

DISSERTATION

POST-INITIATION REGULATORY MECHANISMS OF  
TRANSCRIPTION IN THE ARCHAEA

Submitted by

Breanna Renée Wenck

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

Fall 2023

Doctoral Committee:

Advisor: Thomas Santangelo

Jeffrey C. Hansen  
Erin Osborne Nishimura  
Carol Wilusz

Copyright by Breanna Renée Wenck 2023

All Rights Reserved



## ABSTRACT

### POST-INITIATION REGULATORY MECHANISMS OF TRANSCRIPTION IN THE ARCHAEA

Increasingly sophisticated biochemical and genetic techniques are unraveling the regulatory factors and mechanisms that control gene expression in the Archaea. While some similarities in regulatory strategies are universal, archaeal-specific regulatory strategies are emerging to complement a complex patchwork of shared archaeal-bacterial and archaeal-eukaryotic regulatory mechanisms employed in the archaeal domain. Archaeal systems contain simplified, basal regulatory transcription components and mechanisms homologous to their eukaryotic counterparts, but also deploy tactics common to bacterial systems to regulate promoter usage and influence elongation-termination decisions. Many archaeal genomes are organized with histone proteins that resemble the core eukaryotic histone fold, which permits DNA wrapping through select histone-DNA contacts to generate chromatin-structures that impacts transcription regulation and gene expression. Despite such semblance between the eukaryotic and archaeal core histone folds, archaeal genomes lack the canonical N and C terminal extensions that are abundantly modified to regulate transcription in eukaryotic genomes. With combined and continued efforts across the field, tidbits of information regarding factor-mediated transcription regulation in the Archaea has become available over the last ~45 years but remains limited. This thesis aims to i) delineate how alterations to archaeal histone-based chromatin structures – via perturbations to key histone residues – affect the progression of the transcription apparatus, ii) characterize FttA-mediated transcription termination, and iii) explore chromatin- and TFS-mediated regulatory effects on transcription via global RNA polymerase (RNAP) positions.

## ACKNOWLEDGMENTS

How do you thank all the people in one small section of a dissertation that have played unique, but instrumental parts in your success? Because really, it's not just one person working on a doctorate, but a combination of interactions and support from many individuals that culminates in the realization of such a lengthy achievement. This particular adventure in my life has introduced me to some of the most wonderful people and reminded me of how lucky I am to have the familiar people that were a part of my life before I set out on this path. And so, with that, I would like to attempt to express my appreciation with words that will undoubtedly fail to reveal how grateful I truly am.

To my advisor, mentor, sometimes therapist, and friend, Tom Santangelo – Thank you for always seeing in me what I often find difficult to see in myself. When I first stepped into your lab, I was overwhelmed by the amount of knowledge surrounding me that I craved as my own. When I made the decision that I wanted to stay and continue learning in the Santangelo Lab, it wasn't necessarily because of the science, but because of the mentor. You will never truly know, I don't think, how wonderful you are and how much you have changed my life for the better. Thank you for always being pragmatic when it mattered and the teddy bear when I needed him the most.

To my committee members: Jeff Hansen, Erin Osborne Nishimura, and Carol Wilusz – I chose each of you as mentors because of what you all bring individually to science and to the world. I am so grateful to have the guidance and compassion that each of you exude in your daily interactions with your students. Thank you for challenging me and making my science better and in being not just my mentors, but also my friends.

To my bench mentor, Travis Sanders – Thank you for your patience, time, effort, and dedication in teaching me how to science. Your guidance taught me what it means to be a mentor and how to teach with care, detail, and compassion. I am grateful for you, my friend.

To my amazing lab mates/friends: Jocelyn, Craig, Aldy, Brett, Kristin, Seré, Alex, Alina, and the rest of the Santangelo Lab (past and present) – Thank you for always bringing out the best parts of me and being the most brilliant sounding boards I have ever had the privilege to work with. I will forever cherish each of you.

To the undergraduates I have had the fortune to mentor: Kat, Stavros, Amelia, Adrea, and Robert – Thank you for teaching me what it means to be a good mentor and bringing such delight to this experience. You are all phenomenal humans, and I am incredibly proud of each of my lab kids.

A special shoutout to the Bridges to Baccalaureate program: Paul Laybourn, Heather Matthews, Erin Pitts, and my REU cohort – Without each and every one of you, this wouldn't have been possible. From the bottom of my heart, thank you.

To my therapist, Tiah Terranova – You are the light when everything gets too dark; thank you for taking care of me.

And last, but certainly not least, to my beautiful family:

Mom – Thank you for being my practical fire starter and for teaching me how to be resilient and gentle, yet tough in a hard world.

Dad – Thank you for always keeping me humble and thoughtful.

Ariana, Keara, and Cody – Of all the things I've accomplished in my lifetime, raising my three bears, and witnessing the remarkable people you have become, is the most rewarding.

Josh – Thank you for introducing me to the stars and then encouraging me to reach for them.

To the rest of my family/friends – Thank you for sticking this out with me. It's been long, eventful, and stressful but you all continually encouraged me to be the nerdiest version of myself that I could be. I will be forever grateful.

## TABLE OF CONTENTS

ABSTRACT .....	ii
ACKNOWLEDGEMENTS .....	iii
CHAPTER 1: ARCHAEOAL TRANSCRIPTION .....	1
Introduction .....	1
Archaeal RNA polymerase (RNAP) .....	6
Archaeal transcription initiation .....	7
Archaeal-encoded regulatory transcription factors .....	9
Archaeal transcription elongation .....	10
Archaeal transcription termination .....	13
Concluding remarks .....	17
Thesis rationale .....	19
REFERENCES .....	21
CHAPTER 2: ARCHAEOAL HISTONE-BASED CHROMATIN STRUCTURES REGULATE TRANSCRIPTION ELONGATION RATES .....	28
Summary .....	28
Introduction .....	29
Results .....	32
Discussion .....	47
Materials and Methods .....	51
REFERENCES .....	56
CHAPTER 3: FTTA IS A CPSF73 HOMOLOGUE THAT TERMINATES TRANSCRIPTION IN ARCHAEOA .....	60
Summary .....	60
Introduction .....	61
Results .....	65
Discussion .....	85
Materials and Methods .....	91
REFERENCES .....	97
CHAPTER 4: DEVELOPMENT OF NASCENT ELONGATING TRANSCRIPT SEQUENCING (NET-SEQ) TECHNIQUES IN THE HYPERTHERMOPHILIC ARCHAEON THERMOCOCCUS KODAKARENSIS .....	102
Introduction .....	102
Results and Discussion .....	107
Conclusions and Future Directions .....	117
Materials and Methods .....	120
REFERENCES .....	124
APPENDIX A: CHAPTER 2 SUPPLEMENTARIES .....	127
Supplementary Figures .....	127
Supplementary Data .....	154
APPENDIX B: CHAPTER 3 SUPPLEMENTARIES .....	158
Supplementary Figures .....	158
CHAPTER 3 SUPPLEMENTARY REFERENCES .....	172
APPENDIX C: HISTONES DIRECT SITE-SPECIFIC CRISPR SPACER ACQUISITION IN MODEL ARCHAEOON .....	173
Main .....	173
Results .....	176
Discussion .....	185
Materials and Methods .....	191
Extended Data .....	198
REFERENCES .....	211

## LIST OF TABLES

### CHAPTER 2

Table 2.1. Residue substitutions alter TECs <sub>58</sub> half-life .....	43
---	----

### CHAPTER 4

Table 4.1. Sequences of primers used in the NET-seq library preparations .....	114
--	-----

## LIST OF FIGURES

### CHAPTER 1

Figure 1.1. The archaeal transcription cycle .....	3
Figure 1.2. Structures of archaeal RNAP .....	6
Figure 1.3. Eukaryotic and archaeal transcription initiation .....	18
Figure 1.4. TFS mediates the rescue of backtracked TECs .....	12
Figure 1.5. Factor-mediated transcription termination in the three Domains .....	16

### CHAPTER 2

Figure 2.1. Variant archaeal histone-based chromatin landscapes dramatically alter the rate of RNA synthesis and pausing patterns of transcription elongation complexes .....	33
Figure 2.2. RNA synthesis is slowed by different archaeal histone-based chromatin landscapes .....	35
Figure 2.3. Substitution of key archaeal histone residues have pronounced effects on RNAP elongation rates .....	37
Figure 2.4. Histone variants with decreased DNA affinity and altered dimer interactions form chromatin structures that minimally impact transcription elongation .....	40
Figure 2.5. Archaeal- and eukaryotic-histone tetrasomes coordinate wrapped DNA nearly identically .....	47

### CHAPTER 3

Figure 3.1. FttA is a <i>bona fide</i> archaeal transcription termination factor .....	66
Figure 3.2. FttA-mediated termination shares mechanistic requirements of rho-mediated bacterial transcription termination .....	71
Figure 3.3. FttA-mediated transcription termination is competitive with transcription elongation .....	75
Figure 3.4. Inhibition of FttA activity abolishes transcription termination <i>in vitro</i> and reduced FttA-expression or activity alters steady-state RNA 3'-termini <i>in vivo</i> .....	80
Figure 3.5. FttA-mediated transcription termination completes the archaeal transcription cycle .....	86
Figure 3.6. Cryo-EM structure of the FttA-modulated pre-termination complex .....	89

### CHAPTER 4

Figure 4.1. Archaeal histones bind and organize the genome .....	104
Figure 4.2. Chromatin architecture modulates the transcriptome .....	106
Figure 4.3. Schematic of the novel, archaeal NET-seq protocol .....	108
Figure 4.4. Mutations and strains developed to assess for regulatory effects on gene expression .....	109
Figure 4.5. HTkA and HTkB have distinct regulatory roles in gene expression .....	111
Figure 4.6. RNAP can be captured and isolated via affinity purification from whole cell lysate (WCL) .....	112
Figure 4.7. Captured TECs <sub>+58</sub> remain intact during lysis and immunoprecipitation .....	113
Figure 4.8. Schematic of NET-seq library prep and analysis pipeline .....	115
Figure 4.9. HTkA and HTkB have distinct regulatory roles in gene expression .....	117

## APPENDIX A

Figure A.1. Representations of variant archaeal histone-based chromatin structures formed with purified HTkA <sup>WT</sup> or HTkA <sup>variants</sup> tetramers bound to ~60 bp of DNA.....	127
Figure A.2. 298 bp DNA template .....	128
Figure A.3. Histone residue substitutions have dramatic effects on full-length RNA transcript abundance.....	129
Figure A.4. TEC progression is differentially impacted by distinct archaeal histone-based chromatin landscapes .....	131
Figure A.5. The dynamics of histone exchange on DNA can dramatically alter TEC progression.....	137
Figure A.6. The first point of contact between TECs and downstream chromatin structures is rate limiting for RNA synthesis .....	139
Figure A.7. Dynamic exchange of histones on DNA is slow compared to TEC translocation .....	141
Figure A.8. TFS increases productive elongation in native and most variant archaeal histone-based chromatin environments .....	143
Figure A.9. 3D representation of TEC progression in a chromatin environment with and without TFS.....	144
Figure A.10. TFS modestly improves transcription restart of TECs <sub>+58</sub> .....	149
Figure A.11. TFS is not essential, and deletion has nominal impacts for <i>T. kodakarensis</i> in optimal conditions.....	151
Figure A.12. Individual residue substitutions elicit large impacts on the stability of archaeal histone-based chromatin landscapes .....	152

## APPENDIX B

Figure B.1. Summary of transcription termination mechanisms commonly employed in Bacteria, Eukarya and Archaea .....	158
Figure B.2. FttA-mediated termination prefers C- and U-rich transcripts.....	160
Figure B.3. The RNA cleavage activity of FttA is stimulated by interactions with the archaeal TEC.....	161
Figure B.4. The active center of FttA is critical to mediate cleavage and release of nascent transcripts.....	162
Figure B.5. FttA-mediated termination requires an intact Spt4/5 complex for maximal activity.....	163
Figure B.6. The stalk domain of the archaeal RNAP is necessary for efficient and kinetically competitive FttA-mediated termination <i>in vitro</i> .....	164
Figure B.7. FttA is an abundant protein likely responsible for 3'-end formation in archaeal cells .....	164
Figure B.8. Gentle purification of FttA directly from lysates of <i>T. kodakarensis</i> strain TK1428D.....	165
Figure B.9. Proteins identified as co-eluting partners of FttA from lysates of strain TK1428D .....	165
Figure B.10. FttA is highly conserved and shares structural and sequence homology with eukaryotic CPSF73.....	166
Figure B.11. Clustal-Omega alignment of diverse archaeal (coral), eukaryotic (green), and bacterial (blue) FttA-homologues.....	167

## APPENDIX C

Figure C.1. MNase protection assay to characterize <i>P. furiosus</i> DNA binding proteins.....	177
Figure C.2. DNA binding by histones.....	179



Figure C.3. Histone gene deletions alter spacer integration at the leader-adjacent repeat <i>in vivo</i> .....	181
Figure C.4. <i>In vitro</i> evaluation of spacer integration in the presence of archaeal histones .....	183
Figure C.5. <i>In vitro</i> spacer integration into different DNA substrates in the presence of <i>P. furiosus</i> histones.....	184
Figure C.6. Hypothetical model for involvement of DNA binding proteins in spacer integration in bacterial and archaeal systems.....	189
Extended Data Figure C.1. Steps of integration .....	198
Extended Data Figure C.2. Distributions of aligned DNA sequences from MNase protection assay .....	199
Extended Data Figure C.3. Positions of MNase protected DNA fragments and promoter elements within leaders of the seven CRISPR loci in <i>P. furiosus</i> .....	201
Extended Data Figure C.4. Nucleotide alignment of seven CRISPR array leaders with MNase protected DNA peaks, promoter elements, and integration sites mapped .....	203
Extended Data Figure C.5. Growth, RNA expression, and histone abundance characteristics in wildtype, $\Delta$ histone A, and $\Delta$ histone B <i>P. furiosus</i> strains used in this study ....	204
Extended Data Figure C.6. Conservation of histones in Euryarchaeota and purification of recombinant proteins .....	206
Extended Data Figure C.7. <i>In vitro</i> evaluation of spacer integration into four CRISPR arrays in the presence or absence of <i>P. furiosus</i> histones .....	207
Extended Data Figure C.8. MNase protection assay to characterize binding patterns for <i>P. furiosus</i> histones incubated with the pCR7-long plasmid .....	208
Extended Data Figure C.9. <i>In vitro</i> evaluation of spacer integration into pCR7-long in the presence or absence of DNA-binding proteins .....	209
Extended Data Figure C.10. Unbiased HTS sequencing of <i>in vitro</i> integration of a spacer into pCR7-long .....	210

## CHAPTER 1: ARCHAEOAL TRANSCRIPTION

### Introduction

Archaea thrive in diverse habitats, are abundant and globally distributed, and are increasingly employed in research and biotechnology settings wherein their unique enzymes and metabolisms provide specific advantages. While several archaeal clades primarily reside in extreme environments, other archaeal clades prosper in more moderate environments, including mammalian microbiomes<sup>1–3</sup>. Archaeal metabolic activities are critical to global carbon, nitrogen, methane, phosphorous, sulfur, and ammonia cycles<sup>4–10</sup>. Many evolutionary reconstructions define Archaea as the third domain but increasing evidence suggests archaea as the direct progenitors of eukaryotes<sup>11</sup>. While prokaryotic, archaea are quite distinct from bacteria<sup>12</sup>. The general size of most archaea and most bacteria are similar, and archaeal genomes, like those of bacteria, are typically gene-dense and circular<sup>13</sup>. The genomes of both domains include many genes organized into operons, lack introns (with the exception of a few self-splicing introns<sup>14</sup>) and contain short intergenic spaces that typically regulate only adjacent genes or a small regulon of gene clusters<sup>15</sup>. As prokaryotes lack membrane-bound organelles, immediate association of ribosomes to nascent RNA transcripts couple transcription to translation<sup>16</sup>. Archaeal transcripts are not capped nor polyadenylated, but some are heavily modified to generate a unique epitranscriptome not typical of Bacteria<sup>17</sup>. Archaeal genomes encode many bacterial-like transcription regulators that aid in repressing (steric hindrance) or activating (recruitment of RNA polymerase (RNAP) or basal transcription factors (TFs)) specific genomic loci<sup>18,19</sup>.

---

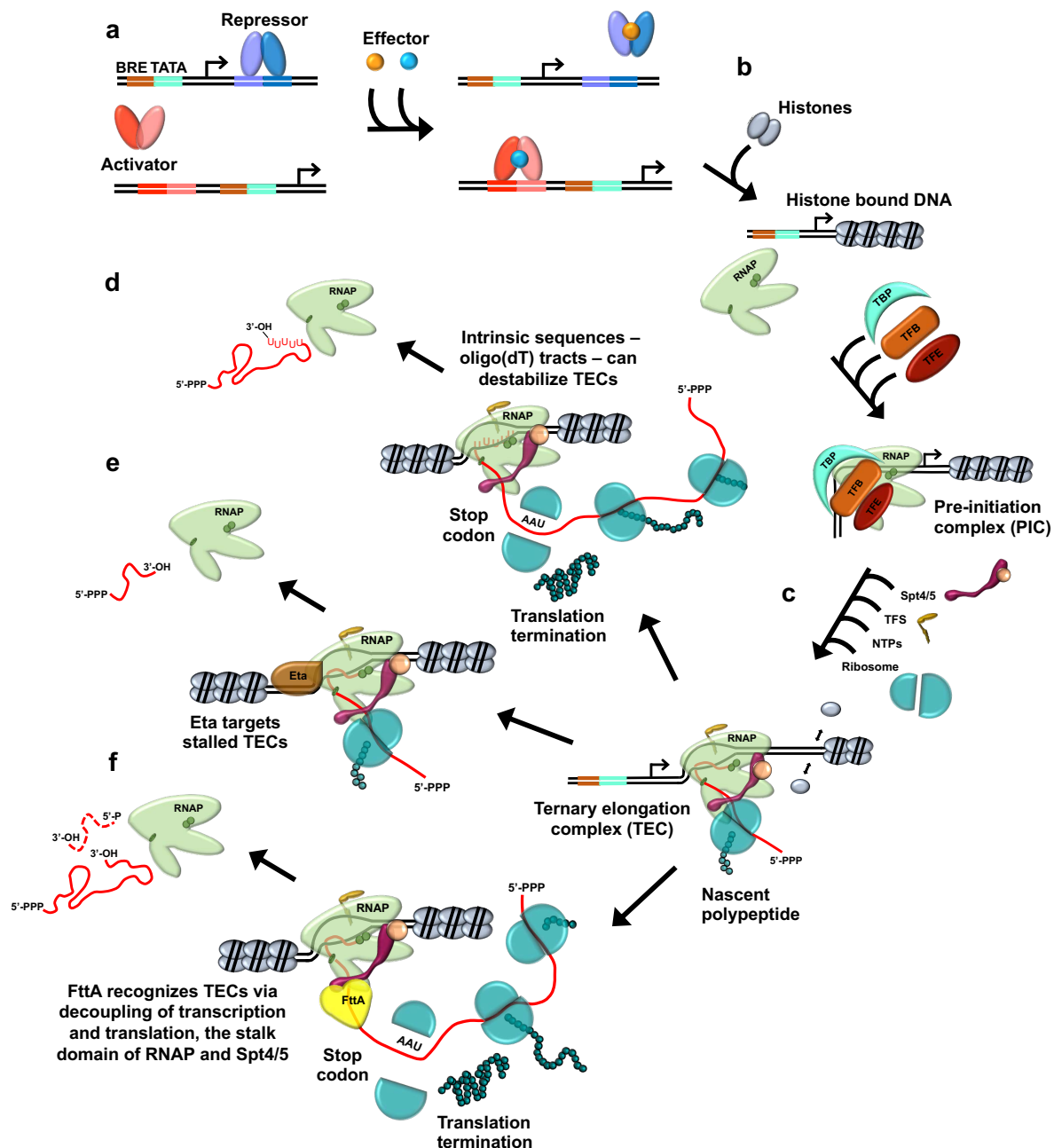
<sup>1</sup> Most of this chapter was previously published under the following title with a few updates: Wenck, B. R. & Santangelo, T. J. Archaeal transcription. *Transcription* **11**, 199–210 (2020).

Long-range interactions mediated by archaeal regulators are known, but silencers, enhancers and post-translational modifications of archaeal-encoded regulators are not dominant regulatory mechanisms to control archaeal transcription<sup>20,21</sup>.

Despite these similarities, archaea and bacteria differ significantly, and perhaps most dramatically at the level of information processing proteins and regulation of such. Archaea universally generate an isoprenoid-derived lipid membrane and employ a plethora of unique and complex metabolic strategies to extract energy from diverse substrates using pathways that are unique to archaeal species. While bioinformatic comparisons have detailed the component simplified but homologous nature of the archaeal transcription apparatus to that of eukaryotes, only more recently have the regulatory mechanisms that control gene expression and support growth in the often rapidly changing niches wherein archaea thrive been examined at the molecular level.

The archaeal transcription cycle can be crudely divided into three main phases: initiation, elongation and termination<sup>22,23</sup> (Fig. 1.1). Although regulatory TFs that influence initiation and direct access of basal TFs to specific genomic loci in archaea tend to be bacterial-like<sup>24</sup>, archaeal genomes encode eukaryotic-like promoter elements and basal TFs to the exclusion of Bacteria<sup>25-27</sup>. Primary among these attributes is the single, multi-subunit archaeal RNAP, typically composed of 11-13 subunits that are often one-to-one homologues of eukaryotic RNA polymerase II (Pol II) subunits<sup>28-30</sup>. In addition to the similarities in the structures of archaeal and eukaryotic RNAPs, both systems are dependent on similar promoter structures and the activities of TATA binding protein (TBP) and Transcription Factor B (TFB; homologous to eukaryotic TFIIB) for formation of pre-initiation complexes; archaeal genomes do not encode sigma factors for promoter recognition<sup>31,32</sup>. Both TBP and TFB recognize shared archaeal-eukaryotic promoter elements, including a TATA-box and B recognition element (BRE), respectively<sup>33</sup>. While TBP,

TFB and the archaeal RNAP are necessary and sufficient to facilitate DNA melting and formation of an open complex in the absence of ATP hydrolysis<sup>26</sup>, contacts within the PIC and between transcription factors and DNA are often optimized by the inclusion of TFE in pre-initiation complexes<sup>34</sup>. All archaea encode the eukaryotic counterpart Transcription Factor E (TFE) – homologous to the TFIIE alpha subunit in eukaryotes – that facilitates formation of open complex and promotes initiation at suboptimal promoter sequences<sup>35</sup>.



**Figure 1.1: The archaeal transcription cycle.** **a.** Binding of archaeal transcription factors to operator sequences near archaeal promoters – defined by BRE and TATA sequences – often inhibits or activates assembly of the pre-initiation complex (PIC). **b.** Promoter-directed archaeal transcription initiation requires only RNAP, TBP, and TFB, although TFE stimulates PIC formation and promoter escape in many instances. Many archaeal genomes encode histone proteins that form a super-helical chromatin structure that can interfere with PIC assembly and hinders processive transcription. **c.** Promoter escape is coupled to the replacement of archaeal initiation factors for the archaeal elongation factors TFS and Spt4/5, which assist RNAP in traversing the chromatin landscape. The absence of membrane-bound compartments permits the immediate association of the translation apparatus with the nascent RNA. **d.** Archaeal TECs can be disrupted by specific DNA sequences – intrinsic termination – wherein weak rU:dA base pairing within the RNA:DNA hybrid lowers TEC stability. **e.** Paused or arrested archaeal TECs can be restored to active complexes or terminated when continued elongation is not possible by the ATP-dependent helicase. **f.** FttA-mediated termination targets TECs uncoupled from the translation apparatus by binding and cleaving the nascent transcript.

Despite the requirements for many additional factors to initiate transcription in eukaryotic systems, archaeal genomes do not encode additional basal TFs nor require ATP-hydrolysis for open complex formation<sup>36</sup>.

Favorable interactions with the promoter and basal TFs are eventually overcome permitting promoter escape<sup>37</sup> and before ternary elongation complexes (TECs) transcribe more than ~50 nts, most TECs exchange contacts with basal initiation factors for contacts with elongation factors<sup>38</sup>. The elongation phase of archaeal transcription consists of a mosaic of bacterial and eukaryotic regulatory strategies with some uniquely archaeal factors<sup>39,40</sup>. Many archaeal genomes encode histone proteins that – akin to eukaryotes – bind DNA to generate a chromatin landscape. Compacted and protein-bound archaeal genomes can impede transcription initiation and elongation<sup>41,42</sup>, and archaeal elongation factors including Spt4/5 (Spt4/5 are conserved in eukaryotes and Spt5 is a homologue of bacterial NusG) and TFS (TFIIS in Eukarya) aid in regulating transcription elongation by assisting RNAP in traversing and maintaining processive elongation through the chromatin environment<sup>43–45</sup>. Despite similarities between some factors that regulate archaeal and eukaryotic elongation, archaea do not appear to encode coactivator complexes nor chromatin-modification or remodeling complexes<sup>46–48</sup>. The patchy distribution of

known bacterial and eukaryotic transcription regulators in archaeal genomes suggests that most transcription can proceed normally without the assistance of factors beyond TFS and Spt4/5, but additional studies are necessary to determine if uniquely archaeal factors can impact transcription in the Archaeal domain.

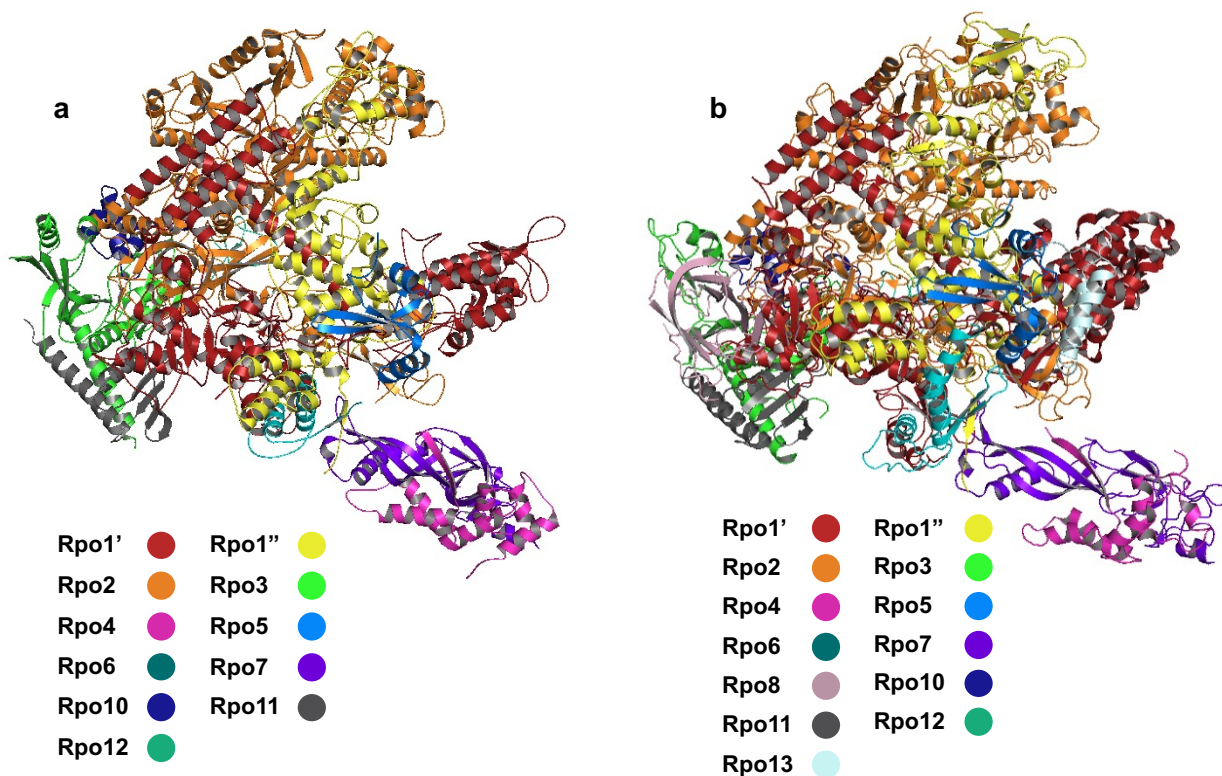
Timely and accurate termination of transcription is critical in gene-dense prokaryotic genomes<sup>49–51</sup>. Archaeal TECs are normally exceptionally stable but transcribing specific sequences or the activity by protein factors can disrupt TECs and terminate transcription. Intrinsic termination has been well-described in each domain<sup>52–54</sup> and intrinsic termination<sup>49,53,55–57</sup> is a prevalent mechanism to disrupt transcription complexes in archaeal species (oligo(dT) tracts in the non-template strand). More recently, factor-dependent archaeal transcription termination mechanisms have been described (Eta and FttA)<sup>58–60</sup>. As was demonstrated for the initiation and elongation phases of transcription, factor-dependent and intrinsic termination mechanisms in archaea share features of bacterial and eukaryotic systems, taking advantage of regulatory mechanisms exploiting coupled transcription and translation and likely promote release of transcription complexes that have arrested.

The entire archaeal transcription cycle has now been defined with purified components<sup>59</sup> and while the collective findings demonstrate that Archaea share regulatory strategies with Bacteria, the molecular machinery is homologous to Eukarya and Archaea utilize exclusive regulatory proteins and approaches not found in bacterial or eukaryotic systems<sup>18,61</sup>. Here we summarize the most impactful and recent findings in archaeal transcription while providing a background of more established parameters for context. The review is not meant as a comprehensive treatise of archaeal transcription but rather as a primer to the newest technologies, findings and approaches surrounding archaeal transcription. The review is focused on the post-initiation regulation of transcription, as such regulation of archaeal transcription has received little

attention in the past and concludes by looking forward towards outstanding questions in the field of archaeal transcription.

### Archaeal RNA polymerase (RNAP)

Multi-subunit RNAPs in each domain share a 5-subunit core with a conserved and central twin-barrel structural fold<sup>28</sup>. In addition to the main core, archaeal RNAPs contain 6-8 additional subunits that surround the central catalytic core and form a stalk domain similar to the stalk domain of eukaryotic polymerase II (Pol II), which is not present in bacterial RNAP<sup>29,32,35,62,63</sup> (Fig. 1.2). Both Bacteria and Archaea encode only a single RNAP that is responsible for transcription of the entire genome, while Eukarya encode several RNAPs with specialized activities that generate gene-family specific transcripts<sup>30</sup>. Despite the homology with Pol II, archaeal RNAP does not encode a C-terminal repeat domain nor has phosphorylation of the large subunit (or any other subunit) of archaeal RNAP as a transcriptional regulation mechanism been demonstrated<sup>64,65</sup>.



**Figure 1.2. Structures of archaeal RNAP.** **a.** The 11-subunit crystal structure of the euryarchaeal RNAP from *Thermococcus kodakarensis* (PDB: 4QIW)<sup>66</sup>. **b.** The 13-subunit crystal structure of RNAP from the crenarchaeal species *Sulfolobus shibatae* (PDB: 2WAQ)<sup>67</sup>.

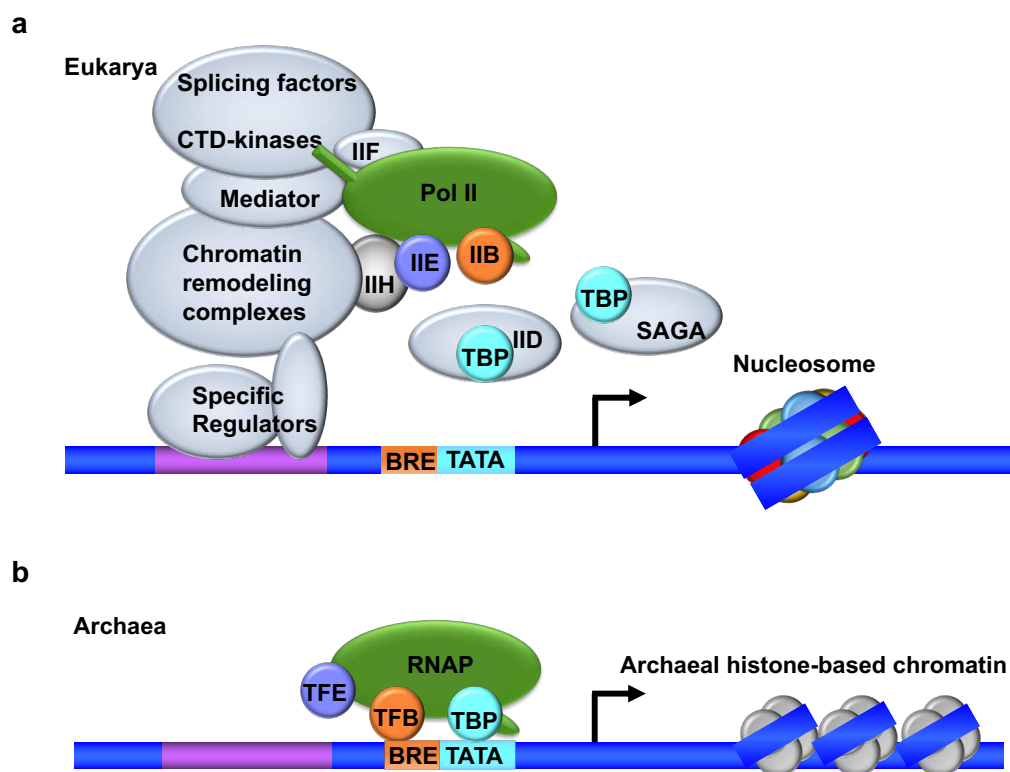
### Archaeal transcription initiation

*In vitro* transcription systems employing archaeal-derived components are often quite robust and efficient<sup>68,69</sup>. Promoter recognition demands only two sequence elements, the TATA-box and the BRE<sup>70</sup>. The A-T rich TATA-box is typically located approximately 25 bp upstream of the transcription start site (TSS) and TBP facilitates recognition of the TATA-box sequences<sup>71</sup>. TSSs in archaea appear to be distinct and thus differ from the more flexible TSSs found in some eukaryotes<sup>72,73</sup>. TBP interaction with the TATA-box results in a relatively symmetrical arrangement that is insufficient to direct the orientation of transcription. Directionality is established by a second sequence element – the BRE located upstream of the TATA-box, recognized by TFB<sup>74</sup>. TBP binding to the TATA-box in eukaryotic systems is often sufficient to stably bend the DNA, stimulating the recruitment of additional basal TFs. Archaeal TBPs form a more dynamic TATA-box associated complex that often is not capable of bending the DNA until TFB recruitment to the BRE<sup>25,33,75</sup>. TBP and TFB binding to the promoter elements forms an oriented and stable conformation for engagement with the archaeal RNAP, permitting subsequent loading and unwinding of dsDNA<sup>26</sup> (Fig. 1.3).

Although TBP and TFB are necessary and sufficient for initiation of near consensus promoters<sup>70</sup>, a third well-conserved archaeal basal TF, transcription factor E (TFE), can facilitate open complex formation and promoter escape at suboptimal promoter sequences or in suboptimal temperatures<sup>34</sup>. All archaeal genomes encode a TFE protein that is a homologue of the eukaryotic TFIIE $\alpha$  subunit from the heterodimer TFIIE $\alpha$ /TFIIE $\beta$  complex, known as TFIIE<sup>76</sup>. Many archaea encode only a single subunit of TFE termed TFE $\alpha$ , which is not strictly required for open complex formation *in vitro*, but does appear essential for viability<sup>36</sup>. A binding partner



for TFE $\alpha$ , termed TFE $\beta$  is encoded in some archaeal clades that functions in concert with TFE $\alpha$  during initiation. Although termed TFE $\beta$ , the archaeal protein has more homology to the Pol III subunit hRCP39 than to TFII $\beta$ <sup>61,77</sup>. In archaeal clades encoding TFE $\alpha/\beta$ , the complex is required for stimulatory effects during initiation, but the impact of TFE $\alpha/\beta$  addition on transcription is still largely dependent on the consensus, or lack thereof of promoter sequences<sup>77</sup>.



**Figure 1.3. Eukaryotic and archaeal transcription initiation.** Transcription initiation in both Eukarya (a) and Archaea (b) utilizes homologous basal factors necessary for initiation in both Domains (matching color components). Archaeal systems are component-simplified when compared to their eukaryotic counterparts.

Some archaea encode more than one TBP and TFB isoform, which provides the possibility of multiple combinations of TBP/TFB to facilitate initiation at various promoters<sup>78</sup>. While the potential for unique TFB/TBP combinations to recognize distinct promoters has been theorized, in systems examined to date, multiple TFB/TBP pairings facilitate recognition of very similar

promoter elements<sup>78</sup>. Although the TATA-box and BRE sequences largely dictate promoter activity, additional elements – the initiator element (INR) located near the transcription start site (TSS)<sup>79</sup> and the promoter proximal element (PPE) located between the BRE/TATA box and the TSS<sup>80</sup> – are known to impact transcription initiation. Neither the INR nor the PPE are required for initiation but can enhance transcriptional output and altering these elements results in reduced promoter strength<sup>80,81</sup>.

### **Archaeal-encoded regulatory transcription factors**

Several excellent reviews regarding archaeal regulatory transcription factors (aRTFs) provide more in-depth information than presented here<sup>18,21,24,82–84</sup>. aRTFs typically function similar to bacterial TFs, directly binding to a sequence-defined operator near critical promoter elements<sup>18</sup>. aRTFs can impede<sup>19</sup> or strengthen<sup>85</sup> interactions between basal TFs and promoter elements. While archaeal repressor operators function to inhibit access of archaeal RNAP and basal TFs to the core promoter – similar to bacteria – activation mechanisms include simple recruitment strategies and may mimic eukaryotic-like mechanisms where several activator binding sites both proximal and distal to the TSS enhance promoter strength<sup>21,86</sup>.

Regulatory/operator sequences in archaea are under the control of families of proteins that primarily consist of one component systems; post-translational modifications of archaeal proteins have been described, however phosphorylation cascades common to eukaryotic and bacterial systems are not common<sup>24,83,84</sup>. The most prevalent aRTFs families are helix-turn-helix DNA-binding proteins containing an N-terminal DNA binding domain and a C-terminal ligand binding domain, commonly responsible for oligomerization following ligand binding<sup>87,88</sup>. The superhelical chromatin architecture in archaeal species encoding histone proteins can impact access to promoter sequences within archaeal genomes, thus interfering with assembly of the

initiation complex<sup>44</sup>. Binding of aRTFs to promoter regions can compete with histone binding to activate initiation<sup>89–92</sup>.

aRTFs are commonly implicated in amino acid metabolism but are also involved in regulating central metabolism, and more recently implicated in regulating cell growth and division<sup>93–95</sup>.

Effectors – such as substrates or ligands – bind to aRTFs and induce conformational changes that allow the basal TFs to access promoter elements<sup>82,96</sup>. Some aRTFs can bind several different effector molecules and contain dual repressor/activator activity depending on the location of the operator sequence<sup>97</sup>.

### **Archaeal transcription elongation**

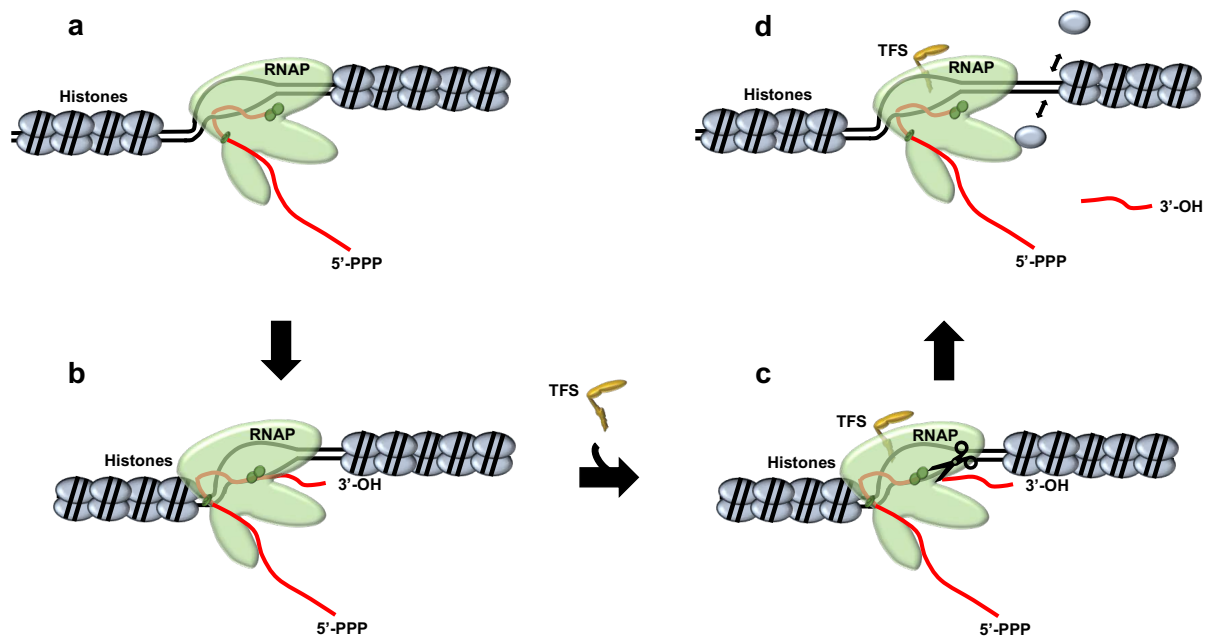
Archaeal transcription elongation is processive but discontinuous due to both sequences and TFs that hinder, pause or delay TEC movement along DNA<sup>98,99</sup>. Exchange of TFE $\alpha$  for the conserved elongation factor Spt5 (Eukarya and Archaea)/NusG (Bacteria) induces conformational changes within the archaeal RNAP complex, encouraging a processive TEC<sup>45,100</sup>. The binding location of Spt5 overlaps with and is in competition with TFE $\alpha$ , and exchange of these factors aids archaeal RNAP in escaping the promoter<sup>45,101</sup>. Spt5/NusG contains a globular N-terminal domain and a (or multiple) C-terminal Kyprides–Onzonis–Woese (KOW) domain(s)<sup>102</sup>. Spt5 in archaea and eukaryotes forms a heterodimer with the small protein Spt4<sup>103</sup>. Escape from the promoter-proximal region from early elongation into productive elongation is rate-limiting and the elongation factors Spt4/5 and Elf1 (Elf1 is a eukaryotic homologue found only in hyperthermophilic Crenarchaea and the Korarchaeon *Candidatus Korarchaeum cryptofilum*<sup>104</sup>) accumulate at these positions with archaeal RNAP<sup>105</sup>. Global occupancy of Spt4/5 correlates distinctly with archaeal RNAP throughout most of the genome, although Spt4/5 is recruited later at a few loci, especially those encoding stable RNAs<sup>106</sup>. This dual mode of Spt4/5 recruitment may yet prove important for shielding TECs on

rRNA and CRISPR loci from attack from transcription termination factors that target the uncoupling of transcription and translation.

Discontinuity of transcription elongation can be due to recent misincorporations, protein roadblocks, sequence directed pausing or arrested (typically backtracked) TECs<sup>40,99,107</sup>. All multi-subunit RNAPs exhibit intrinsic endonuclease activity that can often be stimulated by cleavage induction factors<sup>108,43</sup>. Paused RNAPs are susceptible to backtracking, resulting in an extended RNA 3' end projection through the secondary channel of RNAP, preventing continued polymerization of the nascent RNA<sup>107</sup>. The archaeal cleavage stimulatory factor TFS, homologous to eukaryotic TFIIS and an analogue to bacterial GreA/GreB, interacts with RNAP through the secondary channel and stimulates nascent transcription cleavage to restore the 3' end of the transcript to the active center (Fig. 1.4)<sup>109–111</sup>. Archaeal TFS, like TFIIS proteins, share similar sequences and structures with Rbp9 (a small Pol II subunit implicated in transcriptional proofreading), but only the TFS and TFIIS proteins retain cleavage stimulatory activities and reversibly associate with RNAP<sup>112</sup>. As is true for TFB and TBP, some archaeal genomes encode multiple TFS isoforms. Some TFS paralogues function as a potent inhibitor of archaeal RNAP preventing transcription initiation and elongation and thus play the role of toxins when expressed. As viral infection is correlated with the production of such TFS variants, it is hypothesized that specific TFS variants may form one layer of defense against viral invaders<sup>113</sup>.

Pausing and arrest can be stimulated by downstream DNA protein-barriers<sup>41</sup>. While many proteins transiently associate with DNA, TECs can thus wait for such factors to dissociate before resuming transcription; however, protein barriers with longer half-lives must be overcome to transcribe genes or entire operons in a timely manner. Many archaeal genomes encode and express histone proteins<sup>42</sup> at sufficient levels to saturate the genome with tightly associated histone structures<sup>44</sup>. Purified archaeal histone proteins form chromatin superstructures that dramatically reduce effective transcription rates<sup>114</sup>. Unlike their eukaryotic homologues, archaeal

histones lack obvious N- and C-terminal tails nor are archaeal histones post-translationally modified<sup>115</sup>. The lack of histone remodeling complexes in archaea demands that transcription factors stimulate processive elongation when confronted with histone roadblocks<sup>65</sup>. Spt4/5 aids in processive transcription through the chromatin landscape by reducing TEC pausing while TFS modulates the rescue of backtracked TECs (Fig. 1.4). The activities of Spt4/5 and TFS can act synergistically to stimulate rapid elongation on histone-bound DNA<sup>44</sup>. In addition to their role in reducing transcription rates, archaeal histones are responsible for directing integration of CRISPR spacers to the 5' end of the CRISPR array<sup>116</sup>.



**Figure 1.4. TFS mediates the rescue of backtracked TECs.** Transcription elongation is a discontinuous process as sequences and roadblocks (a), can lead to backtracking where the catalytic site of RNAP becomes disengaged from the 3' end of nascent RNA (b). TFS stimulates the intrinsic endonuclease activity of RNAP (c) and cleavage of the 3' end of the nascent RNA restores the 3' OH to the active site and reactivates TECs (d).

Not all archaea encode histone proteins but all archaea contain nucleoid associating proteins (NAPs) that organize genomes and directly or indirectly impact transcription regulation<sup>41,117</sup>. Alterations to the three-dimensional structure of chromosomes may consequently affect gene expression and archaeal genomes encode many small proteins capable of impacting the

genomic landscape<sup>41</sup>. Many archaeal clades, with the exception of the Crenarchaeota, encode structural maintenance of chromosomes (SMC) complexes that form archaeal condensins implicated in DNA compaction and organization; however, the molecular functions of such remain poorly understood<sup>118,119</sup>. The physical organization of archaeal chromosomes has only been investigated in archaeal species lacking histone proteins but reveals dramatic separation of large regions of the chromosome into distinct domains<sup>120,121</sup>. Genome organization is driven, at least in part and within some members of *Sulfolobus*, by the activities of a coalescing (CIsN) protein, a member of the larger SMC family<sup>122</sup>. Regions of the genome heavily populated by CIsN display higher transcription activity, and it is likely that transcription and genome organization are reciprocally regulated in many archaea<sup>121</sup>.

### **Archaeal transcription termination**

The extremely stable TEC has the ability to remain engaged and transcribe the DNA template for extended periods of time<sup>123</sup>. Archaeal gene-dense genomes have little intergenic space, which demands that the otherwise extremely stable TEC be efficiently disrupted before transcription can continue into neighboring genes or operons resulting in the synthesis of aberrant transcripts<sup>16,124</sup>. Archaeal TECs are sensitive to both sequence-dependent termination mechanisms (i.e. intrinsic termination<sup>53,56</sup>) and to disruption by protein factors (i.e. factor-dependent termination<sup>58-60</sup>). The bacterial RNAP, the archaeal RNAP and eukaryotic RNAP III (Pol III) are all responsive to intrinsic termination signals<sup>50,52,56</sup>. Intrinsic termination signals from each domain utilize conserved oligo(dT) tracts, but the mechanisms and accessory sequences inducing transcription termination vary in each domain. Archaeal TECs are most effectively disrupted by oligo(dT) sequences encoding U-rich 3' ends, but unlike bacterial intrinsic termination sequences, archaeal TECs are easily disrupted in the absence of RNA hairpin structures<sup>52,125</sup>. Genome-wide mapping of steady-state 3' RNA termini in archaeal species supports oligo(dT) tracts as important intrinsic termination features and suggests that many

secondary termination sites or RNA processing events can ensure termination events and conditional coding strategies<sup>55</sup>.

Factor-dependent transcription termination has been described for decades in Eukarya and Bacteria<sup>57</sup>, but factors responsible for termination in archaea have only more recently been described<sup>58–60,105,126,127</sup>. Most euryarchaeal species encode a superfamily II helicase with transcription termination activity, termed the euryarchaeal termination activity, or Eta. Eta-mediated transcription termination shares many attributes of transcription termination modulated by the bacterial Mfd protein<sup>58</sup>. Eta, like Mfd, transiently engages TECs and can restore stalled TECs to active configurations; when continued elongation is not possible, TECs are disrupted<sup>128</sup>. While translocase, ATPase, and helicase activities are a requirement for Eta-mediated termination, these activities are not the direct cause of TEC disengagement resulting in termination. Substitutions of select residues within the CTD of variants that retain translocase, ATPase and helicase activity not only abolish or significantly decrease termination efficiency but also reduces Eta-mediated rescue of backtracked TECs. Complete deletion of the C-terminal domain (CTD) completely abolishes Eta-mediated termination and together with structural modeling, suggests termination requires interactions between the TEC and the Eta CTD, like that of the Mfd-TEC complex<sup>126</sup>. DNA lesions on the template strand that pass through the active center of the archaeal RNAP halt TEC progression at or adjacent to the damage site<sup>129</sup> supporting the need for factors to disrupt TECs and expose the damaged DNA for repair. While Mfd has been well-characterized and has been shown to recruit nucleotide excision repair (NER) machinery as part of the transcription coupled repair (TCR) pathway<sup>130</sup>, detailed TCR mechanisms in archaea remain evasive to date. Eta also likely functions in RNA processing and degradation pathways in the euryarchaeota<sup>131</sup>.

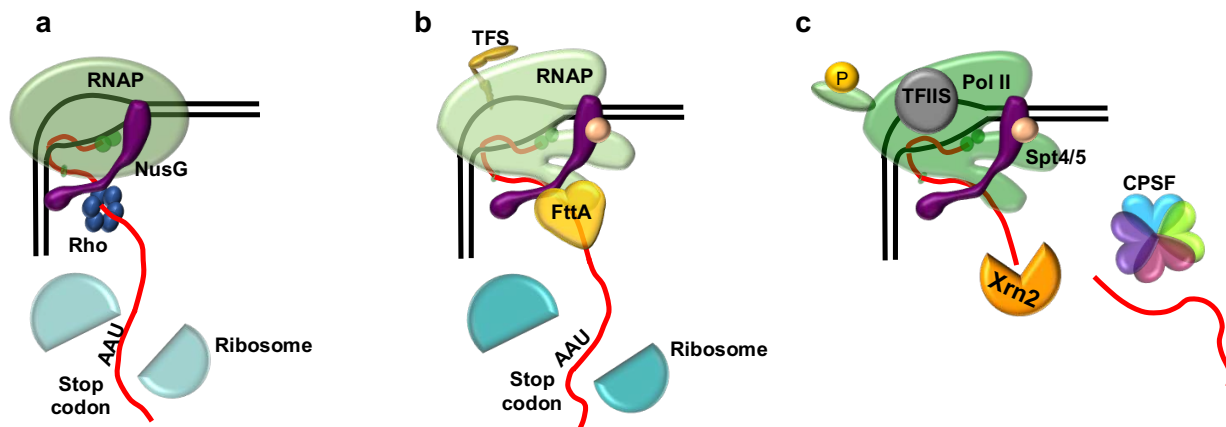
Although Eta can disrupt TECs, Eta-mediated termination is not competitive with normal elongation rates and thus Eta-mediated termination is unlikely to be responsible for polarity or

termination at the ends of genes and operons that do not encode intrinsic terminators. Operons are common in both of the prokaryotic domains (Bacteria and Archaea) and close association of the transcription and translation machineries prevents premature termination of operon transcription<sup>124</sup>; however, in select bacterial species Rho-mediated termination is isolated to G-C rich regions of the genomes and independent of translation machinery<sup>132</sup>. The coupling of transcription and translation in archaea, similar to certain species of bacteria, provides a trailing ribosome to TECs during transcription of an open reading frame that can be exploited as a regulatory signal to influence transcription termination<sup>16,133,134</sup> (Fig 1.5a). When the trailing and protective ribosome dissociates from the nascent transcripts, still elongating TECs are the substrate of a universally-archaeal conserved termination factor, termed FttA (Factor that Terminates Transcription in Archaea)<sup>135,136</sup> (Fig. 1.5b). FttA is a member of the  $\beta$ -CASP ribonuclease family<sup>137</sup> – one of many archaeal RNA-processing protein families responsible for the maturation and degradation of RNA transcripts<sup>17</sup>. FttA, otherwise known as aCPSF1<sup>135</sup>, is an orthologue of the eukaryotic CPSF73 subunit of the cleavage and polyadenylation factor complex associated with eukaryotic Pol I and II transcription termination (Fig. 1.5c).

FttA-mediated termination is competitive with normal transcription rates and is likely the mechanism of termination for genes/operons lacking intrinsic termination sequences and as a backup mechanism due to imperfect activity of intrinsic termination sequences<sup>55,60,127</sup>. FttA-mediated termination is responsible for the robust endonucleolytic cleavage of nascent RNA and the rapid 5'-3' exonuclease activity<sup>135</sup> that completely digests the TEC-associated RNA 3' terminus, which ultimately disrupts TECs<sup>60</sup>. FttA-mediated termination is reliant on interactions between FttA, RNAP and the nascent RNA, and the rate of FttA-mediated termination is increased by Spt4/5<sup>59</sup>, much like bacterial rho activity can be stimulated by NusG<sup>138</sup>. Depletion of FttA results in transcription read-through and an altered transcriptome, implicating FttA as a global transcription termination factor<sup>59,60</sup>. FttA (aCPSF1) was determined to form a dimer in



solution, mediated by the CTD<sup>135,139,140</sup>, where deletion of the CTD as well as removal of the KH domains (N-terminus) prevents FttA-mediated termination activity<sup>127</sup>. While the KH domains of FttA were determined to bind specifically to U-rich tracts (intrinsic terminator sequences), the claim that this binding is vital for aCPSF1 termination activity *in vivo*<sup>127</sup> is invalidated by FttA-mediated transcription termination seen *in vitro* (transcripts were not U-rich)<sup>59</sup> and *in vivo* termination activity at loci that do not encode U-rich tracts. Interestingly, recruitment of aCPSF1 to promoter-proximal regions (in early elongation with no oligo(dT) tracts) is correlated with decreased TEC escape into productive elongation and low RNA levels<sup>105</sup>. Like many aspects of archaeal transcription systems, FttA-mediated termination combines features of bacterial systems (recognition of uncoupling of transcription and translation) with proteins that have eukaryotic homology (CPSF73) but that often act alone or in much less complex assemblies than their eukaryotic counterparts<sup>59</sup>. As both bacterial rho and eukaryotic CPSF73 are critical contributors to RNA processing, it is highly likely that FttA plays a similar role in archaeal cells.



**Figure 1.5. Factor-mediated transcription termination in the three Domains.** The extremely stable TEC can be disrupted via factor-mediated termination. The uncoupling of transcription to translation in both bacteria (a) and archaea (b) provides an exposed nascent transcript for factor-mediated transcription termination. a. Rho-mediated termination is enhanced via interactions with NusG<sup>141</sup>. b. When kinetically coupled to Spt5 and the stalk domain of RNAP, FttA-mediated termination is competitive with elongation. FttA is an orthologue of the eukaryotic subunit CPSF73 (c) and contains similar endonucleolytic cleavage activity of nascent RNA. c. Although not directly responsible for transcription termination, cleavage activity of nascent mRNA in Pol II elongating complexes by CPSF73 is the ultimate signal for termination to occur via Xrn2-mediated activity<sup>142,143</sup>.

## **Concluding remarks**

Archaeal transcription regulation combines aspects from both Bacteria and Eukarya and new findings illuminate uniquely archaeal regulatory strategies that aid to the complexity of gene regulatory systems in extant life. The component simplicity of archaeal systems beguiles the complex and interconnected strategies that control each stage of the transcription cycle. Novel mechanisms regarding archaeal transcription are identified at an increasing rate in conjunction with new technological advances. Next generation sequencing (NGS) techniques have highlighted the complexity of the transcriptome and the delicate interplay of differential gene expression output and the consequence of altering normal protein functions. Although many discoveries of the transcriptome using NGS techniques have enlightened the archaeal community regarding transcription regulatory mechanisms, many aspects of archaeal transcription regulation await further study<sup>15,55,144</sup>.

In particular, the global influence of chromosome architecture and the impacts of chromatin organization undoubtedly hold promise for understanding the origins of the many chromatin-remodeling and histone-modification machineries ubiquitous to eukaryotes, but seemingly absent in archaeal species. In archaeal genomes with no apparent histone genes, NAPs likely regulate DNA compaction and gene expression. Archaeal genomes encode an abundance of small proteins capable of impacting the chromatin landscape that have the potential to impact gene regulatory networks<sup>41</sup>. Further evidence regarding alterations to the archaeal chromatin structure may highlight the role of DNA-bound proteins surrounding gene regulatory mechanisms and provide an extensive view into the importance of archaeal DNA compaction and transcription regulation. As CIsN (SMC superfamily member) proteins play a significant role in genomic organization and higher rates of transcription activity in archaeal species lacking histone proteins, the archaeal SMC superfamily member, condensin or SMC-like proteins, may

support a larger role in chromatin rearrangement, allowing for transcription regulation in archaeal clades encoding histone proteins<sup>121</sup>.

The archaeal transcription cycle has recently been completed with the identification of FttA as a global transcription termination factor, but the discovery and attractive combination of bacterial and eukaryotic characteristics highlights what little is known regarding protein-protein interactions; key factors that influence termination and RNA processing mechanisms. All evidence to date on archaeal TTSs has looked specifically at 3' termini with an emphasis on oligo(dT) tracts as canonical sites of termination<sup>55,60</sup>. While stretches of oligo(dT) can destabilize purified TEC components resulting in termination without the assistance of accessory proteins, not all coding sequences in archaeal genomes encode oligo(dT) tracts<sup>53</sup>. Transcriptional regulation occurs throughout the transcription cycle and determining true 3' termini given specific conditions should include observing the entirety of RNA transcripts to detail individual gene and operon complexity.

The connections between replication and transcription in archaeal clades have yet to be explored, nor have transcription and DNA repair been definitively linked in archaeal species. Finally, no anti-termination complexes nor mechanisms to ensure transcription of stable RNAs been defined in archaeal systems. Establishing the details of archaeal transcription regulation from the organismal to molecular level will surely continue to shed light on shared and unique strategies to control gene expression and modulate the activities of multi-subunit RNAPs.

## Thesis rationale

Archaea are single-celled prokaryotes that are distantly related to Bacteria with a similar general cell structure and S-layer (surface layer) proteins present on the outside of cells walls. Archaeal genomes share a gene-dense, circular structure with Bacteria, encompassing operons and short intergenic regions. While archaeal genomes encode bacterial-like transcription regulators that aid in repressing or activating specific genomic loci, archaeal genomes encode simplified eukaryotic-like promoter elements, basal transcription factors (TFs), and select elongation factors. In addition, many archaeal species encode histone proteins that share a near-identical fold to the core eukaryotic histone proteins that permits DNA wrapping through select histone-DNA contacts to generate chromatin-structures, which must be sequentially traversed by RNA polymerase to generate transcripts. These aspects of archaeal genomes, archaeal transcription systems, and histone-DNA interactions are obvious homologues of eukaryotic systems, to the exclusion of Bacteria.

However, outstanding questions remain regarding regulatory mechanisms of archaeal transcription. Although the physicochemical properties that act on archaeal histone-based chromatin influence the structure and dynamics of histone-based chromatin, our knowledge of *how* changes to the chromatin landscape influences gene expression is incomplete. Additionally, factor-dependent transcription termination has been described for decades in Eukarya and Bacteria, but factors responsible for termination in Archaea have only more recently been described.

The work done in this thesis determined that specific residues within the archaeal histone-DNA complex are essential in maintaining native transcription elongation rates. Altering the histone-based chromatin structure via a single amino acid substitution greatly impacts the progression of RNA polymerase thereby transforming the transcription elongation rate. These results detail

the importance of distinct chromatin structures for archaeal gene expression and provide a unique perspective on the evolution of, and regulatory strategies imposed by, eukaryotic chromatin. Efforts to explore these effects *in vivo* are extensive, ongoing, and will illuminate the influence of the core histone fold on chromatin architecture, transcription regulation, and gene expression. Furthermore, the identification and characterization of the Factor that terminates transcription in Archaea (FttA) completes the archaeal transcription cycle and links transcription regulation between the three Domains. FttA-mediated transcription termination shares many attributes with rho-mediated transcription termination (Bacteria), but FttA retains a structure with obvious homology to the eukaryotic CPSF73 protein. FttA-mediated termination resolves the dichotomy of a prokaryotic gene structure (operons and polarity) and eukaryotic molecular homology (general transcription apparatus) observed in Archaea.

## REFERENCES

1. DeLong, E. F. & Pace, N. R. Environmental Diversity of Bacteria and Archaea. *Syst. Biol.* **50**, 470–478 (2001).
2. Barns, S. M., Fundyga, R. E., Jeffries, M. W. & Pace, N. R. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 1609–1613 (1994).
3. Bintrim, S. B., Donohue, T. J., Handelsman, J., Roberts, G. P. & Goodman, R. M. Molecular phylogeny of Archaea from soil. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 277–282 (1997).
4. Berg, I. A. *et al.* Autotrophic carbon fixation in archaea. *Nature Reviews Microbiology* **8**, 447–460 (2010).
5. Castelle, C. J. *et al.* Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).
6. Evans, P. N. *et al.* An evolving view of methane metabolism in the Archaea. *Nature Reviews Microbiology* **17**, 219–232 (2019).
7. Yadav, A. N. *et al.* Haloarchaea Endowed with Phosphorus Solubilization Attribute Implicated in Phosphorus Cycle. *Sci. Rep.* **5**, 12293 (2015).
8. Ino, K. *et al.* Ecological and genomic profiling of anaerobic methane-oxidizing archaea in a deep granitic environment. *ISME J.* **12**, 31–47 (2018).
9. Prosser, J. I. & Nicol, G. W. Relative contributions of archaea and bacteria to aerobic ammonia oxidation in the environment. *Environmental Microbiology* **10**, 2931–2941 (2008).
10. Seitz, K. W., Lazar, C. S., Hinrichs, K. U., Teske, A. P. & Baker, B. J. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* **10**, 1696–1705 (2016).
11. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nature Reviews Microbiology* **15**, 711–723 (2017).
12. Inagaki, F. *et al.* Archaeology of Archaea: Geomicrobiological record of Pleistocene thermal events concealed in a deep-sea seafloor environment. *Extremophiles* **5**, 385–392 (2001).
13. Edgell, D. R. & Doolittle, W. F. Archaea and the origin(s) of DNA replication proteins. *Cell* **89**, 995–998 (1997).
14. Nawrocki, E. P., Jones, T. A. & Eddy, S. R. Group I introns are widespread in archaea. *Nucleic Acids Res.* **46**, 7970–7976 (2018).
15. Wurtzel, O. *et al.* A single-base resolution map of an archaeal transcriptome. *Genome Res.* **20**, 133–141 (2010).
16. French, S. L., Santangelo, T. J., Beyer, A. L. & Reeve, J. N. Transcription and translation are coupled in Archaea. *Mol. Biol. Evol.* **24**, 893–5 (2007).
17. Clouet-D'Orval, B. *et al.* Insights into RNA-processing pathways and associated RNA-degrading enzymes in Archaea. *FEMS Microbiol. Rev.* **016**, 579–613 (2018).
18. Karr, E. A. Transcription regulation in the third domain. in *Advances in Applied Microbiology* **89**, 101–133 (Academic Press Inc., 2014).
19. Bell, S. D., Cairns, S. S., Robson, R. L. & Jackson, S. P. Transcriptional regulation of an Archaeal operon in vivo and in vitro. *Mol. Cell* **4**, 971–982 (1999).
20. Buffry, A. D., Mendes, C. C. & McGregor, A. P. The Functionality and Evolution of Eukaryotic Transcriptional Enhancers. *Adv. Genet.* **96**, 143–206 (2016).
21. Bylino, O. V., Ibragimov, A. N. & Shidlovskii, Y. V. Evolution of Regulated Transcription. *Cells* **9**, 1675 (2020).
22. Decker, K. B. & Hinton, D. M. Transcription Regulation at the Core: Similarities Among Bacterial, Archaeal, and Eukaryotic RNA Polymerases. *Annu. Rev. Microbiol.* **67**, 113–

- 139 (2013).
23. Fouqueau, T., Zeller, M. E., Cheung, A. C., Cramer, P. & Thomm, M. The RNA polymerase trigger loop functions in all three phases of the transcription cycle. *Nucleic Acids Res.* **41**, 7048–7059 (2013).
24. Lemmens, L., Maklad, H. R., Bervoets, I. & Peeters, E. Transcription Regulators in Archaea: Homologies and Differences with Bacterial Regulators. *Journal of Molecular Biology* **431**, 4132–4146 (2019).
25. Qureshi, S. A. & Jackson, S. P. Sequence-specific DNA binding by the *S. shibatae* TFIIB homolog, TFB, and its effect on promoter strength. *Mol. Cell* **1**, 389–400 (1998).
26. Kosa, P. F., Ghosh, G., Dedecker, B. S. & Sigler, P. B. The 2.1-Å crystal structure of an archaeal preinitiation complex: TATA-box-binding protein/transcription factor (II)B core/TATA-box. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 6042–6047 (1997).
27. Hausner, W. & Thomm, M. Events during initiation of Archaeal transcription: Open complex formation and DNA-protein interactions. *J. Bacteriol.* **183**, 3025–3031 (2001).
28. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nature Reviews Microbiology* **9**, 85–98 (2011).
29. Grohmann, D., Hirtreiter, A. & Werner, F. Molecular mechanisms of archaeal RNA polymerase. *Biochem. Soc. Trans.* **37**, 12–17 (2009).
30. Hsieh, M.-L. & Borger, J. *Biochemistry, RNA Polymerase*. StatPearls (StatPearls Publishing, 2020).
31. Murakami, K. S. Structural biology of bacterial RNA polymerase. *Biomolecules* **5**, 848–864 (2015).
32. Hirata, A. & Murakami, K. S. Archaeal RNA polymerase. *Current Opinion in Structural Biology* **19**, 724–731 (2009).
33. Littlefield, O., Korkhin, Y. & Sigler, P. B. The structural basis for the oriented assembly of a TBP/TFB/promoter complex. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 13668–13673 (1999).
34. Bell, S. D., Brinkman, A. B., Oost, J. van der & Jackson, S. P. The archaeal TFIIE $\alpha$  homologue facilitates transcription initiation by enhancing TATA-box recognition. *EMBO Rep.* **2**, 133–138 (2001).
35. Nagy, J. *et al.* Complete architecture of the archaeal RNA polymerase open complex from single-molecule FRET and NPS. *Nat. Commun.* **6**, 6161 (2015).
36. Blombach, F., Smollett, K. L., Grohmann, D. & Werner, F. 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).
37. Revyakin, A., Liu, C., Ebright, R. H. & Strick, T. R. Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science (80-. ).* **314**, 1139–1143 (2006).
38. Grohmann, D. *et al.* The Initiation Factor TFE and the Elongation Factor Spt4/5 Compete for the RNAP Clamp during Transcription Initiation and Elongation. *Mol. Cell* **43**, 263–274 (2011).
39. Werner, F. Molecular mechanisms of transcription elongation in archaea. *Chemical Reviews* **113**, 8331–8349 (2013).
40. Washburn, R. S. & Gottesman, M. E. Regulation of transcription elongation and termination. *Biomolecules* **5**, 1063–1078 (2015).
41. Sanders, T. J., Marshall, C. J. & Santangelo, T. J. The Role of Archaeal Chromatin in Transcription. *Journal of Molecular Biology* (2019). doi:10.1016/j.jmb.2019.05.006
42. Henneman, B., van Emmerik, C., van Ingen, H. & Dame, R. T. Structure and function of archaeal histones. *PLoS Genetics* **14**, e1007582 (2018).
43. Fish, R. N. & Kane, C. M. Promoting elongation with transcript cleavage stimulatory factors. *Biochimica et Biophysica Acta - Gene Structure and Expression* **1577**, 287–307 (2002).

44. Sanders, T. J. *et al.* TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Mol. Microbiol.* **111**, 784–797 (2019).
45. Hirtreiter, A. *et al.* Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. *Nucleic Acids Res.* **38**, 4040–4051 (2010).
46. Burton, A. *et al.* Heterochromatin establishment during early mammalian development is regulated by pericentromeric RNA and characterized by non-repressive H3K9me3. *Nat. Cell Biol.* **22**, 767–778 (2020).
47. Armache, A. *et al.* Histone H3.3 phosphorylation amplifies stimulation-induced transcription. *Nature* 1–6 (2020). doi:10.1038/s41586-020-2533-0
48. Deal, R. B. & Henikoff, S. Histone variants and modifications in plant gene regulation. *Current Opinion in Plant Biology* **14**, 116–122 (2011).
49. Maier, L.-K. & Marchfelder, A. It's all about the T: transcription termination in archaea. *Biochem. Soc. Trans.* **47**, 461–468 (2019).
50. Ray-Soni, A., Bellecourt, M. J. & Landick, R. Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annu. Rev. Biochem.* **85**, 319–347 (2016).
51. Peters, J. M., Vangeloff, A. D. & Landick, R. Bacterial transcription terminators: The RNA 3'-end chronicles. *Journal of Molecular Biology* **412**, 793–813 (2011).
52. Arimbasseri, A. G., Rijal, K. & Maraia, R. J. Transcription termination by the eukaryotic RNA polymerase III. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1829**, 318–330 (2013).
53. Santangelo, T. J. & Reeve, J. N. Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *J. Mol. Biol.* **355**, 196–210 (2006).
54. Santangelo, T. J. & Artsimovitch, I. Termination and antitermination: RNA polymerase runs a stop sign. *Nat. Rev. Microbiol.* **9**, 319–329 (2011).
55. Dar, D., Prasse, D., Schmitz, R. A. & Sorek, R. Widespread formation of alternative 3' UTR isoforms via transcription termination in archaea. *Nat. Microbiol.* **1**, 1–9 (2016).
56. Santangelo, T. J., Cubonová, L., Skinner, K. M. & Reeve, J. N. Archaeal Intrinsic Transcription Termination In Vivo. *J. Bacteriol.* **191**, 7102–7108 (2009).
57. Porrua, O., Boudvillain, M. & Libri, D. Transcription Termination: Variations on Common Themes. *Trends in Genetics* **32**, 508–522 (2016).
58. Walker, J. E., Luyties, O. & Santangelo, T. J. Factor-dependent archaeal transcription termination. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6767–E6773 (2017).
59. Sanders, T. J. *et al.* FttA is a CPSF73 homologue that terminates transcription in Archaea. *Nature Microbiology* **5**, 545–553 (2020).
60. Yue, L. *et al.* The conserved ribonuclease aCPSF1 triggers genome-wide transcription termination of Archaea via a 3'-end cleavage mode. *Nucleic Acids Res.* (2020). doi:10.1093/nar/gkaa702
61. Gehring, A. M., Walker, J. E. & Santangelo, T. J. Transcription Regulation in Archaea. *J. Bacteriol.* **198**, 1906–1917 (2016).
62. Hirata, A. *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. *Mol. Microbiol.* **70**, 623–33 (2008).
63. Jun, S. H., Reichlen, M. J., Tajiri, M. & Murakami, K. S. Archaeal RNA polymerase and transcription regulation. *Crit. Rev. Biochem. Mol. Biol.* **46**, 27–40 (2011).
64. Schüller, R. *et al.* Heptad-Specific Phosphorylation of RNA Polymerase II CTD. *Mol. Cell* **61**, 305–314 (2016).
65. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* **9**, 85–98 (2011).
66. Jun, S. H. *et al.* The X-ray crystal structure of the euryarchaeal RNA polymerase in an open-clamp configuration. *Nat. Commun.* **5**, (2014).



67. Korkhin, Y. *et al.* Evolution of complex RNA polymerases: The complete archaeal RNA polymerase structure. *PLoS Biol.* **7**, (2009).
68. Gehring, A. M. & Santangelo, T. J. Manipulating archaeal systems to permit analyses of transcription elongation-termination decisions in vitro. *Methods Mol. Biol.* **1276**, 263–79 (2015).
69. Dexl, S. *et al.* Displacement of the transcription factor B reader domain during transcription initiation. *Nucleic Acids Res.* **46**, 10066–10081 (2018).
70. Schulz, S., Kramm, K., Werner, F. & Grohmann, D. Fluorescently labeled recombinant RNAP system to probe archaeal transcription initiation. *Methods* **86**, 10–18 (2015).
71. Soppa, J. Transcription initiation in Archaea: facts, factors and future aspects. *Mol. Microbiol.* **31**, 1295–1305 (1999).
72. Dikstein, R. The unexpected traits associated with core promoter elements. *Transcription* **2**, 201–206 (2011).
73. Blombach, F. & Grohmann, D. Same same but different: The evolution of TBP in archaea and their eukaryotic offspring. *Transcription* **8**, 162–168 (2017).
74. Ouhammouch, M., Hausner, W. & Geiduschek, E. P. TBP domain symmetry in basal and activated archaeal transcription. *Mol. Microbiol.* **71**, 123–131 (2009).
75. Lee, S. & Hahn, S. Model for binding of transcription factor TFIIB to the TBP-DNA complex. *Nature* **376**, 609–612 (1995).
76. Compe, E., Genes, C. M., Braun, C., Coin, F. & Egly, J. M. TFIIE orchestrates the recruitment of the TFIIH kinase module at promoter before release during transcription. *Nat. Commun.* **10**, 1–14 (2019).
77. Blombach, F. *et al.* Archaeal TFE $\alpha$ / $\beta$  is a hybrid of TFIIE and the RNA polymerase III subcomplex hRPC62/39. *Elife* **4**, (2015).
78. Santangelo, T. J., Čuboňová, L. L., James, C. L. & Reeve, J. N. TFB1 or TFB2 Is Sufficient for *Thermococcus kodakaraensis* Viability and for Basal Transcription in Vitro. *J. Mol. Biol.* **367**, 344–357 (2007).
79. Ao, X. *et al.* The *Sulfolobus* initiator element is an important contributor to promoter strength. *J. Bacteriol.* **195**, 5216–5222 (2013).
80. Peng, N., Xia, Q., Chen, Z., Liang, Y. X. & She, Q. An upstream activation element exerting differential transcriptional activation on an archaeal promoter. *Mol. Microbiol.* **74**, 928–939 (2009).
81. Cho, S. *et al.* Genome-wide primary transcriptome analysis of H<sub>2</sub>-producing archaeon *Thermococcus onnurineus* NA1. *Sci. Rep.* **7**, 1–12 (2017).
82. Martinez-Pastor, M., Tonner, P. D., Darnell, C. L. & Schmid, A. K. Transcriptional Regulation in Archaea: From Individual Genes to Global Regulatory Networks. *Annu. Rev. Genet.* **51**, 143–170 (2017).
83. Krell, T. Exploring the (almost) unknown: Archaeal two-component systems. *J. Bacteriol.* **200**, (2018).
84. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Phyletic distribution and lineage-specific domain architectures of archaeal two-component signal transduction systems. *J. Bacteriol.* **200**, 681–698 (2018).
85. Ochs, S. M. *et al.* Activation of archaeal transcription mediated by recruitment of transcription factor B. *J. Biol. Chem.* **287**, 18863–18871 (2012).
86. Flores-Bautista, E. *et al.* Deciphering the functional diversity of DNA-binding transcription factors in Bacteria and Archaea organisms. *PLoS One* **15**, e0237135 (2020).
87. Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiol. Rev.* **29**, 231–262 (2005).
88. Aravind, L. & Koonin, E. V. DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.* **27**, 4658–4670 (1999).

89. Wilkinson, S. P., Ouhammouch, M. & Geiduschek, E. P. Transcriptional activation in the context of repression mediated by archaeal histones. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 6777–6781 (2010).
90. Wierer, S. *et al.* TrmBL2 from *Pyrococcus furiosus* Interacts Both with Double-Stranded and Single-Stranded DNA. *PLoS One* **11**, e0156098 (2016).
91. Efremov, A. K. *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. *J. Biol. Chem.* **290**, 15770–15784 (2015).
92. Maruyama, H. *et al.* Histone and TK0471/TrmBL2 form a novel heterogeneous genome architecture in the hyperthermophilic archaeon *Thermococcus kodakarensis*. *Mol. Biol. Cell* **22**, 386–398 (2011).
93. Rodionova, I. A. *et al.* A novel bifunctional transcriptional regulator of riboflavin metabolism in Archaea. *Nucleic Acids Res.* **45**, 3785–3799 (2017).
94. van de Werken, H. J. G., Verhees, C. H., Akerboom, J., de Vos, W. M. & van der Oost, J. Identification of a glycolytic regulon in the archaea *Pyrococcus* and *Thermococcus*. *FEMS Microbiol. Lett.* **260**, 69–76 (2006).
95. Darnell, C. L. *et al.* The Ribbon-Helix-Helix Domain Protein CdrS Regulates the Tubulin Homolog ftsZ2 To Control Cell Division in Archaea. *MBio* **11**, e01007-20 (2020).
96. Darnell, C. L. *et al.* Systematic Discovery of Archaeal Transcription Factor Functions in Regulatory Networks through Quantitative Phenotyping Analysis Downloaded from. **2**, 32–49
97. Kanai, T. *et al.* A global transcriptional regulator in *Thermococcus kodakaraensis* controls the expression levels of both glycolytic and gluconeogenic enzyme-encoding genes. *J. Biol. Chem.* **282**, 33659–33670 (2007).
98. Mayer, A., Landry, H. M. & Churchman, L. S. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Current Opinion in Cell Biology* **46**, 72–80 (2017).
99. Smirnov, E., Hornáček, M., Vacík, T., Cmarko, D. & Raška, I. Discontinuous transcription. *Nucleus* **9**, 149–160 (2018).
100. Martinez-Rucobo, F. W., Sainsbury, S., Cheung, A. C. & Cramer, P. Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. *EMBO J.* **30**, 1302–1310 (2011).
101. Schulz, S. *et al.* TFE and Spt4/5 open and close the RNA polymerase clamp during the transcription cycle. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1816–E1825 (2016).
102. Kang, J. Y. *et al.* Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell* **173**, 1650-1662.e14 (2018).
103. Klein, B. J. *et al.* RNA polymerase and transcription elongation factor Spt4/5 complex structure. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 546–550 (2011).
104. Daniels, J.-P., Kelly, S., Wickstead, B. & Gull, K. Identification of a crenarchaeal orthologue of Elf1: implications for chromatin and transcription in Archaea. *Biol. Direct* **4**, 24 (2009).
105. Blombach, F., Fouqueau, T., Matelska, D., Smollett, K. & Werner, F. Promoter-proximal elongation regulates transcription in archaea. *Nat. Commun.* **2021 121 12**, 1–15 (2021).
106. Smollett, K., Blombach, F., Reichelt, R., Thomm, M. & Werner, F. A global analysis of transcription reveals two modes of Spt4/5 recruitment to archaeal RNA polymerase. *Nat. Microbiol.* **2**, 17021 (2017).
107. Lange, U. & Hausner, W. Transcriptional fidelity and proofreading in Archaea and implications for the mechanism of TFS-induced RNA cleavage. *Mol. Microbiol.* **52**, 1133–1143 (2004).
108. Orlova, M., Newlands, J., Das, A., Goldfarb, A. & Borukhov, S. Intrinsic transcript

- cleavage activity of RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 4596–600 (1995).
109. Deighan, P. & Hochschild, A. Conformational toggle triggers a modulator of RNA polymerase activity. *Trends in Biochemical Sciences* **31**, 424–426 (2006).
  110. Hausner, W., Lange, U. & Musfeldt, M. Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase. *J. Biol. Chem.* **275**, 12393–12399 (2000).
  111. Marr, M. T. & Roberts, J. W. Function of transcription cleavage factors GreA and GreB at a regulatory pause site. *Mol. Cell* **6**, 1275–1285 (2000).
  112. Knippa, K. & Peterson, D. O. Fidelity of RNA polymerase II transcription: Role of Rbp9 in error detection and proofreading. *Biochemistry* **52**, 7807–7817 (2013).
  113. Fouqueau, T. *et al.* The transcript cleavage factor paralogue TFS4 is a potent RNA polymerase inhibitor. *Nat. Commun.* **8**, 1–13 (2017).
  114. Mattioli, F. *et al.* Structure of histone-based chromatin in Archaea. *Science (80-. )*. **357**, 609–612 (2017).
  115. Brunk, C. F. & Martin, W. F. Archaeal Histone Contributions to the Origin of Eukaryotes. *Trends in Microbiology* **27**, 703–714 (2019).
  116. Watts, E. A. *et al.* Histones direct site-specific CRISPR spacer acquisition in model archaeon. *Nat. Microbiol.* (2023). doi:10.1038/s41564-023-01446-3
  117. Driessen, R. P. C. *et al.* Crenarchaeal chromatin proteins Cren7 and Sul7 compact DNA by inducing rigid bends. *Nucleic Acids Res.* **41**, 196–205 (2012).
  118. Takemata, N. & Bell, S. D. Emerging views of genome organization in Archaea. *Journal of Cell Science* **133**, (2020).
  119. Skibbens, R. V. Condensins and cohesins - one of these things is not like the other! *Journal of cell science* **132**, (2019).
  120. Zhang, Z. *et al.* Archaeal chromatin proteins cren7 and sul7d compact DNA by bending and bridging. *MBio* **11**, (2020).
  121. Takemata, N., Samson, R. Y. & Bell, S. D. Physical and Functional Compartmentalization of Archaeal Chromosomes. *Cell* **179**, 165-179.e18 (2019).
  122. Barillà, D. Driving Apart and Segregating Genomes in Archaea. *Trends Microbiol.* **24**, 957–967 (2016).
  123. Proudfoot, N. J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352**, aad9926 (2016).
  124. Santangelo, T. J. *et al.* Polarity in archaeal operon transcription in *Thermococcus kodakaraensis*. *J. Bacteriol.* **190**, 2244–8 (2008).
  125. You, L. *et al.* Structural basis for transcription antitermination at bacterial intrinsic terminator. *Nat. Commun.* **10**, 1–11 (2019).
  126. Marshall, C. J., Qayyum, M. Z., Walker, J. E., Murakami, K. S. & Santangelo, T. J. The structure and activities of the archaeal transcription termination factor Eta detail vulnerabilities of the transcription elongation complex. *Proc. Natl. Acad. Sci.* **119**, (2022).
  127. Li, J. *et al.* aCPSF1 cooperates with terminator U-tract to dictate archaeal transcription termination efficacy. *Elife* **10**, (2021).
  128. Le, T. T. *et al.* Mfd Dynamically Regulates Transcription via a Release and Catch-Up Mechanism. *Cell* **172**, 344-357.e15 (2018).
  129. Gehring, A. M. & Santangelo, T. J. Archaeal RNA polymerase arrests transcription at DNA lesions. *Transcription* **8**, 288–296 (2017).
  130. Park, J. S., Marr, M. T. & Roberts, J. W. E. coli transcription repair coupling factor (Mfd protein) rescues arrested complexes by promoting forward translocation. *Cell* **109**, 757–767 (2002).
  131. Khanh Phung, D. *et al.* RNA processing machineries in Archaea: the 5-3 exoribonuclease aRNase J of the-CASP family is engaged specifically with the helicase ASH-Ski2 and the 3-5 exoribonucleolytic RNA exosome machinery. *Nucleic Acids Res.* **48**, 3832–3847

- (2020).
132. Johnson, G. E., Lalanne, J. B., Peters, M. L. & Li, G. W. Functionally uncoupled transcription–translation in *Bacillus subtilis*. *Nature* **585**, 124–128 (2020).
  133. Mitra, P., Ghosh, G., Hafeezunnisa, M. & Sen, R. Rho Protein: Roles and Mechanisms. *Annu. Rev. Microbiol.* **71**, 687–709 (2017).
  134. Lo Gullo, G. *et al.* Optimization of an in Vitro Transcription/Translation System Based on *Sulfolobus solfataricus* Cell Lysate. *Archaea* **2019**, (2019).
  135. Phung, D. K. *et al.* Archaeal  $\beta$ -CASP ribonucleases of the aCPSF1 family are orthologs of the eukaryal CPSF-73 factor. *Nucleic Acids Res.* **41**, 1091–103 (2013).
  136. Zhang, C., Phillips, A. P. R., Wipfler, R. L., Olsen, G. J. & Whitaker, R. J. The essential genome of the crenarchaeal model *Sulfolobus islandicus*. *Nat. Commun.* **9**, 1–11 (2018).
  137. Silva, A. P. G. *et al.* Structure and activity of a novel archaeal  $\beta$ -CASP protein with N-terminal KH domains. *Structure* **19**, 622–32 (2011).
  138. Lawson, M. R. *et al.* Mechanism for the Regulated Control of Bacterial Transcription Termination by a Universal Adaptor Protein. *Mol. Cell* **71**, 911–922.e4 (2018).
  139. Mir-Montazeri, B. *et al.* Crystal structure of a dimeric archaeal Cleavage and Polyadenylation Specificity Factor. *J. Struct. Biol.* **173**, 191–195 (2011).
  140. Silva, A. P. G. *et al.* Structure and activity of a novel archaeal  $\beta$ -CASP protein with N-terminal KH domains. *Structure* **19**, 622–632 (2011).
  141. Werner, F. 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).
  142. Eaton, J. D., Francis, L., Davidson, L. & West, S. A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. *Genes Dev.* **34**, 132–145 (2020).
  143. Cortazar, M. A. *et al.* Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a “Sitting Duck Torpedo” Mechanism. *Mol. Cell* **76**, 896–908.e4 (2019).
  144. Laass, S. *et al.* Characterization of the transcriptome of *Haloferax volcanii*, grown under four different conditions, with mixed RNA-Seq. *PLoS One* **14**, e0215986 (2019).

## CHAPTER 2: ARCHAEOAL HISTONE-BASED CHROMATIN STRUCTURES REGULATE TRANSCRIPTION ELONGATION RATES

### Summary

Many archaea encode and express histone proteins at high abundance which wrap, organize, and compact their genomes. Archaeal and canonical eukaryotic histone proteins share a near-identical fold that permits DNA wrapping through select histone-DNA contacts to generate chromatin-structures that must be sequentially traversed by RNA polymerase to generate transcripts. As archaeal histones can spontaneously assemble with just a single histone isoform, single-histone archaeal chromatin variants provide an idealized platform to detail the impacts of distinct histone-DNA contacts and chromatin structures on transcription efficiencies and to detail the role of conserved elongation factors in assisting transcription through chromatin landscapes. We demonstrate that substitution of archaeal histone residues that alter three-dimensional chromatin structures or modify histone-DNA contacts result in chromatin structures that radically alter transcription elongation rates and pausing patterns. Chromatin-barriers slow and pause RNA polymerase elongation, providing regulatory potential, but do not induce substantial backtracking that can be rescued by cleavage of the nascent transcript. The modest impacts of the cleavage stimulatory factor TFS on elongation rates through chromatin landscapes *in vitro* is correlated with our demonstration of TFS-dispensability, without an obvious phenotype, from the model archaeon *Thermococcus kodakarensis*. Our results detail the importance of distinct chromatin structures for archaeal gene expression and provide a unique perspective on the evolution of, and regulatory strategies imposed by, eukaryotic chromatin.

## Introduction

Most Archaea and all Eukarya encode histone proteins that bind DNA to form dynamic chromatin landscapes that compact and organize the genome, thereby impacting transcription and gene expression<sup>1–7</sup>. A few bacterial clades also encode histone-fold containing proteins<sup>8</sup> that interact with DNA very differently from the histone-DNA interactions preserved in archaeal and eukaryotic systems. When sufficiently abundant, archaeal histones spontaneously oligomerize to generate extended archaeal histone-bound chromatin structures that organize the genome and regulate the progression of the transcription apparatus<sup>6,9</sup>. Archaeal RNA polymerase (RNAP) and eukaryotic RNA polymerase II (Pol II) are structurally and functionally homologous and both must overcome nearly identical histone-bound DNA barriers<sup>10–12</sup>. Chromatin architecture can provide regulatory potential during transcription elongation, alter positions of transcription pausing, and is known to impact elongation-termination decisions<sup>13,14</sup>. Much of what is understood about how histone-based chromatin modulates the transcription apparatus comes from studies targeting eukaryotic histone post-translational modifications (PTMs), epigenetic markers, and chromatin remodeling complexes, all of which influence gene expression<sup>15–17</sup>.

Archaea are the likely progenitors of Eukarya<sup>18–23</sup> and the core histone-fold and DNA binding activities of archaeal histones are shared with their eukaryotic counterparts<sup>4,5,24–28</sup>. Archaeal histones retain the canonical histone fold of three  $\alpha$  helices joined by two loops ( $\alpha$ 1-L1- $\alpha$ 2-L2- $\alpha$ 3). Archaeal histones can form both homo- and hetero-dimers that protect ~30 bp of DNA and assemble into an extended, continuous super-helical structure. The geometry of the DNA bound within an archaeal chromatin superhelix nearly exactly matches that of the eukaryotic nucleosomal DNA arrangement<sup>4,24,29–31</sup> and the overall archaeal histone-based extended chromatin structure closely matches chromatin structures found on eukaryotic telomeres<sup>32</sup>. Both archaeal and eukaryotic histone-DNA interactions align to the same

nucleosome positioning code and the specific protein-DNA contacts that stabilize chromatin are conserved<sup>29,33–35</sup>. Archaeal genomes, however, appear devoid of chromatin remodeling complexes. Additionally, archaeal histones typically lack N- and C-terminal extensions common to their eukaryotic counterparts, and PTMs of archaeal histones have not been demonstrated to be either abundant or biologically relevant<sup>36–38</sup>.

Although a single archaeal histone isoform is sufficient to spontaneously form extended chromatin structures *in vitro*, Archaea that encode histone proteins often encode and can differentially express multiple histone isoforms<sup>19,39</sup>. It is possible that chromatin assembled from different archaeal histone isoforms adopts unique structures that differentially impact transcription and genome organization. The model archaeal species *Thermococcus kodakarensis* encodes two histone proteins – HTkA and HTkB – that are individually non-essential; deletion of both HTkA and HTkB is synthetically lethal, indicating histone-based chromatin is essential for informational processing from, and replication of, DNA<sup>38</sup>. Strains lacking HTkA or HTkB can be modified at the sole remaining histone-encoding locus to generate strains with single-histone variant chromatin structures. Substitution of even a single histone residue can radically increase or decrease DNA affinity, disrupt the three-dimensional (3D) structure of archaeal chromatin, or disrupt dimer interactions with dramatic impacts to gene expression, growth rates, and overall fitness<sup>4,40,41</sup>. Outstanding questions remain regarding how the archaeal histone-based chromatin landscape impacts gene expression, what roles conserved transcription factors play in assisting the archaeal RNAP when transcribing histone-based chromatin, and how chromatin organization patterns impact the rate of RNA synthesis and pausing for archaeal RNAP.

Archaeal transcription systems are component simplified but homologous to their eukaryotic counterparts<sup>42–46</sup>. Recapitulation of the archaeal transcription system *in vitro* using histone-bound templates provides an ideal and complementary platform to delineate the regulation,

pausing, and elongation rates of archaeal transcription through varied chromatin landscapes. To detail the role specific individual residues within the archaeal histone-DNA complex have on the progression of the transcription apparatus, we describe *in vitro* RNAP processivity in histone-free, histone-bound, and variant histone-bound environments. Individual histone variants, once assembled into archaeal histone-based chromatin, can elicit dramatic changes in the rate of RNA synthesis and pausing patterns during elongation that resolve the roles of select histone-histone and histone-DNA interactions on transcription elongation.

Addition of the well-conserved elongation factor TFS mildly assists elongation rates and increases full-length RNA production, but TFS activity has very little effect on RNAP processivity in specific chromatin landscapes designed to disrupt the 3D structure of, or stabilize histone-DNA interactions within, archaeal histone-based chromatin. The minimal impacts of TFS *in vitro* adumbrated that TFS-activities may not be essential *in vivo*; the successful yet non-phenotypic deletion of TFS (TK0533) from *T. kodakarensis* suggests that backtracking and rescue of archaeal TECs via TFS-stimulated, RNAP active center mediated endonucleolytic cleavage of nascent transcripts is not a critical component of archaeal transcription regulatory mechanisms.

Our results reveal how changes to specific histone residues alter chromatin structures that regulate transcription elongation rates and pausing patterns. Based on our results, differential expression and assembly of archaeal histone isoforms could be employed as a regulatory mechanism to control gene expression and genome accessibility. Finally, the minimal impact resultant from deletion of TFS, one of only a few well-conserved archaeal factors known to influence post-initiation regulation of the archaeal RNAP, implies that transcription backtracking does not impart significant regulation *in vivo*. The congruence of archaeal and eukaryotic chromatin structures permits extrapolation of our results beyond archaeal systems to detail how the evolution of eukaryotic histone isoforms changed the chromatin landscape and likely led to the requirement for chromatin remodeling and histone modifications in Eukarya.



## Results

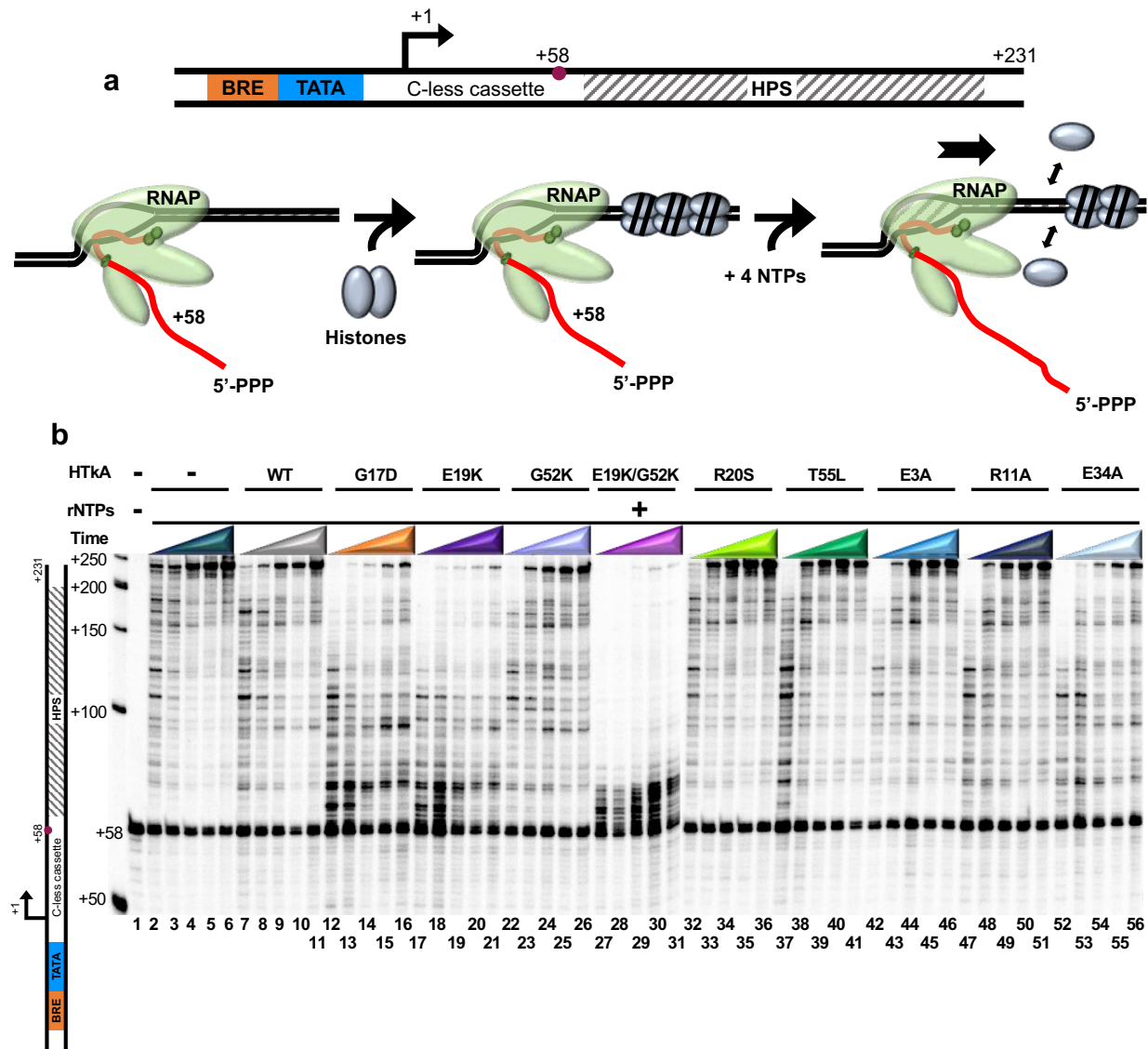
### **Select histone-DNA contacts dramatically alter transcription rates and pausing patterns.**

Translocation and RNA synthesis by archaeal RNAP is slowed by histone-bound DNA but ternary elongation complex (TEC: RNAP, a DNA template, and nascent RNA) stability is not disrupted by chromatin landscapes<sup>9,47</sup>. The ease of spontaneous assembly of chromatin landscapes with single archaeal histone-isoforms permits evaluation of how variant chromatin landscapes regulate transcription elongation rates and pausing patterns.

To examine the impacts of select histone variants and their associated chromatin landscapes on the archaeal transcription apparatus *in vitro*, we recombinantly expressed and purified the well-studied *T. kodakarensis* histone A protein (HTkA<sup>WT</sup>)<sup>4,7,19,41,48</sup> and select HTkA<sup>variants</sup> (Fig. A.1a). Histones with substitutions at residues known to significantly increase (HTkA<sup>E19K</sup>, HTkA<sup>G52K</sup>, and HTkA<sup>E19K/G52K</sup>) or decrease (HTkA<sup>R20S</sup> and HTkA<sup>T55L</sup>) affinity to DNA<sup>40,49</sup>, with a substitution known to interfere with the 3D structure of the archaeal superhelix (HTkA<sup>G17D</sup>)<sup>4,41</sup>, and with substitutions known to disrupt dimer interactions (HTkA<sup>E3A</sup>, HTkA<sup>R11A</sup>, and HTkA<sup>E34A</sup>)<sup>4,31</sup> were prepared to evaluate and compare the impact of specific histone variants on transcription elongation kinetics and pausing. HTkA normally functions at 85-95°C and preparations of recombinant HTkA often retained dimeric-interactions even after extensive heating and SDS-PAGE (Fig. A.1a). Western blotting with anti-HTkA antibodies confirm that the higher order complexes resolved in SDS-PAGE are oligomerized HTkA complexes (Fig. A.1b).

To determine how histone-variant chromatin landscapes impact TEC activities and pausing patterns, we exploited purified HTkA<sup>WT</sup> and HTkA<sup>variants</sup>, basal regulatory archaeal transcription components, archaeal RNAP, and our capacity to monitor the elongation patterns of TECs *in vitro* (Fig. 2.1). To ensure addition of archaeal histones and resultant chromatin structures did not impede transcription initiation, stalled TECs were first formed on histone-free DNA via initiation at a C-less cassette with only ATP, UTP, and GTP. Elongation limited by the absence

of CTP generated TECs with +58 nucleotide nascent transcripts (TECs<sub>+58</sub>) that were captured and washed to remove excess rNTPs, including radiolabeled UTP, thereby ensuring that the specific activity of all transcripts >+58 nts were identical. Templates containing TECs<sub>+58</sub> were then saturated with HTkA<sup>WT</sup> or HTkA<sup>variant</sup> proteins to form a chromatin landscape that TECs must traverse to extend nascent transcripts upon elongation restart<sup>6,9</sup> (Fig. 2.1). DNA templates included a tandem, 60 base pair (bp) SELEX-derived histone positioning sequence (HPS)<sup>49,50</sup>, optimized to bind histones downstream the stalled TECs<sub>+58</sub> (Fig. A.2).



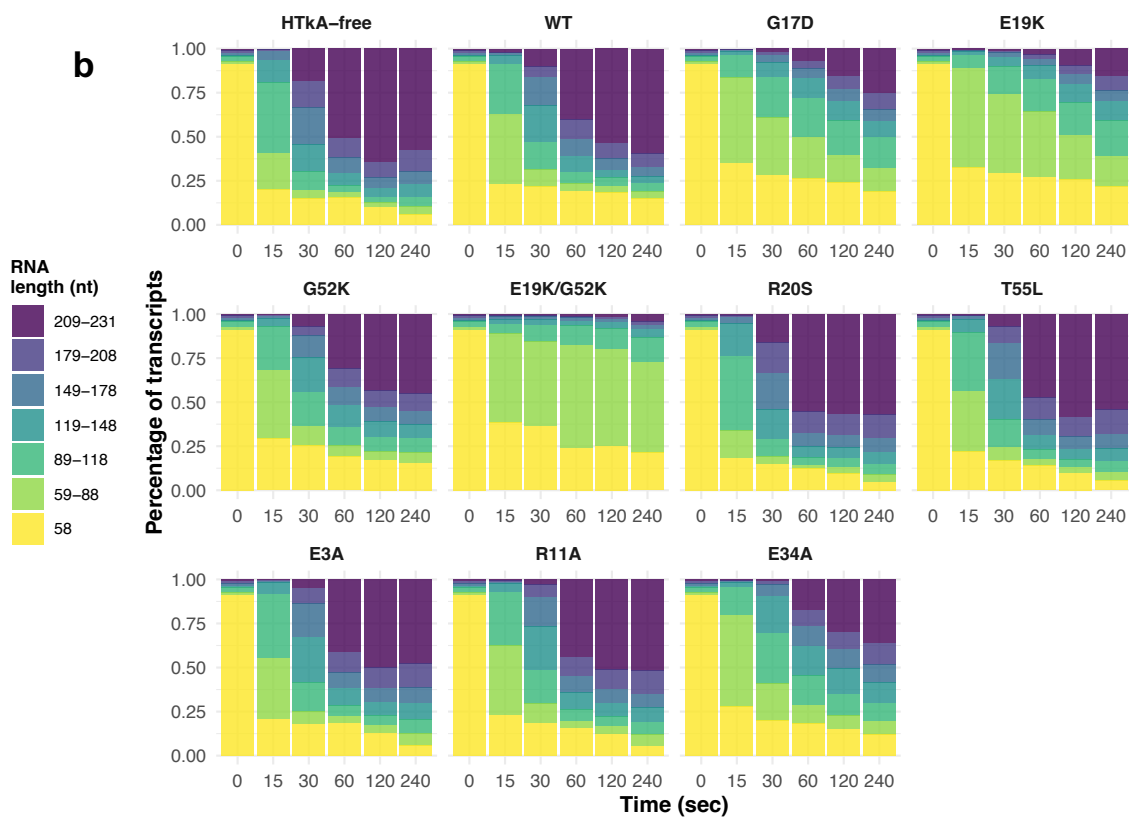
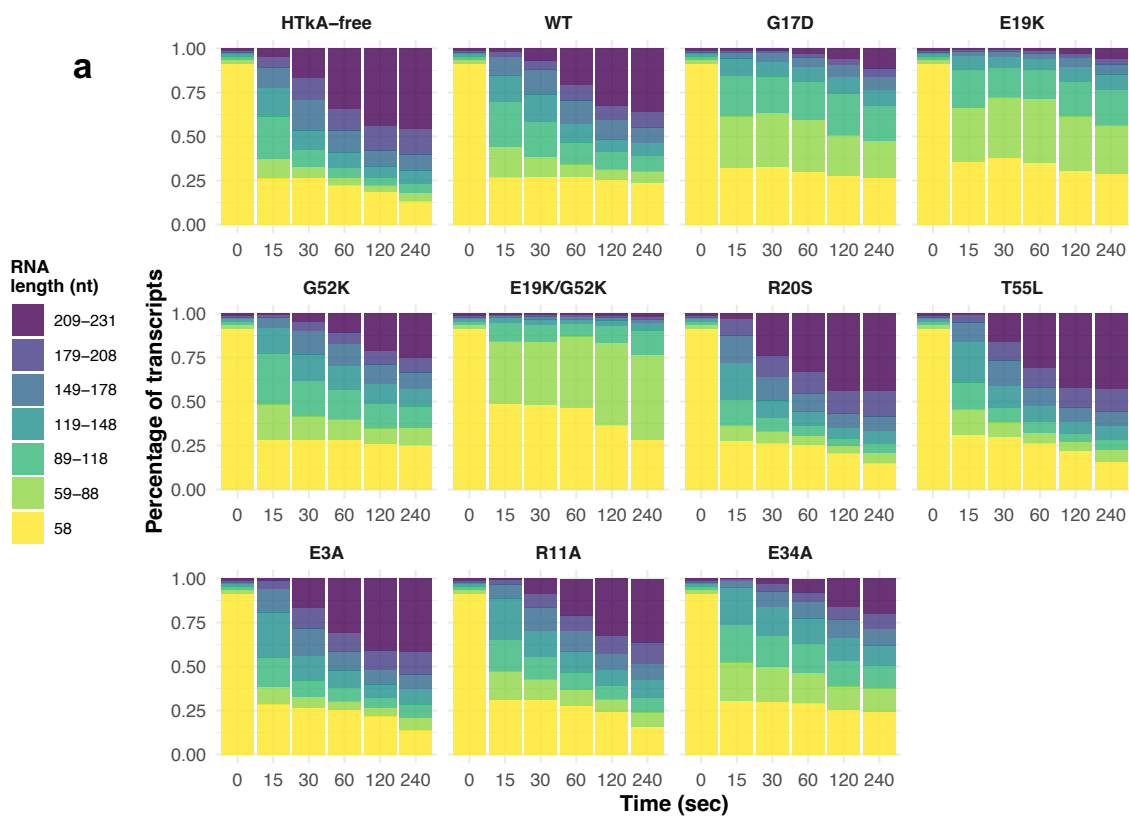
**Figure 2.1. Variant archaeal histone-based chromatin landscapes dramatically alter the rate of RNA synthesis and pausing patterns of transcription elongation complexes.**

**a.** RNA synthesis to +231 was monitored following elongation restart of TECs<sub>+58</sub> on linear templates that were bound by archaeal histones after TEC<sub>+58</sub> formation. The template contained the *hmtB* promoter, a +58 nt C-less cassette, and tandem 60-bp histone-positioning sequence elements (HPS). **b.** Continued RNA synthesis from TECs<sub>+58</sub> (lane 1) was monitored by revealing changes in nascent transcript length in reaction aliquots removed after 15-, 30-, 60-, 120-, and 240-seconds following transcription restart upon rNTP addition. Elongation was permitted along DNA lacking any bound proteins (HTkA-free), or histone-bound templates formed with HTkA<sup>WT</sup> or HTkA<sup>variants</sup>, n = 4. Radiolabeled ssDNA makers provide size standards.

Upon rNTP addition, elongation rapidly restarts but quickly becomes asynchronous (Fig 2.1b).

Monitoring the changes in nascent RNA length over time permits evaluation of the ensemble average activities of archaeal TECs as they navigate both histone-free and histone-bound DNA (Figs. 2.2 & 2.3 and Figs. A.3 & A.4). By binning the percentage of transcripts according to length (seven-bins; Fig. 2.2) and monitoring changes to transcript distribution over time, the totality of TEC elongation kinetics could be visualized (Fig. 2.2 and Fig. A.3). Elongation on histone-free templates (Fig. 2.1b, lanes 2-7) reveals several short-lived sequence-specific pause positions during elongation from +58 to +231. Full-length transcripts are evident after even just 15 seconds and continue to accumulate as more TECs reach the end of the linear template. No individual template position results in a majority of the TECs pausing for any significant timeframe, and the accumulation of full-length transcripts plateaus within two minutes following elongation restart (Figs. 2.2a and A.3a).

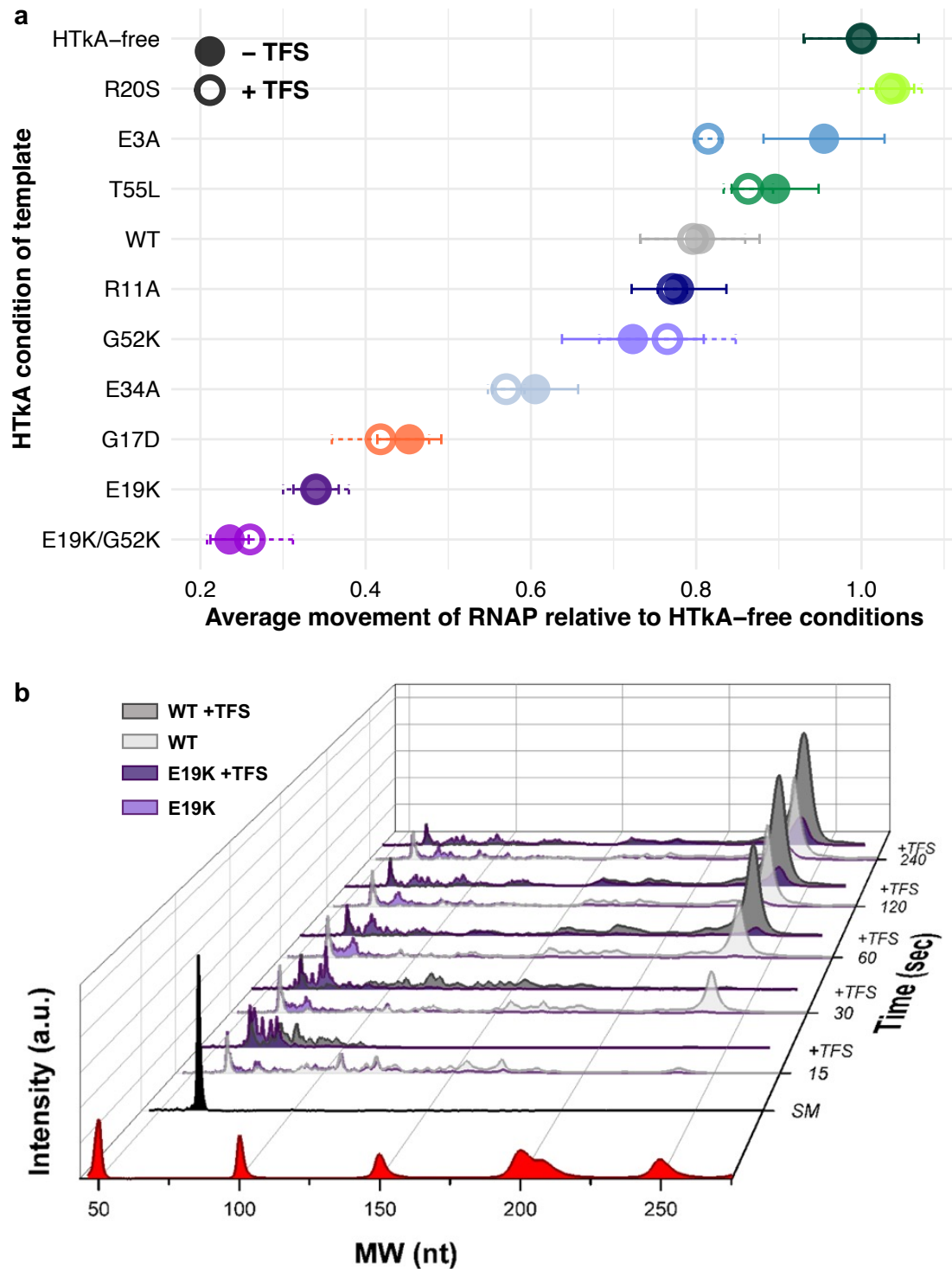
To determine the average RNAP elongation rate on histone-free, histone-bound, and variant histone-bound chromatin landscapes in nucleotides/second (nt/s) we calculated the sum of the average density of RNAs in each bin over time. The mean (nt/s) across all timepoints defines how quickly RNAP can traverse the template in each chromatin landscape relative to HTkA-free conditions (Fig. 2.3a). The comparison of the average TEC progression between histone-free and histone-bound templates provides important insights into the impacts of distinct archaeal histone-based chromatin structures on elongation rates and pausing patterns (Fig. 2.3).



**Figure 2.2. RNA synthesis is slowed by different archaeal histone-based chromatin landscapes.** **a.** Stacked bar plots quantify the changing percentage of differing transcript lengths (y-axis), divided into seven distinct bins as detailed in the key (left) over time (x-axis), during elongation on non-, WT-, and variant- chromatinized templates. **b.** The addition of TFS *in vitro* modestly impacts the percentage of RNA products within each bin.

Addition of HTkA<sup>WT</sup> permits formation of chromatin structures on the DNA template downstream of the stalled TECs<sub>+58</sub>. Upon rNTP restart, a native HTkA-based chromatin landscape results in only modest changes to the pausing patterns and progression of TECs (Fig. 2.1b, lanes 7-11 and Figs. 2.2 & 2.3). Pause positions noted on histone-free DNA are mildly accentuated, and a few additional, albeit short-lived, new pause positions emerge that collectively reduce the ensemble rate of transcription elongation by ~20% (Fig. 2.3 & Fig. A.4). Formation of chromatin structures downstream of TECs<sub>+58</sub> has minimal impacts on the percentage of TECs that restart elongation upon rNTP addition. Therefore, isolated archaeal RNAP, without the aid of transcription factors or remodeling complexes, is proficient at elongation on histone-bound DNA *in vitro*. The minor impacts to elongation rates due to downstream histone barriers likely helps explain the absence of any known chromatin remodeling complexes within archaeal genomes, as archaeal histone-based chromatin (in WT form) does not dramatically impact TEC translocation and RNA synthesis.

While histone-based chromatin formed with HTkA<sup>WT</sup> has modest impacts on elongation kinetics and pausing, changing select residues that impact histone-histone or histone-DNA interactions, or that alter the 3D structure of archaeal histone-based chromatin can elicit significant changes to elongation rates and pausing positions (Figs. 2.1b, 2.2, & 2.3 and Fig. A.4). For example, T55 is positioned appropriately to make a salt bridge with R20 from an adjacent monomer (Fig. 2.4a). Introduction of histone variants R20S (Fig. 2.1b, lanes 32-36) and T55L (Fig. 2.1b, lanes 37-41) decrease DNA affinity and result in minimal elongation conflicts; elongation rates on chromatin formed with HTkA<sup>R20S</sup> or HTkA<sup>T55L</sup> are decreased just ~5-10% from histone-free conditions and thus have less impact on RNAP progression than observed with HTkA<sup>WT</sup>-derived



**Figure 2.3. Substitution of key archaeal histone residues have pronounced effects on RNAP elongation rates.** Relative RNAP elongation rates in the absence (a) and presence (b) of TFS are plotted with respect to the rate of synthesis (nt/s) on templates lacking histone proteins. Closed circles and open circles, respectively, detail the average elongation rates in the absence and presence of TFS addition. Error bars represent the SE from -TFS,  $n = 4$  and +TFS,  $n = 3$  experiments.

chromatin. In stark contrast, introduction of histone variants E19K (Fig. 2.1b, lanes 17-21), G52K (Fig. 2.1b, lanes 22-26), or a variant with both E19K/G52K (Fig. 2.1b, lanes 27-31) that significantly increase DNA affinity result in formation of a chromatin landscape that slows (each single mutant) and nearly impedes (double mutant) transcription elongation upon rNTP restart.

The increased interactions between the phosphate backbone of DNA and the positively charged lysine residue(s) impair TEC movement significantly, reducing elongation rates by ~25% (HTkA<sup>G52K</sup>), ~65% (HTkA<sup>E19K</sup>), and ~75% (HTkA<sup>E19K/G52K</sup>) (Fig. 3A), nearly eliminating full-length transcript production over the initial time course in the HTkA<sup>E19K/G52K</sup>-based environment (Figs. 2.2a and Fig. A.4e). Extending the time course of transcription elongation on histone-free templates or through a HTkA<sup>WT</sup>-based chromatin landscape (Fig. A.5a, lanes 2-8 and lanes 9-15, respectively) demonstrates that essentially all TECs<sub>+58</sub> eventually restart elongation upon rNTP addition and that no prominent pause positions dominate elongation rates (Fig. A.5). While HTkA<sup>E19K/G52K</sup>-based chromatin landscapes do result in a near complete capture of TECs in an extended pause at ~+70 nts, essentially all TECs are eventually capable of independently clearing this pause and elongating towards the end of the template (Figs. A.5a, lanes 16-22; A.5b; and A.5c).

For HTkA<sup>E19K</sup> and HTkA<sup>E19K/G52K</sup>-based chromatin, TECs encounter a long-lived pause at ~+70 nts, corresponding to the position at which the leading edge of the TEC is likely to collide with a well-positioned and tightly bound histone-dimer at the HPS. As the footprint of archaeal RNAP is ~20 bp<sup>51</sup>, forward translocation following elongation restart from +58 on our DNA template puts the 3' end of the nascent RNA in the active site of RNAP at ~70 nts if the TEC is subject to pause near the beginning of the HPS (Fig. A.2). Release from the ~+70 nts pause is rate limiting for full-length transcript production (Fig. A.6). Given that histone dimers bind and protect just 30 bp, a nearly identical barrier should be encountered again as TECs reach ~100, 130, 160, and 190 nts, but only the first barrier represents a significant pause position. The first histone-based

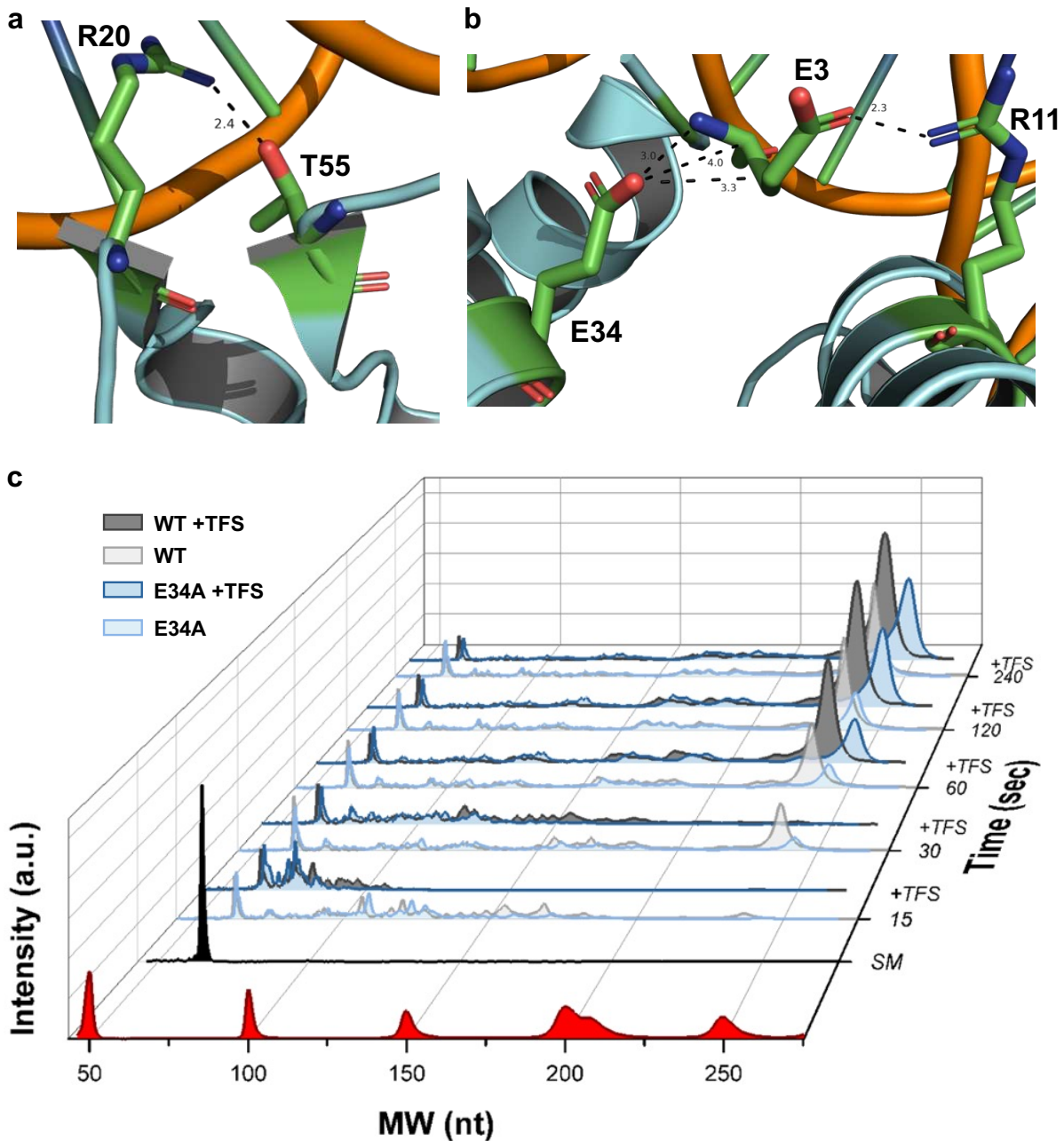
chromatin barrier to continued elongation is thus the prominent position of regulation for continued transcription. It is possible that disrupting the chromatin barrier from the most TEC proximal histone-dimer results in changes to the extended chromatin structure that clear the template of major barriers to continued elongation.

The prominent pause at ~+70 nts on templates containing HTkA<sup>E19K/G52K</sup>-based chromatin implies that TECs must wait for downstream histone-dimers to spontaneously release from the template to permit continued elongation. Stable HTkA<sup>E19K/G52K</sup>-histone-DNA complexes are confirmed by monitoring elongation for extended times (Fig. A.7). After 2 minutes of elongation, most TECs are still paused at ~+70 nts on HTkA<sup>E19K/G52K</sup>-histone bound templates (Fig. A.7a, lanes 2-5). Dilution of the reactions to reduce total histone concentrations (Fig. A.7a, lanes 6-8), or addition of HTkA<sup>WT</sup> to promote exchange of DNA-bound histones (Fig. A.7a, lanes 9-11) was not successful in altering elongation rates in comparison to maintaining the identical landscape through addition of HTkA<sup>E19K/G52K</sup> (Fig. A.7a, lanes 12-14).

As observed for histone-variants that increase DNA affinity, introduction of a histone variant that disrupts the 3D structure of chromatin (HTkA<sup>G17D</sup>; Fig. 2.1b, lanes 12-16) results in impaired TEC progression, substantial pausing, and reduced RNA synthesis rates. Given that G17 does not directly contact DNA, it is perhaps surprising that the elongation kinetics do not match HTkA<sup>WT</sup>; HTkA<sup>G17D</sup>-based chromatin reduces elongation rates by ~55% (Figs. 2.2 & 2.3). A prominent pause site at ~+70 nts demonstrates that TECs still encounter the most TEC-proximal dimer as a major barrier of HTkA<sup>G17D</sup>-based chromatin that is rate limiting for production of full-length transcripts (Fig. A.6). It will thus be critical to evaluate systems wherein histone-isoforms that may impair continued polymer formation are introduced into heterologous archaeal histone-based chromatin structures. Although neither HTkA nor HTkB is predicted to impair polymerization, some histone-isoforms in less genetically tractable systems lacking *in*



*vitro* transcription systems may provide a mechanism to cap the growth of extended histone polymers<sup>52</sup>.



**Figure 2.4. Histone variants with decreased DNA affinity and altered dimer interactions form chromatin structures that minimally impact transcription elongation.** **a.** T55 and R20 from an adjacent monomer form an intermolecular salt bridge. Disrupting this salt bridge decreases histone-DNA affinity, permitting increased RNA synthesis rates. **b.** Histone dimer interactions are facilitated by E34 interactions with R11 and E3. **c.** Waterfall plots demonstrate that chromatin resultant from HTkA<sup>E34A</sup> modestly (~25%) decreases elongation rates compared to HTkA<sup>WT</sup>.

Archaeal histone dimers can themselves dimerize to form tetramers, and polymerization of histone dimers can continue, in theory, indefinitely to form very long extended chromatin structures. The dimer-dimer interface is, in part, coordinated by E3, R11, and E34 (Fig. 2.4b). E34 is situated within the histone-DNA complex to position R11 favorably with E3, allowing a salt bridge to stabilize histone-dimerization<sup>1,4,5,53</sup> (Fig 2.4b). Chromatin generated from HTkA<sup>E3A</sup> (Fig. 2.1b, lanes 42-46) only mildly reduces transcription elongation rates when compared to histone-free conditions and permits faster elongation than seen with HTkA<sup>WT</sup>-chromatin (Figs. 2.2 & 2.3). While HTkA<sup>E3A</sup>-based chromatin does not result in any new pausing patterns compared to HTkA<sup>WT</sup>, chromatin landscapes formed by HTkA<sup>R11A</sup>- or HTkA<sup>E34A</sup>-histone variants do mildly and significantly hinder RNA synthesis, respectively, with pausing patterns at similar positions as those observed with HTkA<sup>WT</sup>, but each with increased duration (Figs. 2.1b & 2.4c). The pausing is more significant for HTkA<sup>E34A</sup>-based chromatin landscapes that reduce elongation rates to just half that of histone-free conditions (Figs. 2.2 & 2.3). Although HTkA<sup>R11A</sup>- and HTkA<sup>E34A</sup>-based chromatin structures increase pausing and decrease elongation rates, neither displays the prominent ~+70 nt pause (Fig. A.6) that controls overall synthesis rates for HTkA<sup>E19K</sup>-, HTkA<sup>E19K/G52K</sup>-, and HTkA<sup>G17D</sup>-based chromatin landscapes.

***Transcription Factor S (TFS) increases productive elongation through chromatin.***

Nearly all Archaea, including those species that do not encode histone proteins, encode Spt4, Spt5, and Transcription Factor S (TFS)<sup>12</sup>. Spt4/5 and TFS are known transcription factors that directly bind the archaeal RNAP and can accelerate transcription through chromatin landscapes<sup>6</sup>. Near-universal retention of each factor in archaeal genomes implies their importance to transcription regulation and fidelity *in vivo*. TFS is known to stimulate the intrinsic endonucleolytic cleavage activities of archaeal RNAP, providing a mechanism to escape pauses that result in retrograde movement (i.e., backtracking) of archaeal TECs that encounter barriers to continued forward translocation<sup>6,54,55</sup>. Collisions between the archaeal TEC and chromatin-

barriers are expected to result in some backtracking; however, TFS addition to *in vitro* transcription reactions has little effect on the average RNAP elongation rate from TECs<sub>+58</sub> that resume elongation upon rNTP addition (Figs. 2.2 & 2.3; Figs. A.6, A.7 & A.8; Table 2.1). While the addition of TFS has negligible impacts on the average TEC elongation rate through HTkA<sup>WT</sup>-chromatin landscapes *in vitro*, TFS addition does result in an increase of full-length transcripts by ~10%, largely due to release of a substantial percentage of +58 complexes into active elongation (Figs. 2.2 & 2.3; Fig. A.10; Table 2.1).

TFS addition to TECs elongating through chromatin landscapes formed with histone-variants produced mixed results (Figs. 2.2 & 2.3; Figs. A.8 & A.9). In every case, addition of TFS has a slight prohibitory effect on TEC progression in early timepoints (Figs. A.8 & A.9), but release from +58 was increased (Fig. A.10; Table 2.1). During elongation through HTkA<sup>WT</sup>- and HTkA<sup>variant</sup>-based chromatin landscapes that only modestly impact elongation rates and do not result in a significant percentage of TECs pausing at ~+70 nts, TFS addition has limited influence on the average RNA synthesis rates but results in an increase of full-length RNA products compared to the reactions without TFS (Figs. 2.2 & 2.3; Figs. A.8 & A.9). Akin to the previous chromatin landscapes, TFS addition to transcription reactions traversing the HTkA<sup>E19K</sup>-, HTkA<sup>G17D</sup>-, or HTkA<sup>E19K/G52K</sup>-based chromatin landscapes had minimal impacts on the average RNA synthesis rate (Figs. 2.2 & 2.3; Figs. A.8 & A.9; Supplementary Data, Appendix A).

TFS addition was impactful at reducing the +70 nts prominent pause on templates with HTkA<sup>E19K</sup>- or HTkA<sup>G17D</sup>-based chromatin structures (Fig. A.6), implying that cleavage-stimulatory activities of TFS rescued some backtracked TECs and permitted more rapid release from the +70 nts prominent pause. However, TFS-mediated activity was not sufficient to reduce the prominent and rate-limiting pause at ~+70 nts of the HTkA<sup>E19K/G52K</sup>-based chromatin structure (Fig. A.6).

**Table 2.1.** Residue substitutions alter TECs<sub>+58</sub> half-life.

<i>Half-life at +58 relative to HTkA<sup>WT</sup> (-TFS)</i>		
Condition of template	<b>t<sub>1/2</sub> (sec)</b>	
	<b>-TFS (n = 4)</b>	<b>+TFS (n = 3)</b>
<i>HTkA-free</i>	0.28	0.19
<b>WT</b>	<b>1.00</b>	<b>0.54</b>
<i>G17D</i>	0.65	0.40
<i>E19K</i>	0.84	0.59
<i>G52K</i>	0.77	0.38
<i>E19K/G52K</i>	0.44	0.41
<i>R20S</i>	0.30	0.18
<i>T55L</i>	0.28	0.17
<i>E3A</i>	0.27	0.19
<i>R11A</i>	0.27	0.17
<i>E34A</i>	0.56	0.29

***TFS (TK0533) is dispensable with negligible impacts on overall fitness.***

The impact of TFS on the percentage of full-length products through histone-based chromatin landscapes are often positive, but the prominent pauses that delay continued elongation on some templates do not behave as would be expected from TECs backtracking due to collisions with chromatin barriers. The absence of TFS from some archaeal genomes, coupled with the minor impacts on transcription rates *in vitro*, led to attempts to generate a strain of *T. kodakarensis* wherein TK0533 (encoding TFS) was deleted. While previous attempts suggested TFS may be an essential protein, continued genetic efforts were successful in deleting TK0533, resulting in strain RLV2 (Fig. A.11). Markerless deletion<sup>56</sup> of the full sequences encoding TFS were first confirmed through diagnostic PCRs using DNA purified from strain RLV2. The exact endpoints of the TK0533 deletion, and the absence of any second site mutations throughout the entire 2.08 Mbp genomes, were confirmed via whole genome sequencing (WGS) with over 100X coverage (Fig. A.11a). Despite the positive effects of TFS on backtracked TECs traversing some chromatin landscapes *in vitro*, the absence of TFS does not

result in any notable growth defects (Fig. A.11b) implying that the activities of the archaeal RNAP, alone or in combination with Spt4/5, suffice to permit normal transcription rates and regulated gene expression in the native chromatin landscape of *T. kodakarensis* in optimal conditions.

### ***Modeling the impacts of histone variants on archaeal chromatin structure.***

HTkA<sup>WT</sup>, like most canonical archaeal histones, is just 67 amino acids long and thus substitution of even a single residue can significantly alter the charge and DNA interactions of the histone dimer (the minimal protein unit capable of stable DNA binding). The HTkA<sup>E19K</sup> and HTkA<sup>G52K</sup> variants have obvious impacts on transcription elongation rates and pausing patterns that are easily explained by additional hydrogen bonding between positively charged surface residues of the histone dimer with the phosphate backbone of the bound and wrapped DNA. The rationale for the impacts of other HTkA<sup>variants</sup> on elongation rates and pausing patterns necessitates that we model the impacts of amino acid substitutions on the totality of histone-histone and histone-DNA interactions within the extended archaeal histone-based chromatin superstructure. We calculated the impacts of select amino acid substitutions within the extended histone-based chromatin structure composed of three histone dimers in complex with ~90 bps of DNA (PDB:5T5K)<sup>4</sup> in PyRosetta-4<sup>57,58</sup> (Fig. A.12 and Supplementary Data, Appendix A).

The HTkA<sup>T55L</sup> and HTkA<sup>R20S</sup> variants were known to reduce DNA affinity<sup>1,40</sup>. An important interaction between these residues - from one monomer to the next within the histone dimer, known as the R-T pair - was first predicted, then demonstrated through structural studies. Interactions of T55 and R20 position R20 for stabilizing interactions with several nucleotides of bound DNA (Fig. A.12a; green nucleotides). Substitution of R20 with serine eliminates interactions with four nucleotides of the bound DNA (Fig. A.12b; grey nucleotides). The increased bulk and hydrophobicity of the T55L substitution disrupts the important salt bridge with R20<sup>40</sup>, compromising the alignment of R20 for idealized DNA interactions and significantly

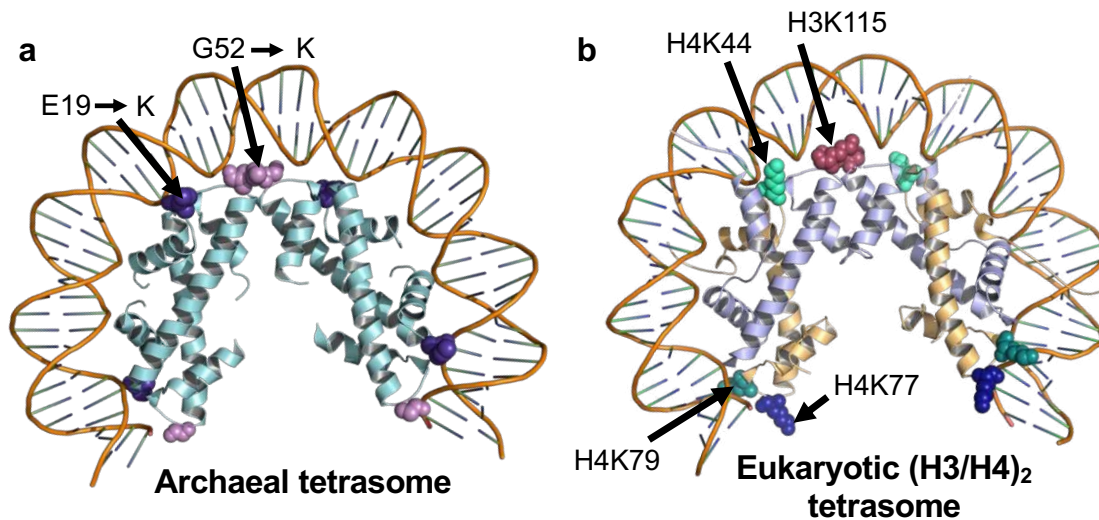
increases the Lennard-Jones repulsive term (in kcal/mol)<sup>58,59</sup> between T55L, R20, and a neighboring nucleotide (Fig. A.12c; red nucleotide).

Elongation rate reduction on templates bound by HTkA<sup>G17D</sup> was predicted to result, at least in part, from the impact of a bulky substitution resulting in the separation of adjacent gyres of the superhelix<sup>4,41</sup>. Introduction of an aspartic acid at position 17 results in steric hindrance with the preceding residue (A16) that is a part of the conserved AGA motif necessary for the tightly packed L1-L1 interface<sup>4,5</sup>. De-compacting archaeal-histone based chromatin might be predicted to facilitate transcription elongation, not hinder such, and modeling revealed additional impactful changes to histone-based chromatin that likely explain the significant challenges HTkA<sup>G17D</sup>-based chromatin presents to TECs. Accommodating the additional bulk of an aspartic acid side chain results in clashes with the prior residue (A16), increasing the Lennard-Jones repulsive term (kcal/mol)<sup>58,59</sup> between D17 and A16 (Figs. A.12d and A.12e). Substitution of G17D not only results in clashes with its direct neighbor but significantly increases the Lennard-Jones repulsive term within the L1-L1 pocket between a native aspartic acid at position 15 (Figs. A.12d and A.12e) and several other residues within this pocket (Supplementary Data, Appendix A). Additionally, the backbone beta carbon of G17D makes a hydrogen bond with the oxygen in the carboxyl backbone group of D15 (Fig. A.12f). The impacts to the local environment between D17 and A16, in combination with the increase in repulsive energy and newly acquired hydrogen bonds within the L1-L1 interface, likely alter not only the entire dimer structure but the stacking interactions of separate gyres of DNA, which would present a significant barrier to transcription elongation (Figs. A.12d, A.12e, and A.12f).

***Substitutions to specific histone residues in an archaeal-eukaryotic histone ancestor likely assisted the evolution of chromatin-remodeling systems and extensive PTMs in eukaryotes.***

Increasing evidence supports eukaryogenesis as an evolved symbiosis between a bacterium (future mitochondria) and an archaeon (future nucleus) with substantial horizontal gene transfer<sup>18–23</sup>. While the engulfing archaeal cell provided information processing mechanisms and proteins – including the core histone fold – the bacterial cell contributed energy and lipids<sup>18,19,21</sup>. Given an expanding genome, gene transfers, and gene duplication events, it is highly probable that variations to key histone residues provided a pathway to the complex regulatory mechanisms we see in all eukaryotes today. Previous studies have found that there are several residues within HMfB that are structural homologues to residues within H3 and H4 essential for tetramerization in both the HMfB tetramer and the (H3/H4)<sub>2</sub> tetramer<sup>60</sup>. Additionally, natural archaeal histone variants with an extended  $\alpha$ 1-L1 region that consist of four additional residues in the C-terminus encode a lysine at the exact position of H3K79 (which is also an inserted sequence) that is a target for PTMs<sup>1</sup>.

Therefore, it is reasonable that once E19K or G52K histone variants emerged, transcription elongation rates would have been significantly compromised. The impact of lysine residue substitutions to archaeal histones is nearly identical to the impacts to transcription when DNA is bound by the unmodified eukaryotic (H3/H4)<sub>2</sub> tetramer *in vitro*, where elongation was essentially blocked, even upon the addition of elongation factors (TFIIF and TFIIIS)<sup>61</sup>. The reduction in transcription rates would likely favor evolution of systems to permit PTMs to selective residues of eukaryotic histone proteins to counter the impacts of increased histone-DNA complex stability due to favorable interactions between phosphates and positively charged lysine residues. The positions of HTkA<sup>E19K</sup> and HTkA<sup>G52K</sup> in archaeal chromatin nearly identically match the positions of H4K79, H4K77, H4K44, and H3K115 in eukaryotic chromatin that are common targets of PTMs to change local chromatin structures in eukaryotes: H3K115ac facilitates nucleosome repositioning<sup>62</sup>, H4K44ac favors open chromatin configurations<sup>63</sup>, and H4K77ac and H4K79ac facilitate DNA unwrapping and transcription factor binding<sup>64</sup> (Fig 2.5).



**Figure 2.5. Archaeal- and eukaryotic-histone tetrasomes coordinate wrapped DNA nearly identically.** **a.** Key residue substitutions that increase DNA affinity within the archaeal tetrasome closely match the positions of well-defined residues within the eukaryotic (H3/H4)<sub>2</sub> tetrasome (**b**) that can be post-translationally modified to regulate gene expression.

## Discussion

The structure of archaeal histone-based chromatin plays a critical role in cellular viability, gene expression, and *in vitro* transcription activities. Our *in vitro* results detail the effects that specific histone residue substitutions have on the processivity of TECs in a chromatin environment.

Significant changes in RNA synthesis rates resultant from altering the chromatin landscape with just a single amino acid substitution in single histone-isoform archaeal chromatin are profound and demonstrate that archaeal-histone based chromatin structures are a major regulatory force for gene expression in *T. kodakarensis*. HTkA<sup>WT</sup>-based chromatin is a modest impediment to transcription elongation, reducing transcription rates ~20%. In comparison, HTkB-based chromatin inhibited TEC production by ~80% *in vitro* and TFS addition increased elongation rates ~4-fold<sup>6</sup>. While these two proteins have ~85% homology, there are 11 amino acids differing between the two (none of which were studied here), resulting in varied isoelectric points that can have dramatic effects on protein-DNA interactions<sup>40,65</sup>.



Chromatin assembled from HTkA<sup>variants</sup> with known reductions in DNA affinity does not present a significant barrier to transcription elongation. On HTkA<sup>WT</sup> or reduced-histone affinity chromatin, pausing of transcription is sporadic, short-lived, and unlikely to provide considerable regulatory potential. In contrast, substitutions that impact histone-histone and histone-DNA interactions to alter or strengthen histone-DNA contacts significantly reduce elongation rates and generate long-lived pause sites that dramatically impact elongation kinetics. As transcription and translation are coupled in archaea<sup>66</sup>, changes in chromatin structure that impact elongation rates are likely to modify translation rates and influence the control afforded by transcription termination mechanisms dependent on access to the TEC via the nascent transcript<sup>67–69</sup>.

The archaeal RNAP can independently traverse all the single-histone isoform chromatin landscapes we generated, albeit at different rates. Archaeal histone-based chromatin structures elicit a series of transcription pauses, particularly when the TEC first encounters DNA-bound by histones, that provide regulatory potential. These pauses are largely not resultant from backtracking of RNAP upon collisions with downstream protein barriers as might be expected. The minor impacts upon addition of TFS, the well-conserved elongation factor that stimulates endonucleolytic cleavage of the nascent transcript by RNAP, *in vitro* are now matched by evidence that TFS is not required for viability of *T. kodakarensis*. Given negligible phenotypic effects due to deletion of TFS, and failure to identify any additional archaeal-encoded factors that rescue backtracked TECs through stimulating cleavage of the nascent transcript, backtracking of TECs *in vivo* due to chromatin is not anticipated to be a significant regulatory property of archaeal-histone based chromatin in optimal conditions. It is likely that Spt4/5 is sufficient to accelerate TECs through chromatin-barriers *in vivo*. However, future evaluation of strains lacking TFS under stress conditions are imperative in understanding the broader implications of TFS not only in an altered chromatin environment, but the potential novel

regulatory effects TFS may have at specific sequence elements within the *T. kodakarensis* genome<sup>70,71</sup>.

Many of the HTkA<sup>variants</sup> investigated here *in vitro* are known to have profound effects on *in vivo* gene expression<sup>4,5,40,41</sup>. All attempts to introduce HTkA<sup>variants</sup> that significantly decrease DNA-binding (HTkA<sup>T55L</sup> and HTkA<sup>R20S</sup>) into *T. kodakarensis* strains that lack HTkB were unsuccessful<sup>4</sup>, implying that chromatinization of the genome from at least one histone protein is required for viability. In contrast, single, histone-isoform encoding strains of *T. kodakarensis* with HTkA<sup>E19K</sup>, HTkA<sup>G52K</sup>, or HTkA<sup>E19K/G52K</sup> were viable but displayed diminished growth and fitness. Compromising the extended, 3D super-helical structure of archaeal chromatin results in fitness challenges and substantial changes to the steady-state transcriptome in *T. kodakarensis*<sup>41</sup>. The significant impact of HTkA<sup>G17D</sup>-based chromatin on pausing and elongation rates implies that the normally tightly compacted, extended archaeal histone-based chromatin structure facilitates elongation, contrary to histone-based chromatin structures that cannot form extended polymers<sup>4,52</sup>.

The impact of 3D structure for elongation rates and gene regulation is likely to have impacts beyond archaeal systems. Eukaryotic telomeric chromatin forms a columnar structure<sup>32</sup> much akin to the extended, super-helical structure of archaeal chromatin<sup>7</sup>, suggesting that eukaryotic chromatin may retain additional features that match the primordial archaeal chromatin systems and regulatory strategies. Cryo-EM of archaeal-histones in complex with DNA often reveal the canonical super-helical extended structure, but also reveal a minor subpopulation of complexes with a 90° bend, altering the stacking of individual gyres of DNA into a lid-like structure<sup>7</sup>. The ability of archaeal chromatin to dynamically breathe into open and closed conformations, like that of telomeric chromatin, could be a key feature in regulatory mechanisms and gene expression.

In contrast to *T. kodakarensis* and several other archaeal clades, some halophilic archaea express histones at moderate levels, limiting their ability to act as general nucleoid associated proteins and subsequently more like site-specific transcription factors, and archaeal chromatin is unlikely histone-based<sup>19,39,69</sup>. In the many archaeal clades wherein histones are present in sufficient quantities to bind and wrap much or all of the genome(s)<sup>6</sup>, and thus more closely mimic eukaryotic genomes, expression of different histone isoforms could significantly alter transcription processes and genomic architectures. Many archaea encode for more than one histone protein, each with predicted and known differences in DNA binding capacity, tetramer formation, and stability<sup>19,52,72</sup>. Histone exchanges are important in eukaryotes and provide crucial regulatory mechanisms at specific stages of development (e.g., exchange of H3.3 and H3.1 that differ in just 5 amino acid residues)<sup>73,74</sup>. Archaeal histone isoform exchange would provide archaeal organisms that rely on histone proteins for DNA compaction to reliably express critical genes under specific circumstances. For example, even closely related HMfA and HMfB, which share ~85% homology, differ in DNA binding affinity and in total abundance throughout the growth phase, suggesting that each isoform has a unique function<sup>60,65,72,75</sup>. It will thus be critical to continue to evaluate the impacts of different archaeal histone isoforms on transcription processes and cellular fitness, as these ancient DNA-binding proteins provide a platform for complex regulatory mechanisms that have come to dominate much of eukaryotic gene expression and regulation.

## Materials and Methods

### *Expression constructs for HTkA and site-directed mutagenesis*

Mutagenesis was performed on the plasmid pTS600 which encodes TK1413 with the QuikChange II XL kit (Agilent Technologies). Codons were exchanged to those that encode for variant residues G17D, E19K, G52K, E19K/G52K, R20S, T55L, E3A, R11A, and E34A.

### *Protein Purifications*

RNAP (RpoL-HA-His<sub>6</sub>), TFB, and TBP were purified as previously described<sup>56</sup>. HTkA<sup>WT</sup>, HTkA<sup>G17D</sup>, HTkA<sup>E19K</sup>, HTkA<sup>G52K</sup>, HTkA<sup>E19K/G52K</sup>, HTkA<sup>R20S</sup>, HTkA<sup>T55L</sup>, HTkA<sup>E3A</sup>, HTkA<sup>R11A</sup>, and HTkA<sup>E34A</sup> were expressed and purified from Rosetta 2 (DE3) cells (Millipore Sigma) cultured in Luria-Bertani broth supplemented with 270  $\mu$ M ampicillin and 77  $\mu$ M chloramphenicol.

Expression was induced by addition of 0.5 mM isopropyl  $\beta$ -D-1-thiogalactopyranoside and cultures were grown for 3 hours at 37° C with shaking (200 rpm). Biomass was harvested via centrifugation and lysed via sonication in 50 mM Tris-HCl pH 8.3 and 100 mM NaCl (3 mL/g pellet). The cell lysate was centrifuged at 30,000 x g for 30 minutes at 4° C, supernatant was removed, and then spun again at 67,600 x g for 30 minutes at 4° C. The supernatant containing the histones was treated with 20  $\mu$ g/mL DNase I and 5 mM MgCl<sub>2</sub> at 37° C for 2 hours and then heat-treated at 85° C for 1 hour. The heat-treated lysate was clarified by centrifugation 67,600 x g for 30 minutes at 4° C. The heat-treated clarified cell lysate was adjusted to a pH 6.0 and loaded onto a 5-mL HiTrap Heparin column (Cytiva) using an AKTA Pure FPLC system (GE Healthcare). Proteins were eluted over a 60-mL gradient to 50 mM Tris-HCl pH 7.0 and 1 M NaCl. Fractions containing histones were identified by sodium dodecyl sulfate (SDS)-PAGE and pooled. The pooled fractions were then concentrated to ~ 2 mL using Vivaspin 20, 3 kDa MWCO centrifugal concentrators (Sartorius). The concentrated pooled material was loaded over a HiPrep 16/60 Sephacryl S-100 HR equilibrated with 3 M NaCl, 50 mM Tris-HCl pH 6.8, and 5 mM 2-mercaptoethanol. Proteins were collected over a 130 mL elution in the same buffer.

Fractions containing histones were identified by sodium dodecyl sulfate (SDS)-PAGE and pooled. The pooled fractions were dialyzed into storage buffer (50 mM NaCl, 20 mM Tris-HCl pH 7, 50% glycerol). Dialyzed proteins were quantified using a Qubit Protein Assay (Invitrogen).

### ***Tris-Tricine SDS-PAGE and Western blot analysis***

Purified histones (1  $\mu$ g) were resolved on 16.5% Mini-PROTEAN® Tris-Tricine gels and stained with Coomassie brilliant blue (Fig 2B) or detected via Western blot using polyclonal anti-HTkA antibodies as previously described (Fig 2C)<sup>6</sup>.

### ***In vitro transcription***

The DNA template used in transcription assays was generated via PCR and gel purified as previously described<sup>76</sup>. Assembly of preinitiation complexes (PICs) and elongation via NTP deprivation was completed as previously described, replacing Tris-HCl pH 8.0 with Tris-HCl pH 7.0<sup>45,76,77</sup>. Stalled TECs<sub>+58</sub> (10 nM) were chilled to 4° C and then captured via RpoL-His<sub>6</sub> affinity with HisPur™ Ni-NTA Magnetic Beads (Thermoscientific). TECs<sub>+58</sub> were washed (x 3) in 180  $\mu$ L WB (20 mM Tris-HCl pH 7.0, 1 mM EDTA, 0.5 M KCl, 10  $\mu$ M ATP, GTP, UTP, 4 mM MgCl<sub>2</sub>, 0.1 mg/mL BSA, 0.2% glycerol) then resuspended in 10 mM Tris-HCl pH 7.0, 125 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM DTT, containing 10  $\mu$ M each of ATP, GTP, and UTP. The resuspended TECs<sub>+58</sub> were incubated with 3.5  $\mu$ g of HTkA<sup>WT</sup> or HTkA<sup>variant</sup> or histone storage buffer for HTkA-free for 20 minutes on ice. Elongation was reinitiated at 85° C with the addition of 25  $\mu$ M ATP, GTP, CTP, UTP (and ~9  $\mu$ M TFS in the +TFS conditions), removing aliquots after 15, 30, 60, 120, and 240 seconds (with the addition of 480 and 960 second timepoints in the extended reactions) directly to 1.2X Stop Buffer (0.6 M Tris-HCl pH 8.0, 12 mM EDTA). Radiolabeled transcripts were recovered by addition of 15  $\mu$ g of GlycoBlue™ coprecipitant (Invitrogen) following an equal volume phenol/chloroform/isoamyl alcohol (25:24:1, v:v:v) extraction, and precipitation of the aqueous phase with 2.6 volumes 100% ethanol. Precipitated transcripts were resuspended in

95% formamide, 0.1% bromophenol blue, 0.1% xylene cyanol, 20 mM EDTA, heated to 95°C for 1 minute, rapidly chilled on ice, loaded, and resolved in a 12% polyacrylamide/8M urea, 1X TBE denaturing gel. Radiolabeled RNA was detected using Typhoon™ FLA 9500 (GE Healthcare). Gel images were analyzed using ImageQuant TL 8.2 software (Cytiva).

### ***In vitro transcription exchange assay***

Chromatinized, stalled TECs<sub>+58</sub> were assembled as above with HTkA<sup>E19K/G52K</sup>. Following elongation restart as described above, 15-, 30-, 60-, and 120-second aliquots were removed, and the reaction was stopped as above. The remaining volume was split into three separate reactions at 85°C before 16.5 µg of HTkA<sup>WT</sup>, HTkA<sup>E19K/G52K</sup>, or the equivalent volume of storage buffer, respectively, was added, followed by 1-, 5-, and 10-minute aliquots directly to 1.2X Stop Buffer and processed as above.

### ***Stacked bar plots of nascent transcript length***

The products of *in vitro* transcription resolved in each lane were parsed into seven bins based on RNA lengths determined using a linear regression of pixel positions of known molecular weight standards from the 1D gel analysis in ImageQuant™ TL 8.2.0.0 (Cytiva). The mean percentage was calculated and used in ggplot2 to create stacked bar plots in both +/- TFS conditions (RStudio 2022.07.2+576 for macOS).

### ***Average RNAP elongation rate calculations***

The percentage of RNA transcripts parsed into the seven bins from the stacked bar plots was used to determine the average RNAP elongation rate (nt/s). The product from the percentage of transcripts and the theoretical RNA length (middle value within each bin) of each bin was used to determine the average transcript length at each timepoint within each environment. The increase in the average RNA length was determined by taking the difference between the average length of RNA and the starting point (+58). The average RNAP elongation rate (nt/s)

was calculated by taking the mean of the increase in the RNA average length divided by each timepoint.

### ***Pause half-life calculations***

To calculate the rate constant ( $k$ ) at position +58 we determined the average percentage of complexes in bin 1 (+58 nt) from the 1D analysis in ImageQuant<sup>TM</sup> TL 8.2.0.0 (Cytiva) from the stacked bar plots and the formula:  $C2 = C1e^{-k(t2-t1)}$ ; where  $C1$  = the average percentage of complexes in bin 1 at 15 seconds,  $C2$  = the average percentage of complexes in bin 1 at 240 seconds,  $t1$  = 15 seconds,  $t2$  = 240 seconds. The rate constant ( $k$ ) was then used in the formula:  $t_{1/2} = \ln 2/k$  to determine the average half-life of complexes at position 58 in +/- TFS conditions.

### ***Strain construction and growth conditions***

*Thermococcus kodakarensis* strains were constructed as previously described<sup>56,78</sup>. Strain RLV2 was constructed via markerless deletion of TK0533 (TFS). Deletion was confirmed by PCR amplification with primers flanking TK0533 and whole genome sequencing (WGS) on our in-house MinION, which contains the sequencing software, MinKNOW. MinKNOW does a post-run analysis that utilizes Guppy for base-calling, minimap2 for alignment with the reference genome, and medaka to call SNPs/indels, which was then visualized using IGV genome browser (2.16.1). Cultures were grown as previously described<sup>78</sup>.

### ***Molecular modeling***

To determine the predicted impacts of select HTkA<sup>variants</sup> on archaeal-histone based chromatin structures, we loaded the 5T5K PDB structure into PyRosetta-4<sup>57</sup>. In addition, we utilized the PyMOL generate symmetry mates function to stack the 5T5K PDB structure and create a new PDB file to observe the energy within the L1-L1 energy pocket. The PDB structure was cleaned using the cleanATOM function and then relaxed using several PyRosetta functions (FastRelax

and MoveMap) with backbone, sidechain, and start coordinate constraints<sup>79,80</sup>. The selected residues were then mutated utilizing the `mutate_residue` module. To determine the local energy contribution of each residue that was substituted to compare to the WT energy at that position, we utilized the ability of PyRosetta to store the total, residue, residue-pair, and residue neighbor energy information. The core scoring energy function was used on each variant to discern the contributing energy terms associated with each substitution at the given position. To observe the most significant energy contributions, we compared the WT and variant `ScoreType` that calculated the energy score of residue pairs at a given residue (`pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(res, pose, score_function, score_type, threshold=0)`)<sup>57,58,79,80</sup>. The `PyRosetta_HtKA_modeling.html` file (Supplementary Data, Appendix A) provides a step-by-step visualization of the input and output to obtain our results.



## REFERENCES

1. Sandman, K. & Reeve, J. N. Archaeal histones and the origin of the histone fold. *Current Opinion in Microbiology* **9**, 520–525 (2006).
2. Sandman, K. & Reeve, J. N. Chromosome packaging by archaeal histones. *Advances in Applied Microbiology* **50**, 73–99 (2001).
3. Sandman, K. & Reeve, J. N. Structure and functional relationships of archaeal and eukaryal histones and nucleosomes. *Archives of Microbiology* **173**, 165–169 (2000).
4. Mattioli, F. *et al.* Structure of histone-based chromatin in Archaea. *Science (80-. )*. **357**, 609–612 (2017).
5. Bhattacharyya, S., Mattioli, F. & Luger, K. Archaeal DNA on the histone merry-go-round. *FEBS J.* **285**, 3168–3174 (2018).
6. Sanders, T. J. *et al.* TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Mol. Microbiol.* **111**, 784–797 (2019).
7. Bowerman, S., Wereszczynski, J. & Luger, K. Archaeal chromatin ‘slinkies’ are inherently dynamic complexes with deflected DNA wrapping pathways. *Elife* **10**, (2021).
8. Hocher, A. *et al.* Histone-organized chromatin in bacteria. *bioRxiv* 2023.01.26.525422 (2023). doi:10.1101/2023.01.26.525422
9. Xie, Y. & Reeve, J. N. Transcription by an archaeal RNA polymerase is slowed but not blocked by an archaeal nucleosome. *J. Bacteriol.* **186**, 3492–3498 (2004).
10. Hirata, A., Klein, B. J. & Murakami, K. S. The X-ray crystal structure of RNA polymerase from Archaea. *Nature* **451**, 851–854 (2008).
11. Fouqueau, T., Blombach, F. & Werner, F. Evolutionary Origins of Two-Barrel RNA Polymerases and Site-Specific Transcription Initiation. *Annu. Rev. Microbiol.* **71**, 331–348 (2017).
12. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* **9**, 85–98 (2011).
13. Werner, F. Molecular mechanisms of transcription elongation in archaea. *Chemical Reviews* **113**, 8331–8349 (2013).
14. Wilkinson, S. P., Ouhammouch, M. & Geiduschek, E. P. Transcriptional activation in the context of repression mediated by archaeal histones. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 6777–6781 (2010).
15. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* **21**, 381–395 (2011).
16. Handy, D. E., Castro, R. & Loscalzo, J. Epigenetic modifications: Basic mechanisms and role in cardiovascular disease. *Circulation* **123**, 2145–2156 (2011).
17. Clapier, C. R., Iwasa, J., Cairns, B. R. & Peterson, C. L. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature Reviews Molecular Cell Biology* **18**, 407–422 (2017).
18. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nature Reviews Microbiology* **15**, 711–723 (2017).
19. Stevens, K. M., Hocher, A. & Warnecke, T. Deep Conservation of Histone Variants in Thermococcales Archaea. *Genome Biol. Evol.* **14**, (2022).
20. Krupovic, M., Dolja, V. V. & Koonin, E. V. The virome of the last eukaryotic common ancestor and eukaryogenesis. *Nat. Microbiol.* **8**, 1008–1017 (2023).
21. Brunk, C. F. & Martin, W. F. Archaeal Histone Contributions to the Origin of Eukaryotes. *Trends in Microbiology* **27**, 703–714 (2019).
22. Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V. & Martin, W. F. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. (2018). doi:10.1371/journal.pgen.1007518
23. Martin, W. F. Early evolution without a tree of life. *Biol. Direct* **6**, 36 (2011).

24. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
25. Kireeva, M. L. *et al.* Nature of the nucleosomal barrier to RNA polymerase II. *Mol. Cell* **18**, 97–108 (2005).
26. Teves, S. S., Weber, C. M. & Henikoff, S. Transcribing through the nucleosome. *Trends Biochem. Sci.* **39**, 577–586 (2014).
27. Ehara, H. *et al.* Structure of the complete elongation complex of RNA polymerase II with basal factors. *Science (80-. )*. **357**, 921–924 (2017).
28. Kujirai, T. *et al.* Structural basis of the nucleosome transition during RNA polymerase II passage. *Science (80-. )*. **362**, 595–598 (2018).
29. Soares, D. J., Sandman, K. & Reeve, J. N. Mutational analysis of archaeal histone-DNA interactions. *J. Mol. Biol.* **297**, 39–47 (2000).
30. Pereira, S. L., Grayling, R. A., Lurz, R. & Reeve, J. N. Archaeal nucleosomes. *Proc. Natl. Acad. Sci.* **94**, 12633–12637 (1997).
31. Bhattacharyya, S., Mattioli, F. & Luger, K. Archaeal DNA on the histone merry-go-round. *FEBS Journal* **285**, 3168–3174 (2018).
32. Soman, A. *et al.* Columnar structure of human telomeric chromatin. *Nat.* 2022 6097929 **609**, 1048–1055 (2022).
33. Seitz, K. W., Lazar, C. S., Hinrichs, K. U., Teske, A. P. & Baker, B. J. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* **10**, 1696–1705 (2016).
34. Nalabothula, N. *et al.* Archaeal nucleosome positioning in vivo and in vitro is directed by primary sequence motifs. *BMC Genomics* **14**, 391 (2013).
35. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772 (2006).
36. Henneman, B., van Emmerik, C., van Ingen, H. & Dame, R. T. Structure and function of archaeal histones. *PLoS Genetics* **14**, e1007582 (2018).
37. Li, W. T., Sandman, K., Pereira, S. L. & Reeve, J. N. 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* **4**, 43–51 (2000).
38. Sanders, T. J., Marshall, C. J. & Santangelo, T. J. The Role of Archaeal Chromatin in Transcription. *Journal of Molecular Biology* (2019). doi:10.1016/j.jmb.2019.05.006
39. Hoher, A. *et al.* Growth temperature and chromatinization in archaea. *Nat. Microbiol.* 2022 711 **7**, 1932–1942 (2022).
40. Soares, D. J., Sandman, K. & Reeve, J. N. Mutational analysis of archaeal histone-DNA interactions. *J. Mol. Biol.* **297**, 39–47 (2000).
41. Sanders, T. J. *et al.* Extended Archaeal Histone-Based Chromatin Structure Regulates Global Gene Expression in Thermococcus kodakarensis. *Front. Microbiol.* **12**, 1071 (2021).
42. Wenck, B. R. & Santangelo, T. J. Archaeal transcription. *Transcription* **11**, 199–210 (2020).
43. Bell, S. D., Cairns, S. S., Robson, R. L. & Jackson, S. P. Transcriptional regulation of an Archaeal operon in vivo and in vitro. *Mol. Cell* **4**, 971–982 (1999).
44. Ouhammouch, M., Hausner, W. & Geiduschek, E. P. TBP domain symmetry in basal and activated archaeal transcription. *Mol. Microbiol.* **71**, 123–131 (2009).
45. Santangelo, T. J., Čuboňová, L. L., James, C. L. & Reeve, J. N. TFB1 or TFB2 Is Sufficient for Thermococcus kodakaraensis Viability and for Basal Transcription in Vitro. *J. Mol. Biol.* **367**, 344–357 (2007).
46. Yakhnin, A. V. & Babitzke, P. NusG/Spt5: Are there common functions of this ubiquitous transcription elongation factor? *Current Opinion in Microbiology* **18**, 68–71 (2014).

47. Rojec, M., Hocher, A., Stevens, K. M., Merckenschlager, M. & Warnecke, T. Chromatinization of escherichia coli with archaeal histones. *Elife* **8**, (2019).
48. Čuboňová, L. *et al.* An archaeal histone is required for transformation of *Thermococcus kodakarensis*. *J. Bacteriol.* **194**, 6864–6874 (2012).
49. Sandman, K., Soares, D. & Reeve, J. N. Molecular components of the archaeal nucleosome. *Biochimie* **83**, 277–281 (2001).
50. Bailey, K. A., Pereira, S. L., Widom, J. & Reeve, J. N. Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution. *J. Mol. Biol.* **303**, 25–34 (2000).
51. Spitalny, P. & Thomm, M. Analysis of the Open Region and of DNA-Protein Contacts of Archaeal RNA Polymerase Transcription Complexes during Transition from Initiation to Elongation. *J. Biol. Chem.* **278**, 30497–30505 (2003).
52. Erkelens, A. M., Henneman, B., van der Valk, R. A., Kirolos, N. C. S. & Dame, R. T. Specific DNA binding of archaeal histones HMfA and HMfB. *Front. Microbiol.* **14**, (2023).
53. Decanniere, K., Babu, A. M., Sandman, K., Reeve, J. N. & Heinemann, U. Crystal structures of recombinant histones HMfA and HMfB, from the hyperthermophilic archaeon *Methanothermus fervidus*. *J. Mol. Biol.* **303**, 35–47 (2000).
54. Hausner, W., Lange, U. & Musfeldt, M. Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase. *J. Biol. Chem.* **275**, 12393–12399 (2000).
55. Lange, U. & Hausner, W. Transcriptional fidelity and proofreading in Archaea and implications for the mechanism of TFS-induced RNA cleavage. *Mol. Microbiol.* **52**, 1133–1143 (2004).
56. Gehring, A. M., Sanders, T. J. & Santangelo, T. J. Markerless Gene Editing in the Hyperthermophilic Archaeon *Thermococcus kodakarensis*. *Bio-protocol* **7**, (2017).
57. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).
58. Alford, R. F. *et al.* The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031 (2017).
59. Naeem, R. 16.6: The Repulsive Term in the Lennard-Jones Potential - Chemistry LibreTexts. *LibreTexts* 13661 (2021). Available at: [https://chem.libretexts.org/Bookshelves/Physical\\_and\\_Theoretical\\_Chemistry\\_Textbook\\_Maps/Physical\\_Chemistry\\_\(LibreTexts\)/16%3A\\_The\\_Properties\\_of\\_Gases/16.06%3A\\_The\\_Repulsive\\_Term\\_in\\_the\\_Lennard-Jones\\_Potential](https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Physical_Chemistry_(LibreTexts)/16%3A_The_Properties_of_Gases/16.06%3A_The_Repulsive_Term_in_the_Lennard-Jones_Potential). (Accessed: 28th April 2023)
60. Marc, F., Sandman, K., Lurz, R. & Reeve, J. N. Archaeal Histone Tetramerization Determines DNA Affinity and the Direction of DNA Supercoiling. *J. Biol. Chem.* **277**, 30879–30886 (2002).
61. Chang, C. H. & Luse, D. S. The H3/H4 tetramer blocks transcript elongation by RNA polymerase II in vitro. *J. Biol. Chem.* **272**, 23427–23434 (1997).
62. Manohar, M. *et al.* Acetylation of Histone H3 at the Nucleosome Dyad Alters DNA-Histone Binding. *J. Biol. Chem.* **284**, 23312–23321 (2009).
63. Hu, J. *et al.* H4K44 acetylation facilitates chromatin accessibility during meiosis. *Cell Rep.* **13**, 1772 (2015).
64. Simon, M. *et al.* Histone fold modifications control nucleosome unwrapping and disassembly. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 12711–12716 (2011).
65. Bailey, K. A., Marc, F., Sandman, K. & Reeve, J. N. Both DNA and histone fold sequences contribute to archaeal nucleosome stability. *J. Biol. Chem.* **277**, 9293–9301 (2002).
66. French, S. L., Santangelo, T. J., Beyer, A. L. & Reeve, J. N. Transcription and translation are coupled in Archaea. *Mol. Biol. Evol.* **24**, 893–5 (2007).
67. Sanders, T. J. *et al.* FttA is a CPSF73 homologue that terminates transcription in

- Archaea. *Nature Microbiology* **5**, 545–553 (2020).
68. Yue, L. *et al.* The conserved ribonuclease aCPSF1 triggers genome-wide transcription termination of Archaea via a 3'-end cleavage mode. *Nucleic Acids Res.* (2020). doi:10.1093/nar/gkaa702
  69. Blombach, F., Fouqueau, T., Matelska, D., Smollett, K. & Werner, F. Promoter-proximal elongation regulates transcription in archaea. *Nat. Commun.* **2021** *121* **12**, 1–15 (2021).
  70. Jeon, C., Yoon, H. & Agarwal, K. The transcription factor TFIIIS zinc ribbon dipeptide Asp-Glu is critical for stimulation of elongation and RNA cleavage by RNA polymerase II. *Proc. Natl. Acad. Sci.* **91**, 9106–9110 (1994).
  71. Szádeczky-Kardoss, I. *et al.* Elongation factor TFIIIS is essential for heat stress adaptation in plants. *Nucleic Acids Res.* **50**, 1927–1950 (2022).
  72. Stevens, K. M. *et al.* Histone variants in archaea and the evolution of combinatorial chromatin complexity. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33384–33395 (2020).
  73. Tagami, H., Ray-Gallet, D., Almouzni, G. & Nakatani, Y. Histone H3.1 and H3.3 Complexes Mediate Nucleosome Assembly Pathways Dependent or Independent of DNA Synthesis. *Cell* **116**, 51–61 (2004).
  74. Franklin, R., Murn, J. & Cheloufi, S. Cell Fate Decisions in the Wake of Histone H3 Deposition. *Front. Cell Dev. Biol.* **9**, 836 (2021).
  75. Sandman, K., Grayling, R. A., Dobrinski, B., Lurz, R. & Reeve, J. N. Growth-phase-dependent synthesis of histones in the archaeon *Methanothermus fervidus*. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 12624–12628 (1994).
  76. Walker, J. E., Luyties, O. & Santangelo, T. J. Factor-dependent archaeal transcription termination. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6767–E6773 (2017).
  77. Gehring, A. M. & Santangelo, T. J. Manipulating archaeal systems to permit analyses of transcription elongation-termination decisions in vitro. *Methods Mol. Biol.* **1276**, 263–79 (2015).
  78. TJ, S. *et al.* Extended Archaeal Histone-Based Chromatin Structure Regulates Global Gene Expression in *Thermococcus kodakarensis*. *Front. Microbiol.* **12**, (2021).
  79. Gray, J. J., Chaudhury, S. & Lyskov, S. The PyRosetta Interactive Platform for Protein Structure Prediction and Design A Set of Educational Modules. (2009).
  80. Gray, J. J., Chaudhury, S., Lyskov, S. & Labonte, J. W. *The PyRosetta Interactive Platform for Protein Structure Prediction and Design: A Set of Educational Modules.* (2013).

## CHAPTER 3: FTTA IS A CPSF73 HOMOLOGUE THAT TERMINATES TRANSCRIPTION IN ARCHAEA

### Summary

Only select sequences or transcription termination factors can disrupt the otherwise extremely stable transcription elongation complex. We demonstrate that one of the last universally conserved archaeal proteins with unknown biological function is the Factor that terminates transcription in Archaea (FttA). FttA is an orthologue of the eukaryotic cleavage and polyadenylation specificity factor subunit 73 (CPSF73) that cleaves nascent transcripts and terminates archaeal transcription. FttA-mediated termination shares mechanistic similarities with bacterial rho-mediated transcription termination and is kinetically coupled to RNA polymerase by the only universally conserved transcription elongation factor, Spt5. FttA preferentially cleaves C- and U-rich RNA and while addition of Spt4/5 tempers sequence requirements, FttA-mediated termination is reliant on the conserved stalk domain retained in eukaryotic and archaeal RNA polymerases. Reduced FttA expression and inhibition of FttA-mediated termination results in altered 3'-end formation *in vivo*. Our results complete the archaeal transcription cycle, provide a 'missing-link' between prokaryotic and eukaryotic transcription regulation, and rationalize the evolution of the processing activities involved in RNA 3'-end formation.

---

<sup>2</sup> This chapter was previously published under the following title, with the addition of Cryo-EM data that is currently in review at the time of this report: Sanders, T. J. *et al.* FttA is a CPSF73 homologue that terminates transcription in Archaea. *Nature Microbiology* **5**, 545–553 (2020).

## Introduction

The extreme stability required of transcription elongation complexes (TECs) to processively transcribe large genomic regions necessitates robust mechanisms to terminate transcription. Efficient transcription termination is particularly critical for gene-dense bacterial and archaeal genomes<sup>1–3</sup> wherein continued transcription would necessarily transcribe immediately adjacent genes, result in conflicts between the transcription and replication apparatuses<sup>4–6</sup> and the coupling of transcription and translation<sup>7,8</sup> would permit loading of ribosomes onto aberrant transcripts. Transcription termination (Fig. B.1), driven either by DNA sequence and encoded RNA structures (e.g. intrinsic termination) or by protein factors (e.g. factor-dependent termination) ensures rapid dissociation of RNA polymerase (RNAP) from the DNA template to maintain efficient transcription, recycle RNAP and generate RNA 3' ends<sup>1,9</sup>. Intrinsic termination can efficiently disrupt bacterial and archaeal transcription<sup>2,10,11</sup>, as well as eukaryotic RNA Pol III complexes<sup>12</sup> while eukaryotic RNA Pol I and Pol II require factor-mediated termination events<sup>13</sup>. While often prevalent within prokaryotic genomes, intrinsic termination sequences are typically not sufficiently abundant, nor efficient enough, to mediate all termination events, and when TECs from all domains become arrested at DNA lesions, specific proteins or protein complexes are required to release RNAP from the DNA template<sup>14–17</sup>.

Transcription termination is often defined as release of the nascent transcript from RNA polymerase (RNAP) into solution. This definition suffices for bacterial, archaeal, and eukaryotic Pol III intrinsic termination, wherein release of the entire transcript necessarily removes the 3'-portion of the transcript paired to the template DNA strand (Fig. B.1; panels a, b, and c). Disrupting the RNA-DNA hybrid within the TEC permits collapse of the transcription bubble and the remaining minimal RNAP-DNA contacts are insufficient to maintain a stable complex. Factor-mediated bacterial transcription termination<sup>10</sup>, driven by rho or Mfd, also directs release of the entire nascent transcript and results in collapse of the TEC and recycling of RNAP (Fig.

B.1; panels d and g). Rho, an ATP-dependent helicase<sup>18–20</sup>, acts to terminate transcription of bacterial TECs upon the uncoupling of transcription and translation by recognizing a ribosome-free, C-rich RNA sequence termed a *rut* site (rho-ut<sup>ilization</sup>)<sup>9,19,21</sup>. Rho-mediated termination is aided by the activity of universally conserved transcription elongation factor NusG (Spt5 in Eukarya)<sup>21,22</sup>; NusG allosterically activates rho and enhances rho activity at sub-optimal *rut*-sites<sup>23–25</sup>. Mfd can either restart backtracked or stalled elongation factors when elongation is possible or terminate transcription when elongation is blocked by template strand DNA lesions<sup>15,26–28</sup>.

However, release of the majority of the nascent transcript cannot be considered a *bona fide* termination event in of itself (Fig. B.1; panel e). RNA processing events, such as the endonucleolytic cleavage of the nascent RNA within eukaryotic Pol II TECs by the cleavage and polyadenylation factor complex (CPSF)<sup>29–33</sup> yield a 5'-fragment that is often further processed – typically by the addition of a 3'-polyA tail for many Pol II transcripts – but also a 3'-fragment that is encapsulated within a still-stable TEC<sup>34,35</sup>. Stable Pol II complexes, now containing a 5'-monophosphate (5'-P) nascent RNA can continue transcription for several hundreds or even thousands of base pairs following endonucleolytic cleavage of the transcript before protein factors mediate true termination of transcription<sup>36,37</sup>. The exonucleolytic activity of Xrn2<sup>34,38–40</sup> is necessary for proper and *bona fide* transcription termination *in vivo*, and although the mechanistic details are unknown<sup>13</sup>, Pol II is ultimately recycled and the extended 3' portion of the nascent transcript is typically rapidly degraded.

Termination factors that serve as global governance regulators of both normal and aberrant transcription must efficiently recognize TECs and compete with continued elongation to mediate release of the nascent transcript from the still elongating TEC. While the identification of Eta provided evidence of factor-dependent transcription termination in Archaea<sup>17</sup>, no kinetically-efficient mechanism of factor-dependent archaeal transcription termination has been described.

The retention of operon-organized archaeal genomes and the sensitivity of the archaeal transcription apparatus to bacterial rho-mediated termination *in vitro*<sup>2</sup> – combined with the normal coupling of transcription and translation<sup>8</sup>, the resultant polar suppression of downstream expression in the absence of such coupling in archaeal cells<sup>41</sup> and the conservation of Spt5/NusG in all genomes – implied the existence of a kinetically-relevant archaeal transcription termination activity that might function akin to the bacterial rho protein. Rho homologues are, however, restricted to Bacteria<sup>42</sup>. Archaea are the progenitors of Eukarya<sup>43,44</sup> and archaeal information processing systems are typically component simplified versions of the more complex eukaryotic systems that dictate replication, repair and transcription within a chromatin landscape<sup>45</sup>, arguing instead that conserved archaeal-eukaryotic or unique archaeal factors may drive factor-dependent archaeal transcription termination.

In Eukarya, the CPSF73 subunit of the cleavage and polyadenylation complex cleaves RNA downstream of the polyadenylation signal, allowing Xrn2/Rat1 to bind and exonucleolytically degrade the uncapped RNA associated with Pol II<sup>30,32,33,38,40,46</sup>. CPSF73, as part of the RNA 3'-maturation machinery, initiates cleavage of the nascent transcript, a necessary step for mRNA maturation that in-of-itself does not direct transcription termination; Pol II can continue transcription for thousands of base pairs beyond the cleavage position<sup>35,38</sup>. Cleavage by the CPSF complex instead generates a 5'-uncapped, monophosphate end that permits Rat1/Xrn2-mediated degradation of the nascent transcript up to the boundary imposed by Pol II<sup>34</sup>. Although Rat1/Xrn2 is not capable of directly disrupting the Pol II TEC<sup>34</sup>, the combined activities of CPSF and Xrn2 are necessary for normal termination patterns in Eukarya<sup>35,39,40</sup>.

Only a core set of 129 genes are conserved in all sequenced archaeal genomes<sup>47</sup> and the vast majority of these have obvious eukaryotic, rather than bacterial homologues. One of these universally conserved and essential archaeal proteins is an obvious orthologue of a subunit of the cleavage and polyadenylation specificity factor (CPSF) complex<sup>32,48,49</sup>. The homology of



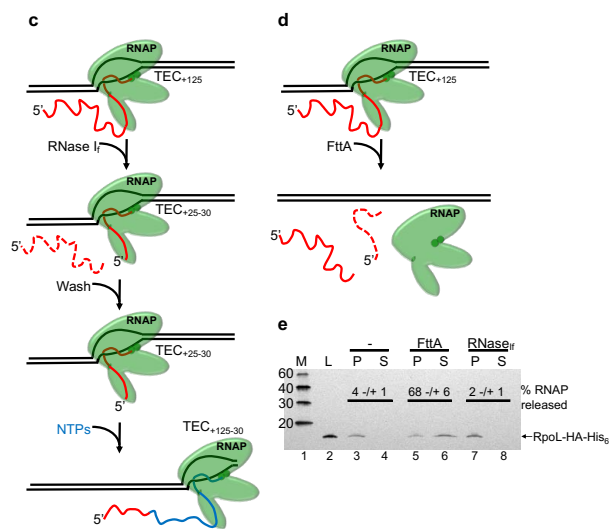
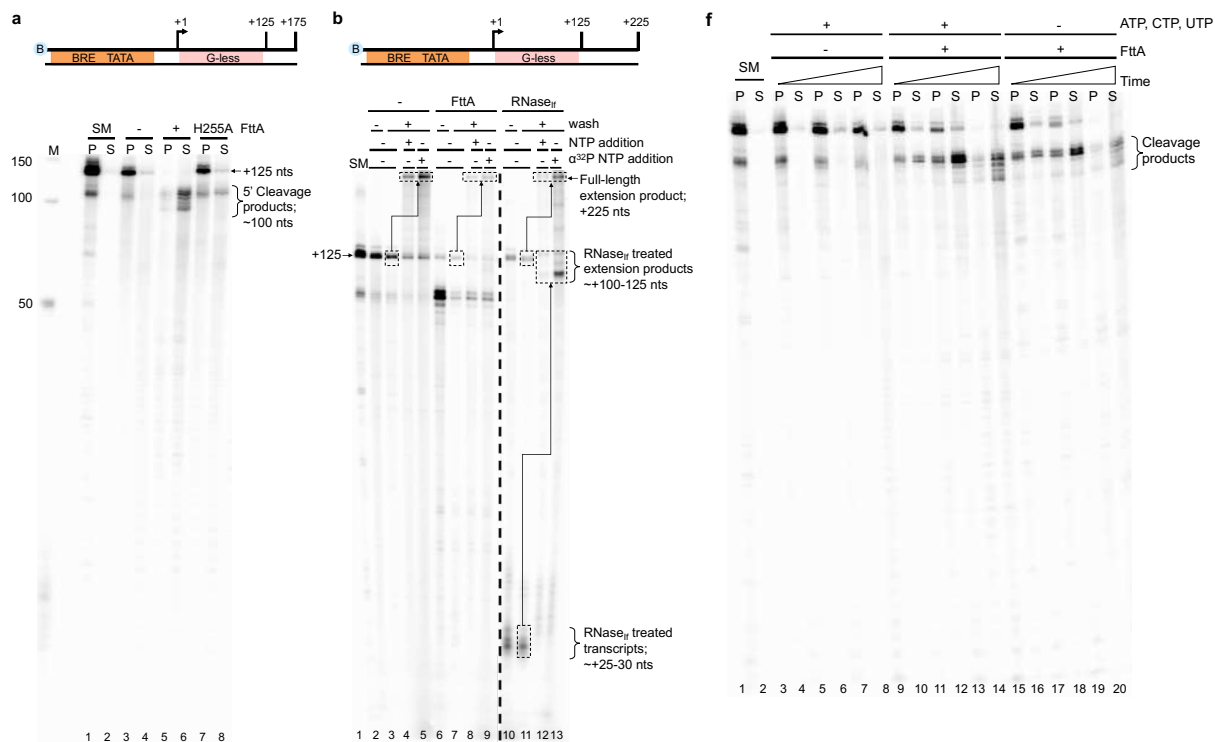
most archaeal transcription components to eukaryotic factors argued that the archaeal homologue (aCPSF1) of eukaryotic CPSF73 might function as the Factor that terminates transcription in Archaea (FttA). Here, using components purified from the model archaeal species *Thermococcus kodakarensis* we establish that FttA can disrupt the otherwise extremely stable TEC and is thus a *bona fide* archaeal transcription termination factor. We provide additional evidence that FttA-mediated transcription termination is competitive with transcription elongation. By manipulating steady-state *in vivo* FttA protein levels or inhibiting FttA activity we demonstrate changes in RNA 3'-end formation, fully supportive of FttA-mediated termination impacting transcription *in vivo*. Identification and characterization of FttA completes the archaeal transcription cycle and links transcription regulation between the three Domains. FttA-mediated transcription termination shares many attributes with rho-mediated transcription termination, but FttA retains a structure with obvious homology to the eukaryotic CPSF73 protein. FttA is kinetically coupled to RNAP by the only universally conserved transcription elongation factor, Spt5 and the archaeal-eukaryotic specific RNAP stalk domain. The 3.9 Å resolution Cryo-EM structure of the archaeal pre-termination complex (TEC, FttA dimer, and Spt4/5) validates the necessary contacts between RNAP, Spt5, and FttA. In addition, the structure confirms the formation of an FttA dimer – one monomer proximal to the TEC (FttA<sup>prox</sup>) and the other monomer distal to the TEC (FttA<sup>dist</sup>). The arrangement of the structure encompasses the Spt5 NGN domain in contact with Spt4 and RNAP, the Spt5 KOW domain interactions with FttA<sup>prox</sup>, and direct contact between FttA<sup>prox</sup> and the mouth of the TEC RNA exit channel<sup>99</sup> (Fig. 3.6). FttA thus resolves the dichotomy of a prokaryotic gene structure (operons and polarity) and eukaryotic molecular homology (general transcription apparatus) observed in Archaea. This missing-link between prokaryotic and eukaryotic transcription regulation provides the most parsimonious link to the evolution of the processing activities involved in RNA 3'-end formation in Eukarya.

## Results

### ***FttA is the “missing” archaeal transcription termination factor that completes the archaeal transcription cycle.***

We hypothesized that the metallo-beta-lactamase fold, beta-CASP domain archaeal homologue of the eukaryotic CPSF73<sup>47,48,50</sup> protein might function as the factor that terminates transcription in Archaea (FttA). The retention of an FttA-homologue in all sequenced archaeal genomes, including the substantially reduced genomes of several nanoarchaeal-species suggested that FttA-activity was essential for proper regulation of transcription *in vivo*. Our *in vitro* transcription system using components from *Thermococcus kodakarensis*<sup>51,52</sup> permits assembly of pre-initiation complexes that can be elongated to specific template positions by addition of NTP subsets resulting in formation of TECs that are stable to extensive washing (Fig. 3.1). By assembling TECs on promoter-directed, biotinylated DNA templates attached to a solid-support, radiolabeled nascent transcripts associated with intact TECs are easily distinguished from equivalent length transcripts that have been released to solution through spontaneous or factor-mediated termination events. By employing an RNAP variant with an epitope- and affinity-tagged RpoL subunit<sup>51,53,54</sup>, the retention of RNAP with the DNA can easily be monitored to follow not only RNA processing and release, but also RNAP release from the DNA template and *bona fide* termination of archaeal TECs.

The product of TK1428 is a ~73.5 Kda protein that is easily expressed and recombinantly purified to homogeneity (Fig. B.1; panel i). TECs stalled by nucleotide deprivation with +125 nucleotide (nt) nascent transcripts (TECs<sub>+125</sub>) remain stably associated with the solid support (P = pellet fraction) in the absence of FttA (Fig. 3.1; panel a, lanes 1-4). Addition of FttA to stalled TECs results in rapid cleavage and release of ~100 nts of the nascent transcript from the TEC into solution (S = supernatant) (Fig. 3.1; panel a, lanes 5&6). Cleavage of a +125 nt nascent transcript labeled with  $\alpha$ -<sup>32</sup>P-UTP to yield a 5'-fragment of ~100 nts should also necessarily



**Figure 3.1. FttA is a *bona fide* archaeal transcription termination factor.** **a.** FttA directs transcript cleavage and transcription termination. The *in vitro* transcription system using components from *T. kodakarensis* permits promoter-directed assembly of pre-initiation complexes with purified RNAP, TFB, and TBP on biotinylated DNA templates that can be attached to streptavidin-coated paramagnetic beads. Transcripts associated with intact TECs are retained in pellet (P) fractions whereas transcripts released from terminated complexes partition into the supernatant (S). Addition of ATP, CTP, UTP and  $^{32}\text{P}$ - $\alpha$ -UTP allows RNAP to elongate to the end of a +125 nt G-less cassette and generate a uniformly radiolabeled RNA transcript. Radiolabeled transcripts within starting material (SM) TECs<sub>+125</sub> and mock treated TECs<sub>+125</sub> are retained in pellet fractions (lanes 1-4), whereas FttA<sup>WT</sup> addition results in cleavage of nascent transcripts and termination of most TECs (lanes 5-6). Addition of a catalytically deficient FttA variant (FttA<sup>H255A</sup>) abrogates cleavage and RNA release (lanes 7-8). Lane M contains  $^{32}\text{P}$ -labeled ssDNA markers. **b.** FttA-mediated termination is distinct from RNase treatment of intact TECs. TECs<sub>+125</sub> (SM, lane 1) are resistant to repeated high-salt buffer washes and readily resume elongation upon NTP addition to generate +225 nt full-length transcripts (lanes 2-5). Dashed boxes and arrows denote +125 transcripts that are elongated to +225 nt transcripts; the specific activity of +225 transcripts can be increased by addition of additional  $^{32}\text{P}$ - $\alpha$ -UTP during elongation from +125 to +225. RNase I<sub>f</sub> digestion of nascent transcripts associated with washed TECs<sub>+125</sub> results in degradation of the nascent transcript to just ~20-30 nts, but TECs with shortened transcripts remain associated with the DNA and thus survive repeated washing (lanes 10-11). TECs<sub>~+25-30</sub> resultant from RNase I<sub>f</sub> treatment of TECs<sub>+125</sub> readily resume elongation upon NTP addition to generate ~+125 nt full-length transcripts (lanes 12-13). Dashed boxes and arrows denote ~+25 transcripts that are elongated to ~+125 nt transcripts; the specific activity of ~+125 transcripts can be increased by addition of additional  $^{32}\text{P}$ - $\alpha$ -UTP during elongation from ~+25 to ~+125. FttA addition to TECs<sub>+125</sub> results in the disruption of most TECs with nascent transcript cleavage (lanes 6-9). FttA-mediated termination results in release of most TECs from the template and cleaved transcripts cannot be extended by NTP addition (lanes 8-9). **c** and **d.** Schematic diagrams of the fate of TECs<sub>+125</sub> follow RNase I<sub>f</sub> and FttA treatment, respectively. **e.** FttA, but not RNase I<sub>f</sub> treatment releases RNAP from the DNA template into solution confirming dissociation of the TEC and *bona fide* FttA-mediated transcription termination. Release of RNAP is tracked and quantified by Western blots (n = 3 independent replicates) with anti-HA antibodies that recognize the modified RpoL-subunit of RNAP. **f.** FttA is not reliant on NTP hydrolysis to quickly inactivate TECs, cleave nascent transcripts and terminate transcription.

generate – at least temporarily – a ~25 nt radiolabeled 3'-fragment. Despite repeated and exhaustive efforts to monitor FttA-mediated cleavage within just seconds of FttA-addition, we were unable to monitor production of a radiolabeled 3'-fragment of the transcript. We were thus initially hesitant to assume that FttA-mediated transcript cleavage was coupled to *bona fide* transcription termination, as a ~25 nt transcript is sufficient to stabilize an archaeal TEC<sup>2,55,56</sup>.

We aimed to conclusively demonstrate if FttA-mediated cleavage of nascent transcripts was directly coupled to the termination of transcription (Fig. 3.1; panels b, c, and d). If TECs remain

intact following FttA-mediated cleavage of the nascent transcript then i) radiolabeled 3'-nascent transcripts should remain associated with TECs and partition with the DNA attached to the solid support (pellet fraction), ii) intact TECs should survive washes designed to remove transcripts not associated with TECs from the solid support, iii) supplementation with NTPs should permit continued elongation of active TECs, allowing the nascent transcript to grow in length, and iv) RNAP should remain within TECs and thus partition to the pellet fraction. In contrast, if FttA-mediated cleavage of the transcripts inactivates and terminates transcription, RNAP should be released to the supernatant and resumed elongation following NTP supplementation will not be possible. Previous studies demonstrated that archaeal TECs could be treated with Rnase I<sub>f</sub> to degrade the nascent transcript to just ~25 nts<sup>17</sup> (Fig. 3.1; panel c). Treatment of TECs<sub>+125</sub> with Rnase I<sub>f</sub> generates stable TECs<sub>~+25</sub> that are easily observed via retention of radiolabeled RNAs in washed TECs that partition to the pellet fraction (Fig. 3.1; panel b, lanes 10&11). Observation of ~+25 nt transcripts in Rnase I<sub>f</sub>-treated TECs confirms that the  $\alpha$ -<sup>32</sup>P-UTP labeling is sufficient to identify intact ~+25 nt transcripts associated with stable TECs. Upon addition of unlabeled NTPs (Fig. 3.1; panel b, lane 12), all TECs<sub>~+25</sub> resume elongation and nascent transcripts are extended until RNAP reaches the end of the template. If additional radiolabeled NTPs are provided during this extension (Fig. 3.1; panel b, lane 13), the specific activity of radiolabeled transcripts increases, but the overall length of the extended transcript does not exceed that permitted by elongation to the end of the DNA template. Finally, examination of RNAP partitioning following Rnase I<sub>f</sub> treatment confirms that essentially all TECs remain intact and RNAP is retained in the pellet fraction (Fig. 3.1; panel e, lanes 7&8). Thus, treatment of TECs<sub>+125</sub> with Rnase I<sub>f</sub> fulfills all of the expectations of transcript processing that is not linked to transcription termination.

The results obtained by treating TECs<sub>+125</sub> with Rnase I<sub>f</sub> contrast completely with the results of treatment of identically prepared TECs<sub>+125</sub> with FttA (Fig. 3.1; panels b-e) and are fully

supportive of FttA-mediated transcription termination. The vast majority of TECs<sub>+125</sub> treated with FttA do not survive washes, instead releasing ~100 nt transcripts into solution that cannot be extended upon the addition of NTPs or radiolabeled NTPs. Again, no evidence of retained ~+25 nt 3'-fragments of the nascent transcript can be visualized, implying the 3'-fragment is rapidly degraded to single nts (or minimally RNA fragments < 10 nts). A small percentage of TECs<sub>+125</sub> escape FttA-action, are retained through washing and readily extend upon NTP addition, confirming conditions support elongation of intact TECs (Fig. 3.1; panel b, lanes 8&9). In addition to FttA-mediated transcript cleavage, FttA-activity releases ~70% of RNAP to solution, demonstrating collapse of the TEC and recycling of RNAP for additional rounds of transcription (Fig. 3.1; panel e, lanes 5&6). Taken together, i) the failure of FttA-treated TECs to remain intact by evidence of association of the DNA, RNAP, and nascent transcripts with the solid support, ii) the failure of FttA-treated TECs to retain a 3'-fragment of RNA, iii) the failure of FttA-treated complexes to resume elongation upon NTP addition and iv) the abundant release of RNAP into solution following FttA-treatment confirm that FttA mediates both transcript cleavage and *bona fide* transcription termination of archaeal TECs. FttA is thus the second archaeal-encoded factor that can mediate transcription termination *in vitro*.

Addition of an FttA variant, wherein an active site histidine of the metallo-beta-lactamase fold<sup>31,48,49,57</sup> was replaced by an alanine (FttA<sup>H255A</sup>), reduced FttA-mediated transcript processing and transcription termination to background levels (Fig. 3.1; panel a, lanes 7&8). Thus, transcription termination activity is linked to FttA-mediated RNA cleavage, rather than FttA-mediated stimulation of the intrinsic cleavage activity of RNAP<sup>58</sup>. The cleavage pattern of the nascent RNA is only compatible with RNAP-mediated transcript cleavage if the entire population of TECs normally reverse-translocated (i.e., backtracked) ~25 nts. Extensive and uniform reverse-translocation of TECs would be unprecedented and backtracked TECs would not be predicted to readily resume elongation upon NTP addition; it is much more parsimonious to

assume the active center of FttA-mediate endonucleolytic cleavage of the nascent transcript. FttA-mediated cleavage of the nascent RNA to yield a ~100 nt 5'-transcript is consistent with FttA stimulating RNA cleavage at the first point of reduced contact between RNAP and the nascent transcript and the ~25 nt of nascent transcript protection are consistent with the results of previous digestions of intact archaeal<sup>17</sup> and eukaryotic TECB with RNA exonucleases.

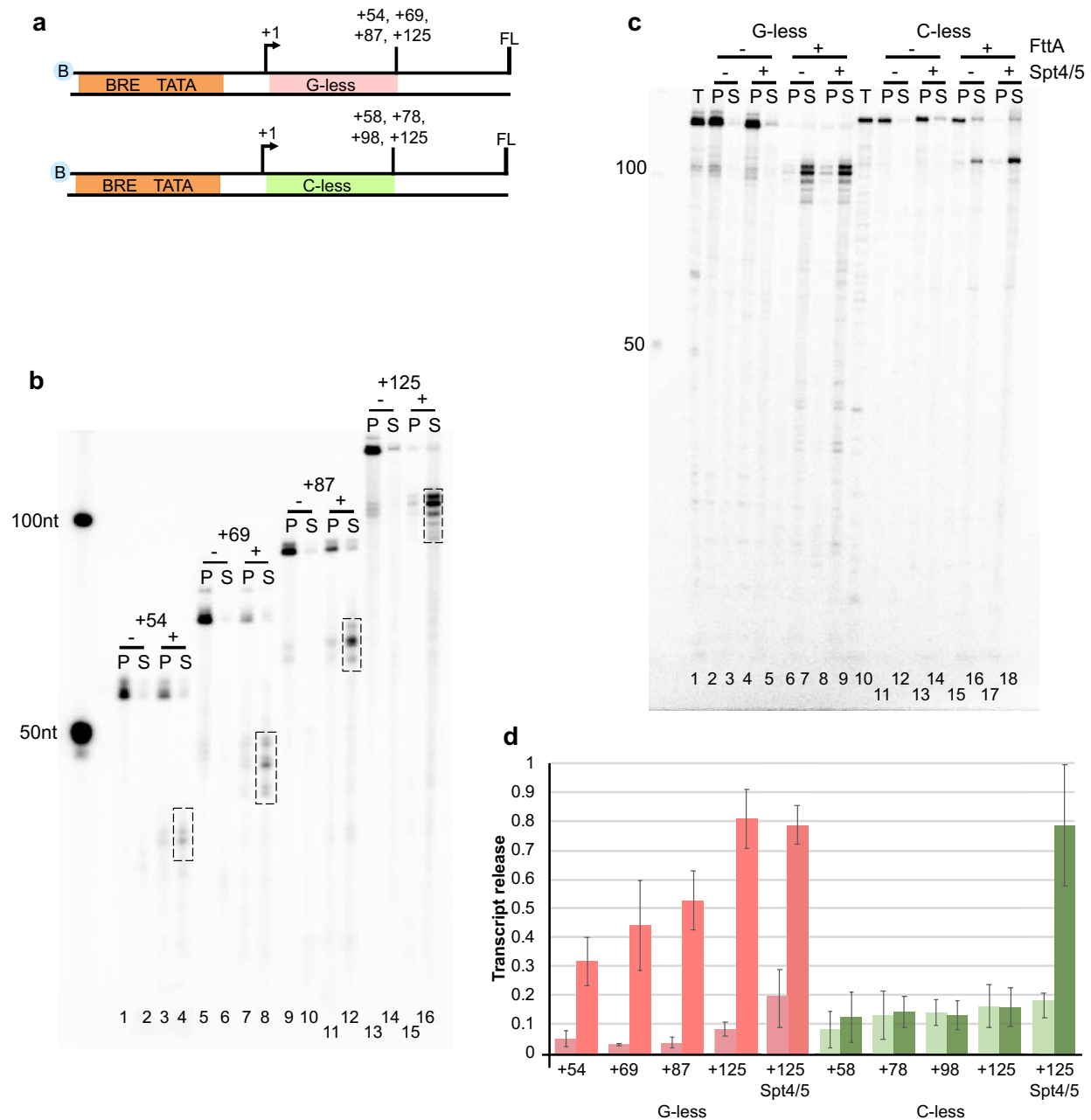
In contrast to other prokaryotic transcription termination factors (*i.e.*, rho, Eta and Mfd), FttA-mediated transcript cleavage and termination is not energy-dependent; equal rates of transcript cleavage and termination of archaeal TECs are observed in reactions lacking and containing free NTPs (Fig. 3.1; panel f). Monitoring transcript cleavage and termination over time reveals that FttA-mediated transcript cleavage yields a ~100 nt 5'-transcript, but at no time is a > 10 nt 3'-fragment identified, again suggesting that FttA-activity results in very rapid degradation of the 3'-fragment and termination of transcription coupled to release of RNAP into solution (Fig. 3.1; panels d-f).

***FttA-mediated termination shares mechanistic requirements of rho-mediated bacterial transcription termination.***

Transcription units (*e.g.* genes or operons) are typically separated by only short (< 100 bp) intergenic regions in gene dense archaeal genomes<sup>1,45,59,60</sup>. For FttA-mediated termination to be effective *in vivo*, FttA must quickly recognize and disrupt TECs before transcription continues into downstream genes or operons. Rho-mediated termination requires ~70 nucleotides (nts) of C-rich unstructured nascent RNA for optimal activity<sup>19</sup>; such sequences typically signal the uncoupling of transcription and translation for productive transcription, or rogue transcription of viral or genomic sequences<sup>22</sup>.

The requirements for FttA-mediated transcription termination revealed very similar requirements to those for rho<sup>9,19,21,61</sup>, suggesting that FttA-activity is likely stimulated by the same biological

signals as rho-mediated termination. FttA has no requirement for substantial upstream or downstream DNA sequences (not shown), and instead recognizes TECs through binding to nascent transcripts (Fig. 3.2). Incubation of TECs formed by nucleotide deprivation on G-less cassettes, and thus with C-rich RNAs, revealed a near-linear relationship between transcript length and FttA-mediated termination (Fig. 3.2; panel b, quantified in panel d). While FttA-mediated termination is possible with only short segments of solvent accessible nascent





**Figure 3.2. FttA-mediated termination shares mechanistic requirements of rho-mediated bacterial transcription termination.** **a.** Promoter-directed transcription of biotinylated templates encoding G-less or C-less cassettes permits formation of TECs with increasing length A-, C-, and U-rich, or A-, G-, and U-rich nascent transcripts, respectively. FL = full-length; all templates permit elongation for 100 nts beyond the G- or C-less cassette. **b.** TECs remain stably associated and transcripts are primarily recovered in the pellet (P) fraction in the absence (-) of FttA. When FttA is present (+), transcripts are cleaved and primarily recovered in the supernatant (S) fraction. Cleavage releases ~20 – 30 nt shorter transcripts (boxed). The left-most lane contains <sup>32</sup>P-labeled ssDNA markers. **c.** Addition of Spt4/5 largely abrogates the RNA sequence-requirements of FttA-mediated transcription termination. T= total reaction = P+S. The left-most lane contains <sup>32</sup>P-labeled ssDNA markers. **d.** Transcript release was quantified with and without FttA addition for TECs with increasing length transcripts on G-less (pink/salmon) and C-less cassettes (mint/green), with and without Spt4/5 addition for TECs<sub>+125</sub> formed on G- and C-less cassettes. Error bars were calculated as standard deviation from the mean (n ≥ 3 replicates).

transcript sequences – a notable discriminating feature between rho- and FttA-mediated termination – the efficiency and rate of FttA-mediated termination are modest in such instances. When transcript sequences are extended to > 100 nts, the efficiency of FttA-mediated transcription termination approaches completion.

Rho-mediated termination is most efficient on C-rich transcripts<sup>21,62</sup>. Incubation of archaeal TECs formed by nucleotide deprivation on C-less cassettes, and thus with G-rich RNAs limited FttA-activity to near background levels (Figs. 3.2 and B.2). Thus, FttA-mediated termination, in the absence of any accessory factors, displays a preference for particular sequences in the nascent transcript, either preferring C-rich RNAs – as is the case for bacterial rho-mediated termination – or discriminating against transcripts that are particular G-rich. Rho-activity can be stimulated by NusG when *rut*-sites are non-optimal, and the archaeal-eukaryotic homologue of NusG, Spt5, together with its common binding partner Spt4, can likewise stimulate FttA-mediated termination when transcript sequences limit FttA-recognition or FttA-activity (Fig. 3.2; panel c, quantified in panel d). When transcripts are G-rich and normally poor substrates for FttA-mediated termination, Spt4/5 addition stimulates FttA-mediated termination to near completion. Spt4/5-stimulated FttA-mediated transcription termination and transcript cleavage still generally result in endonucleolytic cleavage of ~25 nts (Fig. 3.2; panel c, lanes 17&18).

***FttA-mediated termination and RNase activity is stimulated by interaction with the archaeal TEC.***

FttA is a known endo- and 5'-3' exonuclease<sup>48,63</sup>. When offered free RNA, FttA-mediated endonucleolytic cleavage has been mapped at CA and CC dinucleotide sequences and exonuclease activity is seen on transcripts with various 5'-moieties (Fig. B.3). Regardless of transcript length (Fig. 3.2, panels a, b & d), FttA not only mediates transcription termination, but FttA-mediated cleavage reduces transcript length by ~20-30 nts, consistent with FttA-mediated cleavage being dictated by RNAP-RNA interactions near the stalk-domain and RNA exit channel of RNAP. While FttA-mediated cleavage of nascent transcripts is complete within ~1-2 minutes, incubations of FttA with purified RNA under identical conditions require ~30-times longer to generate even mild cleavage patterns (Fig. B.3), consistent with previous *in vitro* results<sup>48,64</sup>. When FttA is incubated with RNA alone, both endo- and exonuclease activities are limited (Fig. B.3). FttA-mediated cleavage of nascent transcripts is stimulated by interactions with the archaeal TEC, but not RNAP alone (Fig. B.3). FttA appears to retain activities of the eukaryotic CPSF complex (*i.e.*, endonucleolytic transcript cleavage, albeit seemingly not dependent on a specific cleavage recognition sequence) and the Xrn2 exonuclease (*i.e.*, rapid exonucleolytic degradation of the remaining RNA 3'-fragment) in a single polypeptide that is stimulated by interactions with the TEC, but not RNAP alone. It should be noted that CPSF does retain exonuclease activity that is required for processing of histone-encoding eukaryotic transcripts<sup>57,65,66</sup>.

***FttA-mediated transcription termination is competitive with transcription elongation.***

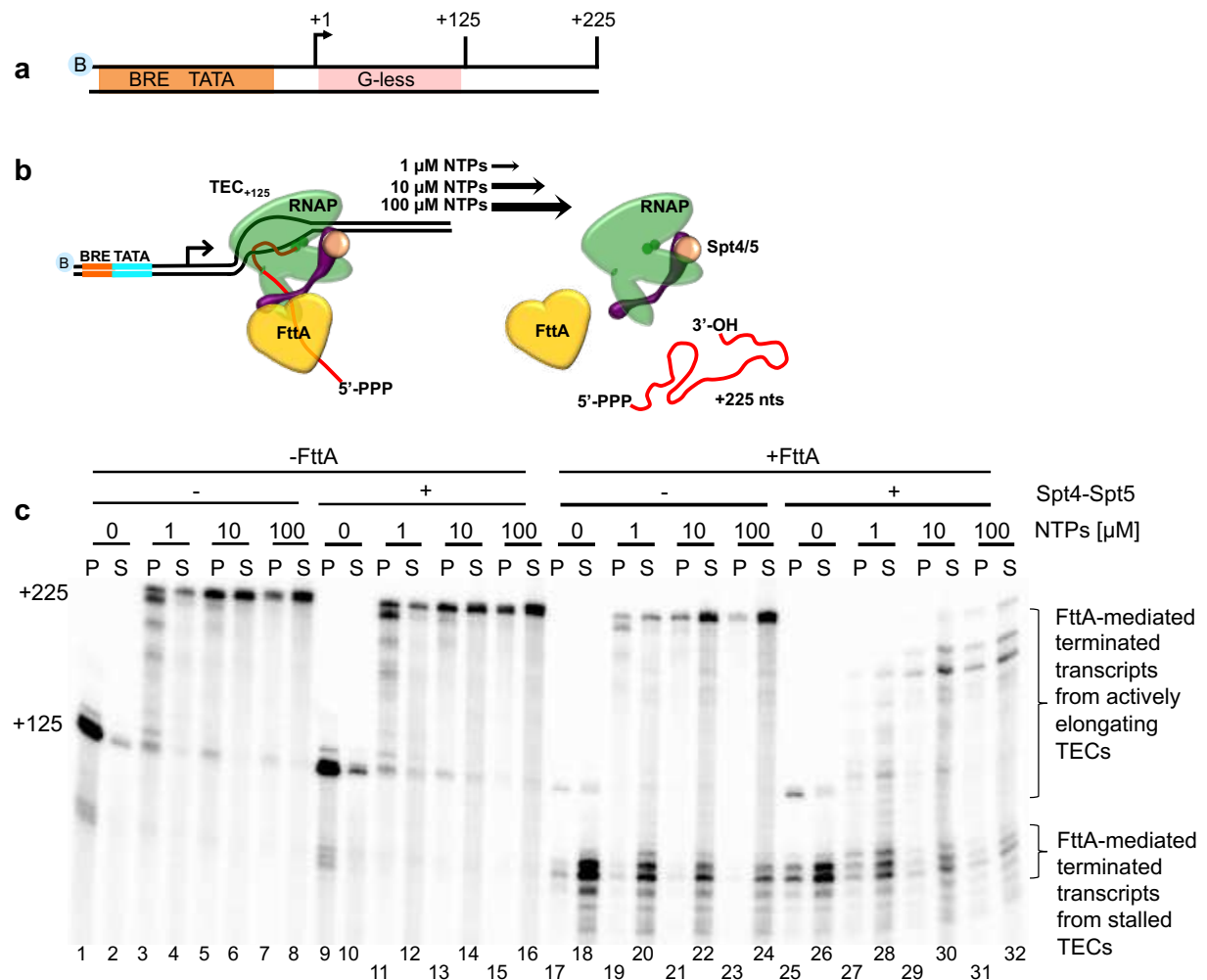
The rate of RNA synthesis by multi-subunit RNAPs (~20 – 50 nt/sec) requires transcription termination factor activity to be meaningfully competitive with continued elongation if the factor is anticipated to participate in 3'-end formation of completed transcripts. Slow acting termination factors involved in DNA repair such as Mfd (Bacteria)<sup>15,26</sup> and Eta (Euryarchaea)<sup>17</sup> are non-

competitive with elongation at even starvingly low nucleotide concentrations. We sought to determine if FttA-activity, alone or stimulated by interactions with Spt4/Spt5<sup>21,67</sup>, was sufficiently rapid enough to terminate actively transcribing TECs (Fig. 3.3).

TECs<sub>+125</sub> stalled by nucleotide deprivation readily resume elongation upon addition of NTPs (Fig. 3.3). By supplying TECs<sub>+125</sub> with a low [NTP], the rate of elongation can be limited by NTP availability, and TECs elongate slowly to the end of the template. Even after several minutes of incubation with 1  $\mu$ M NTPs, many TECs are still transcribing from their initial starting position (+125) to the end of the template (+225) as evidenced by a mixture of transcripts of various lengths between 125-225 nts (Fig. 3.3, panel b, lanes 3&4). Despite slowly elongating, the TECs are stable and retain essentially all transcripts in the pellet fraction to the exclusion of the supernatant (Fig. 3.3, panel b, lanes 3&4). By providing increasingly higher [NTP], the rate of transcription elongation can be steadily increased until TECs are elongating at rates equal to, or perhaps even greater than normal elongation rates *in vivo*. While RNAP can elongate on protein-free DNAs at ~40 nts per second, the deposition of archaeal histones on DNA reduces elongation rates substantially and the rate of synthesis on chromatinized templates likely more accurately reflects *in vivo* elongation rates<sup>52,68</sup>. When 10  $\mu$ M NTPs are provided, the vast bulk of TECs resume elongation and almost all complete synthesis of 100 additional nts within a few minutes (Fig. 3.3, panel b, lanes 5&6). When 100  $\mu$ M NTPs are provided, all TECs not only complete 100 nts of synthesis, but their extended incubation at the end of the template leads to TEC running off the end of the template and releasing most +225 nt transcripts into solution (Fig. 3.3, panel b, lanes 7&8).

Addition of increasing [NTPs] permits stalled TECs<sub>+125</sub> to resume productive elongation, albeit at different rates, and produce +225 nt transcripts with release of transcripts only observed at the end of the template (Fig. 3.3, panel b, lanes 3-8). Spt4-Spt5 addition is known to stabilize TECs

<sup>52,69–71</sup> but had no discernable effect on elongation rate or the efficiency of resumed elongation at any [NTP] (Fig. 3.3, panel b, lanes 9-16).



**Figure 3.3. FttA-mediated transcription termination is competitive with transcription elongation.** **a** and **b**. Washed, NTP-deprived TECs<sub>+125</sub> were assembled on biotinylated templates with a +125 nt G-less cassette. Resumed elongation upon differential [NTP] addition permits transcription to generate +225 nt transcripts, albeit at different rates. **c**. FttA readily terminates stalled or slowly elongating TECs (lanes 17-24) and FttA-mediated termination becomes competitive with transcription elongation even at high [NTP] in the presence of Spt4/5 (lanes 25-32).

The differential elongation rates provide a relative measure of the efficiency of FttA-mediated transcription termination in competition with transcription elongation. As anticipated, addition of FttA to stalled TECs (Fig. 3.3, panel b, lanes 17 & 18) resulted in near complete transcript release, transcript cleavage and RNAP release into solution. As [NTPs] and the rate of

transcription elongation increased, FttA-mediated termination decreased, resulting in substantial production of +225 nt full-length transcripts. Thus, alone FttA-mediated termination is not competitive with transcription elongation and FttA is only able to efficiently recognize and terminate TECs that have yet to resume elongation or are very slowly elongating (Fig. 3.3, panel b, lanes 19-24).

Although not an obligate subcomplex of archaeal RNAP, Spt4/5 engages RNAP *in vivo* quite early during elongation and remains associated with TECs throughout long genes and operons<sup>72</sup>. The ability of Spt4/5 to temper the transcript requirements for FttA-mediated termination (Fig. 3.2) suggested that addition of Spt4/5 may accelerate FttA-recognition of or action towards TECs. In support of this hypothesis, the addition of Spt4/5 greatly increased the termination efficiency of FttA, demonstrated by the release of transcripts > +125 nt but < +225 nt (Fig. 3.3, panel b, lanes 27-32). Even at high [NTPs], the combinatorial activities of FttA and Spt4/5 result in release of essentially all transcripts to solution and essentially no-full length transcripts are generated. As noted earlier, cleavage of the RNA resulted in a stable 5'-fragment, but no evidence of an intact 3'-fragment could be seen. The results demonstrate that FttA is kinetically coupled to RNAP via elongation factors Spt4/5, a striking analogy to the stimulation of the unrelated bacterial rho protein by NusG.

To ensure that FttA mediates transcript cleavage and transcription termination when combined with Spt4/5 – and that termination and cleavage in the presence of all three factors was not a new activity of Spt4/5 – we employed a variant of FttA (FttA<sup>H255A</sup>) that retains only partial activity (Fig. 3.1). FttA<sup>H255A</sup> was not capable of effective termination or RNA cleavage under identical conditions to those employed with FttA<sup>WT</sup>, although the modest RNA cleavage and termination activity of FttA<sup>H255A</sup> was increased ~3-fold upon addition of Spt4-Spt5 to FttA<sup>H255A</sup> containing reactions (Fig. B.5). Despite the increased activity of FttA<sup>H255A</sup> on stalled complexes, FttA<sup>H255A</sup>

was largely not capable of terminating actively elongating TECs in the presence of Spt4/5 with most TECs<sub>+125</sub> successfully resuming elongation and generating +225 nts transcripts.

***Kinetically competitive FttA-mediated termination is dependent on Spt4, Spt5 and the stalk-domain of RNAP.***

In Bacteria, interactions between rho and the C-terminal KOW domain of NusG stimulate rho-mediated termination<sup>18,19,21</sup>. The NusG-KOW domain is normally engaged with the ribosome and only becomes available when transcription becomes uncoupled from translation<sup>73</sup>.

Archaeal transcription and translation are coupled<sup>8</sup>, and we asked whether the isolated KOW domain of Spt5 would suffice to stimulate FttA-mediated transcription termination. Deletion of the Spt5 N-terminal domain (termed the NGN domain) likely blocks Spt5-RNAP interactions<sup>74-77</sup> which appear critical for stimulation of FttA-mediated termination, as addition of the KOW-domain (which remains thermostable) alone does not influence the activities of FttA or RNAP *in vitro* (Fig. B.5, panel b). We noted that addition of Spt5<sup>WT</sup> alone nor Spt5<sup>ΔNGN</sup> alone was sufficient to stimulate kinetically relevant FttA-mediated termination (Fig. B.5, panel b, lanes 17-32). Spt5 is often in a heterodimeric partnership with Spt4, and this partnership is critical to kinetically couple FttA activity to RNAPs, as addition of full-length Spt5 or Spt4 alone is insufficient to stimulate FttA-mediated termination to compete with elongation at high [NTPs] (Fig. B.5, panel c). The combined activities of Spt4/5 are thus necessary to maximize FttA-mediated termination and permit termination of actively transcribing TECs.

Like the nuclear eukaryotic RNAPs, the archaeal RNAP contains a stalk domain (subunits E & F)<sup>67</sup>. The stalk provides binding surfaces for conserved initiation and elongation factors and the nascent transcript<sup>1,45,78-80</sup>. Archaeal cells encoding RNAPs that fail to assemble the stalk domain of RNAP are viable but only at reduced temperatures<sup>53</sup>. Purified stalk-less RNAP (RNAP<sup>ΔE/ΔF</sup>), when combined with TBP and TFB, is competent for transcription initiation,

elongation, and intrinsic termination<sup>53</sup>, but fails to respond correctly to FttA-mediated termination (Fig. B.6). TECs<sub>+125</sub> assembled with RNAP<sup>ΔE/ΔF</sup> resumed elongation to the same extent as TECs<sub>+125</sub> assembled with RNAP<sup>WT</sup> at each [NTP] (Fig. B.6, panel b, lanes 1-8) and like RNAP<sup>WT</sup>, the elongation activities of RNAP<sup>ΔE/ΔF</sup> are largely unchanged by the addition of Spt4/5 (Fig. B.6, panel b, lanes 9-16). Despite the unchanged elongation activities of the TECs lacking the stalk domain, addition of FttA to reactions containing RNAP<sup>ΔE/ΔF</sup> did not result in extensive FttA-mediated RNA cleavage or transcription termination (Fig. B.6, panel b, lanes 17-24). In reactions with RNAP<sup>WT</sup>, the vast majority of transcripts are cleaved and released to solution, whereas reactions with RNAP<sup>ΔE/ΔF</sup> nearly eliminated FttA-mediated RNA cleavage and reduced FttA-mediated termination significantly. Even when continued elongation was prohibited, the cleavage and termination activities of FttA were stunted by the loss of the RNAP-stalk domain (Fig. B.6, panel b, lanes 17 & 18). At higher [NTPs], the dominant RNA species was not terminated nor cleaved, but rather TECs typically generated full-length +225 nt transcripts. Addition of Spt4/5 stimulates FttA-mediated RNA cleavage and FttA-mediated transcription termination of stalled TECs<sub>+125</sub> assembled with RNAP<sup>ΔE/ΔF</sup> (Fig. B.6, panel b, lanes 25 & 26), but unlike the kinetically competitive transcription termination afforded by Spt4/5 addition to RNAP<sup>WT</sup>, addition of Spt4/5 to TECs assembled with RNAP<sup>ΔE/ΔF</sup> does not stimulate FttA-mediated termination to rates that are competitive with continued elongation even at low [NTP] (Fig. B.6, panel b, lanes 27-32). The combined results suggest a cooperative role for both the stalk domain of RNAP and Spt5 – that is further stabilized by Spt4 – in accelerating FttA-mediated termination to permit kinetically competitive transcription termination *in vitro*.

### ***Inhibition of FttA-activity alters transcription termination in vivo and in vitro.***

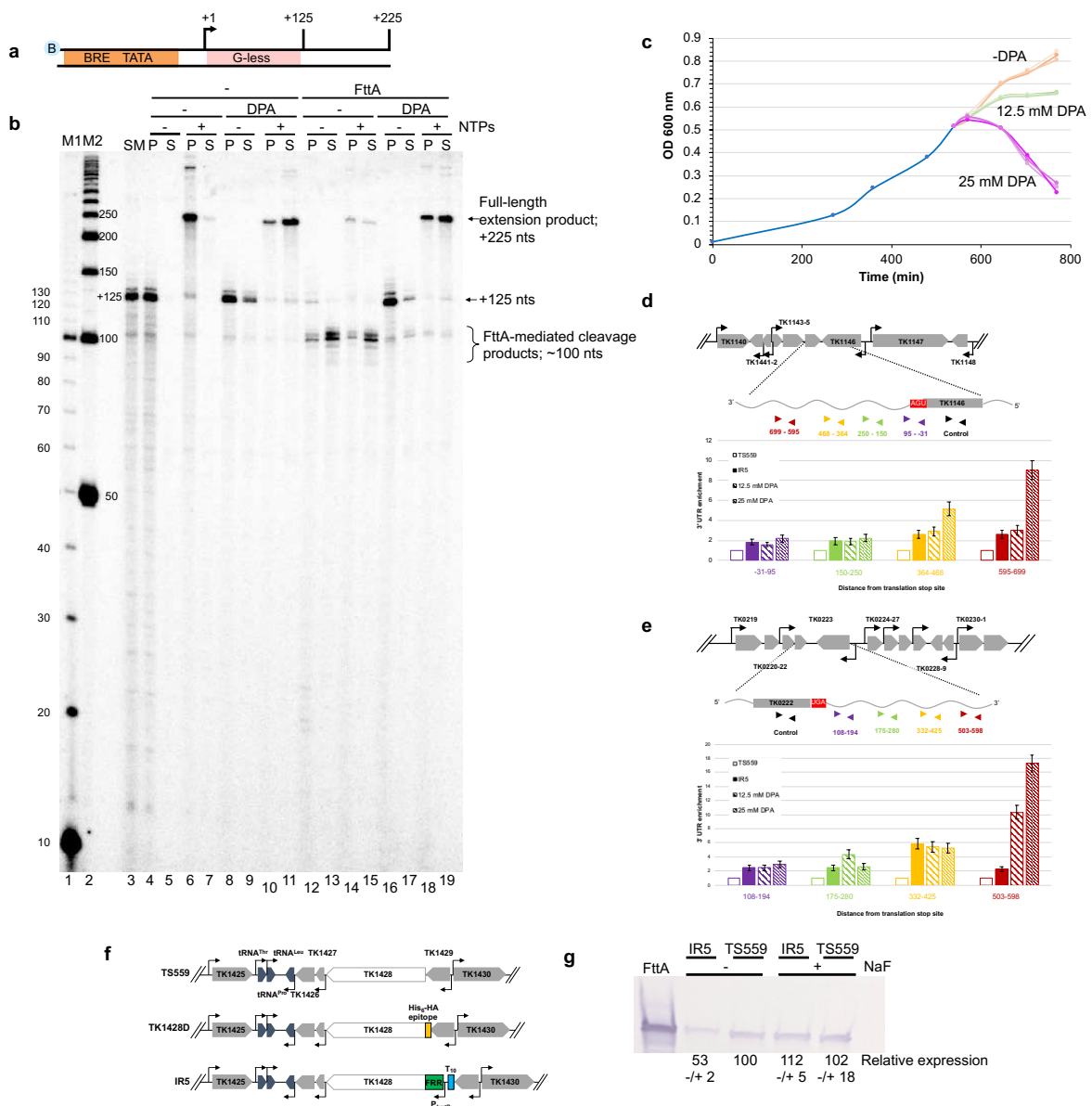
FttA is present at relatively high concentrations (~2,100 +/- 500 molecules/cell; Fig. B.7) and is therefore seemingly sufficiently abundant to monitor global transcription (RNAP is estimated to

be present in ~3,000 molecules/cell)<sup>51</sup>. FttA is a metallo-beta lactamase-fold protein containing a beta-CASP domain<sup>48</sup> and we predicted that general inhibitors of beta-CASP proteins<sup>81</sup> may impair FttA activity *in vitro* and *in vivo*. Testing of many such inhibitors revealed that pyridine carboxylates, and particularly 2,6-pyridine dicarboxylic acid (also termed dipicolinic acid, or DPA) impaired or nearly-completely inhibited FttA-mediated transcription termination *in vitro* without compromising the activities of the archaeal RNAP (Fig. 3.4, panels a and b). When TECs<sub>+125</sub> are permitted to resume elongation upon NTP addition, TECs extend nascent transcripts to produce +225 nucleotide full-length transcripts (Fig. 3.4, panel b, lanes 6&7); if TECs are exposed to DPA, elongation is unimpaired (Fig. 3.4, panel b, lanes 10&11). While FttA is able to cleave most nascent transcripts, disrupt most TECs and prevent most TECs from resuming elongation upon NTP addition in the absence of DPA (Fig. 3.4, panel b, lanes 12-15) addition of inhibitory concentrations of DPA blocks FttA-mediated termination, permitting stable TECs to resume elongation and generate full-length transcripts (Fig. 3.4, panel b, lanes 16-19). As DPA addition does not impact RNAP activity *in vitro*, these results further bolster the active center of FttA, rather than RNAP, in mediating RNA cleavage and transcription termination.

Given the effectiveness of DPA in inhibiting FttA activity *in vitro*, we challenged *T. kodakarensis* cultures with DPA and demonstrated a reduced, then complete inhibition of growth upon addition of sufficient [DPA] (Fig. 3.4, panel c). While DPA may impact the activities of several cellular factors, we rationalized that monitoring RNA 3'-ends following DPA addition may reveal changes due to inactivation of FttA *in vivo*. In support of FttA activity impacting RNA 3'-end formation *in vivo*, quantitative, reverse-transcription-PCR (q-RT-PCR) analyses following DPA addition revealed significant (up to ~17-fold) changes to 3'-ends at several loci *in vivo* (Fig. 3.4, panels d and e). As two examples, transcription of TK1146 (Fig. 3.4, panel d) and the TK0221-TK0222 bicistronic operon (Fig. 3.4, panel e) yield transcripts that can be reversed transcribed



with oligonucleotides complementary to sequences at increasing distances (95, 250, 468 and 699 nts, respectively for TK1146; 194, 280, 425 and 598 nts, respectively for TK0222) downstream of the translation stop codon. Once reverse transcribed, relative cDNA abundance – reflective of the abundance of transcripts containing extended 3'-UTR sequences – can be calculated by quantitative PCR. By normalizing  $C_t$  values from each 3'-UTR amplicon (purple, green, orange, and red) to the abundance of control amplicons (black) generated with a pair of oligonucleotide primers complementary to sequences within the TK1146 and TK0222 open



**Figure 3.4. Inhibition of FttA activity abolishes transcription termination *in vitro* and reduced FttA-expression or activity alters steady-state RNA 3'-termini *in vivo*. a and b.**

TECs<sub>+125</sub> (lane 3) assembly on biotinylated templates with G-less cassettes are stable and readily resume elongation upon NTP addition to generate +225 nt transcripts in the presence or absence of 25 mM DPA (lanes 4-11). FttA addition results in transcript cleavage, termination, and release of most TECs to the supernatant in the absence of DPA (lanes 12-13), thereby inhibiting resumed elongation upon NTP addition (lanes 14-15). Incubation of FttA with 25 mM DPA prior to addition to reactions inhibits FttA-mediated termination and RNA cleavage (lanes 16-17), permitting intact TECs<sub>+125</sub> to resume elongation and generate +225 nt transcripts (lanes 18-19). Lanes M1 and M2 contain <sup>32</sup>P-labeled 10- and 50-nt ssDNA markers, respectively. **c.** Inhibition of beta-CASP protein activity by addition of DPA impairs growth of *T. kodakarensis*. An actively growing culture of *T. kodakarensis* strain TS559 was split after reaching an optical density of 0.5 to nine cultures, with three biological replicates of cultures exposed to 0 mM DPA (peach series), 12.5 mM DPA (green series) or 25 mM DPA (purple series). **d and e.** RNAs recovered one-hour post DPA-addition to cultures of TS559, or from cultures of IR5 grown in the absence of NaF display altered 3'-termini. Trizol extracted RNAs were reverse transcribed with oligonucleotide primers complementary to nascent transcript sequences of TK1146 and TK0222 to generate cDNAs that were quantified and normalized into internal controls. Inhibiting FttA-activity with DPA or lowering steady-state FttA levels by riboswitch-mediated controlled expression impacts the abundance of RNAs with extended 3'-termini *in vivo*. RNA abundance in untreated TS559 cultures (open bars) was set to 1.0, and fold changes in the abundance of amplicons reflecting RNA transcripts with extended 3'-sequences at increasing distances from the translation stop site (purple, green, orange and red) are shown for strain IR5 (solid bars), strain TS559 treated with 12.5 mM DPA (wide stripes) and strain TS559 treated with 25 mM DPA (narrow stripes). Errors are calculated from minimally triplicate technical replicates of at least three biological replicates. **f.** Genome maps of the TK1428 locus in parental (TS559), N-terminally tagged (TK1428D) and riboswitch-regulated expression (IR5) strains of *T. kodakarensis*. **g.** Western blot demonstrating the reduction in steady-state FttA protein levels in strain IR5 upon removal of NaF from the culture medium. n ≥ 6 independent replicates.

reading frames, respectively, the fold-change in abundance of extended transcripts can be determined. Triplicate-technical replicates of biological triplicate samples reveal significant changes in the steady-state 3'-ends of transcripts from both TK1146 and TK0221-TK0222, with the fold-changes generally increasing in magnitude compared to untreated cultures both with respect to the distance from the translation stop codon and with increasing [DPA] treatments.

**Genetic perturbation of FttA-activity impacts transcription termination *in vivo*.**

While DPA treatments of whole cells likely impacts FttA activity *in vivo*, physiological effects and changes to the steady-state abundance of transcripts with altered 3'-termini stemming from exposure to a high concentration of a general beta-CASP inhibitor cannot be definitively attributed to direct inhibition of FttA activity *in vivo*. FttA is universally conserved in all archaeal

genomes<sup>47,48,82,83</sup>, including the severely reduced genomes of several symbiotic Nanoarchaeota, and it was perhaps not surprising that exhaustive attempts to delete or generate variants that radically impair activity of FttA<sup>33,57,84</sup> in *T. kodakarensis* were unsuccessful; full-deletion of the TK1428 locus, addition of sequences encoding C-terminal epitope- and affinity-tags, introduction of single amino acid variants – D364N, E389Q, H255A and H591A – and a H253A/H255A/E256A/H258A quadruple variant were all nonviable. The failure to generate strains lacking TK1428 or encoding FttA-variants suggested that TK1428 may be genetically inaccessible, supported by the essential nature of FttA homologues in related archaeal species<sup>82,83</sup>. This was not the case, as we successfully generated a markerlessly-modified<sup>85</sup> *T. kodakarensis* strain (termed TK1428D) encoded an N-terminal His<sub>6</sub>-affinity and HA-epitope (Fig. 3.4, panel f) as well as a G590T variant that results in cold-sensitive activity of the yeast homologue of FttA, Brr5/Ysh1<sup>84,86</sup>. The failure to recover strains lacking TK1428 or wherein FttA-activity was limited by introduction of mutational variants thus supports the biological essentiality of FttA for *T. kodakarensis* viability.

Gentle purification of *in vivo* assembled complexes based on retention of target proteins via the affinity tag have established interactions between replication, translation, energy-generation and transcription apparatuses<sup>17,80,87–90</sup>. Strain TK1428D growth was indistinguishable from the parental strain TS559<sup>85,91</sup> and N-terminally tagged FttA was easily recovered directly from TK1428D cell lysates in large abundance while no FttA was recovered from identical procedures from TS559 biomass (Fig. B.8). Proteins co-purifying with FttA from TK1428D were identified by MuDPIT<sup>87,90</sup>, returning only a small number of identified proteins (Fig. B.9) that have minimal inferred activity related to transcription and gene expression. No obvious stoichiometric FttA interaction partners were recovered, fully supportive on our *in vitro* demonstration that FttA alone can disrupt archaeal TECs. Affinity purification of FttA does not return RNAP subunits nor Spt4/5.

Given our desire to directly demonstrate that reduced FttA activity *in vivo* impacts transcription termination *in vivo*, coupled with our inability to generate *T. kodakarensis* strains encoding enzymatically impaired FttA variants, we next attempted to reduce FttA activity by limiting FttA expression and altering steady-state FttA<sup>WT</sup> protein levels. TK1428 is normally co-transcribed with TK1429 as the downstream gene in a bicistronic operon<sup>92</sup>. To ensure that introduced and regulated expression of TK1428 did not impact TK1429 expression, we separated TK1428 expression from TK1429 by introduction of a new promoter and intrinsic termination sequence, then placed the TK1428 coding sequences downstream of sequences encoding an archaeal fluoride-responsive riboswitch<sup>93</sup>, thereby generating strain IR5 (Fig. 3.4, panel f).

Construction of IR5 was only possible when cultures were continuously provided with fluoride even though fluoride impairs general growth of *T. kodakarensis*<sup>93</sup>, supporting that very limited expression of TK1428 was not compatible with life. Quantitative Western blots (n = 6 biological replicates) of steady-state FttA levels in IR5 strains grown in the absence and presence of fluoride reveals a modest ~2-fold change in FttA levels *in vivo* when fluoride is removed from cultures (Fig. 3.4, panel g; Coomassie stained gel shown below to demonstrate equal total protein levels were loaded in each lane). The modest change in steady-state FttA levels however does significantly and reproducibly impact transcription termination *in vivo* as determined by qRT-PCR (Fig. 3.4, panels d and e). Even a two-fold decrease in steady-state FttA levels suffices to alter the abundance of transcripts with extended 3'-termini and the magnitude of effects due to reducing FttA levels just two-fold generally mimic the impacts of 12.5 mM DPA treatment of cultures, resulting in ~2-6-fold changes in the abundance of transcripts with extended 3'-termini. The increased abundance of RNA with extended 3'-UTRs in strains with reduced FttA protein abundance is supportive of FttA normally directing transcription termination *in vivo*.

The essentiality of FttA in *T. kodakarensis* and related archaea<sup>82,83</sup>, the complete conservation of FttA in Archaea<sup>48,50,94</sup>, the demonstrated *in vitro* ability of FttA-mediated termination to compete with productive elongation (Fig. 3.3) and the demonstrated changes to RNA 3'-ends in strains wherein FttA activity is reduced by two independent mechanisms (Fig. 3.4) suggests that FttA is likely responsible for 3'-end formation of transcripts that are not directed by intrinsic termination, and further that FttA-mediated termination is likely responsible for polarity in archaeal cells<sup>60</sup>.

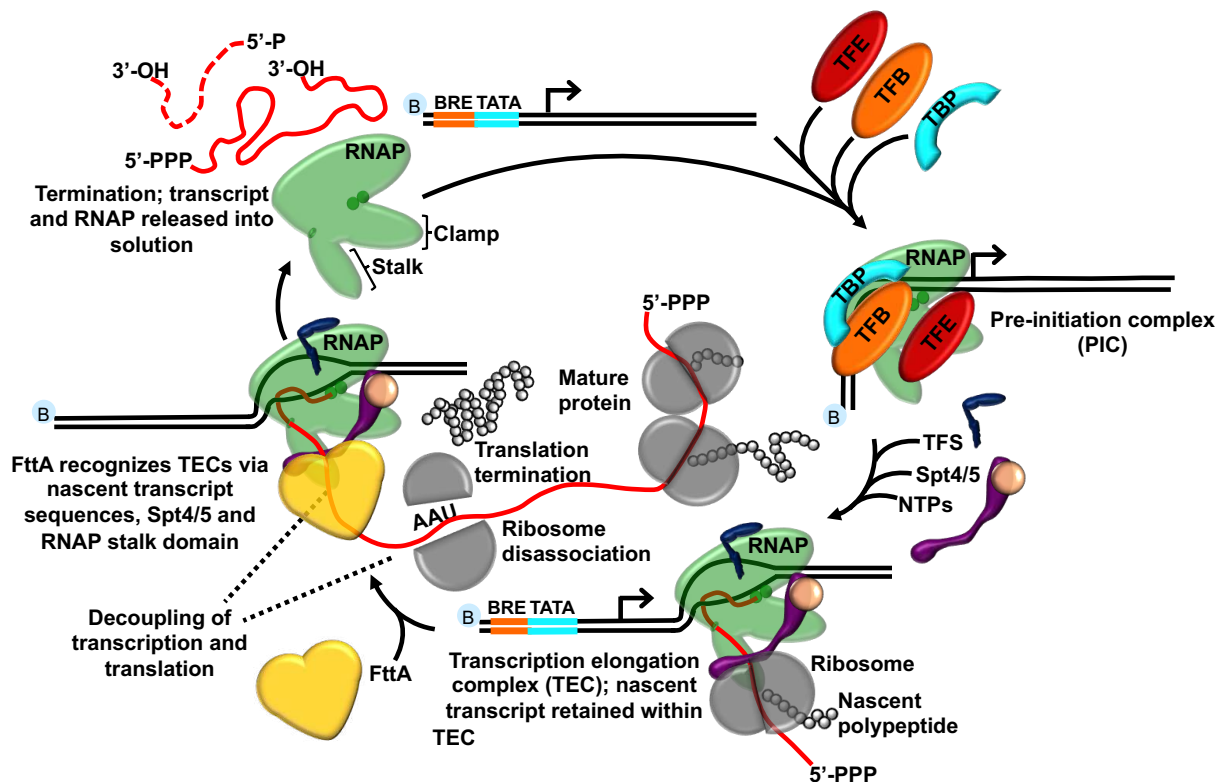
### ***The structure of FttA mimics eukaryotic CPSF73.***

FttA contains two N-terminal KH domains that presumably facilitate RNA recognition and binding<sup>64</sup>, but otherwise the structure of FttA is nearly identical to the CPSF73 subunit of the eukaryotic CPSF complex (Fig. B.10)<sup>30,31,48,49,95</sup>. The overlay of the C-terminal FttA metallo-beta-lactamase, beta-CASP containing fold and the human CPSF73 demonstrates a conservation of overall structure as well as congruence regarding the orientation and alignment of active site residues. CPSF73 is the known endonuclease of the eukaryotic CPSF complex<sup>29,32,33</sup> and can also function as an exonuclease under specific conditions<sup>57,65,66</sup>. While FttA and CPSF73 share an endonucleolytic cleavage activity, CPSF73 activity is limited to specific sequences whereas FttA cleavage is apparently dictated not solely by nascent transcript sequences, but rather proximity to the TEC, with FttA-mediated RNA cleavage generally occurring near the first point of solvent accessible RNA that extends from the TEC into the FttA active center (Fig. 3.2). FttA is also a known exonuclease<sup>48,63,64</sup>, a function seemingly not shared with CPSF73 under most conditions<sup>57,65,66</sup>. Exonucleolytic degradation of the 3'-transcript associated with archaeal TECs appears FttA-mediated, whereas the activity of Xrn2/Rat1<sup>34,38-40</sup> appears responsible and necessary for proper degradation of transcripts and termination of eukaryotic TECs formed with Pol I and Pol II<sup>35,96-98</sup>.

## Discussion

We demonstrate here that the universally conserved archaeal homologue of the eukaryotic CPSF73 protein is the factor that terminates transcription in Archaea (FttA). Purified FttA (Fig. S2) functions as the second *bona fide* archaeal-encoded termination factor, cleaving nascent transcripts, disrupting TECs and recycling RNAP for additional rounds of synthesis (Figs. 3.1 and B.1). Although previously termed aCPSF or aCPSF1 (for archaeal-CPSF)<sup>48,64</sup>, archaeal transcripts are not polyadenylated and thus the retention of the designation aCPSF is inappropriate; FttA more accurately describes the activity of this universally conserved archaeal protein. FttA recognizes TECs through the nascent RNA, favoring C- and U-rich RNAs and displaying maximal activity when nascent transcripts are >~100 nts (Figs. 3.2 and B.2). FttA-mediated termination is kinetically competitive with continued elongation *in vitro* (Fig. 3.3) and by establishing the requirements for FttA-mediated transcription termination (Figs. 3.1-3.4, B.2 and B.4-B.7) we complete the archaeal transcription cycle and describe a new mechanism of 3' end formation (Fig. 3.5). It is important to note that the described activities of FttA suggest that the steady-state 3'-termini of *in vivo* transcripts terminated by FttA do not reflect that actual position of termination of the archaeal RNAP.

Thus, consensus termination sequences derived from next-generation sequencing and Term-seq data<sup>1,99</sup> should be re-evaluated given that FttA-terminated transcripts are likely to be lacking minimally ~20-30 nts from the 3'-terminus; additional RNA processing events are likely to further complicate attempts to map the 3'-termini of transcripts that reflect the true position of TEC dissociation. Even transcripts derived from loci encoding putative intrinsic termination sequences should be reevaluated, as FttA-activity may influence the efficiency of intrinsic termination or serve as a backup mechanism of transcription termination for genes/operons with less-efficient intrinsic termination signals.



**Figure 3.5. FttA-mediated transcription termination completes the archaeal transcription cycle.** Promoter-directed assembly of pre-initiation complexes requires RNAP, TFB and TBP and is often assisted by TFE. *De novo* RNA synthesis permits promoter escape and transcription initiation factors are replaced by transcription elongation factors TFS and Spt4/5. The absence of a nuclear compartment permits translation initiation and the normal coupling of the archaeal transcription and translation apparatuses throughout transcription of the gene or operon, but this coupling is disrupted by translation termination. The exposed nascent transcript likely permits loading of FttA to TECs and FttA activity mediates cleavage of nascent transcripts and release of RNAP to solution.

FttA alone can disrupt TECs in an energy-independent manner (Fig. 3.1). Although abundant (Fig. B.7), no obvious *in vivo* binding partners of FttA could be identified by gentle-purification of FttA directly from cell lysates (Figs. B.6-B.8). Although FttA can function independently, the efficiency of FttA-mediated termination is increased when TECs are bound by Spt4/5 and the nascent transcript is C-rich (Figs. 3.1-3.3, B.2 and B.4-B.7). FttA recognizes TECs using information gained from the stalk domain of RNAP, the length and sequence composition of the nascent transcript and presumptive interactions with the KOW domain of Spt5 in an Spt4/5 complex. FttA possesses RNase activity<sup>48,64</sup> and FttA-mediated RNA cleavage and transcription

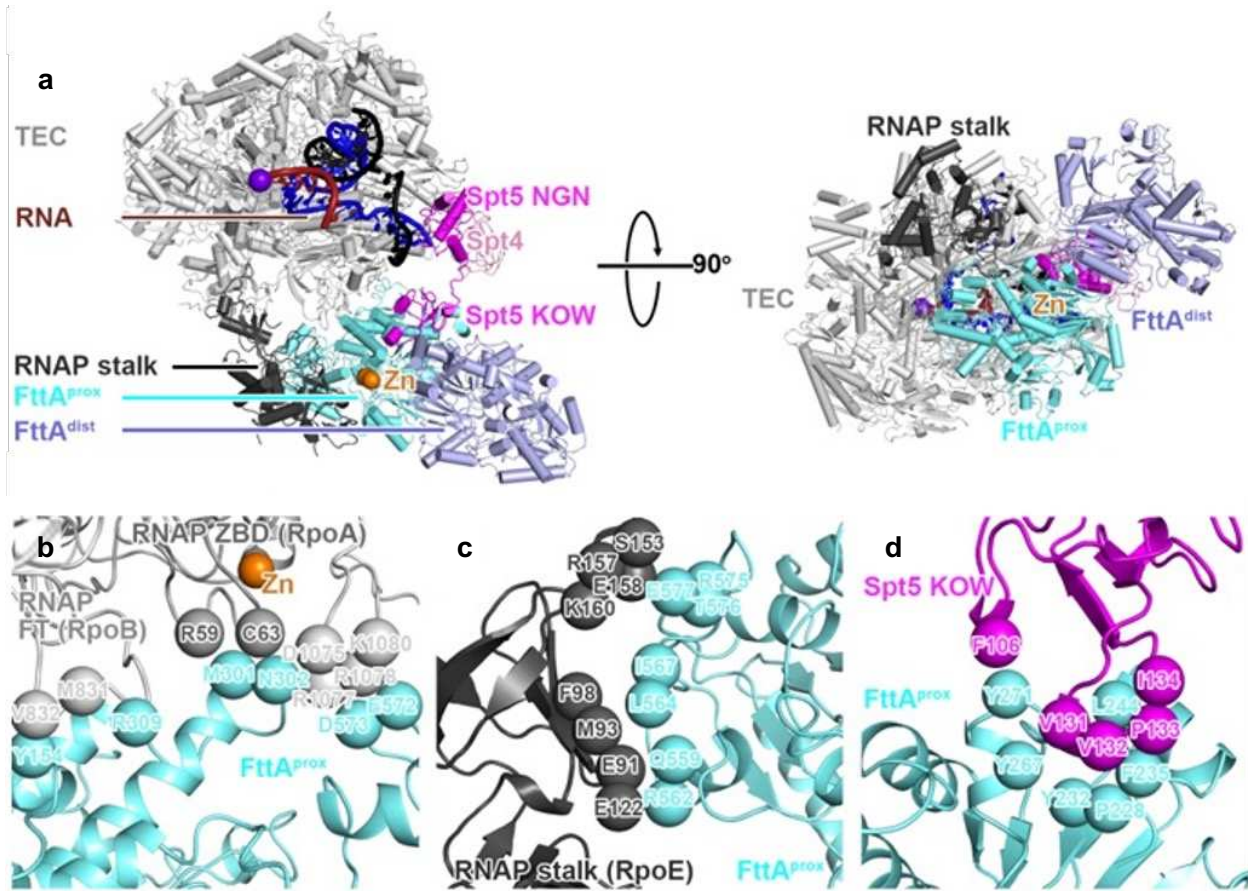
termination are dramatically stimulated by interactions with the TEC but not RNAP alone (Fig. B.3), suggesting that endo- and exo-nuclease activities of FttA are stimulated by likely subtle rearrangements to the C-terminal metallo-beta-lactamase, beta-CASP containing fold of FttA that shares a remarkable degree of structural similarity to the eukaryotic CPSF73 subunit (Fig. B.10). As predicted from the *in vitro* data (Figs. 3.3, B.5, and B.6) recent Cryo-EM of the pre-termination complex (Fig. 3.6) validates direct interactions between FttA, the stalk domain of RNAP (Fig 3.6a, b, & c), and the KOW domain of Spt5 (Fig. 3.6a and d). In addition, FttA forms a dimer within the pre-termination complex, denoted FttA<sup>prox</sup> (proximal to the TEC) and FttA<sup>dist</sup> (distal to the TEC), and the structure highlights an FttA dimer at the RNA exit channel of RNAP (Fig. 3.6a). To prevent FttA-mediated termination of the complex, the structure was formed with the FttA<sup>H255A</sup> mutant (Fig. 3.1; panel a, lanes 7&8). Interestingly, the structure lacked termination activity but retained the ability to perform endonuclease cleavage, noted by the presence of RNA within RNAP and the lack of RNA in the exit channel (Fig 3.6a)<sup>99</sup>.

Altering FttA expression, or the addition of general beta-CASP protein small molecule inhibitors impacts the abundance of RNA species with extended 3'-termini *in vivo*, fully supportive of FttA-mediated termination being an active and important mechanism of termination in archaeal cells. The universal conservation of FttA in all Archaea likely extends our findings in *T. kodakarensis* to the entire archaeal domain. The ability to drive transcription termination alone (Figs. 3.2 and 3.3), but to function more efficiently when interacting with TECs containing Spt5 and Spt4 – and particularly the requirement for the KOW-domain of Spt5 for increased FttA-mediated termination (Fig. 3.4) – argues that FttA functions analogously to rho. FttA and Rho share no obvious homology, either at the primary sequence level or structure level, and while rho is known to function as a hexamer<sup>19</sup>, FttA likely functions as a monomer or dimer<sup>48,64</sup>. FttA and rho share similar but non-identical requirements for length and nucleotide composition of nascent transcripts and given that polarity-mediated repression of downstream genes is efficient



in *T. kodakarensis*<sup>60</sup>, and possibly all Archaea, it is likely that like rho, FttA serves as the governance factor for Archaea. Archaeal transcription and translation are coupled<sup>8</sup> and disrupting the normally tight coupling of the transcription and translation apparatuses through dissociation of the ribosome at a normal stop codon would permit continued transcription to expose sufficient nascent RNA sequences to activate FttA-mediated termination before the archaeal RNAP transcribed significantly into the next gene or operon (Fig. 3.5). Further, if the archaeal TEC is linked to the archaeal ribosome as bacterial TECs and bacterial ribosomes are linked – via an interaction between the KOW domain of NusG/Spt5 and NusE/S10<sup>73</sup> – it is plausible that archaeal TECs with coupled trailing ribosomes are protected from FttA-mediated termination until the translation apparatus uncouples from the TEC.

The requirements for FttA-mediated termination suggest that long 5' UTRs observed for some archaeal transcripts may serve as points of regulation for premature termination upstream of the coding sequences to provide regulatory potential for the downstream gene or operon<sup>92,100</sup>. How transcription of stable RNAs, including rRNAs is protected from FttA-mediated termination will be of interest to determine. Exclusion of Spt4/5 from TECs transcribing stable RNAs, or structures within the nascent transcript may suffice to hinder FttA-loading or FttA-mediated termination of archaeal TECs. While Spt4/5 can be considered general elongation factors and evidence of co-occupancy of Spt5 and RNAP on most genes supports immediate recruitment of Spt4/5 to TECs, a second and delayed mechanism of Spt5 recruitment to rRNA and CRISPR loci has been identified<sup>101</sup>. It is possible that this delayed recruitment of Spt4/5 permits RNA processing machinery to engage the nascent transcript before FttA to permit continued elongation in the absence of translation. It is interesting that full-length FttA homologues are retained in the genomes of several bacterial species, suggesting that FttA may function as a termination factor in multiple domains (Fig. B.11). Bacterial FttA proteins are restricted to a few species, and although initially tempting to envision that FttA would be restricted to bacteria



**Figure 3.6. Cryo-EM structure of the FttA-modulated pre-termination complex.** **a.** The 3.9 Å resolution structure of the pre-termination complex detailing the positions and interactions between the TEC, Spt4, Spt5, and FttA dimers (left). Orthogonal rotation of the structure highlights FttA dimers interacting with the mouth of the RNA exit channel (right). **b.** FttA<sub>prox</sub> contacts RpoA, RpoB, **c.** RpoE, and **d.** the KOW domain of Spt5<sup>99</sup>.

that lack rho, an obvious correlation is not observed<sup>42</sup>. It will be of immediate interest to determine if the bacterial FttA proteins can direct transcription termination, and if they can, whether they cooperate with or can substitute for rho to block aberrant transcription and generate 3' ends. It will be similarly interesting to determine if FttA activity can disrupt eukaryotic TECs formed with Pol I, II and III.

CPSF73, as part of the RNA 3'-maturation machinery, initiates cleavage of the nascent transcript, a necessary step for mRNA maturation that in-of-itself does not direct transcription termination. The combined activities of CPSF and Xrn2 are necessary for normal termination

patterns in Eukarya<sup>35,40</sup>. FttA retains all the necessary activities within a single protein: FttA can bind TECs, mediate cleavage and release of the nascent transcript and use 5'-3' exonuclease activities to degrade the 3'-transcript. We propose that the eukaryotic CPSF complex<sup>29</sup>, which minimally contains 4 homologous but non-identical subunits, arose from archaeal FttA. The ability of the CPSF complex to directly terminate transcription was likely lost during the specialization and partnership with factors that direct RNA 3'-maturation in Eukarya.

## Materials and Methods

***T. kodakarensis* culturing conditions.** *T. kodakarensis* strain TS559 and derivatives of such were grown at 85°C under anaerobic conditions as previously described<sup>1</sup>. DPA (Sigma) was added at neutral pH to either 12.5 or 25 mM as shown in Fig. 3.4. NaF was added to 4 mM when necessary.

**Protein purifications.** Archaeal RNA polymerases (WT and  $\Delta E/F$  variant) containing His<sub>6</sub>-RpoL subunits, TBP, and TFB were purified as previously described<sup>2,3</sup>. *T. kodakarensis* Spt5 and His<sub>6</sub>-Spt4 were purified as previously described<sup>4</sup>. Spt5 <sup>$\Delta$ NGN</sup> was purified as was full-length Spt5. WT and an H255A variant of FttA were purified from Rosetta2 *E. coli* cells carrying pQE-80L (Qiagen) expression vectors carrying the wildtype or variant TK1428 coding sequence (Fig. B.2). Cells were grown in LB medium at 37°C with shaking (~220 rpm) with 30 µg/ml chloramphenicol and 100 µg/ml ampicillin to an optical density at 600 nm of 0.5 before expression was induced with 0.5 mM isopropyl  $\beta$ -D-1-thiogalactopyranoside. Cultures were grown for an additional 3 h at 37°C with shaking before biomass was harvested via centrifugation (~8,000 x *g*, 20 min, 4°C), resuspended and lysed via sonication (3 ml/g of biomass) in 20 mM Tris-HCl pH 8.0, 5 mM 2-mercaptoethanol, 10 mM MgCl<sub>2</sub>, 100 mM NaCl. Cellular lysates were clarified by centrifugation (~20,000 x *g*, 20 min, 4°C), heated to 85°C for 30 min to denature most host proteins, and clarified again by centrifugation (~20,000 x *g*, 20 min, 4°C). Heat-treated clarified cell lysates were resolved through a 5 ml HiTrap-heparin column (GE Healthcare) with a linear gradient from 0.1 – 1.0 M NaCl dissolved in 20 mM Tris-HCl pH 8.0, 5 mM 2-mercaptoethanol, 10 mM MgCl<sub>2</sub>. Fractions containing > 95% pure FttA were identified by SDS-PAGE, pooled, and dialyzed into 25 mM Tris-HCl pH 8.0, 100 mM KCl, 10 mM 2-mercaptoethanol, 50% glycerol before storage at -80°C. All protein concentrations were quantified using a Bradford Assay<sup>5</sup>.

**DNA templates.** Double-stranded DNA templates used in all transcription reactions were PCR amplified from plasmids and gel purified as previously described<sup>2,4,6</sup>. All transcription templates contain a non-template 5'-strand biotin-TEG moiety to provide attachment to streptavidin-coated paramagnetic beads (Promega).

**In vitro transcription assays.** Assembly of preinitiation complexes (PICs) and elongation via NTP deprivation was carried out as described previously<sup>2,4,6</sup>. To obtain stalled TECS on G-less cassette templates, PICs were assembled using 10 nM template, 20 nM RNAP, 40 nM TBP, 40 nM TFB in a 20 µl total volume of transcription buffer (20 mM Tris-HCl pH 8.0, 250 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM DTT) with 75 µM ApC for 3 min at 85°C before addition of 200 µM ATP, 200 µM CTP, 10 µM UTP and 10 µCi [α-<sup>32</sup>P]-UTP for 3 additional min at 85°C, then chilled to 4°C. To obtain stalled TECs on C-less cassette templates, reactions were identical to those above, with the substitution of 200 µM GTP for 200 µM CTP. RNAP bound templates were captured with HisPur™ Ni-NTA magnetic particles (Thermo Fisher Scientific) and washed three times with 100 µl 20 mM Tris-HCl pH 8.0, 1 mM EDTA, 500 mM KCl.

For Fig. 3.1, panel A, washed TECs were resuspended in 10 mM Tris-HCl pH 8.0, 125 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM DTT, with 10 µM each of ATP, CTP, and UTP before addition of 1 µM FttA or FttA<sup>H255A</sup> for 5 min at 85°C. Reactions were chilled to 4°C followed by separation of pellet and supernatant fractions by addition of streptavidin coated paramagnetic particles (Promega). For Fig. 3.1, panel b, washed TECs (lane 1) were resuspended in 10 mM Tris-HCl pH 8.0, 125 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM DTT with 10 µM each of ATP, CTP, and UTP and were incubated at 85°C for 7 min (lane 2) before being chilled on ice, bound to streptavidin-coated paramagnetic beads and washed with 100 µl 20 mM Tris-HCl pH 8.0, 1 mM EDTA, 500 mM KCl (lane 3). Washed TECs were incubated at 85°C for 1 minute before addition of 100 µM NTPs (lane 4) or 100 µM ATP, CTP and GTP, 10µM UTP containing 1µC <sup>32</sup>P-α-UTP (lane 5) and continued incubation at 85°C for 3 minutes. Washed TECs were exposed to 1µM FttA (lane 6) at 85°C for

7 minutes before being chilled on ice, bound to streptavidin-coated paramagnetic beads and washed with 100  $\mu$ l 20 mM Tris-HCl pH 8.0, 1 mM EDTA, 500 mM KCl (lane 7). FttA-treated, washed TECs were incubated at 85°C for 1 minute before addition of 100  $\mu$ M NTPs (lane 8) or 100  $\mu$ M ATP, CTP and GTP, 10 $\mu$ M UTP containing 1 $\mu$ C <sup>32</sup>P- $\alpha$ -UTP (lane 9) and continued incubation at 85°C for 3 minutes. Washed TECs were exposed to 50U RNase I<sub>f</sub> (lane 10) at 37°C for 7 minutes before being chilled on ice, bound to streptavidin-coated paramagnetic beads and washed with 100  $\mu$ l 20 mM Tris-HCl pH 8.0, 1 mM EDTA, 500 mM KCl (lane 11). RNase I<sub>f</sub>-treated, washed TECs were incubated at 85°C for 1 minute before addition of 100  $\mu$ M NTPs (lane 12) or 100  $\mu$ M ATP, CTP and GTP, 10 $\mu$ M UTP containing 1 $\mu$ C <sup>32</sup>P- $\alpha$ -UTP (lane 13) and continued incubation at 85°C for 3 minutes. For Fig. 3.1, panel f, washed TECs were resuspended in 10 mM Tris-HCl pH 8.0, 125 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM DTT, with or without 10  $\mu$ M each of ATP, CTP, and UTP before addition of 1  $\mu$ M FttA at 85°C. Reaction aliquots were removed after 1, 2, or 5 minutes, chilled to 4°C, then pellet and supernatant fractions were separated by addition of streptavidin coated paramagnetic particles (Promega).

For Figs. 3.2, panels b and c, and B, washed TECs were assembled as above on G-less or C-less cassettes of various lengths. Washed TECs were resuspended in 10 mM Tris-HCl pH 8.0, 125 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM DTT, with or without 1  $\mu$ M FttA at 85°C for 3 minutes. Reactions were chilled to 4°C, then pellet and supernatant fractions were separated by addition of streptavidin coated paramagnetic particles (Promega).

For Figs. 3.3, B.5-B.7, washed TECs were assembled with WT or  $\Delta$ E/F RNAP as above on a +125 G-less cassette template. Continued elongation was permitted by the addition of 0, 1, 10 or 100  $\mu$ M NTPs in the presence or absence of combinations of 6  $\mu$ M Spt4, 6  $\mu$ M Spt5, 6  $\mu$ M Spt5 <sup>$\Delta$ NGN</sup>, 1  $\mu$ M FttA or FttA<sup>H255A</sup>. After 5 min at 85°C, reactions were chilled to 4°C followed by

separation of pellet and supernatant fractions by addition of streptavidin coated paramagnetic particles (Promega).

For Fig. 3.4, stalled TECs on a G-less cassette template assembled using 10 nM template, 20 nM RNAP, 40 nM TBP, 40 nM TFB in a 20  $\mu$ l total volume of transcription buffer (20 mM Tris-HCl pH 8.0, 250 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM DTT) with 75  $\mu$ M ApC for 3 min at 85°C before addition of 200  $\mu$ M ATP, 200  $\mu$ M CTP, 10  $\mu$ M UTP and 10  $\mu$ Ci [ $\alpha$ -<sup>32</sup>P]-UTP and incubation for 3 additional min at 85°C, then chilled to 4°C. RNAP bound templates were captured with HisPur<sup>TM</sup> Ni-NTA magnetic particles (Thermo Fisher Scientific) and washed three times with 100  $\mu$ l 20 mM Tris-HCl pH 8.0, 1 mM EDTA, 500 mM KCl. Washed TECs were resuspended in 10 mM Tris-HCl pH 8.0, 125 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM DTT, with 10  $\mu$ M each of ATP, CTP, and UTP before addition of reaction buffer (10 mM Tris-HCl pH 8.0, 120 mM KCl, 8 mM DTT, and 1.25 mM MgCl<sub>2</sub>) +/- 25 mM DPA and/or +/- 1  $\mu$ M FttA at 85°C. Reaction aliquots were incubated for 3 minutes, and then chased with 250  $\mu$ M ATP, CTP, UTP, and GTP for 2 minutes to allow for elongation to +225. Reactions were chilled to 4°C, then pellet and supernatant fractions were separated by addition of streptavidin coated paramagnetic particles (Promega).

Radiolabeled transcripts from Figs. 3.1-3.4, B.2, B.4-B.7 were recovered by addition of 5 volumes STOP buffer (600 mM Tris-HCl pH 8.0, 30 mM EDTA) containing 7  $\mu$ g of tRNA (total), equal volume phenol/chloroform/isoamyl alcohol (25:24:1, V/V) extractions, and precipitations of the aqueous phase with 2.6 volumes 100% ethanol. Precipitated transcripts were resuspended in 95% formamide, 1X TBE, heated to 99°C for 5 minutes, rapidly chilled on ice, loaded and resolved in 10-20% polyacrylamide, 8M urea, 1X TBE denaturing gels. Radiolabeled RNA was detected using phosphorimaging (GE Healthcare). Gel images were analyzed using GE Imagequant 5.2 software.

To generate +125 nt [ $\alpha$ - $^{32}$ P]-UTP labeled transcripts used in Fig. B.3, transcription reactions were assembled and terminated as above, with the substitution of 10  $\mu$ g of glycogen for 7  $\mu$ g of tRNA during reaction clean up. Radiolabeled transcripts were incubated at 85°C with or without 1  $\mu$ M FttA or at 37°C with 5U RNase A (Thermo Fisher Scientific) in *T. kodakarensis* transcription buffer (20 mM Tris-HCl pH 8.0, 250 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM DTT) to monitor cleavage. When present, *T. kodakarensis* RNAP was added to 40 nM final.

**Western blot analysis of RNAP release to solution** (Fig. 3.1, panel e). Anti-HA antibodies (BioLegend 901513) were employed as previously described<sup>1,2,7</sup> to quantify RpoL-levels in P and S fractions.

**Western blot analysis of FttA protein levels** (Fig. 3.4, panel e, and Figs. B.7-B.9). Purified, recombinant full-length FttA was used as an antigen to prepare polyclonal antibodies in mice (Cocalico Biologicals). Known quantities of purified FttA were resolved as comparative quantification standards in adjacent lanes to clarified cell lysates derived from known quantities of cells. Proteins were separated via SDS-PAGE, transferred to PVDF membranes, and probed with primary anti-FttA antibodies. Blots were developed by addition of an IgG-AP conjugated anti-mouse secondary antibody allowed for detection by NBT/BCIP (Roche). A linear regression FttA signal intensity to FttA amount in ng was generated.

**Construction of strains TK1428D and IR5.** *T. kodakarensis* strains used here were constructed from the parental strains TS559<sup>8</sup> as previously described<sup>1</sup>.

**Purification of FttA directly from lysates of strain TK1428D and MuDPIT analysis.**

Purification procedures and MuDPIT analysis was performed as previously described<sup>7,9,10</sup>.

**TRIzol-based RNA purifications from *T. kodakarensis* cultures and q-RT-PCR.** RNA extractions were performed essentially as previously described<sup>11</sup> from strain TS559 prior to, or 1



hour after DPA addition. RNA extractions from IR5 and TS559, in the absence and presence of 4 mM NaF were performed as previously described<sup>11</sup>. qRT-PCR reactions were performed as previously described<sup>12</sup>, except that 500 ng of total RNA was used during cDNA synthesis.

## REFERENCES

1. Maier, L.-K. & Marchfelder, A. It's all about the T: transcription termination in archaea. *Biochem. Soc. Trans.* **47**, 461–468 (2019).
2. Santangelo, T. J. & Reeve, J. N. Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *J. Mol. Biol.* **355**, 196–210 (2006).
3. Sela, I., Wolf, Y. I. & Koonin, E. V. Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11399–11407 (2016).
4. Helmrich, A., Ballarino, M., Nudler, E. & Tora, L. Transcription-replication encounters, consequences and genomic instability. *Nat. Struct. Mol. Biol.* **20**, 412–418 (2013).
5. Washburn, R. S. & Gottesman, M. E. Transcription termination maintains chromosome integrity. *Proc. Natl. Acad. Sci.* **108**, 792–797 (2011).
6. Shin, J.-H., Santangelo, T. J., Xie, Y., Reeve, J. N. & Kelman, Z. Archaeal minichromosome maintenance (MCM) helicase can unwind DNA bound by archaeal histones and transcription factors. *J. Biol. Chem.* **282**, 4908–15 (2007).
7. Miller, O. L., Hamkalo, B. A. & Thomas, C. A. Visualization of Bacterial Genes in Action. *Science (80-. )*. **169**, 392–395 (1970).
8. French, S. L., Santangelo, T. J., Beyer, A. L. & Reeve, J. N. Transcription and translation are coupled in Archaea. *Mol. Biol. Evol.* **24**, 893–5 (2007).
9. Ray-Soni, A., Bellecourt, M. J. & Landick, R. Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annu. Rev. Biochem.* **85**, 319–347 (2016).
10. Roberts, J. W. Mechanisms of Bacterial Transcription Termination. *J. Mol. Biol.* (2019). doi:10.1016/j.jmb.2019.04.003
11. Santangelo, T. J., Cubonová, L., Skinner, K. M. & Reeve, J. N. Archaeal intrinsic transcription termination in vivo. *J. Bacteriol.* **191**, 7102–8 (2009).
12. Mishra, S. & Maraia, R. J. RNA polymerase III subunits C37/53 modulate rU:dA hybrid 3' end dynamics during transcription termination. *Nucleic Acids Res.* **47**, 310–327 (2019).
13. Libri, D. Endless Quarrels at the End of Genes. *Mol. Cell* **60**, 192–194 (2015).
14. Selby, C. P. Mfd Protein and Transcription-Repair Coupling in *Escherichia coli*. *Photochem. Photobiol.* **93**, 280–295 (2017).
15. Le, T. T. *et al.* Mfd Dynamically Regulates Transcription via a Release and Catch-Up Mechanism. *Cell* **172**, 344–357.e15 (2018).
16. Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **29**, 948–60 (2015).
17. Walker, J. E., Luyties, O. & Santangelo, T. J. Factor-dependent archaeal transcription termination. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6767–E6773 (2017).
18. Lawson, M. R. *et al.* Mechanism for the Regulated Control of Bacterial Transcription Termination by a Universal Adaptor Protein. *Mol. Cell* **71**, 911–922.e4 (2018).
19. Mitra, P., Ghosh, G., Hafeezunnisa, M. & Sen, R. Rho Protein: Roles and Mechanisms. *Annu. Rev. Microbiol.* **71**, 687–709 (2017).
20. Roberts, J. W. Termination Factor for RNA Synthesis. *Nature* **224**, 1168–1174 (1969).
21. Lawson, M. R. & Berger, J. M. Tuning the sequence specificity of a transcription terminator. *Curr. Genet.* (2019). doi:10.1007/s00294-019-00939-1
22. Cardinale, C. J. *et al.* Termination Factor Rho and Its Cofactors NusA and NusG Silence Foreign DNA in *E. coli*. *Science (80-. )*. **320**, 935–938 (2008).
23. Shashni, R., Qayyum, M. Z., Vishalini, V., Dey, D. & Sen, R. Redundancy of primary RNA-binding functions of the bacterial transcription terminator Rho. *Nucleic Acids Res.* **42**, 9677–90 (2014).
24. Valabhoju, V., Agrawal, S. & Sen, R. Molecular basis of NusG-mediated regulation of

- Rho-dependent transcription termination in bacteria. *J. Biol. Chem.* **291**, 22386–22403 (2016).
25. Peters, J. M. *et al.* Rho and NusG suppress pervasive antisense transcription in *Escherichia coli*. *Genes Dev.* **26**, 2621–2633 (2012).
  26. Park, J. S., Marr, M. T. & Roberts, J. W. E. coli transcription repair coupling factor (Mfd protein) rescues arrested complexes by promoting forward translocation. *Cell* **109**, 757–767 (2002).
  27. Smith, A. J. & Savery, N. J. Effects of the bacterial transcription-repair coupling factor during transcription of DNA containing non-bulky lesions. *DNA Repair (Amst)*. **7**, 1670–9 (2008).
  28. Adebali, O., Chiou, Y.-Y., Hu, J., Sancar, A. & Selby, C. P. Genome-wide transcription-coupled repair in *Escherichia coli* is mediated by the Mfd translocase. *Proc. Natl. Acad. Sci. U. S. A.* 201700230 (2017). doi:10.1073/pnas.1700230114
  29. Hill, C. H. *et al.* Activation of the Endonuclease that Defines mRNA 3' Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex. *Mol. Cell* **73**, 1217–1231.e11 (2019).
  30. Sun, Y. *et al.* Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1419–E1428 (2018).
  31. Mir-Montazeri, B. *et al.* Crystal structure of a dimeric archaeal Cleavage and Polyadenylation Specificity Factor. *J. Struct. Biol.* **173**, 191–195 (2011).
  32. Mandel, C. R. *et al.* Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* **444**, 953–6 (2006).
  33. Ryan, K., Calvo, O. & Manley, J. L. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. *RNA* **10**, 565–573 (2004).
  34. Dengl, S. & Cramer, P. *Torpedo* Nuclease Rat1 Is Insufficient to Terminate RNA Polymerase II *in Vitro*. *J. Biol. Chem.* **284**, 21270–21279 (2009).
  35. Proudfoot, N. J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352**, aad9926 (2016).
  36. Baejen, C. *et al.* Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes. *Mol. Cell* **66**, 38–49.e6 (2017).
  37. Schaughency, P., Merran, J. & Corden, J. L. Genome-Wide Mapping of Yeast RNA Polymerase II Termination. *PLoS Genet.* **10**, e1004632 (2014).
  38. Fong, N. *et al.* Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Mol. Cell* **60**, 256–67 (2015).
  39. Eaton, J. D. *et al.* Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Genes Dev.* **32**, (2018).
  40. Santangelo, T. J. *et al.* Polarity in archaeal operon transcription in *Thermococcus kodakaraensis*. *J. Bacteriol.* **190**, 2244–8 (2008).
  41. D'Heygere, F., Rabhi, M. & Boudvillain, M. Phyletic distribution and conservation of the bacterial transcription termination factor Rho. *Microbiology* **159**, 1423–1436 (2013).
  42. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nature Reviews Microbiology* **15**, 711–723 (2017).
  43. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
  44. Gehring, A. M., Walker, J. E. & Santangelo, T. J. Transcription Regulation in Archaea. *J. Bacteriol.* **198**, 1906–1917 (2016).
  45. Clerici, M., Faini, M., Aebersold, R. & Jinek, M. Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex. *Elife* **6**, (2017).
  46. Makarova, K., Wolf, Y. & Koonin, E. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales,

- Methanococcales, and Methanobacteriales. *Life* **5**, 818–840 (2015).
47. Phung, D. K. *et al.* Archaeal  $\beta$ -CASP ribonucleases of the aCPSF1 family are orthologs of the eukaryal CPSF-73 factor. *Nucleic Acids Res.* **41**, 1091–103 (2013).
  48. Nishida, Y. *et al.* Crystal structure of an archaeal cleavage and polyadenylation specificity factor subunit from *Pyrococcus horikoshii*. *Proteins Struct. Funct. Bioinforma.* **78**, 2395–2398 (2010).
  49. Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Microbial genome analysis: the COG approach. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbx117
  50. Santangelo, T. J., Čuboňová, L. L., James, C. L. & Reeve, J. N. TFB1 or TFB2 Is Sufficient for *Thermococcus kodakaraensis* Viability and for Basal Transcription in Vitro. *J. Mol. Biol.* **367**, 344–357 (2007).
  51. Sanders, T. J. *et al.* TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Mol. Microbiol.* **111**, 784–797 (2019).
  52. Hirata, A. *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. *Mol. Microbiol.* **70**, 623–33 (2008).
  53. Santangelo, T. J. & Reeve, J. N. Deletion of switch 3 results in an archaeal RNA polymerase that is defective in transcript elongation. *J. Biol. Chem.* **285**, 23908–15 (2010).
  54. Gehring, A. M. & Santangelo, T. J. Archaeal RNA polymerase arrests transcription at DNA lesions. *Transcription* **8**, 288–296 (2017).
  55. Gehring, A. M. & Santangelo, T. J. Manipulating archaeal systems to permit analyses of transcription elongation-termination decisions in vitro. *Methods Mol. Biol.* **1276**, 263–79 (2015).
  56. Kolev, N. G., Yario, T. A., Benson, E. & Steitz, J. A. Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3'-end maturation. *EMBO Rep.* **9**, 1013–8 (2008).
  57. Orlova, M., Newlands, J., Das, A., Goldfarb, A. & Borukhov, S. Intrinsic transcript cleavage activity of RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 4596–600 (1995).
  58. Ray, W. C. & Daniels, C. J. PACRAT: a database and analysis system for archaeal and bacterial intergenic sequence features. *Nucleic Acids Res.* **31**, 109–13 (2003).
  59. Santangelo, T. J. *et al.* Polarity in archaeal operon transcription in *Thermococcus kodakaraensis*. *J. Bacteriol.* **190**, 2244–8 (2008).
  60. Hart, C. M. & Roberts, J. W. Deletion analysis of the lambda tR1 termination region. Effect of sequences near the transcript release sites, and the minimum length of rho-dependent transcripts. *J. Mol. Biol.* **237**, 255–65 (1994).
  61. Hart, C. M. & Roberts, J. W. Rho-dependent transcription termination. Characterization of the requirement for cytidine in the nascent transcript. *J. Biol. Chem.* **266**, 24140–8 (1991).
  62. Phung, D. K. & Clouet-d'Orval, B. Tips and Tricks to Probe the RNA-Degrading Activities of Hyperthermophilic Archaeal  $\beta$ -CASP Ribonucleases. 453–466 (2015). doi:10.1007/978-1-4939-2214-7\_26
  63. Silva, A. P. G. *et al.* Structure and activity of a novel archaeal  $\beta$ -CASP protein with N-terminal KH domains. *Structure* **19**, 622–32 (2011).
  64. Yang, X. -c., Sullivan, K. D., Marzluff, W. F. & Dominski, Z. Studies of the 5' Exonuclease and Endonuclease Activities of CPSF-73 in Histone Pre-mRNA Processing. *Mol. Cell. Biol.* **29**, 31–42 (2009).
  65. Dominski, Z., Yang, X. & Marzluff, W. F. The Polyadenylation Factor CPSF-73 Is Involved in Histone-Pre-mRNA Processing. *Cell* **123**, 37–48 (2005).
  66. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* **9**, 85–98 (2011).

67. Xie, Y. & Reeve, J. N. Transcription by an archaeal RNA polymerase is slowed but not blocked by an archaeal nucleosome. *J. Bacteriol.* **186**, 3492–3498 (2004).
68. Hirtreiter, A. *et al.* Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. *Nucleic Acids Res.* **38**, 4040–4051 (2010).
69. Martinez-Rucobo, F. W., Sainsbury, S., Cheung, A. C. & Cramer, P. Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. *EMBO J.* **30**, 1302–1310 (2011).
70. Kang, J. Y. *et al.* Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell* **173**, 1650–1662.e14 (2018).
71. Smollett, K., Blombach, F., Reichelt, R., Thomm, M. & Werner, F. A global analysis of transcription reveals two modes of Spt4/5 recruitment to archaeal RNA polymerase. *Nat. Microbiol.* **2**, 17021 (2017).
72. Burmann, B. M. *et al.* A NusE:NusG Complex Links Transcription and Translation. *Science (80-. )*. **328**, 501–504 (2010).
73. Guo, G. *et al.* Structural and biochemical insights into the DNA-binding mode of MjSpt4p:Spt5 complex at the exit tunnel of RNAPII. *J. Struct. Biol.* **192**, 418–425 (2015).
74. Ehara, H. *et al.* Structure of the complete elongation complex of RNA polymerase II with basal factors. *Science (80-. )*. **357**, 921–924 (2017).
75. Klein, B. J. *et al.* RNA polymerase and transcription elongation factor Spt4/5 complex structure. *Proc. Natl. Acad. Sci.* **108**, 546–550 (2011).
76. Schulz, S. *et al.* TFE and Spt4/5 open and close the RNA polymerase clamp during the transcription cycle. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1816–E1825 (2016).
77. Nagy, J. *et al.* Complete architecture of the archaeal RNA polymerase open complex from single-molecule FRET and NPS. *Nat. Commun.* **6**, 6161 (2015).
78. Plaschka, C. *et al.* Architecture of the RNA polymerase II–Mediator core initiation complex. *Nature* **518**, 376–380 (2015).
79. Walker, J. E. & Santangelo, T. J. Analyses of in vivo interactions between transcription factors and the archaeal RNA polymerase. *Methods* **86**, 73–79 (2015).
80. Horsfall, L. E. *et al.* Competitive Inhibitors of the CphA Metallo- $\beta$ -Lactamase from *Aeromonas hydrophila*. *Antimicrob. Agents Chemother.* **51**, 2136–2142 (2007).
81. Sarmiento, F., Mrázek, J. & Whitman, W. B. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon *Methanococcus maripaludis*. *Proc. Natl. Acad. Sci.* **110**, 4726–4731 (2013).
82. Zhang, C., Phillips, A. P. R., Wipfler, R. L., Olsen, G. J. & Whitaker, R. J. The essential genome of the crenarchaeal model *Sulfolobus islandicus*. *Nat. Commun.* **9**, 1–11 (2018).
83. Garas, M., Dichtl, B. & Keller, W. The role of the putative 3' end processing endonuclease Ysh1p in mRNA and snoRNA synthesis. *RNA* **14**, 2671–84 (2008).
84. Gehring, A. M., Sanders, T. J. & Santangelo, T. J. Markerless Gene Editing in the Hyperthermophilic Archaeon *Thermococcus kodakarensis*. *Bio-protocol* **7**, (2017).
85. Chanfreau, G., Noble, S. M. & Guthrie, C. Essential Yeast Protein with Unexpected Similarity to Subunits of Mammalian Cleavage and Polyadenylation Specificity Factor (CPSF). *Science (80-. )*. **274**, 1511–1514 (1996).
86. Li, Z., Santangelo, T. J., Cuboňová, L., Reeve, J. N. & Kelman, Z. Affinity purification of an archaeal DNA replication protein network. *MBio* **1**, e00221-10–e00221-19 (2010).
87. Dev, K. *et al.* Archaeal aIF2B interacts with eukaryotic translation initiation factors eIF2 $\alpha$  and eIF2 $\beta$ : Implications for aIF2B function and eIF2B regulation. *J. Mol. Biol.* **392**, 701–22 (2009).
88. Santangelo, T. J., Cuboňová, L. & Reeve, J. N. Deletion of alternative pathways for reductant recycling in *Thermococcus kodakarensis* increases hydrogen production. *Mol. Microbiol.* **81**, 897–911 (2011).
89. Burkhart, B. W., Febvre, H. P. & Santangelo, T. J. Distinct Physiological Roles of the

- Three Ferredoxins Encoded in the Hyperthermophilic Archaeon *Thermococcus kodakarensis*. *MBio* **10**, (2019).
90. Santangelo, T. J., Cubonová, L. & Reeve, J. N. Thermococcus kodakarensis genetics: TK1827-encoded beta-glycosidase, new positive-selection protocol, and targeted and repetitive deletion technology. *Appl. Environ. Microbiol.* **76**, 1044–52 (2010).
  91. Jäger, D., Förstner, K. U., Sharma, C. M., Santangelo, T. J. & Reeve, J. N. Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics* **15**, 684 (2014).
  92. Speed, M. C., Burkhart, B. W., Picking, J. W. & Santangelo, T. J. An Archaeal Fluoride-Responsive Riboswitch Provides an Inducible Expression System for Hyperthermophiles. *Appl. Environ. Microbiol.* **84**, e02306-17 (2018).
  93. Wolf, Y. I., Makarova, K. S., Yutin, N. & Koonin, E. V. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol. Direct* **7**, 46 (2012).
  94. Clerici, M., Faini, M., Muckenfuss, L. M., Aebersold, R. & Jinek, M. Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex. *Nat. Struct. Mol. Biol.* **25**, 135–138 (2018).
  95. El Hage, A., Koper, M., Kufel, J. & Tollervey, D. Efficient termination of transcription by RNA polymerase I requires the 5' exonuclease Rat1 in yeast. *Genes Dev.* **22**, 1069–1081 (2008).
  96. Kawauchi, J., Mischo, H., Braglia, P., Rondon, A. & Proudfoot, N. J. Budding yeast RNA polymerases I and II employ parallel mechanisms of transcriptional termination. *Genes Dev.* **22**, 1082–1092 (2008).
  97. Braglia, P., Kawauchi, J. & Proudfoot, N. J. Co-transcriptional RNA cleavage provides a failsafe termination mechanism for yeast RNA polymerase I. *Nucleic Acids Res.* **39**, 1439–1448 (2011).
  98. Dar, D., Prasse, D., Schmitz, R. A. & Sorek, R. Widespread formation of alternative 3' UTR isoforms via transcription termination in archaea. *Nat. Microbiol.* **1**, 1–9 (2016).
  99. Wang, C. *et al.* Structural basis of archaeal FttA-dependent transcription termination. *bioRxiv Prepr. Serv. Biol.* (2023). doi:10.1101/2023.08.09.552649
  100. Cohen, O. *et al.* Comparative transcriptomics across the prokaryotic tree of life. *Nucleic Acids Res.* **44**, W46–W53 (2016).
  101. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–54 (1976).
  102. Mattioli, F. *et al.* Structure of histone-based chromatin in Archaea. *Science* (80-. ). **357**, 609–612 (2017).

## CHAPTER 4: DEVELOPMENT OF NASCENT ELONGATING TRANSCRIPT SEQUENCING (NET-SEQ) TECHNIQUES IN THE HYPERTHERMOPHILIC ARCHAEON THERMOCOCCUS KODAKARENSIS

### Introduction

The activity of multi-subunit RNA polymerase (RNAP) is conserved across all domains and is responsible for transcribing the messages contained within the genome<sup>1-6</sup>. The synthesis of nascent RNAs via RNAP is a discontinuous process and can be interrupted by periodic pausing due to sequence elements, nucleotide misincorporation, protein roadblocks, or arrested/backtracked ternary elongation complexes (TECs)<sup>7-9</sup>. Backtracked TECs result in disengagement of the 3' end of nascent RNA from the catalytic site of RNAP<sup>10</sup>. The endonucleolytic activity of RNAP can be stimulated by cleavage stimulatory factors (TFS in Archaea, TFIIIS in Eukarya, and GreA/GreB in Bacteria) to cleave and restore the 3' OH to the active site and reactivate TECs<sup>11-15</sup>. While archaeal TFS is highly conserved, some archaea encode several TFS proteins with separate functions that are dispensable (TFS2 is essential but lacks the Asp-Glu dipeptide critical for stimulation of RNAP cleavage activity)<sup>16-19</sup>. The TFS4 paralogue is a potent inhibitor of transcription initiation during viral infections<sup>18</sup>; yet, little is known regarding the role TFS plays in stress conditions. Additionally, the TFIIIS orthologue and GreA/GreB analogues have been shown to have fundamental roles in transcription regulation and the absence of such begets vulnerability during environmental assaults<sup>20-22</sup>. Given the identical stimulatory functions of TFS, TFIIIS, and GreA/GreB, the absence of archaeal TFS presages similar vulnerabilities as seen in the bacterial and eukaryotic studies.

While all Domains contain DNA-binding proteins that may induce RNAP backtracking, many archaea, and all eukaryotes express histone proteins in sufficient quantities to saturate the genome<sup>23</sup>, which can slow and even impede the progression of the transcription apparatus<sup>23-27</sup>. Archaeal histones retain the canonical histone fold that consists of 3 alpha helices joined by two

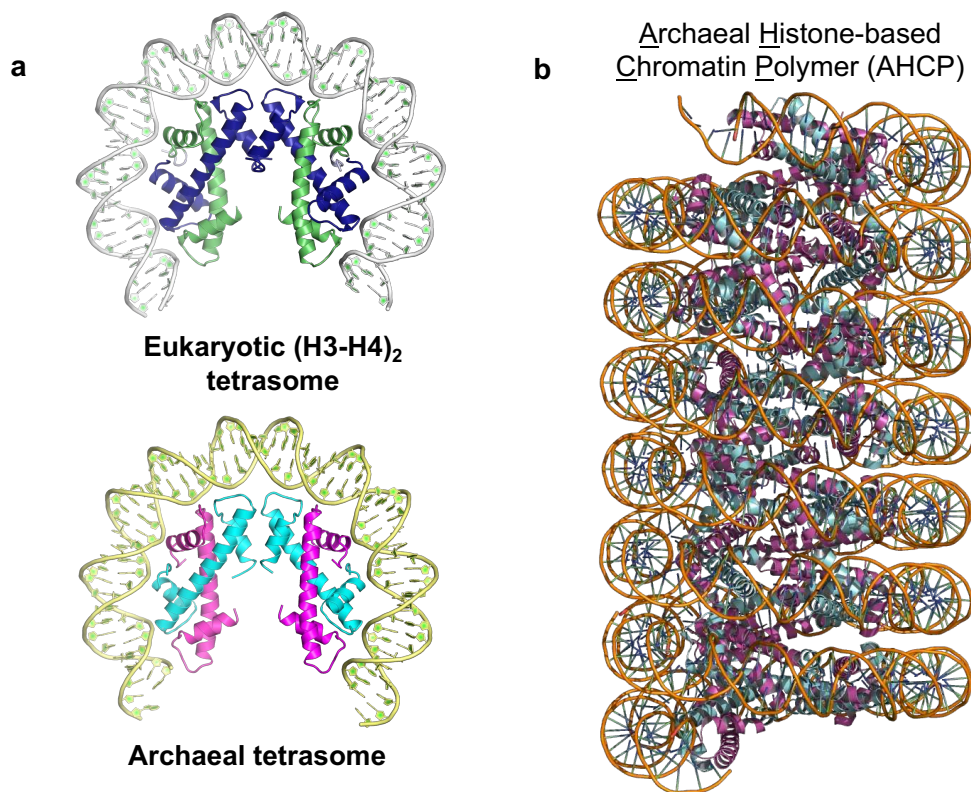
loops ( $\alpha 1$ -L1- $\alpha 2$ -L2- $\alpha 3$ ). The geometry of the archaeal chromatin superhelix resembles that of the eukaryotic nucleosome DNA arrangement, specifically the tertiary structure of the (H3/H4)<sub>2</sub> tetramer (Fig. 4.1a).

Comparable to eukaryotic histones, archaeal histones organize the genome and regulate the progression of the transcription apparatus. Chromatin regulatory factors and modifications to eukaryotic chromatin are extensive and well-studied, which depicts an incredibly complex signaling cascade that tightly regulates gene expression. However, archaeal genomes are devoid of chromatin remodeling complexes, lack the canonical N and C terminal extensions common to eukaryotic organisms, and evidence of PTMs is biologically insufficient. Additionally, archaeal histone tetramers bind and wrap the DNA to form an extended, super-helical structure with stacked gyres of DNA (Fig. 4.1b), which is in significant contrast to the eukaryotic 'beads-on-a-string' nucleosome model but interestingly is a nearly exact super-helical arrangement to eukaryotic telomeres<sup>28</sup>. The lack of eukaryotic histone-regulatory techniques in archaeal genomes suggests that histone-mediated regulation of transcription is circumvented with the aid of transcription factors. Evidence suggests that the elongation factors Spt4/5 (Spt5 is conserved in all domains: Spt5 in Archaea/Eukarya and NusG in Bacteria) and TFS assist RNAP in processive and productive elongation in a histone-based chromatin landscape<sup>23</sup>.

Many archaea encode for more than one histone protein, each with predicted and known differences in DNA binding capacity, tetramer formation, and stability<sup>31–33</sup>. For instance, histone B from *Methanothermobacter feravidus* (HMfB) has a higher affinity for DNA than that of the histone A (HMfA) isoform. Additionally, HMfA and HMfB differ in total abundance throughout the growth phase, suggesting that each isoform has a unique function<sup>32,34–36</sup>. The hyperthermophilic archaeon, *Thermococcus kodakarensis*, encodes two histone proteins – histone A (HTkA) and histone B (HTkB) that differ by just 11 amino acids which result in varied histone-DNA interactions, are both individually non-essential (deletion of both is synthetically lethal), and can



form both homo- and hetero-dimers but strains can be developed that encode only one histone isoform<sup>30,37</sup>. Addition of HTkB to DNA templates *in vitro* decreased transcription elongation rates by ~80%<sup>23</sup>. In contrast, HTkA-based chromatin decreased transcription elongation rates by just ~20% (Chapter 2). Although TFS activity increased elongation rates in the *in vitro* HTkB-based chromatin environment by ~4-fold<sup>23</sup>, the addition of TFS had little impact on elongation rates in the HTkA-based chromatin landscape; however, TFS addition did increase the percentage of full-length RNA transcripts (Chapter 2). Despite the positive effects of TFS on TECs in a histone-based chromatin environment, TFS is dispensable with negligible impacts on overall growth in optimal conditions (Chapter 2).



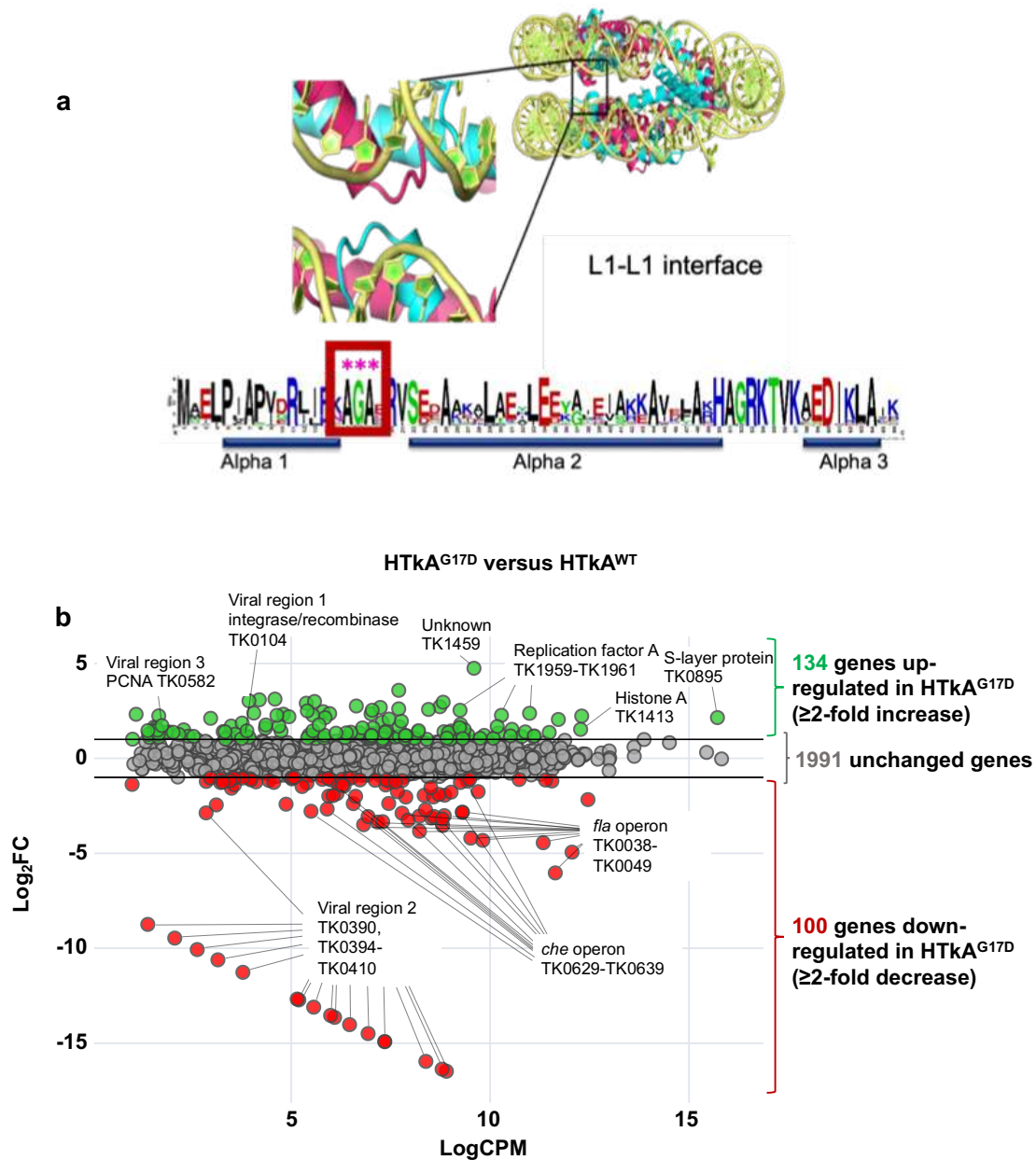
**Figure 4.1. Archaeal histones bind and organize the genome.** **a.** The eukaryotic (H3/H4)<sub>2</sub> tetrasome<sup>29</sup> (top) and the archaeal tetrasome<sup>30</sup> (bottom) share the same geometry of the tertiary structure<sup>30</sup>. The histone-DNA contacts within each tetrasome are conserved. **b.** Archaeal tetrasomes bind and wrap ~60 bp of DNA. Separate gyres of DNA stack upon each other forming an extended, tightly packed, slinky-like structure<sup>30</sup>.

Modifications to histone residues responsible for tight-packing of archaeal histone-based chromatin polymers (AHCPs) disrupt cellular fitness and generate altered chromatin digestion patterns<sup>30</sup>. Within the four helix-bundle of archaeal histones, loop 1 (L1) from dimers 1 and 4 create an L1-L1 interface, in which a conserved glycine (G17) is positioned to create the only region of close contact between adjacent layers of AHCPs (Fig. 4.2a). Introducing a larger side chain at position G17 disrupts the tightly-packed extended chromatin structure<sup>30</sup>, resulting in a significantly altered transcriptome (Fig. 4.2b) and decreased *in vitro* elongation rates (Chapter 2)<sup>38,39</sup>. While these results are indicative of regulatory gene expression mechanisms induced by chromatin dynamics and architecture – they do not reveal the rationale for the observed changes<sup>38</sup>.

Nascent elongating transcript sequencing (NET-seq) is a method to establish nascent transcription profiles at single-nucleotide resolution to monitor genome-wide transcription pausing, targeting only transcripts associated with RNAP, as total RNA-seq largely comprises steady-state and mature transcripts<sup>30,40</sup>. NET-seq protocols have been established in both Eukarya and Bacteria but has never been applied to archaeal cells; therefore, the distribution of TECs in even unperturbed cells will be enlightening and impactful<sup>40,41</sup>. Single-nucleotide resolution of eukaryotic *in vivo* NET-seq transcripts illustrate RNA polymerase II (Pol II) dwell sites that correlate to regions of strong interactions between the nucleosome and DNA<sup>24,42,43</sup>. Archaeal chromatin contains the same arrangement of four consecutive histone fold domains as the eukaryotic nucleosome, suggesting AHCPs share similar interactions.

Our knowledge of how changes to the archaeal chromatin landscape influence transcription and gene expression is incomplete. By establishing the profiles of nascent RNA transcripts at single-nucleotide resolution in +/- TFS conditions we aim to elucidate TFS-mediated transcription regulation in a normal histone-based chromatin environment. In addition, we aim to use NET-seq to determine global RNAP positioning and dwell times when the structure of histone-based

chromatin is disrupted with different histone variants in the presence and absence of TFS. This work will provide insight into the influence TFS has on transcription and gene expression, the regulation chromatin imposes on the transcription apparatus, and elucidate the rate-limiting step of archaeal transcription<sup>40,41,44–46</sup>.



**Figure 4.2. Chromatin architecture modulates the transcriptome.** **a.** The highly conserved alanine-glycine-alanine (AGA) motif at the L1-L1 interface permits proximity of separate gyres of DNA<sup>30</sup>. **b.** Differential gene expression of RNA-seq reveals significant (> 2-fold) changes in transcript abundance attributable to an altered chromatin landscape. Chromatin architecture changes result in up- (green) and down- (red) regulation of steady-state transcripts<sup>38</sup>.

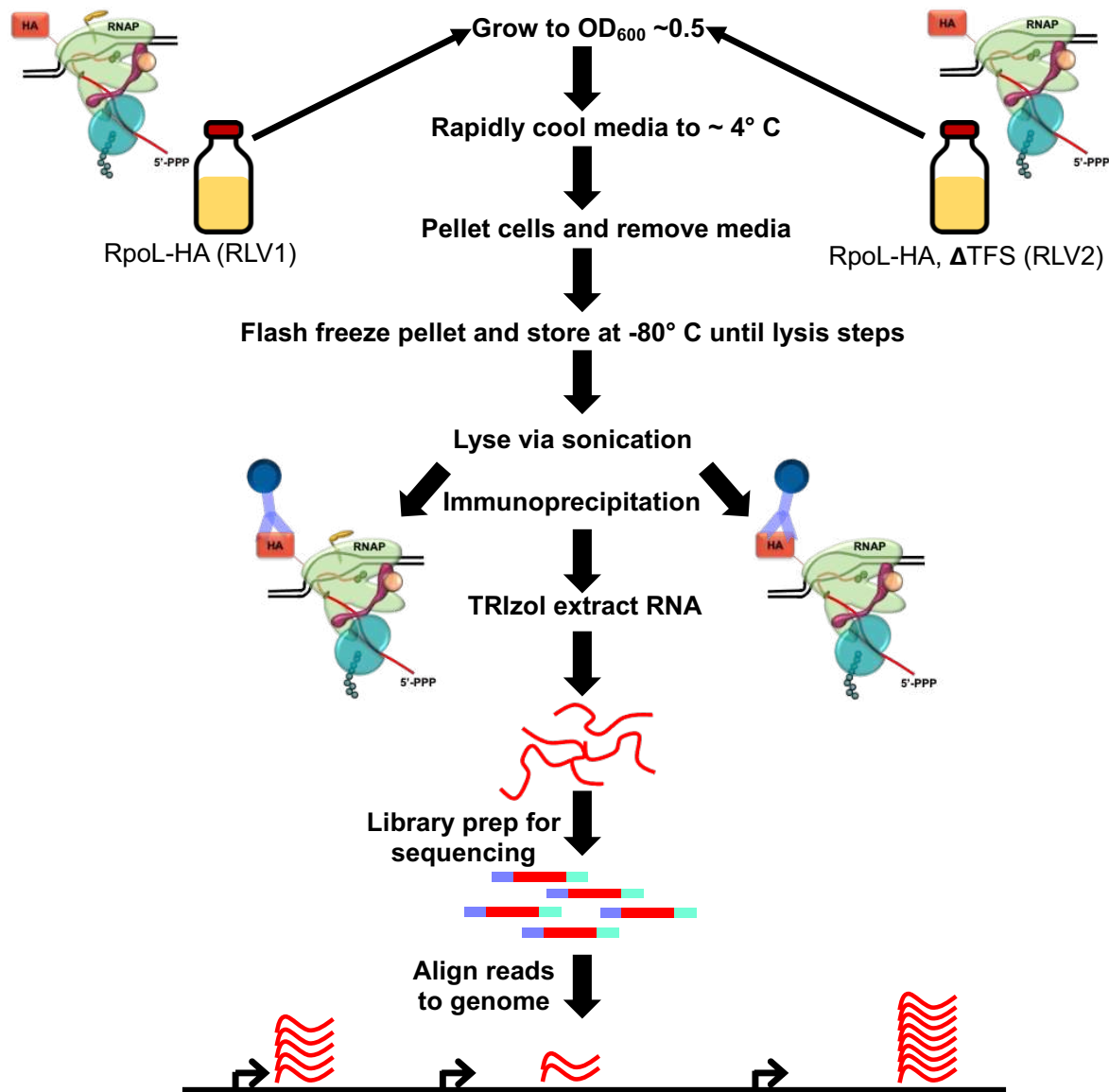
## Results and Discussion

### *Proposed protocol to capture nascent transcripts from in vivo TECs*

Archaeal histone-based chromatin forms a super-helical, slinky-like structure (Fig. 4.1a) with stacked gyres of DNA. RNA-seq data from a *T. kodakarensis* variant strain that disrupts the L1-L1 interface (Fig. 4.2a) of AHCPs reveals the differential expression of steady-state RNA levels (Fig. 4.2b) – indicative of regulatory gene expression mechanisms induced by chromatin dynamics and architecture – but the basis for the observed changes remains unclear<sup>38</sup>. To understand the cause of the steady-state changes, we will exploit the ability of NET-seq to target global positions of TECs<sup>41,42,47,48</sup> in normal and variant chromatin landscapes at single-nucleotide resolution.

NET-seq has been applied to both eukaryotic and bacterial cells but has never been applied to archaeal cells<sup>41,47,48</sup>. The model archaeal organism, *T. kodakarensis*, is a hyperthermophile that thrives at 85°C, which makes rapidly cooling the cells (an essential part of the protocol) problematic, while the number of cells required for sequencing and analysis necessitates large volumes of media. To circumvent these complications, we devised a protocol (Fig. 4.3) to rapidly cool large volumes of media within seconds to enable “freezing” of TECs in their native, global position (materials and methods). Cultures are grown to mid-exponential to ensure that transcription is active, abundant, and capture of TECs represents accuracy in global positions that reflect the first step in archaeal gene expression. Taking a ‘snapshot’ of the cell during high-transcriptional output requires the immediate cessation of cellular activity, one that is commonly achieved with the addition of cross-linkers (e.g., formaldehyde). While this technique is quite useful for mesophilic organisms, reverse cross-linking with formaldehyde will occur ~50 percent

of the time at 85°C<sup>49</sup>. Additionally, the large volume of media required to collect a sufficient amount of nascent RNA associated with TECs for sequencing makes using formaldehyde cross-linking to visualize global RNAP positions undesirable.



**Figure 4.3. Schematic of the novel, archaeal NET-seq protocol.** Cell cultures are grown ( $n = 3$ ) to a mid-exponential growth phase of  $OD_{600} \cong 0.5$ . The cells are rapidly cooled from 85°C to 4°C, pelleted, lysed, the RNA is extracted, and prepped for sequencing. The libraries are prepared for the Illumina NovaSeq sequencing platform.

**Strain construction to identify the TFS- and chromatin-mediated regulation imposed on global RNAP positioning patterns**

Identification of the regulation imposed by TFS, HTkA, HTkB, and chromatin architecture on the global positioning patterns of RNAP requires extensive construction of strains to isolate the regulation modulated by each individually, and in combination with, each factor. Utilizing our well-established genetic techniques<sup>50</sup>, we successfully generated multiple strains to analyze via NET-seq (Fig. 4.4). Most strains have been constructed and verified via whole genome sequencing (WGS), but a few still need to either be constructed or verified at the time of this report.

Mutation	Mutation details					
TS559	Parental strain					
TK1167C	His <sub>6</sub> -affinity and HA-epitope tag on C terminus of RpoL subunit of RNAP					
ΔTK0533	Deleted TFS – transcription elongation factor					
ΔTK2289	Deleted HTkB					
ΔTK1413	Deleted HTkA					
TK2289 <sup>G17D</sup>	HTkB mutation at position 17 from a G (glycine) to a D (aspartic acid)					
TK1413 <sup>G17D</sup>	HTkA mutation at position 17 from a G (glycine) to a D (aspartic acid)					

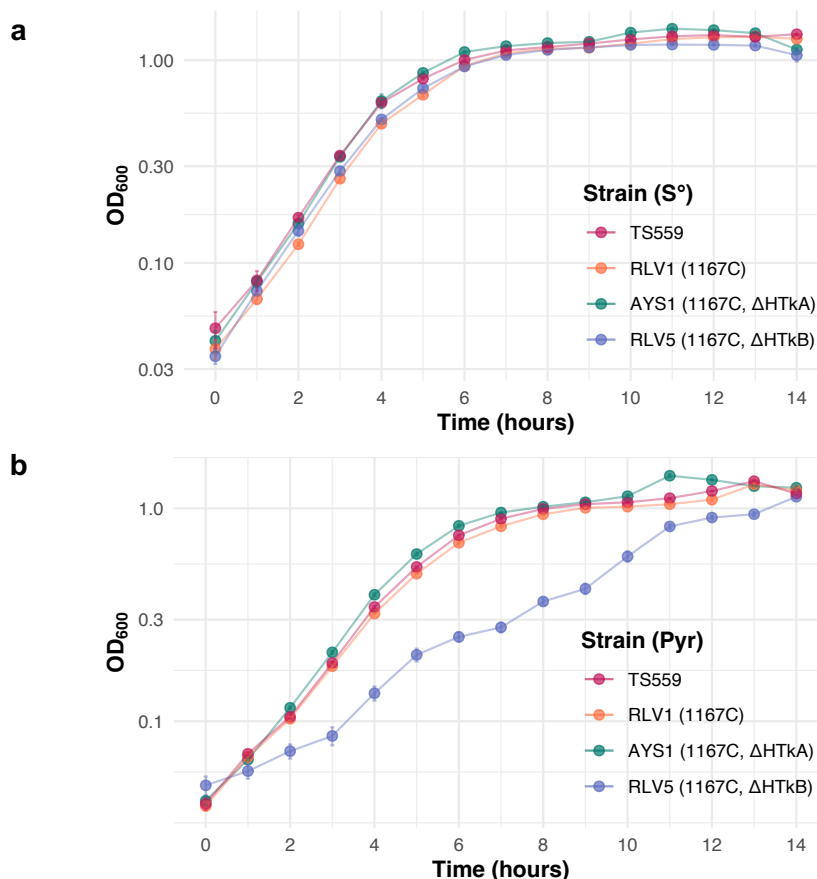
Strain	TK1167C	ΔTK0533	ΔTK2289	ΔTK1413	TK2289 <sup>G17D</sup>	TK1413 <sup>G17D</sup>
RLV1	✓					
RLV2	✓	✓				
RLV3	✓	✓	✓			
RLV4	✓	✓	✓			✓
RLV5	✓		✓			
RLV6	✓		✓			✓
RLV7	✓				✓	
RLV8	✓			✓	✓	
RLV9	✓				✓	✓
AYS1	✓			✓		
AYS2	✓	✓		✓		
AYS3	✓	✓			✓	
PAU1	✓					✓
AME1	✓	✓				✓
AME2	✓	✓			✓	✓
PAU2	✓	✓		✓	✓	

**Figure 4.4. Mutations and strains developed to assess for regulatory effects on gene expression.** The parental strain, TS559, was used to create subsequent strains (bottom) with the corresponding mutations (top) to determine what regulatory effects each factor has on gene expression via NET-seq to visualize perturbations in global RNAP positions.

To obtain the RLV1 strain, a His<sub>6</sub>/HA-tag was introduced to the C-terminus of the RpoL subunit of RNAP in the TS559 parental strain. The RLV2 strain was obtained from the RLV1 strain by deleting the TK0533 locus which encodes for TFS. Subsequent mutations were made by manipulating either RLV1 or RLV2 for the desired genotype. However, any mutations that were in combination with either deletion or mutation of HTkA were done before manipulation of HTkA due to its fundamental role in DNA uptake and recombination<sup>51</sup>.

To identify any preliminary phenotypes associated with select strains, cultures were grown in either sulfur/pyruvate or pyruvate media (Fig 4.5). *T. kodakarensis* can utilize both sulfur and pyruvate components as a final electron acceptor, producing either H<sub>2</sub>S or H<sub>2</sub> by-products, respectively. The production of either metabolite requires distinct metabolic pathways and the expression of different metabolic genes<sup>52</sup>. As expected, addition of the His<sub>6</sub>/HA-tag on the RpoL subunit (RLV1) is non-phenotypic in both sulfur and pyruvate conditions when compared to the TS559 parental strain (Fig. 4.5). Both the  $\Delta$ HTkA strain (AYS1) and  $\Delta$ HTkB strain (RLV5) grow similar to the other strains in the sulfur media (Fig 4.5a). However, in pyruvate media RLV5 has a more severe phenotype, failing to reach a density congruent with the rest of the strains (Fig 4.5b). Interestingly, the absence of HTkA (AYS1) also has little phenotypic effects until the stationary phase in the pyruvate condition. The lack of HTkA at this growth phase seems to provide the cells a brief chance to increase slightly more in overall density than the other strains (Fig. 4.5b). Given the lack of a dramatic phenotype in the  $\Delta$ HTkA strain (AYS1), it could be concluded that HTkA has little to no regulatory role regarding transcription; however, perturbations to the *in vivo* single isoform HTkA-based chromatin landscape significantly alter the transcriptome (Fig. 4.2b). Together, these results suggest that each histone protein plays a

distinct role in regulating gene expression and that the presence of both is crucial for adapting to environmental changes.



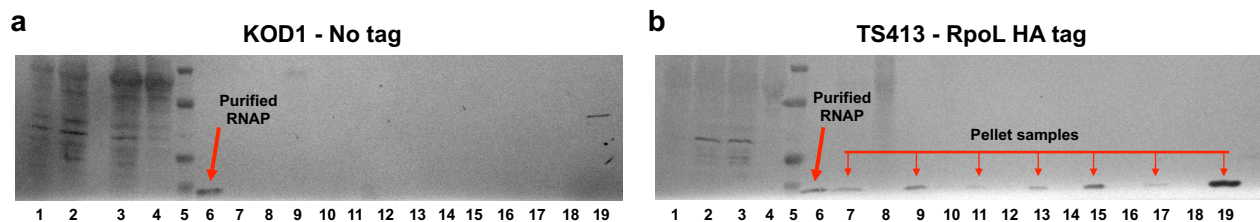
**Figure 4.5. HTkA and HTkB have distinct regulatory roles in gene expression.** *T. kodakarensis* strains TS559 (parental), RLV1 (RpoL-His<sub>6</sub>/HA tag), AYS1 (ΔHTkA), and RLV5 (ΔHTkB) grow nearly identically at 85°C in sulfur conditions, until the stationary phase when AYS1 and RLV5 have distinct phenotypes (a). The absence of sulfur highlights a severe phenotype and distinct role for HTkB in the expression of specific metabolic genes (b). Error bars represent the SE from biological triplicate cultures.

### Verification of TEC capture and RNA isolation of proposed NET-seq protocol

Isolation of *in vivo* TECs relies on capture via a His<sub>6</sub>/HA-tag on the small RpoL subunit of RNAP. To observe our ability to capture RNAP with specificity and maintain TECs throughout the lysis and immunoprecipitation protocol we utilized the His<sub>6</sub> tag and Ni-NTA magnetic beads for affinity purification of strains KOD1 and TS413. We grew cultures of KOD1 (WT strain with no His<sub>6</sub>/HA-



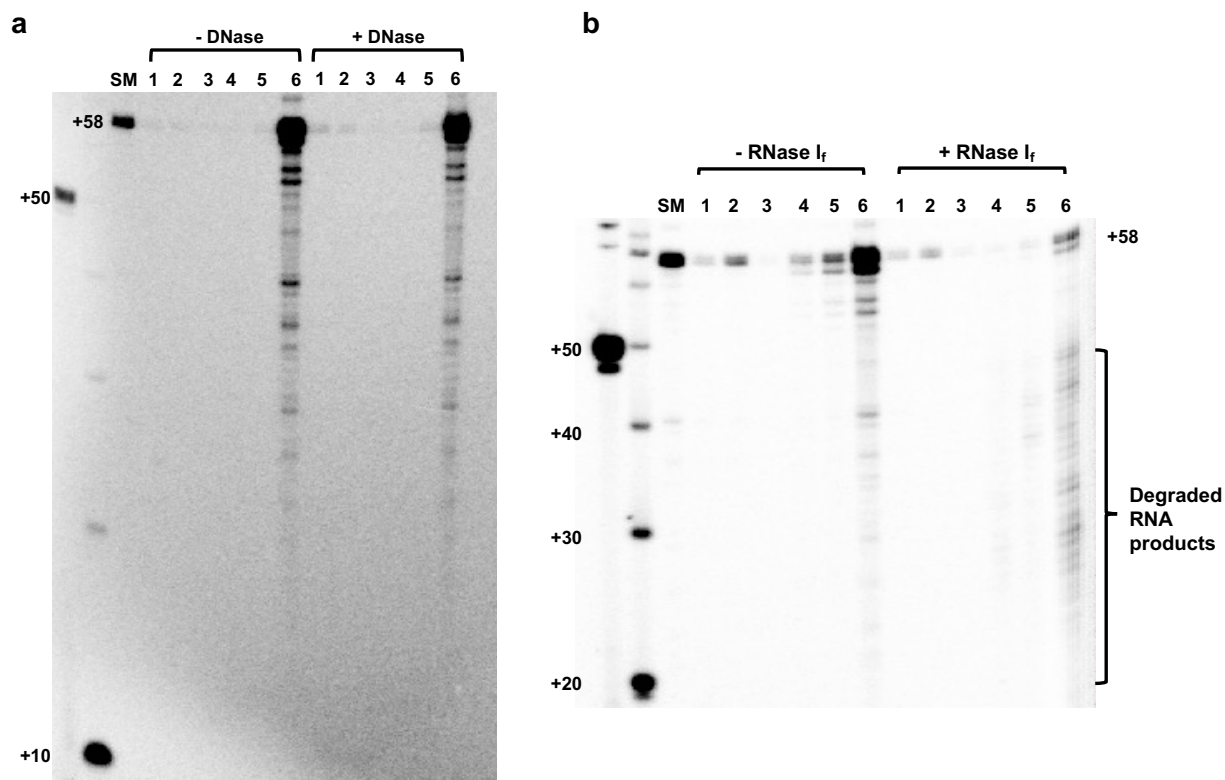
tag) and TS413 (strain with RpoL- His<sub>6</sub>/HA) to mid-exponential, isolated the pellets, and went through our immunoprecipitation protocol (exchanging anti-HA magnetic beads for Ni-NTA magnetic beads) to detect the presence/absence of RNAP utilizing a Western blot and anti-HA antibodies (Fig. 4.6). While capture and maintenance of RNAP was successful throughout the lysis and immunoprecipitation protocol (i.e., RNAP remained in the pellet), the ability to maintain TECs was still obscure (Fig. 4.6b).



**Figure 4.6. RNAP can be captured and isolated via affinity purification from whole cell lysate (WCL).** KOD1 (a) and TS413 (b) cells were lysed via sonication and subjected to the immunoprecipitation protocol. The lysate and washes were collected throughout and ran on a 4-20% Criterion gel. RpoL-HA was detected using anti-HA antibodies. Lane 6 provides a positive control. Lane 1 = whole cells, lane 2 = WCL, lane 3 = WCL + Ni-NTA beads, lane 4 = WCL after Ni-NTA bead incubation, lane 5 = Kaleidoscope marker, lane 6 = positive control, lane 7 = Ni-NTA beads/TECs in Wash Buffer (WB – wash 1), lane 8 = wash 1 (no Ni-NTA/TECs), lane 9 = Ni-NTA beads/TECs in WB (wash 2), lane 10 = wash 2 (no Ni-NTA/TECs), lane 11 = Ni-NTA beads/TECs in Nuclease Buffer (NB x 1), lane 12 = NB x 1 (no Ni-NTA/TECs), lane 13 = Ni-NTA beads/TECs in NB x 2, lane 14 = NB x 2 (no Ni-NTA/TECs), lane 15 = nuclease treatment (NT) + Ni-NTA beads/TECs, lane 16 = NT supernatant (no Ni-NTA/TECs), lane 17 = Ni-NTA beads/TECs in WB (wash 3), lane 18 = wash 3 (no Ni-NTA/TECs), lane 19 = final pellet.

Therefore, to verify that TECs are not disrupted by sonication and that they remain intact during the immunoprecipitation protocol, *in vitro* TECs<sub>+58</sub> containing radiolabeled RNA were subjected to the lysis and immunoprecipitation steps and the collected samples were run on a 12% polyacrylamide/urea gel for visualization (Fig. 4.7). RNAP used in our *in vitro* transcription assays also contains a His<sub>6</sub>/HA tag on the RpoL subunit, which we exploit to affinity capture and purify paused TECs at position 58 on the DNA template. DNase I treatment does not interfere with the stability of TECs<sub>+58</sub> (Fig. 4.7a), and the protocol results in the successful isolation of the nascent RNA associated with RNAP (Fig. 4.7b). As noted in the -RNase I<sub>f</sub> conditions (Fig. 4.7b,

lanes 2-5), high salt buffer conditions result in some loss of TECs<sub>+58</sub>. The salt concentration was decreased in the current buffer conditions to prevent loss of TECs (Materials and Methods).



**Figure 4.7. Captured TECs<sub>+58</sub> remain intact during lysis and immunoprecipitation.** The *in vitro* transcription system using components from *T. kodakarensis* permits promoter-directed assembly of pre-initiation complexes with purified RNAP (His<sub>6</sub>/HA tagged that can be affinity purified or immunoprecipitated), TFB, and TBP. Addition of ATP, CTP, UTP and <sup>32</sup>P- $\alpha$ -UTP allows RNAP to elongate to the end of a +58 nt C-less cassette and generate a uniformly radiolabeled RNA transcript (TECs<sub>+58</sub>). TECs<sub>+58</sub> (SM) are resistant to repeated high-salt buffer washes. **a.** *In vitro* TECs<sub>+58</sub> remain intact throughout lysis via sonication and DNase I treatment has no effect on the stability of TECs<sub>+58</sub>. SM = starting material, lane 1 = supernatant from lysis via sonication, lane 2 = combined wash 1 and 2, lane 3 = combined buffer exchange for nuclease buffer x 2, lane 4 = supernatant from DNase I treatment, lane 5 = wash 3, lane 6 = final pellet. **b.** While the buffer conditions resulted in some loss, the TECs<sub>+58</sub> remained intact throughout the protocol and +/- RNase I<sub>f</sub> treatment shows that the RNA inside of RNAP is protected from degradation and can be successfully isolated. Lane 1 = supernatant from lysis via sonication, lane 2 = combined wash 1 and 2, lane 3 = combined buffer exchange for nuclease buffer x 2, lane 4 = supernatant from RNase I<sub>f</sub> treatment, lane 5 = wash 3, lane 6 = final pellet.

### Proposed NET-seq library preparation

Library preparations of the unknown sequences of RNA isolated from *in vivo* TECs for Illumina sequencing requires an initial step to remove any phosphates on the 5' end of the RNA followed by a kinase reaction to place just one phosphate group on the isolated RNAs to make the 5' end known, uniform, and distinguishable from the 3' OH for adapter ligation. The phosphorylated RNAs are ligated to an RNA adapter that has homology with sequences in the Reverse Transcriptase (RT) primer (Fig. 4.7a). Each RT primer encodes a 5' phosphate, a 3' OH, unique indices (single index – i7) for library identification for multiplex sequencing, Illumina P5 and P7 sequences, read 1 and read 2 primer sequences (for amplification), and an iSP18 spacer to extend the size of the final amplicons for sequencing (Table 4.1).

**Table 4.1.** Sequences of primers used in the NET-seq library preparations.

Oligonucleotide	Sequence (5' → 3')
3' adapter	App-NNNNNTGGAATTCTCGGGTGCCAAGG-ddNTP
RT primer 1	TACGAC/iSp18/TCGGACTGTAGAACTCTGAACGTGTAGATCTCGGTGGTCGCCGTATC ATTGCGC/CAAGCAGAAGACGGCATAACGAGAT [ATCACG] GTGACTGGAGTTCCTTGGCA CCCAGAATTCCA
RT primer 2	TACGAC/iSp18/TCGGACTGTAGAACTCTGAACGTGTAGATCTCGGTGGTCGCCGTATC ATTGCGC/CAAGCAGAAGACGGCATAACGAGAT [CGATGT] GTGACTGGAGTTCCTTGGCA CCCAGAATTCCA
RT primer 3	TACGAC/iSp18/TCGGACTGTAGAACTCTGAACGTGTAGATCTCGGTGGTCGCCGTATC ATTGCGC/CAAGCAGAAGACGGCATAACGAGAT [TTAGGC] GTGACTGGAGTTCCTTGGCA CCCAGAATTCCA
RT primer 4	TACGAC/iSp18/TCGGACTGTAGAACTCTGAACGTGTAGATCTCGGTGGTCGCCGTATC ATTGCGC/CAAGCAGAAGACGGCATAACGAGAT [TGACCA] GTGACTGGAGTTCCTTGGCA CCCAGAATTCCA
RT primer 5	TACGAC/iSp18/TCGGACTGTAGAACTCTGAACGTGTAGATCTCGGTGGTCGCCGTATC ATTGCGC/CAAGCAGAAGACGGCATAACGAGAT [ACAGTG] GTGACTGGAGTTCCTTGGCA CCCAGAATTCCA
RT primer 6	TACGAC/iSp18/TCGGACTGTAGAACTCTGAACGTGTAGATCTCGGTGGTCGCCGTATC ATTGCGC/CAAGCAGAAGACGGCATAACGAGAT [GCCAAT] GTGACTGGAGTTCCTTGGCA CCCAGAATTCCA
Sequencing primers	Forward: AATGATACGGCGACCACCGAGATCTACAC Reverse: CAAGCAGAAGACGGCATAACGAGAT

\* Index sequences encoded in the RT primers for Illumina sequencing are in brackets.

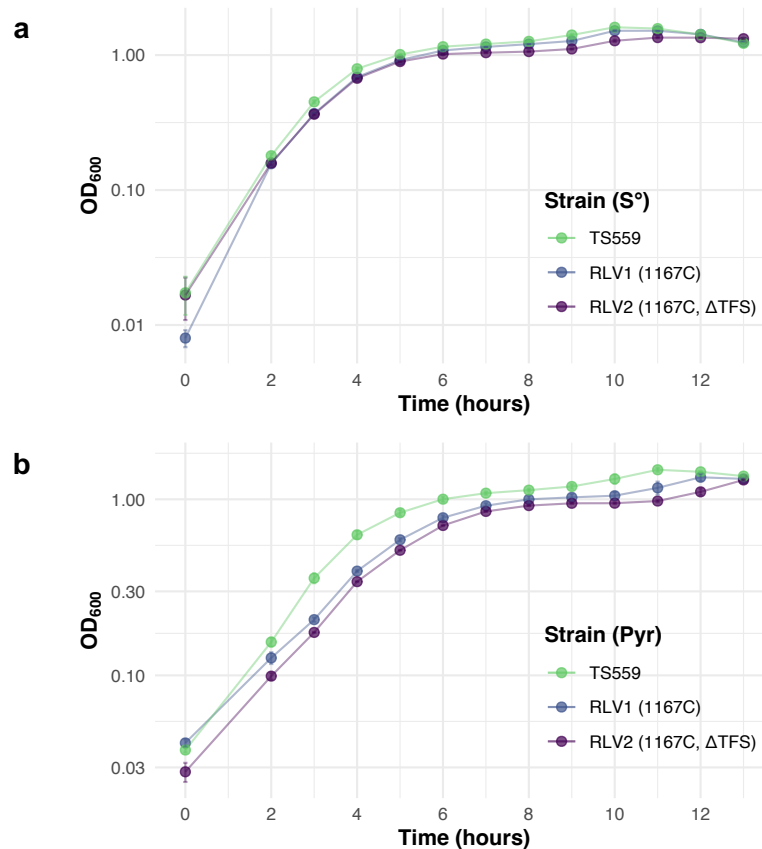
The RT primer and ligated RNA/adapter are reverse transcribed and then circularized utilizing CircLigase II. The unknown sequences and the small size of RNA necessitates circularization to place the P5 and P7 sequences in the correct locations for sequencing on Illumina platforms.



template is then amplified utilizing PCR, the amplicons purified, and sent for Illumina NovaSeq sequencing. Analysis of the Illumina NovaSeq sequencing reads will include fastp to trim adapters and low-quality sequences; HISAT2 to align the reads to the reference genome; custom (developed in Python) and open source (bedtools) software to count reads at defined genomic intervals (i.e., bins); and R software to visualize the results (Fig. 4.8b).

***Deletion of TFS (TK0533) has little impact on the overall growth of T. kodakarensis cells***

TFS in archaeal cells has been shown to rescue backtracked TECs and increase the rate of elongation in an *in vitro* HTkB-based chromatin environment<sup>23</sup>. However, TFS-mediated activity is not limited to archaeal species that encode histone proteins. The Crenarchaeon, *Sulfolobus solfataricus*, does not encode histone proteins yet expresses TFS paralogues that are all dispensable excepting TFS2, which does not appear to have the same behavior as the canonical TFS factor<sup>17,18</sup>. Therefore, the role the canonical TFS factor plays in assisting TECs in a histone-based environment may just be one role for the elongation factor, and in fact, likely has other regulatory roles similar to those seen in eukaryotes and bacteria in stress-induced conditions<sup>19–22</sup>. While the absence of TFS in *T. kodakarensis* seems to be non-phenotypic in normal, optimal growth conditions in media supplemented with both sulfur and pyruvate (Fig. 4.9a), the absence of sulfur results in a decrease in the optical density of cells lacking TFS (Fig. 4.9b). This phenotype does not appear to be severe but investigating the global RNAP positions in the +/- sulfur conditions may highlight TFS-mediated regulation of specific metabolic genes in *T. kodakarensis*. In addition, we will use NET-seq to map the global positions of TECs in +/- TFS in normal chromatin strains comparing stress and non-stress conditions (e.g., temperature) to determine any auxiliary regulatory role TFS may play in archaeal cells.



**Figure 4.9. HTkA and HTkB have distinct regulatory roles in gene expression.** *T. kodakarensis* strains TS559 (parental), RLV1 (RpoL-His<sub>6</sub>/HA tag), and RLV2 (RpoL-His<sub>6</sub>/HA tag & ΔTFS) grow nearly identically at 85°C in sulfur/pyruvate conditions (a). Removal of sulfur highlights a slight phenotype that suggests TFS may have a regulatory role in expression of specific metabolic genes (b). Error bars represent the SE from biological triplicate cultures.

## Conclusions and Future Directions

Little is known regarding the distinct role archaeal histones play in mediating the expression of genetic material via RNAP nor the mechanisms of how histone-mediated regulation is imposed. While archaeal TFS has been described to be a cleavage stimulatory factor and assists backtracked RNAP into productive elongation in a histone-based chromatin environment, not all Archaea encode histone proteins, but TFS is conserved<sup>2</sup>. TFS paralogues TFS1 and TFS4 in *S. solfataricus*, which do not encode histone proteins, are dichotomous in their ability to stimulate or inhibit transcription, respectively<sup>18</sup>. To explore histone- and TFS-mediated regulation of

transcription, we have devised a novel, hyperthermophilic NET-seq protocol (Fig. 4.3) to delineate the regulation imposed by each factor via RNAP positioning patterns.

Our preliminary data suggests that each histone protein has a distinct and regulatory role in transcriptional output (<sup>38</sup> and Fig. 4.5b) and that TFS may be important for the expression of certain metabolic genes (Fig. 4.9b). Current efforts are focused on the completion of all strains (Fig. 4.4) and preparation of strains RLV1 and RLV2 (Fig. 4.4) in triplicate for sequencing. The sequencing results from strains RLV1 and RLV2 are expected to determine the regulatory effects TFS has on global RNAP positioning patterns in a normal chromatin environment to isolate TFS-mediated activity. The sequencing and analytical outcomes from these strains will provide insight into the above proposed NET-seq protocol and any changes necessary for success to determine TFS-mediated regulation in stress conditions and eventually, chromatin-mediated regulation on global RNAP positioning patterns.

Additionally, once established, NET-seq has great potential to map archaeal transcription termination sites<sup>47</sup>. Characterization of factor-mediated archaeal transcription termination has made significant advances over the last few years, yet outstanding questions remain. While FttA (aCPSF1) has been described as a global transcription termination factor in Archaea and has been mapped to oligo(dT) tracts (intrinsic termination sequences) *in vivo*<sup>53–55</sup>, FttA-mediated termination is not restricted to U-rich sequences nor do all genes within archaeal genomes encode intrinsic termination sequences<sup>53,56–58</sup>. Mapping global RNAP positions in conjunction with global FttA positions could clarify the current data and provide greater insight into the regulation imposed by FttA-mediated termination on TECs.

Finally, efforts should be made to produce MNase-seq data regarding strains RLV5, RLV6, AYS1, and RLV8 (Fig. 4.4) to establish the positions of histones in single isoform strains that permit or abolish the extended chromatin structure. NET-seq derived transcripts should be

mapped to the reference genome and compared to the MNase-seq data to determine if histone proteins are positioned at locations of accumulated transcripts, which will reveal the sites of regulation specifically imposed by histone-based chromatin architecture in archaeal cells. Together, the data collected and analyzed from the TFS- and chromatin-mediated experiments will ultimately provide evidence for the influence TFS has on gene expression, the regulation chromatin imposes on the transcription apparatus, and the rate-limiting step of archaeal transcription<sup>40,41,44–46</sup>.



## **Materials and Methods**

### ***Strain construction***

*Thermococcus kodakarensis* strains were constructed as previously described<sup>50,59</sup>. Strain RLV1 and RLV2 were used as the base strain to construct all subsequent strains. RLV2 was constructed from RLV1 via markerless deletion of TK0533 (TFS). Deletion was confirmed by PCR amplification with primers flanking TK0533 and whole genome sequencing (WGS) on our in-house MinION. Subsequent deletions and mutations were confirmed via PCR amplification with primers flanking the target locus and WGS.

### ***In vitro transcription***

RNAP (RpoL-HA-His<sub>6</sub>), TFB, and TBP were purified as previously described<sup>46</sup>. The DNA template used in transcription assays was generated via PCR and gel purified as previously described<sup>60</sup>. Assembly of preinitiation complexes (PICs) and elongation via NTP deprivation was completed as previously described<sup>46,60,61</sup>.

### ***Growth conditions and biomass collection***

Cultures were grown as previously described for both the growth curves and for biomass collection<sup>59</sup>. The biomass collection cultures were grown to an OD<sub>600</sub> = ~0.5. The cultures were rapidly cooled to 4°C by pouring liquid cultures into a glass dish placed in an ice bath. Dry ice was added to the glass dish containing the culture in small and fine enough amounts to avoid cell lysis and the contents were continuously rocked until the temperature reached 4°C. The biomass was harvested via centrifugation, flash frozen with liquid nitrogen, and stored at -20°C.

### ***RNA isolation***

Frozen cell pellets (0.1 grams) were thawed on ice and resuspended by bringing the volume up to 2 mL with Nuclease Buffer (NB – 40 mM KCl, 20 mM Tris-HCl pH 8.0, 10 mM MgCl<sub>2</sub>, 2 mM CaCl<sub>2</sub>, 5% glycerol, 0.05% Tween-20). The cells were lysed via sonication and the cell lysate

was incubated with 24 U of DNase I (Roche), 500 U of RNase I<sub>f</sub> (NEB), and 5 mM MnCl<sub>2</sub> for 1 hour at 37°C. TECs were captured via RpoL-HA immunoprecipitation (100 µg) with Pierce™ Anti-HA Magnetic Beads (Thermo Scientific) or via RpoL-His<sub>6</sub> affinity purification (125 µg) with HisPur Ni-NTA Magnetic Beads (Thermo Scientific) by incubating at room temperature with rotation for 1 hour. Anti-HA Magnetic Beads were pre-incubated with 20 µg UltraPure™ Salmon Sperm DNA (Invitrogen). Captured TECs were washed (x2) in 500 µL Wash Buffer (WB – 20 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.5 M KCl, 4 mM MgCl<sub>2</sub>, 0.1 mg/mL BSA, 0.2% glycerol) and then resuspended in 50 µL NB and 500 U of RNase I<sub>f</sub>, and incubated for 30 minutes at 37°C. The reactions were washed (x1) in 500 µL WB and the RNA was extracted via TRIzol™ Reagent (Invitrogen). The TRIzol extracted pellets were resuspended in 50 µL NB and 10 U DNase I (TV = 100 µL) and incubated for 30 minutes at 37°C, followed by TRIzol extraction. To remove phosphate groups, the TRIzol extracted pellets were resuspended in 8.25 mM Tris-HCl pH 8.0, 1X Antarctic Phosphatase (AnP) reaction buffer, 5 U AnP, and 0.05% Tween-20, and incubated following the manufacturer's protocol (NEB). To add a phosphate group to the 5' end of the RNA transcripts, 2 U T4 Polynucleotide Kinase (PNK), 1X T4 PNK reaction buffer, and 4 mM ATP was added to the reactions, incubated following the manufacturer's protocol (NEB), TRIzol extracted and resuspended in 11 µL 10 mM Tris-HCl pH 8.0. The RNA transcripts (1 µL) with a 5' phosphate and a 3' hydroxyl were quantified via the Qubit™ microRNA Assay Kit (Invitrogen).

### ***Tris-Glycine SDS-PAGE and Western blot analysis***

500 ng of purified RNAP and samples from the KOD1 and TS413 purification were resolved on 4–20% Criterion™ TGX Stain-Free™ protein gels (Bio-Rad) and probed with primary Mouse (IgG1) HA Tag Monoclonal Antibody (Invitrogen). Addition of the secondary alkaline phosphatase-labeled Anti-Mouse IgG (H+L) Antibody (seracare) allowed for detection via 1-Step™ NBT/BCIP Substrate Solution (Thermo Scientific).

## ***Library preparation***

### Ligation reaction and excess 3' adapter removal

The quantified RNA was incubated with the 3' RNA adapter in 10 U T4 RNA Ligase 2, truncated (T4 Rnl2tr – NEB); 1X T4 Rnl2tr Reaction Buffer; 10  $\mu$ L RNA substrate ( $\leq$  200 ng); 14.5% PEG8000; and 6  $\mu$ M 3' adapter overnight at 16°C. 50 U of 5' deadenylase (NEB) and 30 U of RecJ<sub>f</sub> (NEB) were added to the reaction and incubated for 90 minutes at 30°C. The ligated products were purified via P/C/I (25/24/1) extraction and resuspended in 10  $\mu$ L nuclease-free water.

### RTase reaction

The purified ligation reactions were incubated with 3  $\mu$ M RT primer, 0.5 mM dNTPs, 5 mM DTT, 0.6 U SUPERase•In™ RNase Inhibitor (Invitrogen), 10 U PrimeScript RTase (Takara), and 1X PrimeScript Buffer for 30 minutes at 48°C. 2 U RNase H was added to the reactions and incubated for 15 minutes at 37°C.

### Gel purification

The reactions were separated on a 10% Criterion™ TBE-Urea Polyacrylamide Gel (BioRad) and products between ~154 – 178 bp were excised and the cDNA was eluted in 600  $\mu$ L gel elution buffer (0.5 M NaCl, 10 mM Tris-HCl pH 8.0, 1  $\mu$ M EDTA pH 8.0, and 0.05% Tween-20) overnight at 37°C with shaking (200 rpm). The eluant and gel pieces were separated via centrifugation and the eluant was cleaned up using the DNA extraction from polyacrylamide gels protocol and reagents from the NucleoSpin® Gel and PCR Clean-Up Kit (Takara). The cDNA was eluted with 15  $\mu$ L, 10 mM Tris-HCl pH 8.8.

#### CircLigase II reaction

The purified cDNA was incubated with 5U CircLigase™ II ssDNA Ligase (Biosearch), 1X CircLigase Reaction Buffer, 2.5 mM  $\text{MgCl}_2$ , and 1 M Betaine overnight at 60°C. The reactions were inactivated for 10 minutes at 80°C.

#### Restriction digest

The circularized ssDNA reaction was incubated with 1X rCutSmart Buffer and 20 units of HhaI (NEB, total volume = 50  $\mu\text{L}$ ) overnight at 37°C. The reaction was heat inactivated for 20 minutes at 65°C.

#### Sequencing library

The restriction digested cDNA was subjected to PCR via sequencing primers and PrimeSTAR Max DNA Polymerase (Takara). The PCR products were separated and purified following the gel purification protocol from above, except the sequencing library was eluted with 30  $\mu\text{L}$ , 10 mM Tris-HCl pH 8.8 and quantified via the Qubit microRNA Assay Kit.

## REFERENCES

1. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nature Reviews Microbiology* **9**, 85–98 (2011).
2. Grohmann, D., Hirtreiter, A. & Werner, F. Molecular mechanisms of archaeal RNA polymerase. *Biochem. Soc. Trans.* **37**, 12–17 (2009).
3. Hirata, A., Klein, B. J. & Murakami, K. S. The X-ray crystal structure of RNA polymerase from Archaea. *Nature* **451**, 851–854 (2008).
4. Nagy, J. *et al.* Complete architecture of the archaeal RNA polymerase open complex from single-molecule FRET and NPS. *Nat. Commun.* **6**, 6161 (2015).
5. Jun, S. H., Reichlen, M. J., Tajiri, M. & Murakami, K. S. Archaeal RNA polymerase and transcription regulation. *Crit. Rev. Biochem. Mol. Biol.* **46**, 27–40 (2011).
6. Hirata, A. *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. *Mol. Microbiol.* **70**, 623–33 (2008).
7. Mayer, A., Landry, H. M. & Churchman, L. S. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Current Opinion in Cell Biology* **46**, 72–80 (2017).
8. Smirnov, E., Hornáček, M., Vacík, T., Cmarko, D. & Raška, I. Discontinuous transcription. *Nucleus* **9**, 149–160 (2018).
9. Washburn, R. S. & Gottesman, M. E. Regulation of transcription elongation and termination. *Biomolecules* **5**, 1063–1078 (2015).
10. Lange, U. & Hausner, W. Transcriptional fidelity and proofreading in Archaea and implications for the mechanism of TFS-induced RNA cleavage. *Mol. Microbiol.* **52**, 1133–1143 (2004).
11. Fish, R. N. & Kane, C. M. Promoting elongation with transcript cleavage stimulatory factors. *Biochimica et Biophysica Acta - Gene Structure and Expression* **1577**, 287–307 (2002).
12. Orlova, M., Newlands, J., Das, A., Goldfarb, A. & Borukhov, S. Intrinsic transcript cleavage activity of RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 4596–600 (1995).
13. Deighan, P. & Hochschild, A. Conformational toggle triggers a modulator of RNA polymerase activity. *Trends in Biochemical Sciences* **31**, 424–426 (2006).
14. Hausner, W., Lange, U. & Musfeldt, M. Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase. *J. Biol. Chem.* **275**, 12393–12399 (2000).
15. Marr, M. T. & Roberts, J. W. Function of transcription cleavage factors GreA and GreB at a regulatory pause site. *Mol. Cell* **6**, 1275–1285 (2000).
16. Sarmiento, F., Mrázek, J. & Whitman, W. B. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon *Methanococcus maripaludis*. *Proc. Natl. Acad. Sci.* **110**, 4726–4731 (2013).
17. Zhang, C., Phillips, A. P. R., Wipfler, R. L., Olsen, G. J. & Whitaker, R. J. The essential genome of the crenarchaeal model *Sulfolobus islandicus*. *Nat. Commun.* **9**, 1–11 (2018).
18. Fouqueau, T. *et al.* The transcript cleavage factor paralogue TFS4 is a potent RNA polymerase inhibitor. *Nat. Commun.* **8**, 1–13 (2017).
19. Jeon, C., Yoon, H. & Agarwal, K. The transcription factor TFIIS zinc ribbon dipeptide Asp-Glu is critical for stimulation of elongation and RNA cleavage by RNA polymerase II. *Proc. Natl. Acad. Sci.* **91**, 9106–9110 (1994).
20. Szádeczky-Kardoss, I. *et al.* Elongation factor TFIIS is essential for heat stress adaptation in plants. *Nucleic Acids Res.* **50**, 1927–1950 (2022).
21. Jha, R. K. *et al.* Conditional down-regulation of GreA impacts expression of rRNA and transcription factors, affecting *Mycobacterium smegmatis* survival. *Sci. Rep.* **10**, 5802

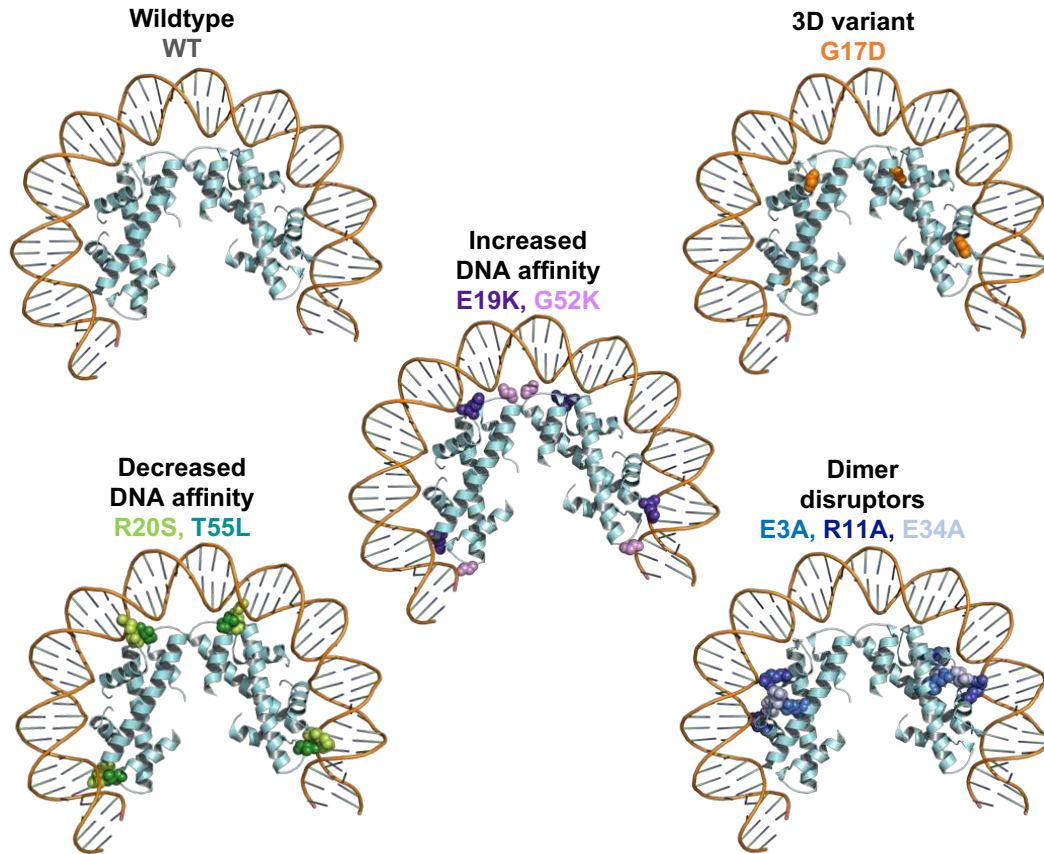
- (2020).
22. Feng, S. *et al.* Involvement of Transcription Elongation Factor GreA in Mycobacterium Viability, Antibiotic Susceptibility, and Intracellular Fitness. *Front. Microbiol.* **11**, (2020).
  23. Sanders, T. J. *et al.* TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Mol. Microbiol.* **111**, 784–797 (2019).
  24. Xie, Y. & Reeve, J. N. Transcription by an archaeal RNA polymerase is slowed but not blocked by an archaeal nucleosome. *J. Bacteriol.* **186**, 3492–3498 (2004).
  25. Chang, C. H. & Luse, D. S. The H3/H4 tetramer blocks transcript elongation by RNA polymerase II in vitro. *J. Biol. Chem.* **272**, 23427–23434 (1997).
  26. Mariño-Ramírez, L., Kann, M. G., Shoemaker, B. A. & Landsman, D. Histone structure and nucleosome stability. *Expert Review of Proteomics* **2**, 719–729 (2005).
  27. Teves, S. S., Weber, C. M. & Henikoff, S. Transcribing through the nucleosome. *Trends Biochem. Sci.* **39**, 577–586 (2014).
  28. Soman, A. *et al.* Columnar structure of human telomeric chromatin. *Nat.* 2022 6097929 **609**, 1048–1055 (2022).
  29. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
  30. Mattioli, F. *et al.* Structure of histone-based chromatin in Archaea. *Science (80-. ).* **357**, 609–612 (2017).
  31. Stevens, K. M., Hoher, A. & Warnecke, T. Deep Conservation of Histone Variants in Thermococcales Archaea. *Genome Biol. Evol.* **14**, (2022).
  32. Stevens, K. M. *et al.* Histone variants in archaea and the evolution of combinatorial chromatin complexity. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33384–33395 (2020).
  33. Erkelens, A. M., Henneman, B., van der Valk, R. A., Kirolos, N. C. S. & Dame, R. T. Specific DNA binding of archaeal histones HMfA and HMfB. *Front. Microbiol.* **14**, (2023).
  34. Sandman, K., Grayling, R. A., Dobrinski, B., Lurz, R. & Reeve, J. N. Growth-phase-dependent synthesis of histones in the archaeon *Methanothermobacter thermautotrophicus*. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 12624–12628 (1994).
  35. Bailey, K. A., Marc, F., Sandman, K. & Reeve, J. N. Both DNA and histone fold sequences contribute to archaeal nucleosome stability. *J. Biol. Chem.* **277**, 9293–9301 (2002).
  36. Marc, F., Sandman, K., Lurz, R. & Reeve, J. N. Archaeal Histone Tetramerization Determines DNA Affinity and the Direction of DNA Supercoiling. *J. Biol. Chem.* **277**, 30879–30886 (2002).
  37. Peeters, E., Driessen, R. P. C., Werner, F. & Dame, R. T. The interplay between nucleoid organization and transcription in archaeal genomes. *Nature Reviews Microbiology* **13**, 333–341 (2015).
  38. Sanders, T. J. *et al.* Extended Archaeal Histone-Based Chromatin Structure Regulates Global Gene Expression in *Thermococcus kodakarensis*. *Front. Microbiol.* **12**, 1071 (2021).
  39. Tagashira, K. *et al.* Genetic studies on the virus-like regions in the genome of hyperthermophilic archaeon, *Thermococcus kodakarensis*. *Extremophiles* **17**, 153–160 (2013).
  40. Nojima, T., Gomes, T., Carmo-Fonseca, M. & Proudfoot, N. J. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat. Protoc.* **11**, 413–428 (2016).
  41. Imashimizu, M. *et al.* Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. *Genome Biol.* **16**, (2015).
  42. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368–373 (2011).

43. Hall, M. A. *et al.* High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nat. Struct. Mol. Biol.* **16**, 124–129 (2009).
44. Nalabothula, N. *et al.* Archaeal nucleosome positioning in vivo and in vitro is directed by primary sequence motifs. *BMC Genomics* **14**, 391 (2013).
45. Fukui, T. *et al.* Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. *Genome Res.* **15**, 352–363 (2005).
46. Gehring, A. M. & Santangelo, T. J. Manipulating archaeal systems to permit analyses of transcription elongation-termination decisions in vitro. *Methods Mol. Biol.* **1276**, 263–79 (2015).
47. Nojima, T. *et al.* Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).
48. Sun, Z., Yakhnin, A. V., FitzGerald, P. C., McIntosh, C. E. & Kashlev, M. Nascent RNA sequencing identifies a widespread sigma70-dependent pausing regulated by Gre factors in bacteria. *Nat. Commun.* **2021 121 12**, 1–14 (2021).
49. Kennedy-Darling, J. & Smith, L. M. Measuring the Formaldehyde Protein–DNA Cross-Link Reversal Rate. *Anal. Chem.* **86**, 5678–5681 (2014).
50. Gehring, A. M., Sanders, T. J. & Santangelo, T. J. Markerless Gene Editing in the Hyperthermophilic Archaeon *Thermococcus kodakarensis*. *Bio-protocol* **7**, (2017).
51. Čuboňová, L. *et al.* An archaeal histone is required for transformation of *Thermococcus kodakarensis*. *J. Bacteriol.* **194**, 6864–6874 (2012).
52. Burkhart, B. W., Febvre, H. P. & Santangelo, T. J. Distinct Physiological Roles of the Three Ferredoxins Encoded in the Hyperthermophilic Archaeon *Thermococcus kodakarensis*. *MBio* **10**, (2019).
53. Sanders, T. J. *et al.* FttA is a CPSF73 homologue that terminates transcription in Archaea. *Nature Microbiology* **5**, 545–553 (2020).
54. Li, J. *et al.* aCPSF1 cooperates with terminator U-tract to dictate archaeal transcription termination efficacy. *Elife* **10**, (2021).
55. Yue, L. *et al.* The conserved ribonuclease aCPSF1 triggers genome-wide transcription termination of Archaea via a 3'-end cleavage mode. *Nucleic Acids Res.* (2020). doi:10.1093/nar/gkaa702
56. Wenck, B. R. & Santangelo, T. J. Archaeal transcription. *Transcription* **11**, 199–210 (2020).
57. Dar, D., Prasse, D., Schmitz, R. A. & Sorek, R. Widespread formation of alternative 3' UTR isoforms via transcription termination in archaea. *Nat. Microbiol.* **1**, 1–9 (2016).
58. Blombach, F., Fouqueau, T., Matelska, D., Smollett, K. & Werner, F. Promoter-proximal elongation regulates transcription in archaea. *Nat. Commun.* **2021 121 12**, 1–15 (2021).
59. TJ, S. *et al.* Extended Archaeal Histone-Based Chromatin Structure Regulates Global Gene Expression in *Thermococcus kodakarensis*. *Front. Microbiol.* **12**, (2021).
60. Walker, J. E., Luyties, O. & Santangelo, T. J. Factor-dependent archaeal transcription termination. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6767–E6773 (2017).
61. Santangelo, T. J., Čuboňová, L. L., James, C. L. & Reeve, J. N. TFB1 or TFB2 Is Sufficient for *Thermococcus kodakaraensis* Viability and for Basal Transcription in Vitro. *J. Mol. Biol.* **367**, 344–357 (2007).

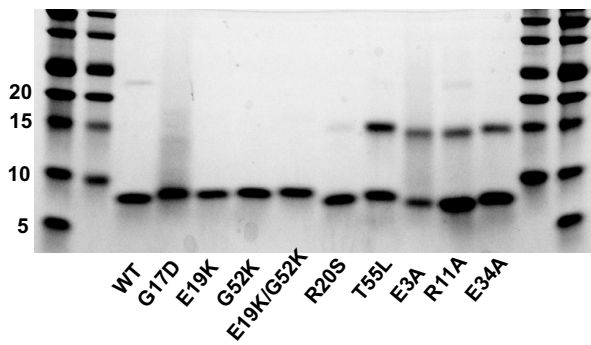
# APPENDIX A: CHAPTER 2 SUPPLEMENTARIES

## Supplementary Figures

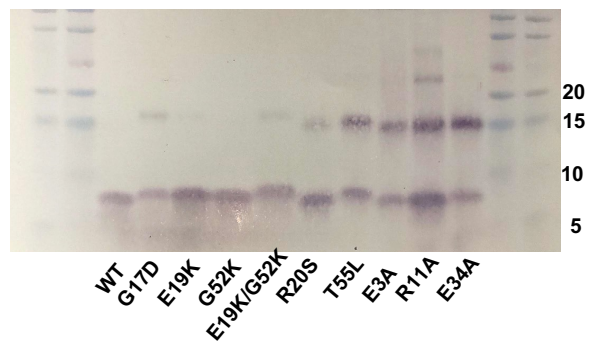
**a**



**b**



**c**



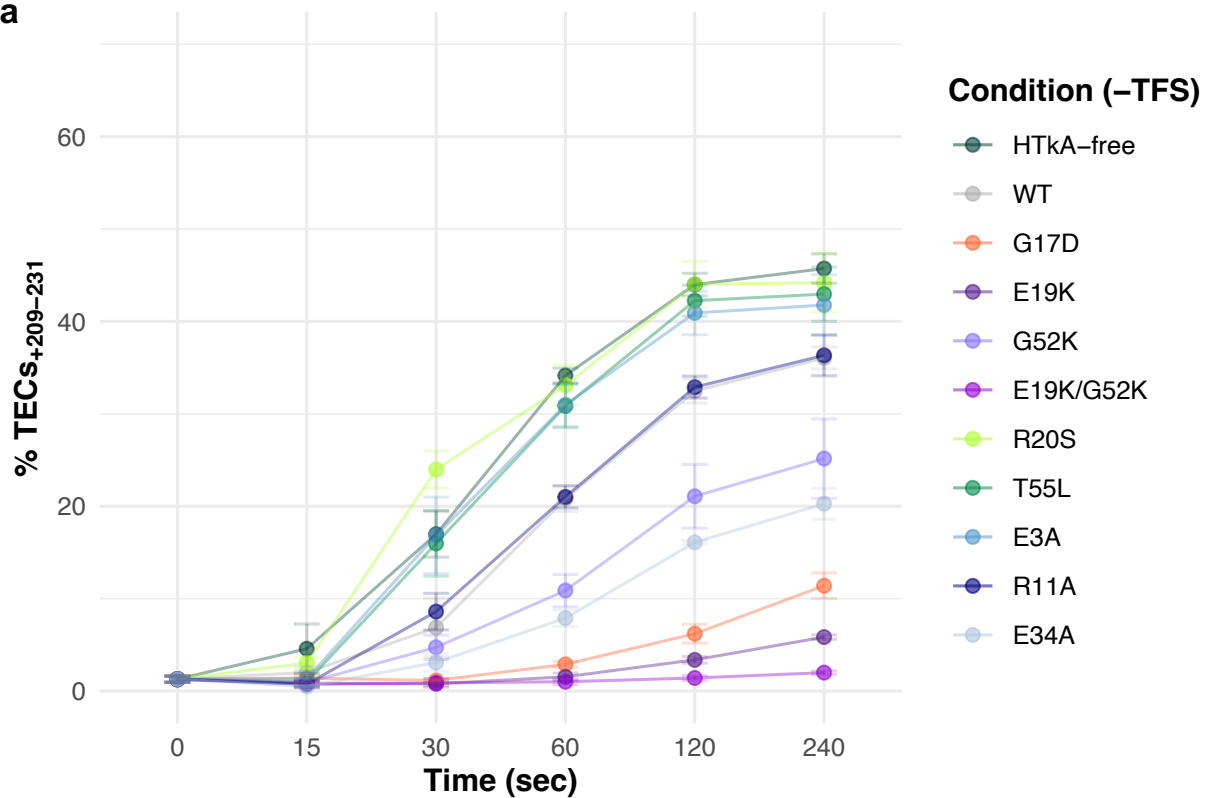
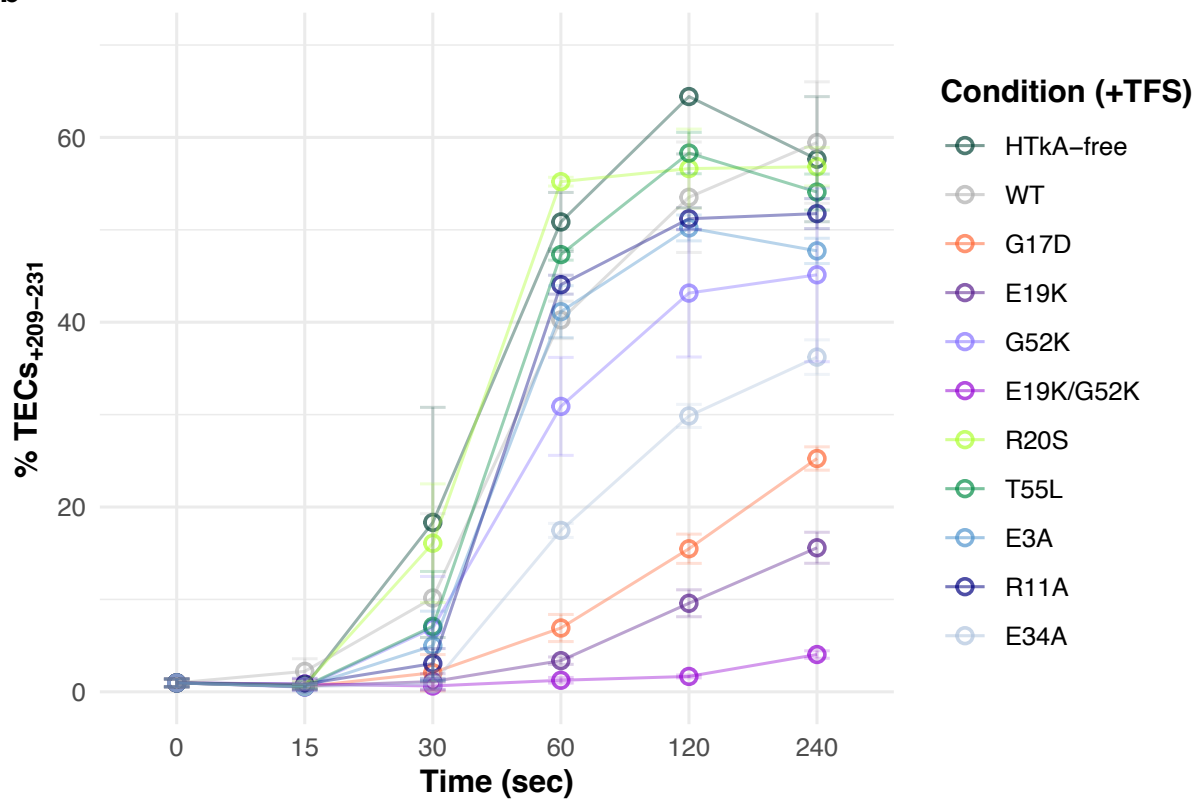


**Figure A.1. Representations of variant archaeal histone-based chromatin structures formed with purified HTkA<sup>WT</sup> or HTkA<sup>variants</sup> tetramers bound to ~60 bp of DNA.**

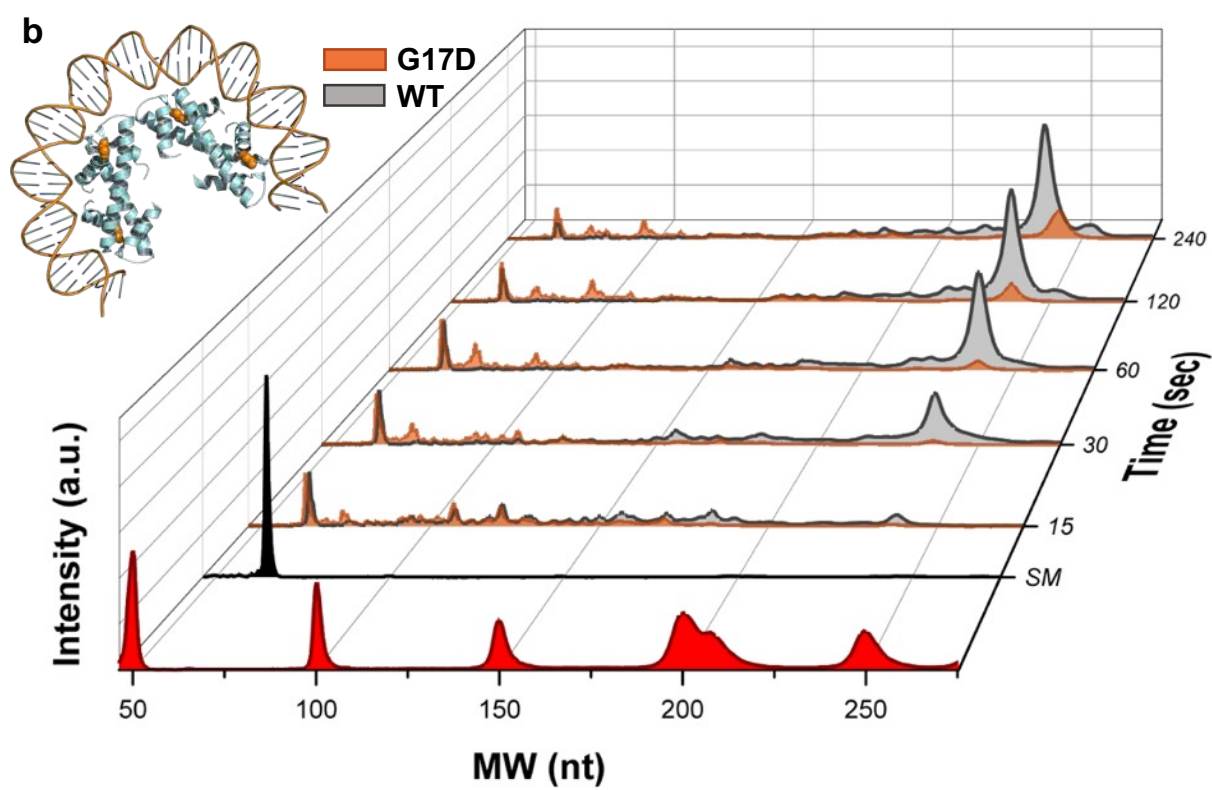
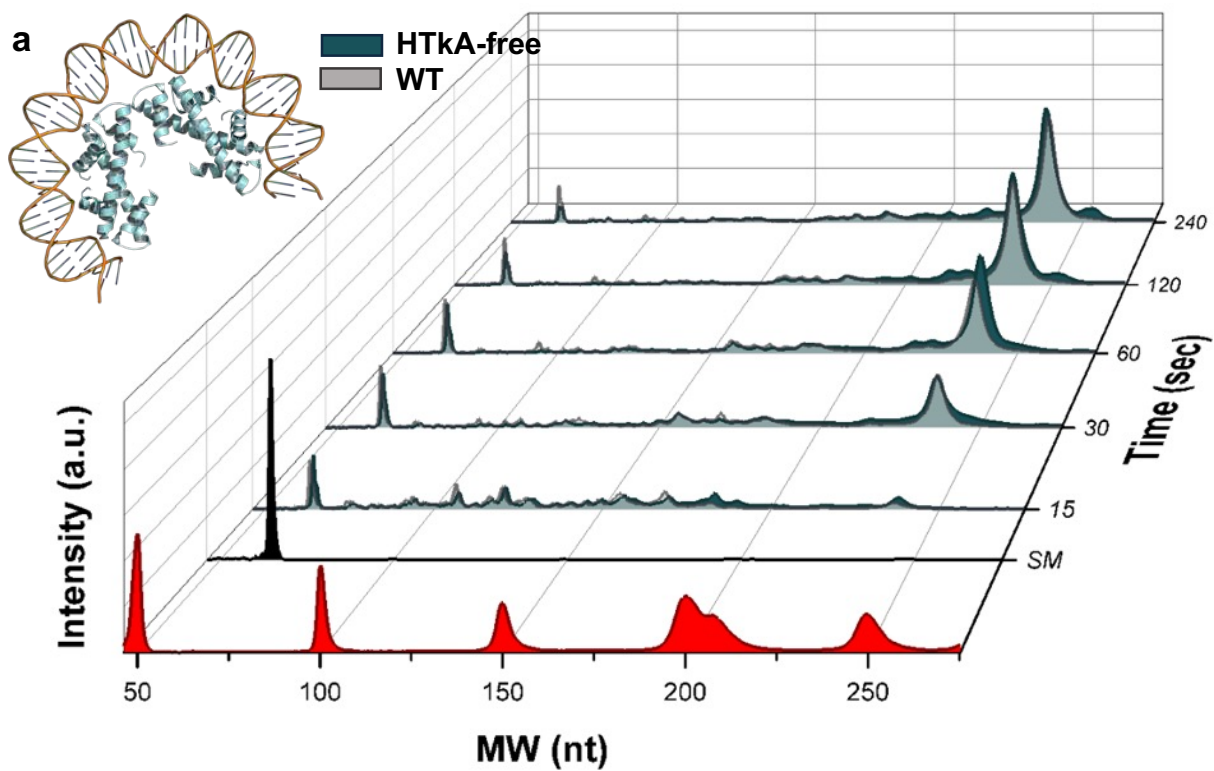
**a.** Cartoon representations of HTkA<sup>WT</sup> or HTkA<sup>variants</sup> tetrasome structures. HTkA-variants were designed to hinder formation of a tightly packed 3D structure of archaeal-histone based chromatin (HTkA<sup>G17D</sup>), increase (HTkA<sup>E19K</sup>, HTkA<sup>G52K</sup>, and HTkA<sup>E19K/G52K</sup>) or decrease (HTkA<sup>R20S</sup> and HTkA<sup>T55L</sup>) histone-DNA affinity, or interfere with internal dimer, or dimer-dimer interactions supporting tetrasome formation (HTkA<sup>E3A</sup>, HTkA<sup>R11A</sup>, and HTkA<sup>E34A</sup>). **b.** Denaturing gels reveal the expected migration of recombinantly expressed and purified HTkA proteins primarily as monomers although some variants resolved as dimers or higher-order complexes during SDS-PAGE visualized with Coomassie brilliant blue R-250. **c.** A Western blot employing polyclonal anti-HTkA antibodies confirms that higher order structures revealed by SDS-PAGE are multimeric histone complexes. Protein molecular weight markers provide standards in kDa.

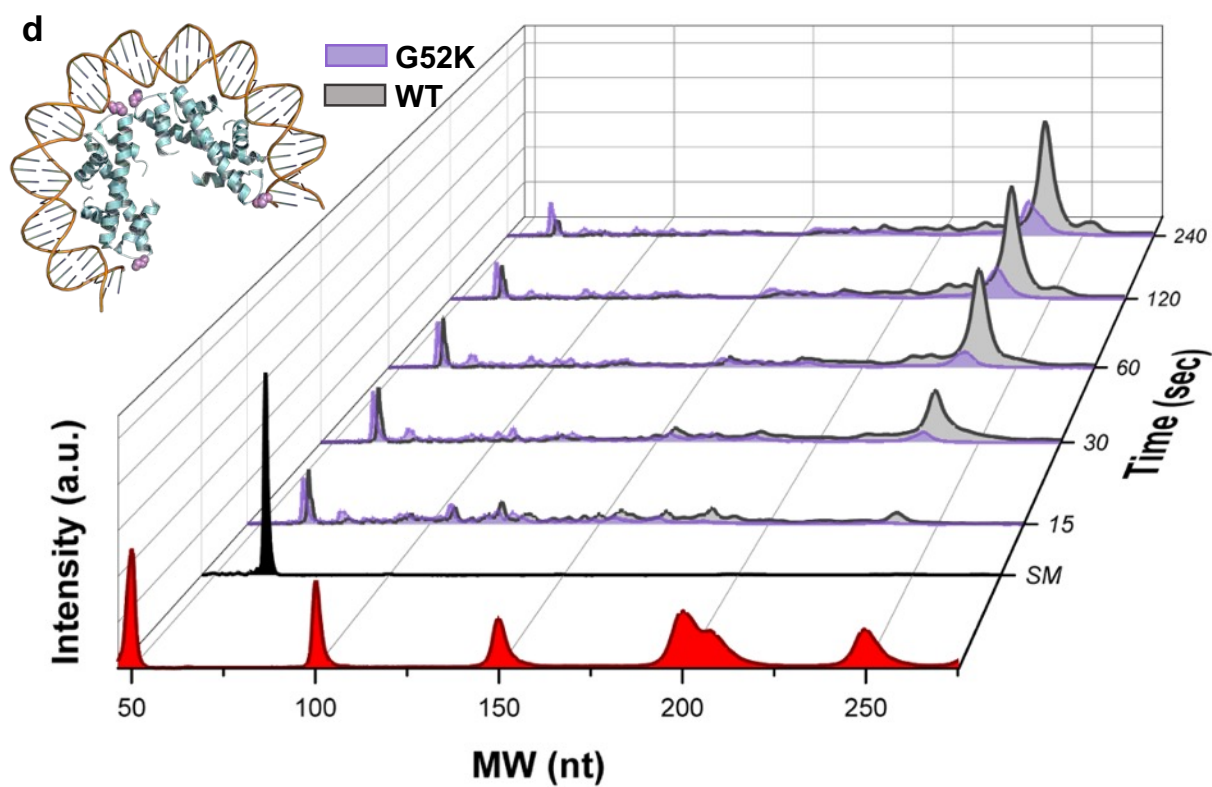
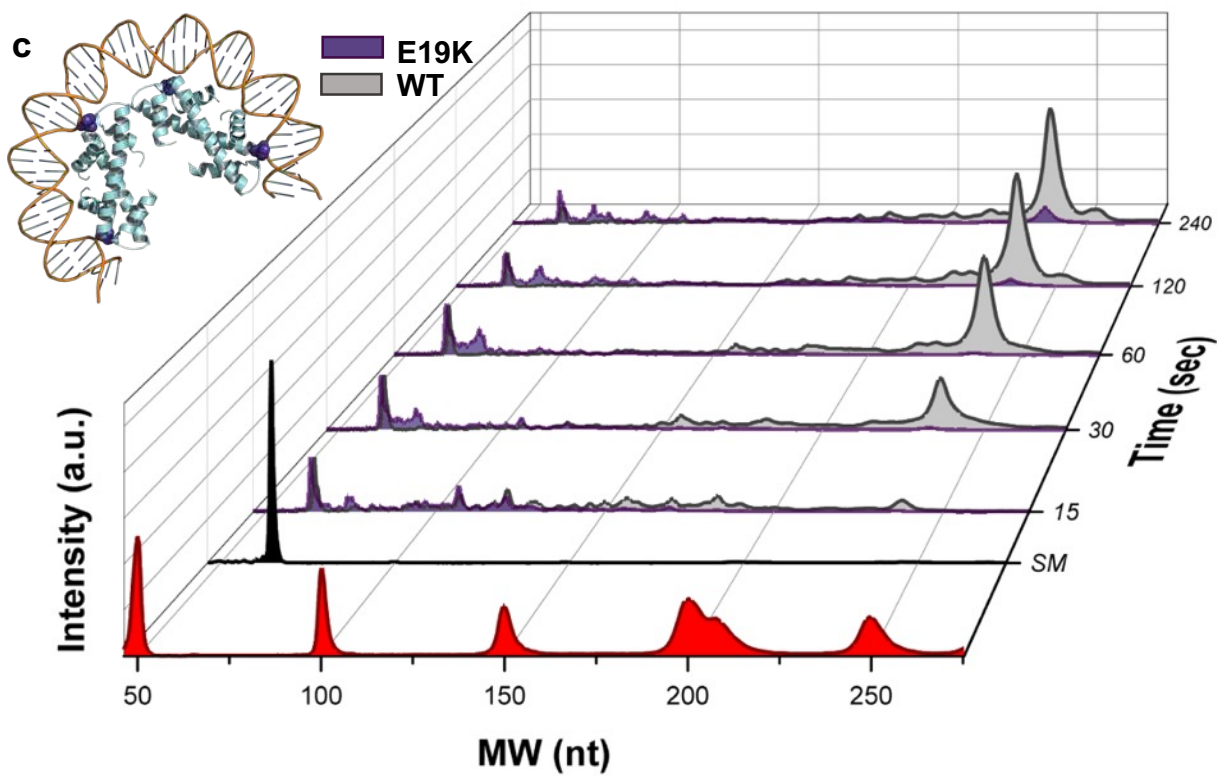
*Promoter/ApC (TSS) /C-less cassette(+58 walk)/ Double SELEX 60 bp HPS*  
TCCGCATGCGAGCTCGGTACCCCGGGGAGCGATATATTTATATAGGGATATAGTAATAGATAATATC  
**ACATGTGTGTTGTGAAATATTGTAGTTGGATGTAGGTTGGAGTGGAGTGTGTGGTGTGCCCGGATCC**  
GATATCAACCGTACTGGTGTTCCTACGCTAATCTAAGCCGTTTACTCGCGATTTTGAAAATAGCT  
TAGGTGGTGGTGTTCCTACGCTAATCTAAGCCGTTTACTCGCGATTTTGAAAATAGCTTAGGTGG  
AGATCTGATATCAAGCTTCCCCCGGGCCCG

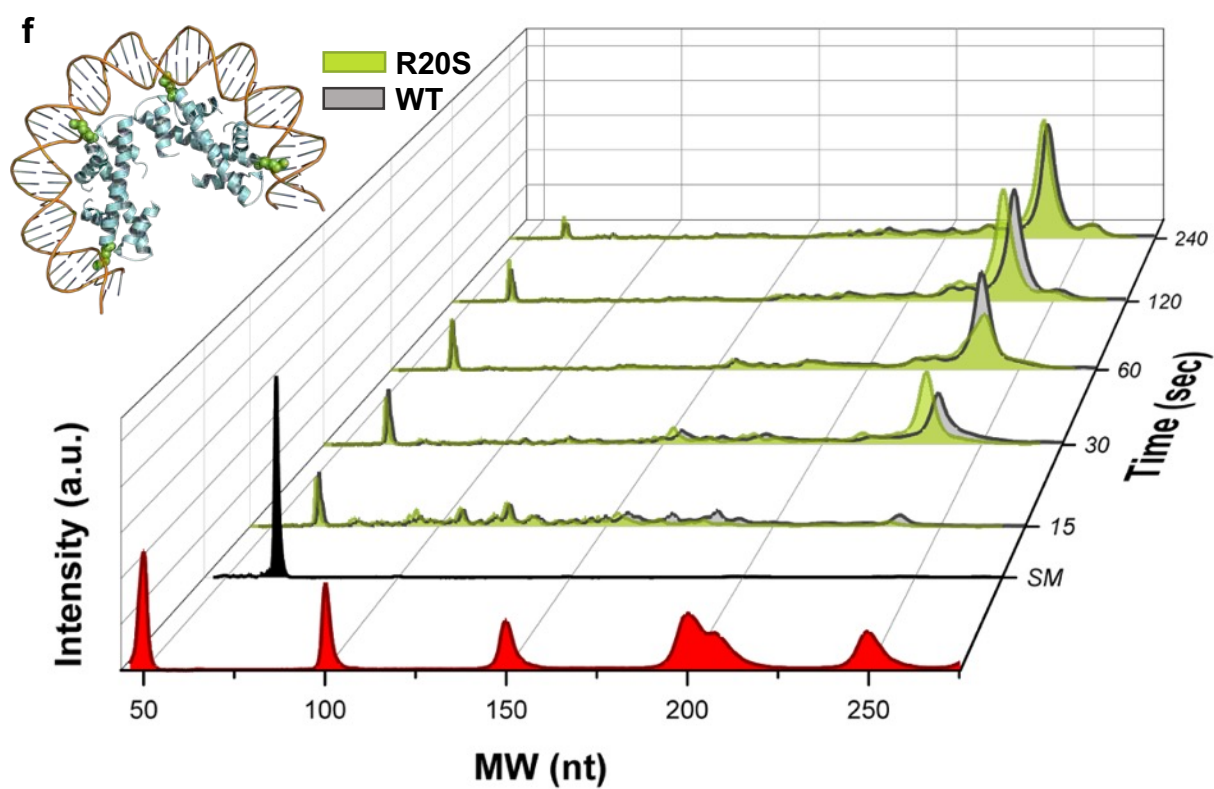
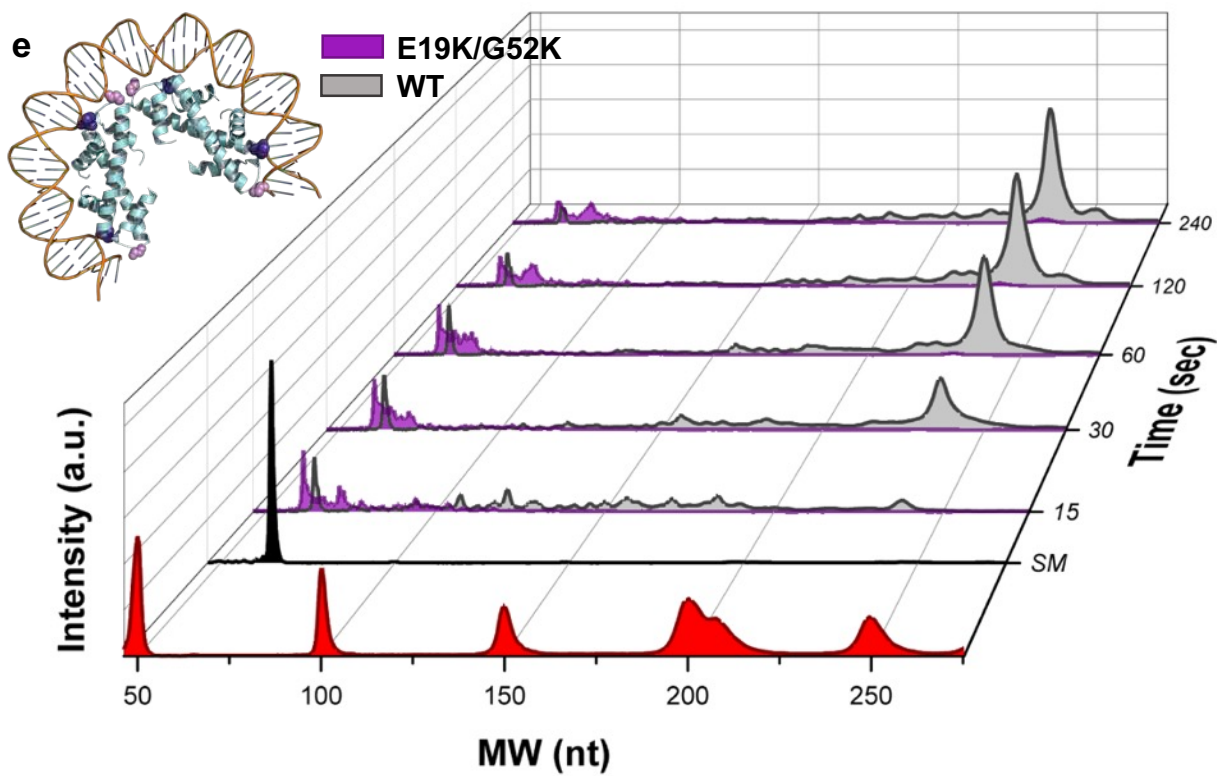
**Figure A.2. 298 bp DNA template.** DNA template designed with a robust promoter sequence (*italics*), a defined transcription start-site (**bold**), a c-less cassette for a +58 walk-out (*italics*), and a SELEX-derived double 60 bp histone positioning sequence (HPS) optimized for histone binding. The DNA template allows for walk-out to +58 bp so as not to interfere with initiation upon histone binding.

**a****b**

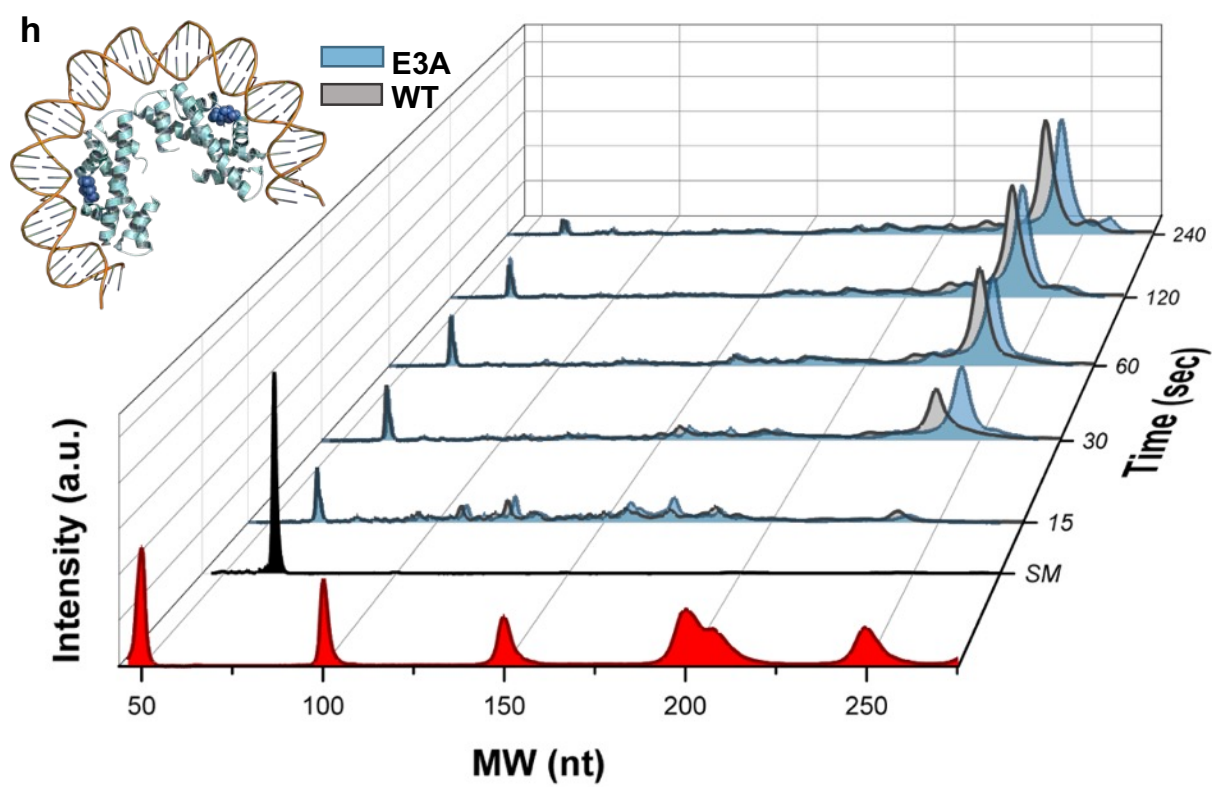
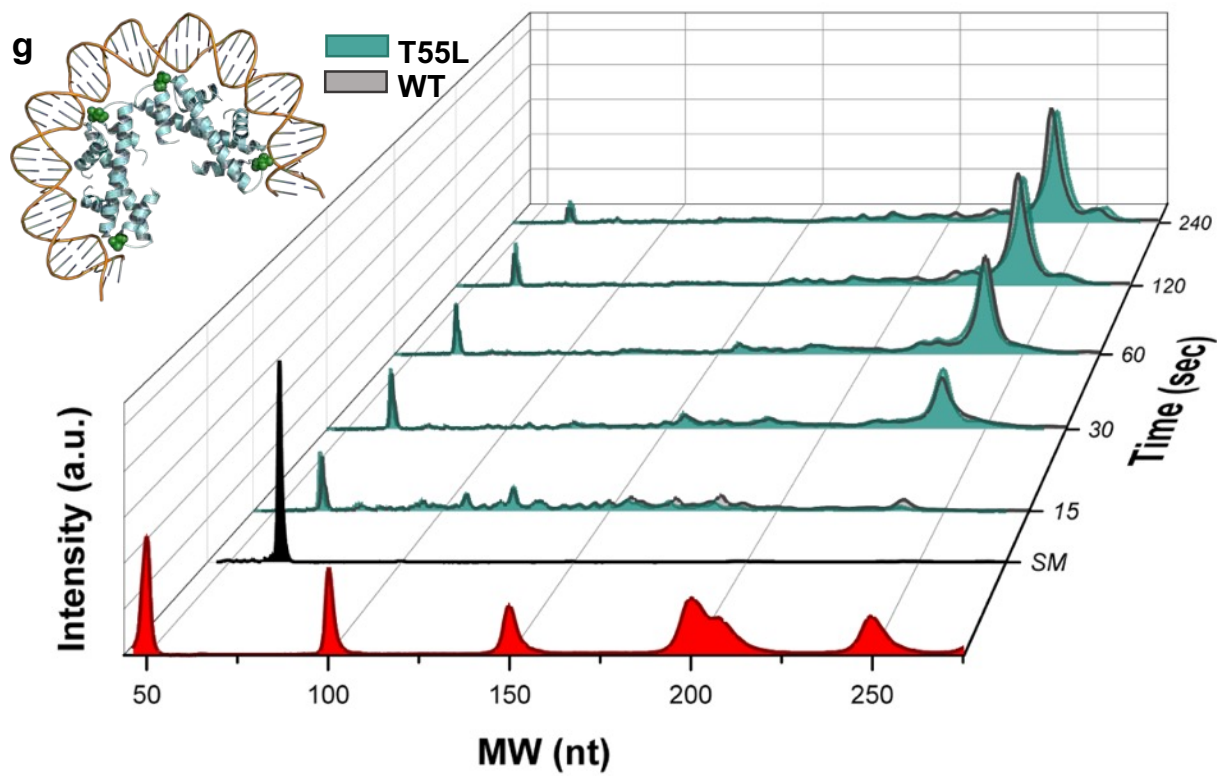
**Figure A.3. Histone residue substitutions have dramatic effects on full-length RNA transcript abundance.** **a.** The percentage of RNA transcripts that progress to full-length varies between HTkA-free, HTkA<sup>WT</sup>, and HTkA<sup>variant</sup> environments. **b.** Addition of TFS has a positive impact on the abundance of full-length RNA transcripts in all conditions but is unable to accelerate TECs through HTkA<sup>E19K</sup>-, HTkA<sup>E19K/G52K</sup>-, and HTkA<sup>G17D</sup>-based chromatin landscapes to match the RNA transcript abundance in HTkA<sup>WT</sup> with no TFS conditions.

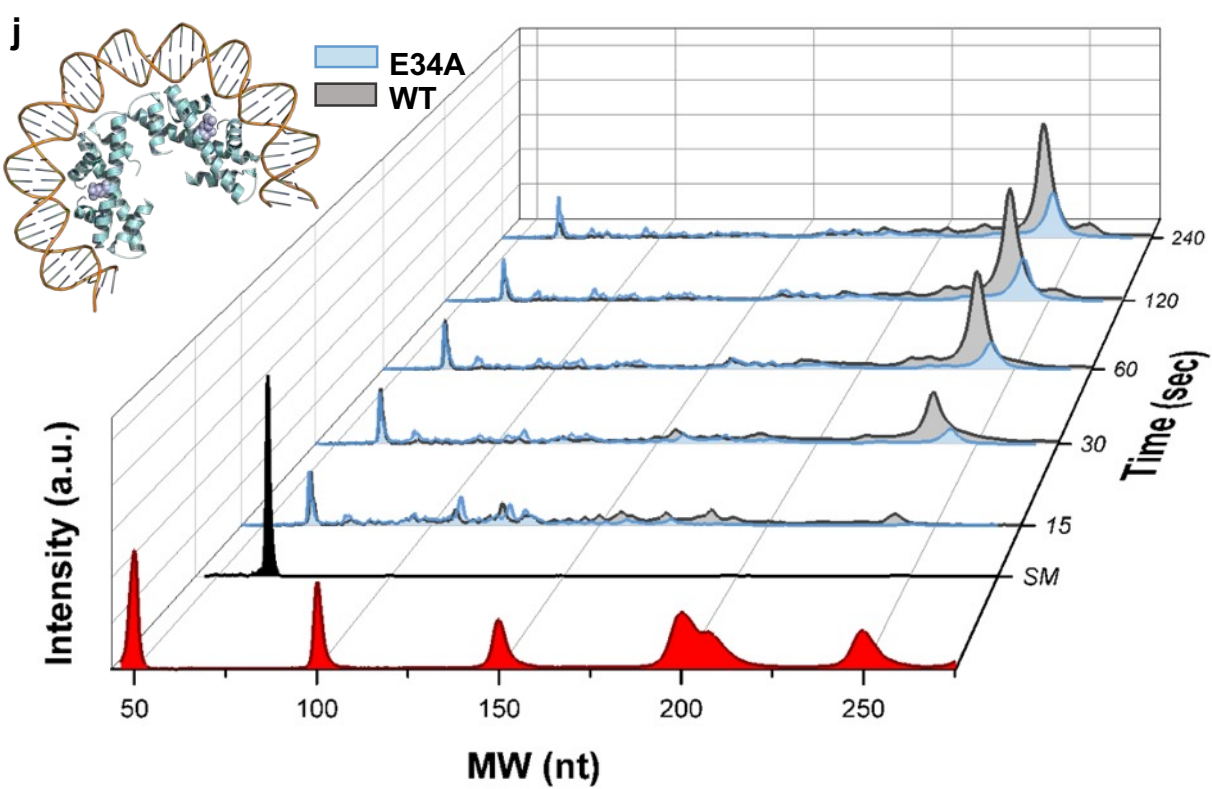
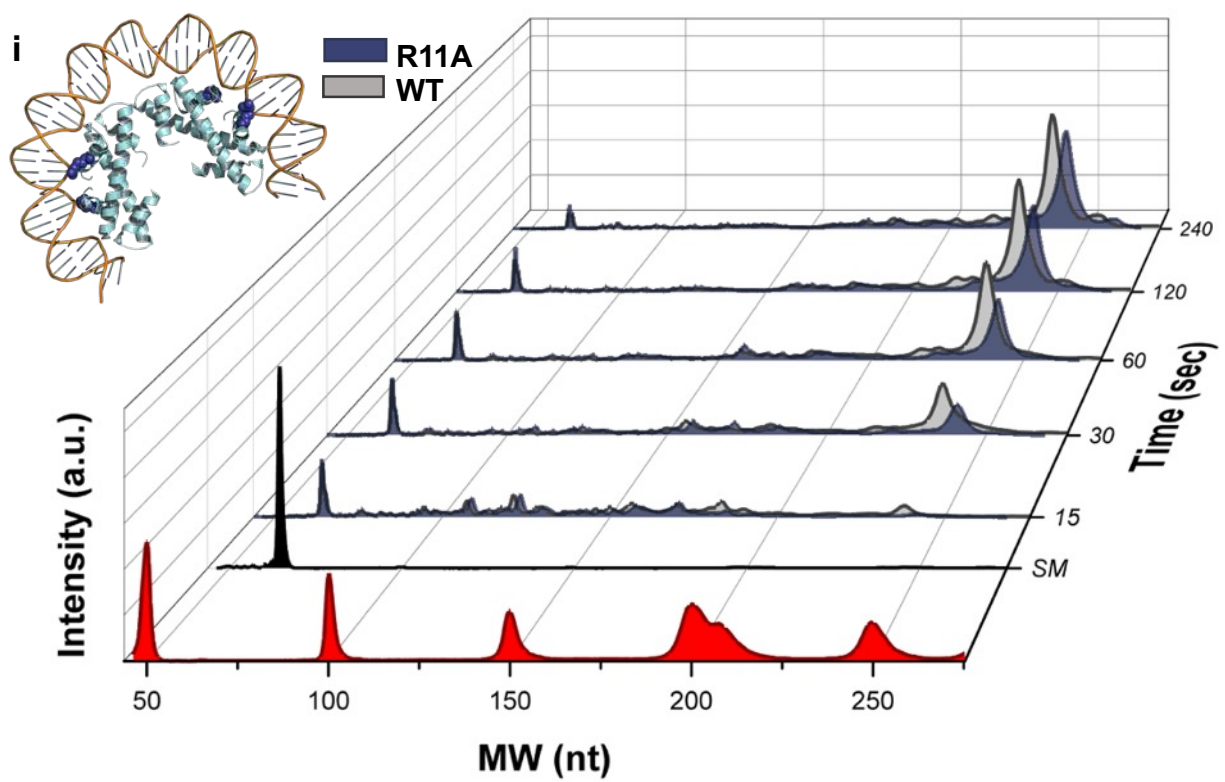






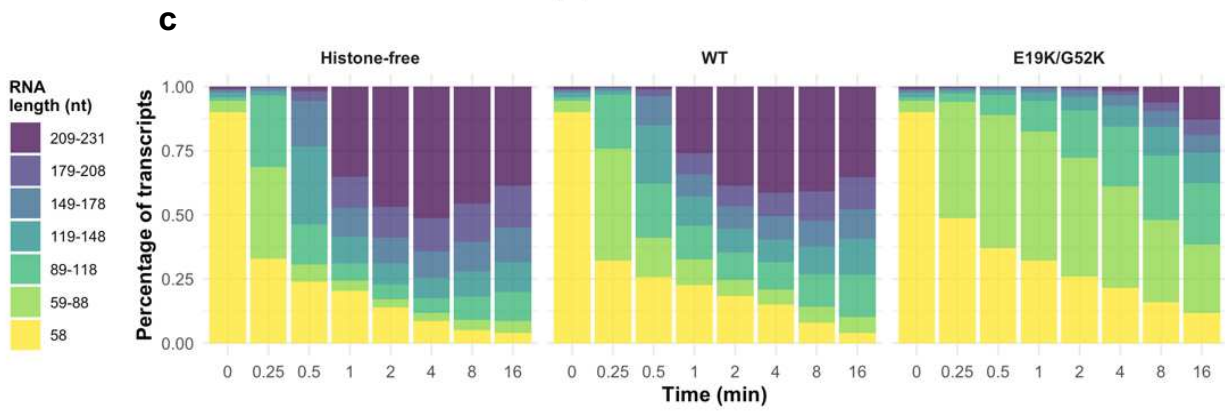
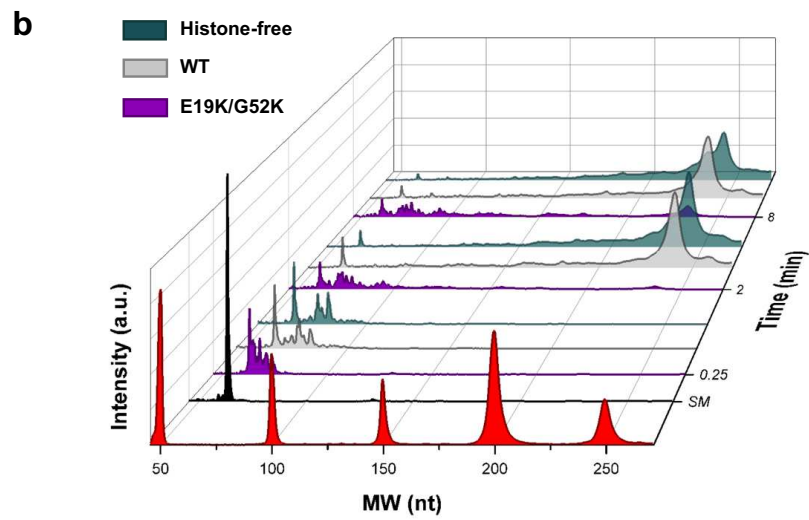
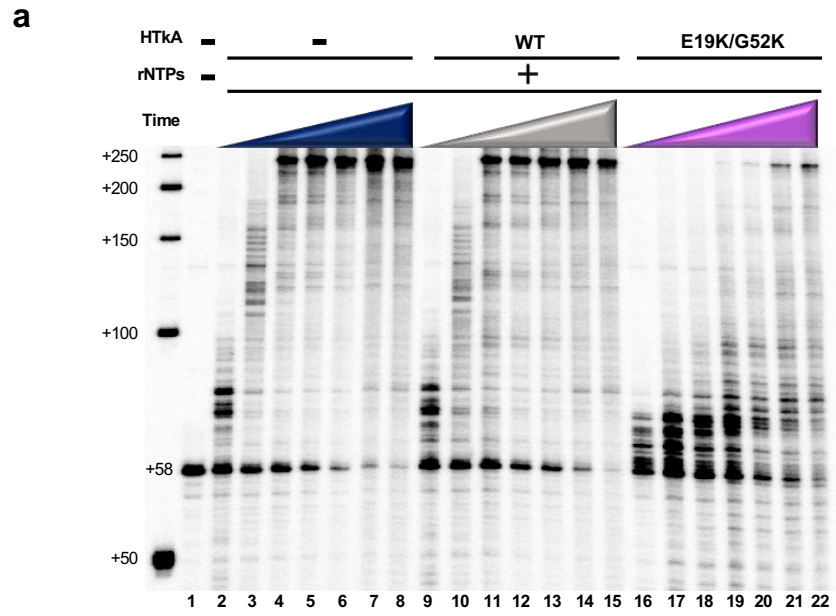




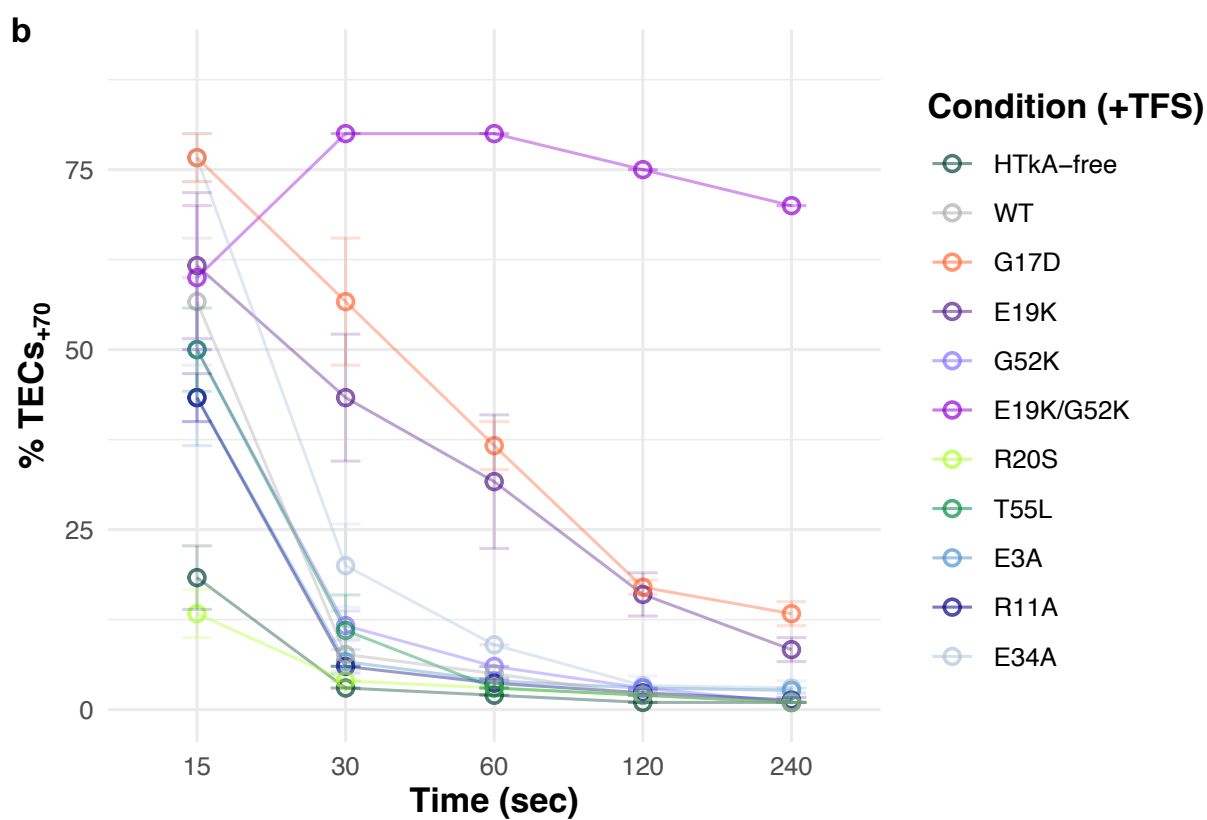
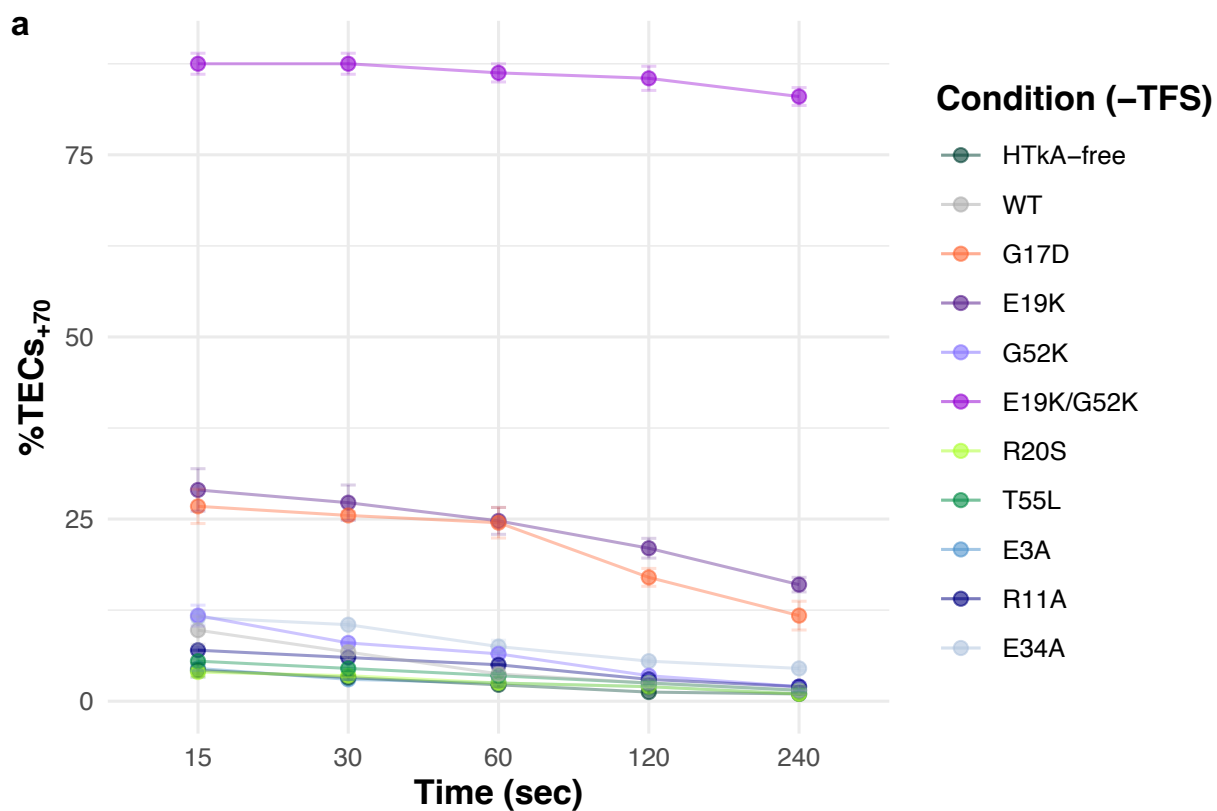




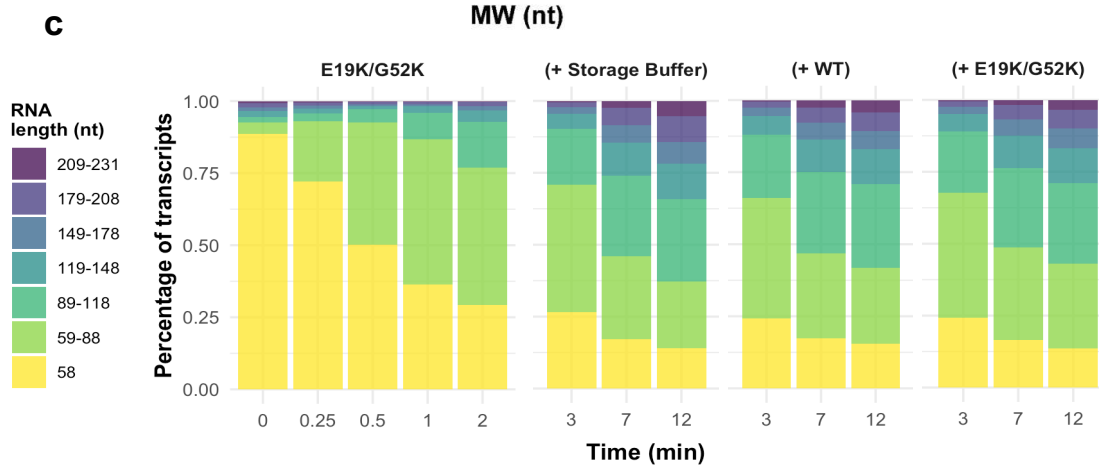
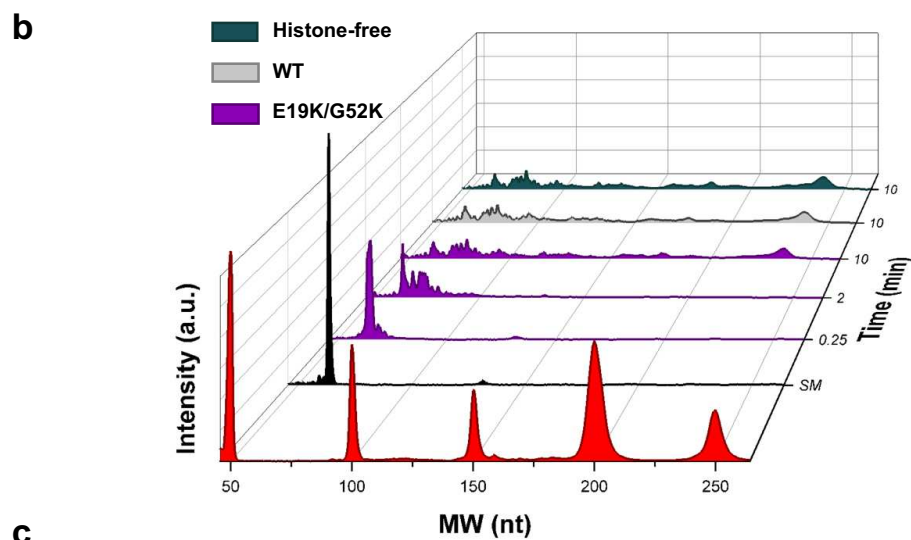
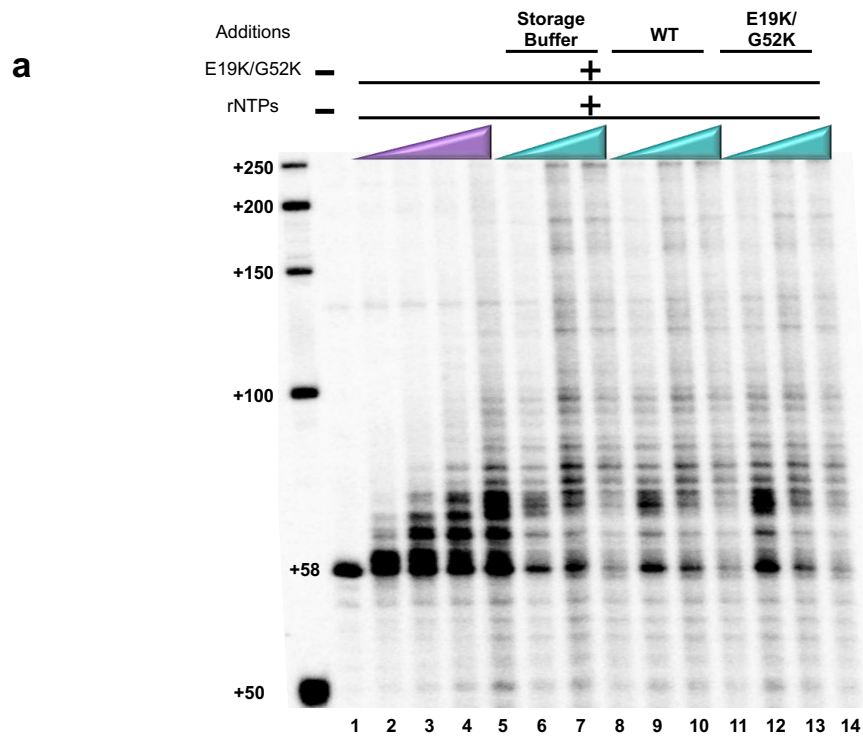
**Figure A.4. TEC progression is differentially impacted by distinct archaeal histone-based chromatin landscapes.** Waterfall plots permit quantification of the distribution of nascent transcript lengths over time, and in the presence or absence of TFS. The relative intensity of different transcript lengths was normalized to the sum of the counts in the starting material (SM) within each lane. Transcript abundance is compared for (a) histone-free (HTkA-free) and wildtype (HTkA<sup>WT</sup>), or between (b – j) HTkA<sup>WT</sup> and HTkA<sup>variants</sup>.



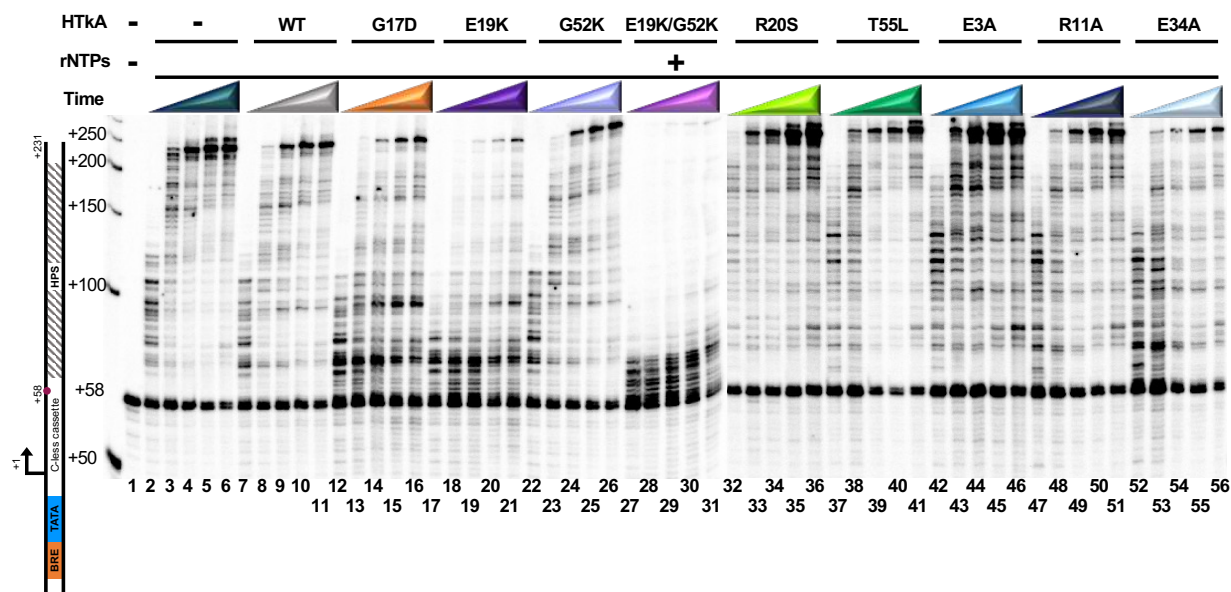
**Figure A.5. The dynamics of histone exchange on DNA can dramatically alter TEC progression.** **a.** Stalled TECs<sub>+58</sub> were incubated at 85°C without HTkA (lanes 2 - 8), with HTkA<sup>WT</sup> (lanes 9 - 15) or with HTkA<sup>E19K/G52K</sup> (lanes 16 – 22) prior to elongation restart. While elongation on histone-free or HTkA<sup>WT</sup>-bound templates permits rapid accumulation of full-length transcripts, HTkA<sup>E19K/G52K</sup> bound DNAs require extended incubation to permit RNAP to overcome the chromatin landscape. Radiolabeled ssDNA makers provide size standards. **b.** Waterfall plots permit quantification of the distribution of nascent transcript lengths over time; the relative intensity of different transcript lengths was normalized to the sum of the counts in the starting material (SM) within each lane. **c.** Transcript distributions are quantified to reveal the impediment to elongation imposed by HTkA<sup>E19K/G52K</sup> bound DNAs.



**Figure A.6. The first point of contact between TECs and downstream chromatin structures is rate limiting for RNA synthesis.** **a.** While TECs can traverse a WT chromatin landscape with minimal pausing, increasing the strength of histone-DNA contacts or disrupting the normally tight 3D packing of archaeal histone-based chromatin transiently pauses the majority of TECs. **b.** Addition of TFS to *in vitro* transcription reactions does not alter the +70-prominent pause significantly, implying TECs are not backtracked due to chromatin-based impediments to translocation.

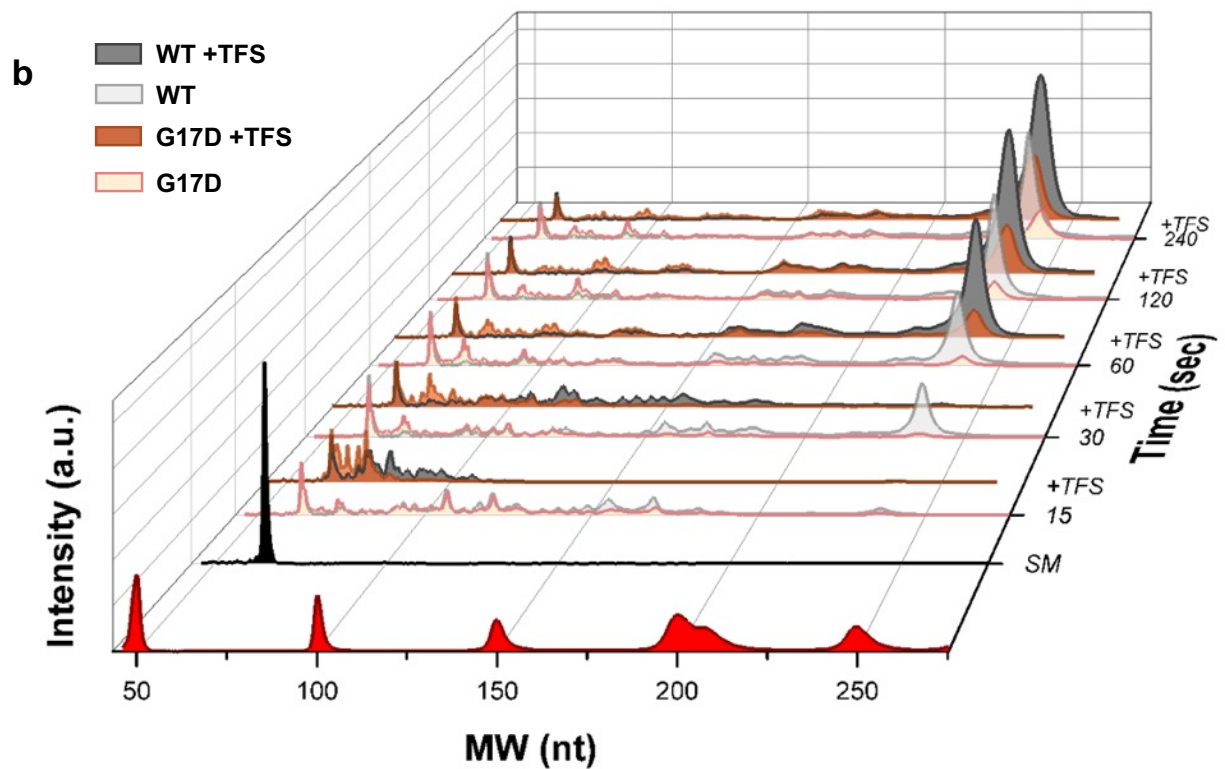
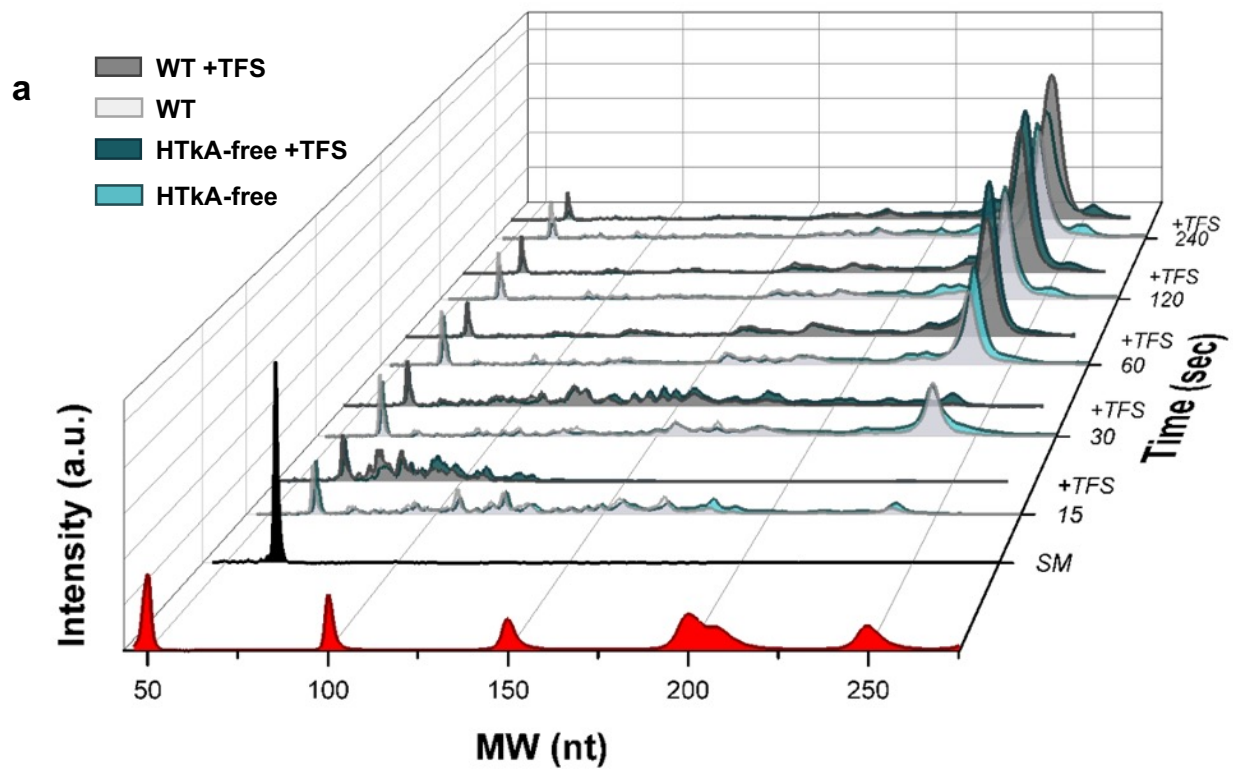


**Figure A.7. Dynamic exchange of histones on DNA is slow compared to TEC translocation.** **a.** Stalled TECs<sub>+58</sub> were incubated with HTkA<sup>E19K/G52K</sup>, then transcription was reinitiated by addition of all 4 rNTPs, and aliquots were collected after 15-, 30-, 60-, and 120-seconds to monitor RNA synthesis. TECs remaining after 2 minutes were split into thirds, and either storage buffer, HTkA<sup>WT</sup>, or additional HTkA<sup>E19K/G52K</sup> was added, before allowing additional time for continued RNA synthesis. Radiolabeled ssDNA makers provide size standards. **(b and c).** Quantitative comparisons of transcript lengths following dilution, exchange with HTkA<sup>WT</sup>, or maintenance of HTkA<sup>E19K/G52K</sup>- based chromatin reveal negligible difference in TEC progression, implying that histone exchange is limited during the time course of TEC translocation through the chromatin landscape.

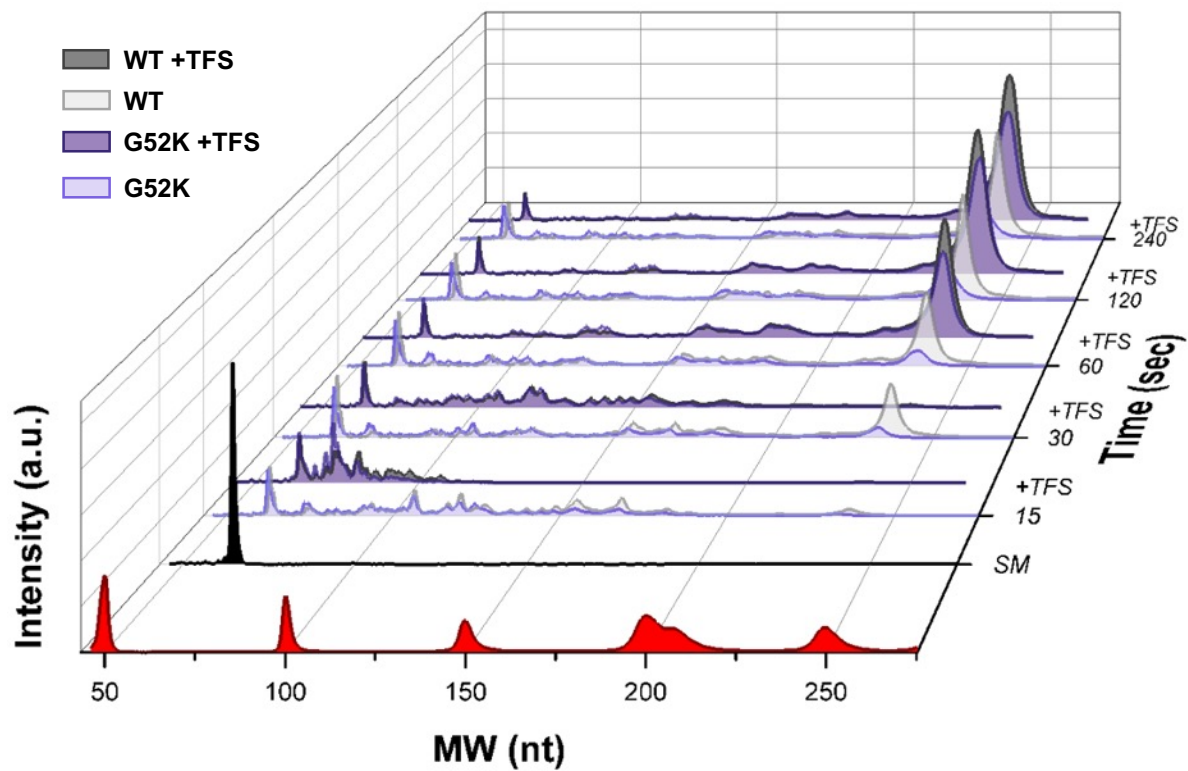


**Figure A.8. TFS increases productive elongation in native and most variant archaeal histone-based chromatin environments.** Continued RNA synthesis from TECs<sub>+58</sub> (lane 1) was monitored identically to Fig. 1, with the addition of TFS following transcription restart upon rNTP addition. Nascent transcript length was detected by collecting 15-, 30-, 60-, 120-, and 240-second aliquots. Radiolabeled ssDNA makers provide size standards. A representative gel image is shown; n = 3.

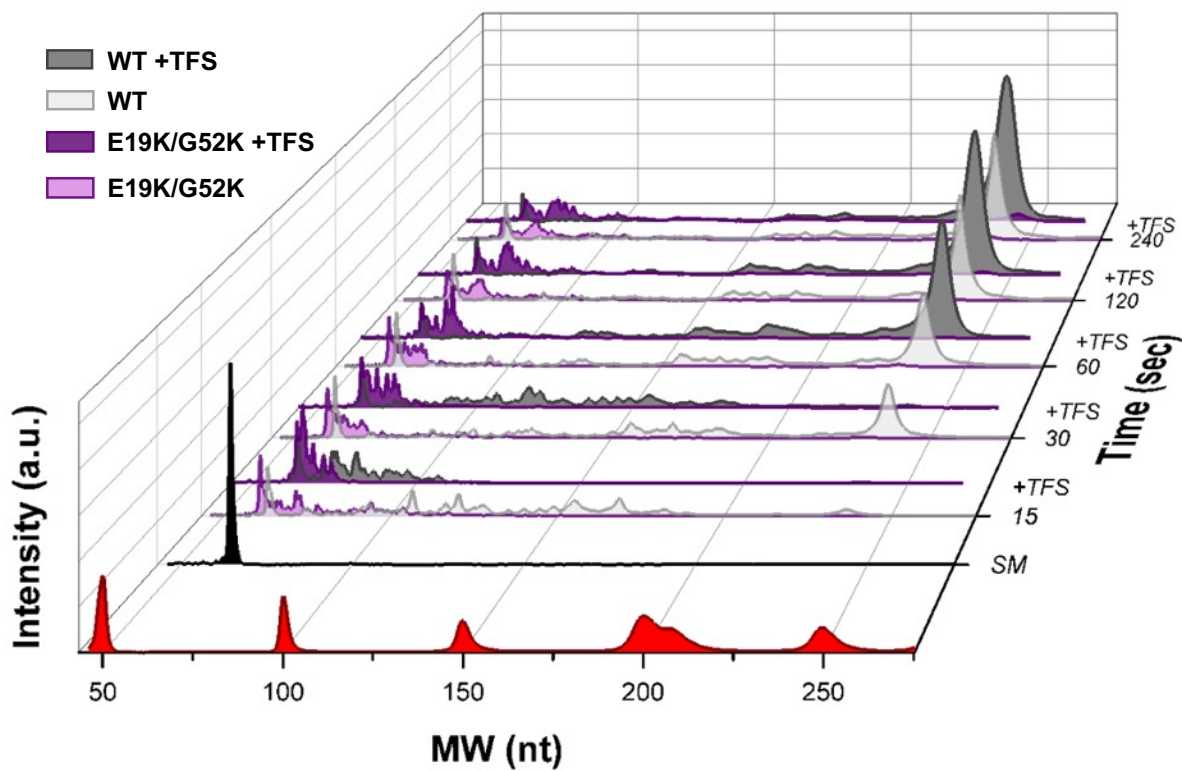




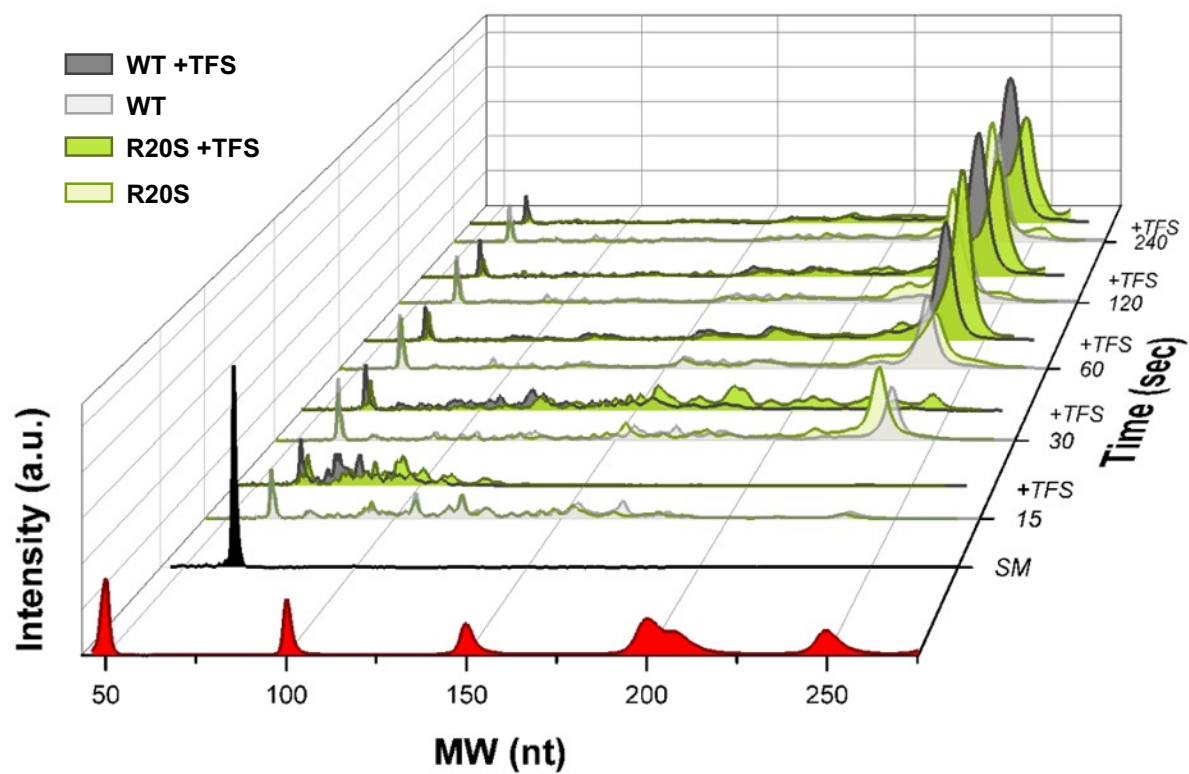
**c**



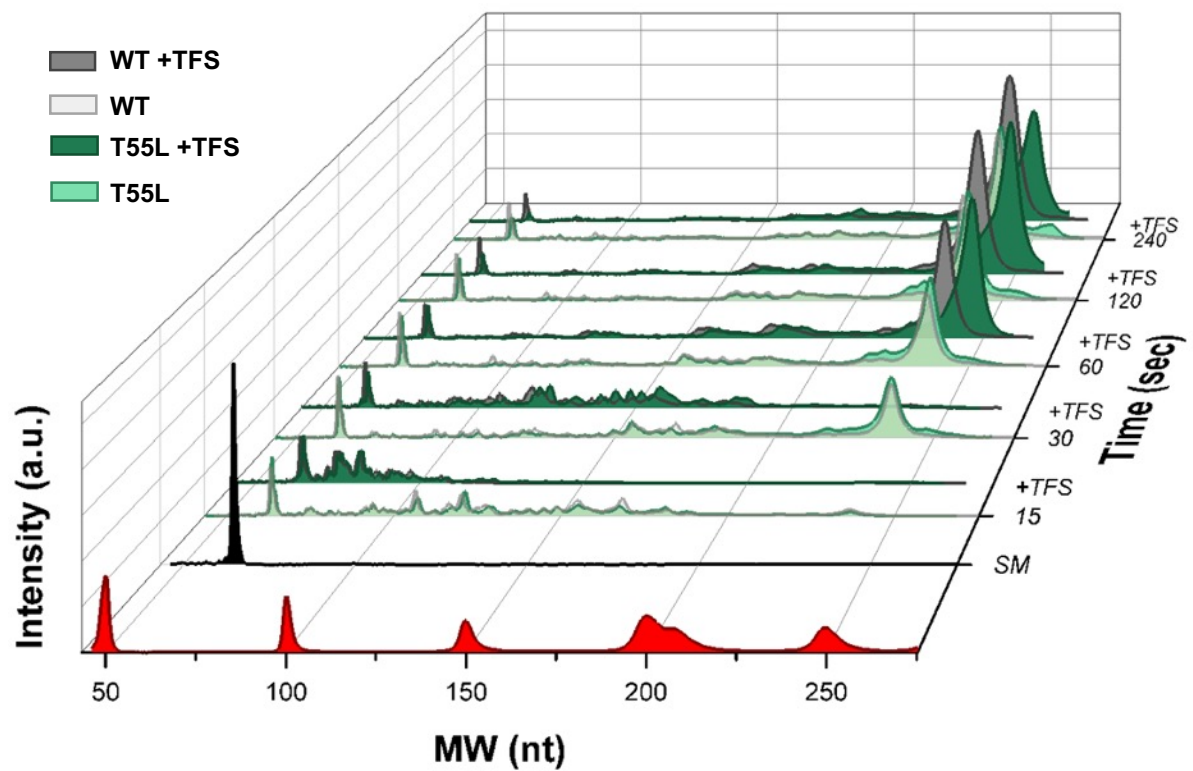
**d**



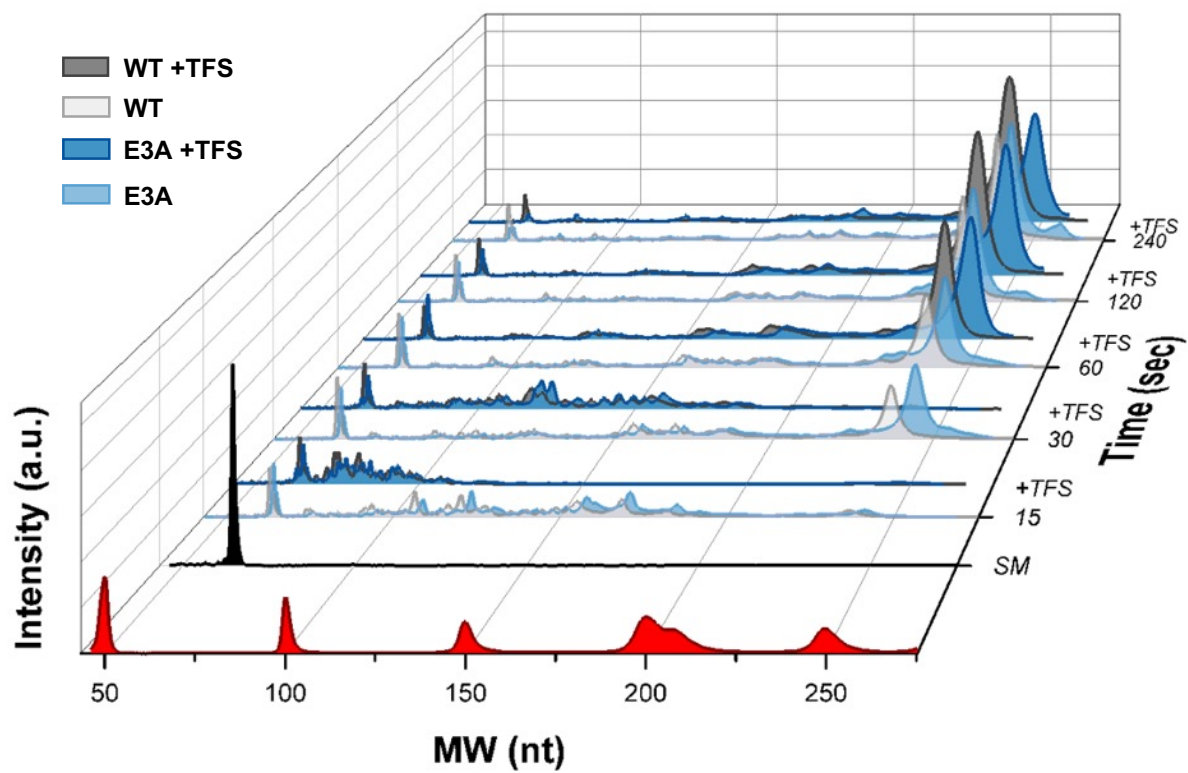
e



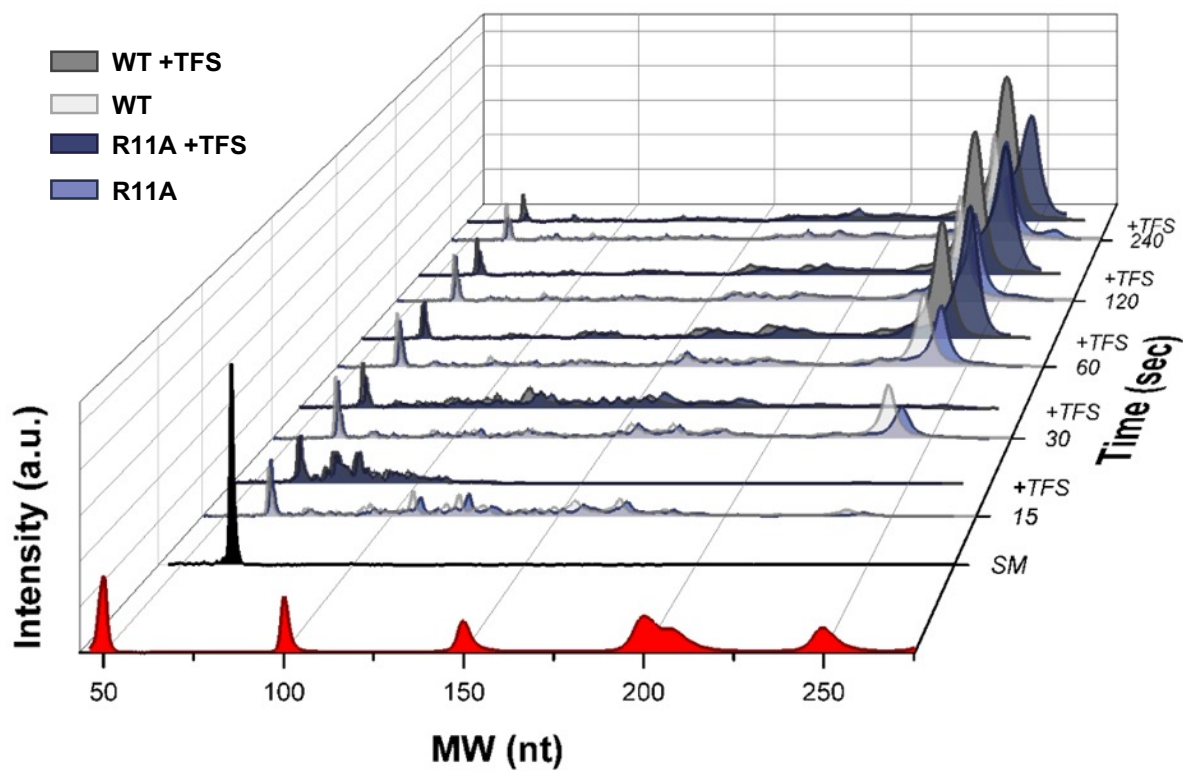
f



g

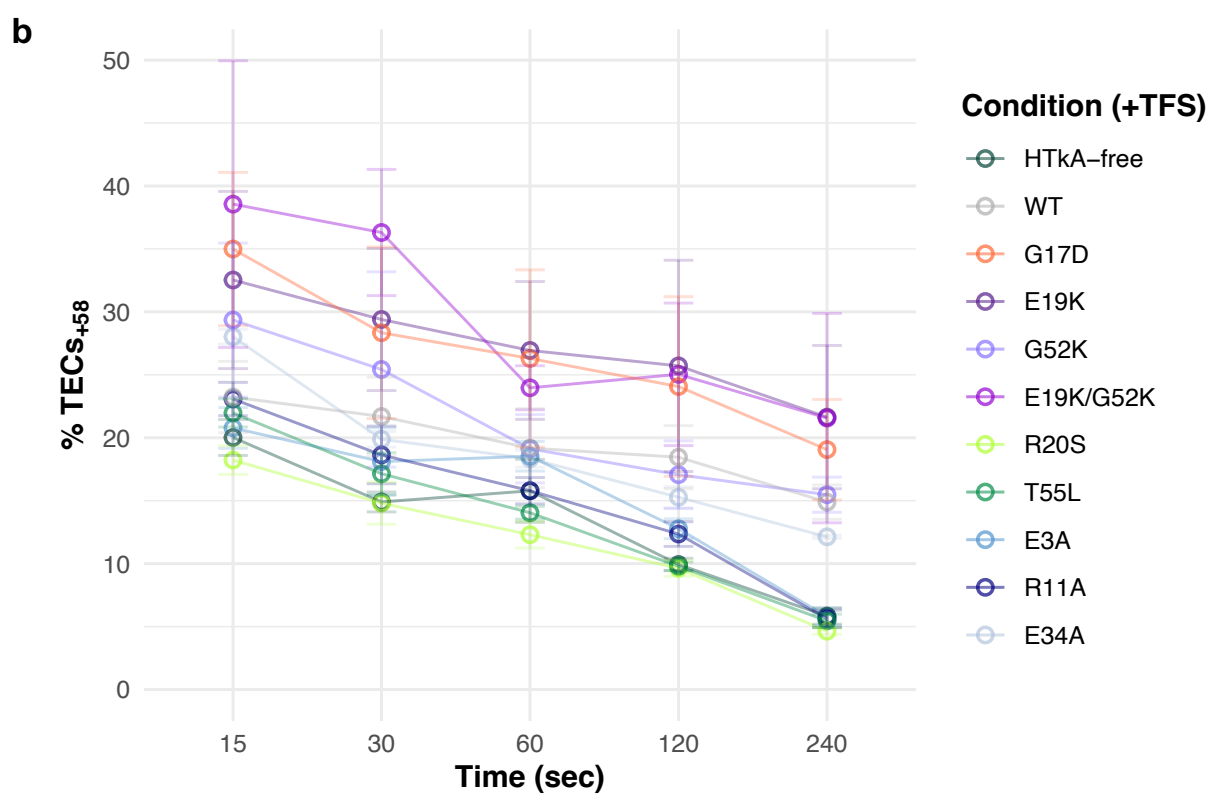
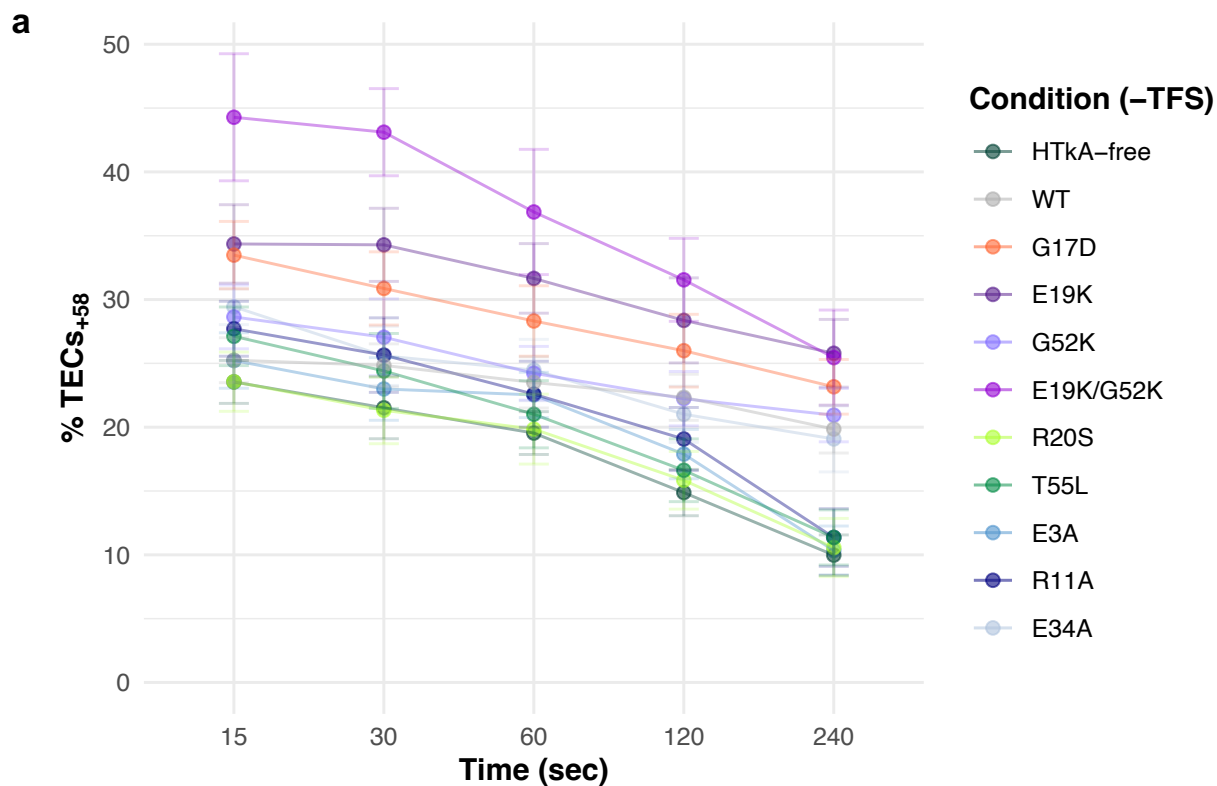


h

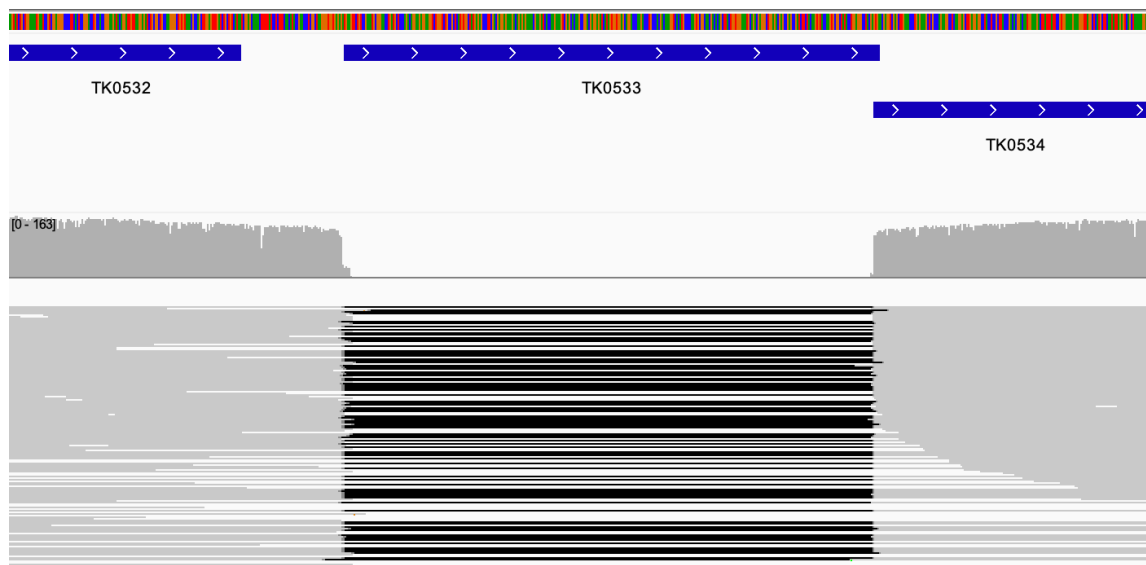
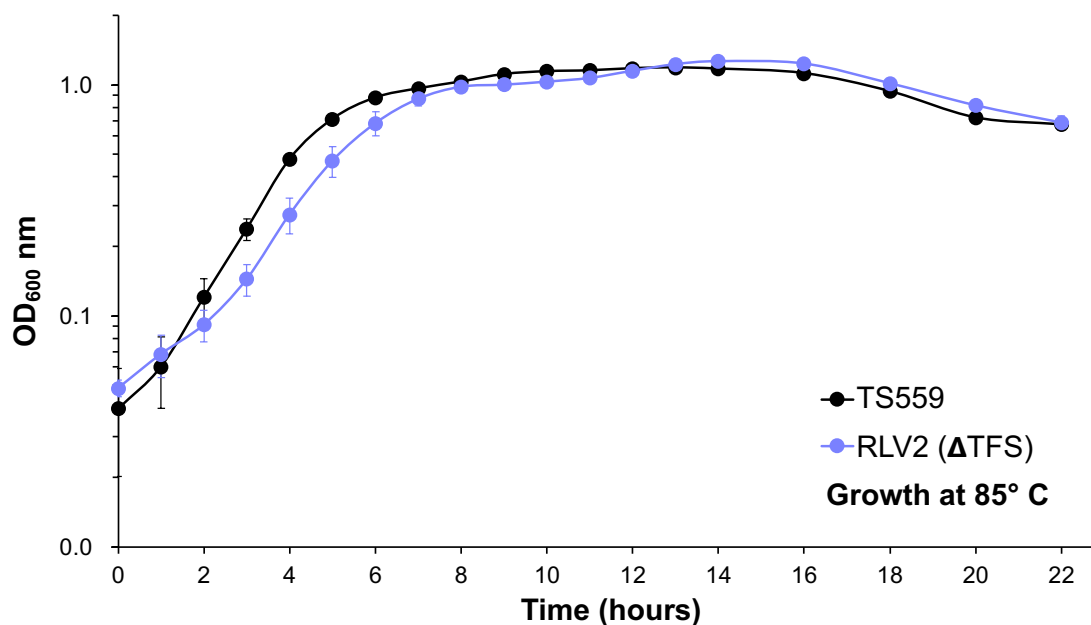


**Figure A.9. 3D representation of TEC progression in a chromatin environment with and without TFS.** (a – h) Waterfall plots permit quantification of the distribution of nascent transcript lengths over time in the presence or absence of TFS. The relative intensity of different transcript lengths was normalized to the sum of the counts in the starting material (SM) within each lane. Transcript abundance is compared for histone-free, HTkA<sup>WT</sup>, and HTkA<sup>variant</sup> landscapes in the presence and absence of TFS.



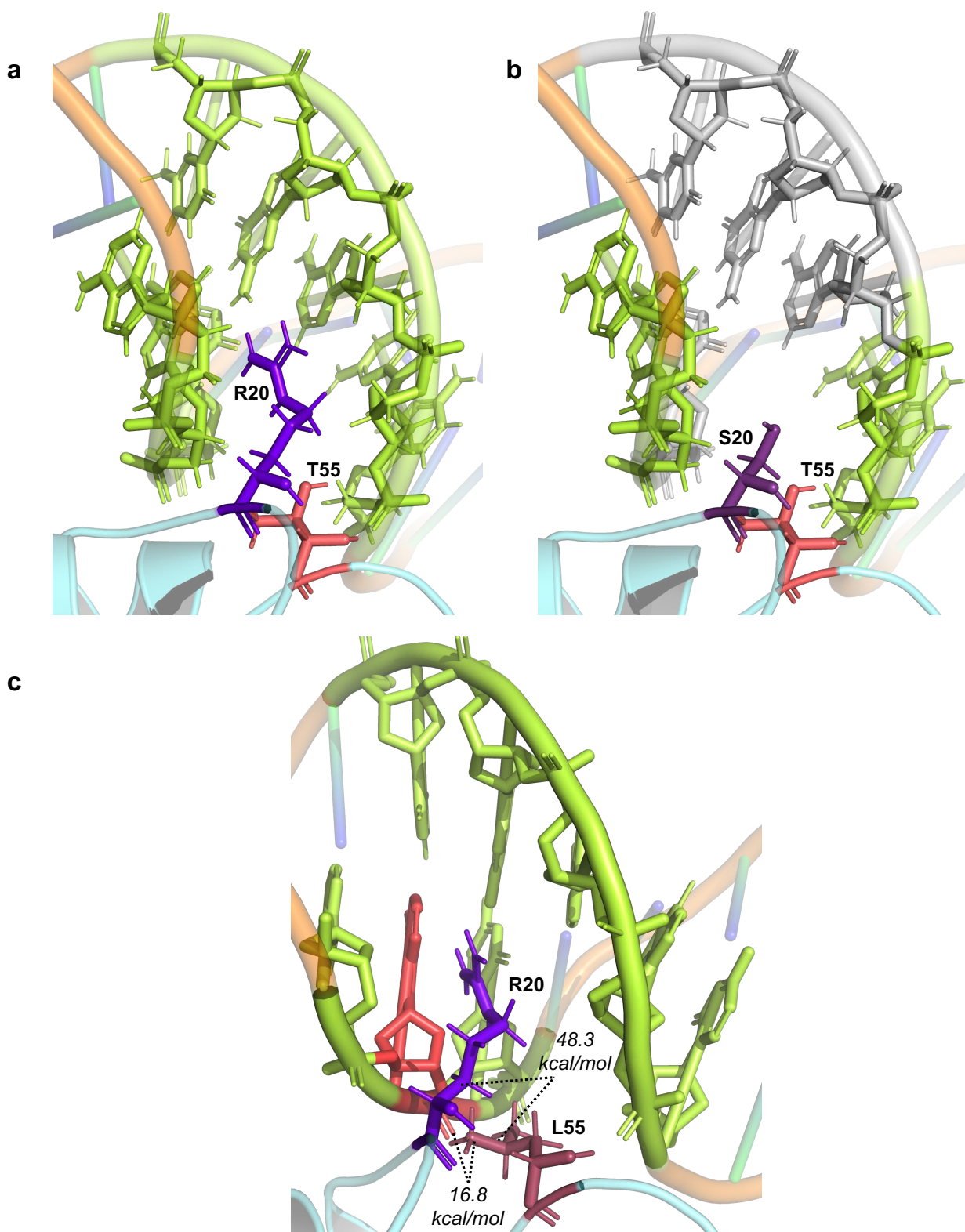


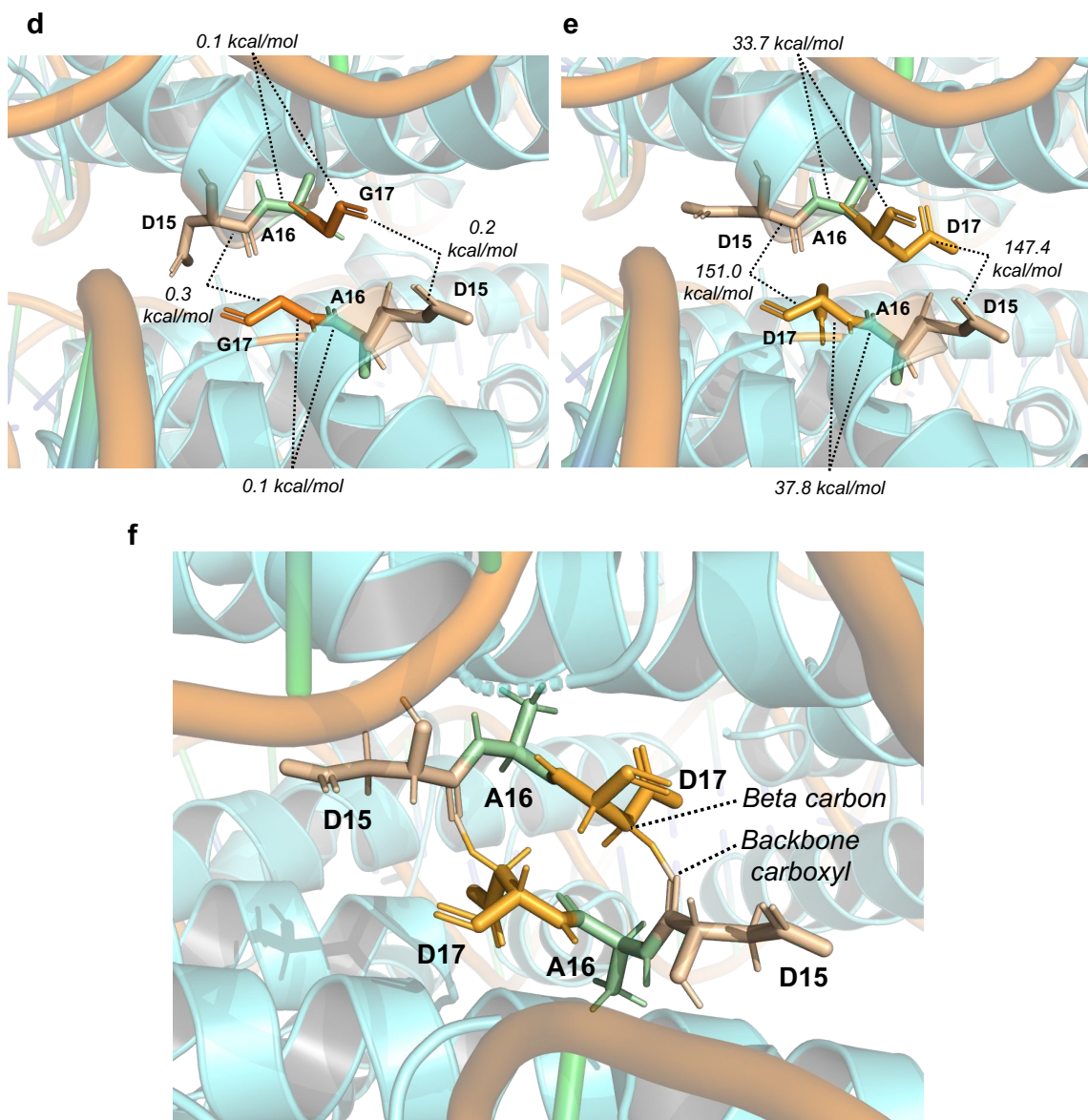
**Figure A.10. TFS modestly improves transcription restart of TECs<sub>+58</sub>.** **a.** The percentage of TECs at position +58 nt decreases as TECs escape into active elongation. **b.** Addition of TFS increases the rate of escape from +58 implying that some TECs were backtracked and could be stimulated into active elongation through TFS rescue. Average with standard error of n = 3 experiments is reported.

**a****b**

**Figure A.11. TFS is not essential, and deletion has nominal impacts for *T. kodakarensis* in optimal conditions.** **a.** TK0533, encoding for TFS, was successfully deleted from the *T. kodakarensis* TS559 genome. The exact endpoints of the deletion were confirmed via >100x whole genome sequencing (WGS). **b.** *T. kodakarensis* strains TS559 (parental) and RLV2 ( $\Delta$ TK0533) grow nearly identically at 85°C. Error bars represent the SD from biological triplicate cultures.







**Figure A.12. Individual residue substitutions elicit large impacts on the stability of archaeal histone-based chromatin landscapes.** **a.** Chromatin formed with HTkA<sup>WT</sup> permits interactions between T55 (pink) and R20 (purple) that position R20 for electrostatic interactions with surrounding nucleotides (green) of the wrapped DNA. **b.** Substitution R20S results in lost interactions with four nucleotides (grey) that normally stabilize the archaeal histone-based chromatin landscape. **c.** Substitution T55L crowds the histone-dimer interface, driving rearrangements that increase the Lennard-Jones repulsive term (kcal/mol) between T55L (maroon), R20 (purple) and an adjacent nucleotide (red). **d.** The absence of a side chain on residue 17 (G17) eliminates conflicts with neighboring residues and permits tight association of the gyres of archaeal histone-based chromatin. **e.** Substitution G17D drives rearrangements of A16 and D15 (from an adjacent monomer at the L1-L1 interface) that increases the Lennard-Jones repulsive energy, which impedes tight gyre association and impacts the 3D structures of archaeal histone-based chromatin. **f.** G17D creates new hydrogen bonds with a native aspartic acid at position 15 on an adjacent gyre at the L1-L1 interface.

## Supplementary Data

*Python code of Jupyter Notebook detailing the code used to create and analyze the attractive and repulsive forces of the wild type and mutant histone-based chromatin structures.*

```
#!/usr/bin/env python
# coding: utf-8
## Archaeal histone-based chromatin structures regulate transcription
elongation rates.
#### Wenck, B. et al. (2023)

# In[1]: #!pip install pyrosettacolabsetup #only needs to be run once
import pyrosettacolabsetup; pyrosettacolabsetup.install_pyrosetta()
import pyrosetta; pyrosetta.init()

# In[2]: from pyrosetta import * init()

# ### After importing PyRosetta, it is important to clean up the pdb
structure.

# In[3]: from pyrosetta.toolbox import cleanATOM
cleanATOM("pdb/5T5K.pdb")

# In[4]: cpose = pose_from_pdb("pdb/5T5K.clean.pdb")

# In[5]: print(cpose.pdb_info())

# In[6]: from pyrosetta.rosetta.protocols.loops import get_fa_scorefxn
sfxn = get_fa_scorefxn()

# In[7]: sfxn.show(cpose)

# In[8]: temp_pose = cpose.clone()
sfxn.show(temp_pose)

# ### We need to relax the structure to optimize the conformations for point
mutations.

# In[9]: movemap = MoveMap()
movemap.set_bb(False)
movemap.set_chi(True)
relax = pyrosetta.rosetta.protocols.relax.FastRelax()
relax.constrain_relax_to_start_coords(True)
relax.coord_constrain_sidechains(True)
relax.ramp_down_constraints(False)
relax.set_scorefxn(sfxn)
relax.set_movemap(movemap)
relax.apply(temp_pose)
temp_pose.dump_pdb('pdb/5T5K.relax.pdb')

# In[10]: rpose = pose_from_pdb('pdb/5T5K.relax.pdb')
sfxn.show(rpose)
```

```
# ### Observation of the scorefunction before and after relaxing the
structure shows a significant reduction in the overall free energy, important
for finding optimal conformations for point mutations.
```

```
# In[11]: from pyrosetta.toolbox import mutate_residue
```

```
# In[12]: R20S = pyrosetta.pose_from_pdb("pdb/5T5K.relax.pdb")
mutate_residue(R20S, 19, "S")
sfxn.show(R20S)
```

```
# In[13]: T55L = pyrosetta.pose_from_pdb("pdb/5T5K.relax.pdb")
mutate_residue(T55L, 54, "L")
sfxn.show(T55L)
T55L.dump_pdb("pdb/5T5K.T55L.pdb")
```

```
# ### Because the free energy associated with the T55L variant is
significantly greater than WT (rpose), we need to relax the structure to
eliminate any incompatible conformations.
```

```
# In[14]: temp_T55L = pyrosetta.pose_from_pdb("pdb/5T5K.T55L.pdb")
movemap = MoveMap()
movemap.set_bb(False)
movemap.set_chi(True)
relax = pyrosetta.rosetta.protocols.relax.FastRelax()
relax.constrain_relax_to_start_coords(True)
relax.coord_constrain_sidechains(True)
relax.ramp_down_constraints(False)
relax.set_scorefxn(sfxn)
relax.set_movemap(movemap)
relax.apply(temp_T55L)
temp_T55L.dump_pdb("pdb/5T5K.relax_T55L.pdb")
```

```
# In[15]: rT55L = pyrosetta.pose_from_pdb("pdb/5T5K.relax_T55L.pdb")
sfxn.show(rT55L)
```

```
# In[16]: G17D = pose_from_pdb('pdb/5T5K.relax.pdb')
mutate_residue(G17D, 16, "D")
sfxn.show(G17D)
```

```
# In[17]: G17D.dump_pdb("pdb/5T5K.G17D.pdb")
```

```
# In[18]: temp_G17D = pyrosetta.pose_from_pdb("pdb/5T5K.G17D.pdb")
movemap = MoveMap()
movemap.set_bb(False)
movemap.set_chi(True)
relax = pyrosetta.rosetta.protocols.relax.FastRelax()
relax.constrain_relax_to_start_coords(True)
relax.coord_constrain_sidechains(True)
relax.ramp_down_constraints(False)
relax.set_scorefxn(sfxn)
relax.set_movemap(movemap)
relax.apply(temp_G17D)
temp_G17D.dump_pdb("pdb/5T5K.relax_G17D.pdb")
```

```
# In[19]: rG17D = pyrosetta.pose_from_pdb("pdb/5T5K.relax_G17D.pdb")
sfxn.show(rG17D)
```

```

# In[20]: scrtypel = pyrosetta.rosetta.core.scoring.ScoreType(1) #fa_atr =
<ScoreType.fa_atr: 1>
scrtypel = pyrosetta.rosetta.core.scoring.ScoreType(2) #fa_rep =
<ScoreType.fa_rep: 2>

# In[21]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(19,
rpose, sfxn, scrtypel, 0)

# In[22]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(19,
R20S, sfxn, scrtypel, 0)

# In[24]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(54,
rpose, sfxn, scrtypel, 0)

# In[25]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(54,
rT55L, sfxn, scrtypel, 0)

# In[26]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(16,
rpose, sfxn, scrtypel, 0)

# In[27]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(16,
rG17D, sfxn, scrtypel, 0)

# ### To look at the energy associated with the L1-L1 interface we stacked
the 5T5K.pdb file using the symmetry mates function in PyMOL.

# In[28]: cleanATOM('pdb/stacked_G17.pdb')

# In[29]: G17_stacked = pose_from_pdb('pdb/stacked_G17.clean.pdb')

# In[30]: movemap = MoveMap()
movemap.set_bb(False)
movemap.set_chi(True)
relax = pyrosetta.rosetta.protocols.relax.FastRelax()
relax.constrain_relax_to_start_coords(True)
relax.coord_constrain_sidechains(True)
relax.ramp_down_constraints(False)
relax.set_scorefxn(sfxn)
relax.set_movemap(movemap)
relax.apply(G17_stacked)
G17_stacked.dump_pdb('pdb/G17_stacked.relax.pdb')

# In[31]: rstacked_G17 = pose_from_pdb('pdb/G17_stacked.relax.pdb')
sfxn.show(rstacked_G17)

# In[32]: print(rstacked_G17.pdb_info())

# In[33]: temp_G17 = rstacked_G17.clone()

# In[34]: mutate_residue(temp_G17, 16, "D")
sfxn.show(temp_G17)

# In[35]: mutate_residue(temp_G17, 218, "D")
sfxn.show(temp_G17)

# In[36]: movemap = MoveMap()
movemap.set_bb(False)

```

```

movemap.set_chi(True)
relax = pyrosetta.rosetta.protocols.relax.FastRelax()
relax.constrain_relax_to_start_coords(True)
relax.coord_constrain_sidechains(True)
relax.ramp_down_constraints(False)
relax.set_scorefxn(sfxn)
relax.set_movemap(movemap)
relax.apply(temp_G17)
temp_G17.dump_pdb('pdb/G17D_stacked.relax.pdb')

# In[37]: rstacked_G17D = pose_from_pdb('pdb/G17D_stacked.relax.pdb')
sfxn.show(rstacked_G17D)

# In[38]: rstacked_G17 = pose_from_pdb('pdb/G17_stacked.relax.pdb')

# In[39]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(16,
rstacked_G17D, sfxn, scrtype2, 0)

# In[40]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(16,
rstacked_G17, sfxn, scrtype2, 0)

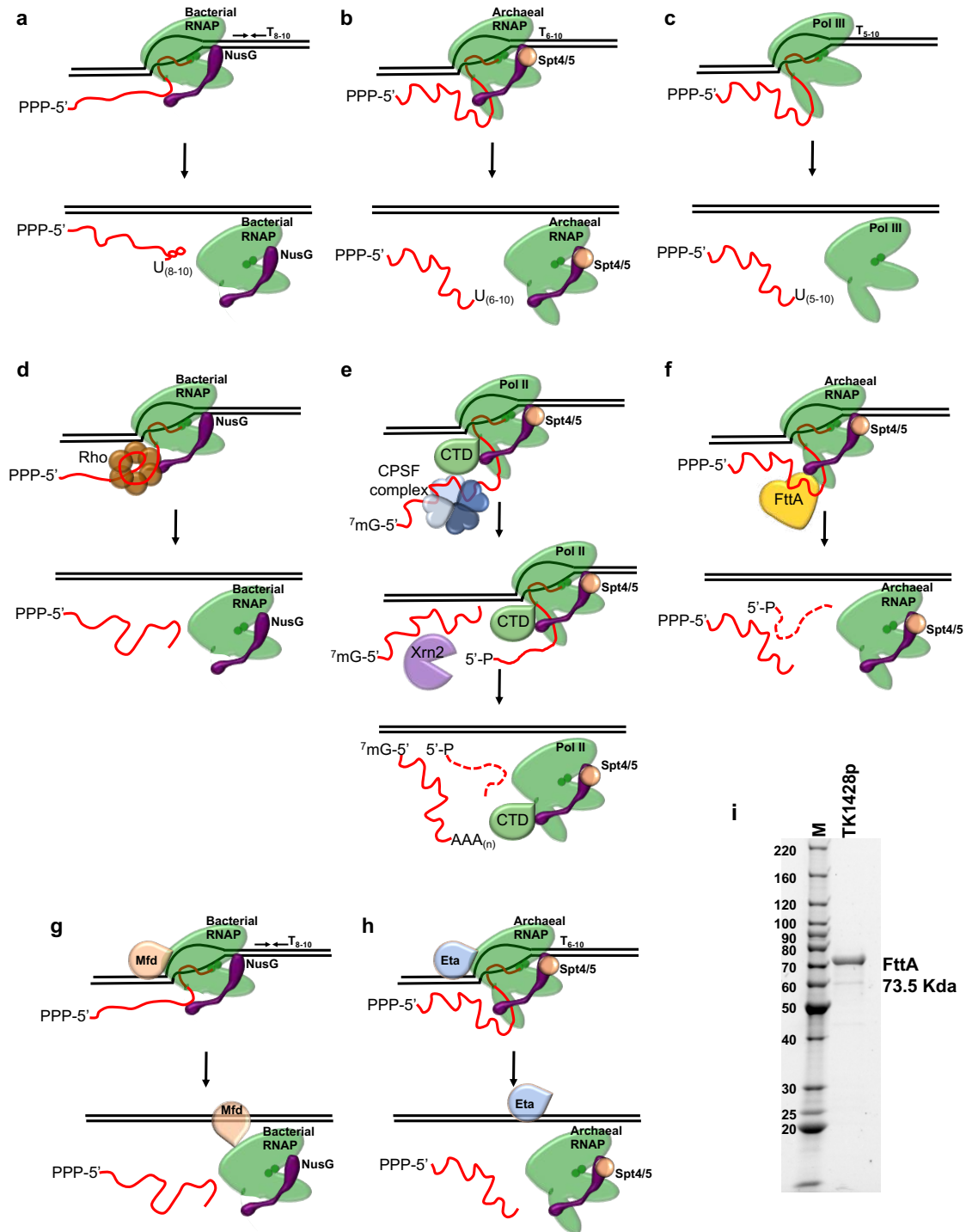
# In[41]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(218,
rstacked_G17D, sfxn, scrtype2, 0)

# In[42]: pyrosetta.toolbox.atom_pair_energy.print_residue_pair_energies(218,
rstacked_G17, sfxn, scrtype2, 0)

```

# APPENDIX B: CHAPTER 3 SUPPLEMENTARIES

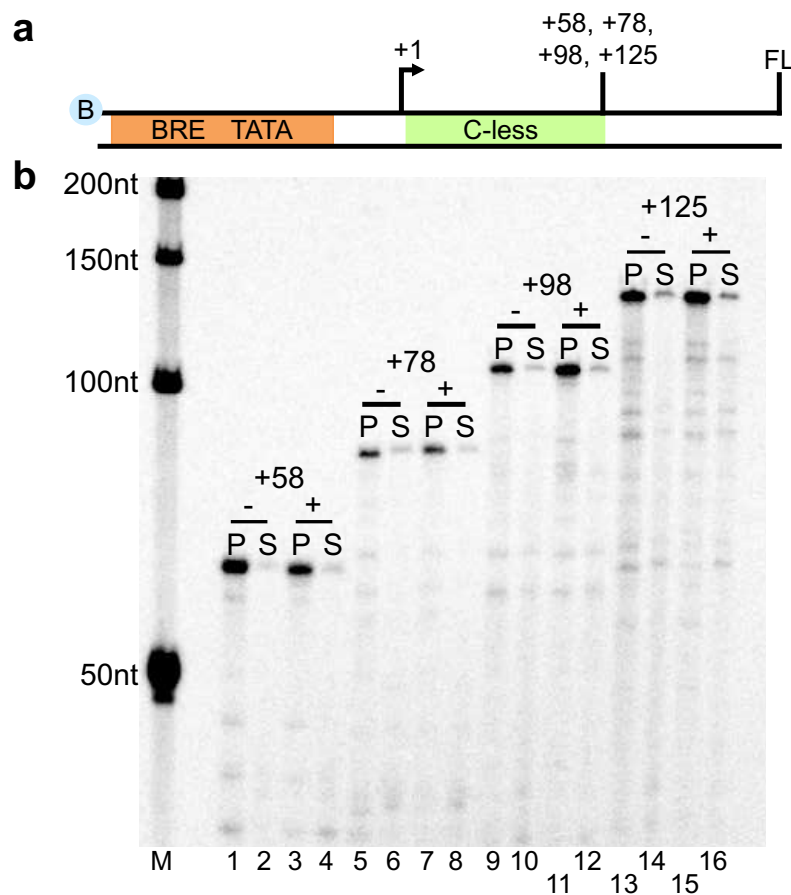
## Supplementary Figures



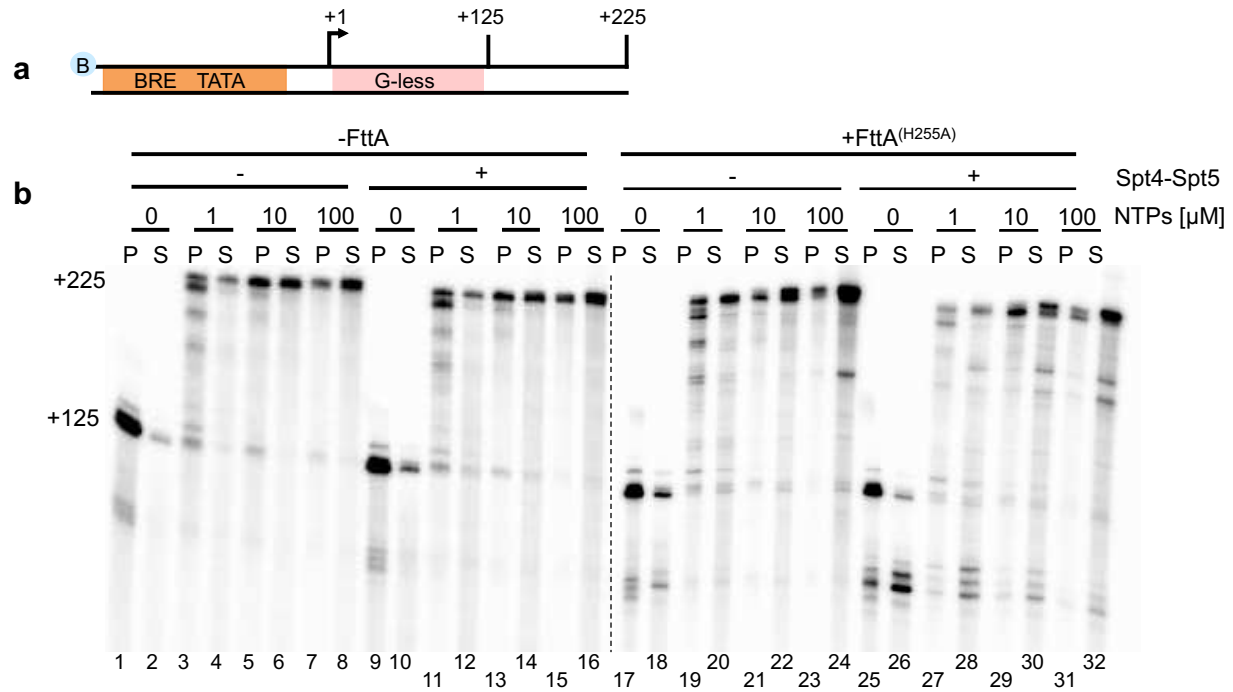
**Figure B.1. Summary of transcription termination mechanisms commonly employed in Bacteria, Eukarya and Archaea.** **a. – c.** Intrinsic transcription termination in Bacteria (a),

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 realize 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. **i.** FttA, an arCOG00543 member, directs transcription termination in Archaea. Recombinant FttA (the protein product of *T. kodakarensis* TK1428; TK1428p) is 73.5 Kda, 85°C thermotolerant and free of contaminating proteins. Lane M contains molecular weight standards in kDa.



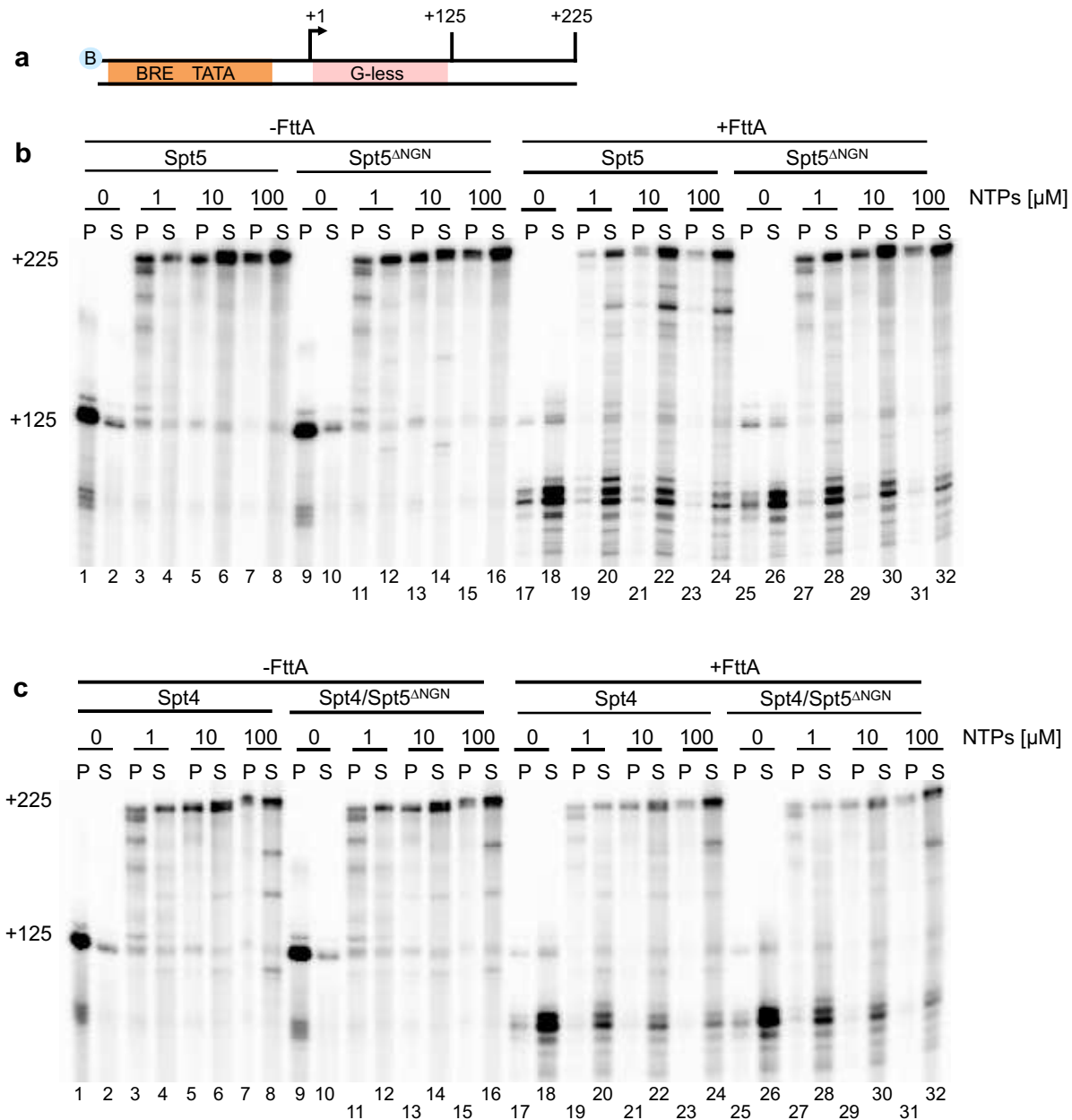


**Figure B.2. FttA-mediated termination prefers C- and U-rich transcripts.** **a.** Promoter-directed transcription of biotinylated templates encoding a C-less cassette permits formation of TECs with increasing length A-, G-, and U-rich nascent transcripts, respectively. FL = full-length; all templates permit elongation for 100 nts beyond the C-less cassette. **b.** TECs remain stably associated and transcripts are primarily recovered in the pellet (P) fraction in the absence (-) of FttA. When FttA is present (+), but nascent transcripts are devoid of CMP, minimal FttA-mediated transcript cleavage or termination occurs, and transcripts are not released to the supernatant (S). Lane M contains  $^{32}\text{P}$ -labeled ssDNA markers. The results are quantified in Fig. 2, panel d.

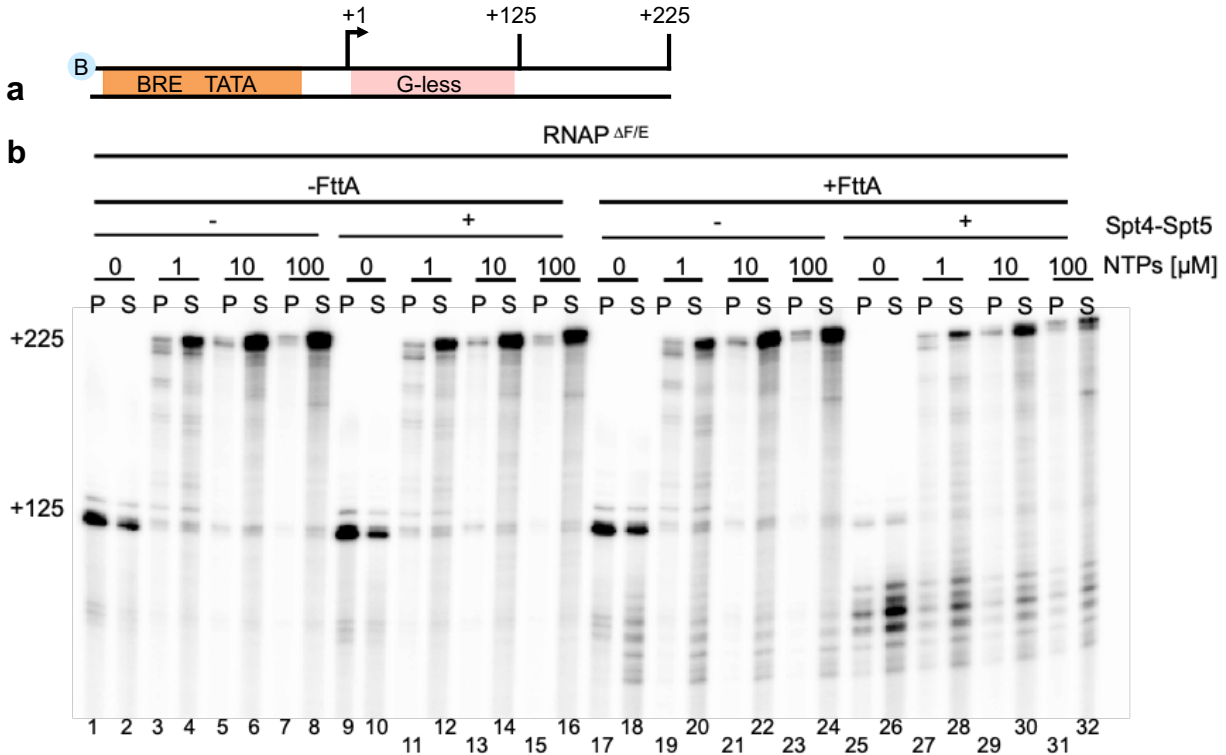


**Figure B.3. The RNA cleavage activity of FttA is stimulated by interactions with the archaeal TEC.** Figs. 3.1-3.4 demonstrate rapid and near complete cleavage of nascent transcripts within minutes of FttA addition. In contrast, **a**. FttA demonstrates minimal activity on the same +125 nt transcript in isolation. Control reactions with RNaseA demonstrate that the purified transcript is not resistant to the activity of RNases. **b**. Addition of *T. kodakarensis* RNAP to reactions containing purified +125 nt transcripts does not stimulate FttA activity over 30 min.

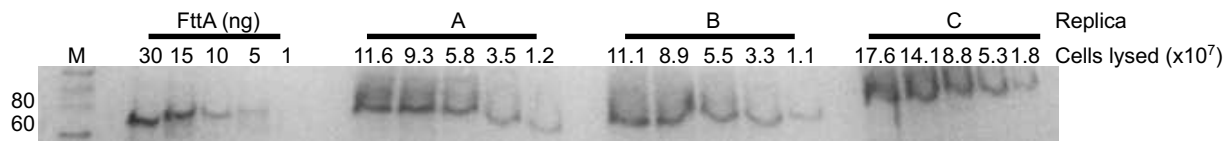




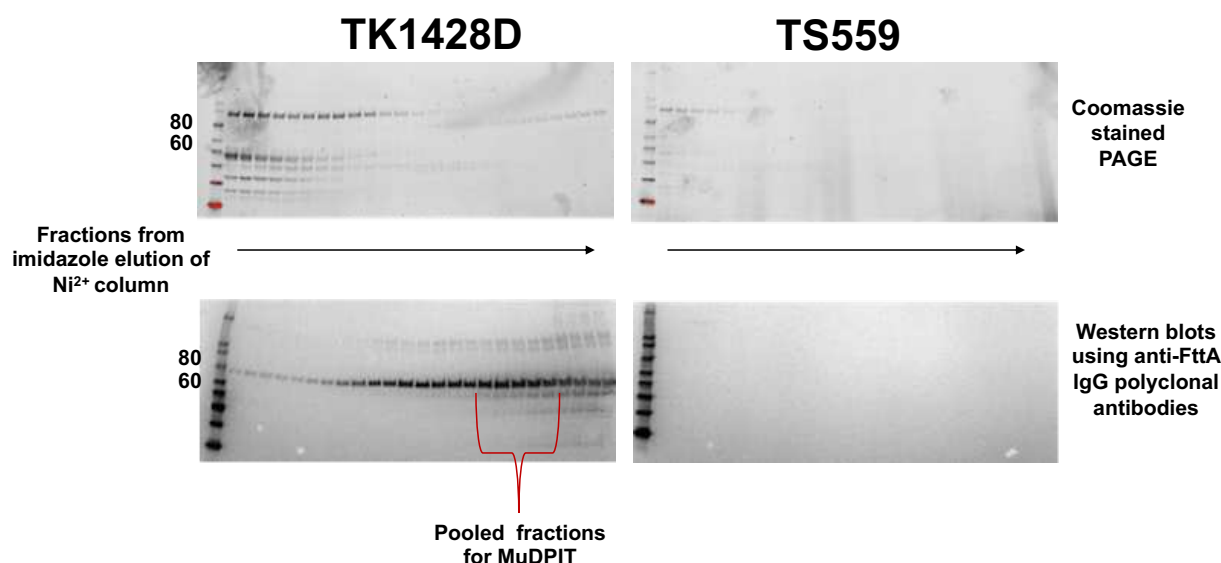
**Figure B.5. FttA-mediated termination requires an intact Spt4/5 complex for maximal activity.** **a.** TECs<sub>+125</sub> were assembled using promoter directed, biotinylated DNA templates. Intact TECs are bound to pellet fractions (P) while released transcripts are recovered from the supernatant (S). **b.** While intact and full-length Spt4/5 complexes stimulate FttA-mediated termination (Fig. 3.3), addition of Spt5 alone, containing (lanes 17-24) or lacking (lanes 25-32) the N-terminal NGN domain, fails to stimulate FttA-mediated termination to be competitive with transcription elongation at high [NTPs]. **c.** Spt4 alone, or together with the KOW domain of Spt5, is insufficient to stimulate FttA-mediated termination to be competitive with transcription elongation at high [NTPs].



**Figure B.6. The stalk domain of the archaeal RNAP is necessary for efficient and kinetically competitive FttA-mediated termination *in vitro*.** **a.** TECs<sub>+125</sub> were assembled using promoter directed, biotinylated DNA templates. Intact TECs are bound to pellet fractions (P) while released transcripts are recovered from the supernatant (S). **b.** While TECs assembled with RNAP<sup>WT</sup> support kinetically competitive FttA-mediated termination (Fig. 3.3), TECs generated with RNAP<sup>ΔF/ΔE</sup> only support FttA-mediated termination of stalled or slowly elongating complexes. The absence of the stalk domain impairs both FttA-mediated cleavage and release of the nascent transcript, and while FttA-activity can be stimulated by the addition of Spt4/5, the hinderance to FttA-mediated termination in the absence of the stalk domain impairs FttA-mediated termination under condition of high [NTP].



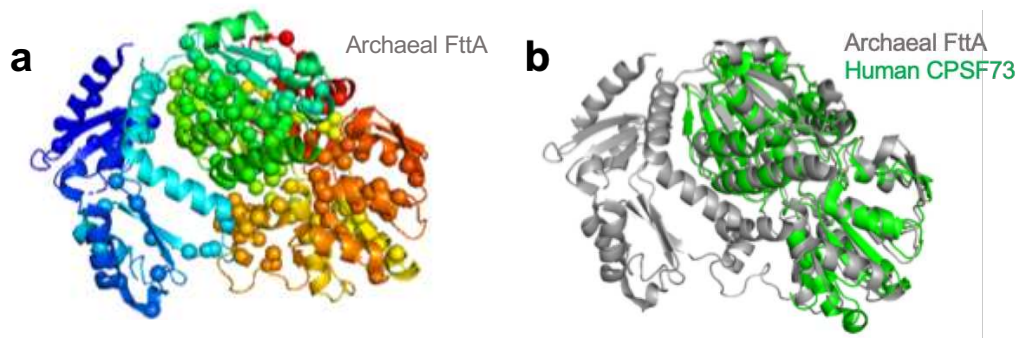
**Figure B.7. FttA is an abundant protein likely responsible for 3'-end formation in archaeal cells.** Quantitative Western blots employing anti-FttA antibodies, purified recombinant FttA, and total cellular lysates derived from known numbers of lysed *T. kodakarensis* cells reveal that FttA is present at ~2,100 +/- 500 copies per cell. Cell counts and protein calculations were performed as described <sup>4</sup>.



**Figure B.8. Gentle-purification of FttA directly from lysates of *T. kodakarensis* strain TK1428D.** Top panels show SDS-PAGE gels of fractions recovered from imidazole elutions of total cell lysates from strains TK1428D (left) and TS559 (right) resolved over 5 ml Ni<sup>2+</sup>-charged chelating columns (GE Healthcare). Bottom panels are Western blots of the same fractions from above probed with anti-HA antibodies to identify fractions within TK1428D lysates that contain FttA. The fractions pooled and analyzed by MuDPIT are identified. Magic Mark protein ladders are identified by molecular weight in Kda to the left of the gels.

Gene	Annotation	Mascot Score	Unique Peptides
TK1428	FttA; cleavage and polyadenylation specificity factor homologue	6388	47
TK2215	tRNA splicing endonuclease	633	12
TK0011	Uncharacterized protein	585	7
TK1557	Predicted dehydrogenase	537	18
TK1165	Predicted AP endonuclease	468	18
TK2250	Serine/Threonine protein kinase	421	17
TK0976	Putative snRNP Sm-like protein	351	8
TK1509	Probable tRNA pseudouridine synthase	202	13
TK0528	Serine hydroxyethyltransferase	168	13
TK0211	Amidophosphoribosyltransferase	147	9
TK1305	Probable translation initiation factor IF-2	133	11

**Figure B.9. Proteins identified as co-eluting partners of FttA from lysates of strain TK1428D.**



**Figure B.10. FttA is highly conserved and shares structural and sequence homology with eukaryotic CPSF73.** FttA is homologous to each subunit of the eukaryotic CPSF complex, with the greatest identity to CPSF73. The C-terminal MBL-fold is conserved in presumptive RNA processing proteins present in most Bacteria. No homology is noted between archaeal FttA and bacterial Rho proteins, but an identifiable homology ( $E_{\text{value}} = 0.046$ ) between *E. coli* NusA and archaeal FttA is intriguing and is likely based solely on the N-terminal KH domains of FttA and the KH domain of NusA. **a.** The crystal structure of FttA from *Pyrococcus horikoshii* (PDB: 3AF5) <sup>13</sup> shown in chainbow coloring (N-terminus in purple to C-terminus in red) reveals two N-terminal KH-domains attached by a linker to a C-terminal MBL fold. Alpha-carbons of highly conserved residues in archaeal FttA homologues are shown in colored spheres. **b.** The MBL-fold of FttA is nearly structural identical to the MBL-fold of the human CPSF73 protein (PDB: 2I7T) <sup>14</sup>.

WP_011250379.1_Thermococcus_kodakarensis	----MIRRETFVDDILKEIREIVQMVPREAGITDVEFGPELVIVYVKNPEAMMDGELI	56
WP_048048997.1_Methanosarcina_mazei	-----MPIEDVLDLKHIEKLNLPAGVTITDVEFGPQLVLYTEEPKRFADDGNLI	51
WP_004044063.1_Haloferax_volcanii	-----MSSVDKQLENLKAETINELPRDISVSDVKYEGPELVIVYTRDKPKFAGNGLI	52
OIO62505.1_Candidatus_Woesearchaeota_archaeon	-----MVLMMNIIKEILSQLP-EEEKISDACFEGANIVLYTKDKKEFFLDNNGLI	49
OLC36512.1_Thaumarchaeota_archaeon	MQRKTOQR-----ELSPAPNIMGTILQSIPKEADVTKIDYEGPRIAVYTKNPRYLEHNEVI	57
OLS12352.1_Candidatus_Lokiarchaeota_archaeon_CR_4	-----MIFVSSDESINAMEIRNTLPPTTDSIQIEYEGPEIAVYTKNPKILEDNGLI	54
TAL51517.1_Nanoarchaeota_archaeon	-----MSKIIKEILKQIP-EDKISDAQFEGANIVLYTKDKKEFFVNNGLI	44
WP_011007526.1_Pyrobaculum_aerophilum	-----MSFVE-IENKVKAILSGIEIVKVNYPEGNLCIYVRRPTED--VIDMI	44
RLI34429.1_Candidatus_Bathyarchaeota_archaeon	MPSPLRGQTRQPETIQSKIQVILQHVPKAEIIRVFEFGPRLAIYTRKPEILLSSQSYIV	60
WP_014513953.1_Sulfolobus_islandicus	-----MNRLASLRNINTISLIYS-ELKDLGITRIEYEGPTIAVYVKKPTMVTEGEVI	52
WP_011277529.1_Sulfolobus_acidocaldarius	-----MSIRLNTVNAIYNGMPKEAGISKIDFEGPEIAVYVRRNPAVV--DGETV	46
RLG13324.1_Candidatus_Nanoarchaeota_archaeon	-----MELEDVKSRLPKTAMITKTLFEGSDIVFYTKNKDFVNGSSDI	43
OLC64863.1_Candidatus_Rokubacteria_bacterium	-----MTIEDILNDARSVVKVVPDHVEITQVDFEGPTIYIYTKNMEVFAESNDIV	51
OEU57241.1_Desulfuromonadales_bacterium_C00003096	-----MPAEATLLALKRRITELLPDVSIVTGVFEFGPELVLYTDTQRVFDVGGALV	51
OGM94118.1_Candidatus_Wolfbacteria_bacterium	-----MNCKLIETILEKVRNALP-QEASIRVLEGSSELIYVTKDRDFFVKHEETV	50
OGZ62538.1_Candidatus_Staskawiczbacteria_bacterium	-----MELLKN-ITDRL-KGKI TEASFEGANVLYTDDAEFFRRSGSKI	42
RKX64688.1_Tenericutes_bacterium	-----MGIEETLSELRRIRKKNLPVGVITISDVEFGPELVITYKEPRKLADNGEV	51
NP_057291.1_Homo_sapiens	-----	0
NP_650738.1_Drosophila_melanogaster	-----	0
NP_502553.2_Caenorhabditis_elegans	-----	0
NP_013379.1_Saccharomyces_cerevisiae	-----	0
WP_011250379.1_Thermococcus_kodakarensis	KNLAKVLKKRISVRPDPDILLPPEKAEELIKQLVPPEAEITNISFDPVSGEVLIEARKPG	116
WP_048048997.1_Methanosarcina_mazei	RNLAKELTRIAMRDPRLVATPEDSISIIIEVVPKESIVSSYFDPDSGEVIEAEKPG	111
WP_004044063.1_Haloferax_volcanii	RKLAKLRKRIIVRPDPDVLSDPREAEPKILSVIPEAGVTDLDPHIDTGEVVEAEKPG	112
OIO62505.1_Candidatus_Woesearchaeota_archaeon	RKIVNDIKKRVELRPDPATLIEDAEKLIKSVIPEAGVDKILFDPQSRVIEAEKPG	109
OLC36512.1_Thaumarchaeota_archaeon	SNMNVVKKRIIVRTESIRKSEESRQILQMLPKDAELQGTFFDTATGEMTIEIKRPW	117
OLS12352.1_Candidatus_Lokiarchaeota_archaeon_CR_4	KNLAKLRKRIIVRSDPSVRRDKEETANYIRTLVTDSEITKVTFDENMGGEVIEAEKPG	114
TAL51517.1_Nanoarchaeota_archaeon	RKIVDDIKKRVELRPDPKITSDPVKAEERIKIMPPEAGVSNVDFPARSQVIEAEKPG	104
WP_011007526.1_Pyrobaculum_aerophilum	GEVAKTLKKRVLLRVPDSNRASEKTAASKVIREVLSD--VEDVVFENN-GDVYIYLAAPL	100
RLI34429.1_Candidatus_Bathyarchaeota_archaeon	SDIVSLIRKRIITLRSDDPEIRMKPEEAERFIRSVIPEAEITNITFDENIGEVVIEAEKPG	120
WP_014513953.1_Sulfolobus_islandicus	KKIAKDIKKRII IKADPSVRDKKEAVEIKKIIVTPEAVIDIKFDDDLGEVLIAKAKPG	112
WP_011277529.1_Sulfolobus_acidocaldarius	KKIAKEIKKRVVVKADPSKRDKEETKEIKNIIVPQEAQIIDIKFDDDLGEVLIAKAKPG	106
RLG13324.1_Candidatus_Nanoarchaeota_archaeon	KALVSEIRKRIINLRPDQSTILAEKDAREKIRTIIVPKEAIIQDILFEPEFGKVIILAQKPG	103
OLC64863.1_Candidatus_Rokubacteria_bacterium	RQIAQQLRRRIIVRPDPSILLASQEDAERIREVIPPQAQITGVYFETETGEVTEALAPG	111
OEU57241.1_Desulfuromonadales_bacterium_C00003096	RTLAKELKKRISVRPSSDILMDPEKASKAHEIPEEGGKIDYFDMDKAEVIEAEKPG	111
OGM94118.1_Candidatus_Wolfbacteria_bacterium	RQVVDQLKKRIEVRPEAAICLSQEATKKRIMELVPAEANKNIYFEERSLVIIEAEKPG	110
OGZ62538.1_Candidatus_Staskawiczbacteria_bacterium	KEIVNEIKKRIELRADKILKDAEETEKEIRKSIPEAEITNIIIPDNYRSIVIEAEKPG	102
RKX64688.1_Tenericutes_bacterium	RSIAKDVKKRIIVIRPKSVLYDPADATAAEKIAPSDAGITDYCFDSDTGEVMEAEKPG	111
NP_057291.1_Homo_sapiens	-----	0
NP_650738.1_Drosophila_melanogaster	-----	0
NP_502553.2_Caenorhabditis_elegans	-----	0
NP_013379.1_Saccharomyces_cerevisiae	-----	0
WP_011250379.1_Thermococcus_kodakarensis	LVIGKNGETLRLITQKVHAPVRVTRTP-----IQSQTIYSIRSILQTESKDRRKFRLQV	171
WP_048048997.1_Methanosarcina_mazei	LVIGKHGATLREITQKIGWIPKVVTRTP-----IKSRTVKNIREFMNRNLKERKEILKV	166
WP_004044063.1_Haloferax_volcanii	MVIGRHGSLTREITQKVGWTEPVVTRTP-----IESSTVSNVRNFKQEREDRRILERT	167
OIO62505.1_Candidatus_Woesearchaeota_archaeon	LAIGQGDQDILREIRKTLWLVLIRRKPA-----IRSQLENIIRAVLYQNSDYRRKFLDKI	164
OLC36512.1_Thaumarchaeota_archaeon	LLNNAEIVNMMVLDLQKTVGRVIRKATT-----AQSQTIQTINYNLKKISSTERSKHLRQV	172
OLS12352.1_Candidatus_Lokiarchaeota_archaeon_CR_4	LVIGKLGANLRISMETLWRPIVVRTP-----IDSKTVKIVRHLKTERGQKEILLRI	169
TAL51517.1_Nanoarchaeota_archaeon	LAIGQGSILREIREKTFVWPFIRRTPA-----IRSVLENIIRAVLYQNSDYRRKFLHSLV	159
WP_011007526.1_Pyrobaculum_aerophilum	R-EKEIKSATREIFVKTGWRVAVIESGVPKDRVKLPHEIVGVRHIFHSAYAQRMELMSK	159
RLI34429.1_Candidatus_Bathyarchaeota_archaeon	LVISKNGATLQELIKVTKWRPRLTRTP-----IPSKIIAHRIHLYQEVKQRIILRSV	175
WP_014513953.1_Sulfolobus_islandicus	LVIGKGSGLQKRIEVRPEAEITFKAEIVREPP-----IRSRTYDSILEHLYNETEYRAKILKV	167
WP_011277529.1_Sulfolobus_acidocaldarius	LVIGKGGSIQKQIFAEYWRPIVREPP-----LKSRTYDGLVTHIYNETEYRAKILKV	161
RLG13324.1_Candidatus_Nanoarchaeota_archaeon	IVIGKTGETLREIKETFWNPEIQRMGP-----FRSKIVNKAREIVHEEAKYRQOFLNKI	158
OLC64863.1_Candidatus_Rokubacteria_bacterium	MVIGRHGSLVLEIKRIGWAPKVVTRTP-----IPSKTVEEIRQYLRTINDERQVFLKQV	166
OEU57241.1_Desulfuromonadales_bacterium_C00003096	GRQIHR--EQLSDDEWVRITLGGCREVGRSFIIVSTPETRILVDCGDKPGSD-----DV	166
OGM94118.1_Candidatus_Wolfbacteria_bacterium	LVIGRGGETFQIRAEFTFWPRIERVPP-----IKSDVIDGIRKVLHQEVKFRKEFLNKI	165
OGZ62538.1_Candidatus_Staskawiczbacteria_bacterium	MVIGKRGIIIDELKHKTFWSPQVQRSPA-----IKSKIVENIREVLVANNYRRRFLNSV	157
RKX64688.1_Tenericutes_bacterium	LVIGQHGSMRLREITRYIGWTPKVVTRTP-----IESSTIKNVRHLRESLDERQIILREL	166
NP_057291.1_Homo_sapiens	-----	0
NP_650738.1_Drosophila_melanogaster	-----MTQAT	5
NP_502553.2_Caenorhabditis_elegans	-----	0
NP_013379.1_Saccharomyces_cerevisiae	-----	0
WP_011250379.1_Thermococcus_kodakarensis	GRNIYR--KSEYKSRWIRITLGGFREVGRSALLVQTDSESYLVDFGVNIAALKDPTKAY	229
WP_048048997.1_Methanosarcina_mazei	GRKIHR--ECTSKDQWVRITLGGCKREVGRSCFLSTPESRILIDCGVNVGSD-----NMT	221
WP_004044063.1_Haloferax_volcanii	GRQIHR--EQLSDDEWVRITLGGCREVGRSFIIVSTPETRILVDCGDKPGSD-----DV	220
OIO62505.1_Candidatus_Woesearchaeota_archaeon	CHRIYDQWIREKKNEWIRVSVLGGARQVGRSCFLIQTPESTRILLDCGVNIAS-D-KDAY	221
OLC36512.1_Thaumarchaeota_archaeon	GEQIFR--PKLSEDAEVLITLGGFSQVGRSCMLLTTHESKILVDCGVNPGA-RSPMEAF	229
OLS12352.1_Candidatus_Lokiarchaeota_archaeon_CR_4	CKRIHR--PVIFKPTAIKLTALGAFREVGRSCILLVQSNSTVILVDCGLNVGS-----PTEPF	224
TAL51517.1_Nanoarchaeota_archaeon	GERIYDSVWRKREGWRLSVLGSAREVGRSCILLQTNESVILLDCGINIAA-NSPOLAY	218
WP_011007526.1_Pyrobaculum_aerophilum	ARYIHQ--EPVVKCEPTITVFLGAAMEVGRSAILVSTTESNVLLDCGLKPAQN-----DEEF	214
RLI34429.1_Candidatus_Bathyarchaeota_archaeon	GDRIFR--APIYKSOEVLVTLGGRFQVGRSAILVTRRESSVLIDCGINPGT-ANPDSF	232
WP_014513953.1_Sulfolobus_islandicus	GERIHR--ETVQDKYVRITLGGFLEVGRSAVLVETPESKVLDDVGLNPSANMFGEKIF	225
WP_011277529.1_Sulfolobus_acidocaldarius	GERIHR--DILFKDRFVRITLGGFNEVGRSAVLVETPESKILLDVGLNPSV-SYGEKLF	218
RLG13324.1_Candidatus_Nanoarchaeota_archaeon	GQKIQL--KKGSGKEGWRFPSALGSGFEGVGRSCIFLQTKESKVLDDCGVNTGS-N---EMN	212
OLC64863.1_Candidatus_Rokubacteria_bacterium	GRRLAR--EVPPGENYVRITLGGFRQVGRSALLSTRDSKVLIDCGVLISED-----NGS	220
OEU57241.1_Desulfuromonadales_bacterium_C00003096	GRRIYR--DKSEDEWLRLITLGGCREVGRNSFLLSTPETRILIDCGVSVGS-----EGT	219
OGM94118.1_Candidatus_Wolfbacteria_bacterium	GEKIFS--ERKTNRDWIRIPLGGFRQVGRSCILLIETPKSKVLIDCGIAPGG-QG-ADAF	221
OGZ62538.1_Candidatus_Staskawiczbacteria_bacterium	GKKIYEDWDPKRTDEWIRLITFLGAARQVGRSCVLLQTPNSKVLDDCGIDVSS-QG-EDKF	215
RKX64688.1_Tenericutes_bacterium	GEKIHR--SITSPDNWIRITLGGCREVGRSCFLSTPETRILVDCGINVGDH--GSGT	222
NP_057291.1_Homo_sapiens	---MSA---IPAESDQLLIRPLGAGQEVGRSCIIIEFKGRKIMLDCGIHPGLE---GMDAL	53
NP_650738.1_Drosophila_melanogaster	---GDAR---MPDEESDQLIKPLGAGQEVGRSCIMLEFKGKIMLDCGIHPGLE---GMDAL	59
NP_502553.2_Caenorhabditis_elegans	---ME---EGGDSNLSLCTPLGSGQEVGRSCHLEYKGRVMLDCGVHPGLH---GDVAL	52
NP_013379.1_Saccharomyces_cerevisiae	-----MERTNTTTFKFFSLGSGNEVGRSCHILQYKGTVMLDAGHAPAQ---GLASL	50



168

WP_011250379.1	Thermococcus kodakarensis	-GMIWEATAIHTAYPEYLSKHIREQIFHEGYNPFLLNPIFKSVANSR-----	ERQDII	507
WP_048048997.1	Methanosarcina mazei	-GMIWEATAIHATHPEYLNNDLRKLIPOKQGNPFLLSECFKPVDSHE-----	ARQKII	496
WP_004044063.1	Haloferax volcanii	-GMIWEATAIHTTYPEYLRDLDLRIFHEDENPFLLAEFNNHDGGE-----	EERQDV	497
OT062505.1	Candidatus Woesearchaeota archaeon	-GSVWDITAIHTAYPEFLNSIVRKQIFHKDQNPFLSEIFKRVGSQK-----	ERTQVI	495
OLC36512.1	Thaumarchaeota archaeon	-GMISEATAIHAYPEYLARDLRQKILETDENPFDSSEYFTNIEHSD-----	AREEPL	503
OLS12352.1	Candidatus Lokiarchaeota archaeon_CR_4	-GLISEATAIHTTHPEYLNKLDRDRIFHQGNPFLLSPFTQVDSNN-----	ARQDPIV	498
TAL51517.1	Nanoarchaeota archaeon	-GMVWDVTAIHTAYPEYLNAAATRLIFHKDQNPFLHEVFEVGSVK-----	ERMQVI	492
WP_011007526.1	Pyrobaculum aerophilum	-GMIVETLNVYLMPHYLNPEVAEEIY--GGVNFPTTSGSVVIVDRAKRVEDRINQVAKIA	493	
RLI34429.1	Candidatus Bathyarchaeota archaeon	-GMVNEVTAIHTAYPEYMSKELQNMILHEDVNPQSEYFVNKNAG-----	ARSEVL	506
WP_014513953.1	Sulfolobus islandicus	-GLVDEVTAIHNAYPEWLGREVRREILYKDNENPTSEHFKEIEGYK-----	EDI	494
WP_011277529.1	Sulfolobus acidocaldarius	-GMVDEITAIHTAYPEWLSRLEAILYRDENPFMSEFFKRIEGYK-----	EDI	487
RLG13324.1	Candidatus Nanoarchaeota archaeon	-GMIWNATAIHTTYPEFLSRRLQNIQIFHNNNPFSSVEFKRVASMN-----	ERKSVI	485
OLC64863.1	Candidatus Rokubacteria bacterium	-GMIWEATAIHTAYPEYLNQLRTIQITQGENPFLSPMFKRVTSD-----	MRSNIC	495
OEU57241.1	Desulfuromonadales bacterium C00003096	-GMIWEATSITHTAYPEYLNKLKNSIFQ--GENPFLSDFVQVDDSE-----	KRKSII	488
OGM94118.1	Candidatus Wolfecubacteria bacterium	-GMIWDANGIFTAYPEYLSRQMQKSIY--GGEDPFKNEIFKRIASQA-----	DREKAW	489
OGZ62538.1	Candidatus Staskawiczbacteria bacterium	-GMIWDINAIHTAYPDLNKLRLALTF--QDKNPFTEIFSRVGSSE-----	ERKNVI	488
RKX64688.1	Tenericutes bacterium	-GMIWEATAIHTTHPEYLNLDLRNIQIFHKGMNPFLLAEFCVVKDSQK-----	MREDLL	497
NP_057291.1	Homo sapiens	SSLAKKCMAYVQTYVNAMNDKIRKQIN--INNPFVFKHISNLKSM-----	DHF	317
NP_650738.1	Drosophila melanogaster	SSLAKKCMAYVQTYINAMNDRIRQIA--VNNPFVFRHISNLKGI-----	DHF	323
NP_502553.2	Caenorhabditis elegans	SSLAKKCMVYQTYTFVNMNSRIQKQIA--VKNPFIFKHVSTLRGM-----	DOF	318
NP_013379.1	Saccharomyces cerevisiae	SNLAKKCMVYQTYVNMMNDIRKFRDSQTNPFIFKNISYLRNL-----	EDF	327
		. . . . . : ** :		
WP_011250379.1	Thermococcus kodakarensis	DSGEPAIIATSGMLVGGPSVEYFKQLAPDPKNSIIFVSYQAECTLGRQVGRLEIPIV	567	
WP_048048997.1	Methanosarcina mazei	QNQPQCVILATSGMNGGVPVEYFKFAEDPRTNLVFGYQADGTIGRRIQKGWKEIPMT	556	
WP_004044063.1	Haloferax volcanii	ADGDPAIILSTSGMVTGPGIMSWLRHVGPDPKSRILVFGYQAQGTIGRRIQNGWDEIPMT	557	
OT062505.1	Candidatus Woesearchaeota archaeon	EETGSCTILATSGMLVGGPSVQYLKQLADNPNTLLPVCYQCGESLGRRIQRGEKMSVD	555	
OLC36512.1	Thaumarchaeota archaeon	-REGCPITMATSGMLGEGPVLEYFKNIAPIQKNKILPVSYQVNGTLGRRVLGSRQVSYM	562	
OLS12352.1	Candidatus Lokiarchaeota archaeon_CR_4	-EGGCPITLATSGMLQGGPSVQYLRLGSDDPNTLLPVSYQVNGTLGRRIQKGRFDSIQ	557	
TAL51517.1	Nanoarchaeota archaeon	EETGCPVILATSGMLTGGPSVEYLKHLGDNAKNSILVPSYQCGESLGRRIQGEREFNPV	552	
WP_011007526.1	Pyrobaculum aerophilum	QSEEPAVIIPAGHMLNGGVPVDFYFAQLAHDENKILPVSYQAECTLGRRIINGEREFPVIK	553	
RLI34429.1	Candidatus Bathyarchaeota archaeon	-EGSCPITLATSGMLGEGPAIEYFRELAPDPKNTVLVSYQIECTLGRRIQSGAKDVITI	565	
WP_014513953.1	Sulfolobus islandicus	AKGEPISILATSGMLNGGPAVEFFKTMADPPKNAIIFVSYQAECTLGRKVRDGAKEVQIL	554	
WP_011277529.1	Sulfolobus acidocaldarius	QAQEPSIILATSGMLNGGPAVEFFKTMADHTNRSIIFVSYQAECTLGRKVRDGAKEVQIL	547	
RLG13324.1	Candidatus Nanoarchaeota archaeon	ENKEPCVILITSGMITGGPIMEYLKCLAYDNKNTLVFIQYQAECTMGSRIQKGWDRIPMT	545	
OLC64863.1	Candidatus Rokubacteria bacterium	DDVEPCVILATSGMNGGVPVLEYFKSWADNPLHALLPVGFQSECSLGRRIQREARETIT	555	
OEU57241.1	Desulfuromonadales bacterium C00003096	GDGEPISILATSGMLNGGVPVLEHLKGLAGDEKNTLVFGYQAECTLGRRIQKGWDEIOLA	548	
OGM94118.1	Candidatus Wolfecubacteria bacterium	-ENKPCVILITSGMLTGGPAIEHLKQLAENPNLLFPASFQAECTLGKKIQKGWTEIPIG	548	
OGZ62538.1	Candidatus Staskawiczbacteria bacterium	-EGGCPVILATSGMLAGGASVEYFKFAVNEKNSIIFVSYQCGESLGRQVQDGAKEVQIL	547	
RKX64688.1	Tenericutes bacterium	SNIDPAIILITSGMNGGPIMEYLKAFAPDEKSTLIFVGYQAEQ-----	541	
NP_057291.1	Homo sapiens	DDIGPSVMAASPGMMQSGLSRELFSWCTDKRNGVILAGYCVETGLAKHIMSEPEITITM	377	
NP_650738.1	Drosophila melanogaster	EDIGPCVIMASPGMMQSGLSRELFSWCTDKRNGVILAGYCVETGLAKAVLSEPEITITL	383	
NP_502553.2	Caenorhabditis elegans	SGNDGCPVILATPGMLQSGFSRELFSWCPDTKNGCIIAGYCVETGLAKHILSEPEITVSL	378	
NP_013379.1	Saccharomyces cerevisiae	QDQGPSVMLASPGMLQSGLSRELFSWCPEDKNLVLITGYSIGTMAKFTMLEPDTIPSI	387	
		.... : * : . : : : *		
WP_011250379.1	Thermococcus kodakarensis	-GED--GRTEVINVMMEVHTIDGFSGHADRRRLMSYVARVRPRERIITVHGEAHHKCLDL	624	
WP_048048997.1	Methanosarcina mazei	-GKN--GSTEMLKMMMEVQVVDGFSGHSDRRQLMEYVKRMQPRPRVTEHGDCAKCVDL	613	
WP_004044063.1	Haloferax volcanii	-GRDSMGRSDTLKMDVETVDGFSGHADRRQLLENFVKTMNPRPEKVLVCHGDESRVQDL	616	
OT062505.1	Candidatus Woesearchaeota archaeon	-G-----SKEITQIKMEVDTMEGFTGSHGRKELLNFYVRCDDPKPKKIIVNHGESSRCLDL	609	
OLC36512.1	Thaumarchaeota archaeon	-GKE--GKIENVINSGTEKLDGFSGHSDYQNLMSYVHLRPLKLRVIVNHGERRKCENL	619	
OLS12352.1	Candidatus Lokiarchaeota archaeon_CR_4	-DAR--GRSQAVRMKLDVQITIEGFSGHSSRSQIMSFIRKVKPKPRVLTNHHGEAAKCVGL	614	
TAL51517.1	Nanoarchaeota archaeon	-GGQ--GRQETTFVKVMEIHTIEGFSGHSDNRQLLNFYVYRATPKPKKVLVNHGENSRCLDL	609	
WP_011007526.1	Pyrobaculum aerophilum	-SL--VGGESEKINMMREVVSIIPGFSGHSDRRRLMKYVEHMEPKPKKIIVLHGEPSKIIISL	610	
RLI34429.1	Candidatus Bathyarchaeota archaeon	-NPE--GKVEVINVMKRVESVEGFSGHSDNRQLLNFYVYRATPKPKKVLVNHGEPSKIIISL	622	
WP_014513953.1	Sulfolobus islandicus	-DRD--GRVESIQINMEVEAVEGFSGHSDRRQLNLFNLTIEPKPKNIILNHGEASATKAF	611	
WP_011277529.1	Sulfolobus acidocaldarius	-DRD--GRVENIKVMEIIGVEGFSGHSDRRQLLAFRLNLFNLTIEPKPKNIILNHGEASATKAF	604	
RLG13324.1	Candidatus Nanoarchaeota archaeon	-END--GTHMLKIEAEVETVEGFSGHSDNRQLNLYVAKLSKPKDRIFLNHGERSKSLNL	602	
OLC64863.1	Candidatus Rokubacteria bacterium	-DR--GQPLTLPLIKLDVETIDGFSGHSDRLQLLNFYVYRATPKPKKVLVNHGEFFKCSDL	611	
OEU57241.1	Desulfuromonadales bacterium C00003096	-D-E--GKIRNVKIKMDIHTVDGFSGHSDNRQLNLFVFRGIIKPLFSKVMCMHGENNKLAL	604	
OGM94118.1	Candidatus Wolfecubacteria bacterium	SGNN--GKPTTIQLKMQVQITIEGLSGHSDNRQLMSFIHRLAARPDVIVCHGEATKAVEF	606	
OGZ62538.1	Candidatus Staskawiczbacteria bacterium	-TD--GKDEFVKNVMEVHTSKGLTAHAGRNELMAFFKKNPKPRKVMVMHGEVSKCLDL	603	
RKX64688.1	Tenericutes bacterium	-----NPEITITPRCQVEEI-SFAAHVDQENLEFIEKIS--APNIILVHGEANPMGR	541	
NP_057291.1	Homo sapiens	S-----GQ--KLPLKMSVDYI-SFSAHTDYQQTSEFIRALK--PPHVILVHGEQNMESRL	427	
NP_650738.1	Drosophila melanogaster	S-----GQ--KLPLNMSVDYI-SFSAHTDYQQTSEFIRLLK--PTHVVLVHGEQNMESRL	433	
NP_502553.2	Caenorhabditis elegans	S-----GE--KLPMRMQGVYV-SFSAHTDYQQTSEFIRKALK--PPHVILVHGEQNMESRL	428	
NP_013379.1	Saccharomyces cerevisiae	N-----NPEITITPRCQVEEI-SFAAHVDQENLEFIEKIS--APNIILVHGEANPMGR	439	
WP_011250379.1	Thermococcus kodakarensis	SSSIHK-----KFGISTRAPNNLDAIRLK-----	648	
WP_048048997.1	Methanosarcina mazei	ASSVYK-----KLKIETRALNLETIVRLI-----	637	
WP_004044063.1	Haloferax volcanii	SSALYH-----DYNMRTFAPKNLETIFRFK-----	640	
OT062505.1	Candidatus Woesearchaeota archaeon	ASSIHK-----LNRVETAAPRNLEAIRLK-----	633	
OLC36512.1	Thaumarchaeota archaeon	ASSIHR-----MFRIQTCTPQVQEGRLIL-----	643	
OLS12352.1	Candidatus Lokiarchaeota archaeon_CR_4	ASVIHK-----RLKIDTRSLVNGETIVLR-----	638	
TAL51517.1	Nanoarchaeota archaeon	ASSIHK-----MFRIETVAPRNLDVRLK-----	633	
WP_011007526.1	Pyrobaculum aerophilum	ATSIEL-----KYKITTIIIPKVGIRIRAL-----	634	
RLI34429.1	Candidatus Bathyarchaeota archaeon	SSTIQR-----IFKIESRAPENLETIRLY-----	646	
WP_014513953.1	Sulfolobus islandicus	ANYIRDDR-----LGYPFIITPAILDSLRVA-----	638	
WP_011277529.1	Sulfolobus acidocaldarius	KRLVESGK-----IKGIDNVYAPKILESIRTI-----	631	
RLG13324.1	Candidatus Nanoarchaeota archaeon	ASTLHR-----MKFVETSVPSNLDVRLK-----	626	
OLC64863.1	Candidatus Rokubacteria bacterium	ASGLYK-----KFGLEAPAPNNLETIRLK-----	635	
OEU57241.1	Desulfuromonadales bacterium C00003096	ASGIHK-----KNIDTVPAPNNLETIRLV-----	628	
OGM94118.1	Candidatus Wolfecubacteria bacterium	ARAVHK-----NFRVETHAPRNLEAIRLK-----	630	
OGZ62538.1	Candidatus Staskawiczbacteria bacterium	ASSIYR-----INRMETNVPRLLETTINLR-----	627	
RKX64688.1	Tenericutes bacterium	-----	541	
NP_057291.1	Homo sapiens	KAALIREYEDNDE--VHIEVHNPRNTEAVTLNFRGEKLAKVMGFLADKKPEQGRVSGIL	485	
NP_650738.1	Drosophila melanogaster	KLALQREYEDAS--TDIKFYNPRNTHAVDLYFRGEKTAKVMGSLAANKSEVSGSLSGVL	491	
NP_502553.2	Caenorhabditis elegans	KSGIERQFQDD--N--IPIEVHNPRNTERLQLOFRGEKTAKVIGKLAQRPVENNETISGVL	481	
NP_013379.1	Saccharomyces cerevisiae	KSALLSNFASLKGTDNDEVHVNPRNCVDEVLEFQ-----	474	

WP_011250379.1	Thermococcus kodakarensis	-----	648
WP_048048997.1	Methanosarcina mazei	-----	637
WP_004044063.1	Haloferax volcanii	-----	640
OIO62505.1	Candidatus Woesearchaeota archaeon	-----	633
OLC36512.1	Thaumarchaeota archaeon	-----	643
OLS12352.1	Candidatus Lokiarchaeota archaeon_CR_4	-----	638
TAL51517.1	Nanoarchaeota archaeon	-----	633
WP_011007526.1	Pyrobaculum aerophilum	-----	634
RLI34429.1	Candidatus Bathyarchaeota archaeon	-----	646
WP_014513953.1	Sulfolobus islandicus	-----	638
WP_011277529.1	Sulfolobus acidocaldarius	-----	631
RIG13324.1	Candidatus Nanohaloarchaeota archaeon	-----	626
OLC64863.1	Candidatus Rokubacteria bacterium	-----	635
OEU57241.1	Desulfuromonadales bacterium C00003096	-----	628
OGM94118.1	Candidatus Wolfecubacteria bacterium	-----	630
OGZ62538.1	Candidatus Staskawiczbacteria bacterium	-----	627
RKX64688.1	Tenericutes bacterium	-----	541
NP_057291.1	Homo sapiens	VKRNFNHYHILSPCDLSNYTDLAMSTVKQTQAIPYTGPFNLLCYQLQKLTGDVEELEIQE-	544
NP_650738.1	Drosophila melanogaster	VKRDFFKHYLLAPSDLGKYTDMSSSVVTQRQSIWPGSSLSLTLELLLDRIAGGCVFV-LEA-	549
NP_502553.2	Caenorhabditis elegans	VKNFNFSYIMVPEELGYSYTLRISLSLEQRMSVHYSGSLKLLIFNLQQLNDDACLINIKL	545
NP_013379.1	Saccharomyces cerevisiae	-----	474

WP_011250379.1	Thermococcus kodakarensis	-----	648
WP_048048997.1	Methanosarcina mazei	-----	637
WP_004044063.1	Haloferax volcanii	-----	640
OIO62505.1	Candidatus Woesearchaeota archaeon	-----	633
OLC36512.1	Thaumarchaeota archaeon	-----	643
OLS12352.1	Candidatus Lokiarchaeota archaeon_CR_4	-----	638
TAL51517.1	Nanoarchaeota archaeon	-----	633
WP_011007526.1	Pyrobaculum aerophilum	-----	634
RLI34429.1	Candidatus Bathyarchaeota archaeon	-----	646
WP_014513953.1	Sulfolobus islandicus	-----	638
WP_011277529.1	Sulfolobus acidocaldarius	-----	631
RIG13324.1	Candidatus Nanohaloarchaeota archaeon	-----	626
OLC64863.1	Candidatus Rokubacteria bacterium	-----	635
OEU57241.1	Desulfuromonadales bacterium C00003096	-----	628
OGM94118.1	Candidatus Wolfecubacteria bacterium	-----	630
OGZ62538.1	Candidatus Staskawiczbacteria bacterium	-----	627
RKX64688.1	Tenericutes bacterium	-----	541
NP_057291.1	Homo sapiens	-----KPALKVF-----KNITVIQEPGMVLEWLANPSNDMYADTVTTVILEVQSNPK	592
NP_650738.1	Drosophila melanogaster	-----ERKLRFV-----GCIELTVEQKIIIVMEWQATHVNDVYADAVLACIMQSELGGT	597
NP_502553.2	Caenorhabditis elegans	KEISKKGVSVTQAITVFQGVNVTVVYGNHVVVVRWDSNPVYDMYADSVVAAILHAQANPV	605
NP_013379.1	Saccharomyces cerevisiae	-----	474

WP_011250379.1	Thermococcus kodakarensis	-----	648
WP_048048997.1	Methanosarcina mazei	-----	637
WP_004044063.1	Haloferax volcanii	-----	640
OIO62505.1	Candidatus Woesearchaeota archaeon	-----	633
OLC36512.1	Thaumarchaeota archaeon	-----	643
OLS12352.1	Candidatus Lokiarchaeota archaeon_CR_4	-----	638
TAL51517.1	Nanoarchaeota archaeon	-----	633
WP_011007526.1	Pyrobaculum aerophilum	-----	634
RLI34429.1	Candidatus Bathyarchaeota archaeon	-----	646
WP_014513953.1	Sulfolobus islandicus	-----	638
WP_011277529.1	Sulfolobus acidocaldarius	-----	631
RIG13324.1	Candidatus Nanohaloarchaeota archaeon	-----	626
OLC64863.1	Candidatus Rokubacteria bacterium	-----	635
OEU57241.1	Desulfuromonadales bacterium C00003096	-----	628
OGM94118.1	Candidatus Wolfecubacteria bacterium	-----	630
OGZ62538.1	Candidatus Staskawiczbacteria bacterium	-----	627
RKX64688.1	Tenericutes bacterium	-----	541
NP_057291.1	Homo sapiens	IRKGAQVQVSKKLEMHVYSKRLEIMLQDIFGEDCVSV-KDDSLSVTV-DGKTANLNLET	650
NP_650738.1	Drosophila melanogaster	NLKGATKQT--KSEDSRFRECLIEITLQDTFGDNCVPMFEGDLLPVTV-SGKRAEINLET	654
NP_502553.2	Caenorhabditis elegans	PDKYLP-----SNSSFPQFNTAIEGMVKHICGDDVSIVMSEGLLAQFEEDGRRLIVEGSS	661
NP_013379.1	Saccharomyces cerevisiae	-----	474

WP_011250379.1	Thermococcus kodakarensis	-----	648
WP_048048997.1	Methanosarcina mazei	-----	637
WP_004044063.1	Haloferax volcanii	-----	640
OIO62505.1	Candidatus Woesearchaeota archaeon	-----	633
OLC36512.1	Thaumarchaeota archaeon	-----	643
OLS12352.1	Candidatus Lokiarchaeota archaeon_CR_4	-----	638
TAL51517.1	Nanoarchaeota archaeon	-----	633
WP_011007526.1	Pyrobaculum aerophilum	-----	634
RLI34429.1	Candidatus Bathyarchaeota archaeon	-----	646
WP_014513953.1	Sulfolobus islandicus	-----	638
WP_011277529.1	Sulfolobus acidocaldarius	-----	631
RIG13324.1	Candidatus Nanohaloarchaeota archaeon	-----	626
OLC64863.1	Candidatus Rokubacteria bacterium	-----	635
OEU57241.1	Desulfuromonadales bacterium C00003096	-----	628
OGM94118.1	Candidatus Wolfecubacteria bacterium	-----	630
OGZ62538.1	Candidatus Staskawiczbacteria bacterium	-----	627
RKX64688.1	Tenericutes bacterium	-----	541
NP_057291.1	Homo sapiens	--RTVECEEGSEDDSLREMVLEAAQLRYEALTPVH-----	684
NP_650738.1	Drosophila melanogaster	--LAISCAEDDV---LRQMLNTTVQKLHQTLVSAL-----	684

**Figure B.11. Clustal-Omega alignment of diverse archaeal (coral), eukaryotic (green), and bacterial (blue) FttA-homologues.** BLASTp searches of the *T. kodakarensis* FttA protein (WP\_011250379.1; TK1428) against all bacterial genomes revealed full-length homologues – containing both KH domains and the metallo-beta-lactamase (MBL) fold – in candidate species of *Rokubacterium*, *Wolfebacterium*, and *Staskawiczbacterium*, as well as *Tenericutes bacterium* and *Desulfuromonadales bacterium* C00003096. It should be noted that all Mollicutes (e.g. *Tenericutes* spp.), all Cyanobacteria, and some Firmicutes are devoid of obvious rho homologues<sup>15</sup>. Many eukaryotic CPSF73 proteins contain C-terminal extensions not found in archaeal or bacterial homologues; an exception is the YSH1 protein of *S. cerevisiae*. Conserved (.), well-conserved (:), and invariant (\*) residues are noted. An active center histidine (H255 in *T. kodakarensis*) is highlighted in red.

## CHAPTER 3 SUPPLEMENTARY REFERENCES

1. Gehring, A., Sanders, T. & Santangelo, T. J. Markerless Gene Editing in the Hyperthermophilic Archaeon *Thermococcus kodakarensis*. *BIO-PROTOCOL* **7**, (2017).
2. Santangelo, T. J., Cubonová, L., James, C. L. & Reeve, J. N. TFB1 or TFB2 is sufficient for *Thermococcus kodakarensis* viability and for basal transcription in vitro. *J. Mol. Biol.* **367**, 344–57 (2007).
3. Hirata, A. *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. *Mol. Microbiol.* **70**, 623–33 (2008).
4. Sanders, T. J. *et al.* TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Mol. Microbiol.* doi:10.1111/mmi.14191
5. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–54 (1976).
6. Gehring, A. M. & Santangelo, T. J. Manipulating Archaeal Systems to Permit Analyses of Transcription Elongation-Termination Decisions In Vitro. in *Methods in molecular biology* (Clifton, N.J.) **1276**, 263–279 (2015).
7. Walker, J. E., Luyties, O. & Santangelo, T. J. Factor-dependent archaeal transcription termination. *Proc. Natl. Acad. Sci.* **114**, E6767–E6773 (2017).
8. Santangelo, T. J., Cubonová, L. & Reeve, J. N. *Thermococcus kodakarensis* genetics: TK1827-encoded beta-glycosidase, new positive-selection protocol, and targeted and repetitive deletion technology. *Appl. Environ. Microbiol.* **76**, 1044–52 (2010).
9. Li, Z., Santangelo, T. J., Cubonová, L., Reeve, J. N. & Kelman, Z. Affinity purification of an archaeal DNA replication protein network. *MBio* **1**, e00221-10-e00221-19 (2010).
10. Burkhart, B. W., Febvre, H. P. & Santangelo, T. J. Distinct Physiological Roles of the Three Ferredoxins Encoded in the Hyperthermophilic Archaeon *Thermococcus kodakarensis*. *MBio* **10**, (2019).
11. Jäger, D., Förstner, K. U., Sharma, C. M., Santangelo, T. J. & Reeve, J. N. Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis*. *BMC Genomics* **15**, 684 (2014).
12. Mattioli, F. *et al.* Structure of histone-based chromatin in Archaea. *Science* (80-. ). **357**, 609–612 (2017).
13. Nishida, Y. *et al.* Crystal structure of an archaeal cleavage and polyadenylation specificity factor subunit from *Pyrococcus horikoshii*. *Proteins* **78**, 2395–8 (2010).
14. Mandel, C. R. *et al.* Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* **444**, 953–6 (2006).
15. D'Heygere, F., Rabhi, M. & Boudvillain, M. Phyletic distribution and conservation of the bacterial transcription termination factor Rho. *Microbiology* **159**, 1423–1436 (2013).

## APPENDIX C: HISTONES DIRECT SITE-SPECIFIC CRISPR SPACER ACQUISITION IN MODEL ARCHAEON

### Main

Prokaryotic CRISPR-Cas systems provide heritable, adaptive immunity against viruses, plasmids and other parasitic mobile genetic elements (MGEs) by maintaining a record of previously-encountered MGEs in a repository called the CRISPR array<sup>1</sup>. The array is a series of short (typically 30-40 bp) direct repeats separated by similarly-sized fragments derived from invading MGE genomes, termed spacers. The organization of this array gave rise to the descriptive name for these systems: Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). Adjacent to the CRISPR array is a regulatory DNA region called the leader which encodes promoter elements used for CRISPR RNA expression as well as sequence elements important for directing spacer integration<sup>2-5</sup>. The CRISPR array is transcribed and processed into mature CRISPR RNAs (crRNA)<sup>6</sup> which guide CRISPR ribonucleoprotein (crRNP) effector complexes to recognize and degrade nucleic acid of invading MGEs in a sequence-specific manner<sup>7-10</sup>.

Uniquely among prokaryotic MGE defense systems, CRISPR-Cas immunity is adaptive: the array can be continuously updated through addition of new spacers against newly encountered MGEs<sup>11</sup>. DNA fragments are sampled and integrated into CRISPR arrays by integrase complexes composed of highly conserved Cas1 and Cas2 proteins<sup>7,12,13</sup>, together with accessory proteins and factors which vary among the different CRISPR-Cas subtypes and host organisms<sup>14-20</sup>.

---

<sup>3</sup>This appendix was previously published under the following title: Watts, E. A. *et al.* Histones direct site-specific CRISPR spacer acquisition in model archaeon. *Nat. Microbiol.* (2023).

The mechanism of spacer integration involves two sequential transesterification reactions whereby the 3' OH groups of the spacer DNA make specific nucleophilic attacks at the 5' borders of the repeat, followed by resolution of the integration site and DNA repair<sup>21,22</sup> (Extended Data Fig. C.1). The end result is the expansion of the array by one new spacer-repeat unit. A common feature of diverse CRISPR systems in both bacterial and archaeal organisms is that new spacers are integrated into the CRISPR array in a polarized manner with a strong preference for the leader-adjacent repeat rather than downstream repeats within the array<sup>11,23</sup>.

Work with different CRISPR systems has revealed that polarized spacer integration proceeds by one of two different modes *in vivo*. In the first mode, certain Cas1-2 integration complexes exhibit intrinsic specificity for the leader-proximal repeat: the complex directly recognizes sequence elements in the leader-proximal repeat as well as leader sequences located immediately adjacent to that repeat<sup>12,24,25</sup>. This intrinsic directionality is exemplified by the Type II-A CRISPR-Cas systems of *Streptococcus thermophilus*, *Streptococcus pyogenes* and *Enterococcus faecalis* where purified Cas1-Cas2 complexes specifically integrate new spacers at the first repeat *in vitro* without additional factors<sup>12,24,25</sup>.

The second, and perhaps dominant mode of polarized integration, is factor-mediated rather than sequence intrinsic. This is typified by Type I-E and I-F CRISPR-Cas systems of *E. coli* and *P. aeruginosa*, where the polarity of integration is guided by a host-encoded chromosomal DNA binding protein, Integration Host Factor (IHF){Santiago-Frangos, 2021 #22366;Yoganand, 2017 #6482;Fagerlund, 2017 #6615;Nunez, 2016 #8293;Wright, 2017 #22368;Santiago-Frangos, 2023 #22672}. *In vivo*, IHF is required for any detectable new spacer acquisition<sup>27,29</sup>. *In vitro*, purified Cas1-2 integration complexes are capable of spacer integration at CRISPR repeat sequences, but integration occurs at all repeats in an array with no preference for the leader

end. When purified IHF is added to the *in vitro* integration reaction, spacer acquisition occurs almost exclusively at the leader-proximal repeat, demonstrating its importance in *bona fide* CRISPR spacer acquisition<sup>21,26,28-30</sup>. IHF binds at one or more conserved sites within the leader and induces a sharp horse shoe-like bend in the leader DNA which brings an upstream leader motif into close proximity to the repeat-bound Cas1-2 complex, strongly promoting spacer integration at the leader-proximal repeat<sup>26,27,29-31</sup>.

Genes encoding IHF are restricted to mostly Proteobacteria<sup>26</sup>, so it remains unclear how factor-mediated polarized integration would occur in other bacteria and in Archaea. In the Crenarchaeon, *Saccharolobus solfataricus*, *in vitro* reconstituted Cas1-2 complexes catalyze spacer integration into purified plasmid DNA, but without leader or repeat specificity. When cell lysate is added to the reaction, spacer integration is directed to the leader-adjacent repeat, but the lysate component responsible for this effect remains unknown<sup>32</sup>.

CRISPR spacer uptake in the model hyperthermophilic archaeon, *Pyrococcus furiosus*, is relatively well-characterized, and as with most other species, spacer addition occurs preferentially at the leader-adjacent repeat<sup>14,33,34</sup>. The *P. furiosus* genome harbors 7 separate CRISPR arrays that each contain the same repeat sequences and highly conserved leaders<sup>35</sup>. Spacer integration at all 7 arrays is catalyzed by a single Cas1-2 integrase complex<sup>14,36,37</sup>. Previous work suggests that *P. furiosus* utilizes a factor-mediated, polarized spacer integration mechanism since reconstituted Cas1-Cas2 integrase from this organism adds spacers to CRISPR repeats without specificity for the first repeat<sup>37</sup> *in vitro* while new spacers are only observed being added at the first repeat *in vivo*<sup>14,33,34</sup>. These data suggest an unidentified factor is required to promote leader-proximal spacer integration in *P. furiosus*. Here we present evidence that archaeal histones are responsible for guiding directional spacer integration into CRISPR arrays in *P. furiosus*.



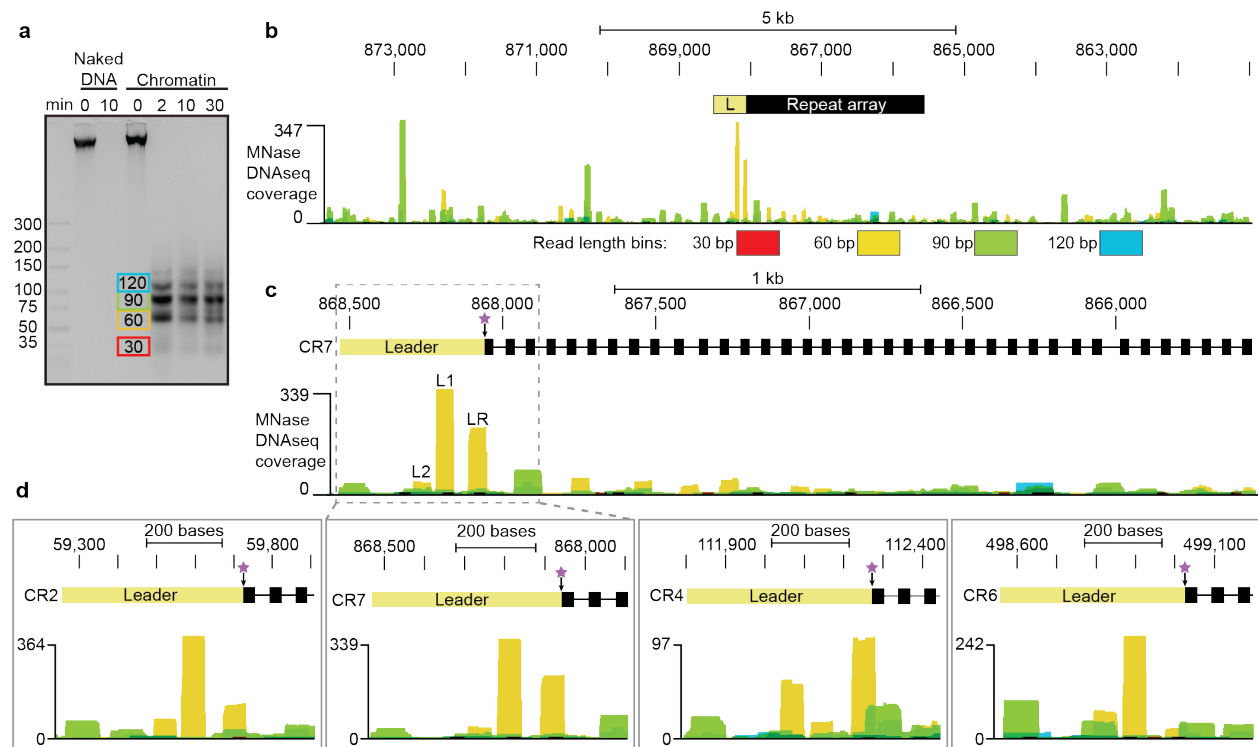
## Results

### ***MN-Seq reveals histone occupancy at CRISPR leaders***

With the aim of identifying the hypothesized factor(s) responsible for directing leader-proximal integration in *P. furiosus* CRISPR arrays<sup>37</sup>, we conducted a genome-wide analysis of DNA binding proteins and then looked for evidence of DNA binding at CRISPR loci. For this, we carried out micrococcal nuclease digestion of chromatin from *P. furiosus* COM1 strain followed by high-throughput sequencing (MN-Seq) of protected (protein-bound) DNA fragments (Fig. C.1). Upon digestion, we obtained a ladder-like pattern of protected DNA fragments, with particularly strong bands for 60, 90, and 120 bp sized products (Fig. C.1a). To visualize patterns of MNase-protected DNA across the genome, the protected DNA fragments were excised and sequenced and resulting reads were aligned to the reference *P. furiosus* genome (Fig. C.1b-d). We generated color-coded sequencing tracks to distinguish the four main fragment sizes of 30, 60, 90 and 120 bp. In general, most of the genome had some level of coverage, with more pronounced peaks interspersed throughout (Fig. C.1b; Extended Data Fig. C.2). There were few reads in the 30 bp size range, consistent with the lower intensity of that band in gel images (Fig. C.1a).

Interestingly, we observed relatively high protection peaks in the leaders of the CRISPR arrays (Fig. C.1b-d). In each of the seven CRISPR arrays, three 60 bp leader peaks were evident (Fig. C.1b-d). The three peaks were not equally pronounced in each array but were consistently observed in different growth conditions and replicates (Extended Data Fig. C.3). We named these peaks LR, L1, and L2. The LR peak partially overlapped the Leader-Repeat junction and site of spacer integration. The LR peak was bounded on the left side by the promoter for CRISPR array expression and the right side by the first CRISPR repeat (Extended Data Fig. C.4). The L1 peak was located ~100 bp upstream from the leader-repeat junction and was generally the site with the greatest level of protection as detected by MN-seq. The L2 peak,

located further upstream, was the least pronounced at most CRISPR loci (Fig. C.1, Extended Data Fig. C.3, C.4).



**Figure C.1. MNase protection assay to characterize *P. furiosus* DNA binding proteins.** (a) Gel image of DNA fragments generated through micrococcal nuclease digestion of *P. furiosus* chromatin. Colored boxes highlight major bands of protected DNA. (b-d) DNA bands highlighted in A were sequenced by HTS and mapped to *P. furiosus* genome; genome browser tracks show the distribution of aligned reads across representative CRISPR loci. Overlapping bars are color-coded according to the size of the DNA fragment (red: 30 +/- 5 bp; golden yellow: 60 +/- 5 bp; green: 90 +/- 5 bp; blue 120 +/- 5 bp). The y-axis indicates the cumulative depth of read coverage. Six replicates were sequenced and yielded similar patterns; a representative replicate is shown here. Panel b shows a wide, approximately 10 kb, view of the genome space surrounding CRISPR7 (leader in pale yellow, repeat array in black), while panel c shows a closer view of full CRISPR7 array. (d) Insets below show the leader and first three repeats for CRISPR7, CRISPR2, CRISPR4 and CRISPR6.

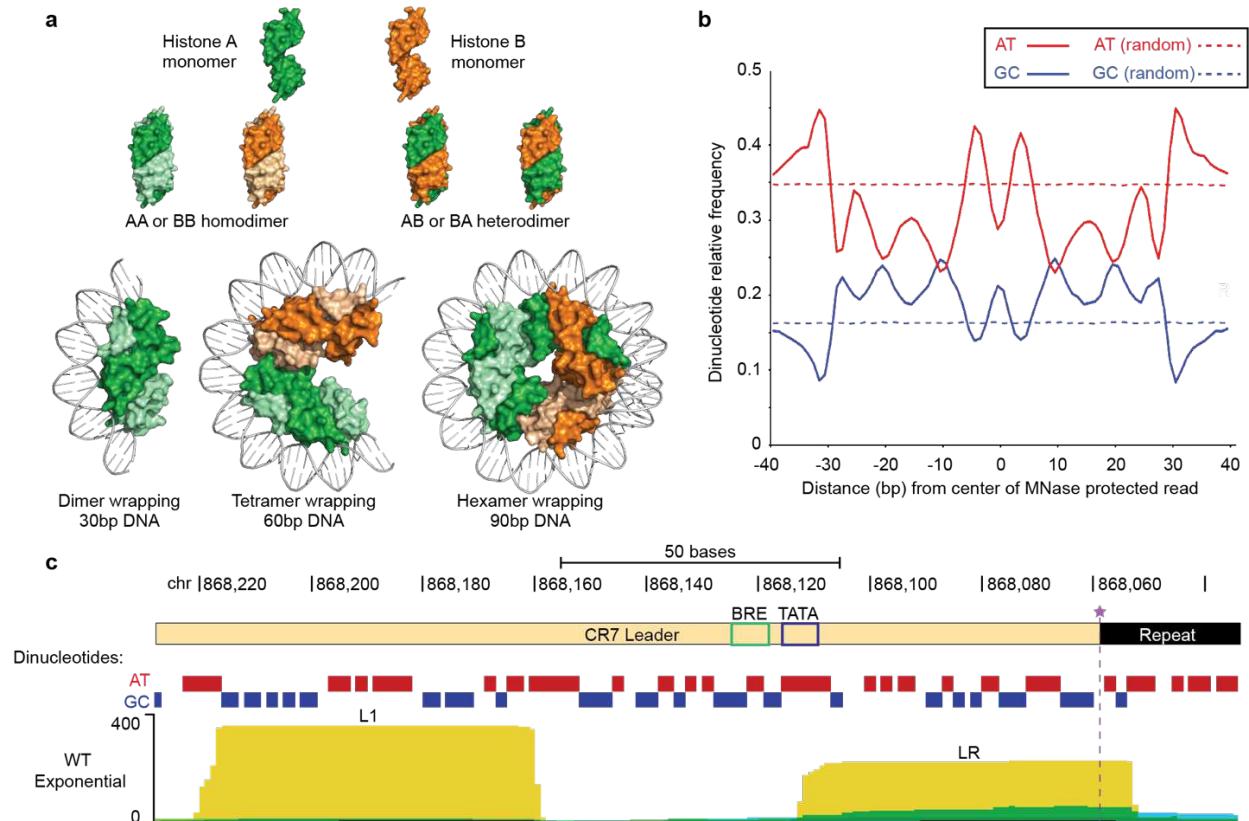
We noted that the leader peaks were almost exclusively comprised of 60 bp reads (Fig. C.1b-d, golden yellow). In contrast, upstream and downstream from the arrays, and generally elsewhere in the genome, yellow peaks, corresponding to 60 bp aligned reads, were usually much lower (Fig. C.1b; Extended Data Fig. C.2). We evaluated genome-wide 60 bp read distributions by examining read depth in 60 bp sliding windows across the genome and found that the tallest

leader peaks were in the 99.9<sup>th</sup> percentile, indicating they are outliers (Extended Data Fig. C.3). We also noted that most of the other relatively tall peaks (outside of the leaders) were comprised of 90 bp reads (green), and many prominent 60 bp read peaks also had some 90 bp read overlap (Fig. C.1b, Extended Data Fig. C.2). In contrast, for the LR and L1 peaks, 60 bp read depth was about 15 times higher than 90 bp read depth, on average, a bias which was relatively rare elsewhere in the genome (Extended Data Fig. C.3). Other than the peaks of 60 bp MN-seq coverage in the leaders, CRISPR arrays did not display unusually dense coverage (Fig. C.1b, Extended Data Fig. C.2). These interesting features prompted us to consider that MNase footprints in CRISPR leaders may reflect unique functions for DNA-binding proteins in this region.

Like most archaea<sup>38-40</sup>, *P. furious* encodes histone proteins that bind and wrap DNA, thus protecting the DNA from MNase digestion<sup>41,42</sup> (Fig. C.2a). *P. furious* encodes two similar but distinct histone isoforms (A and B) predicted to form both homodimers and heterodimers in solution<sup>41,43,44</sup>. These histones are expected to bind and wrap differing lengths of DNA, depending on the number of histone subunits in a complex: 30 bp for a dimer, 60 bp for a tetramer (or dimer of dimers), 90 bp for a hexamer (trimer of dimers) and so on (Fig. C.2a), which would result in the observed 30 bp laddering pattern of protected DNA when chromatin is treated with MNase (Fig. C.1a). Histone tetramers (which would protect 60 bp of DNA like that observed at the CRISPR leaders (Fig. C.2c)) are predicted to induce sharp, horseshoe-like or U-turn bends in the DNA<sup>38,41</sup> (Fig. C.2a).

Histone binding patterns have been studied in a wide range of organisms, leading to the identification of sequence features that promote or repel stable histone binding along stretches of DNA<sup>45,46</sup>. We examined the global characteristics of 60 bp MNase-protected reads in our data set and observed the hallmark phased helical repeats of AT and GC dinucleotides expected for histone binding<sup>45,46</sup>. Focusing on CRISPR leaders, the sequences underlying LR, L1 and L2

were consistent with this pattern of alternating AT and GC dinucleotides (Fig. C.2c). Together, these presumed 60 bp histone tetramer binding footprints in CRISPR leaders appeared to be a unique and interesting feature, suggesting the possibility of a functional role of histone binding in polarized spacer integration.

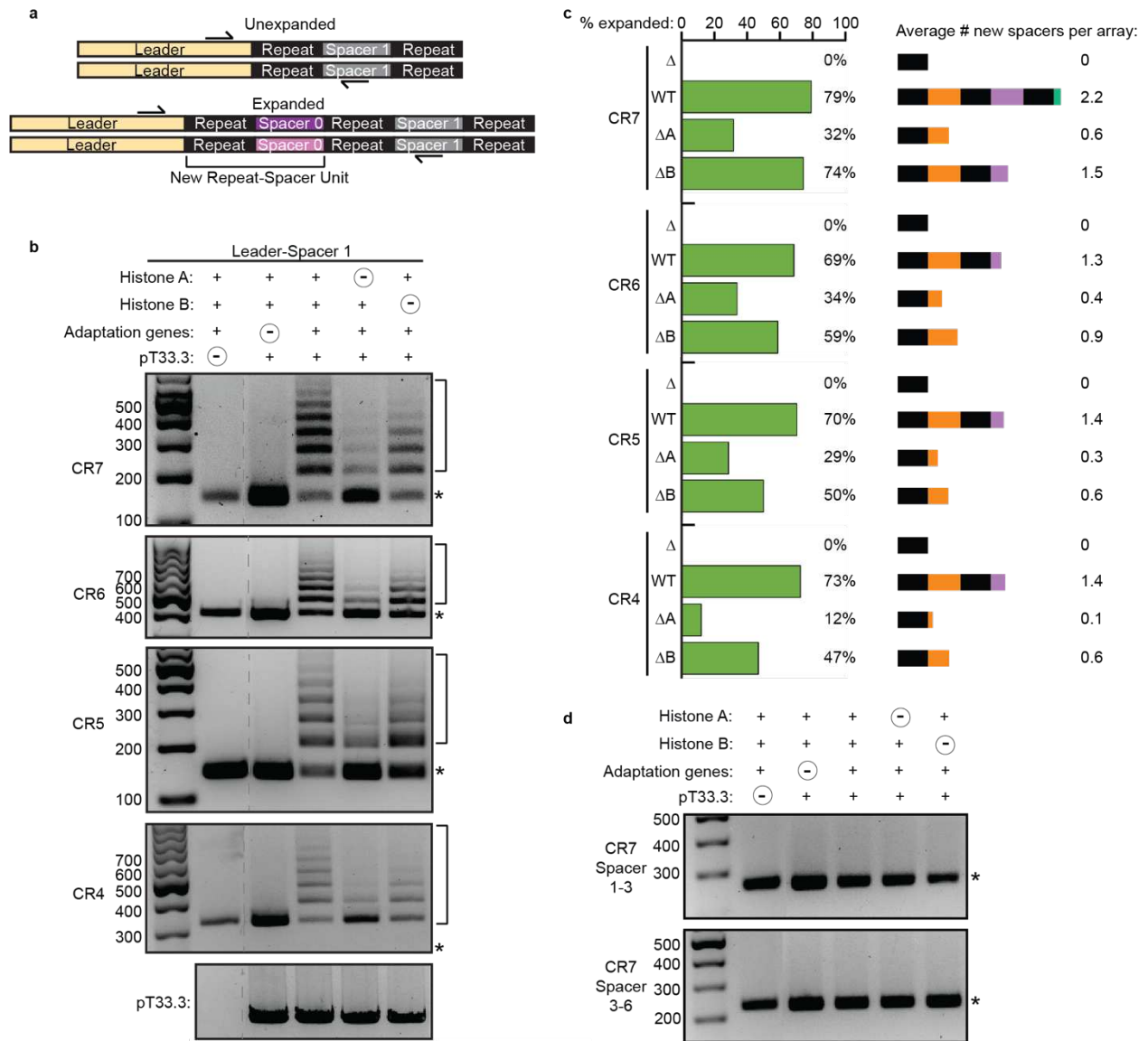


**Figure C.2. DNA binding by histones.** (a) Crystal structure of DNA-bound histones from *Methanothermobacter fervidus* (PDB 5T5K)<sup>65</sup>. (b) Globally, MNase-protected DNA fragments in *P. furiosus* had a non-random distribution of AT and GC dinucleotides along their lengths, with the 10bp periodicity observed for histones in eukaryotes and other archaea. (c) Global AT and GC dinucleotide periodicity trends are also observed over the 60 bp protected peaks in CRISPR leaders.

### ***Histones promote adaptation in vivo***

Based on the prediction that histone-complexes were specifically bound to CRISPR loci, we tested whether histones were important for spacer integration *in vivo* (Fig. C.3). Specifically, we transformed *P. furiosus* strains with a recently identified conjugative plasmid, called pT33.3, that was found to greatly stimulate spacer uptake into CRISPR arrays<sup>47</sup>. We used our established

PCR-based approach (Fig. C.3a) to test the ability of *P. furiosus* strains to acquire new spacers into four separate CRISPR arrays in a strain containing both histones A and B (the *P. furiosus* COM1 parental strain, referred to here as wildtype) as well as in strains lacking either histone A or histone B (Fig. C.3b). We were unable to generate a strain lacking both histones A and B, presumably because at least one histone is required for cell viability<sup>48</sup>. Both the  $\Delta$ histone A and  $\Delta$ histone B strains were viable, showed typical growth rates, and showed largely unchanged gene expression profiles by RNA-seq as compared to wildtype (Extended Data Fig. C.5a-c). As a negative control, we also examined spacer uptake in a strain wherein Cas1 and Cas2 are deleted and adaptation does not occur. Relative to wildtype, deletion of genes for either histone A or B led to a decrease in both the percentage of cells taking up spacers and the number of integrated spacers per array for each of the four arrays examined (Fig. C.3b, c). Ectopic spacer integrations at downstream repeats were not observed irrespective of histone availability (Fig. C.3d). It appeared that loss of histone A had a greater impact on spacer uptake than did loss of histone B (Fig. C.3b, c). To determine whether this effect could be due to differing levels of total histone abundance in these strains, we used Western blotting to estimate the concentration of histone proteins from crude cell lysates. Blots showed that total histone expression was not reduced for the  $\Delta$ histone A strain (Extended Data Fig. C.5d), implying that total histone abundance is not responsible for the greater impact histone A deletion had on adaptation. Western blots also showed that the three strains had similar chromatin organization, as larger multimeric complexes were visible for all (Extended Data Fig. C.5d). Together, these results show that histones A and B promote spacer uptake *in vivo*.

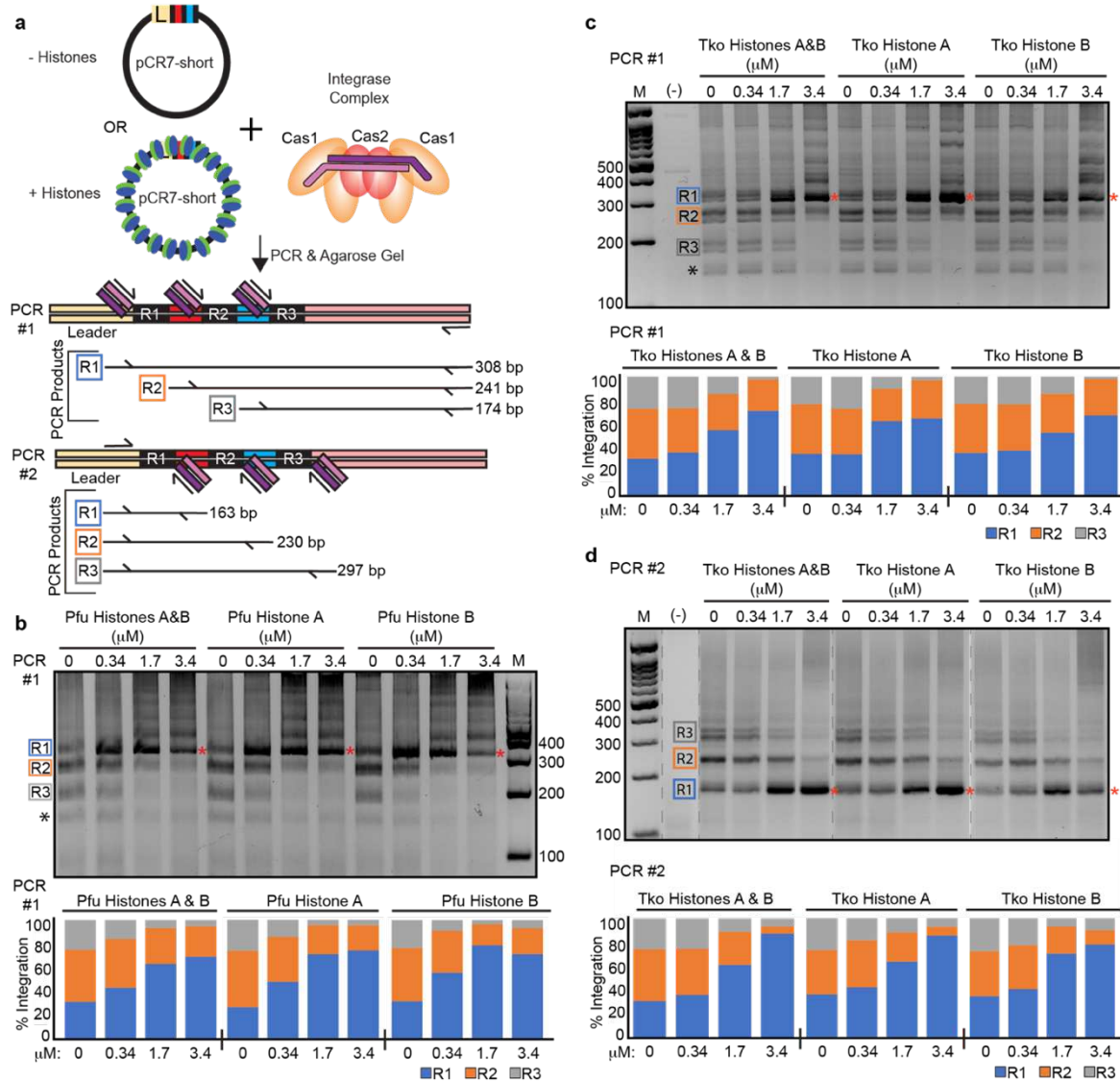


**Figure C.3. Histone gene deletions alter spacer integration at the leader-adjacent repeat *in vivo*.** (a) Schematic diagram of a PCR-based, *in vivo* adaptation assay to detect integration of new spacers at the leader-adjacent repeat. The relative position of PCR primers is labelled by arrows (b) Gel image of PCR products from the assay shown in a. Strains used here contained only the type I-B effector complex. Where indicated, strains were transformed with the conjugative plasmid pT33.3 to enhance adaptation. \* = unexpanded array, [ ] = expanded arrays. The assay was performed with primers for all seven CRISPR arrays to evaluate consistency of the phenotype (CR2, CR4, CR5, and CR6 are shown here). (c) Intensity of all PCR bands, as in B, was quantified using ImageJ. The proportion of expanded to unexpanded products was determined, along with the average number of new spacers in an array, as calculated based on relative band intensity.  $\Delta$  = spacer integration genes deleted ( $\Delta cas1$ ,  $\Delta cas2$ ,  $\Delta cas4-1$ ,  $\Delta cas4-2$ ) (d) Primers that amplify downstream from the leader-adjacent repeat (spacer 1-3, or 3-6, as labeled) were used to detect ectopic (ie internal to the array) spacer integrations.

### ***Histones guide proper spacer integration in vitro***

We turned to *in vitro* studies to further characterize the role of histone proteins in specifying spacer integration (Fig. C.4a). We had previously reconstituted CRISPR spacer integration *in vitro* by incubating purified *P. furiosus* Cas1-Cas2 integrase complexes with a synthetic spacer and a DNA substrate, a plasmid (called pCR7-short) encoding a CRISPR array containing a leader and three repeat-spacer units<sup>37</sup> (Fig. C.4a). Sites of spacer integration into pCR7-short are detected via PCR using a primer complementary to the newly acquired spacer and a primer downstream of the CRISPR array (Fig. C.4a). Previously, we reported that Cas1-Cas2 integrate spacers at each of the three repeats without major preference (<sup>37</sup> and Fig. C.4a). Here we added purified histones from either *P. furiosus* or from *Thermococcus kodakarensis*, a related species of Thermococcales with two histone genes which are well-studied and highly similar to those of *P. furiosus* (Extended Data Fig. C.6). We compared the spacer integration results for histone-free plasmid DNA vs. the same plasmids following incubation with histones A, B, or both combined (i.e. using chromatin CRISPR DNA substrates; Fig. C.4a). As expected from previous results<sup>37</sup>, in the absence of histones, Cas1-Cas2 complexes catalyzed integration at each of the three repeats with no apparent preference (Fig. C.4b). In contrast, addition of purified histones (A alone, B alone, or a mixture of A and B) led to a concentration-dependent shift in spacer integration to the leader proximal repeat (i.e. repeat 1) (Fig. C.4b). We observed the same effect when either *P. furiosus* (Fig. C.4b, c) or the *T. kodakarensis* histones proteins were used (Fig. C.4d-g). PCR products sometimes appeared as doublets, which we believe is due to variable electrophoretic mobility of the same PCR product in the ethidium bromide-stained agarose gels (Fig. C.4d). Quantification of relative PCR band intensities showed that histones A and B from both archaeal species increased preference for the leader-proximal repeat 1 from ~30% (histone-free conditions) to between 60 and 80% (at the highest histone concentration tested) (Fig. C.4c, e, g). These results provide evidence that both histone A or B proteins *per se*,

promote integration of spacers by the Cas1-Cas2 complex at the leader-adjacent repeat, mirroring the integration site used *in vivo*.

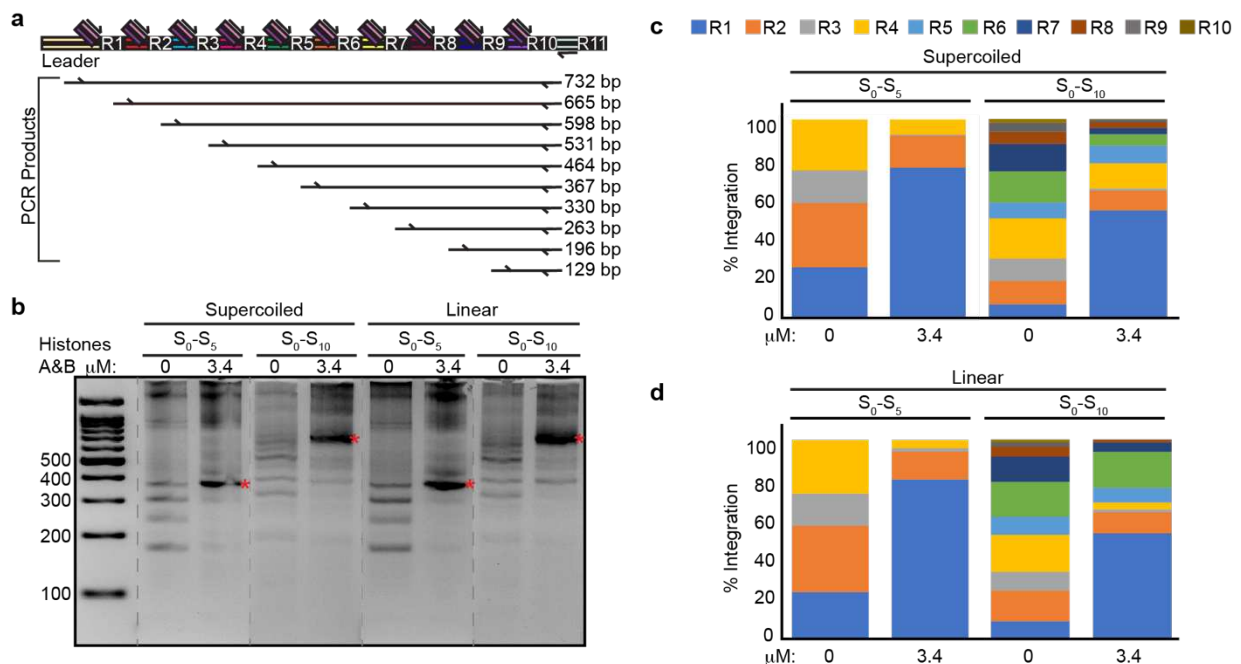


**Figure C.4. *In vitro* evaluation of spacer integration in the presence of archaeal histones.**

(a) Diagram of experimental setup; spacers are integrated into pCR7-short by the Cas1/Cas2 adaptation complex in the presence or absence of histones A and/or B (from either *P. furiosus* or *T. kodakarensis*). PCR was then used to assess the positions and relative efficiencies of *in vitro* integration events into pCR7-short; primers spanned from the inserted spacer to a downstream plasmid position (PCR #1) or from the leader to the inserted spacer (PCR #2). (b, c) Gel image shows products from PCR #1 for assay done with *P. furiosus* (Pfu, b) or *T. kodakarensis* (Tko, c) histones. The expected sizes for PCR products resulting from integrations at repeat 1, repeat 2, and repeat 3 are marked with R1, R2, and R3, respectively. Red asterisks mark the band for integration at repeat 1, which is the natural, preferred point of integration *in vivo*. Histograms below each gel image show relative band intensities, quantified using ImageJ. (d) Gel image shows products from PCR #2 for integration assay done with *T. kodakarensis* histones. Histograms below the gel image show band intensity quantifications, as above.



Several additional experiments further confirmed a specific role for histones in guiding or confining spacer integration to the leader-adjacent repeat *in vitro*. First, when the plasmid-encoded CRISPR array was expanded from three spacer-repeat units (pCR7-short) to eleven spacer-repeat units (pCR7-long; Fig. C.5a), we observed the same phenomenon: integrations occurred at all repeats in the absence of histones but were mostly at the first repeat in the presence of histones (Fig. C.5b; lane 2 versus 3, 4 versus 5, C.5c). The effect was observed with both supercoiled plasmid DNA or linearized plasmid DNA (Fig. C.5b – lane 6 versus 7, and lane 8 versus 9, Fig. C.5d) implying that histones don't simply act by inducing changes to DNA supercoiling. We also found that the effect was not unique to the CR7 array. Histones increased the preference for repeat 1 in all of the three additional arrays that we tested *in vitro* (plasmids bearing CR5, CR6, and CR8; Extended Data Fig. C.7). To confirm that purified *P. furiosus* histones did indeed bind the plasmid-encoded leader, we conducted an MNase protection assay on pCR7-long mixed with purified histones A and B, followed by sequencing. L1 and LR protection peaks were observed in the same relative positions as seen *in vivo* (Extended Data Fig. C.8).



**Figure C.5. *In vitro* spacer integration into different DNA substrates in the presence of *P. furiosus* histones.** (a) Diagrammatic representation of the PCR used to assess *in vitro* integration into pCR7-long. Primer one binds the spacer ( $S_0$ ), while primer two binds either spacer 5 ( $S_5$ ) or spacer 10 ( $S_{10}$ ); in this diagram spacer 10 products are outlined (b) Representative gel image shows products from the PCR outlined in a; pCR7-long was either supercoiled or linearized by inverse PCR. The  $S_0$ - $S_5$  label indicates PCR products when the spacer 5 primer was used;  $S_0$ - $S_{10}$  label indicates PCR products when spacer 10 primer was used. Red asterisks highlight the repeat 1 band. (c-d) Intensity of all PCR bands in b was quantified using ImageJ and the proportion of integration events at all repeats was determined for both supercoiled (c) and linear (d) plasmid DNA.

To further address the specificity of the histone-mediated effects on shaping the site of spacer integration, *in vitro* spacer integration assays were carried out with two other abundant DNA binding proteins found in *P. furiosus*, TrmBL2<sup>49</sup> and Alba<sup>38,50</sup>. Titration of either protein failed to reproduce the changes in integration observed for histones A and B, although higher concentrations of TrmBL2 appeared to suppress integration at all repeats (Extended Data Fig. C.9; purified proteins shown in Extended Data Fig. C.6). Finally, we verified the PCR-based results using an established sequencing-based approach that does not rely on array-specific primers<sup>37</sup>. Without histones, numerous integration events were detected at each CRISPR 5' repeat boundary (Extended Data Fig. C.10). However, with histones there was a clear shift in the number of integrations at the leader-adjacent repeat and that preference scaled with histone concentration (Extended Data Fig. C.10). In contrast, addition of the Alba protein did not alter the sites of integration across repeat junctions at any concentration (Extended Data Fig. C.10). Taken together, the *in vitro* studies provide compelling evidence that histones play a specific role in directing spacer integrations to the leader-adjacent repeat.

## Discussion

Prokaryotic CRISPR arrays serve as memory banks of past infection that are used to immunize an organism against potentially deleterious viruses and other MGEs. This immunity is passed on to future generations, making CRISPR the only example of adaptive immunity found in prokaryotes. The polarized acquisition of new spacers to the leader-adjacent CRISPR repeat is ubiquitous throughout diverse CRISPR-Cas systems of Bacteria and Archaea<sup>51</sup>. To date, only

one bacterial host factor has been shown to drive leader-proximal spacer integration – the proteobacterial IHF protein{Santiago-Frangos, 2021 #22366;Yoganand, 2017 #6482;Fagerlund, 2017 #6615;Nunez, 2016 #8293;Wright, 2017 #22368;Santiago-Frangos, 2023 #22672}}.

Beyond this example, the identity of host factors underlying spacer integration polarity remains unknown for the majority of CRISPR-containing species. Here, *in vivo* and *in vitro* findings reveal that histones play an important role in CRISPR immunity by serving as a key host factor that directs site-specific CRISPR spacer acquisition in the model archaeon, *P. furiosus*.

While the molecular details of how the tested *P. furiosus* and *T. kodakarensis* histones act to promote spacer integration at the leader-adjacent repeat are not yet understood, we propose a general model which mirrors that of IHF (Fig. C.6). Both IHF and histones are abundant host chromosomal binding proteins which wrap DNA, and in doing so, induce sharp bends and topological changes. In the case of IHF-mediated integration, one or two IHF dimers bind conserved IHF binding motifs present in the CRISPR leader<sup>26</sup> and bend the DNA such that upstream leader motifs contact the Cas1-2 spacer integration complex (Fig. C.6). In the *E. coli* system, this IHF binding has been proposed to expose the leader-repeat phosphate for nucleophilic attack by the 3'-OH of the incoming spacer present within the Cas1-Cas2 integrase complex<sup>25,29</sup>. In our model, histone tetramers bind to one, two or even three 60 base-pair conserved leader-encoded motifs and are predicted to bend the DNA in a manner similar to IHF (Fig. C.6).

There are several possibilities for how DNA binding and bending by histones could promote integration at the leader-proximal repeat in *P. furiosus*. Leader bending may allow a secondary upstream motif to contact Cas1-2 complex, as in the *E. coli* IHF system{Santiago-Frangos, 2021 #22366;Yoganand, 2017 #6482;Wright, 2017 #22368;Santiago-Frangos, 2023 #22672}}.

Alternatively, the Cas1-2 complex may directly contact the histone tetramer and some aspect of this physical interaction could aid integration. A third possibility is that the bending or torsion

induced by the histones may promote integration by making the leader-repeat phosphate more exposed for nucleophilic attack by the spacer DNA. We expect that the site of integration at the leader-repeat junction would be readily accessible on the outer surface of the LR positioned histone tetramer (Fig. C.6). Additional experiments and structural data will be necessary to understand if one or more of these possible mechanisms underlie directional spacer integration in *P. furiosus*.

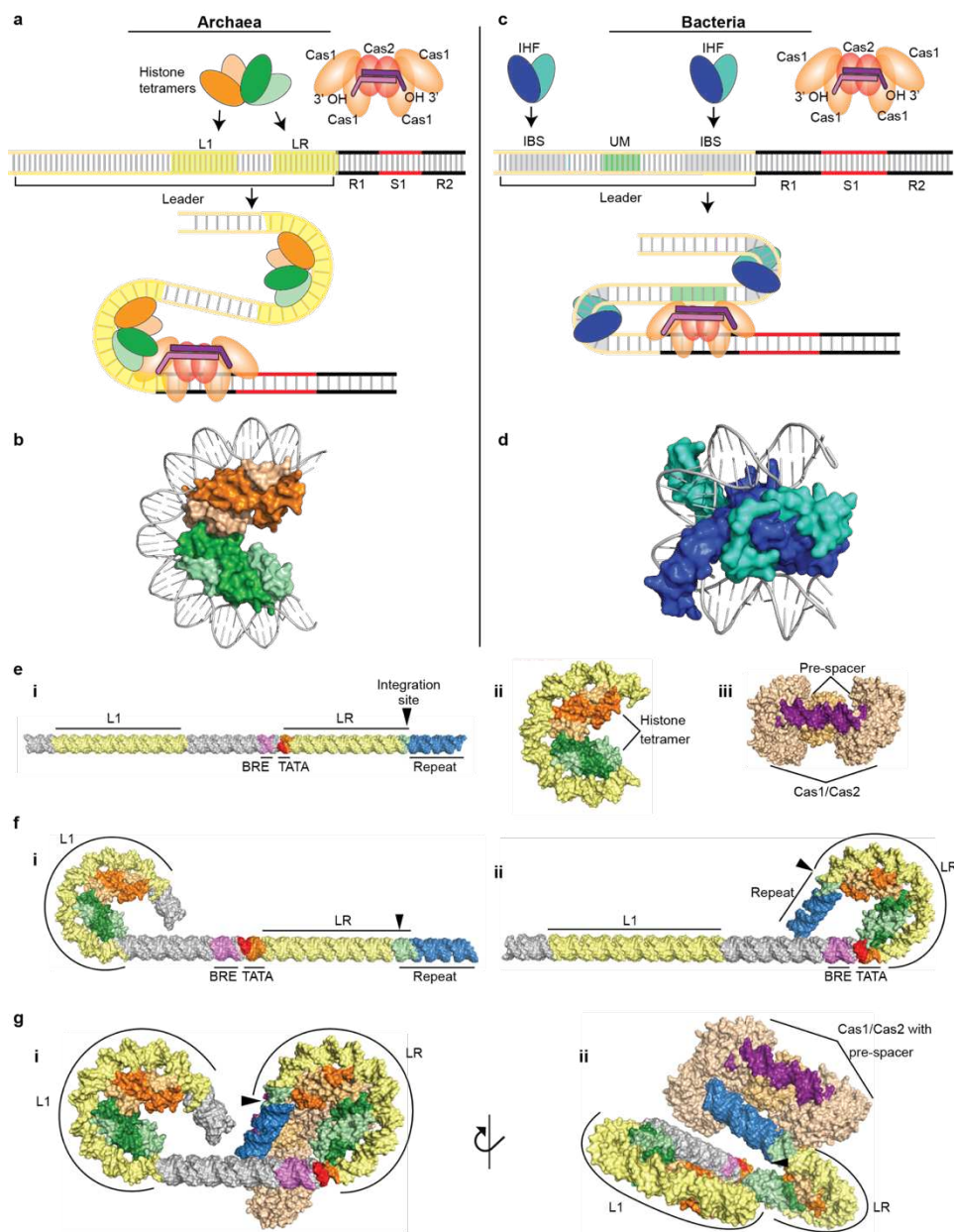
Another interesting parallel between IHF and histones in our study is that both are required for detectable integration in their respective systems *in vivo*, but for these same systems *in vitro*, Cas1 and Cas2 efficiently carry out integration at array repeats without leader-end specificity. This discrepancy between *in vivo* and *in vitro* effects suggests that there may be other factors or cellular features *in vivo* that suppress integration at downstream repeats, even in the absence of leader-binding proteins like IHF and histones. Overall, we find it remarkable that CRISPR spacer acquisition appears to rely on similar modes of DNA binding and bending across the great evolutionary distance between Bacteria and Archaea (Fig. C.6).

Our results could indicate a conserved role for histones in archaeal CRISPR immunity since histones are widespread in archaeal lineages<sup>38-40</sup>. Notable exceptions are the Crenarchaeota, and the Diaforarchaea clades<sup>52</sup>, which lack histones. Intriguingly, histones are not present in the Crenarchaeon *S. solfataricus*, where factor-mediated spacer integration is also known to exist<sup>32</sup>, so polarized spacer integration in Crenarchaeon CRISPR systems may be ensured by yet another co-opted host chromosomal DNA binding protein.

Previous work aimed at understanding the determinants for histone binding to DNA showed that underlying sequences are the primary predictor for archaeal histone occupancy<sup>45,46</sup>. Important DNA features that recruit archaeal histones include helically phased AT and GC dinucleotides and certain GC-rich stretches (and the absence of certain AT-rich DNA stretches<sup>45,46</sup>). The

regions of *P. furiosus* CRISPR leaders with peaks of MNase-Seq read coverage contain these DNA features and are bordered by histone-repelling AT-rich sequences that comprise elements involved in transcription initiation (the BRE/TATA elements) and the CRISPR repeat (Fig. C.1, C.2 and Extended Data Fig. C.4). Histone binding at these same sites was also observed *in vitro* at the leader of a plasmid-encoded CRISPR locus (Extended Data Fig. C.8), confirming that histone binding is largely mediated by sequence features, rather than by other protein(s) or DNA topology. Such a sequence requirement suggests that, as with IHF sites in Proteobacteria, Archaeal CRISPR leaders co-evolved with host-associated chromosomal DNA binding proteins.

Like many archaea, *P. furiosus* and *T. kodakarensis* encode two highly similar histone isoforms, referred to as A and B (Extended Data Fig. C.6). Consistent with previous findings with *T. kodakarensis*<sup>48</sup>, we found that either *P. furiosus* A or B histone could be deleted but a double histone knockout (A + B) was not possible, indicating that at least one histone is necessary for viability. It is not clear why two (or more) related histones commonly occur in archaea<sup>44</sup>. Our *in vivo* data indicated that loss of histone A disrupted spacer uptake more than loss of histone B, potentially indicating a specialized role for histone A in integration (Fig. C.3). In contrast, our *in vitro* work, that both histone A or B (from either *P. furiosus* and *T. kodakarensis*) promoted spacer integration to the first repeat with approximately equal efficiencies (Fig. C.4). Together, the results show that both histone A and B, which likely bind DNA with similar properties<sup>45</sup>, are competent in conferring integration specificity to the Cas1-Cas2 integrase complex but leave open the possibility that additional layers of regulation make histone A more effective at directing new spacer uptake *in vivo*.



**Figure C.6. Hypothetical model for involvement of DNA binding proteins in spacer integration in bacterial and archaeal systems.** Cartoon diagrams, based on recent structural and functional data, showing a general model for how histones (**a**) and IHF (**c**) bind CRISPR leaders and guide integration to the leader-adjacent repeat. (**b**) Crystal structure of histone B from *Methanothermobacter fervidus* PDB 5T5K<sup>65</sup> (**d**) Crystal structure of IHF from *Escherichia coli* PDB 1OWG<sup>66</sup>. (**e**) Components of the integration reaction: DNA structure with positions of LR and L1 histone binding sites (yellow) relative to promoter elements (BRE and TATA in purple and red, respectively), as well as CRISPR repeat (blue) and spacer integration site at leader-repeat junction; Histone tetramer (dimers labelled in orange and green) bound to 60 bp of DNA; Cas1-2 integration complex with bound pre-spacer (purple). (**f**) Binding of histone tetramers to L1 (left), or LR (right) induces a strong bend in CRISPR leader sequence. (**g**) Binding of histone tetramers to both L1 and LR with Cas1-2 bound at the leader-repeat spacer integration site.

It is now clear that factor-mediated mechanisms (histones in Archaea and IHF in Bacteria) ensure that new spacer integration is localized to the leader-adjacent end of the repeat array. However, it is still unclear what evolutionary driving forces promote this localization. Although studies have shown that spacers throughout an array are competent for interference, spacers in the leader-adjacent position are often more highly expressed and give rise to more potent protection than leader-distal spacers<sup>5,51,53</sup>. Additionally, the numerous repeats in a CRISPR array may make central spacers less stable (subject to loss through recombination). Finally, the combinatorial requirement of leader-repeat sequence elements together with leader-bound protein factors may ensure specificity toward CRISPR arrays, thereby minimizing ectopic integration of foreign DNA fragments at repeat-like sequences elsewhere in the genome.

The fact that new spacers are consistently added to host CRISPR arrays has been leveraged into powerful technologies that harness CRISPR arrays as 'molecular recording devices'<sup>54-58</sup>. These technologies exploit the capacity of diverse Cas1-Cas2 integrases to continuously capture and directionally store DNA or RNA information over time in CRISPR loci of living cells. Ideally, sequencing and analysis of spacer location within the molecular recording arrays will allow researchers to derive time-ordered histories of biological processes, including transcription events and temporal gene expression pathways. CRISPR-based recorders may be optimized and/or their capacities expanded by considering the contribution of the protein factors like histones in new spacer uptake.

## Materials and Methods

### ***Micrococcal nuclease DNA digestion and high-throughput sequencing analysis.***

*Pyrococcus furiosus* COM1 strain<sup>59</sup> was grown in 250 mL cultures in defined media to either exponential or stationary phase, pelleted, and flash frozen. Pellets were resuspended in MNase buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl and 1 mM CaCl<sub>2</sub>) then lysed mechanically by freeze/thawing with liquid nitrogen while grinding with a mortar and pestle. Lysate was clarified by centrifugation to remove cellular debris. RNA was removed by digestion with RNase A (Qiagen) at a final concentration of 4 mg/mL and incubation at 37°C for one hour. 500 Units of MNase (ThermoFisher Scientific) was then added to the lysate, and 100 µL fractions were taken out at various timepoints. MNase reactions were stopped by the addition of phenol/chloroform/isoamyl alcohol (25:24:1) followed by vigorous mixing and centrifugation at 4°C for 5 minutes at 15,000 rpm. The aqueous layer was transferred into a 1.5 mL tube with 200 µL of 1 M Tris-HCl pH 8.0. DNA was ethanol precipitated, washed, and resuspended in 10 mM Tris-HCl pH 8.0.

MNase digested DNA was visualized by ethidium bromide staining after separation on a 4% Nusieve agarose gel. Bands corresponding to digested products 30, 60, 90, and 120 bp in length were excised from the gel and purified using the Zymo Gel DNA Recovery Kit (Zymo Research, Irvine, CA, USA). Purified DNA was then prepared for Illumina platform sequencing using the NEBNext Ultra II DNA library prep kit (New England Biolabs, Ipswich, MA, USA). Since MNase digested DNA was already fragmented, the fragmentation steps of the library prep protocol were excluded. Completed libraries were analyzed for size distribution and DNA concentration using TapeStation (Agilent, Santa Clara, CA, USA) and Qubit (ThermoFisher Scientific, Waltham, MA, USA) instruments, respectively, then pooled and sequenced on a NextSeq 550 (Illumina Inc, San Diego, CA, USA) using a 2 x 75 paired end protocol and high output kit.



Following sequencing, reads were demultiplexed by index, adapter trimmed, and aligned to a the *P. furiosus* reference genome and pT33.3 plasmid by bowtie2 version 2.5.0<sup>60</sup>. To visualize patterns of MNase-protected DNA, alignment outputs were processed using samtools version 1.9 and bedtools version 2.29.0 to generate genome coverage profiles (coverage BED files) for each sample<sup>61,62</sup> and then custom genome browser tracks were made with publicly available tools provided and described by the University of California Santa Cruz ( kent-tools, <https://genome.ucsc.edu>). To examine genome-wide trends in read depth (as in Extended Data Fig. 3), custom python scripts were written to take genome coverage depth BED files as input and determine the depth of 60 bp and 90 bp read coverage in sliding 60 bp windows across the genome (1 bp increments). This generated a set of densities which was then binned, tallied, and analyzed to determine 1) the distribution of 60 bp read densities, 2) the ratio of 60 bp to 90 bp read coverage in each sliding window, and 3) the distribution of 60 bp to 90 bp read coverage ratios. These ratios were output as matrix files which could be viewed in Microsoft Excel version 16.69.1.

**Conjugative transfer of pT33.3 plasmid.** Conjugative transfer of pT33.3 plasmid to *P. furiosus* strains were carried out as described<sup>47</sup>. Briefly, *Thermococcus* donor strains were cultured in complex media at 85°C and the *Pyrococcus* recipient strains were cultured in defined selective media at 95°C, all grown overnight for ~16 hours. Five mL of each fresh, overnight culture was centrifuged at ~3,800 x g for 20 minutes. Supernatant was discarded and all remaining liquid media removed. Using a wide-bore pipette tip, cells were resuspended in 50 µL of 1X Pfu Base Salts<sup>63</sup>. Recipient cells were mixed with the appropriate donor in a 1.5 mL tube, for a total of 100 µL of cell resuspension. The entire cell mixture was spotted in the center of selective media plates. Plates were grown anaerobically for three nights at 95°C. The spot of growth was

scraped using the flat side of a micropipette tip into 5 mL of defined selective liquid media and grown overnight at 95°C.

***P. furiosus strain construction.*** Strains with histone A or B gene deletions were generated by homologous recombination with PCR products<sup>64</sup>.

***Growth of P. furiosus cultures retaining both or one histone-encoding loci.*** Strains were inoculated 1:100 from overnight cultures and growth monitored by changes in optical density at 600 nm at 95°C. Anaerobic media was prepared and reduced as previously described<sup>65</sup>. Culture density was recorded hourly for ~32 hours. Average cultures densities with standard error of the mean are reported for at least three biological replicates of each strain. Uninoculated cultures were also monitored to provide a baseline since small changes in media absorbance occur during extended incubation at 95°C.

***Analyses of histone protein levels in total cell lysate from P. furiosus cultures retaining both or one histone-encoding loci by Western blotting.*** Biomass from anaerobic 95°C cultures was harvested via centrifugation at mid-exponential (E) and early-stationary (S) phase, respectively, resuspended and lysed in 25 mM Tris-HCl pH 8.0, 500 mM NaCl, 10% glycerol, 2% SDS, and total protein concentrations of each sample were quantified using a Qubit Protein Assay (Invitrogen). 3 µg of total protein from each strain harvested either from exponential or stationary phase cells was boiled and resolved through 4-20% Criterion™ TGX Stain-Free™ Protein gels alongside 500 ng of purified, recombinant HPfA and/or HPfB. Despite boiling samples in 2% SDS before loading the gels, the hyperthermophilic histones from *P. furiosus* retain dimer, tetramer, and larger multimeric complexes even when resolved in SDS-PAGE, resulting in a ladder of histone complexes revealed by Western blotting. Gels were transferred, probed with polyclonal antibodies raised against *T. kodakarensis* HTkA (that recognize HTkA, HTkB, HPfA, and HPfB), and developed as previously described<sup>66</sup>. Six independent biological

replicates (culturing, biomass preparations, and Western analyses of total histone levels from each strain) were done with similar results.

**RNA-seq analysis of wildtype and histone deletion strains.** For each of the three strains (wildtype,  $\Delta$ histone A,  $\Delta$ histone B), five separate 250 mL volumes of liquid culture grown to stationary phase were centrifuged to pellet cells. Cell pellets were frozen and later resuspended in 5 ml TRI Reagent® RT (Molecular Research Center), incubated at room temperature for 5 min at room temperature, and then placed on ice. 250  $\mu$ l of 4-bromoanisole (BAN; Molecular Research Center) was added, samples were mixed by inversion and then centrifuged at 12,000 X g for 15 min at 4°C. The resulting 3 ml of aqueous layer was removed and added to 4.5 ml of 100% isopropyl alcohol and moved to -20°C for at least 15 min prior to being centrifuged again for 30 min. Pelleted material was washed with 75% ethanol, resuspended in 450  $\mu$ l nuclease-free water (ThermoFisher) + 50  $\mu$ l DNase I Reaction Buffer (NEB) + 4  $\mu$ l DNase I (NEB), and incubated at 37°C for 30 min. 500  $\mu$ l of 4°C acid-phenol:chloroform IAA (ThermoFisher) was added, the sample was mixed by inversion and centrifuged at 16,000 X g for 10 min at 4°C. 450  $\mu$ l of aqueous layer was removed and added to 45  $\mu$ l 3M sodium acetate + 750  $\mu$ l 100% isopropyl alcohol and moved to -20°C for at least 15 min prior to being centrifuged at 16,000 X g again for 30 min. Pelleted material was washed with 75% ethanol, resuspended in 40  $\mu$ l nuclease-free water and isolated RNA was quantified using a NanoDrop™ spectrophotometer (ThermoFisher).

The NEBNext® rRNA Depletion Kit (NEB; E6310) was utilized to deplete *P. furiosus* rRNAs prior to making libraries for RNA sequencing. The “rRNA Depletion Solution” provided in the kit was replaced with a mixture of 82 custom oligonucleotides (IDT) with complementary sequences to *P. furiosus* rRNA sequences and containing an equimolar concentration of 1  $\mu$ M for each oligo. The protocol in the manual for the kit was followed using an input amount of 840 ng *P. furiosus* RNA for each sample with the only exception being that after the enzymatic steps

the RNA was purified by using a Monarch® RNA Cleanup Kit (NEB; T2030) rather than magnetic beads. rRNA-depleted RNA was eluted in 7 µl and then 5 µl of that volume was used to make libraries via using a NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina® (NEB; E7760) and the protocol section in the manual for intact rRNA-depleted RNA. The libraries were amplified via 10 cycles of PCR with NEBNext® Multiplex Oligos for Illumina® (NEB; E6440), quantified by analyzing on a 2100 Bioanalyzer (Agilent), mixed and paired-end sequenced 2 X 150 bp on an Illumina® NovaSeq 6000 instrument. Sequencing was conducted on two separate runs. For the first run, sequencing resulted in between 109,066 and 140,590 reads (average 124,032) for each library. For the second run, sequencing resulted in between 105,874 and 44,868,432 reads (average 29,067,816) for each library. For both sequencing runs, average insert size was between 142 and 159 bp.

Following sequencing, reads were demultiplexed by index, adapter trimmed, and aligned to the *P. furiosus* COM1 reference genome by bowtie2 in local mode, default settings<sup>60</sup>. To look for differences in transcript abundances between wildtype and  $\Delta$ histone A or  $\Delta$ histone B strains, we took the aligned RNA-seq data and determined per-gene read counts using bedtools intersect version 2.29.0<sup>62</sup> and the available gene annotations for our reference genome. A custom python script was used to parse output and generate a read count matrix which was then imported into DESeq2<sup>67</sup>. Samples were evaluated for differences in gene expression in two separate comparisons (wildtype versus  $\Delta$ histone A, wildtype versus  $\Delta$ histone B).

***In vivo spacer integration assay.*** After the conjugation of the pT33.3 plasmid into *P. furiosus* strains, a population of cells (~20 colonies) was added to 5 mL of defined, selective, liquid media and allowed to grow at 95°C overnight (~16 hours). One mL of overnight culture was then used for genomic DNA extraction using the Quick-DNA miniprep kit (Zymo Research). 10 ng of genomic DNA was used in each PCR reaction. To detect addition of a new leader-adjacent spacer, PCR primers were designed to amplify between the leader and the first known spacer in

the array. PCR products were then run on a 2.5% agarose gel to separate unexpanded bands (no integration of new spacers between the leader and the first known spacer) from expanded bands (integration of a new leader adjacent spacer-repeat unit) and the DNA was detected by ethidium bromide staining.

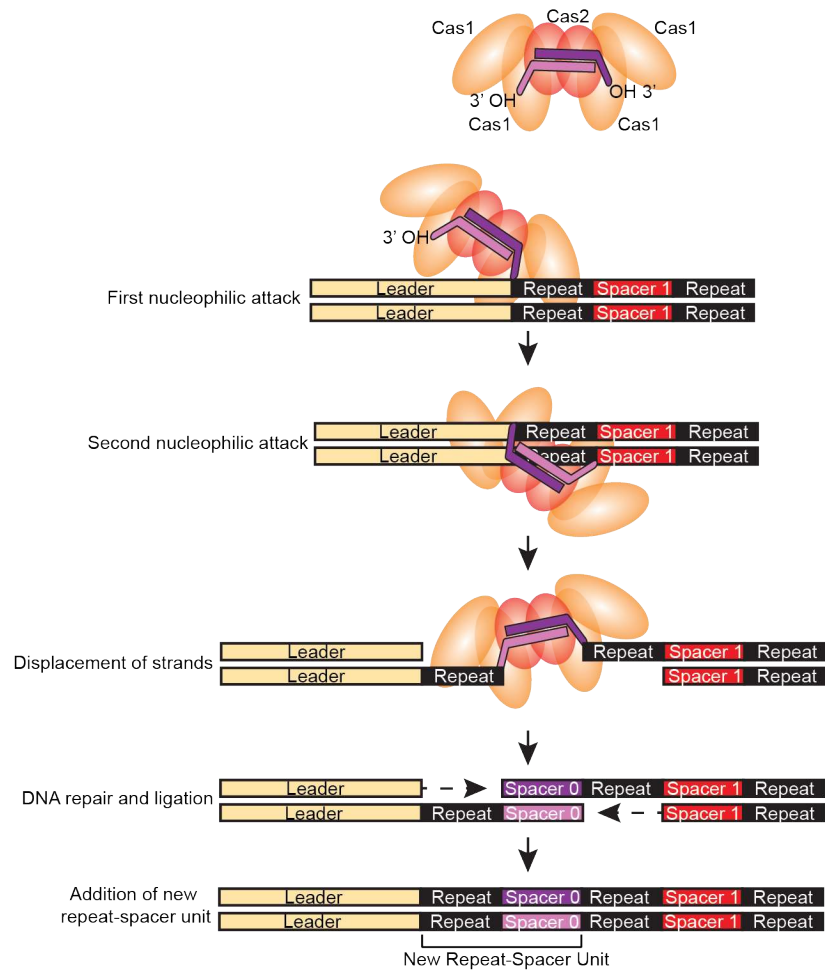
**Recombinant protein expression and purification.** *P. furiosus* Cas1, Cas2, Alba, Trmbl2, histone A and B genes were PCR amplified from genomic DNA, cloned into pET21d and transformed into *E. coli* BL21-RIPL expression strains. The C-terminal 6 x histidine tag proteins (Cas1, Cas2, Alba, Trmbl2) were expressed and purified by heat precipitation and nickel chromatography as described<sup>37</sup>. Histone A and B genes (from *P. furiosus* and *T. kodakarensis*) were expressed as native (untagged) proteins and purified by heat precipitation and heparin chromatography as described<sup>45</sup>.

**In vitro integration assay.** A full CRISPR leader and an array consisting of either three repeats and two spacers (pCRISPR-short) or 11 repeats and 10 spacers (pCRISPR-long) were added to Blunt II TOPO vectors via blunt end ligation. These CRISPR containing plasmids were used in *in vitro* integration assays to monitor the position of new spacer integration events. A final concentration of 1  $\mu$ M Cas1, 1  $\mu$ M Cas2, 1 mM DTT, 10 mM MgCl<sub>2</sub>, and 50 nM of spacer DNA (with 5 nt overhangs) was incubated in reaction buffer (20 mM Tris, 100 mM KCl, 5% glycerol, pH 7.5) for 1 hour at 4°C. Histone concentrations were selected to range from fully saturated (i.e., one histone dimer for every 30 bp of DNA in the reaction), to half saturation, to one tenth saturation, which was 3.4  $\mu$ M, 1.7  $\mu$ M, and 0.34  $\mu$ M respectively. A final concentration of 50 nM pCRISPR and varying concentrations as indicated of histones, Alba, or TrmBL2 (0-5.1  $\mu$ M) were added and incubated at 70°C for 1 hour. After incubation, the Cas1-Cas2 mixture was added and the reaction was incubated for one hour at 70°C. The integration reactions were then stopped by addition of 25 mM EDTA and proteins were degraded using 10  $\mu$ g proteinase K (Life

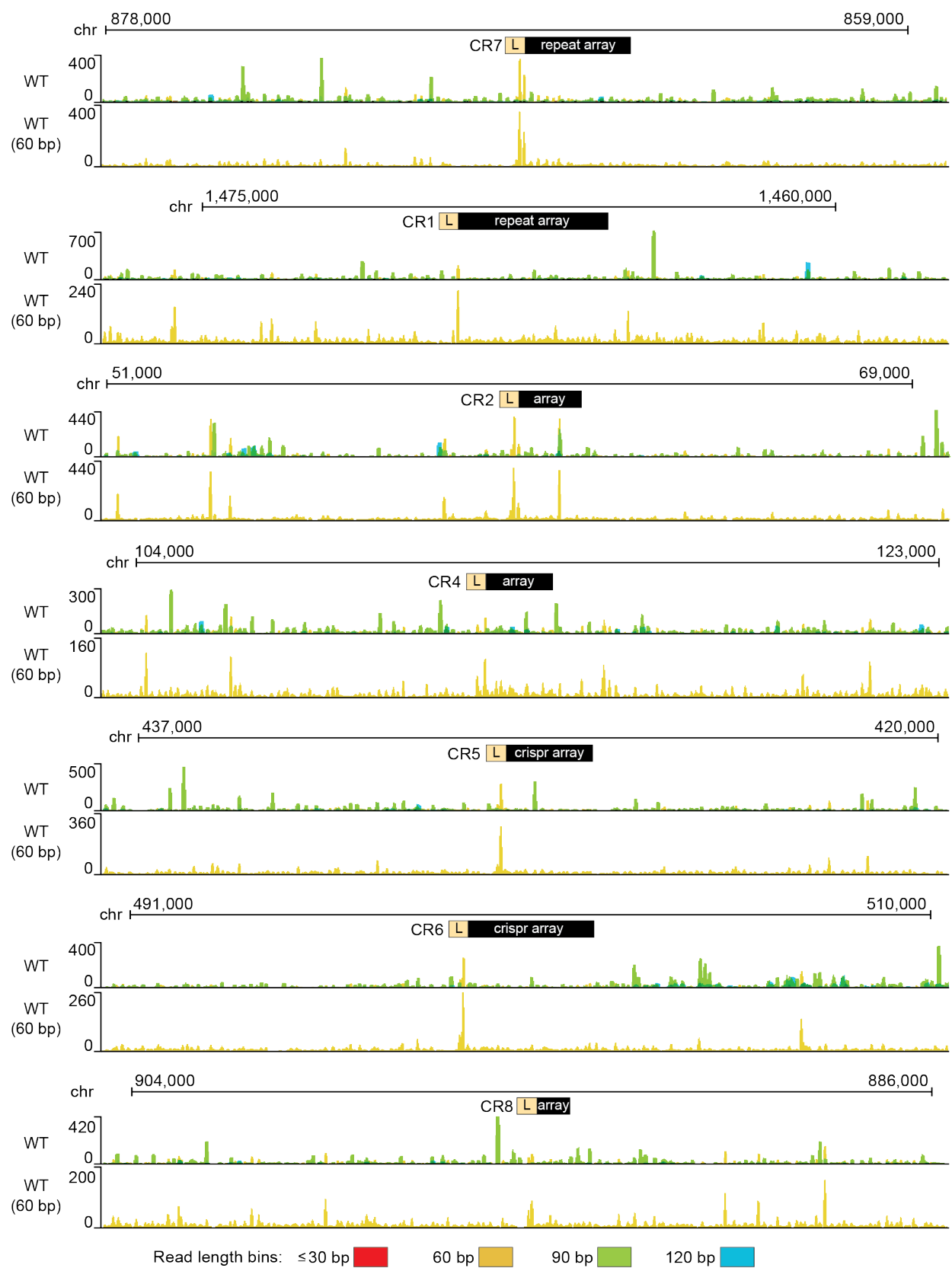
Technologies). DNA was purified from each reaction with the DNA Clean and Concentrator Kit (Zymo Research, Irvine, CA, USA), and eluted in a volume of 15  $\mu$ L of TE Buffer. Integration reaction products were then assayed by PCR: 1  $\mu$ L of the reaction product was added to a PCR with primers designed to amplify from the CRISPR leader to the incoming spacer, or from spacer 5 or 10 on the CRISPR array to the incoming spacer. PCR products were then run on a 4% Nusieve agarose gel at 120 V for 45-50 minutes for detected by ethidium bromide staining.

***ImageJ and high-throughput sequencing analysis of in vitro integration products.*** The average intensity of each band (corresponding to an integration event at a repeat) in each lane was acquired using ImageJ version 1.53 (NIH). The normalized intensity of those bands was then determined: for a given band, the average intensity of that band was divided by the total band intensities observed for the whole lane multiplied by 100. For high-throughput sequencing, *in vitro* integration products were treated, prepared for sequencing, and analyzed as previously described<sup>37</sup>. Briefly, the *in vitro* spacer integration assay was performed and DNA was isolated as described above. Excess un-integrated pre-spacer was removed using Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN). Illumina adapter sequence with an N10 random primer was annealed to the plasmid DNA and extended (thermocycler conditions: 98°C for 30s, 25°C for 30s, 35°C for 30s, 45°C for 30S, and 72°C for 5 min). Excess Illumina adapter was then removed using AMPure beads. A PCR was done to specifically amplify plasmid DNA that contained integrated pre-spacer; the forward primer targeted the pre-spacer, while reverse primers targeted the Illumina adapter introduced with the random anneal and extension step. Illumina barcodes and additional adapter sequences were added with a final PCR and the resulting library was separated on a 1% agarose gel to select for DNA in a 400 to 700 bp size range. DNA was isolated using the Zymo Gel DNA Recovery Kit (Zymo Research, Irvine CA) and sequenced on an Illumina MiSeq in a 100 by 50 paired end run.

## Extended Data

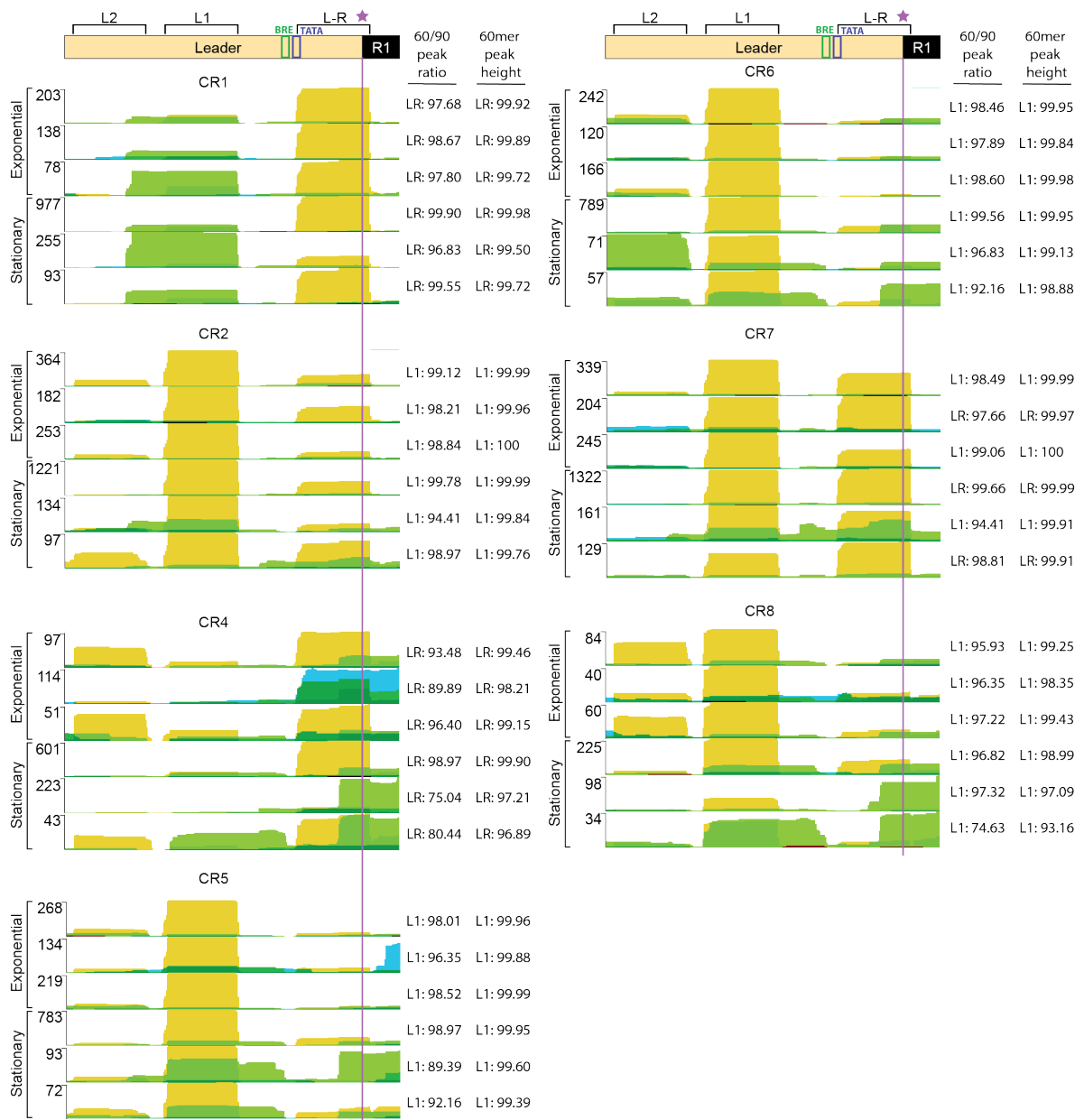


**Extended Data C.1. Steps of integration.** Cartoon diagram shows the steps by which a spacer is integrated into the CRISPR array at the leader-adjacent repeat.

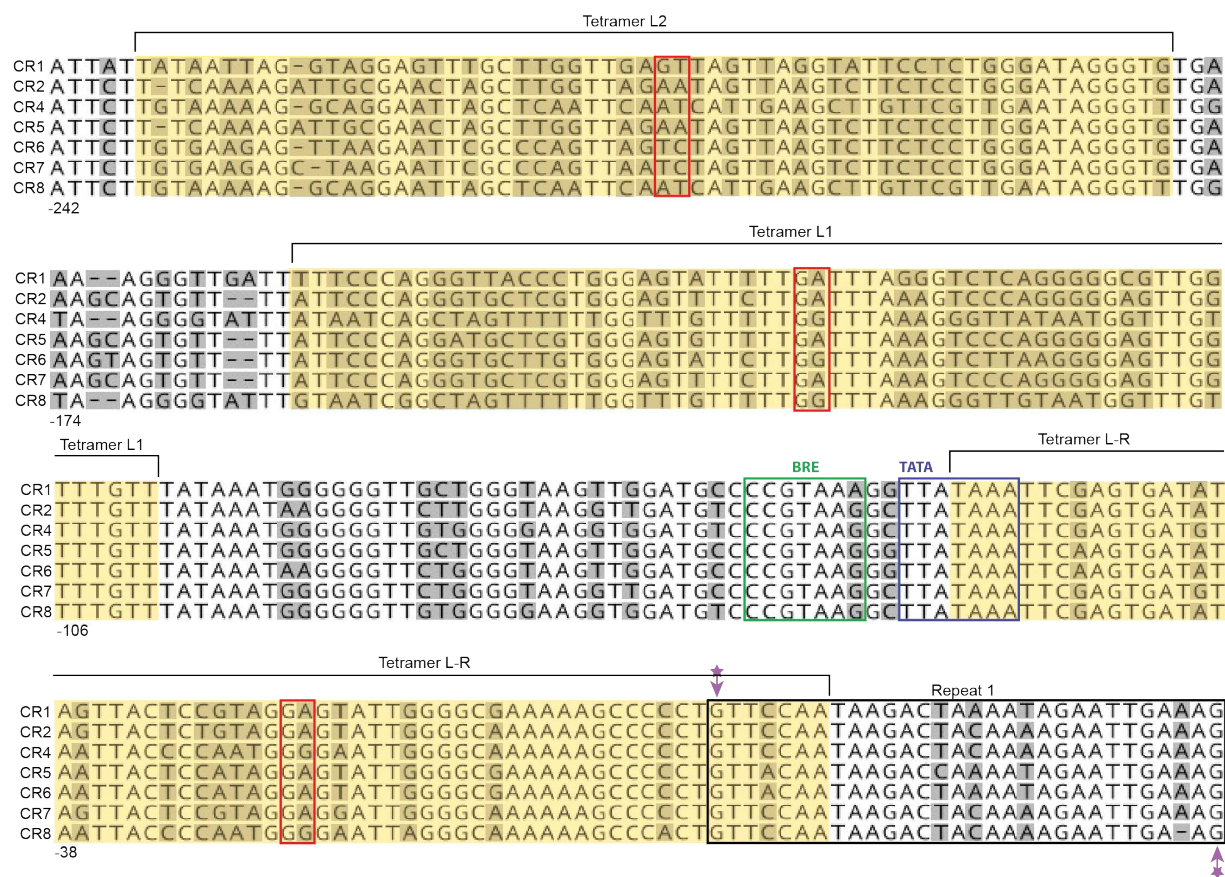




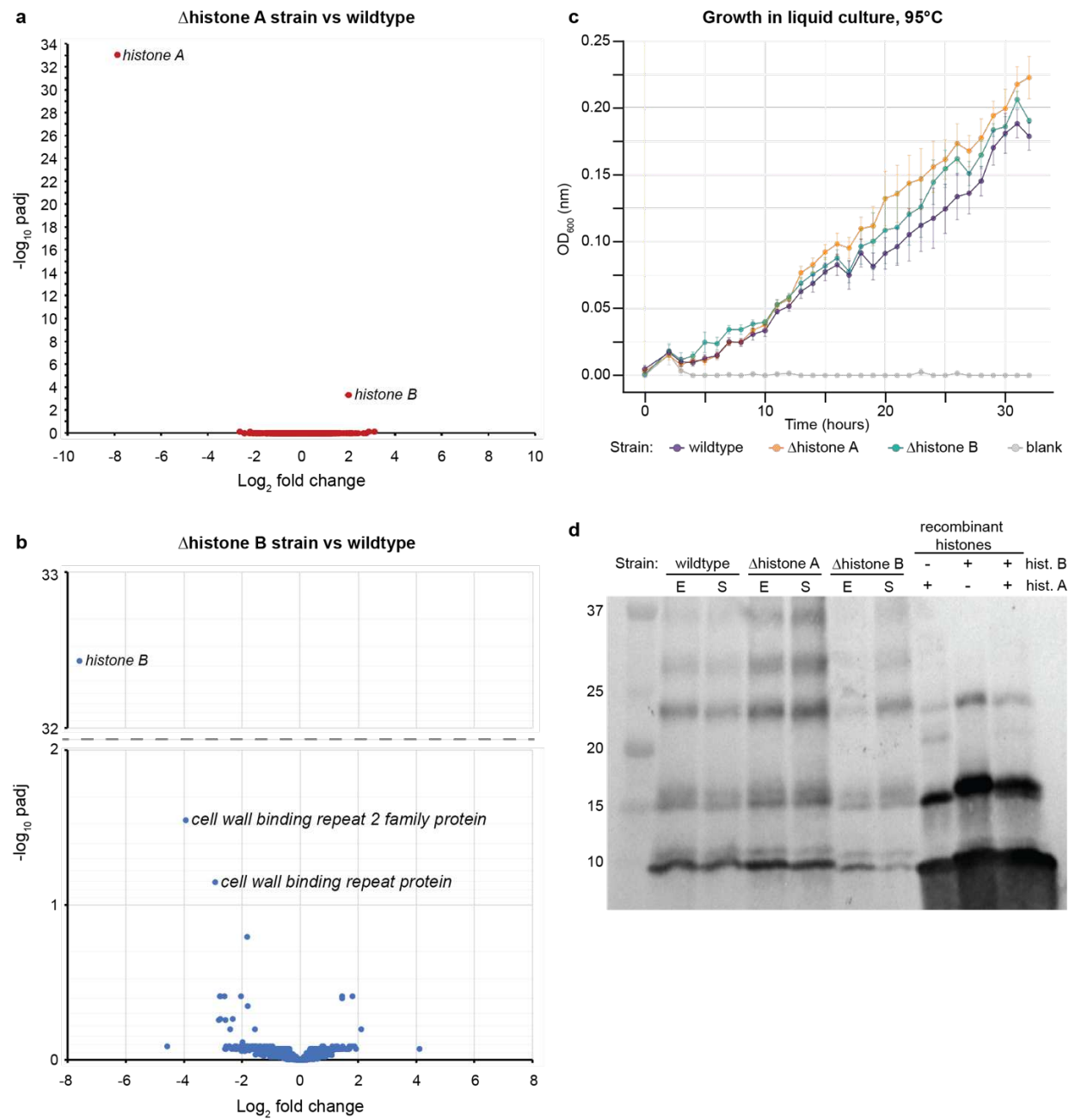
**Extended Data C.2. Distributions of aligned DNA sequences from MNase protection assay.** MNase protected DNA (Fig. C.1a) was sequenced by HTS, aligned, and genome browser tracks were generated to examine the genome-wide distribution of aligned DNA fragments. Overlapping subtracks are color-coded according to the size of the DNA fragment (red: 30 bp +/- 5; golden yellow: 60 bp +/- 5; green: 90 bp +/- 5; blue 120 bp +/- 5). Positions of CRISPR loci are indicated (leader in pale yellow, repeat array in black). Approximately 10 kb of genome space upstream and downstream from each CRISPR locus is shown. Six replicates (three stationary growth, three exponential) were sequenced and yielded similar patterns; a representative replicate is shown here.



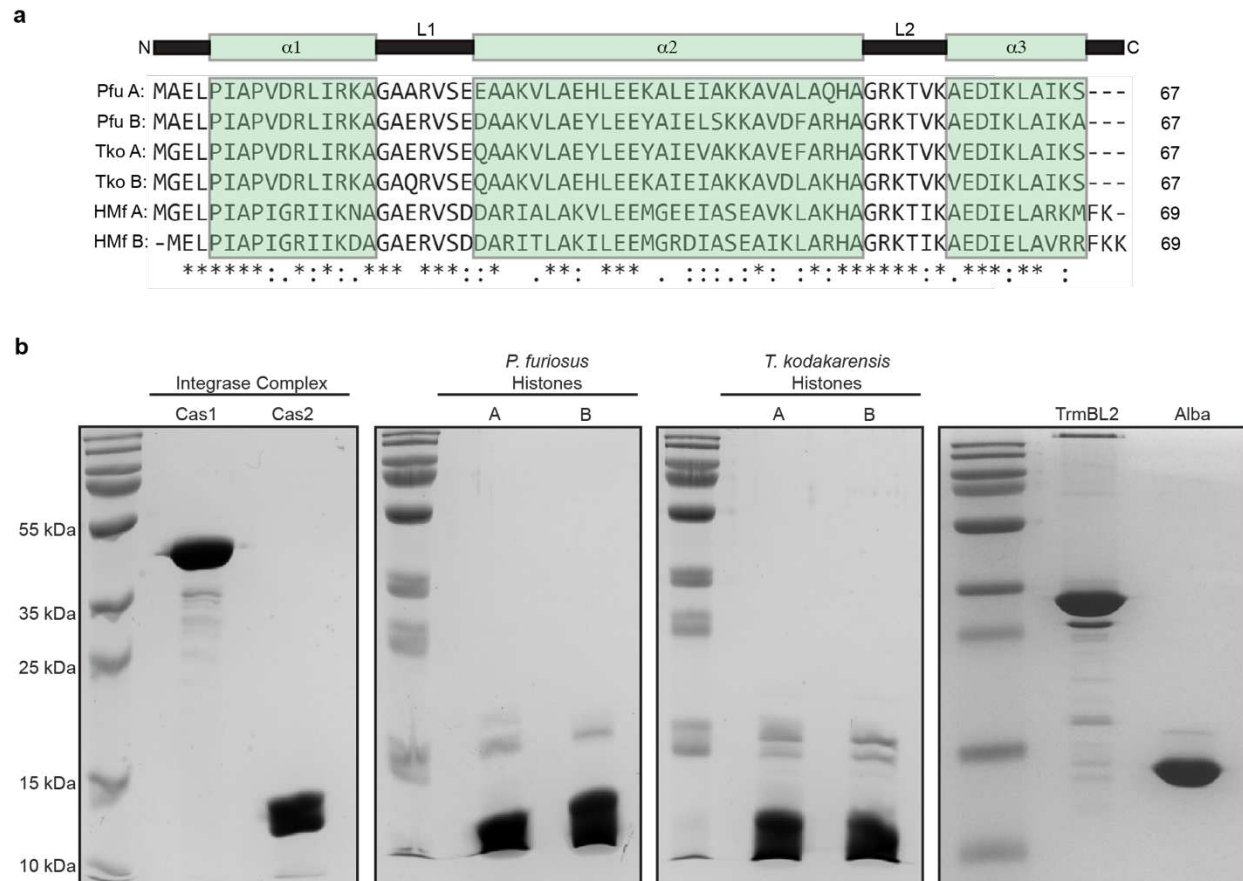
**Extended Data C.3. Positions of MNase protected DNA fragments and promoter elements within leaders of the seven CRISPR loci in *P. furiosus*.** Genome browser tracks were examined to determine the distribution of aligned MNase-protected DNA fragments within CRISPR leaders. Overlapping subtracks are color-coded according to the size of the DNA fragment (red: 30 bp +/- 5; golden yellow: 60 bp +/- 5; green: 90 bp +/- 5; blue 120 bp +/- 5). The x-axis indicates the cumulative depth of read coverage. Six replicates (three stationary growth (St), three exponential (Ex)) are shown. We noted that the 60 bp read coverage (peak height) over L1 and LR was unusually high compared to 60 bp coverage elsewhere in the genome. To quantify this, we analyzed sliding 60 bp windows across the genome, determined the density of 60 bp reads in those sliding windows, then binned and tallied the results to determine the distribution of 60 bp read densities. The values found under the label “60mer peak height” indicate the percentile of the L1 or LR peaks for each replicate; for example, the LR peak in CR1 exponential replicate 1 is higher than 99.92% of all 60 bp windows in the genome. We also noted that L1 and LR peaks usually had more 60mer coverage than 90mer coverage, even though 90mer reads were more abundant elsewhere in the genome. We used the sliding windows to quantify the 60mer to 90mer read density ratios across the genome. The values found under the label “60/90 peak ratio” indicate how biased towards 60mer read coverage the L1 or LR peak is; for example, the LR peak in CR1 exponential replicate 1 is more biased towards 60mer read coverage than 97.68% of all 60 bp windows in the genome.



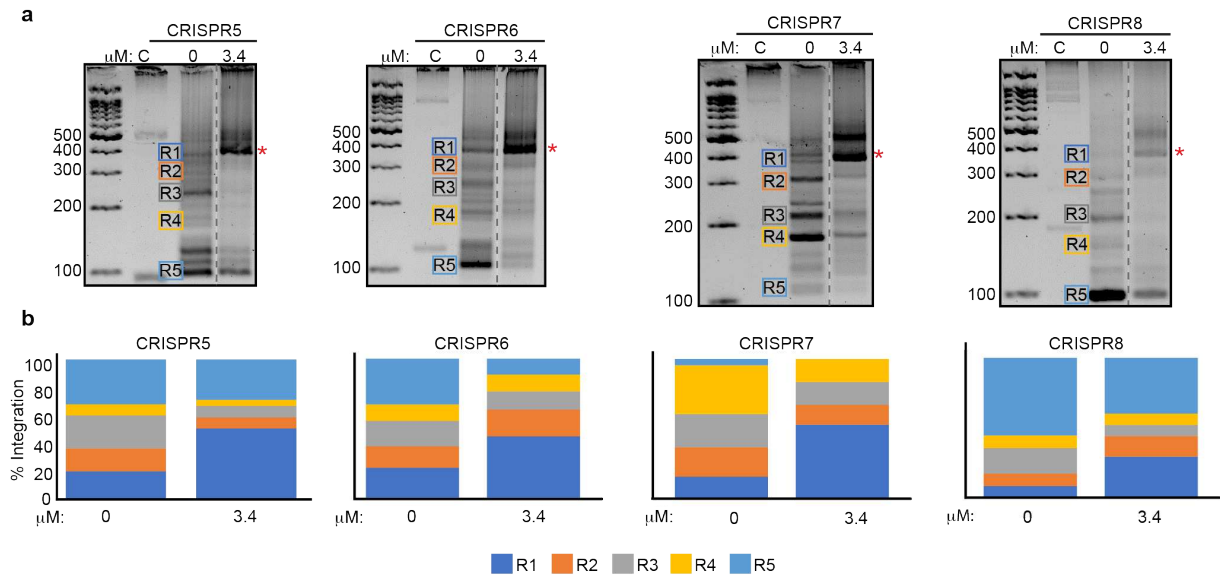
**Extended Data C.4. Nucleotide alignment of seven CRISPR array leaders with MNase protected DNA peaks, promoter elements, and integration sites mapped.** Nucleotide sequences of seven leaders were aligned; conserved nucleotide regions are highlighted in gray. Peaks of MNase protected DNA are highlighted in yellow; the purple star and line at the leader-repeat junction show the position of typical, *in vivo* spacer integration. Promoter elements in the leader are labeled (BRE and TATA).



**Extended Data C.5. Growth, RNA expression, and histone abundance characteristics in wildtype,  $\Delta$ histone A, and  $\Delta$ histone B *P. furiosus* strains used in this study.** Total, ribo-depleted RNA was sequenced for wildtype,  $\Delta$ histone A, and  $\Delta$ histone B strains (five replicates each) and analyzed using the DESeq2 package to identify annotated transcripts with differential expression between (a) wildtype and  $\Delta$ histone A and (b) wildtype and  $\Delta$ histone B. A difference in transcript abundance was considered significant if the log 2-fold difference was equal or greater than 0.58 and the adjusted p-value was equal or less than 0.1. (c) Wildtype,  $\Delta$ histone A, and  $\Delta$ histone B strains were grown in liquid medium supplemented with pyruvate and Na<sub>2</sub>S at 95°C and optical density (OD<sub>600nm</sub>) was measured every hour as a proxy for cell growth. Blank sample contained medium but no inoculum. Error bars show standard error of the mean for three biological replicates. (d) Western blots, employing polyclonal antibodies raised against HTkA (that recognize HTkA, HTkB, HPfA, and HPfB), were done on total protein from each strain harvested either from exponential (E, 10 hour time point) or stationary (S, 30 hour time point) phase cells. Recombinantly purified HPfA and HPfB were resolved on identical gels to provide size references. Six biological replicates were run with similar results; a representative blot image is shown here. Although samples were boiled in 2% SDS before gel loading, histones retained dimer, tetramer, and larger multimeric complexes, giving rise to the multi-band ladder appearance visible on the gel.

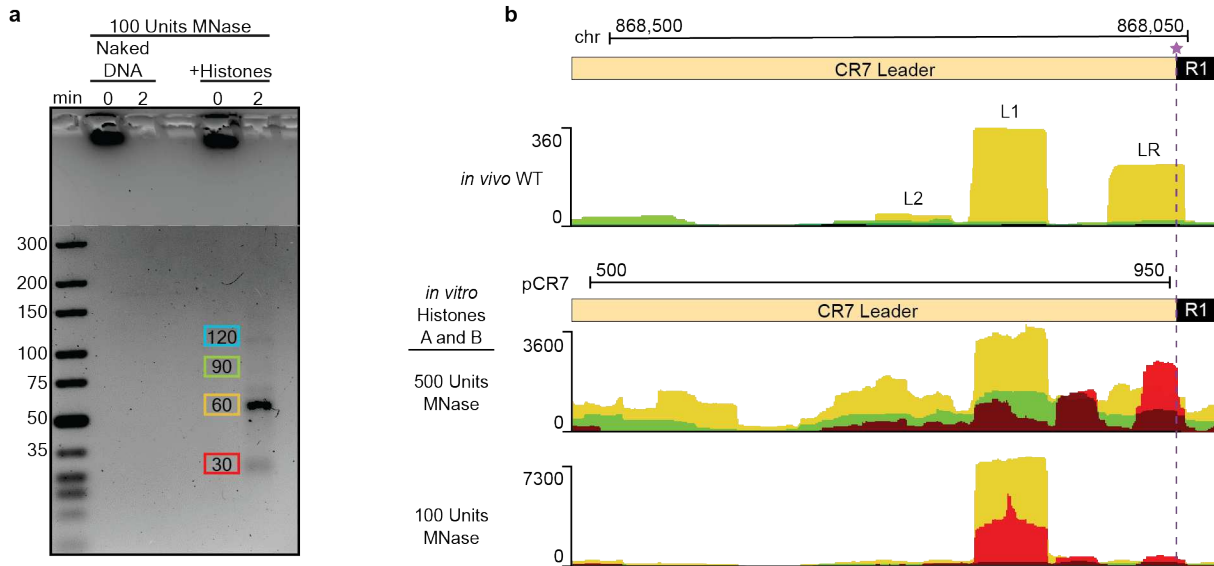


**Extended Data C.6. Conservation of histones in Euryarchaeota and purification of recombinant proteins.** (a) Clustal Omega Multiple Sequence Alignment of histones from *P. furiosus* (Pfu A and Pfu B), *Thermococcus kodakarensis* (Tko A and Tko B), and *Methanothermobacter fervidus* (HMf A and HMf B). \* = completely conserved residues, : = conservation between groups with strongly similar properties, . = conservation between groups with weakly similar properties. (b) SDS PAGE gels with the purified proteins used in the *in vitro* integration assays. Cas1=37.5 kDa, Cas2=10 kDa, *P. furiosus* histones A=7.4 kDa and B=7.3 kDa, *T. kodakarensis* histones A=7.3 kDa and B=7.1 kDa, TrmBL2=30.6 kDa.

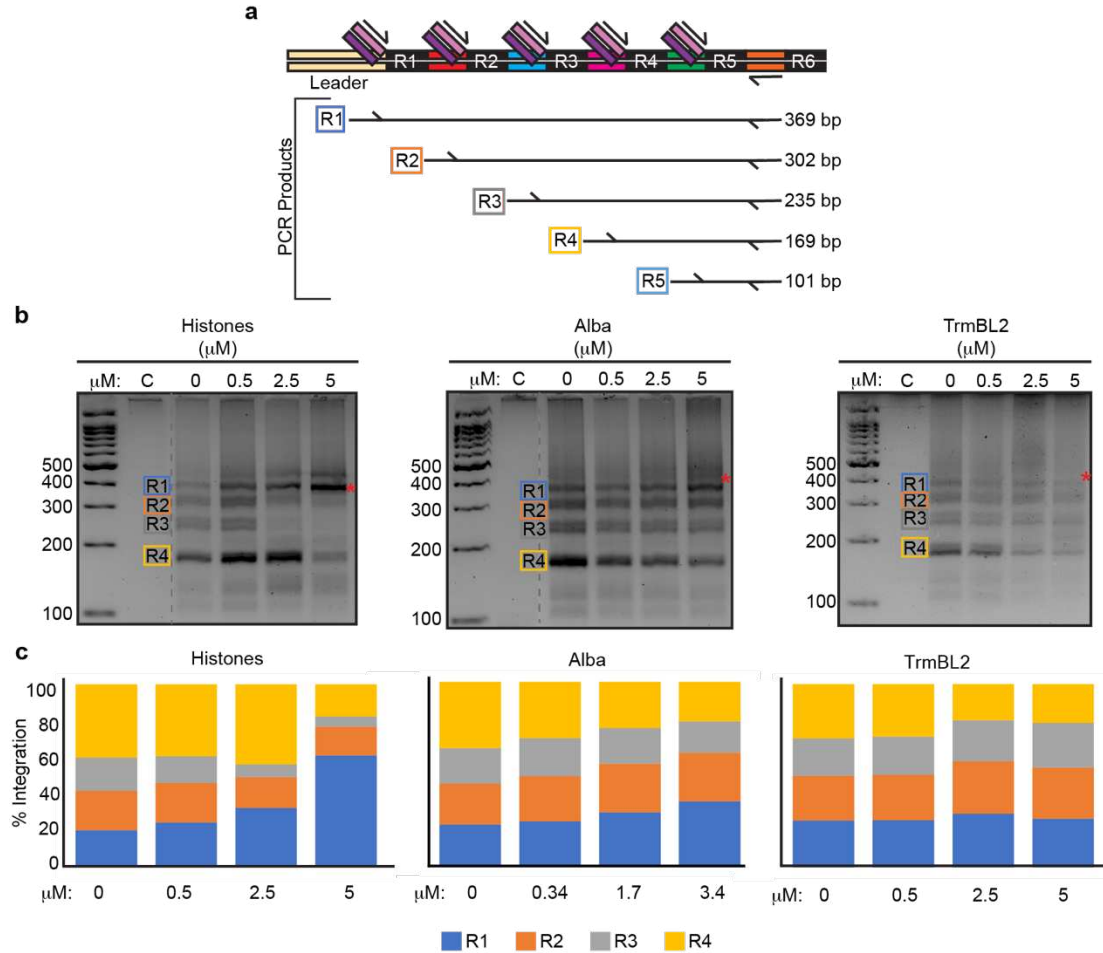


**Extended Data C.7. *In vitro* evaluation of spacer integration into four CRISPR arrays in the presence or absence of *P. furiosus* histones.** (a) Gel images show representative results from PCR (carried out as in Fig. C.5d) with primers targeting four CRISPR arrays: CRISPR5, CRISPR6, CRISPR7, and CRISPR8. The expected sizes for PCR products resulting from integrations at repeats 1 - 5 are marked with R1 – R5, respectively. Red asterisks identify the band corresponding to integrations at repeat 1, which is the natural, preferred point of integration *in vivo*. (b) Intensity of all PCR bands was quantified using ImageJ and the proportion of integration events at the five repeats was determined.

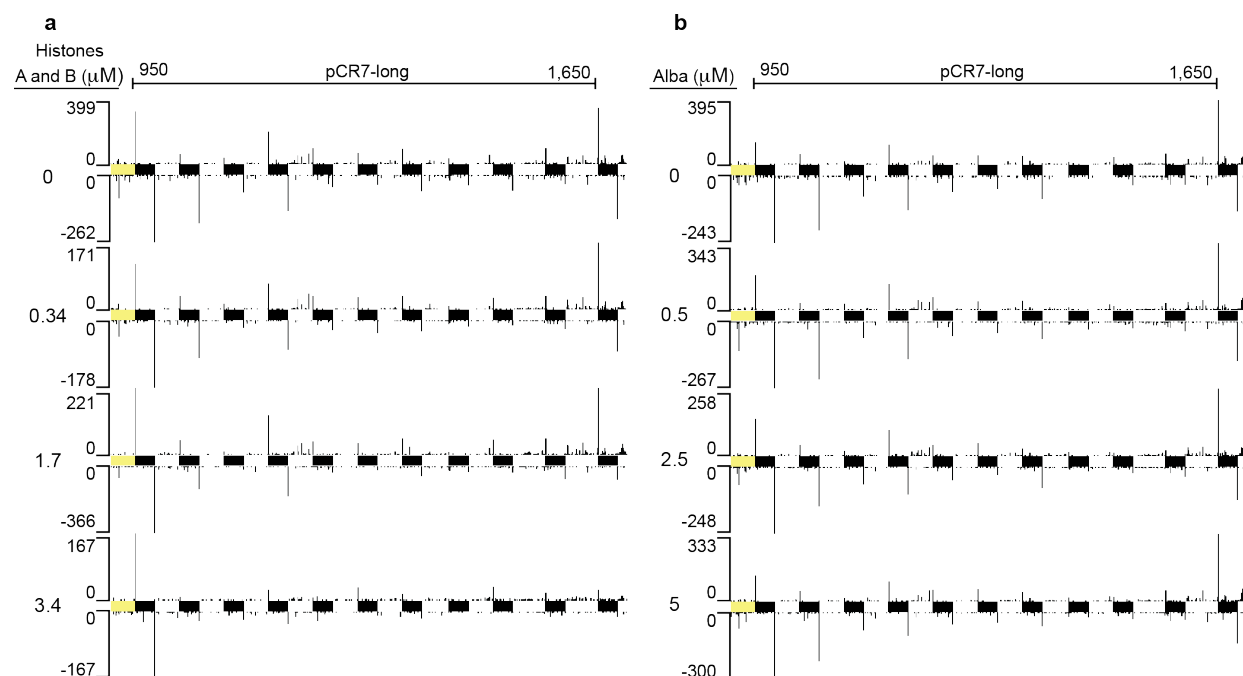




**Extended Data C.8. MNase protection assay to characterize binding patterns for *P. furiosus* histones incubated with the pCR7-long plasmid. (a)** Gel image of DNA fragments generated by micrococcal nuclease digestion of pCR7-long plasmid incubated with purified recombinant *P. furiosus* histones A and B. Colored boxes highlight major bands of protected DNA. The gel image is for a digest containing 100 units of MNase; a 500 unit digestion was also done and resulting DNA bands were much fainter. **(b)** DNA bands highlighted in A were sequenced and genome browser tracks were generated. Overlapping subtracks are color-coded according to the size of the DNA fragment (red: 30 bp +/- 5; golden yellow: 60 bp +/- 5; green: 90 bp +/- 5; blue 120 bp +/- 5). The full CRISPR7 and part of one repeat are shown, with the purple line indicating the position where integration would typically occur *in vivo*. The x-axis indicates the cumulative depth of read coverage.



**Extended Data C.9. *In vitro* evaluation of spacer integration into pCR7-long in the presence or absence of DNA-binding proteins.** (a) Diagrammatic representation of the PCR used to assess *in vitro* integration into pCR7-long. (b) Gel images show representative results from the PCR in (a) when recombinant purified histones, Alba, or TrmBL2 were added to the reaction at the indicated concentrations. The expected sizes for PCR products resulting from integrations at repeats 1 - 4 are marked with R1 - R4, respectively. Red asterisks identify the band corresponding to integrations at repeat 1, which is the natural, preferred point of integration *in vivo*. (c) Intensity of all PCR bands was quantified using ImageJ and the proportion of integration events at the five repeats was determined.



**Extended Data C.10. Unbiased HTS sequencing of *in vitro* integration of a spacer into pCR7-long.** The *in vitro* integration assay with and without histones (a) or alba (b) was carried out, as before, and integration events were then sequenced using a two-step, semi-degenerate PCR protocol which targets all DNA fragments bearing the spacer. Sequencing results were used to generate genome browser tracks; the amplitude of the black peaks in these tracks indicate the total number of reads supporting an integration event at that nucleotide position. The positions of the leader (pale yellow) and repeats (black) are shown.

## REFERENCES

- 1 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712 (2007).
- 2 Alkhnbashi, O. S. *et al.* Characterizing leader sequences of CRISPR loci. *Bioinformatics* **32**, i576-i585 (2016).
- 3 Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic acids research* **43**, 1749-1758 (2015).
- 4 Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic acids research* **40**, 5569-5576 (2012).
- 5 Hale, C. R. *et al.* Essential Features and Rational Design of CRISPR RNAs that Function with the Cas RAMP Module Complex to Cleave RNAs. *Molecular cell* **45**, 292-302 (2012).
- 6 Carte, J., Wang, R. Y., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes & development* **22**, 3489-3496 (2008).
- 7 Makarova, K. S. *et al.* Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nature reviews. Microbiology* (2019).
- 8 Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964 (2008).
- 9 Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945-956 (2009).
- 10 Nussenzweig, P. M. & Marraffini, L. A. Molecular Mechanisms of CRISPR-Cas Immunity in Bacteria. *Annual review of genetics* **54**, 93-120 (2020).
- 11 McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nature reviews. Microbiology* **17**, 7-12 (2019).
- 12 Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* **550**, 137-141 (2017).
- 13 Nunez, J. K. *et al.* Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nature structural & molecular biology* **21**, 528-534 (2014).
- 14 Shiimori, M., Garrett, S. C., Graveley, B. R. & Terns, M. P. Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Molecular cell* **70**, 814-824 e816 (2018).
- 15 Hu, C. *et al.* Mechanism for Cas4-assisted directional spacer acquisition in CRISPR-Cas. *Nature* **598**, 515-520 (2021).
- 16 Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199-202 (2015).
- 17 Wei, Y. Z., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes & development* **29**, 356-361 (2015).
- 18 Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505-510 (2015).
- 19 Jakhanwal, S. *et al.* A CRISPR-Cas9-integrase complex generates precise DNA fragments for genome integration. *Nucleic acids research* **49**, 3546-3556 (2021).
- 20 Ivancic-Bace, I., Cass, S. D., Wearne, S. J. & Bolt, E. L. Different genome stability proteins underpin primed and naive adaptation in *E. coli* CRISPR-Cas immunity. *Nucleic acids research* **43**, 10821-10830 (2015).
- 21 Nunez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193-198 (2015).

- 22 Rollie, C., Schneider, S., Brinkmann, A. S., Bolt, E. L. & White, M. F. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *eLife* **4** (2015).
- 23 Barrangou, R. & Dudley, E. G. CRISPR-Based Typing and Next-Generation Tracking Technologies. *Annual review of food science and technology* **7**, 395-411 (2016).
- 24 Kim, J. G., Garrett, S., Wei, Y., Graveley, B. R. & Terns, M. P. CRISPR DNA elements controlling site-specific spacer integration and proper repeat length by a Type II CRISPR-Cas system. *Nucleic acids research* **47**, 8632-8648 (2019).
- 25 Wright, A. V. & Doudna, J. A. Protecting genome integrity during CRISPR immune adaptation. *Nature structural & molecular biology* **23**, 876-883 (2016).
- 26 Santiago-Frangos, A., Buyukyoruk, M., Wiegand, T., Krishna, P. & Wiedenheft, B. Distribution and phasing of sequence motifs that facilitate CRISPR adaptation. *Current biology : CB* **31**, 3515-3524 e3516 (2021).
- 27 Yoganand, K. N., Sivathanu, R., Nimkar, S. & Anand, B. Asymmetric positioning of Cas1-2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. *Nucleic acids research* **45**, 367-381 (2017).
- 28 Fagerlund, R. D. *et al.* Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E5122-E5128 (2017).
- 29 Nunez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Molecular cell* **62**, 824-833 (2016).
- 30 Wright, A. V. *et al.* Structures of the CRISPR genome integration complex. *Science* **357**, 1113-1118 (2017).
- 31 Rice, P. A., Yang, S., Mizuuchi, K. & Nash, H. A. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* **87**, 1295-1306 (1996).
- 32 Rollie, C., Graham, S., Rouillon, C. & White, M. F. Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic acids research* **46**, 1007-1020 (2018).
- 33 Garrett, S. *et al.* Primed CRISPR DNA uptake in *Pyrococcus furiosus*. *Nucleic acids research* **48**, 6120-6135 (2020).
- 34 Shiimori, M. *et al.* Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in *Pyrococcus furiosus*. *Nucleic acids research* (2017).
- 35 Terns, R. M. & Terns, M. P. The RNA- and DNA-targeting CRISPR-Cas immune systems of *Pyrococcus furiosus*. *Biochemical Society transactions* **41**, 1416-1421 (2013).
- 36 Tang, D. *et al.* A distinct structure of Cas1-Cas2 complex provides insights into the mechanism for the longer spacer acquisition in *Pyrococcus furiosus*. *Int J Biol Macromol* **183**, 379-386 (2021).
- 37 Grainy, J., Garrett, S., Graveley, B. R. & M, P. T. CRISPR repeat sequences and relative spacing specify DNA integration by *Pyrococcus furiosus* Cas1 and Cas2. *Nucleic acids research* (2019).
- 38 Laursen, S. P., Bowerman, S. & Luger, K. Archaea: The Final Frontier of Chromatin. *Journal of molecular biology* **433**, 166791 (2021).
- 39 Henneman, B., van Emmerik, C., van Ingen, H. & Dame, R. T. Structure and function of archaeal histones. *PLoS genetics* **14**, e1007582 (2018).
- 40 Hoher, A. *et al.* Growth temperature and chromatinization in archaea. *Nature microbiology* (2022).
- 41 Mattioli, F. *et al.* Structure of histone-based chromatin in Archaea. *Science* **357**, 609-612 (2017).
- 42 Maruyama, H. *et al.* An alternative beads-on-a-string chromatin architecture in *Thermococcus kodakarensis*. *EMBO reports* **14**, 711-717 (2013).

- 43 Sandman, K., Krzycki, J. A., Dobrinski, B., Lurz, R. & Reeve, J. N. HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon *Methanothermus fervidus*, is most closely related to histones. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 5788-5791 (1990).
- 44 Stevens, K. M., Hocher, A. & Warnecke, T. Deep Conservation of Histone Variants in Thermococcales Archaea. *Genome biology and evolution* **14** (2022).
- 45 Nalabothula, N. *et al.* Archaeal nucleosome positioning in vivo and in vitro is directed by primary sequence motifs. *BMC genomics* **14**, 391 (2013).
- 46 Bailey, K. A., Pereira, S. L., Widom, J. & Reeve, J. N. Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution. *Journal of molecular biology* **303**, 25-34 (2000).
- 47 Watts, E. A. *et al.* Hyper-stimulation of *Pyrococcus furiosus* CRISPR DNA uptake by a self-transmissible plasmid. *Extremophiles : life under extreme conditions* (2022).
- 48 Cubonovaa, L. *et al.* An archaeal histone is required for transformation of *Thermococcus kodakarensis*. *Journal of bacteriology* **194**, 6864-6874 (2012).
- 49 Efremov, A. K. *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**, 15770-15784 (2015).
- 50 Crnigoj, M. *et al.* Interactions of archaeal chromatin proteins Alba1 and Alba2 with nucleic acids. *PloS one* **8**, e58237 (2013).
- 51 Garrett, S. C. Pruning and Tending Immune Memories: Spacer Dynamics in the CRISPR Array. *Frontiers in microbiology* **12**, 664299 (2021).
- 52 Adam, P. S., Borrel, G., Brochier-Armanet, C. & Gribaldo, S. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *The ISME journal* **11**, 2407-2425 (2017).
- 53 McGinn, J. & Marraffini, L. A. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Molecular cell* **64**, 616-623 (2016).
- 54 Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345-349 (2017).
- 55 Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).
- 56 Bhattarai-Kline, S. *et al.* Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature* **608**, 217-225 (2022).
- 57 Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380-385 (2018).
- 58 Schmidt, F. *et al.* Noninvasive assessment of gut function using transcriptional recording sentinel cells. *Science* **376**, eabm6038 (2022).
- 59 Bridger, S. L., Lancaster, W. A., Poole, F. L., 2nd, Schut, G. J. & Adams, M. W. Genome sequencing of a genetically tractable *Pyrococcus furiosus* strain reveals a highly dynamic genome. *Journal of bacteriology* **194**, 4097-4106 (2012).
- 60 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012). <https://doi.org/10.1038/nmeth.1923>
- 61 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
- 62 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
- 63 Adams, M. W. *et al.* Key role for sulfur in peptide metabolism and in regulation of three hydrogenases in the hyperthermophilic archaeon *Pyrococcus furiosus*. *Journal of bacteriology* **183**, 716-724 (2001).

- 64 Elmore, J., Deighan, T., Westpheling, J., Terns, R. M. & Terns, M. P. DNA targeting by the type I-G and type I-A CRISPR-Cas systems of *Pyrococcus furiosus*. *Nucleic acids research* **43**, 10353-10363 (2015).
- 65 Liman, G. L. S., Stettler, M. E. & Santangelo, T. J. Transformation Techniques for the Anaerobic Hyperthermophile *Thermococcus kodakarensis*. *Methods in molecular biology* **2522**, 87-104 (2022).
- 66 Sanders, T. J. *et al.* TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Molecular microbiology* **111**, 784-797 (2019).
- 67 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
- 68 Bhattacharyya, S., Mattioli, F. & Luger, K. Archaeal DNA on the histone merry-go-round. *The FEBS journal* **285**, 3168-3174 (2018).
- 69 Lynch, T. W., Read, E. K., Mattis, A. N., Gardner, J. F. & Rice, P. A. Integration host factor: putting a twist on protein-DNA recognition. *Journal of molecular biology* **330**, 493-502 (2003).