander, M. F. White, Equal rates of repair of DNA photoproducts in transcribed and non-transcribed strands in Sulfolobus solfataricus. *Molecular Microbiology* **63**, 521–529 (2007).

38. N. Stantial, J. Dumpe, K. Pietrosimone, F. Baltazar, D. J. Crowley, Transcription-coupled repair of UV damage in the halophilic archaea. *DNA Repair* **41**, 63–68 (2016).

39. A. M. Gehring, T. J. Santangelo, Archaeal RNA polymerase arrests transcription at DNA lesions. *Transcription* **8**, 288–296 (2017).

40. A. M. Deaconescu, M. M. Suhanovsky, From Mfd to TRCF and Back Again—A Perspective on Bacterial Transcription-coupled Nucleotide Excision Repair. *Photochemistry and Photobiology* **93**, 268–279 (2017).

41. A. M. Deaconescu, A. Sevostyanova, I. Artsimovitch, N. Grigorieff, Nucleotide excision repair (NER) machinery recruitment by the transcription-repair coupling factor involves unmasking of a conserved intramolecular interface. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3353–3358 (2012).

42. O. Adebali, Y. Y. Chiou, J. Hu, A. Sancar, C. P. Selby, Genome-wide transcription-coupled repair in Escherichia coli is mediated by the Mfd translocase. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E2116–E2125 (2017).

43. B. J. Schalow, C. T. Courcelle, J. Courcelle, Mfd is required for rapid recovery of transcription following UV-induced DNA damage but not oxidative DNA damage in Escherichia coli. *Journal of Bacteriology* **194**, 2637–2645 (2012).

44. J. Y. Kang, *et al.*, Structural basis for transcription complex disruption by the Mfd translocase. *eLife* **10** (2021).

45. J. E. Walker, O. Luyties, T. J. Santangelo, Factor-dependent archaeal transcription termination. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E6767–E6773 (2017).

46. O. Adebali, A. Sancar, C. P. Selby, Mfd translocase is necessary and sufficient for Transcription-coupled repair in Escherichia coli. *Journal of Biological Chemistry* **292**, 18386–18391 (2017).

47. T. T. Le, *et al.*, Mfd Dynamically Regulates Transcription via a Release and Catch-Up Mechanism. *Cell* **172**, 344-357.e15 (2018).

48. D. Crowley, *et al.*, The uvrA, uvrB and uvrC genes are required for repair of ultraviolet light induced DNA photoproducts in Halobacterium sp. NRC-1. *Saline Systems* **2**, 11 (2006).

49. R. Fujikane, S. Ishino, Y. Ishino, P. Forterre, Genetic analysis of DNA repair in the hyperthermophilic archaeon, Thermococcus kodakaraensis. *Genes and Genetic Systems* **85**, 243–257 (2010).

50. C. Rouillon, M. F. White, The XBP-Bax1 helicase-nuclease complex unwinds and cleaves DNA: Implications for eukaryal and archaeal nucleotide excision repair. *Journal of Biological Chemistry* **285**, 11013–11022 (2010).

51. L. Feng, *et al.*, The trimeric Hef-associated nuclease HAN is a 3′→5′ exonuclease and is probably involved in DNA repair. *Nucleic Acids Research* **46**, 9027–9043 (2018).

52. R. Lestini, F. Delpech, H. Myllykallio, DNA replication restart and cellular dynamics of Hef helicase/nuclease protein in Haloferax volcanii. *Biochimie* **118**, 254–263 (2015).

53. K. M. Zatopek, *et al.*, RADAR-seq: A RAre DAmage and Repair sequencing method for detecting DNA damage on a genome-wide scale. *DNA Repair* **80**, 36–44 (2019).

54. J. J. Truglio, D. L. Croteau, B. van Houten, C. Kisker, Prokaryotic nucleotide excision repair: The UvrABC system. *Chemical Reviews* **106**, 233–252 (2006).

55. N. M. Haines, Y. I. T. Kim, A. J. Smith, N. J. Savery, Stalled transcription complexes promote DNA repair at a distance. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 4037–4042 (2014).

56. A. Gehring, T. Sanders, T. J. Santangelo, Markerless Gene Editing in the Hyperthermophilic Archaeon Thermococcus kodakarensis. *BIO-PROTOCOL* **7**, e2604 (2017).

57. T. Borodina, J. Adjaye, M. Sultan, “A strand-specific library preparation protocol for RNA sequencing” in *Methods in Enzymology*, (Academic Press Inc., 2011), pp. 79–98.

APPENDIX A

ARCHAEAL DNA REPAIR MECHANISMS[3]

A.1 Introduction

DNA is threatened by a plethora of intrinsic, internal and environmental insults. Endogenous damage has many sources: genomic material can spontaneously deaminate or break down under mechanical stress, cellular machineries will occasionally incorporate mismatch errors or ribonucleotides (rNMPs) into newly synthesized DNA, and many metabolic enzymes produce reactive species which can act upon DNA bases acting as oxidizing, alkylating and hydrolyzing agents. Cells must also tolerate exogenous sources of DNA damage depending on the external environment. Chemical crosslinkers, environmentally-generated reactive species, ultraviolet light and ionizing radiations from within or which penetrate the atmosphere all have mutagenic effects on DNA. Lacking the appropriate metabolic responses to DNA damage increases mutation rates, causing adverse effects on cellular health, with often drastic consequences to cellular and organismal fitness.

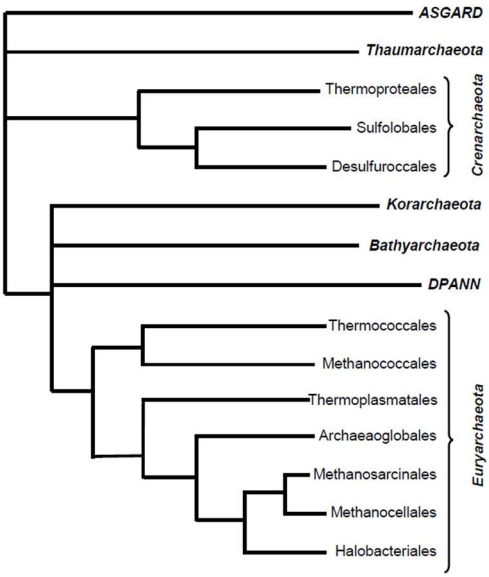
Many archaea thrive within niche and extreme environments which can amplify rates of DNA damage. Many halophilic archaea, for example, thrive in shallow salt plains and endure extreme levels of UV radiation (1), while some hyperthermophilic species persist at temperatures that would easily denature purified DNA (2, 3), and yet the increased rates of deamination, depurination and oxidation are somehow tolerated (4–6). In addition to growth in

[3]This appendix was previously published under the following title: Marshall, C. J. and Santangelo, T. J. (2020) 'Archaeal DNA repair mechanisms', *Biomolecules*. MDPI AG, pp. 1–23.

the extremes, many archaeal species maintain genomic stability sufficiently to display similar rates of spontaneous mutation to mesophilic prokaryotes such as *Escherichia coli* (7–9). It is perhaps surprising then that only DNA repair strategies conserved in Bacteria or Eukarya have yet been recognized in Archaea. While DNA repair strategies vary by clade, no new uniquely archaeal DNA repair pathway has been described (Table A.1). Insight into how Archaea detect and convert damaged DNA bases into repairable substrates has begun to reveal how genomic integrity is preserved *in extremis.* Here, we review our current knowledge of archaeal DNA repair pathways and examine both discrepancies and outstanding questions in the field.

A multitude of strategies to identify modified nucleotides or damaged DNA structures (here collectively termed recognition) and initiating repair are encoded in most genomes, with processes for recognition and repair perhaps best studied within mesophilic bacteria and eukarya. While direct DNA repair typical of photoreactivation of thymine-thymine dimers by photolyases and repair of methylation adducts by alkyltransferases(10, 11) is known, DNA repair more commonly involves sequential steps of recognition, initiation of repair and final ligation of repaired or resynthesized DNA. Collectively, the cycle of recognition-, initiation- and ligation-based DNA repair (Figure A.1) dominates the conserved DNA repair pathways that account for the majority of DNA repair, be it double-strand break (DSB) repair, mismatch repair (MMR), ribonucleotide excision repair (RER), base excision repair (BER), or both global and transcription coupled nucleotide excision repair (GG-NER, TC-NER). The core DNA repair pathways generally consist of recognition factors that more often than not cleave the DNA backbone and or glycosidic linkage to the nucleotide base, a repair DNA polymerase (DNAP) for strand re-synthesis, a nuclease, or the exonuclease activity of DNAP for removal of damaged bases/strands displaced during re-synthesis and DNA ligase to seal nicks generated during repair.

Table A.1: Predicted distribution of pathway-specific archaeal DNA repair proteins by clade (10), according to KEGG (Kyoto Encyclopedia of Genes and Genomes) orthologies. Many pathways appear conserved, with most variation found in distribution of mismatch repair (MMR) and nucleotide excision repair (NER) proteins.



Mismatch repair	DSB Repair		Ribonucleotide Excision Repair	Base Excision Repair	Nucleotide Excision Repair
	NHEJ/MMEJ	HR			
MutS, MutL, EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
EndoMS	Ku	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
MutL-MutS	?	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
MutL-MutS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
MutS, MutH, EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrB, UvrD, XPB, XPD, XPF
EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABC, XPB, XPD, XPF
MutS, MutH	Ku	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD
?	Ku	Mre11/Rad50	RNaseH2	Multiple	XPB, XPD, XPF
MutS, EndoMS	Ku	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPD, XPF
MutS, EndoMS	Ku	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPD, XPF
MutL/S/H, EndoMS	?	Mre11/Rad50	RNaseH2	Multiple	UvrABCD, XPB, XPF

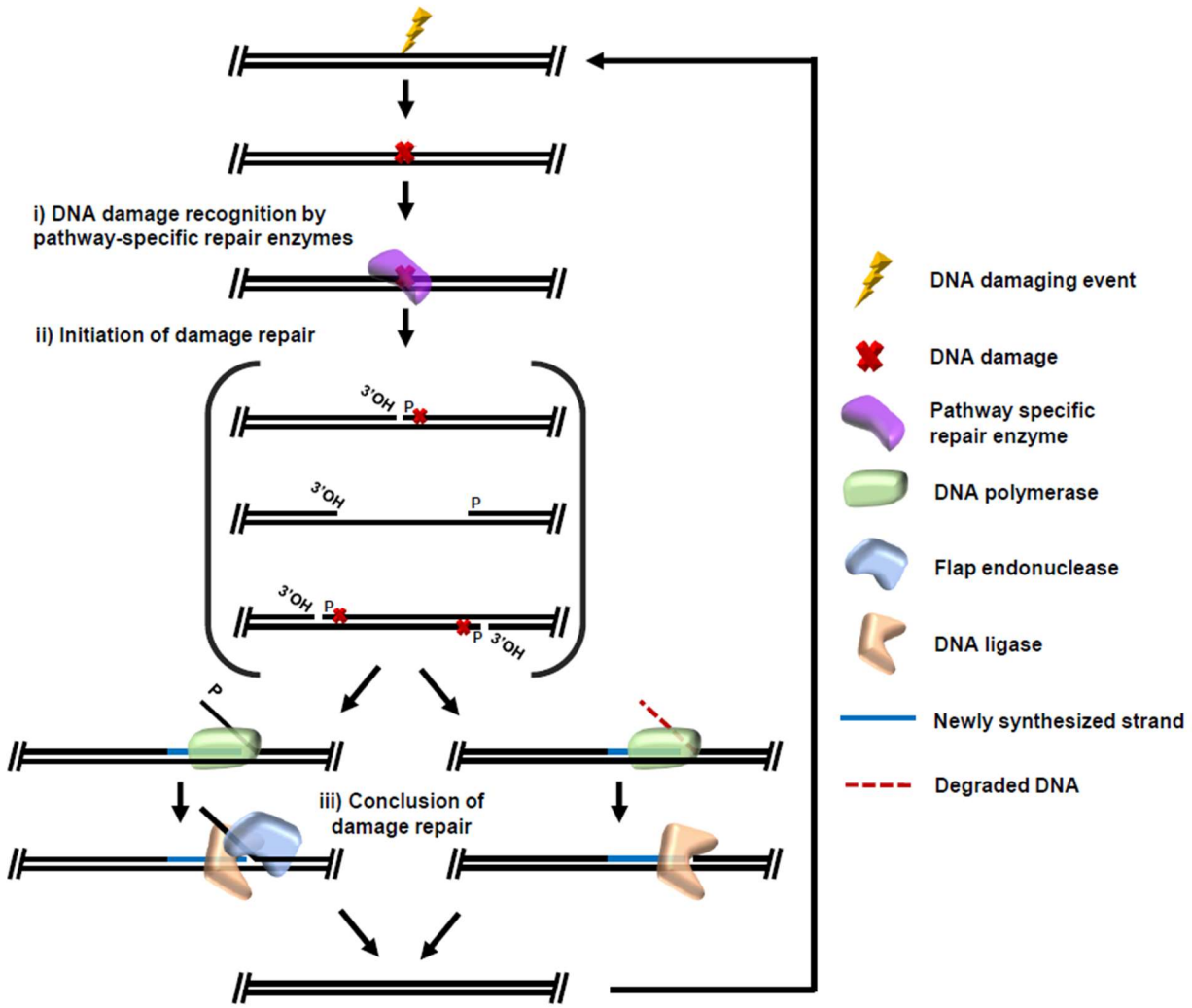


Figure A.1: Archaeal DNA repair pathways follow similar generalized steps. (i) Recognition of DNA damage by pathway-specific enzymes. (ii) Initiation of repair by conversion of DNA damage into appropriate and repairable substrate. (iii) Conclusion of repair by resynthesis of damaged DNA from a complementary undamaged strand, degradation of damaged strand by flap endonuclease of intrinsic DNA polymerase exonuclease activity, and nick ligation by DNA ligase.

A.2 Archaeal double-strand break repair (DSB Repair)

Double strand breaks (DSBs) are potentially the most mutagenic of all DNA damaging events. As the name suggests, DSBs involve a co-localized break in the phosphodiester backbones of both DNA strands, permitting regions of the genome to separate and offering the potential that the wrong ends, or trimmed ends of the DNA will be linked with the loss or repositioning of genetic information. While DSBs can be generated 'accidentally' by missteps of information processing machineries - i.e. by mistiming of replication, replication-transcription complex conflicts, and replication or transcription through existing DNA damage/secondary structures(12–15)- DSBs are also purposefully generated as essential intermediates of many nucleic acid metabolism pathways(16–19) and if such pathways are aborted prematurely, intermediate complexes may be released inappropriately. Detection and repair of DSBs is a top priority, as the resulting disruption of genomic architecture presents immediate danger to cellular health by severely compromising essential cellular processes such as replication and transcription.

The severe nature of the damage generated by DSBs has led to the evolutionary-retention of several mechanisms of DSB repair. Two conserved DSB repair methods - Microhomology Mediated End Joining (MMEJ) and Non-Homologous End Joining (NHEJ)(20, 21)- are relatively rapid and simple but both pathways are prone to loss of genetic material. The retention of multiple, in some cases many tens of genomes in some archaeal species facilitates a more accurate DSB repair mechanism, dependent on homologous recombination (HR-DSB)(22–24). HR-mediated methods for repairing DSBs have a higher energetic cost but are generally error-free because an undamaged template strand is made available without the need for strand resectioning (Figure A.2). HR-DSB is considered accurate but it is not without consequence, as crossover events or gene conversions are common results of HR- likely playing a significant role in the evolution of archaeal genomes.

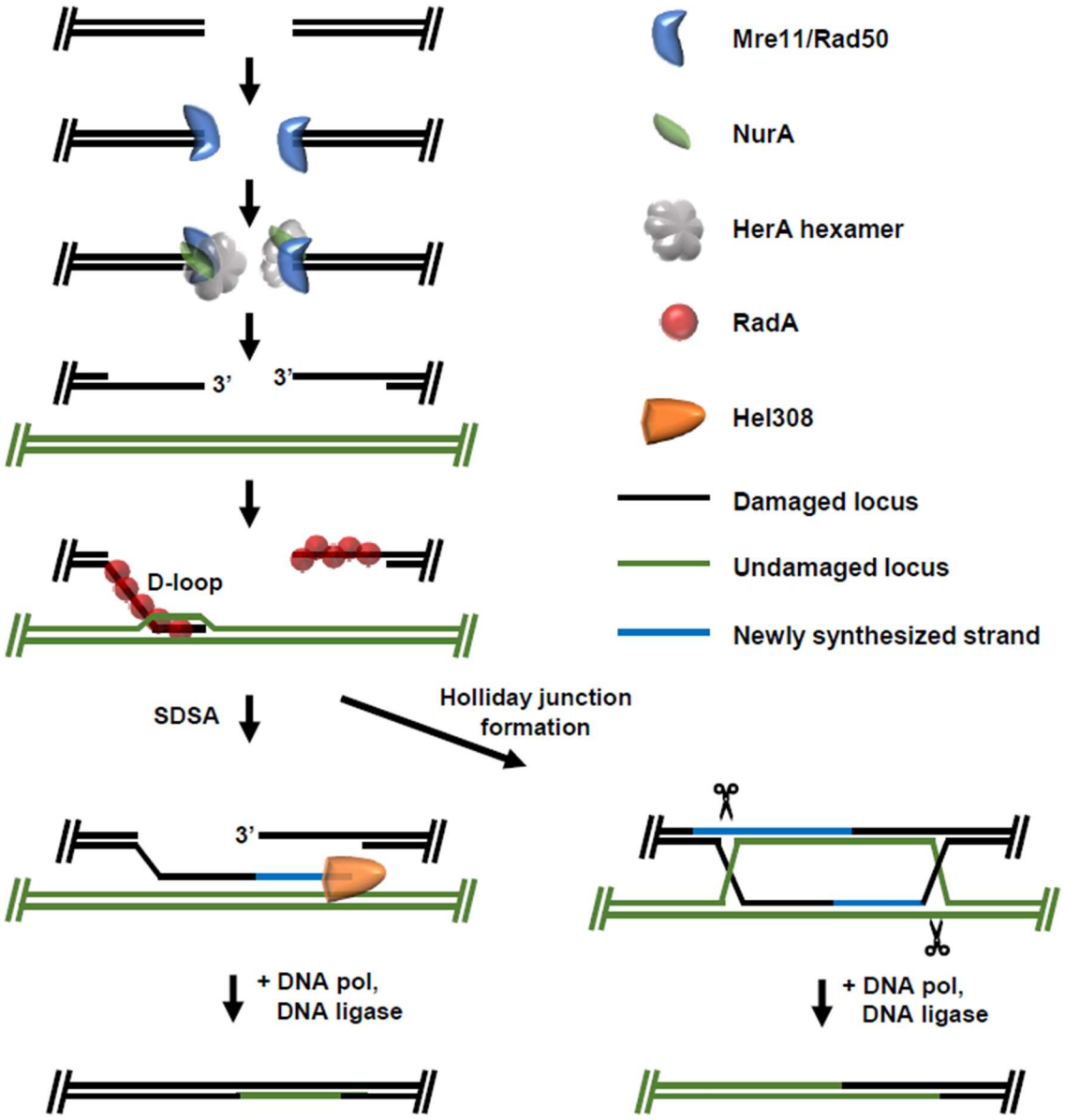


Figure A.2: Homologous recombination-based DSB repair in Archaea. Broken end recognition by the Mre11/Rad50 complex allows formation of 3′ overhangs by the HerA hexamer. RadA forms a nucleoprotein filament on the 3′overhangs and facilitates initiated homologous recombination through strand invasion. In the case of just one strand invasion event, synthesis-dependent strand annealing (SDSA) can occur before repair conclusion, a non-crossover event. If both strands are involved in local strand invasion events, a Holliday junction may form, the resolution of which may lead to crossover events.

A.2.1 HR-DSB repair

Recognition of DSB ends and subsequent ‘resectioning’ by exonucleases to produce single stranded 3’ ends is initiated by the universally conserved Mre11/Rad50 complex (SbcC/SbcD in Bacteria)(25–27). Resectioning steps allowing formation of 3’ ssDNA overhangs have historically been unclear(28, 29). Mre11/Rad50 genes are commonly encoded in operons with both a bipolar helicase HerA and a novel nuclease NurA in hyperthermophilic archaea, implying a functional link of these three enzymes to drive resectioning activities(30, 31). Current models suggest HerA and NurA are responsible for activities that generate the 3’ ssDNA ends after recruitment by the DSB localized Mre11/Rad50 complex(32, 33). In *Sulfolobus,* HerA resectioning is required for cell viability with the functional HerA-complex existing as a mixture of hexameric and heptameric states bound around strand of dsDNA. The nuclease NurA is thought to preferentially bind on the outside of the hexameric HerA-dsDNA substrate(34, 35), where ATP-dependent helicase activity of the HerA ring is thought to stimulate NurA activity, likely by coupling translocation and ssDNA substrate presentation for NurA to degrade(36–39). How this complex is specifically activated by Mre11/Rad50 after recognition of DSBs to produce appropriate resectioning remains elusive and is vitally important information for understanding the initiation of DSB repair by HR.

After resectioning, free ssDNA 3’ ends are recognized by the archaeal recombinase RadA which polymerizes along the length of the ssDNA region(40–42). The resulting dynamic RadA nucleoprotein filament then binds to local dsDNA and searches for a homologous sequence. Once located, the resulting intermediate structure is referred to as the ‘D-loop’, the primary initiation point for HR-DSB repair. D-loop formation permits two alternative and divergent pathways to complete repair. In some cases, only one 3’ end of the re-sectioned DSB is captured into a D-loop and is subsequently used as a starting point for DNA synthesis using the invaded, undamaged DNA strand as a template in a process termed synthesis-dependent strand annealing (SDSA)(43, 44). The newly synthesized strand is then unwound from the

invaded stand, where it can anneal with homologous sequences on the other side of the DSB to accurate repair the lesion. Unwinding of the newly synthesized strand is facilitated by the helicase Hel308, which uses a winged-helix domain in a ratchet mechanism to translocate 3'-5', simultaneously separating DNA strands in an ATP-dependent manner(45, 46). SDSA HR-DSBR does not result in a crossover event but can result in gene conversion if the *invaded* strand used as a template and the *invading* strand are heterozygous(47).

Alternatively, both ends of the DSB can be captured giving rise to a Holliday junction. Once generated, the Holliday junction must be resolved before repair can be completed. The archaeal Holliday junction resolvase Hjc specifically recognizes four-way junctions of DNA and uses nuclease activity to resolve the junction(48, 49). The resultant newly formed junctions can have significant impacts on genomic integrity and Holliday junction resolution is likely an important point of regulation for HR-DSBR. The cleavage activity of Hjc is repressed by phosphorylation in *Sulfolobus islandicus,* which is line with bacterial and eukaryotic resolvases(50, 51) and cells are also more resistant to high doses of DNA damaging agents when the phosphomimetic version of Hjc is expressed(52).

DSBs are a likely consequence of replication apparatuses reaching nicks or damaged DNAs, thus it is perhaps not surprising that many of the DSBR enzymes maintain interactions with known components of the replicative apparatus. Hjc, Mre11/Rad50 and Hel308 are all known to interact with DNA replication proteins, reinforcing the link between double strand break repair proteins and locating to areas of active replication.

A.2.2 Error-prone DSB repair pathways in Archaea

It is likely, especially in Archaea with low or varying ploidy (53, 54), that HR is not always a readily available pathway for the efficient repair of DSBs. This is highlighted by the evolution of alternative methods of DSB repair which do not require an undamaged template strand for repair. Microhomology Mediated End Joining (MMEJ) (Figure A.3), repair is dependent on

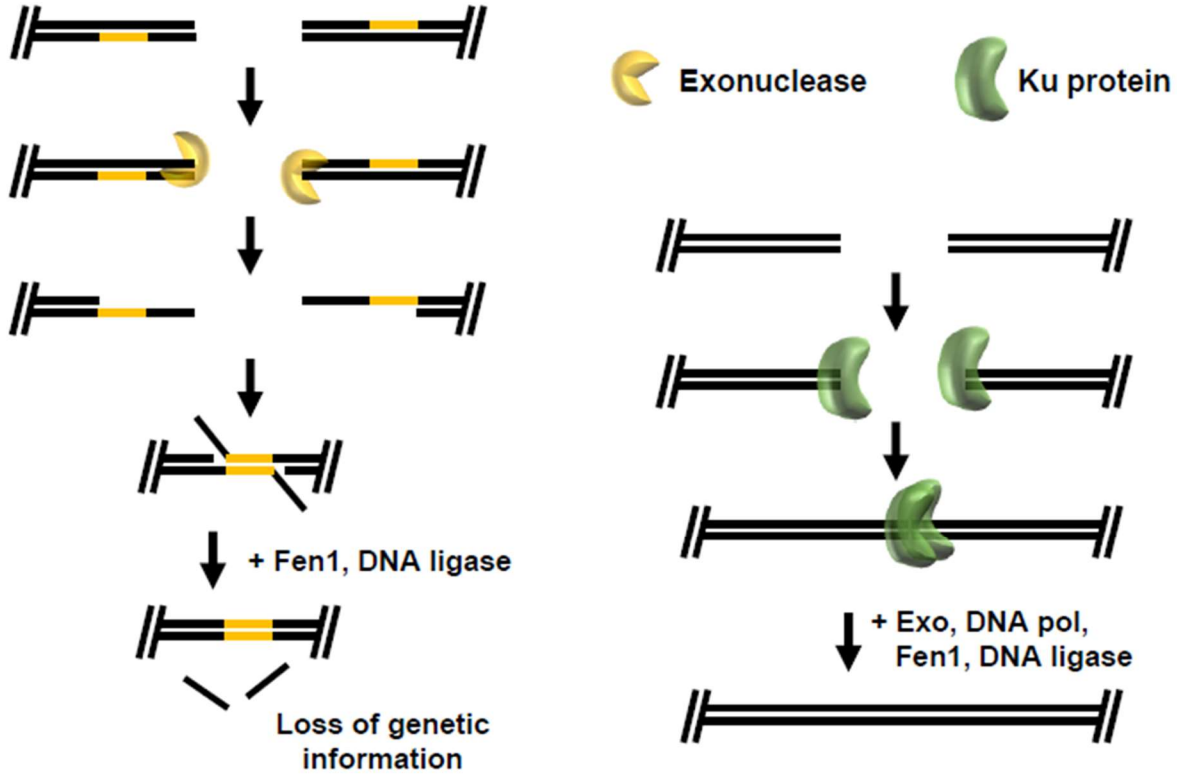


Figure A.3: "Error-prone" double-strand break (DSB) repair pathways in Archaea. (a) In microhomology-mediated end joining (MMEJ), small regions of microhomology (yellow) are revealed by exonuclease activity; annealing and subsequent processing by flap endonuclease and DNA ligase often results in the loss of genetic information. (b) Non-homologous end joining (NHEJ) in some archaeal species relies on recognition of broken ends by Ku which brings broken ends together, where exonuclease activity produces complementary ends for conclusion of DNA repair. The proteins that mediate NHEJ in many archaeal clades have not yet been defined.

short regions of close homology between sequences upstream and downstream of the DSB. These microhomologies are revealed by cellular exonucleases, allowing complementary sequences to anneal, producing a flapped substrate which is likely trimmed by flap endonuclease (Fen1) or the GINS-associated nuclease (GAN) before DNA ligase seals the final nick(s)(55). The unfortunate consequence of dependence on areas of microhomology is that they can sometimes be located far from the site of damage, and often intervening sequence are lost during repair (56). Many details of archaeal MMEJ require additional studies, but DNA repair products reminiscent of MMEJ pathways have been observed in both crenarchaea and euryarchaea(57).

Non-Homologous End Joining (NHEJ) (Figure A.3) does not require large- or even micro-regions of homology for repair of DSBs. Instead, broken ends are brought together in a protein-mediated complex involving the DNA end-binding Ku protein and a multitude of likely dynamically-associated DNA repair enzymes. Although the molecular details have not been determined, Ku bound ends are exonucleolytically processed to generate 3' ends that can be extended by strand-displacement synthesis by DNA polymerase(58, 59). Synthesis by DNA polymerase bridges the DSB, allowing DNA ligase to seal resulting nicks. Archaeal NHEJ relies on exonuclease activity to produce a template for strand re-synthesis and can thus result small deletions of genetic information.

How cells commit to an accurate or error-prone DSB repair pathway has significant consequences for gene conversion, genomic stability and crossover events. Competition between repair pathways is likely, and manipulation of one pathway can result in surprising impacts on another. When mutations to the Mre11/Rad50 complex in *Haloferax volcanii* were introduced that were predicted to recruit resectioning enzymes essential for HR, instead of activating HR-DSBR, the rates of HR-DSBR decreased(60). Post-translational modification of DSB repair components, including methylation the Mre11/Rad50 complex in *Sulfolobus*

acidocaldarius likely also contribute to the efficiency and rates of different DSBR pathways(61, 62).

A.2.3 New resources emerging from DSB repair pathways

As more molecular details of both HR and error-prone archaeal DSB repair mechanisms emerge, opportunities abound. The induced and natural competency of many archaeal species permit genetic manipulations, most dependent on HR-directed gene conversion and integrations of new DNA. The archaeal Hel308 enzyme, believed to be responsible for strand displacement during SDSA, has been extensively studied for use in nanopore sequencing(63). DSB repair is also an essential process for CRISPR viral defense systems found in ~85% of Archaea, in which Cas enzymes generate guided double-strand breaks which are subsequently repaired by non-HR DSB repair pathways, NHEJ and MMEJ(64). Knowledge of conserved non-HR DSB repair has allowed for development of the first type II CRISPR-Cas based genomic editing systems in archaea(56).

A.3 Mismatch Repair (MMR) in archaea

DNA polymerases must not only perform replication with high-fidelity, but also with physiologically relevant high speeds to avoid disruption of proper gene expression. The necessity for fast DNA synthesis inevitably leads to errors by replicative DNAPs, with incorrect base incorporations once every 10^6–10^{10} nucleotides under normal conditions. In general, misincorporating a purine for purine (or pyrimidine for pyrimidine) occurs more readily, resulting in transitions (A:T to/from G:C) rather than transversions (i.e. A:T to/from C:G)(65, 66). Failure to efficiently recognize and repair the resulting mismatches leads to increased mutation rates (67). The canonical pathway of DNA mismatch repair (MMR) is the MutL/MutS/MutH pathway which has been well characterized in Bacteria and Eukarya(68, 69), but many Archaea do not encode obvious homologs of these enzymes. The apparent lack of MutL/MutS in many Archaea

drove efforts to describe an alternative pathway for mismatched base recognition and resulted in identification of the novel EndoMS nuclease. Here, we summarize the MutL/MutS pathway and recent insights into potential EndoMS-based MMR.

A.3.1 MutL/MutS

The MMR machinery in Bacteria is likely localized to nascent DNA strands during DNA replication, where mis-matched bases are first recognized by MutS. Once bound to mismatched DNA, MutS subsequently recruits MutL, and the MutS-MutL complex can then stimulate nuclease activity of MutH. MutH specifically nicks at unmethylated GATC methylation sites allowing discrimination between the template and nascent DNA strands. Cutting at unmethylated GATC sites ensures the nick (and subsequent degradation of mismatched DNA) is performed on the newly synthesized strand which likely contains the error. The helicase UvrD is then thought to perform strand displacement, with subsequent degradation of the damaged strand by generic cellular exonucleases. This allows DNA polymerase to re-synthesize from the undamaged strand and DNA ligase to seal the resulting nick. Eukaryotic MMR is similar but does not contain MutH or UvrD(70, 71). Instead, it is thought that asymmetric loading of MutS/MutL-mediated by interactions with replisome components- directs MutL nuclease activity to the newly synthesized strand. The eukaryotic repair polymerase contains both replication and exonuclease activities which are believed to facilitate removal and degradation of the damaged strand during re-synthesis.

Studies of methanogenic Archaea which encode MutS/MutL homologs indicate that the initial steps of this pathway are likely comparable to eukaryotic-like MMR. *Methanosaeta thermophila* MutS1 binds mismatched dsDNA but has low affinity for perfectly matched duplexes. The corresponding archaeal MutL makes single stranded nicks at the site of mismatches which are assumed to be directed to a specific strand in a similar manner to eukaryotic homologs(72). The importance of MutS/MutL for MMR, however, does not seem to

be ubiquitous, as homologs from halophilic Archaea are readily deleted with no apparent effect on mutation rate(73), suggesting an alternative MMR pathway was present in these species.

A.3.2 EndoMS

To identify potential MMR enzymes in species apparently lacking MutL/MutS, cosmid-expressed *Pyrococcus furiosis* genome regions were screened for the ability to cleave the DNA backbone at the site of mismatches and resulted in the identification of EndoMS- a novel nuclease with homologs found in Bacteria (NucS) (74). Deletion of EndoMS/NucS in *Mycobacterium smegmatis* and *Corynebacterium glutamicum* resulted in an increased mutation rate, with observed mutations that match transitions- the expected result of MMR deficiency. In fact, the *C. glutamicum* EndoMS/beta-clamp interaction is required for high fidelity DNA replication(75)- cementing the role of the enzyme in an alternatively initiated MMR pathway.

Biochemical and structural characterizations of archaeal EndoMS revealed MMR-like activities differing significantly from the Mut enzymes, offering dual activities of mismatch recognition and backbone cleavage in a single enzyme. A mismatched DNA:EndoMS complex structure suggests the enzyme forms a functional dimer and uses a 'base-flipping' mechanism for recognition of mismatched bases, eventually activating extending nuclease domains (76, 77). Perhaps most striking is the offset cuts made on opposite strands *in vitro* by EndoMS resulting in a substrate akin to a DSB with two 5' overhangs. If this activity is maintained *in vivo,* the detrimental consequences of DSBs (re. DSBR in Archaea) must then be dealt with. While it is possible that the activity of EndoMS is somehow directed *in vivo*- both bacterial and archaeal EndoMS interact with replisome clamp domains (beta-clamp in bacteria, PCNA in archaea) (78)- the downstream consequences of a DSB-like product remain unclear (Figure A.4). HR based double strand break repair is a probable compliment pathway, but this is less likely in Archaea which spend significant time in a diploid state. Furthermore, there remain Archaea without a

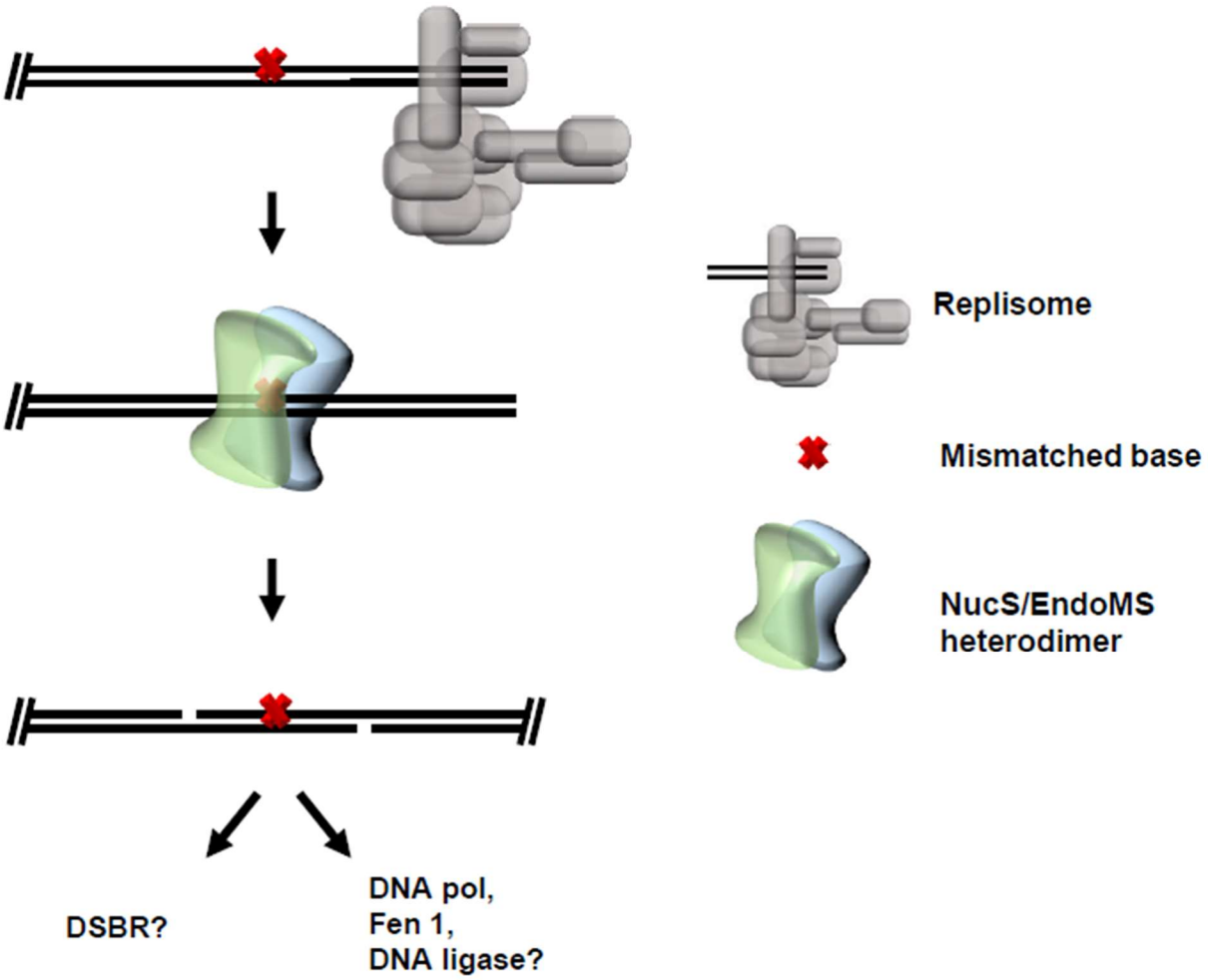


Figure A.4: Potential mismatch repair pathways through NucS/EndoMS. NucS/EndoMS may surveil newly synthesized areas of the genome for mismatch incorporations. If a dual cut is made as *in vitro*, a DSB-like substrate would be formed, requiring DSB repair pathways or more immediate repair conclusion by DNA Polymerase, Flap endonuclease, and DNA ligase.

characterized MutL/MutS or EndoMS pathway for MMR- suggesting undiscovered avenues for MMR in these species.

A.4 Ribonucleotide Excision Repair (RER) in archaea

The replicative DNA polymerase must not only reduce mismatches by distinguishing between DNA bases but must also monitor the usage of dNTPs vs rNTPs(79). Cellular concentrations of rNTPs can be magnitudes higher than that of dNTPs, and thus inappropriate incorporations of rNTPs into dsDNA are inevitable. Archaeal D family DNA polymerases have been shown to incorporate 1 rNTP every ~1000 bases, and archaeal B family DNA polymerases every ~2500 bases(80, 81)- but not all incorporations are erroneous. Purposefully incorporated ribonucleotides are common, i.e. RNA primers for DNA replication, and it is posited that many rNTP incorporation events by DNA polymerases are evolutionarily conserved[76]. Whether rNTP incorporation into DNA is accidental or purposeful, the lack of efficient removal of rNTPs has detrimental effects on genome stability- specifically by altering DNA-from and enhancing hydrolytic activity brought by the rNTP 2'OH which is lacking in dNTPs(82).

Ribonucleotide excision repair (RER) is the universally conserved pathway for removal of rNTPs incorporated into dsDNA. RER is initiated by RNaseH2, generating a nick on the 5' end of the embedded rNTP. In eukaryotes, the 3' end generated by this nick is used in strand displacement synthesis by DNA polymerase δ/ε, and the resulting flap (with rNTP incorporation) is cleaved by flap exonuclease Fen1(83). In Bacteria, the strand displacement synthesis and flap cleavage are both carried out by DNA polymerase I(84). Archaeal RER activities were tracked in *Thermococcus* lysates lacking computationally annotated homologs of RER enzymes on dsDNA substrates with a single embedded rNTP. An archaeal RNaseH2 homolog recognizes the rNTP incorporation, nicks at the 5' end, allowing strand displacement synthesis by the B family (eukaryotic DNA polymerases α, δ, and ε) repair DNA polymerase in *Thermococcus* (Figure A.5). Consistent with a eukaryotic-like RER mechanism, the repair

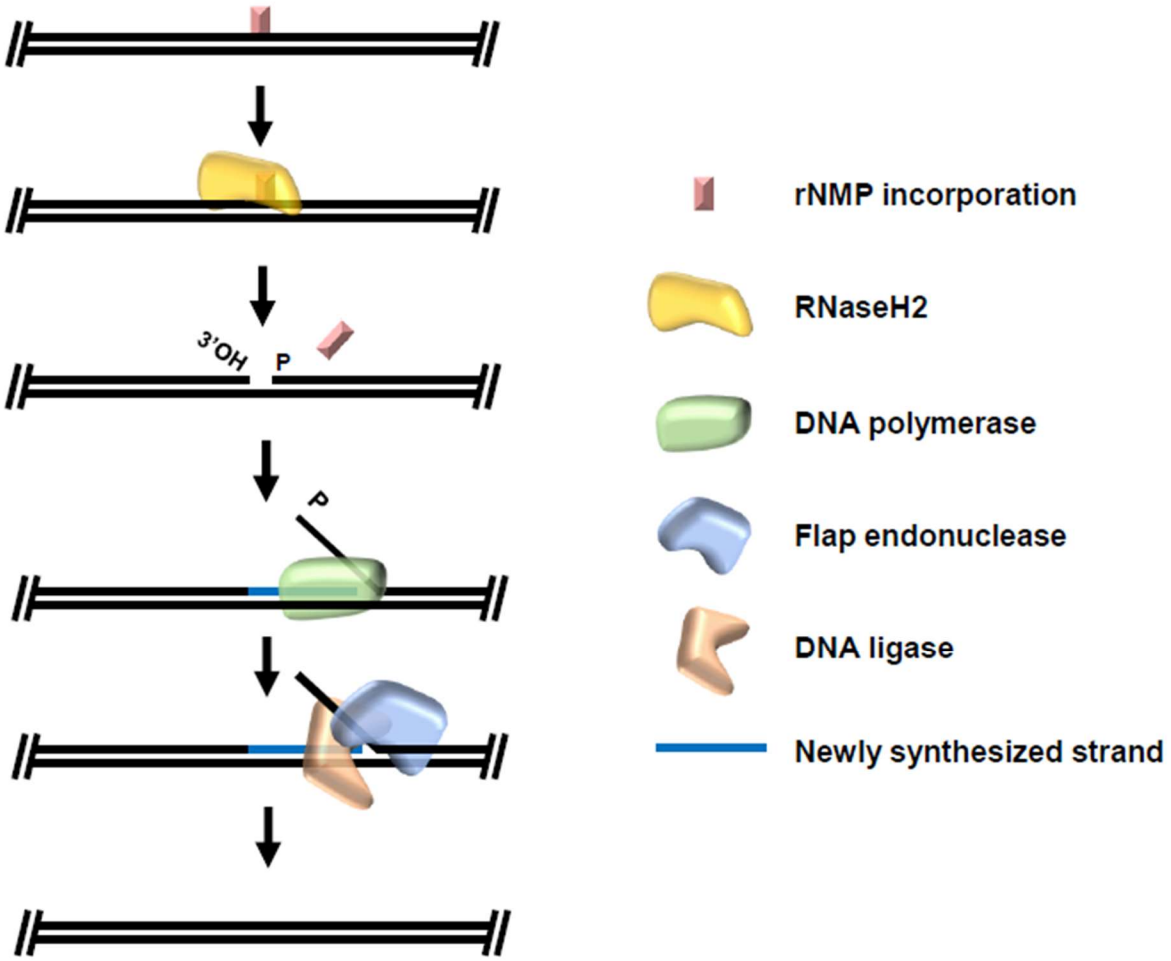


Figure A.5: **Archaeal ribonucleotide excision repair**. Embedded ribonucleotide monophosphates (rNMPs) are recognized and specifically excised by RNaseH2, resulting in one nucleotide gap with 3′-hydroxyl. Repair is concluded when DNA polymerase performs strand-displacement synthesis and the activities of flap endonuclease and DNA ligase remove the original strand and seal the resulting nick.

polymerase Pol B does not have flap exonuclease activity which is instead performed by a homolog of eukaryotic Fen 1.

Single rNTPs repaired by RER can be misincorporated by DNA polymerases but are sometimes a cause of inefficient removal during Okazaki fragment maturation(85). In *Thermococcus,* Okazaki fragment maturation resembles RER and involves DNA polymerase using strand displacement synthesis to remove RNA primers used to initiate DNA synthesis on the lagging strand during replication(86). In some cases, the 5'-3' exonuclease activity of GAN is necessary to remove the displaced RNA flap- but in the absence of GAN, the RER enzymes Fen 1 and RNaseH2 are reported to function together to remove the displaced RNA flap. Cells need either GAN or both Fen1/RNaseH2 for survival- not only suggesting that RER is possibly sufficient for Okazaki fragment maturation but also the activity of GAN exonuclease during DNA replication is sufficient to maintain viable levels of rNTP:dNTP in cellular DNA. *Thermococcales* are generally polyploid, raising the possibility that increased homologous recombination events often expose DNA strands to enzymes responsible for maintaining genomic maintenance – allowing retained DNA repair pathways to compensate for repair activity lost when deleting enzymes involved in another repair pathway. In diploid mammals, deletion of RNaseH2a- the homolog of archaeal RNaseH2- results in embryonic lethality(87).

A.5 BER in archaea

Not all DNA damage arises from mistakes made by cellular machineries- in fact, single base modifications (i.e. alkylations, deaminations, and oxidations) are the most common DNA damage. Such damages can be exemplified by both the internal and external environments of the cell and are correlated with high mutation rates incompatible with sustained life if left unchecked(88–90). Archaea often occupy niche and extreme environments which increase prevalence of exogenous damage sources but nonetheless utilize the universally conserved base excision repair (BER) pathway to remove chemically modified bases from DNA(91). The

canonical BER pathway involves recognition of specific base modifications by glycosylases specific to each modification. The glycosylases cleave the glycosidic bond between the base and phosphodiester backbone- 'base excision'. This leaves an abasic (AP) site, allowing AP-specific nucleases to create a free 3'OH for re-synthesis of the damaged strand. The damaged strand can then be removed at a junction site by Flap endonuclease- or by direct removal of the abasic nucleotide- and the resulting nicks sealed by DNA ligase. One recently characterized example of archaeal BER involves the Ogg-subfamily archaeal GO glycosylase (AGOG) of *Thermococcus kodakarensis* (92). AGOG specifically recognizes 8-oxo-guanine (8oxoG) modifications which result from the oxidation of guanine and acts as a dual glycosylase and AP lyase to perform base excision. The activity of AGOG-like BER enzymes leave a 1nt gap with a 5' phosphate and 3' unsaturated aldehyde after recognition of a chemically modified base. The 3' unsaturated aldehyde must be chemically converted to a 3'OH by cellular endonucleases (Endo IV in *T. kodakarensis*) before strand-displacement synthesis by DNA polymerase B, flap cleavage by Fen1, and ligation of the resultant nick by DNA ligase (Figure A.6). Studies of AGOG have also provided insight into structural bases of specificity- as structural analyses have determined specific recognition and cleavage of damaged substrates by AGOG is mediated by a conserved proline and phenylalanine motif allowing appropriate conformational freedom(93).

While the downstream steps of repair are thought to be similar, recognition of chemically modified DNA in Archaea has also been shown to be mediated directly by endonucleases rather than glycosylases. In hyperthermophilic archaeal species, where temperature-dependent chemical modifications are presumably more common, the existence of multiple damage repair initiating enzymes is likely advantageous. Alternative excision repair (AER) pathways do not rely on excision of the damaged base and subsequent abasic site recognition as separate steps, potentially accelerating repair of specific damage types. Endonuclease V has been shown to recognize all deaminated bases in *Ferroplasma acidarmanus,* and specifically hypoxanthine (deaminated adenine) bases in *Pyroccocus furiosis* and *Thermococcus barophilus*(94–96).

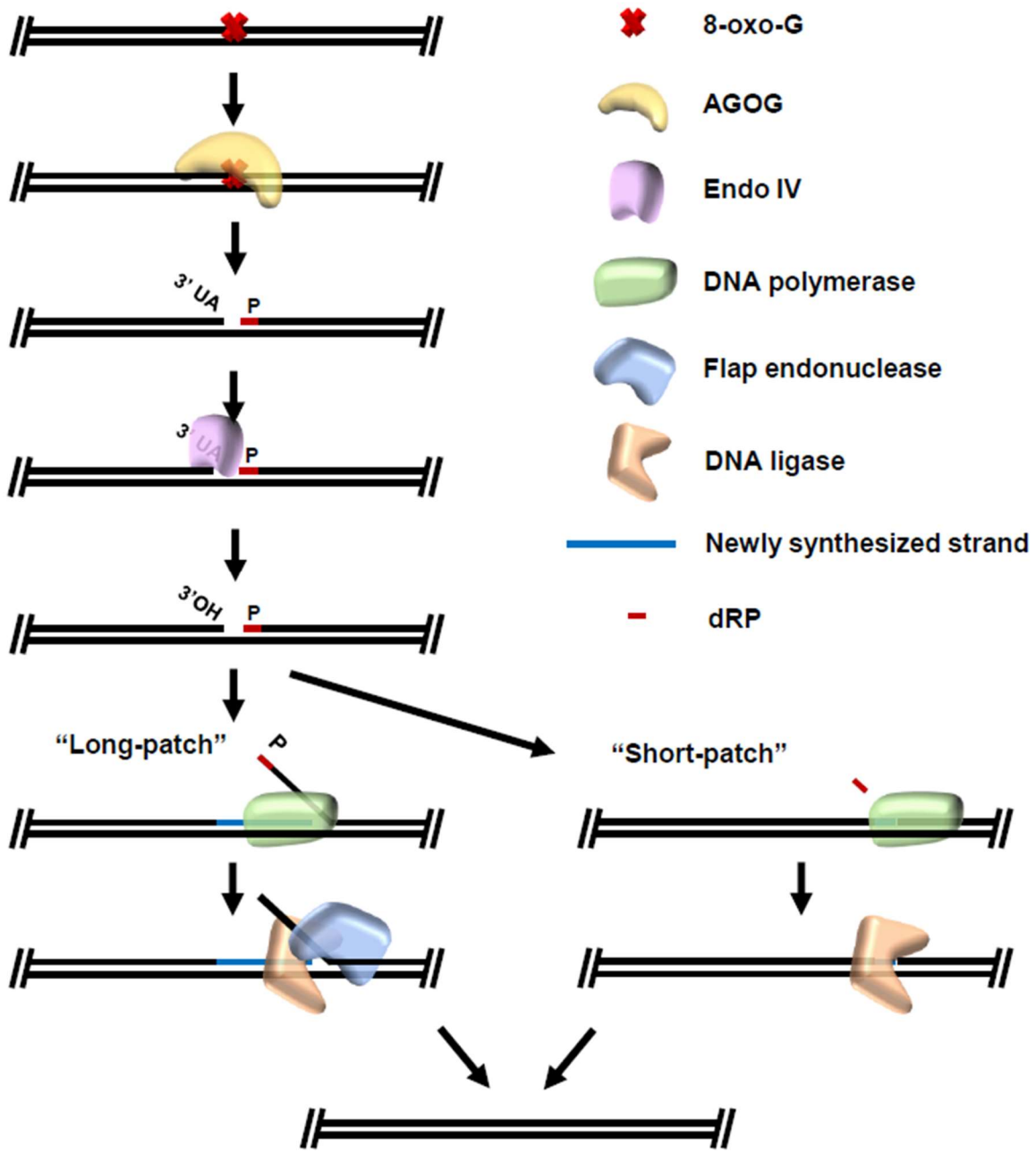


Figure A.6: Reconstituted archaeal base excision repair from *Thermococcus kodakarensis*. AGOG recognizes 8-oxo-G modifications and acts as a bifunctional glycosylase, both excising the damaged base and cleaving the DNA backbone at the site of damage. The resulting substrate contains a 3′ unsaturated aldehyde (UA) and 5′ dRP. Damage repair is initiated by the activity of Endonuclease IV, which converts the 3′-UA to an extendable 3′-hydroxyl group. In long-patch base excision repair (BER), strand displacement activity of DNA polymerase during synthesis is used in tandem with flap endonuclease and DNA ligase to conclude repair. In short-patch BER, dRP lyase activity intrinsic to DNA polymerase simply removes the dRP moiety while synthesizing the correct base from the undamaged strand, and DNA ligase seals the nick.

Endonuclease V binds and cuts at the 3' end of the specific damage type, initiating downstream repair processes. Another novel nuclease, Endonuclease Q, was recently discovered in *P. furiosis* and shown to cleave the DNA backbone at deaminated bases and abasic sites(97). Similar to EndoMS, endonuclease Q also uses a 'base-flipping' mechanism, placing bases in an active-site adjacent pocket which allows for cleavage in the event of improper base pairing resulting from oxidized bases such as 5-hydroxyuracil and 5-hydroxycytosine(98, 99). The wide substrate range of endonuclease Q, coupled with its studied interactions with PCNA, suggest that endonuclease Q may also localize to newly synthesized DNA, catching DNA damages that may not lead to misincorporations by DNA polymerase.

The characterization of a large selection of archaeal BER enzymes specific to particular damage types has provided an advantageous protocol for biochemical analyses of DNA damages on a genome-wide scale. In RADAR-seq(100), enzymes specific to a DNA damage type are used to make lesion-dependent nicks on sequencing libraries prepared from purified genomic DNA. DNA repair enzymes which create an extendable 3'OH at the nick site are then utilized, followed by strand-displacement synthesis by DNA polymerase in the presence of methylated dNTPs. Methylated bases are thus incorporated the site of DNA damage repair and can be detected via PacBio SMRT sequencing(101) allowing genome wide coverage of the locations of a single DNA-damage type. RADAR-seq has been used to exhibit the increase in genomic rNTP incorporation after deletion of RNaseH-which is essential in rNTP removal- and will likely continue to be established as an accepted method of assessing genome-wide DNA damage.

A.6 NER in archaeal species

Some DNA damages, i.e. UV induced photoproducts, result in a distortion of the dsDNA helix which has stalling effects on critical processes such as replication and transcription. DNA repair mechanisms have evolved which detect the general distortions rather than the actual

modification to detect a broad range of DNA damages. Global nucleotide excision repair (NER) in Bacteria and Eukarya relies on enzymes to recognize the 'bulky lesion' and direct strand specific cuts on the damaged DNA strand(102). The DNA damage, now between two nicks, is thus primed for 'excision' from the DNA allowing resynthesis from the undamaged strand, and nick ligation to complete repair. In Bacteria, NER is initiated by recognition of helix distortion by the UvrA dimer and subsequent verification of actual damage by UvrB(103, 104). The activity of UvrB converts general strand distortion detection by UvrA into damage and strand specific detection, which directs the nuclease activity of UvrC either side of the DNA damage(105)- which can then be excised by the UvrD helicase. Global NER in eukarya is a more complicated system- but the core steps remain the same (Figure A.7). DNA helix distortions are first recognized by the XPC repair protein, and then damage is verified by the XPA protein to form a pre-incision complex. Helicases XPB and XPD then separate DNA strands at the site of damage- and the orientation of the resulting complex allows strand-specific cuts by XPG and XPF on either side of the site of DNA damage(106, 107). The damage containing strand is excised in complex with TFIIH(108).

Some Archaea encode homologs of bacterial Uvr proteins, but the majority encode homologs of critical eukaryotic NER proteins- in particular helicases XPB/XPD and endonuclease XPF (Figure/Table 1). No explicit NER pathway, however, has yet been defined in Archaea, and research has focused on activities individual enzymes with human counterparts, as they exist outside the context of a multi-protein complex. For example, independent structures of XPD from *Thermoplasma acidophilum* and *Sulfolobus acidocaldarius* revealed a 4 domain structure and disease causing mutants from human XPD could be mapped to functionally critical sites of the structure(109–111).

While research into archaeal XP homologs has been structurally fruitful, establishing the NER pathway in Archaea has remained challenging and elusive. Perturbations in the UvrA, UvrB, and UvrC homologs found in *Halobacterium* resulted in almost total loss of resistance to

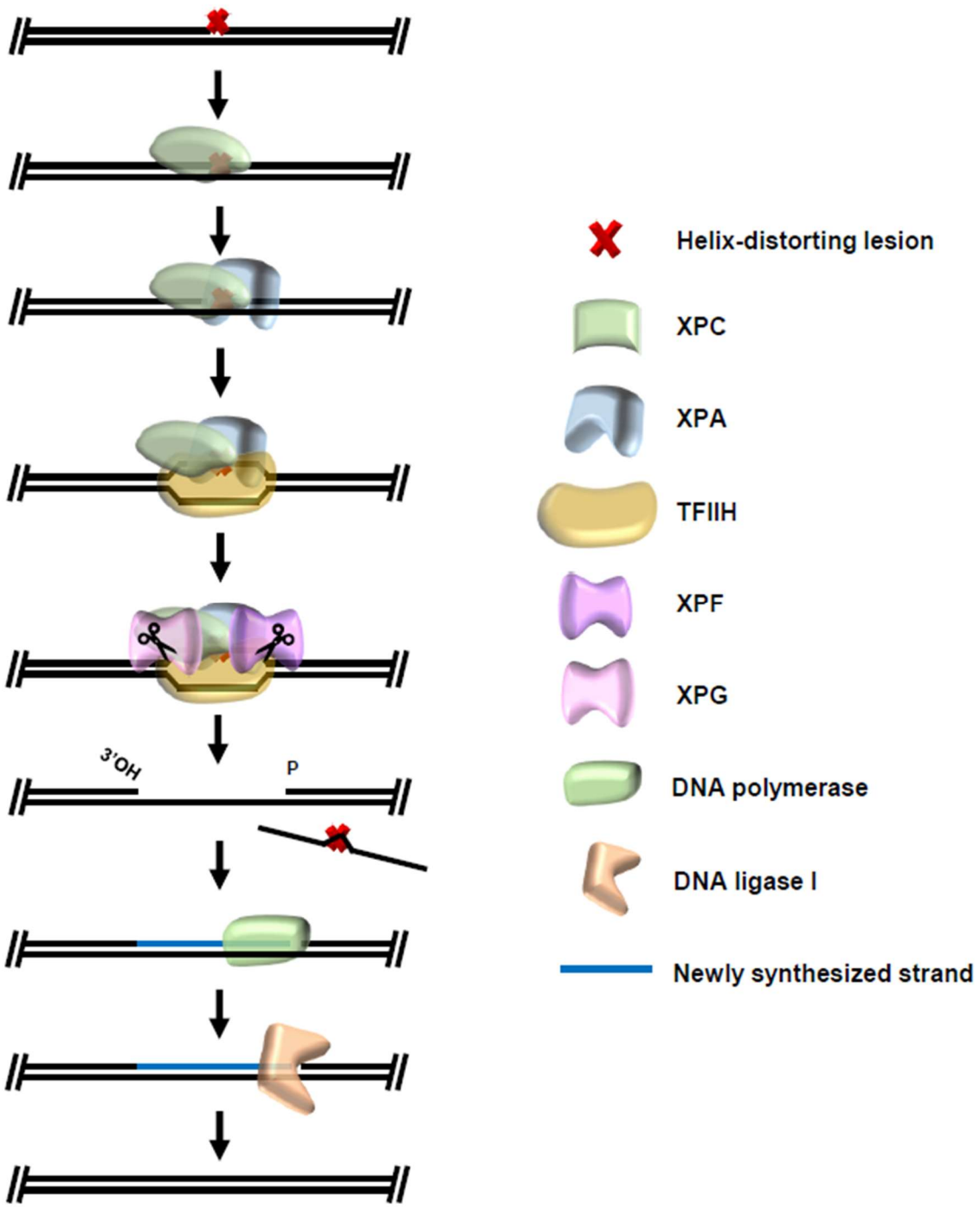


Figure A.7: Eukaryotic global genomic nucleotide excision repair. DNA damages which distort the DNA double helix are recognized by XPC, which recruits the damage recognition XPA and TFIIH complex. Components of the TFIIH complex melt strands of DNA around a verified DNA lesion, allowing cuts of the damaged strand by XPG and XPF. The TFIIH complex uses helicase activity to "excise" the damaged strand, allowing conclusion of repair by DNA polymerase and DNA ligase I.

UV exposure- but it remains unseen if these homologs function in a recognized NER pathway, and the Uvr proteins are only found in a minority of Archaea(112). Conversely, deletions of XPB, XPD, and XPF from *Thermococcus kodakarensis* resulted in only slight sensitivity to moderate doses of UV irradiation(113), suggesting these enzymes are involved in- *but not required for*- UV damage response. Additional factors could potentially play a role in archaeal NER, and in some cases the eukaryotic-like NER enzymes are paired with auxiliary nucleases. XPB helicase is sometimes found encoded in an operon with a nuclease named Bax1 and these enzyme act in concert to open a DNA bubble and make cuts(114). In many XPF encoding species, the 3'-5' exonuclease HAN is often encoded, potentially recapitulating *in vitro* experiments where XPF and HAN form a functional nuclease complex(115). Recent biochemical examinations of archaeal XPF have investigated the enzyme in the context of replication restart and Holliday junction formation(116–118), but it is possible that XPF performs multiple functions within the cell- including one in an NER pathway.

The lack of direct evidence for NER in most Archaea has led speculation there is no conserved NER pathway in the Domain and this deficiency is simply made up by increased activity of repair enzymes during stalled replisome restart after DNA polymerase is stalled by helix distortions. If a conserved NER pathway involving eukaryotic-like enzymes exists in Archaea, there remain several unrevealed details and it likely differs significantly from the eukaryotic pathway. Eukaryotic-like nucleotide excision involves strand nicking by two distinct exonucleases, XPF and XPG, but the latter is not found in Archaea. It is feasible that XPF is responsible for both cuts- or it only makes one cut in an MMR-like mechanism-but of prime importance is the question of damage recognition. There are no known homologs of the eukaryotic NER damage recognition enzyme XPC in Archaea- and thus elucidating how bulky helix-distorting lesions are detected will be of great value in establishing archaeal NER.

A.7 TC-NER in archaeal species

In Bacteria and Eukarya NER can be initiated by recognition of transcription elongation complexes (TECs) which stall upon DNA lesions entering the active site during transcription- a process termed transcription coupled DNA repair (TCR). Utilizing actively transcribing RNAPs to sense DNA damage offers an evolutionary advantage in that actively transcribed regions of the genome are actively monitored for lesions. Akin to global NER, TCR has yet to be described in Archaea but current evidence suggests it is an active pathway in some clades. While studies in crenarchaea have revealed no significant change in DNA repair of transcribed versus non-transcribed strands(119), euryarchaeal species have displayed preferential repair of transcribed DNA strands- a hallmark of TCR(120). Additionally, the archaeal RNAP- which closely resembles eukaryotic RNAPII- has been shown to stall specifically at template strand DNA damage(121).

In eukaryotes, the CSB protein acts as the transcription repair coupling factor (TRCF)- initially recognizing stalled TECs and allowing localization of TFIIH and other NER enzymes directly to the site of damage(122). In Bacteria, the transcription termination factor Mfd acts as the TRCF- simultaneously recruiting the Uvr family of NER enzymes and terminating transcription to prevent formation of mutant transcripts(123–125). There are no homologs of either CSB of Mfd found in the archaeal Domain, suggesting a potential archaeal TCR pathway- and TRCF- evolved separately. Recently, the first enzyme with transcription termination activity was reported in Archaea- but is intimately linked with other nucleic acid metabolic pathways and is a candidate for acting as the archaeal TRCF(126, 127). Euryarchaeal termination activity (Eta) requires DNA sequences upstream of RNAP, aids backtracked RNAPs, is ATP-dependent and non-competitive with elongation- all attributes shared with the bacterial TRCF Mfd. Mfd catches up to backtracked or stalled polymerases by ‘autonomously’ patrolling DNA upstream of TECs(128). Deletion of both Mfd and Eta produce a UV sensitivity phenotype, further suggesting they share an analogous role(129). However, species which encode Eta also

encode eukaryotic XP NER enzymes which have yet to be implicated in an NER pathway. Without an obvious damage recognition NER enzyme encoded, it is attractive to think of damage stalled RNAP fulfilling this role. If Eta acts as an archaeal TRCF analogous to Mfd, but recruits eukaryotic-like NER enzymes, another intriguing example of an archaeal physiological pathway with both bacterial and eukaryotic-like elements would be presented (Figure A.8), and implicate TCR as a universally conserved DNA repair pathway for the first time.

A.8 Discussion

Archaeal DNA repair-based research has offered inspiring mechanistic insight into the strategies of preserving DNA stability in extremes once thought inhospitable. At face value, such strategies of recognition of DNA damages and their preparation for the core resynthesis machinery (i.e. DNAP, Fen1, DNA Ligase) are intrinsically fascinating, but perhaps the most alluring facet of archaeal DNA repair has recently been the development of new techniques at the protein and whole-genome level as archaeal species have become more genetically accessible. Novel archaeal DNA repair enzymes will likely continue to be characterized and find new roles in the exponentially growing biotechnology world. Bioinformatic approaches, such as RADAR-seq, will continue to provide population/genomic levels DNA repair activities, and the continuously developing knowledge of DSBR in relation to CRISPR systems will surely yield more tools for geneticists. Super-resolution microscopy, once thought over-encumbered by the small size of archaeal cells, has recently become optimized and used to image foci of DSB sites in *H. volcanii*(130), offering the DNA repair research as a platform for development of more broadly applicable procedures. The continued development of these (and new) technologies, however, will only be progressed *alongside* our understanding of archaeal DNA repair as a whole- and thus identifying and answering the most pressing questions in the field must remain a priority.

Once thought a detriment to cellular health, the double strand break is appearing more

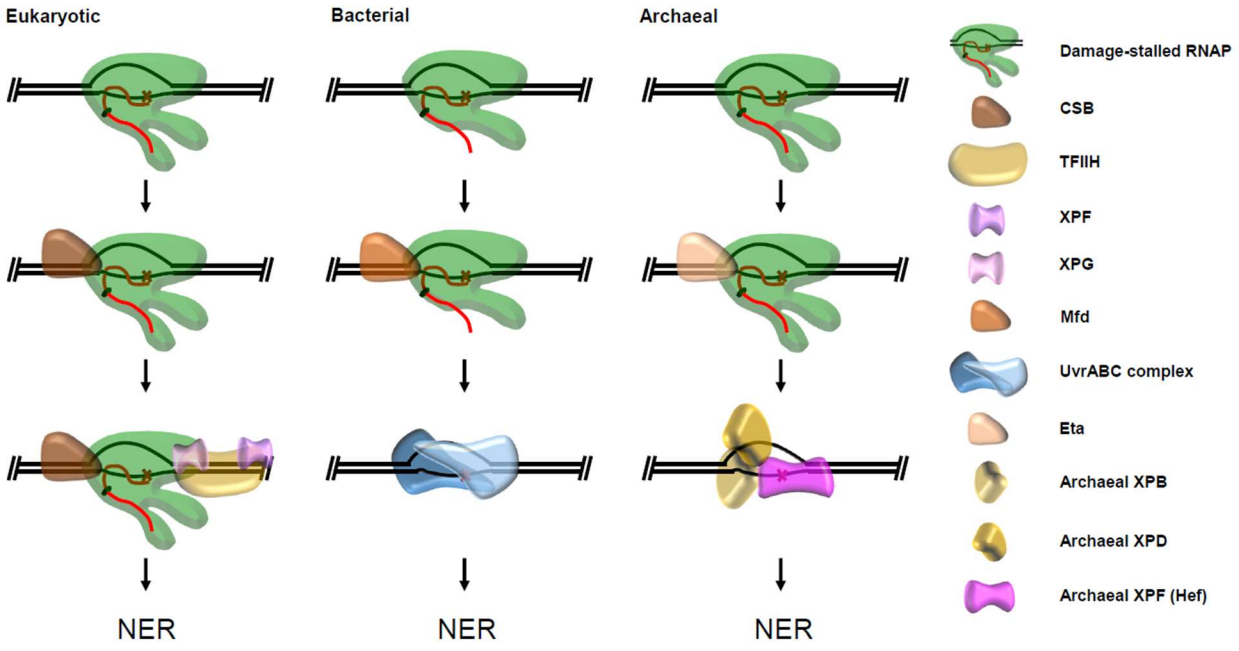


Figure A.8: Current eukaryotic and bacterial models of transcription coupled nucleotide excision repair (TC-NER) and a hypothetical archaeal model. In all cases, RNA polymerase (RNAP) is arrested at template-strand DNA damage and recognized by the TRCF-CSB in Eukarya, Mfd in Bacteria, and potentially Eta in Archaea. The TRCF either backtracks RNAP or terminates transcription while recruiting NER enzymes directly to the site of damage. Homologs of the eukaryotic XP proteins found in many Archaea act in our archaeal model.

of an essential intermediate to many metabolic processes outside of replication- potentially altering our view of archaeal metabolic biology. How Archaea deal with such an intermediate has been resolved through multiple pathways (NHEJ, MMEJ, HR), but the next challenge is understanding how cells 'decide' which of these pathways is most appropriate in a given context- and it the ploidy state influences rate of HR. One such context may be resultant DSBs from EndoMS activity during archaeal MMR which, if verified, will allow us to probe how cells use DSB substrates purposefully outside of the replisome. Generating depth to our current understanding is of great importance- but there still remain significant 'unknowns' in the field which have yet to be resolved. Are there alternative pathways for MMR of BER yet undiscovered- and can recognition enzymes be repurposed? Does NER or the transcription coupled sub-pathway (TCR) exist in Archaea, and do they more closely mirror a prokaryotic or eukaryotic system? Finally, as interconnectedness or repair and replication systems becomes more apparent, how are repair pathways regulated, segregated, or organized in the context of the prokaryotic cell? The answers to these questions will not only provide a clearer picture of DNA maintenance *in extremis,* but likely hold intriguing insights into our own ancestral metabolic history.

REFERENCES

1. D. L. Jones, B. K. Baxter, DNA repair and photoprotection: Mechanisms of overcoming environmental ultraviolet radiation exposure in halophilic archaea. *Frontiers in Microbiology* **8** (2017).

2. H. Atomi, J. Reeve, Microbe profile: *Thermococcus kodakarensis*: The model hyperthermophilic archaeon. *Microbiology (United Kingdom)* **165**, 1166–1168 (2019).

3. C. Vieille, G. J. Zeikus, Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. *Microbiology and Molecular Biology Reviews* **65**, 1–43 (2001).

4. T. Lindahl, Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).

5. C. A. Lewis, J. Crayle, S. Zhou, R. Swanstrom, R. Wolfenden, Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 8194–8199 (2016).

6. M. Ehrlich, K. F. Norris, R. Y. Wang, K. C. Kuo, C. W. Gehrke, DNA cytosine methylation and heat-induced deamination. *Bioscience Reports* **6**, 387–393 (1986).

7. Y. Ishino, I. Narumi, DNA repair in hyperthermophilic and hyperradioresistant microorganisms. *Current Opinion in Microbiology* **25**, 103–112 (2015).

8. D. W. Grogan, G. T. Carver, J. W. Drake, Genetic fidelity under harsh conditions: Analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus

acidocaldarius. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 7928–7933 (2001).

9. A. Palud, *et al.*, Intrinsic properties of the two replicative DNA polymerases of pyrococcus abyssi in replicating abasic sites: Possible role in DNA damage tolerance? *Molecular Microbiology* **70**, 746–761 (2008).

10. C. Yi, C. He, DNA repair by reversal of DNA damage. *Cold Spring Harbor Perspectives in Biology* **5** (2013).

11. A. Sancar, Mechanisms of DNA Repair by Photolyase and Excision Nuclease (Nobel Lecture). *Angewandte Chemie - International Edition* **55**, 8502–8527 (2016).

12. A. K. Sinha, *et al.*, Broken replication forks trigger heritable DNA breaks in the terminus of a circular chromosome. *PLoS Genetics* **14** (2018).

13. B. Michel, A. K. Sinha, D. R. F. Leach, Replication Fork Breakage and Restart in Escherichia coli . *Microbiology and Molecular Biology Reviews* **82** (2018).

14. G. Keijzers, *et al.*, Human exonuclease 1 (EXO1) regulatory functions in DNA replication with putative roles in cancer. *International Journal of Molecular Sciences* **20** (2019).

15. B. A. M. Bouwman, N. Crosetto, Endogenous DNA double-strand breaks during DNA transactions: Emerging insights and methods for genome-wide profiling. *Genes* **9** (2018).

16. M. Shiimori, *et al.*, Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in Pyrococcus furiosus. *Nucleic Acids Research* **45**, 11281–11294 (2017).

17. F. A. Khan, S. O. Ali, Physiological Roles of DNA Double-Strand Breaks. *Journal of Nucleic Acids* **2017**, 1–20 (2017).

18. W. J. Cannan, D. S. Pederson, Mechanisms and Consequences of Double-Strand DNA Break Formation in Chromatin. *Journal of Cellular Physiology* **231**, 3–14 (2016).

19. S. Gray, P. E. Cohen, Control of Meiotic Crossovers: From Double-Strand Break Formation to Designation. *Annual Review of Genetics* **50**, 175–210 (2016).

20. J. H. Seol, E. Y. Shim, S. E. Lee, Microhomology-mediated end joining: Good, bad and ugly. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **809**, 81–87 (2018).

21. H. H. Y. Chang, N. R. Pannunzio, N. Adachi, M. R. Lieber, Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews Molecular Cell Biology* **18**, 495–506 (2017).

22. A. M. Gehring, et al., Genome replication in Thermococcus kodakarensis independent of Cdc6 and an origin of replication. *Front. Microbiol.* **8**, 2084 (2017).

23. L. M. Kelman, Z. Kelman, Do Archaea Need an Origin of Replication? *Trends in Microbiology* **26**, 172–174 (2018).

24. M. F. White, Homologous recombination in the archaea: The means justify the ends in *Biochemical Society Transactions*, (Biochem Soc Trans, 2011), pp. 15–19.

25. E. Mladenov, S. Magin, A. Soni, G. Iliakis, DNA double-strand-break repair in higher eukaryotes and its role in genomic instability and cancer: Cell cycle and proliferation-dependent regulation. *Seminars in Cancer Biology* **37–38**, 51–64 (2016).

26. L. Käshammer, *et al.*, Mechanism of DNA End Sensing and Processing by the Mre11-Rad50 Complex. *Molecular Cell* **76**, 382-394.e6 (2019).

27. A. Quaiser, F. Constantinesco, M. F. White, P. Forterre, C. Elie, The Mre11 protein interacts with both Rad50 and the HerA bipolar helicase and is recruited to DNA

following gamma irradiation in the archaeon Sulfolobus acidocaldarius. *BMC Molecular Biology* **9** (2008).

28. G. Hogrel, *et al.*, Physical and functional interplay between PCNA DNA clamp and Mre11–Rad50 complex from the archaeon Pyrococcus furiosus. *Nucleic Acids Research* **46**, 5651–5663 (2018).

29. S. Sung, *et al.*, DNA end recognition by the Mre11 nuclease dimer: insights into resection and repair of damaged DNA . *The EMBO Journal* **33**, 2422–2435 (2014).

30. F. Constantinesco, A bipolar DNA helicase gene, herA, clusters with rad50, mre11 and nurA genes in thermophilic archaea. *Nucleic Acids Research* **32**, 1439–1447 (2004).

31. C. F, F. P, E. C, NurA, a Novel 5'-3' Nuclease Gene Linked to rad50 and mre11 Homologs of Thermophilic Archaea. *EMBO reports* **3** (2002).

32. B. B. Hopkins, T. T. Paull, The P. furiosus Mre11/Rad50 Complex Promotes 5′ Strand Resection at a DNA Double-Strand Break. *Cell* **135**, 250–260 (2008).

33. A. Quaiser, F. Constantinesco, M. F. White, P. Forterre, C. Elie, The Mre11 protein interacts with both Rad50 and the HerA bipolar helicase and is recruited to DNA following gamma irradiation in the archaeon Sulfolobus acidocaldarius. *BMC Molecular Biology* **9** (2008).

34. Q. Huang, *et al.*, Efficient 5'-3' DNA end resection by HerA and NurA is essential for cell viability in the crenarchaeon Sulfolobus islandicus. *BMC Molecular Biology* **16** (2015).

35. Z. Ahdash, *et al.*, Mechanistic insight into the assembly of the HerA–NurA helicase–nuclease DNA end resection complex. *Nucleic Acids Research* **45**, 12025–12038 (2017).

36. N. J. Rzechorzek, *et al.*, Structure of the hexameric HerA ATPase reveals a mechanism of translocation-coupled DNA-end processing in archaea. *Nature Communications* **5** (2014).

37. M. de Falco, F. Catalano, M. Rossi, M. Ciaramella, M. de Felice, NurA is endowed with endo- and exonuclease activities that are modulated by HerA: New insight into their role in DNA-end processing. *PLoS ONE* **10** (2015).

38. J. Chae, Y. C. Kim, Y. Cho, Crystal structure of the NurA–dAMP–Mn2+ complex. *Nucleic Acids Research* **40**, 2258–2270 (2012).

39. J. K. Blackwood, *et al.*, Structural and functional insights into DNA-end processing by the archaeal HerA helicase–NurA nuclease complex. *Nucleic Acids Research* **40**, 3183–3196 (2012).

40. K. Komori, *et al.*, Both RadA and RadB are involved in homologous recombination in Pyrococcus furiosus. *Journal of Biological Chemistry* **275**, 33782–33790 (2000).

41. D. S. Shin, Full-length archaeal Rad51 structure and mutants: mechanisms for RAD51 assembly and control by BRCA2. *The EMBO Journal* **22**, 4566–4576 (2003).

42. S. Haldenby, M. F. White, T. Allers, RecA family proteins in archaea: RadA and its cousins. *Biochemical Society Transactions* **37**, 102–107 (2009).

43. L. J, *et al.*, Pathways and Assays for DNA Double-Strand Break Repair by Homologous Recombination. *Acta biochimica et biophysica Sinica* **51** (2019).

44. T. Miura, *et al.*, Homologous recombination via synthesis-dependent strand annealing in yeast requires the IRC20 and SRS2 DNA helicases. *Genetics* **191**, 65–78 (2012).

45. C. P. Guy, Archaeal Hel308 helicase targets replication forks in vivo and *in vitro* and unwinds lagging strands. *Nucleic Acids Research* **33**, 3678–3690 (2005).

46. S. J. Northall, *et al.*, DNA binding and unwinding by Hel308 helicase requires dual functions of a winged helix domain. *DNA Repair* **57**, 125–132 (2017).

47. K. Xu, X. Wu, J. D. Tompkins, C. Her, Assessment of anti-recombination and double-strand break-induced gene conversion in human cells by a chromosomal reporter. *Journal of Biological Chemistry* **287**, 29543–29553 (2012).

48. K. Komori, S. Sakae, H. Shinagawa, K. Morikawa, Y. Ishino, A holliday junction resolvase from Pyrococcus furiosus: Functional similarity to Escherichia coli RuvC provides evidence for conserved mechanism of homologous recombination in Bacteria, Eukarya, and Archaea. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 8873–8878 (1999).

49. K. Komori, Biochemical characterization of the Hjc Holliday junction resolvase of Pyrococcus furiosus. *Nucleic Acids Research* **28**, 4544–4551 (2000).

50. P. M. Dehé, *et al.*, Regulation of Mus81-Eme1 Holliday junction resolvase in response to DNA damage. *Nature Structural and Molecular Biology* **20**, 598–603 (2013).

51. M. G. Blanco, J. Matos, S. C. West, Dual Control of Yen1 Nuclease Activity and Cellular Localization by Cdk and Cdc14 Prevents Genome Instability. *Molecular Cell* **54**, 94–106 (2014).

52. Q. Huang, *et al.*, Phosphorylation of the archaeal holliday junction resolvase hjc inhibits its catalytic activity and facilitates DNA repair in Sulfolobus islandicus REY15A. *Frontiers in Microbiology* **10** (2019).

53. C. Hildenbrand, T. Stock, C. Lange, M. Rother, J. Soppa, Genome copy numbers and gene conversion in methanogenic archaea. *Journal of Bacteriology* **193**, 734–743 (2011).

54. K. A. Dulmage, C. L. Darnell, A. Vreugdenhil, A. K. Schmid, Copy number variation is associated with gene expression change in archaea. *Microbial Genomics* **4** (2018).

55. S. Sharma, *et al.*, Homology and enzymatic requirements of microhomology-dependent alternative end joining. *Cell death & disease* **6**, e1697 (2015).

56. D. D. Nayak, W. W. Metcalf, Cas9-mediated genome editing in the methanogenic archaeon Methanosarcina acetivorans. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 2976–2981 (2017).

57. C. Zhang, R. J. Whitaker, Microhomology-mediated high-throughput gene inactivation strategy for the hyperthermophilic crenarchaeon Sulfolobus islandicus. *Applied and Environmental Microbiology* **84** (2018).

58. E. J. Bartlett, N. C. Brissett, A. J. Doherty, Ribonucleolytic resection is required for repair of strand displaced nonhomologous end-joining intermediates. *Proceedings of the National Academy of Sciences of the United States of America* **110** (2013).

59. E. J. Bartlett, N. C. Brissett, P. Plocinski, T. Carlberg, A. J. Doherty, Molecular basis for DNA strand displacement by NHEJ repair polymerases. *Nucleic Acids Research* **44**, 2173–2186 (2016).

60. S. Delmas, L. Shunburne, H. P. Ngo, T. Allers, Mre11-Rad50 promotes rapid repair of DNA damage in the polyploid archaeon Haloferax volcanii by restraining homologous recombination. *PLoS Genetics* **5** (2009).

61. T. T. Paull, M. Gellert, A mechanistic basis for Mre11-directed DNA joining at microhomologies. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 6409–6414 (2000).

62. A. Kish, J. C. Gaillard, J. Armengaud, C. Elie, Post-translational methylations of the archaeal Mre11: Rad50 complex throughout the DNA damage response. *Molecular Microbiology* **100**, 362–378 (2016).

63. J. M. Craig, *et al.*, Determining the effects of DNA sequence on Hel308 helicase translocation along single-stranded DNA using nanopore tweezers. *Nucleic Acids Research* **47**, 2506–2513 (2019).

64. U. Gophna, T. Allers, A. Marchfelder, Finally, Archaea Get Their CRISPR-Cas Toolbox. *Trends in Microbiology* **25**, 430–432 (2017).

65. A. Castañeda-Garciá, *et al.*, A non-canonical mismatch repair pathway in prokaryotes. *Nature Communications* **8** (2017).

66. N. Takemoto, I. Numata, M. Su'etsugu, T. Miyoshi-Akiyama, Bacterial EndoMS/NucS acts as a clamp-mediated mismatch endonuclease to prevent asymmetric accumulation of replication errors. *Nucleic Acids Research* **46**, 6152–6165 (2018).

67. B. Meier, *et al.*, Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. *Genome Research* **28**, 666–675 (2018).

68. P. Modrich, Mechanisms in E. coli and Human Mismatch Repair (Nobel Lecture). *Angewandte Chemie - International Edition* **55**, 8490–8501 (2016).

69. J. S. Lenhart, M. C. Pillon, A. Guarné, J. S. Biteen, L. A. Simmons, Mismatch repair in Gram-positive bacteria. *Research in Microbiology* **167**, 4–12 (2016).

70. Z. Li, A. H. Pearlman, P. Hsieh, DNA mismatch repair and the DNA damage response. *DNA Repair* **38**, 94–101 (2016).

71. T. A. Kunkel, D. A. Erie, Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annual Review of Genetics* **49**, 291–313 (2015).

72. A. Minobe, *et al.*, Biochemical characterization of mismatch-binding protein MutS1 and nicking endonuclease MutL from a euryarchaeon Methanosaeta thermophila. *DNA Repair* **75**, 29–38 (2019).

73. C. R. Busch, J. DiRuggiero, MutS and MutL are dispensable for maintenance of the genomic mutation rate in the halophilic archaeon Halobacterium salinarum NRC-1. *PLoS ONE* **5** (2010).

74. S. Ishino, *et al.*, Identification of a mismatch-specific endonuclease in hyperthermophilic Archaea. *Nucleic Acids Research* **44**, 2977–2986 (2016).

75. S. Ishino, *et al.*, Activation of the mismatch-specific endonuclease EndoMS/NucS by the replication clamp is required for high fidelity DNA replication. *Nucleic Acids Research* **46**, 6206–6217 (2018).

76. M. Ariyoshi, K. Morikawa, A Dual Base Flipping Mechanism for Archaeal Mismatch Repair. *Structure* **24**, 1859–1861 (2016).

77. S. Nakae, *et al.*, Structure of the EndoMS-DNA Complex as Mismatch Restriction Endonuclease. *Structure* **24**, 1960–1971 (2016).

78. N. Takemoto, I. Numata, M. Su'etsugu, T. Miyoshi-Akiyama, Bacterial EndoMS/NucS acts as a clamp-mediated mismatch endonuclease to prevent asymmetric accumulation of replication errors. *Nucleic Acids Research* **46**, 6152–6165 (2018).

79. M. Lemor, *et al.*, Differential Activities of DNA Polymerases in Processing Ribonucleotides during DNA Synthesis in Archaea. *Journal of Molecular Biology* **430**, 4908–4924 (2018).

80. M. R. Heider, B. W. Burkhart, T. J. Santangelo, A. F. Gardner, Defining the RNaseH2 enzyme-initiated ribonucleotide excision repair pathway in Archaea. *Journal of Biological Chemistry* **292**, 8835–8845 (2017).

81. A. Gardner, Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Research* **27**, 2545–2553 (1999).

82. C. Ban, B. Ramakrishnan, M. Sundaralingam, A Single 2′-Hydroxyl Group Converts B-DNA to A-DNA. *Journal of Molecular Biology* **236**, 275–285 (1994).

83. B. Hiller, *et al.*, Ribonucleotide excision repair is essential to prevent squamous cell carcinoma of the skin. *Cancer Research* **78**, 5917–5926 (2018).

84. A. Vaisman, *et al.*, Investigating the mechanisms of ribonucleotide excision repair in Escherichia coli. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **761**, 21–33 (2014).

85. J. S. Williams, T. A. Kunkel, Ribonucleotides in DNA: Origins, repair and consequences. *DNA Repair* **19**, 27–37 (2014).

86. B. W. Burkhart, *et al.*, The GAN exonuclease or the flap endonuclease Fen1 and RNase HII are necessary for viability of Thermococcus kodakarensis. *Journal of Bacteriology* **199** (2017).

87. M. A. M. Reijns, *et al.*, Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. *Cell* **149**, 1008–1022 (2012).

88. N. Chatterjee, G. C. Walker, Mechanisms of DNA damage, repair, and mutagenesis. *Environmental and Molecular Mutagenesis* **58**, 235–263 (2017).

89. M. Dizdaroglu, P. Jaruga, Mechanisms of free radical-induced damage to DNA. *Free Radical Research* **46**, 382–419 (2012).

90. A. P. Schuch, N. C. Moreno, N. J. Schuch, C. F. M. Menck, C. C. M. Garcia, Sunlight damage to cellular DNA: Focus on oxidatively generated lesions. *Free Radical Biology and Medicine* **107**, 110–124 (2017).

91. S. S. Wallace, Base excision repair: A critical player in many games. *DNA Repair* **19**, 14–26 (2014).

92. A. M. Gehring, *et al.*, Biochemical reconstitution and genetic characterization of the major oxidative damage base excision DNA repair pathway in Thermococcus kodakarensis. *DNA Repair* **86** (2020).

93. L. Zhang, *et al.*, Biochemical characterization and mutational studies of the 8-oxoguanine DNA glycosylase from the hyperthermophilic and radioresistant archaeon Thermococcus gammatolerans. *Applied Microbiology and Biotechnology* **103**, 8021–8033 (2019).

94. S. Kanugula, G. T. Pauly, R. C. Moschel, A. E. Pegg, A bifunctional DNA repair protein from Ferroplasma acidarmanus exhibits O6-alkylguanine-DNA alkyltransferase and endonuclease V activities. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3617–3622 (2005).

95. Y. Wang, *et al.*, Biochemical characterization of a thermostable endonuclease V from the hyperthermophilic euryarchaeon Thermococcus barophilus Ch5. *International Journal of Biological Macromolecules* **117**, 17–24 (2018).

96. S. Kiyonari, Y. Egashira, S. Ishino, Y. Ishino, Biochemical characterization of endonuclease V from the hyperthermophilic archaeon, Pyrococcus furiosus. *Journal of Biochemistry* **155**, 325–333 (2014).

97. M. Shiraishi, *et al.*, A novel endonuclease that may be responsible for damaged DNA base repair in Pyrococcus furiosus. *Nucleic Acids Research* **43**, 2853–2863 (2015).

98. K. Miyazono, *et al.*, Crystal structure of the novel lesion-specific endonuclease PfuEndoQ from Pyrococcus furiosus. *Nucleic Acids Research* **46**, 4807–4818 (2018).

99. M. Shiraishi, S. Iwai, Molecular basis of substrate recognition of endonuclease Q from the euryarchaeon pyrococcus furiosus. *Journal of Bacteriology* **202** (2020).

100. K. M. Zatopek, *et al.*, RADAR-seq: A RAre DAmage and Repair sequencing method for detecting DNA damage on a genome-wide scale. *DNA Repair* **80**, 36–44 (2019).

101. J. Eid, *et al.*, Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

102. M. G. Kemp, “Damage removal and gap filling in nucleotide excision repair” in *Enzymes*, (Academic Press, 2019), pp. 59–97.

103. J. J. Truglio, D. L. Croteau, B. van Houten, C. Kisker, Prokaryotic nucleotide excision repair: The UvrABC system. *Chemical Reviews* **106**, 233–252 (2006).

104. M. Jaciuk, E. Nowak, K. Skowronek, A. Tañska, M. Nowotny, Structure of UvrA nucleotide excision repair protein in complex with modified DNA. *Nature Structural and Molecular Biology* **18**, 191–198 (2011).

105. M. Jaciuk, *et al.*, A combined structural and biochemical approach reveals translocation and stalling of UvrB on the DNA lesion as a mechanism of damage verification in bacterial nucleotide excision repair. *DNA Repair* **85** (2020).

106. A. Sancar, J. T. Reardon, Nucleotide excision repair in E. coli and man. *Advances in Protein Chemistry* **69**, 43–71 (2004).

107. G. Spivak, Nucleotide excision repair in humans. *DNA Repair* **36**, 13–18 (2015).

108. M. G. Kemp, J. T. Reardon, L. A. Lindsey-Boltz, A. Sancar, Mechanism of release and fate of excised oligonucleotides during nucleotide excision repair. *Journal of Biological Chemistry* **287**, 22889–22899 (2012).

109. L. Fan, *et al.*, XPD Helicase Structures and Activities: Insights into the Cancer and Aging Phenotypes from XPD Mutations. *Cell* **133**, 789–800 (2008).

110. J. Kuper, S. C. Wolski, G. Michels, C. Kisker, Functional and structural studies of the nucleotide excision repair helicase XPD suggest a polarity for DNA translocation. *EMBO Journal* **31**, 494–502 (2012).

111. S. C. Wolski, *et al.*, Crystal structure of the FeS cluster-containing nucleotide excision repair helicase XPD. *PLoS Biology* **6**, 1332–1342 (2008).

112. D. Crowley, *et al.*, The uvrA, uvrB and uvrC genes are required for repair of ultraviolet light induced DNA photoproducts in Halobacterium sp. NRC-1. *Saline Systems* **2**, 11 (2006).

113. R. Fujikane, S. Ishino, Y. Ishino, P. Forterre, Genetic analysis of DNA repair in the hyperthermophilic archaeon, Thermococcus kodakaraensis. *Genes and Genetic Systems* **85**, 243–257 (2010).

114. C. Rouillon, M. F. White, The XBP-Bax1 helicase-nuclease complex unwinds and cleaves DNA: Implications for eukaryal and archaeal nucleotide excision repair. *Journal of Biological Chemistry* **285**, 11013–11022 (2010).

115. L. Feng, *et al.*, The trimeric Hef-associated nuclease HAN is a 3′→5′ exonuclease and is probably involved in DNA repair. *Nucleic Acids Research* **46**, 9027–9043 (2018).

116. R. Lestini, *et al.*, Intracellular dynamics of archaeal FANCM homologue Hef in response to halted DNA replication. *Nucleic Acids Research* **41**, 10358–10370 (2013).

117. R. Lestini, F. Delpech, H. Myllykallio, DNA replication restart and cellular dynamics of Hef helicase/nuclease protein in Haloferax volcanii. *Biochimie* **118**, 254–263 (2015).

118. R. Lestini, Z. Duan, T. Allers, The archaeal Xpf/Mus81/FANCM homolog Hef and the Holliday junction resolvase Hjc define alternative pathways that are essential for cell viability in Haloferax volcanii. *DNA Repair* **9**, 994–1002 (2010).

119. R. Dorazi, D. Götz, S. Munro, R. Bernander, M. F. White, Equal rates of repair of DNA photoproducts in transcribed and non-transcribed strands in Sulfolobus solfataricus. *Molecular Microbiology* **63**, 521–529 (2007).

120. N. Stantial, J. Dumpe, K. Pietrosimone, F. Baltazar, D. J. Crowley, Transcription-coupled repair of UV damage in the halophilic archaea. *DNA Repair* **41**, 63–68 (2016).

121. A. M. Gehring, T. J. Santangelo, Archaeal RNA polymerase arrests transcription at DNA lesions. *Transcription* **8**, 288–296 (2017).

122. J. Xu, *et al.*, Structural basis for the initiation of eukaryotic transcription-coupled DNA repair. *Nature* **551**, 653–657 (2017).

123. A. M. Deaconescu, M. M. Suhanovsky, From Mfd to TRCF and Back Again—A Perspective on Bacterial Transcription-coupled Nucleotide Excision Repair. *Photochemistry and Photobiology* **93**, 268–279 (2017).

124. A. M. Deaconescu, A. Sevostyanova, I. Artsimovitch, N. Grigorieff, Nucleotide excision repair (NER) machinery recruitment by the transcription-repair coupling factor involves unmasking of a conserved intramolecular interface. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3353–3358 (2012).

125. O. Adebali, Y. Y. Chiou, J. Hu, A. Sancar, C. P. Selby, Genome-wide transcription-coupled repair in Escherichia coli is mediated by the Mfd translocase. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E2116–E2125 (2017).

126. J. E. Walker, O. Luyties, T. J. Santangelo, Factor-dependent archaeal transcription termination. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E6767–E6773 (2017).

127. D. K. Phung, *et al.*, RNA processing machineries in Archaea: the 5′-3′ exoribonuclease aRNase J of the β-CASP family is engaged specifically with the helicase ASH-Ski2 and the 3′-5′ exoribonucleolytic RNA exosome machinery. *Nucleic Acids Research* **48**, 3832–3847 (2020).

128. T. T. Le, *et al.*, Mfd Dynamically Regulates Transcription via a Release and Catch-Up Mechanism. *Cell* **172**, 344-357.e15 (2018).

129. B. J. Schalow, C. T. Courcelle, J. Courcelle, Mfd is required for rapid recovery of transcription following UV-induced DNA damage but not oxidative DNA damage in Escherichia coli. *Journal of Bacteriology* **194**, 2637–2645 (2012).

130. F. Delpech, *et al.*, Snapshots of archaeal DNA replication and repair in living cells using super-resolution imaging. *Nucleic Acids Research* (2018) https:/doi.org/10.1093/nar/gky829.

APPENDIX B

THE ROLE OF ARCHAEAL CHROMATIN IN TRANSCRIPTION[4]

B.1 Introduction

The regulation imposed on gene expression by chromatin or nucleoid structures in Eukarya and Bacteria, respectively, has a long and rich history (1–10). Organization of the genome can facilitate or impair the ability of the transcription apparatus to recognize promoter elements, to form an open complex and to transition into stable elongation. Once transcription elongation complexes (TECs) are established, they must traverse a protein-bound template (11–14). The dynamic associations of DNA-bound proteins and the resultant larger structures formed by cooperative interactions of such hinder translocation. Both bacterial nucleoid and eukaryotic chromatin structures involve the formation of loops, connecting spatially distant locations on the genome via protein-DNA interactions (15, 16), and the formation and stability of such topologically-constrained regions can be controlled to alter expression of single loci or very large regions of the genome. Regulation of gene expression through alteration of genomic architecture offers the potential to tailor gene expression to maximize fitness gains in changing environments.

The role of genomic architecture in modulating gene expression in archaeal species has only more recently been investigated with the scrutiny applied to bacterial and eukaryotic systems. Archaeal genomes are typically circular, small (< 5 Mbp), gene dense (~80-90%

[4]This appendix was previously published under the following title: Marshall, C. J., Sanders, T. J., and Santangelo, T. J. (2019) 'The role of archaeal chromatin in transcription.', *J Mol Biol*. 431(20):4103-4115.

coding sequence), and many genes are organized within operons (17–21). Despite sharing many hallmarks of typical bacterial genomes, archaeal genomes are expressed with a single RNA polymerase (RNAP) that shares more similarity in overall structure, subunit composition, and basal-transcription factor requirements with eukaryotic RNAPs, in particular Pol II (22–34). The archaeal RNAP lacks the C-terminal repeats found on Pol II and is not known to be post-translationally modified, but the archaeal RNAP is still directed to the transcription start site in a manner comparable to Pol II. The archaeal transcription system is a component simplified version of the Pol II apparatus, requiring only interactions with Transcription Factor B (TFB - TFIIB in Eukarya), Transcription Factor E (TFE - TFIIE in Eukarya) and TATA Binding Protein (TBP) to recognize core archaeal promoter elements – TATA box and BRE – that share sequence conservation with eukaryotic promoter elements (31, 35–38). Minimal evidence for long-range interactions between transcription factors and promoter elements is known, and substantial evidence has instead emerged that demonstrates that most archaeal promoters are regulated by bacterial-like repressors or activators that bind immediately adjacent to or overlapping core promoter elements (39–49). Studies suggest that core promoters are generally devoid of organized chromatin structures (50, 51), and that when present, the binding affinity of transcription regulators outcompetes the binding of histones or nucleoid-associated proteins to permit regulation within an organized and protein-bound genome (40, 42, 46, 52).

Following transcription initiation, TECs must stably associate with and transcribe the template for long periods (e.g. minutes or hours at ~40 nt/sec), necessarily displacing DNA-bound proteins that impede translocation. Transcription initiation and elongation in eukaryotes is facilitated by the combinatorial activities of transcription factors and chromatin remodeling and modification machinery. Given the absence of obvious chromatin remodeling and modification machinery in archaeal genomes, transcription factors likely play the dominant role in aiding archaeal transcription during initiation and elongation. The rates of elongation and pausing of archaeal transcription are regulated by conserved archaeal-eukaryotic factors Spt4/Spt5 and

TFS (TFIIS in eukaryotes) (13). Spt5, homologous to bacterial NusG, is the only universally conserved transcription factor. Spt5-RNAP interactions facilitate formation of the closed-clamp configuration of RNAP that aids in processive elongation. Pausing is inevitable, and when collisions with DNA-bound proteins stalls forward translocation of RNAP, reverse translocation can inactivate RNAP. The cleavage-stimulatory activity of TFS/TFIIS (53) – analogous to the cleavage stimulatory activities of GreA and GreB in Bacteria (54, 55)– helps rescue such backtracked complexes, and the activity of TFS is essential for archaeal species (13).

In this review, we discuss recent advancements in archaeal chromatin and genome organization in the context of transcription regulation. We first examine the architectural mechanisms and regulatory implications of genome compaction dominated by archaeal histone proteins. Most archaeal clades encode histone proteins that generate DNA structures remarkably similar to eukaryotic nucleosomes, albeit with only the core histone-fold and often with only a single histone isoform. We next identify and outline important advancements in the identification of transcription factors and basal transcription mechanisms that facilitate transcription in the context of an archaeal histone-based chromatin landscape. We then focus on the archaeal clades that lack histone proteins and instead encode a suite of small basic proteins that presumably function like bacterial nucleoid-associated proteins to condense and organize the archaeal genome. Finally, we consider the major bottlenecks within the archaeal transcription field in the context of chromatin organized genomic architectures. We conclude with discussion of current debates within the field and highlight the future potential of studies investigating the influence of genomic architecture on archaeal gene expression.

B.2 Archaeal histone-based chromatin

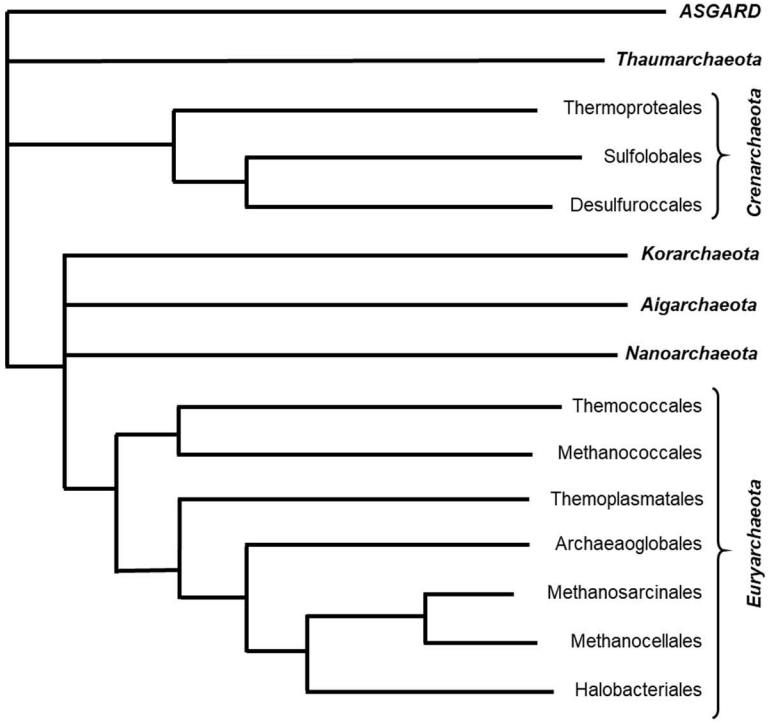
B.2.1 Structure of archaeal histone-based chromatin

Whole genome sequencing of many cultured and many-more environmentally-isolated, but not yet cultured Archaea suggests that most archaeal lineages encode one or more histone proteins

(Figure B.1) (56–64) – six histone isoforms can be identified in *Methanocaldococcus jannaschii* (65) – that are likely to organize the genome into structures that mimic DNA organization by eukaryotic nucleosomes (56, 66, 67). Although not universally encoded (typically to the exclusion of *Crenarchaeota* (56, 64)), in archaeal species with histone proteins, a chromatin landscape presents barriers to initiation (41, 52, 65, 68–70), elongation (12, 13), and likely influences termination. Archaeal histones are composed of only the core-histone fold and lack the N- and C-terminal tails and extensions common to the canonical eukaryotic histones (66, 67, 71–75) (e.g. H2A, H2B, H3, and H4). Archaeal genomes do not encode obvious linker histones (e.g. H1), nor chromatin-remodeling complexes that are abundant and essential for gene expression in eukaryotes. Unlike the mandatory eukaryotic histone heterodimer partnerships, archaeal chromatin can be spontaneously assembled with a single histone protein (51, 66, 67, 72, 76, 77), and there is currently no evidence for post-translational modification of archaeal histones.

Despite this minimalist approach to histone-based chromatin architecture, archaeal histone-DNA interactions align to the same nucleosome positioning code that was established for Eukarya (10, 51, 63), and the constrained structure of DNA bound by archaeal histones is nearly identical to the structure of DNA in the eukaryotic nucleosome (Figure B.2) (66, 67, 78, 79). The superhelically-wrapped DNA shares the geometry, diameter, pitch, and writhe of the eukaryotic nucleosomal superhelix, and specific protein-DNA contacts that stabilize archaeal histone-based chromatin are conserved in eukaryotes (56, 66, 67, 80). The structure of archaeal histone-based chromatin suggests the architectural function of histones (*i.e.* the ability to bend DNA into the nucleosomal superhelix) was established long (>1 bya) ago, and that the 'signaling functions' (*i.e.* addition of histone extensions and epigenetic modifications) were a secondary addition that came with the expansion to four canonical histones in eukaryotes (66, 67).

Local histone-binding is known to sterically compete with binding of transcription components and offers regulatory potential (12, 40–42, 81), and the extended structure of



	Histone(s)	Nucleoid Associated Proteins					
		Alba	Sul7	Cren7	Sso10a	CC1	TrmBL
ASGARD	✓	✓					✓
Thaumarchaeota	✓	✓					✓
Thermoproteales (Crenarchaeota)	✓	✓		✓	✓	✓	✓
Sulfolobales (Crenarchaeota)		✓	✓	✓	✓		✓
Desulfuroccales (Crenarchaeota)		✓		✓	✓	✓	✓
Korarchaeota	✓	✓					✓
Aigarchaeota	✓						✓
Nanoarchaeota	✓	✓					✓
Themococcales (Euryarchaeota)	✓	✓					✓
Methanococcales (Euryarchaeota)	✓	✓					✓
Themoplasmatales (Euryarchaeota)	✓	✓					✓
Archaeaoglobales (Euryarchaeota)	✓	✓					✓
Methanosarcinales (Euryarchaeota)	✓						✓
Methanocellales (Euryarchaeota)	✓	✓					✓
Halobacteriales (Euryarchaeota)	✓						✓

Figure B.1: Distribution of chromatin associated proteins identified across the Archaea. Histone proteins and nucleoid-associated proteins (NAPs; right) encoded in each phylum according to the schematic evolutionary tree of Archaea (left).

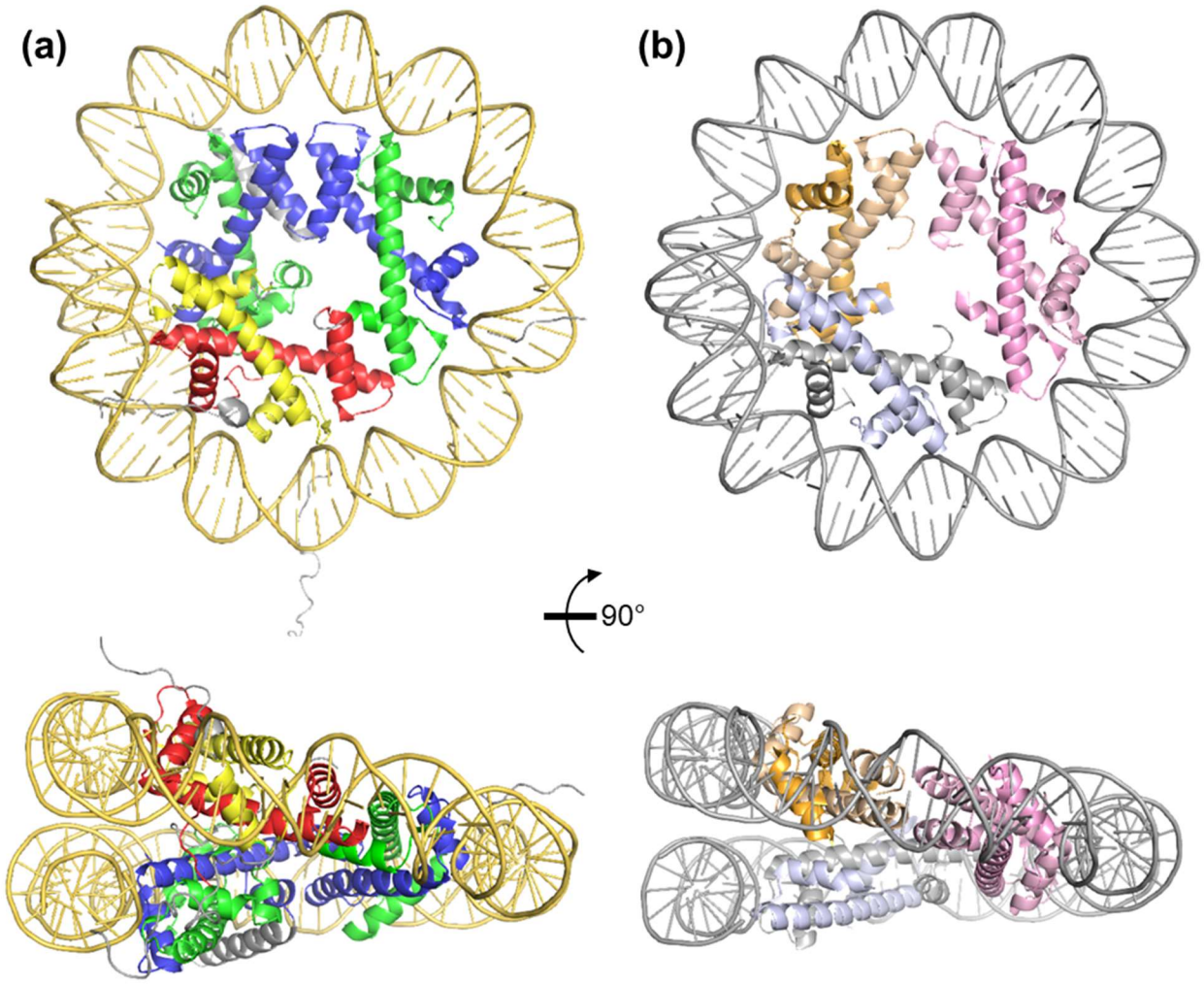


Figure B.2: The structure of histone-based chromatin in Archaea mirrors that of the eukaryotic nucleosome.
(a) The eukaryotic nucleosome hexamer containing two H3-H4 dimers (blue, green respectively) and one H2A-H2B dimer (yellow, red respectively) with wrapped DNA (gold) from a top-down and side view. N and C terminal extensions, specific to eukaryotic histones, are shown in grey. (b) Histone based-chromatin in Archaea can form from varied numbers of histone dimers (three dimers are shown here for comparison to the eukaryotic hexasome), with wrapped DNA (silver) from a top-down and side view. The archaeal histone-based chromatin structure formed with three histone dimers is almost identical to the eukaryotic hexasome without the N- and C-terminal extensions.

archaeal histone-based chromatin may also offer regulatory potential. Perhaps the most striking feature of the structure of archaeal histone-based chromatin is the continuous helical ramp of histone dimers and the close association of adjacent layers of the complex that result in a tightly-packed 3D chromatin structure (66, 67). The extensions common to eukaryotic histones normally radiate into solution and facilitate nucleosome-nucleosome interactions. The absence of such extensions on archaeal histones in part permits the close association of adjacent layers of archaeal chromatin. The resultant superstructure places the L1 loops of histone-dimers 1 and 4 along the helical ramp in closest-proximity to each other. Apart from four helix-bundles that link the histone dimers, the only region of close contact between the adjacent layers of archaeal chromatin is where the L1 regions of dimers 1 & 4 meet. L1 sequences almost always retain a central glycine at the point of closest approach and substitution of this glycine with larger side-chains impedes tight packing of archaeal chromatin, impairs gene expression *in vivo*, and reduces overall fitness (67).

Extension of the structure by one additional histone-dimer extends the length of DNA protection by ~30 bp, resulting in extended polymers that protect DNA from minimally ~60 bp (two histone dimers) to ~480 bp, in 30bp increments (82). Comparisons of archaeal histone sequences with the atomic-resolution structure of archaeal chromatin reveals that most archaeal histones retain the residues that directly interact with the DNA backbone, use nearly identical residues to stabilize histone-histone and histone-DNA interactions, and that close association of chromatin gyres is likely possible due to minimal side chains in the L1-L1 interface. The eukaryotic nuclear RNA polymerases (RNAPs) and the archaeal RNAP thus regularly encounter – and must overcome – nearly identical histone-DNA contacts that present barriers to transcription elongation (67, 83–86).

In contrast to eukaryotic histones, there is no evidence of post-translational modifications to archaeal histones. Although there are many acetyl- and methyl-transferases encoded throughout the Archaea, no activity towards histone proteins has been reported, and the bulk of

characterized acetyl- and methyl-transferases are active on DNA or RNA (87–89). A minority of archaeal organisms encode histones which contain sequences beyond the core histone fold. Excluding single histone isoforms that contain a fused second histone fold (effectively a histone-dimer within a single polypeptide) extended histone sequences are rarely observed (56). Such extensions are not homologous to those found in eukaryotes but are 'eukaryote-like' in being rich in charged residues, especially lysine. Investigation of one such extended archaeal histone variant, MJ1647, a C-terminal extension-containing histone in *M. jannaschii*, demonstrated that the C-terminal extension was critical for DNA binding and the formation of higher order structures (90). Modeling the C-terminal extension of MJ1647 into the atomic structure of archaeal histone-based chromatin suggests that the C-terminal extension might impair continued polymerization and impact the global structure of archaeal chromatin. The discovery of new histone variants and histone proteins with extensions in the *Heimdallarchaeota* and *ASGARD* archaeal clades hints at the expansion to four canonical histones, the exchange of a histone-polymer for discrete nucleosome particles, and the regulation imposed by post-translational modifications of the histone proteins in all Eukarya (56, 57, 61, 66, 67). Structural modeling of *Heimdall LC 3* histones, which contain tails similar in length and sequence composition to extensions on eukaryotic H4, suggests that the extended archaeal histone-based chromatin structure will not be impacted by inclusion of such tails (56). It will now be important to elucidate the expression, abundance, and function of these archaeal histone variants, included extended-histone variants, in controlling genomic architecture and gene expression.

B.2.2 Global regulation of transcription by archaeal histone-based chromatin

A consensus surrounding the role of archaeal histones in transcription regulation is dubious. This is highlighted by the varying essentiality of histone proteins across archaeal species. Controversy on the role of archaeal histones in controlling gene expression exists at the total transcriptome level when genetically-accessible archaeal species have their histone-encoding

loci deleted or modified. In the euryarchaeaon *Thermococcus kodakarensis,* two histone variants are encoded, and while each individually is not essential, attempts at deleting both histones have been unsuccessful indicating histone-based chromatin is critical for regulation of cellular processes. The importance of regulated genomic architecture was revealed by changes – up to ~10-fold – in the expression of ~5% of genes upon deletion of either histone isoform (91). The importance of tightly-packed 3D archaeal histone-based chromatin was demonstrated by introduction of histone variants with specific mutations to residues in the L1-L1 interface (67). Replacing G17 with bulkier amino acid residues does not disrupt local DNA binding but does disrupt extended chromatin structures that in turn impact gene expression. Disruption of extended histone-based chromatin structures also abrogates adaptive gene expression necessary to respond to changing environmental conditions.

Histone-proteins are not encoded in all species (Figure B.1) and histones are not essential for some extant histone-encoding archaeal species. The sole histone encoded in the methanogen *Methanosarcina mazei* is dispensable but deletion results in reduced growth, increased sensitivity to DNA damaging agents, reduced overall transcription for many genes, and an altered overall transcriptome (92). Thus, although non-essential, deletion of the histone-encoding locus, and thus the presumptive loss of histone-based genomic organization – does significantly impact global transcription. Changes to global gene expression and growth were restored upon complementation of *M. mazei* strains with exogenously produced histone protein, suggesting that histone-based genomic architecture is important, but not essential in some archaeal species. A potentially different view of the role of archaeal histones emerges from studies of halophilic (e.g. salt-loving) archaea. *Halobacterium salinarum* encodes just one histone protein with several unique attributes. Unlike the typical basic pI of most histone proteins, the high-intracellular salt concentrations of halophiles (~ 4 M) has likely resulted in retention of a histone with an acidic pI. The halophilic histone proteins are also typically a single polypeptide containing two tandemly-repeated histone folds. The single *H. salinarum* histone,

like the single *M. mazei* histone, is dispensable and deletion results in globally significant, but mild fold-changes in gene expression (92). Interestingly, these mild changes are growth-phase dependent and, although often small at the transcriptome level, result in significant changes in overall cell morphology. These results were interpreted as indicating a transcription factor-like function of histone proteins in *Halobacterium*, with global architecture imparted by histone-proteins as largely unimportant to regulating transcriptome-wide expression but select loci with critical histone-binding positions displaying differential expression due to loss of histone production in deletion strains.

B.2.3 Regulation of transcription initiation and elongation with archaeal histone-based chromatin.

Genome-wide impacts of archaeal histone-based chromatin on regulation of gene expression implies that histones are important, often essential, and that changes in histone expression, or histone-induced genomic architecture, impact cellular fitness (9). To determine how the histone-based landscape directly impacts gene expression, most studies have taken advantage of purified transcription systems and the capacity of archaeal histones to spontaneously bind DNAs *in vitro* at the same positions utilized *in vivo* and to form structures that match *in vivo* 3D chromatin architectures. Early *in vitro* transcription experiments using components from *Methanothermobacter thermautotrophicus* demonstrated a repressive effect of histone-addition on transcript production, with complete inhibition of transcription when histone proteins were provided at levels that would theoretically saturate DNA binding (~1 histone dimer per 30 bp of DNA) (12, 52, 72, 93). These *in vitro* results were later extended and confirmed using components from *Pyrococcus furiosus* (69).

Transcription regulation must normally occur within a chromatin landscape. Most archaeal transcription regulators mimic bacterial transcription regulators and bind within or immediately adjacent to core promoter elements to impact formation of initiation complexes.

DNA binding positions upstream of the *rb2* gene in *M. jannaschii* were shown to act as histone-nucleating sites, localizing histones whose binding reduces transcription by blocking formation of pre-initiation complexes (65). Histones are non-specific DNA binding proteins, and unsurprisingly, precision *in vitro* hydroxyl radical footprinting revealed that the site-specific DNA binding transcription factor Ptr2 effectively competes with localized histone binding – even at saturating histone levels – to activate transcription.

Transcription elongation is also affected by archaeal histone-based chromatin. *In vitro* transcription assays have been used to establish that the archaeal RNAP is unable to achieve elongation rates that are physiologically relevant through an archaeal chromatin barrier (12, 13). Using DNA templates capable of binding *M. thermautotrophicus* histone proteins, the *M. thermautotrophicus* RNA polymerase transcribed template DNA at a rate of ∼20 nts/sec in the absence of histone, but just ~2-5 nts/sec when archaeal histones were added to template. The initial collision between the TEC and the histone-barrier results in the greatest obstacle, causing RNA polymerase to pause and likely backtrack. The duration of the initial pause is much greater than subsequent pauses which occur every ~10-15 bp after the transcription elongation complex (TEC) escapes the initial collision. The rate limiting step of transcription through these archaeal histone-based barriers is translocation through the initial DNA-histone contacts.

The first data supporting factors that facilitate elongation through chromatin barriers is supportive of the congruent nature of the simplified archaeal transcription system and the more component complex Pol II apparatus (13). *In vitro* transcription experiments, using factors purified from *T. kodakarensis*, demonstrate that the activities of the conserved transcription factor TFS (TFIIS in Eukarya), and an Spt4/5 complex (also termed Spt4/5 in Eukarya) accelerate the archaeal transcription apparatus through histone-bound templates. The archaeal RNAP often backtracks due to downstream chromatin barriers, and archaeal TFS stimulated endonucleolytic cleavage of transcripts within backtracked complexes results in formation of a new RNA 3'-OH in the active center of RNAP (94–97). Reactivation of backtracked TECs

permits elongation restart and another opportunity for the TEC to transcribe up to and through a downstream chromatin barrier. The Spt4/5 complex, but neither factor individually, also aided *in vitro* transcription through archaeal histone-based chromatin, presumably due to their stabilizing effects of a closed-clamp configuration of the TEC in aiding proper alignment and retention of the 3'-OH in the RNAP active center (29, 96, 98–100).

Given the observations of archaeal histone-based chromatin controlling the initiation and elongation aspects of transcription, it is likely that the local chromatin environment also plays a role in termination and proper 3' end formation of transcripts.

B.3 Nucleoid-Associated Proteins (NAPs) in Archaea

The regulation imposed by genomic architecture in archaeal species that do not encode histone proteins has also been investigated in diverse clades. Perhaps the best studied protein is the well-conserved Alba (Sac10b-homologues), but abundant small basic proteins are encoded in both histone- and non-histone encoding archaea that likely impact genomic architectures. We focus first on Alba, then on more recently identified and emerging NAPs in diverse species.

B.3.1 Alba, a conserved chromatin protein, with controversial roles in genomic architecture

Substantial and contentious debate surrounds the Sac10b family of proteins, commonly termed Alba for 'acetylation lowers binding affinity', which dominates studies of the non-histone-based organization and regulation of archaeal genomes (101). Sac10b is a general nucleic-acid binding protein, with affinity for both single-stranded and double-stranded RNA and DNA. Evidence for Sac10b-mediated roles in DNA compaction and organization are recognized, although near equal evidence supports a role for Sac10b in RNA metabolism and binding. A contentious debate surrounds Sac10b, its role in DNA versus RNA binding, and whether acetylation or methylation is the post-translational modification that may impact function of Sac10b proteins *in vivo*. The focus of many studies was the modification of lysine 16, a well-

conserved residue in Sac10b homologues, and identification of proteins that could add or remove a reported acetyl group to impact Sac10b activity. Post-translational modification of K16 within Sac10b proteins was initially described as an acetylation event, hence the common Alba acronym (acetylation lowers binding affinity), but this modification has more recently been identified as a trimethylation (102). Due to the limited research regarding other nucleoid associated proteins, examination of this paradox is presented here from a historical perspective in the context of newer findings and argues for the further examination of other potential chromatin protein targets.

The Sac10b family of nucleic acid-binding proteins are highly conserved within Archaea, especially species that thrive in (hyper)thermophilic environments. Sac10b family members are encoded in both histone-encoding and non-histone-encoding archaea and are thought to play a major structural role in archaeal chromatin. Most research has focused within the *Crenarchaeota*, specifically the *Sulfolobales*. Much of the initial biochemical analyses focused on Alba-DNA interactions. The Sac10b homologue from *Sulfolobus shibatae* (Ssh10b) is a highly abundant protein (~4% of total protein), was shown to bind dsDNA and influence DNA topology at physiological temperatures (103). Both electron microscopy (EM) and atomic force microscopy (AFM) experiments revealed an Alba concentration-dependent compaction of archaeal DNA (104–106).

Sac10b proteins are typically encoded in archaeal genomes in the form of Alba1 but some species encode an additional paralog (Alba2) that is typically expressed at lower steady-state protein levels (104). More detailed investigations detailed that Sac10b bound DNA as a homodimer, and when Alba2 isoforms were present, that Alba heterodimers could also bind and compact DNA (104–106); Alba2 forms obligate heterodimers with Alba1 and is found exclusively associated with Alba1 *in vivo*. At lower Alba:DNA ratios, Alba1 homodimers bridge DNA duplexes, slightly compacting DNA by promoting the formation of loop structures (105, 106). At higher concentrations Alba1 homodimers form rigid protein-bound DNA structures (106). Much

like Alba1 homodimers at low concentration, Alba1/Alba2 heterodimers form looped, slightly contracted DNA structures (104). However, at higher Alba:DNA ratios, the Alba1/Alba2 heterodimers induced highly compacted DNA structures that differed significantly from the rigidified linear chromatin structure of Alba1 homodimers(106). Crystal structures of Sac10b protein homologues from *Aeropyrum pernix K1*, *Sulfolobus solfataricus,* and *Pyrococcus horikoshii OT3* all confirm a dimeric mode of nucleic acid interaction (107–112).

In addition to forming distinct protein:DNA complexes that impact DNA topology based on concentration and dimeric partnerships, Sac10b proteins were shown to have high affinity for RNA (113, 114). In Eukarya, Alba-like proteins have diverse RNA metabolism roles (113), suggesting Sac10b proteins may be involved in RNA stability or degradation pathways. Localization of Sac10b to the cytoplasm with no observable association with the nucleoid suggested interaction with RNA rather than DNA *in vivo* (115). This suggestion was corroborated by *in vivo* cross-linking studies with Ssh10b that resulted in the co-purification of primarily ribosomal RNA and mRNA over DNA (114). Finally, addition of Ssh10b was demonstrated to directly destabilize RNA secondary structure *in vitro* (116). The *in vitro* binding affinity of Sac10b is comparable between RNA, ssDNA, and dsDNA and Sac10b can protect both RNA and DNA from RNase and DNase digestion.

Phyla specific modes of action have also been observed for Sac10b homologues, and particular notice should be taken to studies in mesophilic species versus (hyper)thermophilic archaea. Current evidence suggests that the biological role of Sac10b proteins may have diverged between mesophilic and thermophilic archaea. In contrast to the abundance of Sac10b in (hyper)thermophiles, studies of the Sac10b protein homolog Mmo10b in the mesophilic species *Methanococcus maripaludis* revealed that Mmo10b is present only in low abundance and bound specific DNA sequences rather than displaying general DNA affinity (117, 118). Deletion of a Sac10b homolog from *Methanococcus voltae* resulted in changes to protein expression patterns that overlapped with a histone B deletion in the same species (117) and in

T. kodakarensis deletion of histone B resulted in altered Sac10b homolog expression (91). Taken together, these results suggest Sac10b homologues may share an overlapping regulatory role with histones in archaea, and that the presence of histones may reduce the impact of Sac10b regulation of genomic architecture.

B.3.2 Post-translational modification of Alba may impact genomic architecture and gene expression in vivo.

The post-translational modification (PTM) of Sac10b was shown to impact DNA binding affinity and was extrapolated to suggest that PTM of Alba provided regulation akin to PTMs of histone residues common in eukaryotes (101, 108). Recombinant preparations of Alba lacking PTMs displayed greater affinity for DNA than natively purified, PTM-Alba populations. The increased affinity of unmodified Alba also impeded transcription elongation to a greater extent than native, PTM-Alba preparations, consistent with Alba-mediated regulation of genomic structure based on PTM of Alba.

Initial MALDI-TOF mass spectrometry analysis identified lysine 16 (K16) in the Sac10b protein from *Sulfolobus solfataricus* P2 as the primary site of acetylation. In vitro acetylation by protein acetyltransferase 1 (Pat1) and *in vitro* deacetylation by the silent information regulator (Sir2) were shown to modify Sac10b imparting a mechanism of Sac10b binding control (101, 109, 119). However, K16 is not well-conserved in Sac10b homologues (118), and the initial identification of K16 as the site of modification, and even the PTM itself are now in question. More recent studies have identified Sac10b as a target for both methylation and N-terminal acetylation, but not K16 acetylation (120). Post-translational modification of the N-terminus of Sac10b by N-acetyl transferase (NAT) has been demonstrated *in vitro* and is proposed to be the primary site of Sac10b acetylation *in vivo*. Recent mass-spectrometry (NanoLC-MS-MS) data of a Sac10b homologue from *S. islandicus* has revealed methylation, acetylation, and deamination of this protein (102). Strikingly, K16 was trimethylated, not acetylated. The improvements in

mass spectrometry and identification of K16 trimethylation challenges the core assertion of Sac10b:DNA interactions being controlled by acetylation at K16. Taken together, the conflicting information on the PTM status of K16, the likely role of Sac10b homologues in binding DNA and RNA, and the differential abundance and importance of Sac10b homologues in diverse species argues that PTM(s) of Sac10b members may also be diverse and likely impact aspects of both RNA and DNA binding.

B.3.3 Variety in archaeal NAPs may shape genomes in diverse environments.

While the biological importance of mechanisms governing Sac10b nucleic acid interactions are heavily debated, it is important to consider the roles of the many other NAPs encoded in archaeal genomes. In addition to Sac10b, most crenarchaea encode small ~7 kDa proteins, with Sul7 and Cren7 dominating the literature. Cren7 is a 7kDa, basic protein that has been found associated with DNA *in vivo*. The abundant and basic Cren7 protein has high affinity for double-stranded DNA, suggesting a primary role in genomic organization (121). Although no obvious relationship is present at the primary amino acid level, Sul7 is structurally homologous to Cren7, and both are known to induce DNA compaction *in vitro* (122). Crenarchaeal species such as *Pyrobaculum aerophilum* and *Thermoproteus tenax* lack obvious Cren7 or Sul7 homologues and instead encode the chromatin protein CC1. Like Sac10b proteins, CC1 is able to bind double stranded and single stranded DNA, suggesting a role in chromatin organization (123).

In the euryarchaeal *Thermococcales* the TrmBL2 family is an abundant DNA-associated protein (124). At likely physiological salt concentrations (~300mM KCl) TrmBL2 binds DNA in a site-specific manner, while displaying non-specific DNA binding at lower salt concentrations. Non-specific DNA binding results in a filamentous structure that can compete with histone-binding (125). In *T. kodakarensis* the abundance of TrmBL2 changes with the growth phase, and the interplay/competition between histones and TrmBL2 may offer an additional path to

regulate genomic architecture and thus gene expression in response to environmental conditions. TrmBL2 occupancy of promoter regions can impact transcription, whereas TrmBL2 minimally impacts transcription elongation (126). Deletion of TrmBL2 is possible and results in reduced condensation of chromatin and altered expression of approximately the same percentage of genes as deletion of a histone isoform (126).

B.4 Conclusions and future perspectives

Archaea are ecologically and metabolically diverse and thus it is perhaps not surprising that substantial differences in genomic architecture and regulation are imposed in different clades. Most species encode proteins with the core histone-fold, and archaeal chromatin thus dominates the landscape of regulation in archaeal species. The overall structural similarities between archaeal histone-based chromatin and eukaryotic chromatin are obvious, but the regulatory potential of the latter far exceeds the potential of the former. Archaeal chromatin is often formed with only one histone isoform, and given the absence of identifiable PTMs, it is likely that archaeal histones are not subject to repositioning or changes in DNA affinity that could increase or decrease transcription levels at specific loci. Significant questions remain for species that encode multiple histone isoforms and whether regulated assembly or binding of unique heterodimers impacts genomic architecture and thus gene regulation. The identification of PTMs or factors that could impact the normally tight association of adjacent gyres of archaeal histone-based chromatin may provide a route to regulate chromatin structure and transcriptional output. Identification of any such factors may help reveal the evolutionary origin of remodeling and modification machineries found ubiquitously in eukaryotes.

The identification of archaeal species that encode extensions on the core histone fold is an exciting new revelation in the context of histone-based regulation of gene expression. The expansion beyond the core histone-fold, and the retention of discrete histone isoforms in many archaeal species provides tantalizing evidence in support of the expansion that must have

occurred to provide all extant eukaryotes with the canonical four histones. The length and stability of extended nucleosome-like structures formed with archaeal histones is likely impacted by histone isoforms and the presence of extensions beyond the core histone-fold. Given that disrupting the tight-association of archaeal histone-based chromatin results in massive fitness defects, it is plausible to predict that more fine-tuned and regulated mechanisms may exist to control and adjust chromatin formation or limit the length of the extended histone-based polymers to control gene expression *in vivo*. The timing of and expansion to defined heterodimeric histone partnerships that lead to the transition from an extended histone-based polymer structure to the discrete particles that define the eukaryotic nucleosome is a major outstanding question.

In addition to chemical modification machinery, eukaryotes encode a wealth of complexes to reposition nucleosomes. Repositioning nucleosomes or altering histone-DNA affinity may help or hinder transcription in eukaryotic cells. It remains possible that archaeal encoded modification or repositioning complexes exist, but current evidence suggests instead that archaeal TECs are reliant on conserved transcription factors to aid in overcoming histone-induced barriers to transcription elongation. To fully illustrate the evolution of the transcription apparatus, the roles of other conserved and potentially novel transcription factors and effectors will need to be characterized. The noted effects of Spt4/5 and TFS suggest that direct modification of the transcription apparatus may suffice for unmodified and relatively uniform histone-based chromatin structures, but that more powerful chromatin remodeling complexes and modification machinery are required for the diverse landscape of extant eukaryotic chromatin landscapes.

Despite many archaea encoding both NAPs and histone proteins, only limited information is available regarding the combinatorial regulation provided by the interplay of architectures produced by binding of both classes of proteins (126). It is logical to predict that the length and

stability of extended histone-based structures may be regulated by NAP binding or NAP-mediated formation of DNA loops that impact overall topology and DNA flexibility (Figure B.3). While minimally conserved at an amino acid sequence level, the structural conservation and functionality of archaeal NAPs suggests a conserved strategy for organizing DNA structure (127). Clarity surrounding the role of the nearly ubiquitous Sac10b family of proteins with respect to RNA versus DNA binding – and clarification of the locations, identity and impacts of potential PTMs – should illuminate the role of this often-abundant protein in organizing and providing dynamic regulation of archaeal genomes.

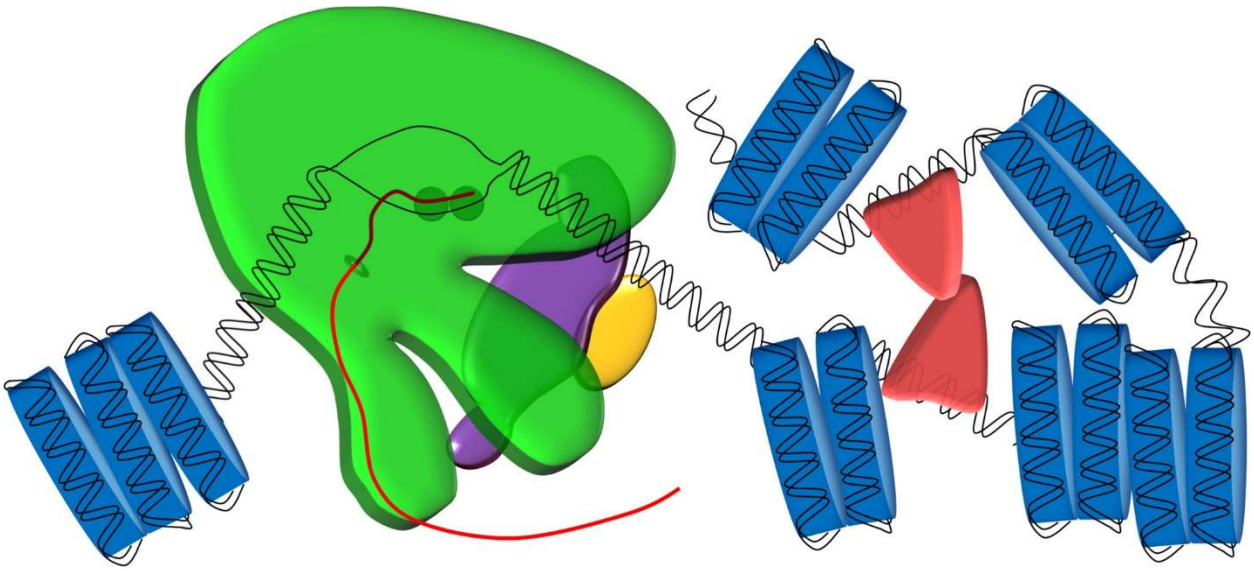


Figure B.3: The archaeal chromatin landscape is dynamic.
a) Wrapping of DNA by archaeal histones forms various sizes of extended histone-based chromatin structures. The regulation and depositions of these structures is unknown, but nucleoid associated proteins (NAPs) may play a role in both looping of DNA and size restriction of extended histone polymers. b) Transcription initiation factors TFB and TBP compete with histone proteins for the promoter element in archaea allowing transcription initiation upstream of a chromatinized gene body. c) RNAP must traverse a chromatinized gene body. Spt4-Spt5 permit the transition from initiation to early elongation by displacing TFE and facilitating processive elongation through a chromatin landscape.

REFERENCES

1. K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, T. J. Richmond, Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–60 (1997).

2. R. T. Dame, The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Molecular microbiology* **56**, 858–70 (2005).

3. L. Farnung, S. M. Vos, P. Cramer, Structure of transcribing RNA polymerase II-nucleosome complex. *Nature communications* **9**, 5432 (2018).

4. T. J. Krogh, J. Møller-Jensen, C. Kaleta, Impact of Chromosomal Architecture on the Function and Evolution of Bacterial Genomes. *Frontiers in microbiology* **9**, 2019 (2018).

5. S. C. Dillon, C. J. Dorman, Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nature reviews. Microbiology* **8**, 185–95 (2010).

6. R. Spurio, M. Falconi, A. Brandi, C. L. Pon, C. O. Gualerzi, The oligomeric structure of nucleoid protein H-NS is necessary for recognition of intrinsically curved DNA and for DNA bending. *The EMBO journal* **16**, 1795–805 (1997).

7. S. T. Arold, P. G. Leonard, G. N. Parkinson, J. E. Ladbury, H-NS forms a superhelical protein scaffold for DNA condensation. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 15728–32 (2010).

8. P. B. Becker, W. Hörz, ATP-dependent nucleosome remodeling. *Annual review of biochemistry* **71**, 247–73 (2002).

9. M. J. E. Koster, B. Snel, H. T. M. Timmers, Genesis of chromatin and transcription dynamics in the origin of species. *Cell* **161**, 724–36 (2015).

10. E. Segal, *et al.*, A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).

11. C. H. Chang, D. S. Luse, The H3/H4 tetramer blocks transcript elongation by RNA polymerasem II *in vitro*. *Journal of Biological Chemistry* **272**, 23427–23434 (1997).

12. Y. Xie, J. N. Reeve, Transcription by an archaeal RNA polymerase is slowed but not blocked by an archaeal nucleosome. *Journal of Bacteriology* **186**, 3492–3498 (2004).

13. T. J. Sanders, *et al.*, TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Molecular Microbiology* (2018) https:/doi.org/10.1111/mmi.14191.

14. M. V Kotlajich, *et al.*, Bridged filaments of histone-like nucleoid structuring protein pause RNA polymerase and aid termination in bacteria. *eLife* **4** (2015).

15. R. A. van der Valk, J. Vreede, F. Crémazy, R. T. Dame, Genomic looping: a key principle of chromatin organization. *Journal of molecular microbiology and biotechnology* **24**, 344–59 (2014).

16. W. A. Bickmore, B. van Steensel, Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270–84 (2013).

17. I. Sela, Y. I. Wolf, E. V Koonin, Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 11399–11407 (2016).

18. C. R. Woese, O. Kandler, M. L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 4576–9 (1990).

19. P. P. Chan, A. D. Holmes, A. M. Smith, D. Tran, T. M. Lowe, The UCSC Archaeal Genome Browser: 2012 update. *Nucleic Acids Research* **40** (2012).

20. K. L. Schneider, The UCSC Archaeal Genome Browser. *Nucleic Acids Research* **34**, D407–D410 (2006).

21. C. Brochier-Armanet, P. Forterre, S. Gribaldo, Phylogeny and evolution of the Archaea: one hundred genomes later. *Current opinion in microbiology* **14**, 274–81 (2011).

22. M. N. Wojtas, N. G. A. Abrescia, Archaeal transcription: making up for lost time. *Biochemical Society Transactions* **41**, 356–361 (2013).

23. M. N. Wojtas, M. Mogni, O. Millet, S. D. Bell, N. G. A. Abrescia, Structural and functional analyses of the interaction of archaeal RNA polymerase with DNA. *Nucleic Acids Research* **40**, 9941–9952 (2012).

24. S.-H. Jun, *et al.*, The X-ray crystal structure of the euryarchaeal RNA polymerase in an open-clamp configuration. *Nature Communications* **5**, 5132 (2014).

25. A. Hirata, B. J. Klein, K. S. Murakami, The X-ray crystal structure of RNA polymerase from Archaea. *Nature* **451**, 851–854 (2008).

26. A. Hirata, *et al.*, Archaeal RNA polymerase subunits E and F are not required for transcription *in vitro*, but a Thermococcus kodakarensis mutant lacking subunit F is temperature-sensitive. *Molecular microbiology* **70**, 623–33 (2008).

27. T. J. Santangelo, L. Cubonová, C. L. James, J. N. Reeve, TFB1 or TFB2 is sufficient for Thermococcus kodakaraensis viability and for basal transcription *in vitro*. *Journal of molecular biology* **367**, 344–57 (2007).

28. F. Blombach, K. L. Smollett, D. Grohmann, F. Werner, Molecular Mechanisms of Transcription Initiation-Structure, Function, and Evolution of TFE/TFIIE-Like Factors and Open Complex Formation. *Journal of molecular biology* **428**, 2592–2606 (2016).

29. F. Blombach, *et al.*, Archaeology of RNA polymerase: factor swapping during the transcription cycle. *Biochemical Society transactions* **41**, 362–7 (2013).

30. T. Fouqueau, F. Blombach, F. Werner, Evolutionary Origins of Two-Barrel RNA Polymerases and Site-Specific Transcription Initiation. *Annual review of microbiology* **71**, 331–348 (2017).

31. J. N. Reeve, Archaeal chromatin and transcription. *Molecular microbiology* **48**, 587–98 (2003).

32. F. Werner, Structural evolution of multisubunit RNA polymerases. *Trends in microbiology* **16**, 247–50 (2008).

33. F. Werner, D. Grohmann, Evolution of multisubunit RNA polymerases in the three domains of life. *Nature Reviews Microbiology* **9**, 85–98 (2011).

34. A. M. Gehring, J. E. Walker, T. J. Santangelo, Transcription Regulation in Archaea. *Journal of bacteriology* **198**, 1906–1917 (2016).

35. A. M. Gehring, T. J. Santangelo, "Manipulating Archaeal Systems to Permit Analyses of Transcription Elongation-Termination Decisions In Vitro" in *Methods in Molecular Biology (Clifton, N.J.)*, (2015), pp. 263–279.

36. D. Grohmann, F. Werner, Recent advances in the understanding of archaeal transcription. *Current opinion in microbiology* **14**, 328–34 (2011).

37. A. Gietl, *et al.*, Eukaryotic and archaeal TBP and TFB/TF(II)B follow different promoter DNA bending pathways. *Nucleic acids research* **42**, 6219–31 (2014).

38. L. Aravind, E. V Koonin, DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic acids research* **27**, 4658–70 (1999).

39. E. Peeters, D. Charlier, The Lrp Family of Transcription Regulators in Archaea. *Archaea* **2010**, 1–10 (2010).

40. E. Peeters, R. P. C. Driessen, F. Werner, R. T. Dame, The interplay between nucleoid organization and transcription in archaeal genomes. *Nature reviews. Microbiology* **13**, 333–41 (2015).

41. M. A. Pritchett, S. P. Wilkinson, E. P. Geiduschek, M. Ouhammouch, Hybrid Ptr2-like activators of archaeal transcription. *Molecular microbiology* **74**, 582–93 (2009).

42. S. P. Wilkinson, M. Ouhammouch, E. P. Geiduschek, Transcriptional activation in the context of repression mediated by archaeal histones. *Proceedings of the National Academy of Sciences* **107**, 6777–6781 (2010).

43. M. Krug, S.-J. Lee, W. Boos, K. Diederichs, W. Welte, The three-dimensional structure of TrmB, a transcriptional regulator of dual function in the hyperthermophilic archaeon *Pyrococcus furiosus* in complex with sucrose. *Protein Science* **22**, 800–808 (2013).

44. T. J. Lie, *et al.*, Overlapping repressor binding sites regulate expression of the Methanococcus maripaludis glnK(1) operon. *Molecular microbiology* **75**, 755–62 (2010).

45. G. L. Lipscomb, *et al.*, SurR: a transcriptional activator and repressor controlling hydrogen and elemental sulphur metabolism in Pyrococcus furiosus. *Molecular microbiology* **71**, 332–49 (2009).

46. E. A. Karr, “Transcription Regulation in the Third Domain” in *Advances in Applied Microbiology*, (2014), pp. 101–133.

47. S.-J. Lee, M. Surma, W. Hausner, M. Thomm, W. Boos, The role of TrmB and TrmB-like transcriptional regulators for sugar transport and metabolism in the hyperthermophilic archaeon Pyrococcus furiosus. *Archives of microbiology* **190**, 247–56 (2008).

48. T. Kanai, *et al.*, A global transcriptional regulator in Thermococcus kodakaraensis controls the expression levels of both glycolytic and gluconeogenic enzyme-encoding genes. *The Journal of biological chemistry* **282**, 33659–70 (2007).

49. S. D. Bell, S. S. Cairns, R. L. Robson, S. P. Jackson, Transcriptional regulation of an archaeal operon in vivo and *in vitro*. *Molecular cell* **4**, 971–82 (1999).

50. H. Maruyama, *et al.*, An alternative beads-on-a-string chromatin architecture in Thermococcus kodakarensis. *EMBO reports* **14**, 711–717 (2013).

51. N. Nalabothula, *et al.*, Archaeal nucleosome positioning in vivo and *in vitro* is directed by primary sequence motifs. *BMC genomics* **14**, 391 (2013).

52. J. N. Reeve, K. Sandman, C. J. Daniels, Archaeal histones, nucleosomes, and transcription initiation. *Cell* **89**, 999–1002 (1997).

53. R. N. Fish, C. M. Kane, Promoting elongation with transcript cleavage stimulatory factors. *Biochimica et biophysica acta* **1577**, 287–307 (2002).

54. O. Laptenko, J. Lee, I. Lomakin, S. Borukhov, Transcript cleavage factors GreA and GreB act as transient catalytic components of RNA polymerase. *The EMBO journal* **22**, 6322–34 (2003).

55. M. Orlova, J. Newlands, A. Das, A. Goldfarb, S. Borukhov, Intrinsic transcript cleavage activity of RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 4596–600 (1995).

56. B. Henneman, C. van Emmerik, H. van Ingen, R. T. Dame, Structure and function of archaeal histones. *PLOS Genetics* **14**, e1007582 (2018).

57. K. Zaremba-Niedzwiedzka, *et al.*, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

58. L. Eme, A. Spang, J. Lombard, C. W. Stairs, T. J. G. Ettema, Archaea and the origin of eukaryotes. *Nature Reviews Microbiology* **15**, 711–723 (2017).

59. J. N. Reeve, *et al.*, Archaeal histones: structures, stability and DNA binding. *Biochemical Society transactions* **32**, 227–30 (2004).

60. K. Sandman, J. N. Reeve, Chromosome packaging by archaeal histones. *Advances in applied microbiology* **50**, 75–99 (2001).

61. L. Mariño-Ramírez, *et al.*, The Histone Database: an integrated resource for histones and histone fold-containing proteins. *Database : the journal of biological databases and curation* **2011**, bar048 (2011).

62. V. Visone, *et al.*, Chromatin structure and dynamics in hot environments: architectural proteins and DNA topoisomerases of thermophilic archaea. *International journal of molecular sciences* **15**, 17162–87 (2014).

63. R. Ammar, *et al.*, Chromatin is an ancient innovation conserved between Archaea and Eukarya. *eLife* **1**, e00078 (2012).

64. L. Cubonova, K. Sandman, S. J. Hallam, E. F. DeLong, J. N. Reeve, Histones in Crenarchaea. *Journal of Bacteriology* **187**, 5482–5485 (2005).

65. S. P. Wilkinson, M. Ouhammouch, E. P. Geiduschek, Transcriptional activation in the context of repression mediated by archaeal histones. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 6777–81 (2010).

66. S. Bhattacharyya, F. Mattiroli, K. Luger, Archaeal DNA on the histone merry-go-round. *The FEBS Journal* **285**, 3168–3174 (2018).

67. F. Mattiroli, *et al.*, Structure of histone-based chromatin in Archaea. *Science* **357**, 609–612 (2017).

68. M. Ouhammouch, R. E. Dewhurst, W. Hausner, M. Thomm, E. P. Geiduschek, Activation of archaeal transcription by recruitment of the TATA-binding protein. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 5097–102 (2003).

69. D. Soares, *et al.*, Archaeal histone stability, DNA binding, and transcription inhibition above 90 degrees C. *Extremophiles : life under extreme conditions* **2**, 75–81 (1998).

70. K. A. Bailey, F. Marc, K. Sandman, J. N. Reeve, Both DNA and histone fold sequences contribute to archaeal nucleosome stability. *The Journal of biological chemistry* **277**, 9293–301 (2002).

71. L. Bintu, *et al.*, Nucleosomal elements that control the topography of the barrier to transcription. *Cell* **151**, 738–49 (2012).

72. R. A. Grayling, K. Sandman, J. N. Reeve, Histones and chromatin structure in hyperthermophilic Archaea. *FEMS microbiology reviews* **18**, 203–13 (1996).

73. M. R. Starich, K. Sandman, J. N. Reeve, M. F. Summers, NMR structure of HMfB from the hyperthermophile, Methanothermus fervidus, confirms that this archaeal protein is a histone. *Journal of molecular biology* **255**, 187–203 (1996).

74. K. Decanniere, K. Sandman, J. N. Reeve, U. Heinemann, Crystallization and preliminary X-ray characterization of the Methanothermus fervidus histones HMfA and HMfB. *Proteins* **24**, 269–71 (1996).

75. K. Decanniere, A. M. Babu, K. Sandman, J. N. Reeve, U. Heinemann, Crystal structures of recombinant histones HMfA and HMfB from the hyperthermophilic archaeon Methanothermus fervidus. *Journal of Molecular Biology* **303**, 35–47 (2000).

76. K. Sandman, D. Soares, J. N. Reeve, Molecular components of the archaeal nucleosome. *Biochimie* **83**, 277–81 (2001).

77. K. Sandman, J. N. Reeve, Archaeal histones and the origin of the histone fold. *Current opinion in microbiology* **9**, 520–5 (2006).

78. D. J. Soares, K. Sandman, J. N. Reeve, Mutational analysis of archaeal histone-DNA interactions. *Journal of Molecular Biology* **297**, 39–47 (2000).

79. K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, T. J. Richmond, Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–60 (1997).

80. K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, T. J. Richmond, Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–60 (1997).

81. M. Ouhammouch, F. Werner, R. O. J. Weinzierl, E. P. Geiduschek, A fully recombinant system for activator-dependent archaeal transcription. *The Journal of biological chemistry* **279**, 51719–21 (2004).

82. F. Marc, K. Sandman, R. Lurz, J. N. Reeve, Archaeal histone tetramerization determines DNA affinity and the direction of DNA supercoiling. *The Journal of biological chemistry* **277**, 30879–86 (2002).

83. T. Kujirai, *et al.*, Structural basis of the nucleosome transition during RNA polymerase II passage. *Science*, eaau9904 (2018).

84. H. Ehara, *et al.*, Structure of the complete elongation complex of RNA polymerase II with basal factors. *Science* **357**, 921–924 (2017).

85. S. S. Teves, C. M. Weber, S. Henikoff, Transcribing through the nucleosome. *Trends in Biochemical Sciences* **39**, 577–586 (2014).

86. M. L. Kireeva, *et al.*, Nature of the nucleosomal barrier to RNA polymerase II. *Molecular cell* **18**, 97–108 (2005).

87. K. Höfer, A. Jäschke, Epitranscriptomics: RNA Modifications in Bacteria and Archaea. *Microbiology Spectrum* **6** (2018).

88. M. Ouellette, J. Gogarten, J. Lajoie, A. Makkay, R. Papke, Characterizing the DNA Methyltransferases of Haloferax volcanii via Bioinformatics, Gene Deletion, and SMRT Sequencing. *Genes* **9**, 129 (2018).

89. M. Couturier, A.-C. Lindås, The DNA Methylome of the Hyperthermoacidophilic Crenarchaeon Sulfolobus acidocaldarius. *Frontiers in microbiology* **9**, 137 (2018).

90. W. T. Li, K. Sandman, S. L. Pereira, J. N. Reeve, MJ1647, an open reading frame in the genome of the hyperthermophile Methanococcus jannaschii, encodes a very thermostable archaeal histone with a C-terminal extension. *Extremophiles : life under extreme conditions* **4**, 43–51 (2000).

91. L. Čuboňováa, *et al.*, An archaeal histone is required for transformation of Thermococcus kodakarensis. *Journal of bacteriology* **194**, 6864–74 (2012).

92. K. Weidenbach, *et al.*, Deletion of the archaeal histone in *Methanosarcina mazei* Gö1 results in reduced growth and genomic transcription. *Molecular Microbiology* **67**, 662–671 (2008).

93. K. Sandman, J. N. Reeve, Archaeal chromatin proteins: different structures but common function? *Current opinion in microbiology* **8**, 656–61 (2005).

94. W. Hausner, U. Lange, M. Musfeldt, Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase. *Journal of Biological Chemistry* **275**, 12393–12399 (2000).

95. U. Lange, W. Hausner, Transcriptional fidelity and proofreading in Archaea and implications for the mechanism of TFS-induced RNA cleavage. *Molecular Microbiology* **52**, 1133–1143 (2004).

96. T. J. Sanders, *et al.*, TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Molecular Microbiology* (2019) https:/doi.org/10.1111/mmi.14191.

97. T. Fouqueau, *et al.*, The transcript cleavage factor paralogue TFS4 is a potent RNA polymerase inhibitor. *Nature communications* **8**, 1914 (2017).

98. A. Hirtreiter, *et al.*, Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. *Nucleic Acids Research* **38**, 4040–4051 (2010).

99. F. W. Martinez-Rucobo, S. Sainsbury, A. C. Cheung, P. Cramer, Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. *The EMBO Journal* **30**, 1302–1310 (2011).

100. G. Guo, *et al.*, Structural and biochemical insights into the DNA-binding mode of MjSpt4p:Spt5 complex at the exit tunnel of RNAPII. *Journal of structural biology* **192**, 418–425 (2015).

101. S. D. Bell, C. H. Botting, B. N. Wardleworth, S. P. Jackson, M. F. White, The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation. *Science (New York, N.Y.)* **296**, 148–51 (2002).

102. J. Cao, Q. Wang, T. Liu, N. Peng, L. Huang, Insights into the post-translational modifications of archaeal Sis10b (Alba): lysine-16 is methylated, not acetylated, and this does not regulate transcription or growth. *Molecular microbiology* **109**, 192–208 (2018).

103. H. Xue, R. Guo, Y. Wen, D. Liu, L. Huang, An abundant DNA binding protein from the hyperthermophilic archaeon Sulfolobus shibatae affects DNA supercoiling in a temperature-dependent fashion. *Journal of bacteriology* **182**, 3929–33 (2000).

104. C. Jelinska, *et al.*, Obligate heterodimerization of the archaeal Alba2 protein with Alba1 provides a mechanism for control of DNA packaging. *Structure (London, England : 1993)* **13**, 963–71 (2005).

105. R. Lurz, M. Grote, J. Dijk, R. Reinhardt, B. Dobrinski, Electron microscopic study of DNA complexes with proteins from the Archaebacterium Sulfolobus acidocaldarius. *The EMBO journal* **5**, 3715–21 (1986).

106. N. Laurens, *et al.*, Alba shapes the archaeal genome using a delicate balance of bridging and stiffening the DNA. *Nature communications* **3**, 1328 (2012).

107. T. Tanaka, S. Padavattan, T. Kumarevel, Crystal structure of archaeal chromatin protein Alba2-double-stranded DNA complex from Aeropyrum pernix K1. *The Journal of biological chemistry* **287**, 10394–402 (2012).

108. B. N. Wardleworth, R. J. M. Russell, S. D. Bell, G. L. Taylor, M. F. White, Structure of Alba: an archaeal chromatin protein modulated by acetylation. *The EMBO journal* **21**, 4654–62 (2002).

109. K. Zhao, X. Chai, R. Marmorstein, Structure of a Sir2 substrate, Alba, reveals a mechanism for deacetylation-induced enhancement of DNA binding. *The Journal of biological chemistry* **278**, 26071–7 (2003).

110. T. Kumarevel, *et al.*, Crystal structure of an archaeal specific DNA-binding protein (Ape10b2) from Aeropyrum pernix K1. *Proteins* **71**, 1156–62 (2008).

111. Q. Cui, *et al.*, Two conformations of archaeal Ssh10b. The origin of its temperature-dependent interaction with DNA. *The Journal of biological chemistry* **278**, 51015–22 (2003).

112. K. Hada, *et al.*, Crystal structure and functional analysis of an archaeal chromatin protein Alba from the hyperthermophilic archaeon Pyrococcus horikoshii OT3. *Bioscience, biotechnology, and biochemistry* **72**, 749–58 (2008).

113. L. Aravind, L. M. Iyer, V. Anantharaman, The two faces of Alba: the evolutionary connection between proteins participating in chromatin structure and RNA metabolism. *Genome biology* **4**, R64 (2003).

114. R. Guo, H. Xue, L. Huang, Ssh10b, a conserved thermophilic archaeal protein, binds RNA in vivo. *Molecular microbiology* **50**, 1605–15 (2003).

115. B. Bohrmann, … E. K.-J. of S., undefined 1994, Localization of histone-like proteins in thermophilic Archaea by immunogold electron microscopy. *Elsevier*.

116. L. Guo, *et al.*, Biochemical and structural insights into RNA binding by Ssh10b, a member of the highly conserved Sac10b protein family in Archaea. *The Journal of biological chemistry* **289**, 1478–90 (2014).

117. I. Heinicke, J. Müller, M. Pittelkow, A. Klein, Mutational analysis of genes encoding chromatin proteins in the archaeon Methanococcus voltae indicates their involvement in the regulation of gene expression. *Molecular genetics and genomics : MGG* **272**, 76–87 (2004).

118. Y. Liu, *et al.*, The Sac10b homolog in Methanococcus maripaludis binds DNA at specific sites. *Journal of bacteriology* **191**, 2315–29 (2009).

119. V. L. Marsh, S. Y. Peak-Chew, S. D. Bell, Sir2 and the acetyltransferase, Pat, regulate the archaeal chromatin protein, Alba. *The Journal of biological chemistry* **280**, 21122–8 (2005).

120. E. A. Vorontsov, E. Rensen, D. Prangishvili, M. Krupovic, J. Chamot-Rooke, Abundant Lysine Methylation and N-Terminal Acetylation in Sulfolobus islandicus Revealed by Bottom-Up and Top-Down Proteomics. *Molecular & cellular proteomics : MCP* **15**, 3388–3404 (2016).

121. L. Guo, *et al.*, Biochemical and structural characterization of Cren7, a novel chromatin protein conserved among Crenarchaea. *Nucleic acids research* **36**, 1129–37 (2008).

122. R. P. C. Driessen, *et al.*, Crenarchaeal chromatin proteins Cren7 and Sul7 compact DNA by inducing rigid bends. *Nucleic acids research* **41**, 196–205 (2013).

123. X. Luo, *et al.*, CC1, a novel crenarchaeal DNA binding protein. *Journal of bacteriology* **189**, 403–9 (2007).

124. S. Wierer, *et al.*, TrmBL2 from Pyrococcus furiosus Interacts Both with Double-Stranded and Single-Stranded DNA. *PloS one* **11**, e0156098 (2016).

125. A. K. Efremov, *et al.*, Transcriptional Repressor TrmBL2 from Thermococcus kodakarensis Forms Filamentous Nucleoprotein Structures and Competes with Histones for DNA Binding in a Salt- and DNA Supercoiling-dependent Manner. *The Journal of biological chemistry* **290**, 15770–84 (2015).

126. H. Maruyama, *et al.*, Histone and TK0471/TrmBL2 form a novel heterogeneous genome architecture in the hyperthermophilic archaeon Thermococcus kodakarensis. *Molecular biology of the cell* **22**, 386–98 (2011).

127. R. P. C. Driessen, R. T. Dame, Nucleoid-associated proteins in Crenarchaea. *Biochemical Society transactions* **39**, 116–21 (2011).