DISSERTATION

DNA REPLICATION IN THE ENVIRONMENTAL EXTREMES

Submitted by

Geraldy Lie Stefanus Liman

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

Summer 2024

Doctoral Committee:

Advisor: Thomas J. Santangelo

Steven Markus
Grant Schauer
Daniel Sloan

ABSTRACT

DNA REPLICATION IN THE ENVIRONMENTAL EXTREMES

DNA replication is an essential biological process across all life on Earth. For the prokaryotic Archaea domain, which contains organisms that can thrive in inhospitable environments like hydrothermal vents or salt deposits in the Dead Sea, the cell machinery for these conserved processes have acclimated over the course of evolution to encourage survival. While the origin of replication (*ori*), a predetermined position within the genome where DNA replication starts, is conserved in all Domains, its significance is not equal between them. Surprisingly, the model hyperthermophilic archaeon, *T. kodakarensis*, replicates its genome without relying on origin-dependent replication (ODR), and instead, relies mostly on recombination-dependent replication (RDR). In fact, the *ori* in *T. kodakarensis* is dispensable from the organism without much phenotypic consequence. Although dispensable, *ori* persists after millions of years of evolution in this organism, suggesting some functional significance under certain conditions. Not to mention, archaeal replisomes are comprised of unique components that are distinct from the other two domains of life, though surprisingly more similar to those found in Eukarya. Central to all replisomes is the activity of the DNA polymerase (DNAP). Most archaeal organisms, except for the Creanarchaea, encode two main replicative DNAPs, the eukaryotic-like B-family DNAP (PolB) and the archaeal-specific D-family DNAP (PolD). In *T. kodakarensis,* PolD is the essential replicative DNAP while PolB is dispensable. This thesis aims to (1) characterize the activity and regulation of RadA, the main recombinase in Archaea, (2) characterize the exaptation of inteins to regulate DNA replication, (3) delineate the *in vivo* function(s) of PolB. Furthermore, I hope to further characterize DNA replication in the context of evolutionary biology and how it relates to the three Domains of life.

ACKNOWLEDGMENTS

Nobody thought that I would graduate, but here I am writing the last document of my journey as a PhD candidate and a student. These last six years have been the most tumultuous in my almost 30 years of existence. I express my deepest gratitude to all members of the Santangelo lab and everyone who helped me during this chapter of my life.

Tom – Thank you for giving me a chance to not only develop as a scientist but also as an individual. Your guidance and constant push helped me maintain focus on my aims. You are like a second father to me (don't let it go to your head). It was a great adventure!

Aaron, Jaylin, Sarah, Lina, Marina, and Maddie - Thank you for dealing with me as your mentor. I could not ask more from my undergraduate research assistants. I may not show it directly, but I am proud of every single one of you. You will do great in the future!

Mom and Dad - Thank you for supporting me through my journey. Allowing me to move halfway across the world away from home must be hard. I might not always answer your calls or messages but both of you are always in my thoughts.

Oma (grandma) and Opa (grandpa) - I have never seen both of you cry before except before I left for the US. I did not understand it before but now I do. It was the last time I will ever hug and see the both of you in person, at least in this life. Both of you are always in my heart. I was heartbroken after learning about both of your passing, three days before my preliminary exam for Oma and last year for Opa. You are my strength and motivation, thank you.

Gentry and Sesame - Thank you for dealing with me during the most hectic part of my life. I am excited for our next chapter; it might be a doozy too.

All of my collaborators - I would like to extend my gratitude to everyone from the DNA repair / replication consortium (NEB / NIST), the archaeal lipid consortium (Stanford University), and the archaeal imaging consortium (Janelia Research Campus) for all of their expertise that helped me with my research.

TABLE OF CONTENTS

CHAPTER 1: LIFE AS WE KNOW THROUGH THE LENSE OF DNA REPLICATION

Introduction

The central dogma of molecular biology is regarded as the transmutation of genetic material into enzymatic effectors, but for this to occur, the DNA code must be present. When one considers this notion, it is easy to see that DNA replication is an underlying backbone to the dynamics employed within this dogma. For each Domain, the capacity to copy the genetic code is vital. The parental cell must be able to replicate its DNA and pass this information to the resulting daughter cells for the continuity of biological lineages. This notion is intrinsically true for basic cell proliferation or for zygote progression to multicellular status, so we see a process that is universal regardless of organismal complexity.

Unsurprisingly, stages of DNA replication are highly conserved, which owes to the integrity and efficiency of evolution in shaping this critical cell function. All DNA replications can be organized into three distinct steps: initiation, DNA synthesis, and termination[1]. However, there are variations in the machinery involved between Bacteria, Eukarya, and Archaea (Table 1.1)[2], so it is important to understand these differences to better characterize evolutionary descent and what these differences mean as adaptive forces for organism survival.

Where, When, and How: DNA replication initiation

Where, when, and how. Where is the assembly location, at what point in the cell cycle does this occur, and how does the replisome assemble? The answers to these questions are the rhythm by which this biological process moves, and while there is a great amount of overlap between the domains, each march to the beat of its own tune in respect to cell machinery and regulation. Boiled down to the very basic, DNA replication initiation in each Domains require three main

steps prior assembly of the replisome: (1) Origin recognition, (2) helicase recruitment, and (3) DNA unwinding (Figure 1.1).

Table 1.1. DNA replication machinery comparison between the three Domains. Attributes were separated into five different categories: (1) general, (2) prereplication complex (pre-RC), (3) preinitiation complex (pre-IC), (4) elongation complex, and (5) okazaki fragment maturation. Attributes highlighted in red, blue, and black are attributes associated with bacterial / bacterial-like replication machinery, eukaryotic / eukaryotic-like replication machinery, and archaeal specific replication machinery respectively. Adapted and modified from Kelman & Kelman (2014).

Attributes	Bacteria	Eukarya	Archaea	
			General	*T. kodakarensis*
Chromosome	Circular	Linear	Circular	Circular
Replication origin	Single	Multiple	Single/Multiple	Single
Prereplication complex (pre-RC)				
Origin recognition	DnaA	ORC	Cdc6/WhiP	Cdc6
Helicase	DnaB	MCM	MCM	MCM1/2/3
Helicase loader	DnaC	Cdc6/Cdt1	Cdc6	Cdc6
Preinitiation complex (pre-IC)				
Cdc45	-	Cdc45	Cdc45/GAN	GAN
GINS	-	GINS	GINS	GINS
SSB	SSB	RPA	RPA/SSB	RPA
Elongation complex				
Primase	DnaG	Polα/Primase	Primase	Primase
Sliding clamp	β-subunit	PCNA	PCNA	PCNA1/2
Clamp loader	τ-complex	RFC	RFC	RFC
DNA polymerase	PolC	PolB	PolB/D	PolB/D
Okazaki fragment maturation				
Primer removal	PolI	Fen1	Fen1/GAN/RnaseHII	Fen1/GAN/RnaseHII
Gap filling	PolI	PolB	PolB/D	PolB/PolD
DNA ligase	NAD^+-dependent	ATP-dependent	ATP-dependent	ATP-dependent

First, let us explore the “where” of DNA replication. When viewing this topic through the context of domain identity, we see that the way DNA is stored and structured is pivotal to the location of replication initiation. Bacteria and Archaea are quite kindred in this respect since both possess compact, circularized genome architecture; juxtaposed, eukaryotic genomes are linear, full of intergenic regions, and stored in nuclei. However, distinct from the canonically singular origin of

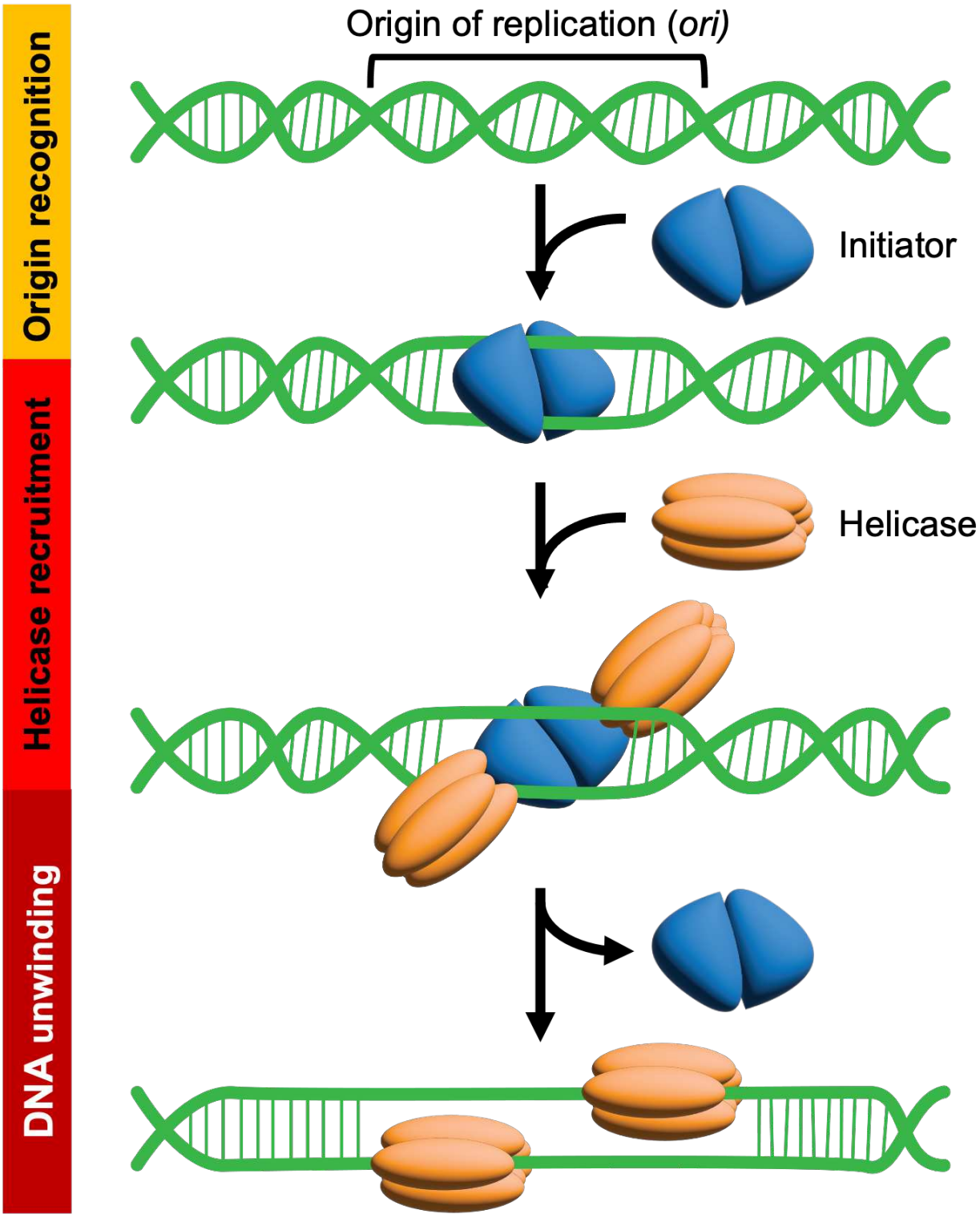


Figure 1.1. Three steps of DNA replication initiation.

replication (*ori*) of Bacteria, certain archaeal organisms can encode for multiple *ori* (Figure 1.2A)[3–7]. Furthermore, DnaA is the sole recognition protein in bacterial organisms while each of the multiple *ori* sites in archaeal organisms have their own unique recognition proteins[8]. Interestingly, these archaeal origin recognition proteins are often referred to as Cdc6/Orc/WhiP because of their homology to the eukaryotic Cdc6, Orc, and Cdt1 proteins, respectively, so we see that Archaea as a hybrid of eukaryotic and prokaryotic traits[8–12].

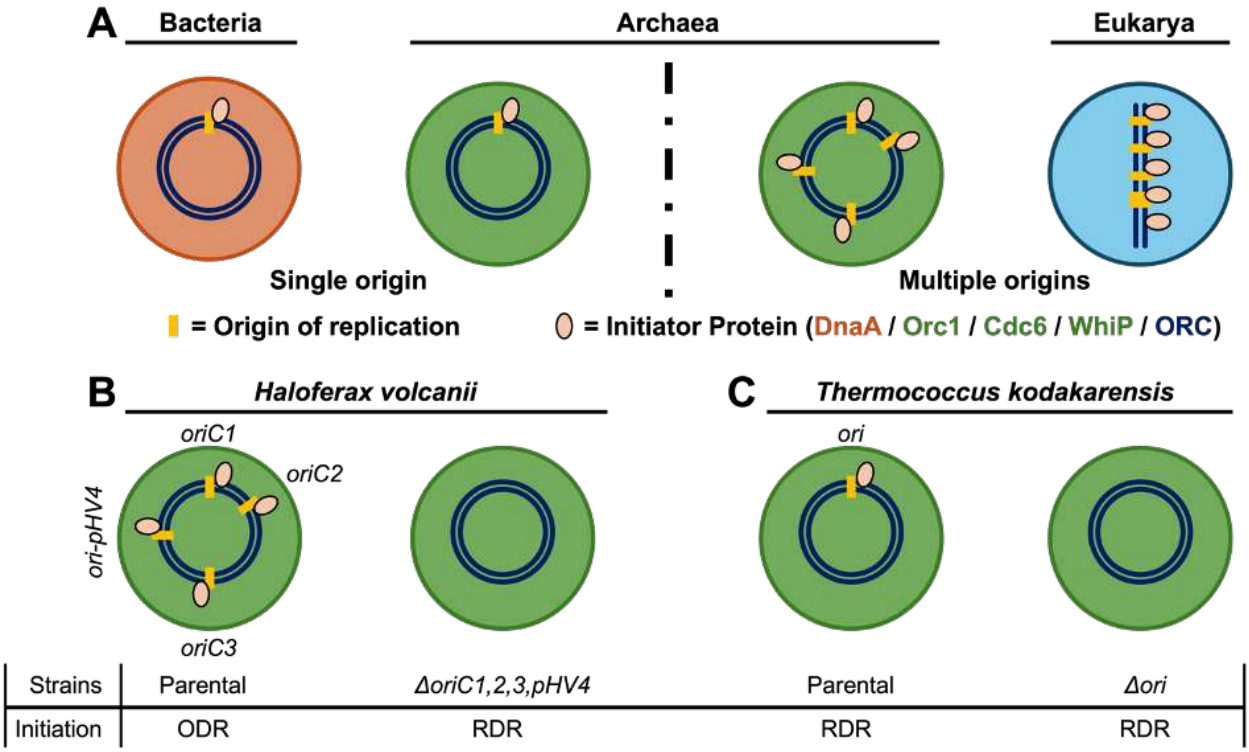


Figure 1.2. DNA replication initiation strategies in the three Domains. (**A**) The origin of replication(s) (*ori)* in each domain represented in orange bars are recognized by their domain-specific origin recognition protein (initiator protein). (**B**) Parental strain of *H. volcanii* initiates DNA replication through ODR and deletion of all four *ori* forces the strain to utilize RDR as the main mode of initiation. (**C**) Both parental strain and deletion of the presumptive *ori* strain of *T. kodakarensis* initiates DNA replication through RDR.

The Orc, Cdc6, and DnaA are all belonging to the initiator class of AAA+ family ATPases[9,12], but the differences in the enzymatic regulation of these proteins is what lends variation to the "when" of DNA replication through the domains of life. The question of "when" replication

initiation occurs in Eukarya is answered by a fine-tuned series of checkpoints in an organized cell cycle controlled by cyclin-dependent kinsases (CDKs) and expression patterns of the cyclins that direct their activities[13]. While cyclin homologues have not been identified in Archaea to this point, the dramatic changes in expression for Cdc6/ Orc in the archaeal *Sulfobulus* genus appear akin to the cyclin alterations in eukaryotic cell cycles, even if archaea do not possess as distinct of phases and checkpoints. In Bacteria, however, initiation of replication is enough to trigger a negative feedback loop that temporarily inactivates the area of the *oriC* chromosome, and the replisome itself converts active DnaA-ATP to inactive DnaA-ADP[13–15].

Finally, we explore the "how" in which replication is initiated. The Bacterial DnaA initiates DNA replication via binding of the origin binding box (ORB) sequences surrounding the *ori* followed by the unwinding of the double-stranded DNA (dsDNA) and allowing recruitment of the bacterial replicative helicase, DnaB, on the *ori* (Figure 1.1)[16–19]. While the archaeal and eukaryotic Orc/Cdc6 have been shown to bind ORBs with higher preference to binding dsDNA instead of single-stranded DNA (ssDNA), much is still unknown regarding how the unwinding of the *ori* dsDNA happens[20,21].

The recruitment of the archaeal and eukaryotic Orc/Cdc6 is followed with the assembly of the hexameric replicative helicases, minichromosome maintenance proteins (MCM) (Figure 1.1). The archaeal MCM is homologous to the eukaryotic MCM, both translocate on ssDNA in 3' to 5' direction, but not to the bacterial DnaB replicative helicase, which distinctly translocate on ssDNA from 5' to 3' direction (Table 1.2)[22–30]. In most cases, replicative helicases form ring-like structures with each other to form homohexamer oligomeric structures but archaeal replicative helicases have been shown previously to form homododecamer structures unlike the other domains[24,25]. Overall, although as a whole, archaeal DNA replication initiation appears to share

some similarities to the bacterial and eukaryotic, it also has unique properties inherent to the archaeal domain.

Table 1.2. Distinct characteristics of replicative helicases from all domains. Adapted from Kelman et al. (2020).

Characteristic	Bacteria	Eukarya	Archaea	
			General	*T. kodakarensis*
Protein(s)	DnaB	MCM2 to MCM7	MCM	MCM1/2/3
Essential for viability	Yes	Yes	Yes	Yes
Oligomeric structure	Homohexamer	Heterohexamer	Homododecamer	Homohexamer
Direction of translocation on ssDNA	5' to 3'	3' to 5'	3' to 5'	3' to 5'
Factors required for activity *in vitro*	None	Cdc45 and GINS	None	None
Binds to ssDNA and dsDNA	Yes	Yes	Yes	Yes
Translocates on ssDNA and dsDNA	Yes	Yes	Yes	Yes
Unwinds DNA-RNA hybrid	Yes	Yes	Yes	Yes

A closer look at the origin of replication

Focusing on the replication origin, when looking at the replisome components, the main replicative components in archaeal organisms are homologous to those found in Eukaryotes even though their optimum living conditions contrast dramatically (Table 1.1)[2,8,11,31]. Most canonically, DNA replication starts at a predetermined position(s) in the genome termed the origin of replication (*ori*)[2]. Bacterial organisms tend to utilize a single *ori* while Eukaryotic organisms utilize multiple *ori* to replicate their genomes (Figure 1.2A). Archaeal organisms have been shown to be a mixture of both where some only utilize a single *ori* like Bacteria and some can utilize multiple *ori* like Eukarya (Figure 1.2A). These more commonly accepted DNA replication initiation strategies through an *ori(s)* are termed **origin-dependent replication** (ODR). However, there are some documented instances where ODR is not the only viable method to begin DNA replication (Figure 1.2B-C).

Previous studies have documented DNA replication can be initiated in the absence of ODR, namely through **recombination-dependent replication** (RDR), wherein replisome assembly is initiated through recombination intermediates[32,33]. In eukaryotic organisms, RDR is employed to restart DNA replication after the collapse of DNA replication forks due to replication stress[34–41]. RDR has also been documented in viral DNA replication mechanisms, such as the one implemented by Papillomaviruses[42,43], as well as organellar DNA replication seen in mitochondrial and chloroplast DNA replication[44–46]. Although RDR is a ubiquitous phenomenon, only the archaeal domain of life has been shown to rely on RDR to sustain normal, or in some cases, accelerated growth of an organism[32]. The main consequence for relying on RDR is the need to maintain multiple copies of the genome, known as polyploidy, to initiate DNA replication. *T. kodakarensis* have been shown to maintain between 7-19 copies of genome per cell which allows for RDR[33,47]. Maintaining multiple copies of genome could be seen as a disadvantage, but some studies have proposed possibility of evolutionary advantages associated with polyploid organism which are, but not limited to: 1) as storage polymer, 2) as repair template to prevent mutation, and 3) allows for gene redundancy under unfavorable environment[48].

Based on previous MFA, RDR is thought to be the main DNA replication initiation strategy for *T. kodakarensis*[33]. Consistent with the MFA results, Cdc6 protein in *T. kodakarensis* have been shown to have minimal interaction with known essential DNA replication related proteins, such as PolD, when subjected to affinity purification experiment at 85°C[49]. In *H. volcanii*, deletion of *ori(s)* resulted in higher reliance to the availability of the conserved recombinase protein, RadA[32]; for *T. kodakarensis*, however, RadA is essential with or without intact Cdc6 or *ori*, both supporting the use of RDR strategy[33]. Although DNA replication initiation in *H. volcanii* and *T. kodakarensis* have been shown to utilize RDR, both organisms are fully equiped to initiate DNA replication through ODR pathway, where ODR is the preferred strategy in *H. volcanii* and RDR

is the preferred strategy in *T. kodakarensis*[32,33]*.* The main questions still persist in the field revolves around the regulation of the two major initiation pathways in these organisms.

Duality of Archaeal Replisome

Formation of the replication bubble follows a successful licensing and firing of *ori.* Archaeal replisome components share high homology to the eukaryotic replisome and much less compared to the bacterial replisome. Our model organism, *Thermococcus kodakarensis*, is native to thermal ocean vents where extremely high temperatures and other harsh environmental variables present conditions that are inhospitable to most other lifeforms[50]. Therefore, we must ask what differences in extremophiles allow them to assemble a replisome containing eukaryotic-like machinery and synthesize DNA in the backdrop of such labile locations. Surprisingly, some replication, recombination, and repair related proteins (RRR) in *T. kodakarensis* have been shown to be surprisingly dispensable from the genome[51–55].

Replicative DNA polymerases are arguably the central enzyme in DNA replication, and to date there are four known main families of replicative DNA polymerases: A, B, C, and D family DNA polymerase (Figure 1.3)[56]. B family DNA polymerases are conserved in both archaeal and eukaryotic domains, while D family DNA polymerases are only conserved within the archaeal domain[31,57]. In *T. kodakarensis* specifically, DNA polymerase B (Tk-PolB) is dispensable and DNA polymerase D (Tk-PolD) seems to be essential for survival of the cell[54,55]. Furthermore, the Tk-PolB deletion strain of *T. kodakarensis* has shown to demonstrate minimal to no phenotypic growth defects[54,55] but showed sensitivity towards DNA-damaging agents[4]. Furthermore, *T. kodakarensis* encodes for three minichromosome maintenance (Mcm) proteins, encoded by TK0096 (Tk-Mcm1), TK1361 (Tk-Mcm2), and TK1620 (Tk-Mcm3), and two proliferating cell nuclear antigen proteins (Tk-PCNA) encoded by TK0535 (Tk-PCNA1) and TK0582 (Tk-PCNA2).

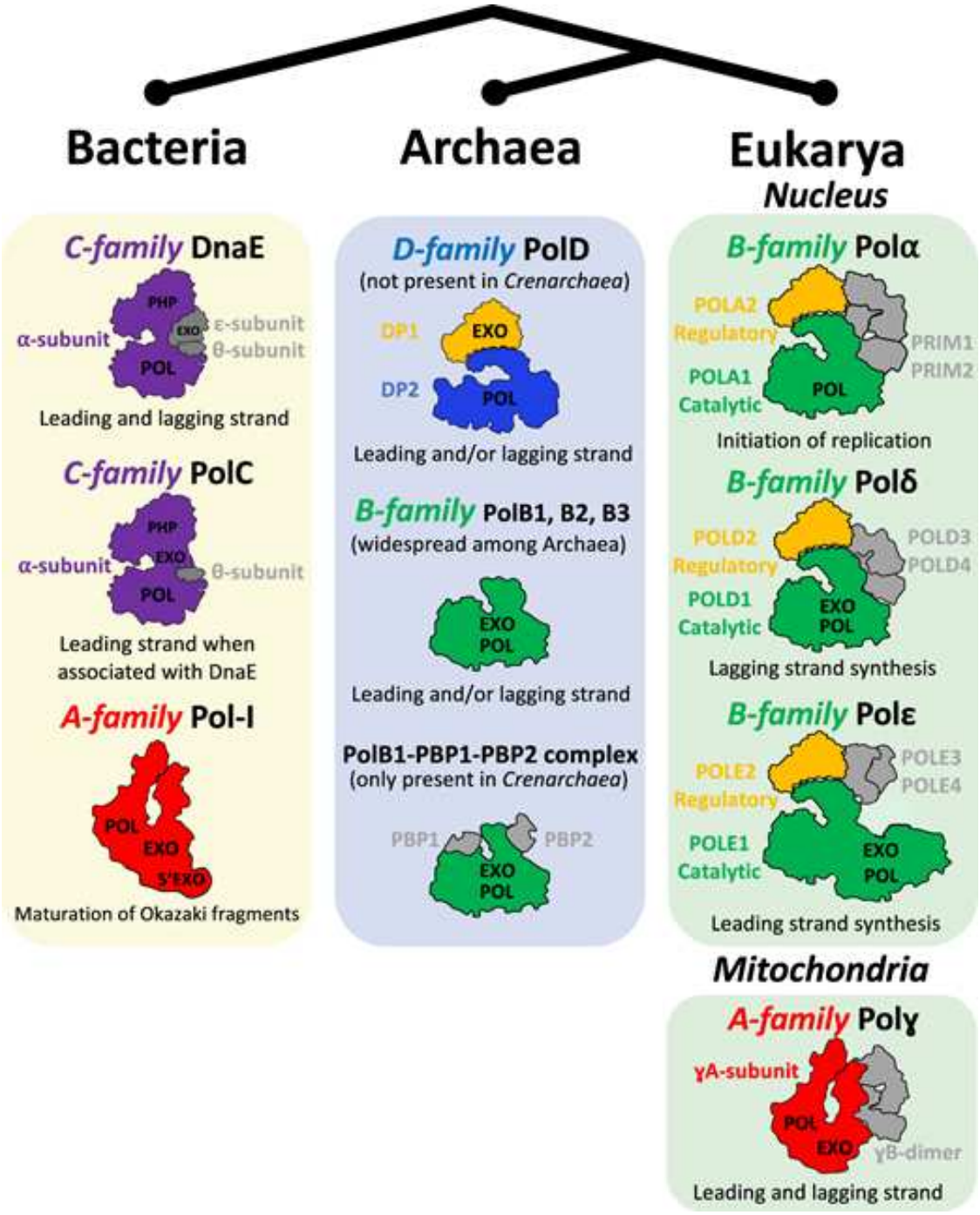


Figure 1.3. Adapted from Raia et al. (2019)

Only Tk-MCM3[51,58] and Tk-PCNA1[53] have been shown to be essential under optimized laboratory conditions.

Having a robust, efficient, and faithful DNA polymerase to maintain genetic integrity within the subsequent populations is integral for all species. Both DNA polymerases, PolB and PolD, are shown to be capable of 5' to 3' DNA synthesis and 3' to 5' exonuclease proofreading[59–61]. The most striking difference between the two polymerases is PolB being a single subunit enzyme whereas both polymerase and exonuclease activities are contained within one gene, PolD requires two subunits, the small and large subunits, to function as either a polymerase, contained within the large subunit or exonuclease contained within the small subunits[61]. Initially, Tk-PolB is thought to be the main replicative polymerase in the model archeon *T. kodakarensis*, based on *in vitro* evidence of PolB showing not only higher fidelity but also more robust activity compared to PolD[60,61].

The ability to delete Tk-PolB from the *T. kodakarensis* genome without lethality[54,55] caused us to speculate if this protein is critical for specific circumstances, and if so, is there a regulatory mechanism controlling the function(s) of each polymerases? In our literature analyses, we discovered the existence of mobile genetic elements termed **in**tervening pro**teins** (inteins) in multiple RRR-related proteins, including Tk-PolB, TK-PolD, and MCM-3 (Table 1.3), the latter two being essential genes. Therefore, we decided to focus our exploration around the "inteins" and how they might have been exaptated to direct DNA replication in respect to environmental circumstances for our model organism.

Table 1.3. Distribution of inteins in *T. kodakarensis* genome. Highlighted in the orange box are the RRR related proteins in *T. kodakarensis* that have been invaded by intein(s).

PROTEIN	FUNCTION	RRR	INT
MCM-3	Replicative DNA helicase	Yes	2
PolB	DNA polymerase	Yes	2
PolD	DNA polymerase II	Yes	1
RadA	Homologous recombinase	Yes	1
Ski2-like	DNA repair and recombination	Yes	1
RFC	Clamp loader	Yes	1
TopA	DNA topoisomerase 1	Yes	1
Rgy	Reverse gyrase	Yes	1
LHR	Large helicase-related protein	Yes	1
RNR	Ribonucleotide reductase	No	2
IF2	Initiation factor 2	No	1
KlbA	type II/IV secretion system ATPase	No	1

What is intein? A selfish genetic element or something else

Inteins (**In**tervening pro**teins**) are internal protein segments inserted in-frame within the coding region of another gene[62,63]. They are capable of autocatalytically removing themselves from a precursor protein. This robust post-/co-translational reaction is completed by (1) the release of the intein and (2) the formation of a native peptide bond ligating the external flanking protein segments (exteins). The underlying mechanism in protein splicing is self-contained and analogous to intron self-splicing.

The first discovery of intein was published in the 1990s by a group of researchers studying the vacuolar H+-ATPase (VMA1) gene from *Saccharomyces cerevisiae,* Brewer's yeast[64,65]. Gene alignment study between the VMA1 gene from *S. cerevisiae* showed not only high sequence

homology to the catalytic subunit of the vacuolar H+-ATPase genes from both carrot and *Neurospora crassa*, but also a large 50-kDa (454 amino-acid residues) in-frame insert of the VMA1-derived endonuclease (VMA1 intein) (Figure 1.4A). The VMA1 gene in S. cerevisiae is expressed as a 120-kDa precursor protein containing all 1071 aa encoded within the singular open reading frame. The post-translational excision of the 50-kDa (454 aa) VMA1 intein and the subsequent ligation of the N- and C-terminal exteins led to the formation of the mature 70-kDa catalytic subunit of the vacuolar H+-ATPase[64,66]. Since then, bioinformatics analysis of known genomes have identified inteins spread across Bacteria, Eukarya, and Archaea (Figure 1.4B)[67,68].

Naturally occurring inteins are classified into two major classes, which are the standard inteins and the split inteins[67,69]. Standard inteins are resulted from the transcription and translation of a single gene followed by protein splicing from a single polypeptide (precursor protein) to form the mature protein and release the intein from the larger precursor *in cis* (Figure 1.5A). Some inteins can be the results of two independently transcribed and translated genes within a single cell, termed the split inteins, wherein two precursor proteins need to come together through non-covalent interactions and subsequently catalyze protein splicing *in trans* (Figure 1.5B). Although both classes are naturally occurring, the vast majority of identified inteins are known to fall into the canonical standard intein class, less than 5% are classed into split-intein[69].

Phylogenetic distribution of inteins suggests not only the primordial origins of these genetic elements but also the prevalence of horizontal gene transfers (HGT) propagating the spread of inteins[69,70]. Inteins are found to have invaded a wide variety of host genes, not just orthologous/paralogous/homologous genes suggesting their primordial origins. Orthologous genes in closely related organisms showed the presence of intein while others showed the

absence of intein suggesting the HGT of these inteins. Many intein-containing alleles, just like the VMA1 inteins, encodes for a homing endonuclease domain (HEN-domain / HED) and it

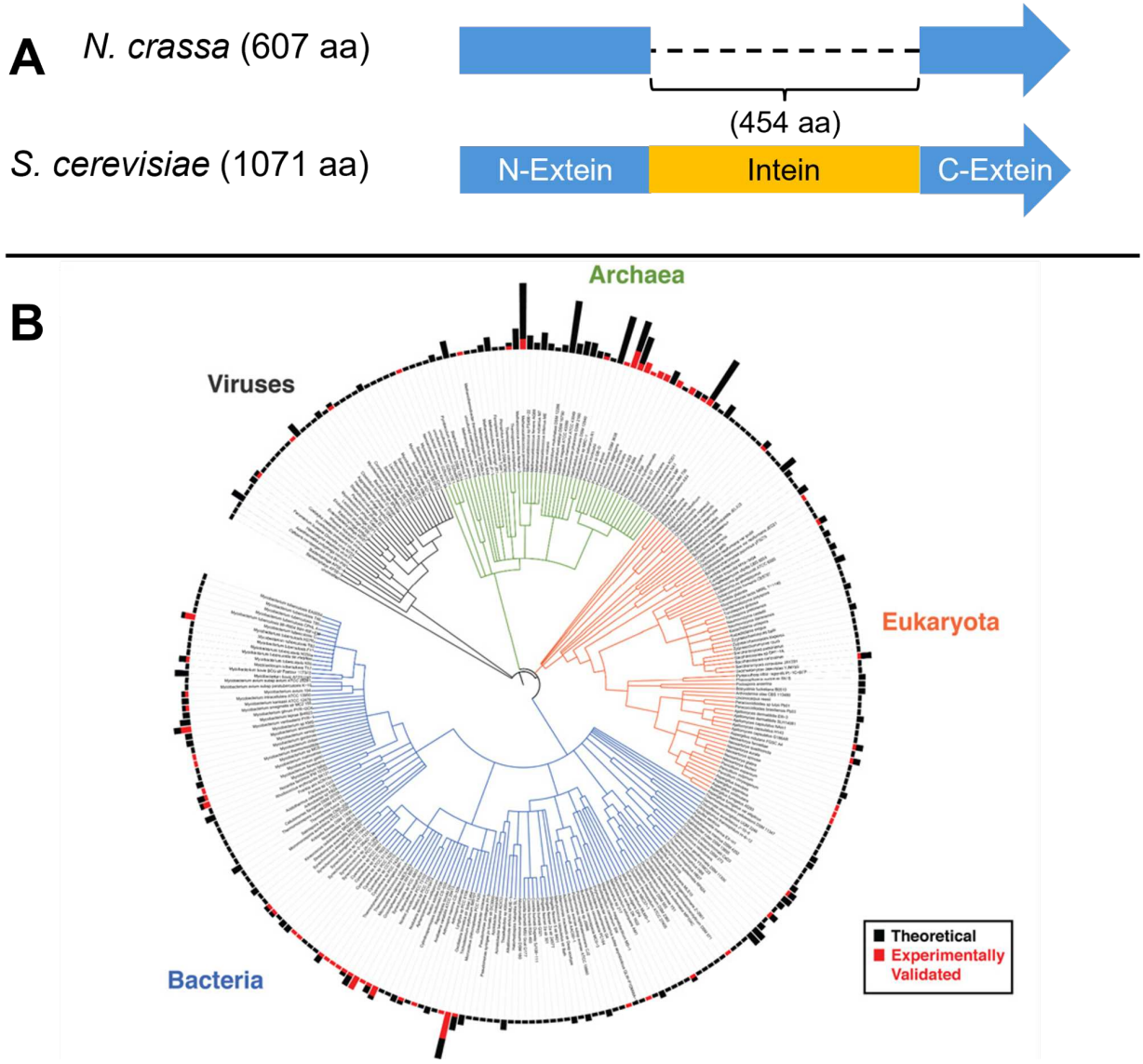


Figure 1.4. Discovery of intein splicing. (**A**) Bioinformatic sequence alignment between the catalytic subunit of vacuolar H+-ATPase genes from *Neurospora crassa* and *Saccharomyces cerevisiae*. Both identical residues and conserved amino acids replacements are depicted in blue. Dashed line represents the main gap introduced from the intein invasion in the *S. cerevisiae* VMA1 gene. (**B**) The phylogenetic distribution of intein-containing organisms adapted from Shah et al. (2020).

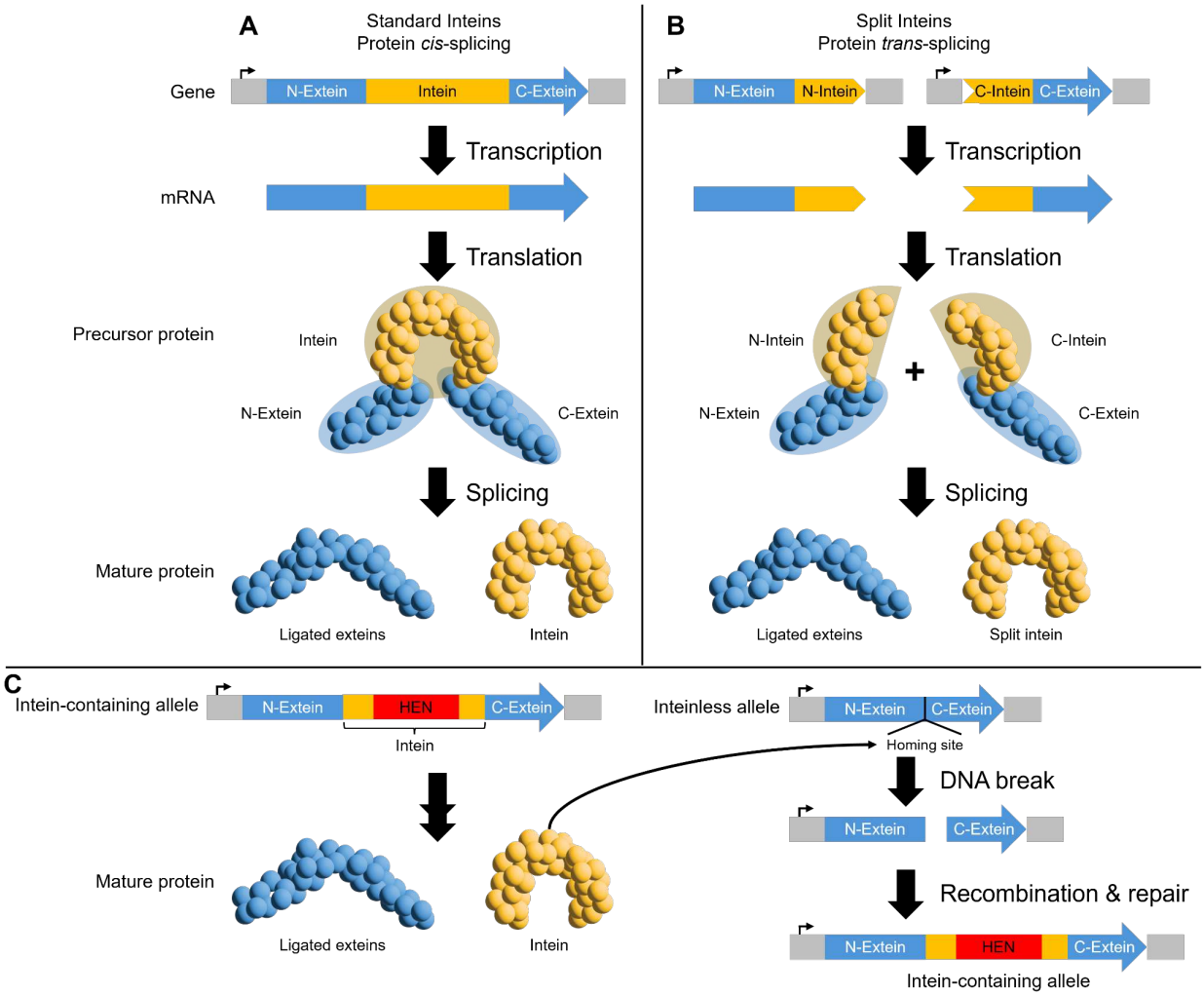


Figure 1.5. Intein protein splicing and intein invasion. (**A-B**) Diagrams depicting the expression and processing of a standard intein-invaded gene (**A**) and a split-intein invaded gene (**B**). The 5' and 3' UTR of the intein containing genes are represented in light gray boxes. The translated protein is depicted in the blue and orange circles denoting the exteins and inteins respectively. (**C**) A diagram depicting the propagation of intein from an intein-containing allele into an inteinless (intein-lacking) allele.

enables HEN-containing inteins to spread horizontally to different species containing the inteinless allele[71–73]. HEN-containing inteins have been shown to have a relatively long and asymmetrical recognition sequence (homing-site), between 14-40 bp, and able to induce a double stranded brake within the homing site of the inteinless allele[71]. The subsequent repair of the double-stranded break induced by the HEN-domain results in the integration of the intein into the inteinless allele (Figure 1.5C). The double-stranded DNA cleavage resulting in 3'-hydroxyl overhangs is catalyzed by the LAGLIDADG motif consisting of a dodecapeptide sequence within the HEN-domain, but also homologous to HO endonucleases, group I and archaeal introns[63,74].

Although the presence of HEN is thought to be essential for the inter-/intra-species spread of inteins, some inteins have evolved to function without their HEN-domain. This group of HEN-less inteins are termed "mini-intein" and do not face the same selective pressure, especially for the instances they have invaded essential proteins[63]. Therefore, we must ask ourselves if there is no selective pressure through the presence of a homing endonuclease, is it possible they're playing a role *in vivo*?

One study explored this question by evaluating the Clusters of Orthologous Groups of proteins (COGs), which gave insight towards the bias of inteins clustering in their host genes active site and also revealed the strong bias towards the type of host gene[68]. Interestingly, a majority of all inteins invade proteins classified into DNA replication, recombination, and repair functional category; specifically, 61.5% in Bacteria and 66.9% in Archaea. Inteins are also observed to invade nucleotide transport and metabolism genes to a lesser extent, but overwhelmingly, they correlate with DNA-related genes. Comparing the intein-containing genes in Bacteria and Archaea, you can identify many non-orthologous genes with high overlap in function that are also invaded by inteins like RecA (Bacteria)/ RadA (Archaea) and DnaB (Bacteria) / MCM

(Archaea). This suggests an ancient origin to intein involvement in DNA repair, maintenance, and recombination.

When exploring the potential advantages of intein presence, it must be asked if these post-translational regulators can be responsive to environmental cues. Termed conditional protein splicing (CPS), these inteins are able to modulate their splicing efficiency within a certain set of parameters. Many *in vitro* studies on inteins from archaeal and bacterial organisms have shown examples of CPS in response to altered temperatures, salt concentrations, oxidative stress, DNA damage, and metal ion levels[74–83]. *Pyrococcus horikoshii* and *Thermococcus sibericus*, which are in a similar vein to our model organism, contain inteins in their RadA proteins that demonstrate CPS behaviors in respect to temperature changes *in vitro*[75]. This evidence seems to point to the exaptation of inteins from a mere mobile genetic element into a functional regulatory element of the invaded gene. However, the field of intein studies has yet to provide *in vivo* evidence of this notion.

REFERENCES

1. O'Donnell, M., Langston, L. & Stillman, B. Principles and Concepts of DNA Replication in Bacteria, Archaea, and Eukarya. *Cold Spring Harb Perspect Biol* **5,** (2013).

2. Kelman, L. M. & Kelman, Z. Archaeal DNA replication. *Annu Rev Genet* **48,** 71–97 (2014).

3. Lundgren, M., Andersson, A., Chen, L., Nilsson, P. & Bernander, R. Three replication origins in Sulfolobus species: Synchronous initiation of chromosome replication and asynchronous termination. *Proc Natl Acad Sci U S A* **101,** 7046–7051 (2004).

4. Zhang, R. & Zhang, C. T. Multiple replication origins of the archaeon Halobacterium species NRC-1. *Biochem Biophys Res Commun* **302,** 728–734 (2003).

5. Matsunaga, F., Norais, C., Forterre, P. & Myllykallio, H. Identification of short 'eukaryotic' Okazaki fragments synthesized from a prokaryotic replication origin. *EMBO Rep* **4,** 154 (2003).

6. Myllykallio, H., Lopez, P., López-García, P., Heilig, R., Saurin, W., Zivanovic, Y., Philippe, H. & Forterre, P. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science (1979)* **288,** 2212–2215 (2000).

7. Matsunaga, F., Forterre, P., Ishino, Y. & Myllykallio, H. In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. *Proc Natl Acad Sci U S A* **98,** 11152–11157 (2001).

8. Greci, M. D. & Bell, S. D. Archaeal DNA Replication. *Annu Rev Microbiol* **74,** 65–80 (2020).

9. Samson, R. Y., Xu, Y., Gadelha, C., Stone, T. A., Faqiri, J. N., Li, D., Qin, N., Pu, F., Liang, Y. X., She, Q. & Bell, S. D. Specificity and function of archaeal DNA replication initiator proteins. *Cell Rep* **3,** 485–496 (2013).

10. Raymann, K., Forterre, P., Brochier-Armanet, C. & Gribaldo, S. Global Phylogenomic Analysis Disentangles the Complex Evolutionary History of DNA Replication in Archaea. *Genome Biol Evol* **6,** 192–212 (2014).

11. Makarova, K. S. & Koonin, E. V. Archaeology of Eukaryotic DNA Replication. *Cold Spring Harb Perspect Biol* **5,** a012963 (2013).

12. Bleichert, F., Botchan, M. R. & Berger, J. M. Mechanisms for initiating cellular DNA replication. *Science (1979)* **355,** (2017).

13. Margolin, W. & Bernander, R. How Do Prokaryotic Cells Cycle? *Curr Biol* **14,** R768 (2004).

14. Skarstad, K., Boye, E. & Steen, H. B. Timing of initiation of chromosome replication in individual Escherichia coli cells. *EMBO J* **5,** 1711 (1986).

15. Skarstad, K. & Katayama, T. Regulating DNA Replication in Bacteria. *Cold Spring Harb Perspect Biol* **5,** 1–17 (2013).

16. Abe, Y., Jo, T., Matsuda, Y., Matsunaga, C., Katayama, T. & Ueda, T. Structure and function of DnaA N-terminal domains: specific sites and mechanisms in inter-DnaA interaction and in DnaB helicase loading on oriC. *J Biol Chem* **282,** 17816–17827 (2007).

17. Messer, W. The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication. *FEMS Microbiol Rev* **26,** 355–374 (2002).

18. Seitz, H., Weigel, C. & Messer, W. The interaction domains of the DnaA and DnaB replication proteins of Escherichia coli. *Mol Microbiol* **37,** 1270–1279 (2000).

19. Sutton, M. D., Carr, K. M., Vicente, M. & Kaguni, J. M. Escherichia coli DnaA protein. The N-terminal domain and loading of DnaB helicase at the E. coli chromosomal origin. *J Biol Chem* **273,** 34255–34262 (1998).

20. Bell, S. D. Archaeal orc1/cdc6 proteins. *Subcell Biochem* **62,** 59–69 (2012).

21. Costa, A., Hood, I. V. & Berger, J. M. Mechanisms for Initiating Cellular DNA Replication. *Annu Rev Biochem* **82,** 25 (2013).

22. Stelter, M., Gutsche, I., Kapp, U., Bazin, A., Bajic, G., Goret, G., Jamin, M., Timmins, J. & Terradot, L. Architecture of a dodecameric bacterial replicative helicase. *Structure* **20,** 554–564 (2012).

23. Remus, D., Beuron, F., Tolun, G., Griffith, J. D., Morris, E. P. & Diffley, J. F. X. Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139,** 719–730 (2009).

24. Sakakibara, N., Kelman, L. M. & Kelman, Z. Unwinding the structure and function of the archaeal MCM helicase. *Mol Microbiol* **72,** 286–296 (2009).

25. Sakakibara, N., Kelman, L. M. & Kelman, Z. How is the archaeal MCM helicase assembled at the origin? Possible mechanisms. *Biochem Soc Trans* **37,** 7–11 (2009).

26. Fernandez, A. J. & Berger, J. M. Mechanisms of hexameric helicases. *Crit Rev Biochem Mol Biol* **56,** 621 (2021).

27. Krupovič, M., Gribaldo, S., Bamford, D. H. & Forterre, P. The evolutionary history of archaeal MCM helicases: a case study of vertical evolution combined with hitchhiking of mobile genetic elements. *Mol Biol Evol* **27,** 2716–2732 (2010).

28. Bell, S. D. & Botchan, M. R. The Minichromosome Maintenance Replicative Helicase. *Cold Spring Harb Perspect Biol* **5,** 12807–12808 (2013).

29. Costa, A. & Onesti, S. Structural biology of MCM helicases. *Crit Rev Biochem Mol Biol* **44,** 326–342 (2009).

30. Brewster, A. S. & Chen, X. S. Insights into MCM functional mechanism: lessons learned from the archaeal MCM complex. *Crit Rev Biochem Mol Biol* **45,** 243 (2010).

31. Makarova, K. S., Krupovic, M. & Koonin, E. V. Evolution of replicative DNA polymerases in archaea and their contributions to the eukaryotic replication machinery. *Front Microbiol* **5,** Preprint at https://doi.org/10.3389/fmicb.2014.00354 (2014)

32. Hawkins, M., Malla, S., Blythe, M. J., Nieduszynski, C. A. & Allers, T. Accelerated growth in the absence of DNA replication origins. *Nature* **503,** 544–547 (2013).

33. Gehring, A. M., Astling, D. P., Matsumi, R., Burkhart, B. W., Kelman, Z., Reeve, J. N., Jones, K. L. & Santangelo, T. J. Genome replication in Thermococcus kodakarensis independent of Cdc6 and an origin of replication. *Front Microbiol* **8,** (2017).

34. Mizuno, K., Lambert, S., Baldacci, G., Murray, J. M. & Carr, A. M. Nearby inverted repeats fuse to generate acentric and dicentric palindromic chromosomes by a replication template exchange mechanism. *Genes Dev* **23,** 2876–2886 (2009).

35. Jalan, M., Oehler, J., Morrow, C. A., Osman, F. & Whitby, M. C. Factors affecting template switch recombination associated with restarted DNA replication. *Elife* **8,** (2019).

36. Nguyen, M. O., Jalan, M., Morrow, C. A., Osman, F. & Whitby, M. C. Recombination occurs within minutes of replication blockage by RTS1 producing restarted forks that are prone to collapse. *Elife* **4,** (2015).

37. Lambert, S., Watson, A., Sheedy, D. M., Martin, B. & Carr, A. M. Gross chromosomal rearrangements and elevated recombination at an inducible site-specific replication fork barrier. *Cell* **121,** 689–702 (2005).

38. Lambert, S., Mizuno, K., Blaisonneau, J., Martineau, S., Chanet, R., Fréon, K., Murray, J. M., Carr, A. M. & Baldacci, G. Homologous recombination restarts blocked replication forks at the expense of genome rearrangements by template exchange. *Mol Cell* **39,** 346–359 (2010).

39. Miyabe, I., Mizuno, K., Keszthelyi, A., Daigaku, Y., Skouteri, M., Mohebi, S., Kunkel, T. A., Murray, J. M. & Carr, A. M. Polymerase δ replicates both strands after homologous recombination-dependent fork restart. *Nat Struct Mol Biol* **22,** 932–938 (2015).

40. Petermann, E., Orta, M. L., Issaeva, N., Schultz, N. & Helleday, T. Hydroxyurea-stalled replication forks become progressively inactivated and require two different RAD51-mediated pathways for restart and repair. *Mol Cell* **37,** 492–502 (2010).

41. Conti, B. A. & Smogorzewska, A. Mechanisms of direct replication restart at stressed replisomes. *DNA Repair (Amst)* **95,** 102947 (2020).

42. Jeske, H., Lütgemeier, M. & Preiß, W. DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *EMBO J* **20,** 6158–6167 (2001).

43. Sakakibara, N., Chen, D. & McBride, A. A. Papillomaviruses Use Recombination-Dependent Replication to Vegetatively Amplify Their Genomes in Differentiated Cells. *PLoS Pathog* **9,** e1003321 (2013).

44. Cheng, N., Lo, Y. S., Ansari, M. I., Ho, K. C., Jeng, S. T., Lin, N. S. & Dai, H. Correlation between mtDNA complexity and mtDNA replication mode in developing cotyledon mitochondria during mung bean seed germination. *New Phytologist* **213,** 751–763 (2017).

45. Morley, S. A., Ahmad, N. & Nielsen, B. L. Plant Organelle Genome Replication. *Plants* **8,** (2019).

46. Maréchal, A. & Brisson, N. Recombination and the maintenance of plant organelle genome stability. *New Phytologist* **186,** 299–317 (2010).

47. Spaans, S. K., van der Oost, J. & Kengen, S. W. M. The chromosome copy number of the hyperthermophilic archaeon Thermococcus kodakarensis KOD1. *Extremophiles* **19,** 741–750 (2015).

48. Soppa, J. Evolutionary advantages of polyploidy in halophilic archaea. in *Biochem Soc Trans* **41,** 339–343 (Portland Press, 2013).

49. Li, Z., Santangelo, T. J., Čuboňová, L., Reeve, J. N. & Kelman, Z. Affinity purification of an archaeal DNA replication protein network. *mBio* **1,** 221–231 (2010).

50. Atomi, H., Fukui, T., Kanai, T., Morikawa, M. & Imanaka, T. Description of Thermococcus kodakaraensis sp. nov., a well studied hyperthermophilic archaeon previously reported as Pyrococcus sp. KOD1. *Archaea* **1,** 263 (2004).

51. Pan, M., Santangelo, T. J., Li, Z., Reeve, J. N. & Kelman, Z. Thermococcus kodakarensis encodes three MCM homologs but only one is essential. *Nucleic Acids Res* **39,** 9671–9680 (2011).

52. Burkhart, B. W., Cubonova, L., Heider, M. R., Kelman, Z., Reeve, J. N. & Santangelo, T. J. The GAN exonuclease or the flap endonuclease Fen1 and RNase HII are necessary for viability of Thermococcus kodakarensis. *J Bacteriol* **199,** (2017).

53. Pan, M., Santangelo, T. J., Čuboňová, Ľ., Li, Z., Metangmo, H., Ladner, J., Hurwitz, J., Reeve, J. N. & Kelman, Z. Thermococcus kodakarensis has two functional PCNA homologs but only one is required for viability. *Extremophiles* **17,** 453–461 (2013).

54. Cubonová, L., Richardson, T., Burkhart, B. W., Kelman, Z., Connolly, B. A., Reeve, J. N. & Santangelo, T. J. Archaeal DNA polymerase D but not DNA polymerase B is required for genome replication in Thermococcus kodakarensis. *J Bacteriol* **195,** 2322–8 (2013).

55. Kushida, T., Narumi, I., Ishino, S., Ishino, Y., Fujiwara, S., Imanaka, T. & Higashibata, H. Pol B, a Family B DNA Polymerase, in Thermococcus kodakarensis is Important for DNA Repair, but not DNA Replication. *Microbes Environ* **34,** 316 (2019).

56. Raia, P., Delarue, M. & Sauguet, L. An updated structural classification of replicative DNA polymerases. *Biochem Soc Trans* **47,** 239–249 Preprint at https://doi.org/10.1042/BST20180579 (2019)

57. Tahirov, T. H., Makarova, K. S., Rogozin, I. B., Pavlov, Y. I. & Koonin, E. V. Evolution of DNA polymerases: An inactivated polymerase-exonuclease module in Pol ε and a chimeric origin of eukaryotic polymerases from two classes of archaeal ancestors. *Biol Direct* **4,** (2009).

58. Ishino, S., Fujino, S., Tomita, H., Ogino, H., Takao, K., Daiyasu, H., Kanai, T., Atomi, H. & Ishino, Y. Biochemical and genetical analyses of the three mcm genes from the hyperthermophilic archaeon, Thermococcus kodakarensis. *Genes to Cells* **16,** 1176–1189 (2011).

59. Takashima, N., Ishino, S., Oki, K., Takafuji, M., Yamagami, T., Matsuo, R., Mayanagi, K. & Ishino, Y. Elucidating functions of DP1 and DP2 subunits from the Thermococcus kodakarensis family D DNA polymerase. *Extremophiles* **23,** 161–172 (2019).

60. Greenough, L., Kelman, Z. & Gardner, A. F. The roles of family B and D DNA polymerases in thermococcus species 9°N Okazaki fragment maturation. *Journal of Biological Chemistry* **290,** 12514–12522 (2015).

61. Henneke, G., Flament, D., Hübscher, U., Querellou, J. & Raffin, J. P. The Hyperthermophilic Euryarchaeota Pyrococcus abyssi Likely Requires the Two DNA Polymerases D and B for DNA Replication. *J Mol Biol* **350,** 53–64 (2005).

62. Perler, F. B. & Allewell, N. M. Evolution, Mechanisms, and Applications of Intein-mediated Protein Splicing. *J Biol Chem* **289,** 14488 (2014).

63. Lennon, C. W. & Belfort, M. Inteins. *Current Biology* **27,** R204–R206 (2017).

64. Hirata, R., Ohsumi, Y., Nakano, A., Kawasaki, H., Suzuki, K. & Anraku, Y. Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of Saccharomyces cerevisiae. *Journal of Biological Chemistry* **265,** 6726–6733 (1990).

65. Anraku, Y. & Satow, Y. Reflections on protein splicing: structures, functions and mechanisms. *Proc Jpn Acad Ser B Phys Biol Sci* **85,** 409 (2009).

66. Kane, P. M., Yamashiro, C. T., Wolczyk, D. F., Neff, N., Goebl, M. & Stevens, T. H. Protein Splicing Converts the Yeast TFP1 Gene Product to the 69-kdDSubunit of the Vacuolar H+-Adenosine Triphosphatase. *Science (1979)* **250,** 651–657 (1990).

67. Shah, N. H. & Muir, T. W. Inteins: Nature's Gift to Protein Chemists. *Chem Sci* **5,** 446–461 (2014).

68. Novikova, O., Jayachandran, P., Kelley, D. S., Morton, Z., Merwin, S., Topilina, N. I. & Belfort, M. Intein clustering suggests functional importance in different domains of life. *Mol Biol Evol* **33,** 783–799 (2016).

69. Pavankumar, T. L. Inteins: Localized Distribution, Gene Regulation, and Protein Engineering for Biological Applications. *Microorganisms* **6,** (2018).

70. Tharappel, A. M., Li, Z. & Li, H. Inteins as Drug Targets and Therapeutic Tools. *Front Mol Biosci* **9,** (2022).

71. Nishioka, M., Fujiwara, S., Takagi, M. & Imanaka, T. Characterization of two intein homing endonucleases encoded in the DNA polymerase gene of Pyrococcus kodakaraensis strain KOD1. *Nucleic Acids Res* **26,** 4409 (1998).

72. Chevalier, B. S. & Stoddard, B. L. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res* **29,** 3757 (2001).

73. Robinzon, S., Cawood, A. R., Ruiz, M. A., Gophna, U., Altman-Price, N. & Mills, K. V. Protein Splicing Activity of the Haloferax volcanii PolB-c Intein Is Sensitive to Homing Endonuclease Domain Mutations. *Biochemistry* **59,** 3359–3367 (2020).

74. Wood, D. W., Belfort, M. & Lennon, C. W. Inteins-mechanism of protein splicing, emerging regulatory roles, and applications in protein engineering. *Front Microbiol* **14,** (2023).

75. Lennon, C. W., Stanger, M., Banavali, N. K. & Belfort, M. Conditional protein splicing switch in hyperthermophiles through an intein-extein partnership. *mBio* **9,** (2018).

76. Yalala, V. R., Lynch, A. K. & Mills, K. V. Conditional Alternative Protein Splicing Promoted by Inteins from Haloquadratum walsbyi. *Biochemistry* **61,** 294 (2022).

77. Belfort, M. Mobile self-splicing introns and inteins as environmental sensors. *Curr Opin Microbiol* **38,** 51–58 Preprint at https://doi.org/10.1016/j.mib.2017.04.003 (2017)

78. Green, C. M., Li, Z., Smith, A. D., Novikova, O., Bacot-Davis, V. R., Gao, F., Hu, S., Banavali, N. K., Thiele, D. J., Li, H. & Belfort, M. Spliceosomal Prp8 intein at the crossroads of protein and RNA splicing. *PLoS Biol* **17,** (2019).

79. Lennon, C. W., Stanger, M. J. & Belfort, M. Mechanism of Single-Stranded DNA Activation of Recombinase Intein Splicing. *Biochemistry* **58,** 3335 (2019).

80. Reitter, J. N., Cousin, C. E., Nicastri, M. C., Jaramillo, M. V. & Mills, K. V. Salt-Dependent Conditional Protein Splicing of an Intein from Halobacterium salinarum. *Biochemistry* **55,** 1279–1282 (2016).

81. Lennon, C. W., Stanger, M. & Belfort, M. Protein splicing of a recombinase intein induced by ssDNA and DNA damage. *Genes Dev* **30,** 2663–2668 (2016).

82. Topilina, N. I., Green, C. M., Jayachandran, P., Kelley, D. S., Stanger, M. J., Piazza, C. L., Nayak, S. & Belfort, M. SufB intein of Mycobacterium tuberculosis as a sensor for oxidative and nitrosative stresses. *Proc Natl Acad Sci U S A* **112,** 10348–10353 (2015).

83. Chiarolanzio, K. C., Pusztay, J. M., Chavez, A., Zhao, J., Xie, J., Wang, C. & Mills, K. V. Allosteric influence of extremophile hairpin motif mutations on the protein splicing activity of a hyperthermophilic intein. *Biochemistry* **59,** 2459 (2020).

CHAPTER 2: TRANSFORMATION TECHNIQUES FOR THE ANAEROBIC HYPERTHERMOPHILE *THERMOCOCCUS KODAKARENSIS*

Summary

Genetic manipulation is an essential tool to investigate complex microbiological phenomena. In this chapter we describe the techniques required to transform the model hyperthermophilic, anaerobic archaeon *Thermococcus kodakarensis*. *T. kodakarensis* can support two modes of genetic manipulation, dependent either on homologous recombination into the genome or through retention of autonomously replicating plasmids. The robust genetic system developed in *T. kodakarensis* offers a variety of selectable and counter-selectable markers for complex, accurate and iterative genetic manipulations offering greater flexibility to probe gene function *in vivo*.

Introduction

Many Archaea can survive and often thrive in the environmental extremes, ranging from saturating salinities to high temperatures within hot springs and near marine vents[1–7]. Archaeal systems are increasingly employed for production of specialized and commodity bio-products, bioremediation efforts, and archaeal enzymes that function at the extremes of pH, temperature and salinity are often used in molecular biology and biotechnological applications[5,8–15]. Efforts to understand and take advantage of the unique characteristics of archaeal organisms demand a reliable means of genetic manipulation for rational and iterative strain construction.

Thermococcus kodakarensis is a naturally competent archaeal organism that thrives in an anaerobic, hyperthermophilic (85 °C optimal growth temperature) environment[16–19]. The genetic

[1] Most of this chapter was previously published under the following title with a few updates: Liman, G. L. S., Stettler, M. E., & Santangelo, T. J. (2022). Transformation Techniques for the Anaerobic Hyperthermophile Thermococcus kodakarensis. *Methods in molecular biology (Clifton, N.J.)*, *2522*, 87–104. https://doi.org/10.1007/978-1-0716-2445-6_5

system for *T. kodakarensis* has been continuously refined through the combined efforts of multiple laboratories and these foundational and improved techniques are now regularly employed for many euryarchaeal species[20–26]. Initial efforts focused on homologous-repetitive and markerless-alterations. The absence of naturally occurring autonomously replicating plasmids lead to the development of novel shuttle vectors that provide a complementary method to alter the genotypes and phenotype of *T. kodakarensis.*

Although multiple strains of *T. kodakarensis* have been used as parental strains, the genetic system described here is optimized for, and reliant on strain TS559 (ΔTK2276; ΔTK0254::TK2276; ΔTK0149; ΔTK0664) [23], which is an agmatine and tryptophan auxotroph strain of *T. kodakarensis*. We describe two techniques to manipulate TS559. The first is based on integration and subsequent excision of nonreplicative plasmids into the genome of TS559 to restore the agmatine biosynthesis pathway via reintroduction of TK0149, encoding a pyruvoyl-dependent arginine decarboxylase, and counterselective pressures afforded by TK0664, a hypoxanthine guanine phosphoribosyltransferase [23]. The second transformation procedure uses a shuttle vector system that restores the tryptophan biosynthesis pathway via reintroduction of TK0254, which encodes for the large subunit of anthranilate synthase, and expresses PF1848 (a hydroxymethylglutaryl coenzyme A reductase gene from *Pyrococcus furiosus*) providing resistance to statin-based antibiotics.

2. Materials

2.1 Microbial Culture

All solutions are sterilized by autoclaving unless noted otherwise. Solutions are made with water purified to 18 MΩ-resistance. All solutions are adjusted to their final volume after the addition of dry ingredients.

2.1.1 *Escherichia coli* media

1. Luria-Bertani (Ec-LB-Amp) liquid media (1 L): 10 g tryptone, 10 g NaCl, 5 g yeast extract, 100 µg/mL ampicillin.
2. LB solid media (Ec-LB-Amp plates): 10 g tryptone, 10 g NaCl, 5 g yeast extract, 20 g agar, 100 µg/mL ampicillin.
3. 25 mg/mL ampicillin

2.1.2 *Thermococcus kodakarensis* media

1. KOD Vitamins (*see* **Note 1, 6**) (1000x) (1 L): 0.2 g niacin, 0.08 g biotin, 0.2 g pantothenate, 0.2 g lipoic acid, 0.08 g folic acid, 0.2 g thiamine, 0.2 g riboflavin, 0.2 g pyridoxine, 0.2 g cobalamin.
2. 1 M Agmatine sulfate (1000x)
3. Wolfe's Trace Minerals (200x) (1 L): 0.5 g $MnSO_4.H_2O$, 0.1 g $CoCl_2.6H_2O$, 0.1 g $ZnSO_4.7H_2O$, 0.01 g $CuSO_4 \cdot 5H_2O$, 0.01 g $AlK(SO_4)_2.12H_2O$, 0.01 g H_3BO_3, 0.01 g $Na_2MoO_4.2H_2O$
4. Artificial Sea Water (ASW)-YT (1 L): 5 g tryptone (*see* **Note 2**), 5 g yeast extract (*see* **Note 3**), 1x Wolfe's trace mineral solution, 1x KOD vitamins, 20 g NaCl, 3 g $MgCl_2.6H_2O$, 6 g $MgSO_4.7H_2O$, 1 g $(NH4)_2SO_4$, 0.2 g $NaHCO_3$**,** 0.3 g $CaCl_2.2H_2O$, 0.5 g KCl, 0.420 g KH_2PO_4, 0.050 g NaBr, 0.020 g $SrCl_2.6H_2O$, 0.010 g $Fe(NH_4)_2(SO_4)_2.6H_2O$
5. Elemental Sulfur (flowers of sulfur; powdered)
6. 2x ASW (*see* **Note 4**) (1 L): 40 g NaCl, 6 g $MgCl_2.6H_2O$, 12 g $MgSO_4.7H_2O$, 2 g $(NH4)_2SO_4$, 0.4 g $NaHCO_3$, 0.6 g $CaCl_2.2H_2O$, 1 g KCl, 0.840 g KH_2PO_4, 0.100 g NaBr, 0.040 g $SrCl_2.6H_2O$, 0.020 mg $Fe(NH_4)_2(SO_4)_2.6H_2O$
7. Gelzan™ CM
8. Polysulfides solution (500x) (*see* **Note 5**) (15 mL): 10 g $Na_2S.9H_2O$, 3 g sulfur.
9. 0.8x ASW (1 L): 16 g NaCl, 2.4 g $MgCl_2.6H_2O$, 4.8 g $MgSO_4.7H_2O$, 0.800 g $(NH_4)_2SO_4$, 0.160 g $NaHCO_3$, 0.240 g $CaCl_2.2H_2O$, 0.400 g KCl, 0.336 g KH_2PO_4, 0.040 g NaBr, 0.016 g $SrCl_2.6H2O$, 0.008 mg $Fe(NH_4)_2(SO_4)_2.6H_2O$

10. 20 amino acid solution (20x) (*see* **Note 6**) (200 mL): 1 g cysteine, 1 g glutamic acid, 1 g glycine, 0.500 g arginine, 0.500 g proline, 0.400 g asparagine, 0.400 g histidine, 0.400 g isoleucine, 0.400 g leucine, 0.400 g lysine, 0.400 g threonine, 0.400 g tyrosine, 0.300 g alanine, 0.300 g methionine, 0.300 g phenylalanine, 0.300 g serine, 0.300 g tryptophan, 0.200 g aspartic acid, 0.200 g glutamine, 0.200 g valine.
11. 19 amino acid solution lacking tryptophan (20x) (*see* **Note 6**) (200 mL): 1 g cysteine, 1 g glutamic acid, 0.500 g arginine, 0.500 g proline, 0.400 g asparagine, 0.400 g histidine, 0.400 g isoleucine, 0.400 g leucine, 0.400 g lysine, 0.400 g threonine, 0.400 g tyrosine, 0.300 g alanine, 0.300 g methionine, 0.300 g phenylalanine, 0.300 g serine, 0.200 g aspartic acid, 0.200 g glutamine, 0.200 g valine.
12. 100 µM 6-Methylpurine
13. 25 mM mevinolin
14. *T. kodakarensis* solid complete media (Tk-ASW-YT): 0.5 g yeast extract, 0.5 g tryptone, 1.0 g Gelzan™, 500 µL Wolfe's Trace minerals (200x), 200 µL polysulfide solution (500x), 5.0 mL 20 amino acid mixture (20x), 100 µL KOD vitamins (1000x), 50 mL H_2O, 50 mL 2x ASW
15. *T. kodakarensis* solid minimal media (Tk-ASW-min): 500 µL Wolfe's trace minerals (200x), 1.0 g Gelzan™, 50 mL H_2O, 50 mL 2x ASW, 200 µL polysulfide solution (500x), 100 µL KOD vitamins (1000x), 5.0 mL 20 amino acid mixture (20x)
16. *T. kodakarensis* solid minimal 19 amino acid media (TK-ASW-min-(-Trp)): 500 µL Wolfe's trace minerals (200x), 1.0 g Gelzan™, 50 mL H_2O, 50 mL 2x ASW, 200 µL polysulfide solution (500x), 100 µL KOD vitamins (1000x), 5.0 mL 19 amino acid mixture (20x).
17. *T. kodakarensis* solid minimal mevinolin media (TK-ASW-min-mev): 500 µL Wolfe's trace minerals (200x), 1.0 Gelzan™, 50 mL H_2O, 50 mL 2x ASW, 200 µL polysulfide solution (500x), 100 µL KOD vitamins (1000x), 12.5 µM mevinolin, 5.0 mL 20 amino acid mixture (20x).

2.1.3 Equipment

1. 1 mL syringe
2. Microcentrifuge
3. 125 mL serum bottles
4. Large aluminum pot with loose-fitting lid
5. Autoclave
6. Anaerobic chamber (Coy Labs)
7. 10% hydrogen, 90% nitrogen (v/v) compressed gas mix
8. 100% nitrogen compressed gas
9. Glass Petri plates (*see* **Note 7**)
10. Plastic Petri plates (*see* **Note 8**)
11. Floor model centrifuge and compatible rotor(s)
12. Thermal cycler
13. GasPak EZ Anaerobe Container System
14. Pipettes
15. Enzyme Cooler, Isotherm System
16. Forced air incubator (37 °C and 85 °C)
17. Dry block heater
18. Anaerobic canister
19. 1.7 mL microcentrifuge tubes
20. 0.2 mL PCR tubes
21. Two-leg lyophilization serum bottle septums
22. 20 mm aluminum seals
23. 20 mm Crimper, Standard Seal
24. 20 mm Decapper
25. Polycarbonate centrifuge tubes

26. Cell spreader

27. 10 mL serum bottles

28. Face shields, lab coats, nitrile gloves, and autoclave gloves

29. Paper towels

30. *T. kodakarensis* strain TS559[23] (*see* **Note 9**)

31. Non-replicative plasmid pTS700 (*see* **Note 9**)

32. Autonomously replicating plasmid pLC70 (*see* **Note 9**)

33. ZR Plasmid Miniprep™ Classic Kit

34. Qubit™ ds DNA BR Assay Kit

35. NucleoSpin Gel and PCR Clean-up, Mini kit

36. In-Fusion® Snap Assembly

37. Agilent Quikchange-II Kit

3. Methods

3.1 Microbial Methods.

T. kodakarensis transformation techniques are reliant on plasmid DNAs isolated from *Escherichia coli*. Standard growth and DNA preparations from *E. coli* yield plasmid DNA of sufficient quality for successful transformations of *T. kodakarensis*. Details are provided to generate i) non-replicative plasmids, based on pTS700, for genomic modifications and ii) autonomously replicating plasmids, based on pLC70, for ectopic expression, in Subheading 3.2.1 and 3.3.1, respectively. The choice of *E. coli* strain has minimal impacts on successful plasmid isolations.

3.1.1 *E. coli* media and cultivation

E. coli LB medium containing ampicillin: Ec-LB-Amp

1. Prepare Ec-LB-Amp and dispense 5 mL aliquots to sterile culture tubes.

2. Inoculate cultures with plasmid containing *E. coli* strains, incubate at 37 °C for 12-16 hours with agitation (~200 rpm).

E. coli LB plate medium with ampicillin: Ec-LB-Amp plate

1. Prepare Ec-LB-Amp plate media, autoclave, cool to ~50-55 ˚C (*see* **Note 10**), and aliquot ~25 mL per plastic petri plate (*see* **Note 11**) and allow to solidify at room temperature. Plates should be inverted and left to partially dry overnight before storing in a sealed bag at 4 °C.

3.1.2 *Thermococcus kodakarensis* media and cultivation

T. kodakarensis strains are typically grown in 100 mL aliquots in nutrient-rich media (ASW-YT) (*see* **Note 12**) to provide large numbers of cells for transformations. Strain TS559 must be supplemented with agmatine even when grown in nutrient-rich media. Smaller (~3-5 mL) cultures of *T. kodakarensis*, prepared with the same media formulations, suffice to yield sufficient DNA from transformants for use in diagnostic PCR to confirm the genotypes of transformed cells.

1. Prepare *T. kodakarensis* complete liquid medium by combining ASW-YT and Wolfe's Trace Minerals inside an anaerobic chamber (*see* **Note 13**). Aliquot 100 mL of mineral-supplemented ASW-YT media into 125 mL serum bottles, then seal with septums and aluminum seals, autoclave (*see* **Note 14)**, and store sterile media at room temperature.
2. Immediately prior to inoculation with *T. kodakarensis* cultures, and inside an anaerobic chamber, add 0.2 g of elemental sulfur and 100 µL KOD vitamins (1000x). When appropriate, add 100 µL 1 M agmatine sulfate solution (1000x).
3. Using a syringe, withdraw 1 mL of an active culture of TS559 (or other *T. kodakarensis* strain), inject the inoculum into the now completely supplemented media, seal the culture with a septum and aluminum seal (*see* **Note 15**), and incubate at 85 °C without agitation

for ~13 hours (cultures entering stationary phase yield the highest percentage of transformants).

Thermococcus kodakarensis complete solid medium: Tk-ASW-YT plate

Clonal populations of transformants from the transformations must be selected on solid media. Special protocols are necessary to generate solid media that remains solid at 85 ˚C; solid media is prepared from two independently autoclaved solutions that are mixed and rapidly distributed (~20-30 seconds) into glass petri plates before setting at room temperature. Depending on the transformation procedure and selective pressures applied, transformed cells are typically plated on nutrient-rich solid media lacking agmatine supplementation (for genomic modifications) or minimal, 19 amino acid-based solid media containing agmatine supplementation (for ectopic modifications).

Nutrient-rich *T. kodakarensis* solid media: Tk-ASW-YT plate

1. Inside an anaerobic chamber, combine 50 mL 2x ASW, 0.5 mL Wolfe's Trace Minerals (200x), 0.5 g tryptone, and 0.5 g yeast extract in a 125 mL serum bottle. Seal the bottle with a septum and aluminum seal, then sterilize via autoclaving.
2. Prepare a second 125 mL serum bottle by mixing 50 mL water and 1 g Gelzan™ CM. Seal the bottle with a septum and aluminum seal, then sterilize via autoclaving.
3. Immediately after autoclaving, bring both halves of the media formulations inside the anaerobic chamber. Remove septums and aluminum seals from both bottles. To the bottle containing 2x ASW add 100 µL KOD vitamins (1000x) and 200 µL polysulfides (500x) (*see* **Note 16**), then immediately combine with the bottle containing the dissolved Gelzan (*see* **Note 17**). Swirl to mix then aliquot the contents into 4 glass petri plates (~25 mL/plate). Plates will solidify within seconds and can be used immediately or stored inverted, within the anaerobic chamber at room temperature for 1-2 days (*see* **Note 18**).

Thermococcus kodakarensis minimal 20 amino acid solid medium: Tk-ASW-min plate

1. Inside an anaerobic chamber, combine 50 mL 2x ASW and 0.5 mL Wolfe's Trace Minerals (200x) in a 125 mL serum bottle. Seal the bottle with a septum and aluminum seal, then sterilize via autoclaving.
2. Prepare a second 125 mL serum bottle by mixing 50 mL water and 1 g Gelzan™ CM. Seal the bottle with a septum and aluminum seal, then sterilize via autoclaving.
3. Immediately after autoclaving, bring both halves of the media formulations inside the anaerobic chamber. Remove septums and aluminum seals from both bottles. To the bottle containing 2x ASW add 100 µL KOD vitamins (1000x), 100 µL 1M agmatine sulfate solution (1000x), 5 mL 20 amino acid solution (20x), 100 µL 100 µM 6-Methylpurine or 50 µL 25 mM mevinolin and 200 µL polysulfides (500x), then immediately combine with the bottle containing the dissolved Gelzan. Swirl to mix, then aliquot the contents into 4 glass petri plates (~25 mL/plate). Plates will solidify within seconds and can be used immediately or stored inverted, within the anaerobic chamber at room temperature for 1-2 days (*see* **Note 18**).

Thermococcus kodakarensis minimal 19 amino acid solid medium lacking tryptophan: Tk-ASW-min-(-Trp) plate

1. Inside an anaerobic chamber, combine 50 mL 2x ASW and 0.5 mL Wolfe's Trace Minerals (200x) in a 125 mL serum bottle. Seal the bottle with a septum and aluminum seal, then sterilize via autoclaving.
2. Prepare a second 125 mL serum bottle by mixing 50 mL water and 1 g Gelzan™ CM. Seal the bottle with a septum and aluminum seal, then sterilize via autoclaving.
3. Immediately after autoclaving, bring both halves of the media formulations inside the anaerobic chamber. Remove septums and aluminum seals from both bottles. To the bottle containing 2x ASW add 100 µL KOD vitamins (1000x), 100 µL 1M agmatine sulfate solution (1000x), 5 mL 19 amino acid solution (20x), 100 µL 100 µM 6-Methylpurine and 200 µL polysulfides (500x), then immediately combine with the bottle

containing the dissolved Gelzan. Swirl to mix then aliquot the contents into 4 glass petri plates (~25 mL/plate). Plates will solidify within seconds and can be used immediately or stored inverted, within the anaerobic chamber at room temperature for 1-2 days (*see* **Note 18**).

3.2 Design and construction of plasmid DNAs used to transform *Thermococcus kodakarensis.* We first detail the steps to generate pTS700-based plasmids that permit desired modifications to the genome of TS559 (*see* **Fig. 2.1, 2.2, and 2.4**). pTS700 contains a unique SwaI cutsite that permits linearization and insertion of amplicons (*see* **Fig. 2.1A**). Given that many genomic targets are ultimately targeted both for deletion and modification, it is practical to generate a common plasmid (termed an "A"-plasmid; *see* **Fig. 2.1**) from which two unique plasmids can be generated, one for deletion of the target locus (termed a "B"-plasmid; *see* **Fig. 2.2B**) and another for modification of the target locus (termed either a "C"- or "D"-plasmid; *see* **Fig. 2.2 C/D**). Standard plasmid preparations of "B"-, "C"-, and "D"-plasmids from most *E. coli* strains yield high-quality constructs that can be transformed into TS559 (*see* **Fig. 2.4**) to first generate transformants with intermediate genomes that can be resolved to generate the desired final genomic modifications (*see* **Fig. 2.4B-D**).

Following the use of pTS700-based vectors for genomic modifications, we then detail the use of autonomously replicating plasmids, based on the pLC70 vector, to generate transformants wherein selectable phenotypes are based on retention of and expression from ectopic vectors (*see* **Fig. 2.3**).

3.2.1 Generating linearized pTS700 to accept amplicons for genomic modifications. Construction of "A"-plasmids.

1. Linearize ~2 µg of pTS700 plasmid using SwaI endonuclease, resolve the reactions via agarose gel electrophoresis, and purify the linear products using the NucleoSpin Gel

extraction protocols (*see* **Fig. 2.1A**). Recovered products are quantified using the Qubit™

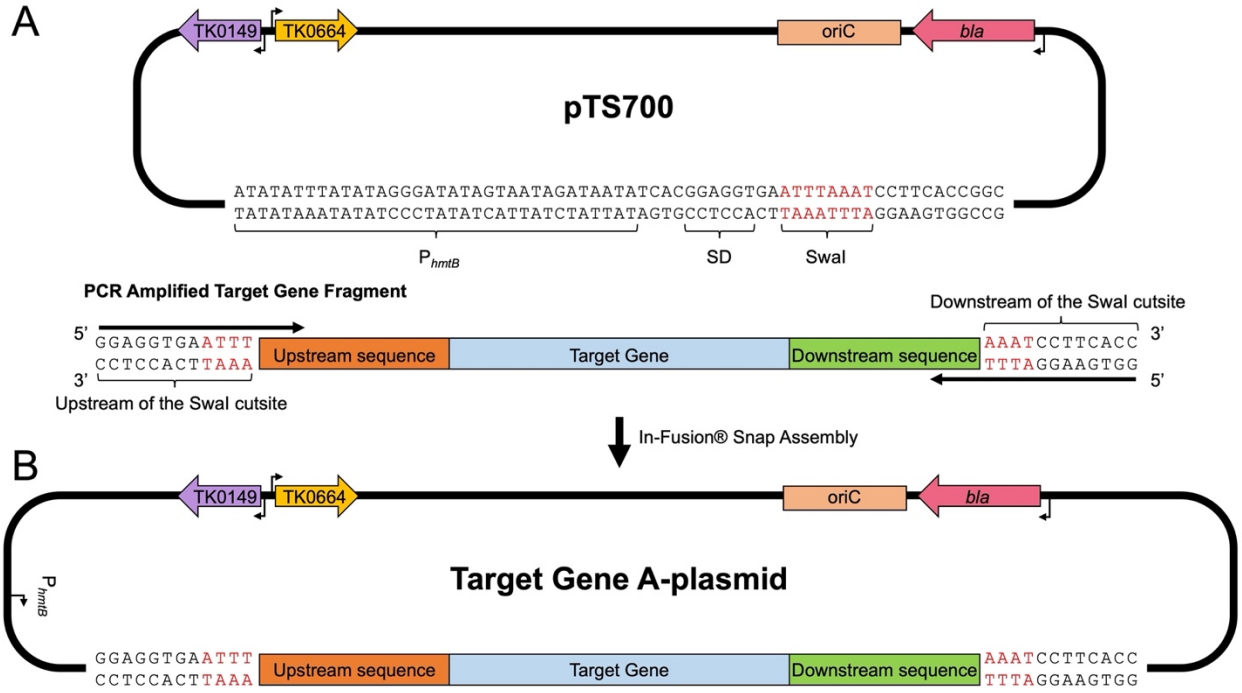


Figure 2.1. Generating an A-plasmid for the targeted *T. kodakarensis* genomic locus. (**A**, top) The pTS700 plasmid encoding TK0149, an agmatine prototrophic selectable marker (purple left arrow), TK0664, a hypoxanthine-guanine-phosphoribosyltransferase, that serves as a 6-methylpurine sensitive counter selectable marker (orange right arrow), oriC, bacterial plasmid origin of replication (salmon box), and β-lactamase (*bla*), ampicillin resistance gene (pink left arrow). The strong constitutive archaeal promoter (P_{hmtB}) and the Shine-Dalgarno sequence (SD) are included to promote expression in instances where recombination of the plasmid into the genome of *T. kodakarensis* displaces genes from their native promoters. (**A**, bottom) The PCR amplified target gene fragment contains 12 bp sequences complementary to both the upstream and downstream regions of the SwaI cutsite in pTS700. The amplicon contains ~700 bp of upstream sequences (red box), ~700 bp of downstream sequences (green box), and the target gene sequence (blue box). (**B**) In-Fusion cloning results in production of a complete A-plasmid for the target gene.

dsDNA BR Assay Kit. SwaI-linearized pTS700 can be prepared in bulk, purified and stored at -20 °C for use in construction of “A”-plasmids for multiple genomic targets. (*see* **Note 19**)

2. Generate a PCR-amplicon, using *T. kodakarensis* genomic DNA as a template, containing the target gene with ~700 bp of upstream and downstream genomic sequences (*see* Fig. 1A). The primer complementary to sequences upstream of the target locus should begin with the sequence, 5' GGAGGTGAATT (*see* **Note 20**) followed by ~ 25 nucleotides of complementarity to the genome. The primer complementary to sequences downstream of the target locus should begin with the sequence, 5' GGTGAAGGATT (*see* **Note 20**) followed by ~ 25 nucleotides of complementarity to the genome. Resolve the reactions via agarose gel electrophoresis, and purify the linear products using the NucleoSpin Gel extraction protocols. Recovered products are quantified using the Qubit™ dsDNA BR Assay Kit.
3. Use In-Fusion® Snap Assembly to insert the target gene fragment into pTS700-SwaI in a 2:1 ratio (gene fragment : linearized vector) (*see* **Fig. 2.1B**). Transform 2.5 µL of the In-Fusion® Snap Assembly reaction into Stellar (*see* **Note 21**) competent cells and plate the cells on Ec-LB-Amp plate. Incubate at 37 °C overnight (12-14 hours). Successful insertion of the target gene into pTS700 is typically confirmed by colony PCR using primers specific to pTS700 that flank the SwaI cutsite (*see* **Note 22**). Once colonies containing the presumptive "A"-plasmid are identified, they are grown in 5 mL Ec-LB-Amp overnight at 37 ˚C with shaking. 5 mL Ec-LB-Amp grown cultures are sufficient to yield sufficient plasmid DNAs first for sequencing and ultimately for transformation into *T. kodakarensis*.
4. Purify the target gene A-plasmid from the *E. coli* cells using ZR Plasmid Miniprep™ - Classic and quantify the concentration of the plasmid using Qubit™ dsDNA BR Assay Kit. Store at -20 °C in a freezer. The full sequence of the amplicon should be confirmed via Sanger sequencing (or alternative sequencing techniques) prior to transformation into *T. kodakarensis.*

3.2.2. Generating “B”- and “C”-plasmids from “A”-plasmids

Mutagenic PCR for deletion of the target gene: generation of the B-plasmid (*see* **Fig. 2.2A-B**).

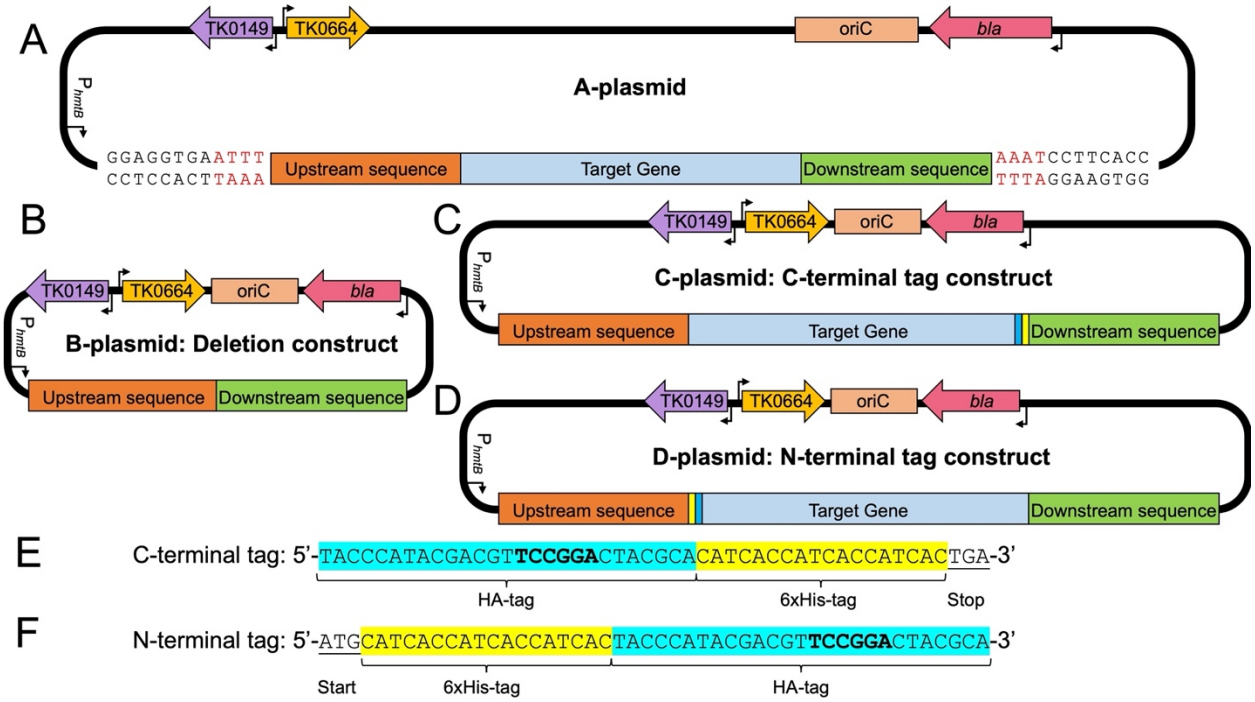


Figure 2.2. Design of integrative vectors to delete or modify genomic targets in *Thermococcus kodakarensis*. (**A**) The parental A-plasmid containing ~700 bp of upstream sequence (red box), the target gene (blue box), and ~700 bp of downstream sequence (green box). (**B**) B-plasmids are used to generate deletion strains of *T. kodakarensis*. The B-plasmid is generated from the A-plasmid by deleting the sequences encoding the target gene while retaining the upstream- and downstream-sequences. (**C** & **D**) The C- and D-plasmids are used to generate strains of *T. kodakarensis* wherein the genomic target locus is extended to encode for HA- and 6xHis-tags. The C-plasmid is generated from the A-plasmid by the inclusion of sequences encoding the HA-tag (cyan box) followed by the 6xHis-tag (yellow box) before the stop codon. The D-plasmid is generated from the A-plasmid by the inclusion of sequences encoding the 6xHis-tag (yellow box) followed by the HA-tag (cyan box) after the start codon. (**E** & **F**) The nucleotide sequences and positions of the C- and N-terminal tags highlighting the introduced BspEI sites (bold).

1. Design a primer pair that permit the deletion of the target gene sequence for use in Quikchange-II reactions. Primers typically are ~60 nt in length, with 30 nt of complementarity to sequences immediately upstream and downstream of the target locus. Use the primer pairs in a Quikchange-II reaction using the A-plasmid as the

template DNA following manufacturer's instructions, inclusive of the transformation into *E. coli* and plating of Ec-LB-Amp plates. Allow colonies to form during overnight incubation at 37 ˚C. As an alternative, the target gene sequence can be deleted via inverse PCR and ligation of the linear product.

2. The success of removing the target gene sequence can be evaluated via colony PCR as in Subheading 3.2.1, step 3. Once *E. coli* clones harboring presumptive B-plasmids are identified, growth of small cultures, plasmid recovery and sequencing should be completed as in Subheading 3.2.1, step 4. Plasmids can be stored at -20 ˚C for extended periods prior to transformation into *T. kodakarensis*.

Mutagenic PCR for tagging of the target gene: generation of C-plasmid and D-plasmids (*see* **Fig. 2.2C-F**).

C-plasmids are generated by addition of sequences that encode HA- and 6xHis-tags immediately prior to the stop codon of the target gene (see **Fig. 2.2C and 2.2E**). D-plasmids are generated by addition of sequences that encode HA- and 6xHis-tags immediately downstream of the start codon of the target gene (see **Fig. 2.2D and 2.2F**).

1. Design a primer pair for use with Quikchange-II reactions that permit the addition of the tag-encoding sequences to either the start or end of the target genes. Primers are typically ~95 nt in length, with 25 nt of complementarity immediately upstream and downstream of the insertion site (50 nt total) and an additional 45 nt encoding the 9 amino acid HA-tag and 6xHis- tags. Use the primer pairs in Quikchange-II reactions using the A-plasmid as the template DNA following manufacturer's instructions, inclusive of transformation into *E. coli* and plating of Ec-LB-Amp plates. Allow colonies to form during overnight incubation at 37 ˚C.
2. The success of adding sequences to extend the target gene at the 5' or 3' end can be evaluated via colony PCR as in Subheading 3.2.1, step 3 using primer pairs that flank

the site of the sequence insertion. Amplicons from desirable C- or D-plasmids harbored in *E. coli* transformants are ~45 bp longer than amplicons generated from A-plasmid that lack the tag-encoding sequences. Once *E. coli* clones harboring presumptive C- and D-plasmids are identified, growth of small cultures, plasmid recovery and sequencing should be completed as in Subheading 3.2.1, step 4. Plasmids can be stored at -20 ˚C for extended periods prior to transformation into *T. kodakarensis*. As an alternative method to identify successful addition of tag-encoding sequences, amplicons generated with primers that flank the site of insertion will generate a product that has a new BspEI endonuclease recognition site that will permit identification of C- and D-plasmids from A-plasmids following digestion of the amplicons with BspEI.

3.3.1. Generating pLC70-based expression plasmids.

pLC70 contains selectable markers to phenotypically select desired transformants (*see* **Fig. 2.3A**). Typically, pLC70 vectors are modified to express genes of interest by insertion of an expression cassette – a *T. kodakarensis* promoter and associated open reading frame – into pLC70 prior to transformation into *T. kodakarensis* strains (*see* **Fig. 2.3B-D**). The expression cassette (*see* **Fig. 2.3C**) can be constructed through various procedures but should ultimately be cloned into pLC70 taking advantage of unique SalI and NotI sites. Cloning procedures similar those detailed for amplicon insertion to pTS700 yield consistent results.

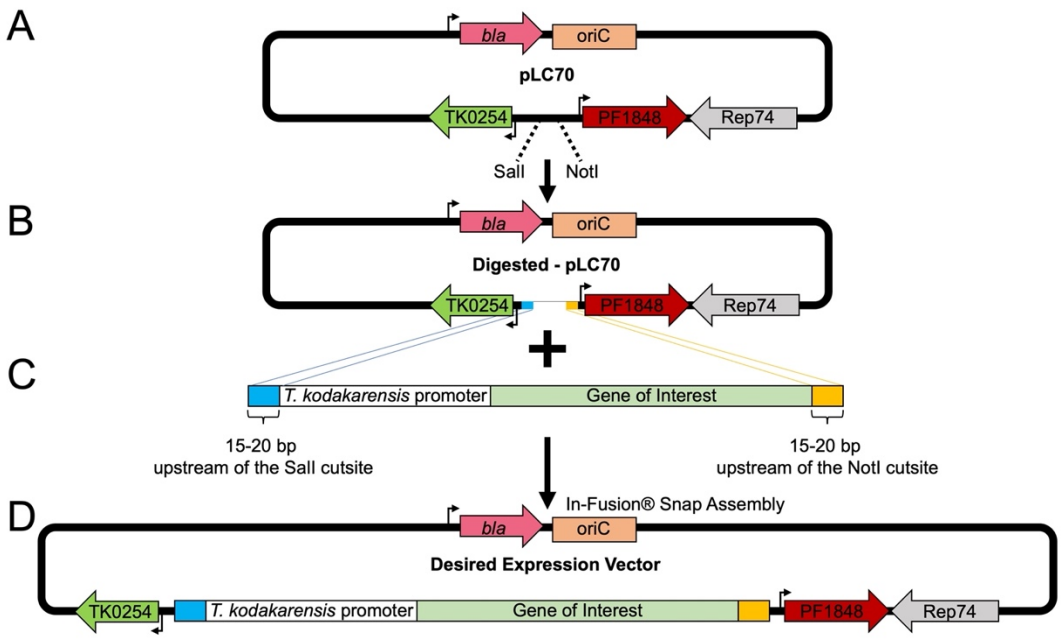


Figure 2.3. The autonomously-replicating pLC70 vector permits ectopic gene expression in *Thermococcus kodakarensis*. (**A**) The pLC70 vector contains the β-lactamase (*bla*) ampicillin resistance gene (pink left arrow), the bacterial origin of replication, oriC (salmon box), the archaeal DNA replication protein Rep74 (grey left arrow), the *Pyrococcus furiosus* PF1848 gene conferring resistance to statin-based antibiotics (red right arrow), and the TK0254 gene which provides a tryptophan prototrophic selectable marker (green left arrow). (**B**) The pLC70 vector is linearized using SalI and NotI endonucleases to accept expression cassettes. (**C**) An expression cassette composed of the sequence for a *T. kodakarensis* promoter (white box), sequences for the target gene of interest (sage box), and terminal flanking sequences homologous to regions near the SalI cutsite (blue box) and the NotI cutsite (orange box) for insertion into the linearized pLC70 vector. (**D**) The complete vector with inserted expression cassette can be transformed into *T. kodakarensis* and selected based on restoration of tryptophan prototrophy and resistance to statin-based antibiotics.

3.4.1. *Thermococcus kodakarensis* strain construction

Initial transformation to confirmation of final strains typically takes 2-3 weeks

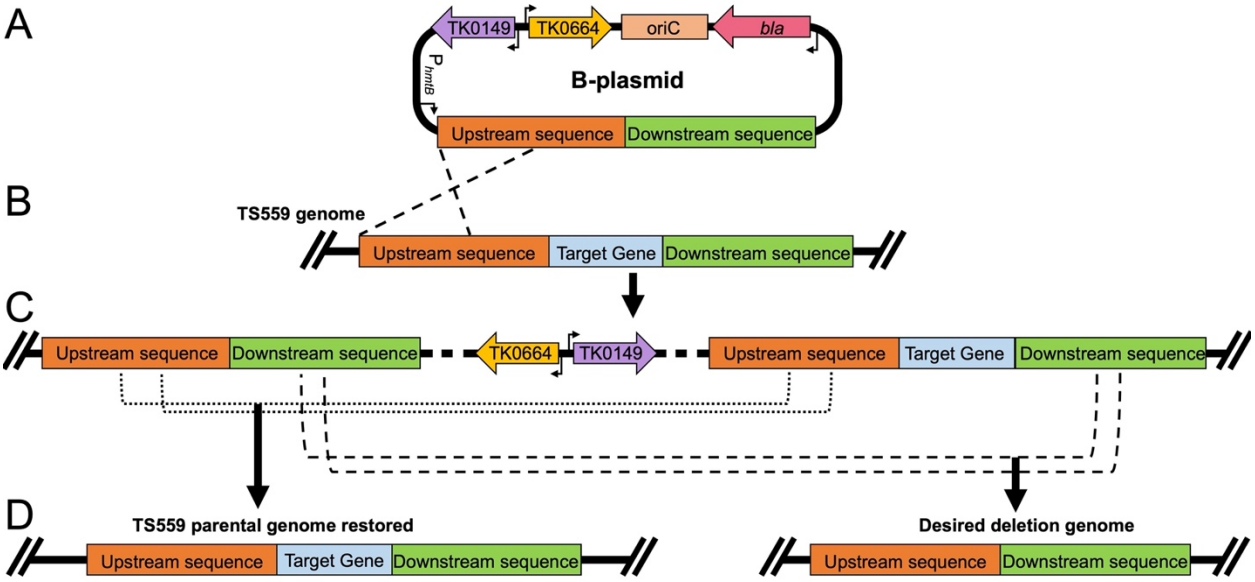


Figure 2.4. Example of target gene locus modification in *Thermococcus kodakarensis* using a B-plasmid to generate a deletion strain. (**A**) The sequence-confirmed B-plasmid of the target gene is transformed into the parental strain (TS559) and integrates into the chromosome via homologous recombination. (**B**) Homologous recombination (black dashed lines) of the B-plasmid into the parental (TS559) genome integrates the entire B-plasmid into the genome, resulting in restoration of agmatine prototrophy and introduction of 6-methylpurine sensitivity. Recombination through upstream sequences (red boxes) is shown, however it is equally probable that the B-plasmid will recombine with the genome via homologous recombination through downstream sequences (green boxes). (**C**) The genotype of the resulting intermediate strain is shown, highlighting the duplication of upstream- and downstream-sequences that ultimately permit the spontaneous excision of the integrated plasmid sequences. Transformants containing the intermediate genome are selected based on agmatine prototrophy. Intermediate strains whose genotype is confirmed by diagnostic PCR are grown with agmatine supplementation to permit for spontaneous excision of the B-plasmid sequence, then plated on media containing 6-methylpurine to select for transformants that have lost TK0664. (**D**) Excision of the B-plasmid from the genome can result in two potential homologous recombination events: upstream recombination (black dotted line) resulting in the restoration of the parental (TS559) genome, or downstream recombination (black dashed line) resulting in the desired deletion of the target locus in the *T. kodakarensis* genome. 6-methylpurine resistant strains are picked to liquid medium (ASW-YT + Agm) and their genotypes are confirmed via diagnostic PCR or whole genome sequencing.

1. Grow strain TS559 in ASW-YT-Agm as directed in Subheading 3.1.2. Incubate at 85 °C for 12-13 hours.
2. Harvest TS559 biomass by anaerobically transferring the culture into high-speed centrifuge tubes followed by centrifugation at ~18,500 *x g* for 10 minutes at 4°C. (*see* **Note 23**) Inside the anaerobic chamber, carefully decant the supernatant and keep the undisturbed pellet.
3. Directly in the anaerobic chamber, gently resuspend the TS559 cell pellet in 3 mL of ice-cold 0.8x ASW. ~200 µL of resuspended cells are typically used per transformation, and by preparing several aliquots, multiple transformations can be performed simultaneously. Aliquot 200 µL of the resuspended cells into 1.7 mL microcentrifuge tubes and incubate anaerobically on ice for 30 minutes. (*see* **Note 24)**
4. Add ~2-4 µg of the transformation plasmid (either a pTS700- or pLC70-based plasmid) to the 200 µL of incubated cells and extend incubation anaerobically on ice for an additional 60 minutes.
5. Heat shock cells at 85 °C for 45 seconds in a dry heat block. Immediately chill on ice for 10 minutes following heat shock (*see* **Note 25**).
6. Carefully spread the now transformed cells to Tk-ASW-YT plates (for pTS700 based plasmids) or Tk-ASW-min-(-Trp) plates (for pLC70 based plasmids). Gently spread cells with a sterile cell spreader inside the anaerobic chambers. Once the transformation media has been absorbed into the plates, invert the plates and transfer to an air-tight anaerobic metal canister, packed with paper towels (*see* **Note 26**) and a Gas Pak EZ anaerobic container system packet. Seal the canister inside of the anaerobic chamber, then remove and incubate the sealed vessel in an 85 °C incubator for 48-96 hours.
7. Return the sealed anaerobic vessel to the anaerobic chamber and remove the incubated plates, being careful to remove condensation. Identify colonies formed from successful transformations; *T. kodakarensis* colonies are nearly transparent and may be difficult to

identify at first (*see* **Note 27**). Transformants from the transformation with pLC70-based plasmids are likely to contain the desired final genotype and phenotype but are typically re-plated on Tk-ASW-min-mev plates to ensure the transformants are resistant to mevinolin before being transferred to liquid media for experimental use (*see* **Fig. 2.3**). In contrast, colonies resultant from transformation with pTS700-based plasmids only contain an intermediate genome (*see* **Fig. 2.4C**) and require additional manipulations to generate the desired modified final genome.

8. Initial transformants from pTS700-based transformations that contain the intermediate genome wherein the entire pTS700-plasmid has integrated via homologous recombination into the TS559 genome must be picked from the plates and used to inoculate 3 ml Tk-ASW-YT liquid cultures containing agmatine in 5 mL serum bottles. Seal the serum bottles anaerobically using septums and aluminum seals, followed with an overnight (12-13 hours) incubation at 85 °C. Growth in the presence of agmatine permits the spontaneous excision of the integrated vector (*see* **Fig. 2.4D**). Regardless of the nature of the excision event, the resulting genomes will lack both TK0149 and TK0664, rendering cells agmatine auxotroph that are resistant to 6-methylpurine. If the excision event captures the target gene, the resulting cells will also lack the target gene (for B-plasmid based transformations) or encode a modified target gene sequences (for C- and D-plasmid based transformations) (*see* **Fig. 2.4D**).
9. 1 mL of overnight ASW-YT+Agm grown intermediate cultures should be anaerobically harvested, concentrated via centrifugation 5-fold, then plated on Tk-ASW-min plates containing 6-methylpurine. Gently spread cells with a sterile cell spreader inside the anaerobic chambers. Invert plates with transformants, transfer to an air-tight anaerobic metal canister, packed with paper towels (*see* **Note 26**) and a Gas Pak EZ anaerobic container system packet. Seal the canister inside the anaerobic chamber, then remove and incubate the sealed vessel in an 85 °C incubator for 48-96 hours.

10. Return the sealed anaerobic vessel to the anaerobic chamber and remove the incubated plates, being careful to remove condensation. Identify colonies formed from successful excision events from the genome. Colonies must be picked from the plates and used to inoculate 3 mL Tk-ASW-YT liquid cultures containing agmatine in 5 mL serum bottles. Seal the serum bottles anaerobically using septums and aluminum seals, followed with an overnight (12-13 hours) incubation at 85 °C.
11. 1 mL of overnight cultures is anaerobically removed to prepare genomic DNA for diagnostic PCRs and sequencing to identify strains wherein the excision event resulted in generation of the desired genotype. Sequencing of PCR amplicons provides nucleotide level confidence of the desired modifications. We also recommend whole genome sequencing of all strains to ensure that spontaneous modifications were not accidentally introduced elsewhere in the genome.

4. Notes

1. This solution is light-sensitive and should be protected by either wrapping a transparent tube in aluminum foil, or by storing in an opaque, amber conical tube.
2. *T. kodakarensis* requires casein peptone that is enzymatically digested using pancreatic enzymes. Other sources of tryptone are suitable for *E. coli* media.
3. AMRESCO, catalog number: J850; For *E. coli* media, any yeast extract is suitable, however *T. kodakarensis* requires this source of yeast extract.
4. 2x ASW can be stored outside of the anaerobic chamber.
5. Dissolve solution using heat, it will be a deep red color when both sulfur and sodium sulfide are completely dissolved.
6. Autoclaving will degrade this solution.
7. Plastic petri plates may melt at *T. kodakarensis* growth temperatures (85° C), requiring the use of glass petri plates.
8. Plastic petri plates will not melt at *E. coli* growth temperature (37 °C).

9. Please contact corresponding author to obtain plasmid. Details of the plasmid sequences can be found at references[16,18,26].
10. Ampicillin is unstable at temperatures above 50 °C, do not add before autoclaving.
11. There must be enough media to cover the bottom of the plate.
12. Water used in the preparation of ASW-YT should be boiled before it is combined with dry ingredients. Boiling releases dissolved oxygen.
13. ASW-YT must be prepared anaerobically.
14. Autoclave liquid cycle, 30 minutes sterilization time.
15. *T. kodakarensis* released H_2S and H_2 gasses as part of its metabolism. Using a septum alone to seal the bottle is inadequate, as the gasses produced will push the septum out of the bottle.
16. Polysulfides replace elemental sulfur in solid *T. kodakarensis* media.
17. Most selective agents for solid media should be added to the ASW solution after autoclaving. Supplemented ASW-based solutions should then be poured into the autoclaved gelzan solution before pouring plates. This order of combination ensures homogeneous solid media.
18. For pouring multiple sets of plates at once, keep autoclaved solutions on a hot plate to prevent premature solidification.
19. SwaI endonuclease creates a blunt cutsite, reducing the likelihood of the linearized vector reannealing to itself, allowing for stable long-term storage.
20. This sequence is homologous to regions near the SwaI cutsite.
21. This strain is the recommended cell line by Takara; Stellar cells do not contain a plasmid prior to transformation, increasing confidence in isolated products after transformation.

22. 700 Forward and 700 Reverse are the primers typically used in this reaction. Further details can be found at references[16,18,26].
23. Centrifuge bottles cannot be opened outside of the chamber. Ensure centrifuge bottles have a rubber seal in their cap to maintain anaerobicity. When transferring centrifuge bottles between the centrifuge and the anaerobic chamber, maintain the bottles in an inverted configuration to prevent the supernatant from washing over the pellet and reducing overall yield.
24. Excess cells cannot be stored for future transformations.
25. This step increases transformation efficiency.
26. Packing the anaerobic canister with paper towels reduces shifting of the canister's contents when being moved, reducing the risk of breaking a glass petri plate and compromising cell cultures. Additionally, condensation will build up in the canister during incubation, and paper towels serve to absorb some moisture in the canister.
27. *T. kodakarensis* colonies are transparent puncta on the surface of media; most colonies are no larger than 1 mm in diameter and appear similar to bubbles on the surface. Occasionally, salt precipitates from media will look like *T. kodakarensis* colonies on solid media.

REFERENCES

1. Martínez-Espinosa, R. M. Microorganisms and Their Metabolic Capabilities in the Context of the Biogeochemical Nitrogen Cycle at Extreme Environments. *International journal of molecular sciences* **21,** (2020).

2. Quehenberger, J., Shen, L., Albers, S.-V., Siebers, B. & Spadiut, O. Sulfolobus – A Potential Key Organism in Future Biotechnology. *Frontiers in Microbiology* **8,** 2474 (2017).

3. Belilla, J., Moreira, D., Jardillier, L., Reboul, G., Benzerara, K., López-García, J. M., Bertolino, P., López-Archilla, A. I. & López-García, P. Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area. *Nature Ecology and Evolution* **3,** 1552–1561 (2019).

4. Mayer, F. & Müller, V. Adaptations of anaerobic archaea to life under extreme energy limitation. *FEMS microbiology reviews* **38,** 449–72 (2014).

5. Poli, A., Finore, I., Romano, I., Gioiello, A., Lama, L. & Nicolaus, B. Microbial Diversity in Extreme Marine Habitats and Their Biomolecules. *Microorganisms* **5,** 25 (2017).

6. Tehei, M. & Zaccai, G. Adaptation to extreme environments: Macromolecular dynamics in complex systems. *Biochimica et Biophysica Acta - General Subjects* **1724,** 404–410 Preprint at https://doi.org/10.1016/j.bbagen.2005.05.007 (2005)

7. Efremov, A. K., Qu, Y., Maruyama, H., Lim, C. J., Takeyasu, K. & Yan, J. 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. *Journal of Biological Chemistry* **290,** 15770–15784 (2015).

8. Hegazy, G. E., Abu-Serie, M. M., Abo-Elela, G. M., Ghozlan, H., Sabry, S. A., Soliman, N. A. & Abdel-Fattah, Y. R. In vitro dual (anticancer and antiviral) activity of the carotenoids produced by haloalkaliphilic archaeon Natrialba sp. M6. *Scientific reports* **10,** 5986 (2020).

9. Patel, A. B., Shaikh, S., Jain, K. R., Desai, C. & Madamwar, D. Polycyclic Aromatic Hydrocarbons: Sources, Toxicity, and Remediation Approaches. *Frontiers in microbiology* **11,** 562813 (2020).

10. Cabrera, Ma. Á. & Blamey, J. M. Biotechnological applications of archaeal enzymes from extreme environments. *Biological Research* **51,** 37 (2018).

11. Crosby, J. R., Laemthong, T., Lewis, A. M., Straub, C. T., Adams, M. W. & Kelly, R. M. Extreme thermophiles as emerging metabolic engineering platforms. *Current Opinion in Biotechnology* **59,** 55–64 (2019).

12. Straub, C. T., Counts, J. A., Nguyen, D. M. N., Wu, C.-H., Zeldes, B. M., Crosby, J. R., Conway, J. M., Otten, J. K., Lipscomb, G. L., Schut, G. J., Adams, M. W. W. & Kelly, R. M. Biotechnology of extremely thermophilic archaea. *FEMS Microbiology Reviews* **42,** 543–578 (2018).

13. Dumorné, K., Córdova, D. C., Astorga-Eló, M. & Renganathan, P. Extremozymes: A potential source for industrial applications. *Journal of Microbiology and Biotechnology* **27,** 649–659 Preprint at https://doi.org/10.4014/jmb.1611.11006 (2017)

14. Kanai, T., Imanaka, H., Nakajima, A., Uwamori, K., Omori, Y., Fukui, T., Atomi, H. & Imanaka, T. Continuous hydrogen production by the hyperthermophilic archaeon, Thermococcus kodakaraensis KOD1. *Journal of biotechnology* **116,** 271–82 (2005).

15. Atomi, H., Sato, T. & Kanai, T. Application of hyperthermophiles and their enzymes. *Current opinion in biotechnology* **22,** 618–26 (2011).

16. Hileman, T. H. & Santangelo, T. J. Genetics techniques for Thermococcus kodakarensis. *Frontiers in Microbiology* **3,** 195 (2012).

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

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

19. Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S. & Imanaka, T. Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes. *Genome Research* **15,** 352–363 (2005).

20. Sato, T., Fukui, T., Atomi, H. & Imanaka, T. Improved and versatile transformation system allowing multiple genetic manipulations of the hyperthermophilic archaeon Thermococcus kodakaraensis. *Applied and environmental microbiology* **71,** 3889–99 (2005).

21. Sato, T., Fukui, T., Atomi, H. & Imanaka, T. Targeted gene disruption by homologous recombination in the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1. *Journal of bacteriology* **185,** 210–20 (2003).

22. Matsumi, R., Manabe, K., Fukui, T., Atomi, H. & Imanaka, T. Disruption of a sugar transporter gene cluster in a hyperthermophilic archaeon using a host-marker system based on antibiotic resistance. *Journal of Bacteriology* **189,** 2683–2691 (2007).

23. Santangelo, T. J., Čuboňová, L. & Reeve, J. N. Thermococcus kodakarensis Genetics: Tk1827-encoded β-glycosidase, new positive-selection protocol, and targeted and repetitive deletion technology. *Applied and Environmental Microbiology* **76,** 1044–1052 (2010).

24. Farkas, J. A., Picking, J. W. & Santangelo, T. J. Genetic techniques for the Archaea. *Annual Review of Genetics* **47,** 539–561 Preprint at https://doi.org/10.1146/annurev-genet-111212-133225 (2013)

25. Catchpole, R., Gorlas, A., Oberto, J. & Forterre, P. A series of new E. coli–Thermococcus shuttle vectors compatible with previously existing vectors. *Extremophiles* **22,** 591–598 (2018).

26. Santangelo, T. J., Čuboňová, L. & Reeve, J. N. Shuttle vector expression in Thermococcus kodakaraensis: Contributions of cis elements to protein synthesis in a hyperthermophilic archaeon. *Applied and Environmental Microbiology* **74,** 3099–3104 (2008).

CHAPTER 3: INTEIN-SPLICING CAN CONTROL ARCHAEAL DNA REPLICATION

Summary

Inteins, mobile genetic elements removed through splicing, often interrupt proteins required for DNA replication, recombination, and repair. An abundance of in vitro evidence implies inteins may act as regulatory elements, whereby reduced splicing inhibits production of the mature protein lacking the intein, but in vivo evidence of regulatory intein-excision is absent. The model archaeon Thermococcus kodakarensis encodes fifteen inteins and we, for the first time, establish the impacts of intein-splicing inhibition on host physiology and replication in vivo. We report that a decrease in intein-splicing efficiency of the recombinase RadA has widespread physiological consequences, including a general growth defect, increased sensitivity to DNA damage, and remarkably, a switch in the mode of DNA replication from recombination-dependent replication towards origin-dependent replication.

Introduction

The replisome is a collection of proteins that carry out DNA replication. Its functions and concentrations are regulated, intertwined, and coordinated to facilitate synthesis of both leading and lagging strands. Perhaps surprisingly, critical replication activities are often performed by domain-specific non-orthologous proteins[1,2] and are best epitomized by the use of evolutionary-unrelated DNA polymerase families (Pol C, Pol D, and Pol B) for replication of bacterial, archaeal, and eukaryotic genomes, respectively[3,4]. The complex systems that regulate the initiation of DNA replication are especially diverse and beguiling[3,5–13]. DNA replication in most species is controlled by domain-specific initiator protein complexes that assemble at specific locations in the genome termed origins of replication (*ori*). Decades of work demonstrate that origin-dependent replication (ODR) is the dominant mechanism of replication in each Domain[11].

In some Archaea, origin-sequences are necessary for replication and origin-recognition proteins (typically Cdc6) initiate replisome assembly by recruiting and helping load the MCM helicase[14,15]. ODR facilitates accurate replication of the genome, but the sophisticated regulatory strategies used to control the initiation of DNA replication were unlikely to be present in the first cells. Recombination-dependent replication (RDR) initiation is an alternative and effective strategy to replicate archaeal genomes[16]. *Thermococcus kodakarensis,* a hyperthermophilic Archaea, is naturally polyploid[17] and can replicate its genome independently of any *ori* and the initiator protein Cdc6[3,13], demonstrating that some modern species may preferentially initiate replication via recombination at many stochastic sites distributed around the genome. The evolutionary retention of predicted origin sequences and Cdc6 argue for selective usage of ODR versus RDR, but how ODR or RDR is selected as the replicative mechanism is unknown.

Under optimal growth conditions *T. kodakarensis* predominately utilizes RDR, but once removed from such, a more judicious use of resources may dictate that an entirely different replication strategy (e.g. ODR) be used to ensure long-term survival[13]. RDR requires the retention of multiple genomes and the activity of a recombinase; therefore, control of ploidy and/or the activities of recombinase proteins provide a plausible mechanism to switch between RDR and ODR. Archaea encode a Rad51- (eukarya), RecA- (bacteria) homologue termed RadA that initiates nucleoprotein filament formation and strand invasion[18–20] that can support DNA synthesis through recombination intermediates *in vitro*[21]. Abundant, active RadA is predicted to be necessary to establish the recombination intermediates that would permit RDR *in vivo*. Mechanisms that control the active levels of RadA in archaeal cells are critical, as changes to RadA protein levels are likely to alter replicative, recombination, and repair mechanisms *in vivo*. Archaeal loci encoding RadA are often interrupted by intein-encoding sequences that provide a potentially radical mechanism – via conditional intein splicing[22–35] – to control active RadA levels

in vivo. Inteins (intervening proteins) are mobile genetic elements (MGEs) spliced from host proteins following translation from a precursor protein to yield the isolated intein protein and ligated exteins that form mature protein[22,34–38].

Given that protein splicing (i.e. the totality of reactions involved in intein excision and extein-ligation) is known to be impacted by environmental conditions[22–35], we sought to understand the mechanism(s) controlling the production of mature RadA in *T. kodakarensis* and whether intein-splicing efficiency could dictate a foundational change in replicative strategy, with high RadA levels supporting RDR and low RadA levels supporting use of ODR. Given the volatile nature of hydrothermal environments, employing an environmentally controlled, intein-based protein switch to regulate DNA replication strategies might prove advantageous. Using otherwise isogenic strains of *T. kodakarensis* with either mutation to, or deletion of, the chromosomal intein sequence within RadA (TK1899), we demonstrate for the first time that differential RadA protein splicing can influence the physiology of an intein-containing host organism. Quantification of protein splicing efficiency *in vitro* and mature RadA levels *in vivo*, combined with marker-frequency analysis (MFA) to monitor potential origin usage at multiple growth temperatures that alter intein splicing efficiency, demonstrates that the degree of RadA protein splicing directs the choice of ODR versus RDR in *T. kodakarensis*. While RDR dominates under idealized growth conditions, mutations that result in inefficient Tk-RadA-intein splicing and in turn, diminished mature Tk-RadA levels, reduce recombination frequencies to tip replication preferences towards ODR.

Our results imply that conditional protein splicing of RadA – and, by inference, other intein-invaded replicative machinery components – is likely to shift the *in vivo* concentrations of mature, spliced factors that drive DNA synthesis, recombination, and repair throughout much of microbial life. An evolutionary shift from intein excision as an initial burden due to the parasitic

MGE invasion to an exaptated, regulated, and fitness-relevant controlled excision event offers regulatory functions to inteins beyond the originally evolved functions of spreading and splicing. It is likely that many intein-encoding species have exaptated spontaneous-parasitic intein splicing events to regulatory-splicing events that assist in the complex mechanisms controlling key features of recombination, repair, and replication strategies.

Results

Tk-RadA-intein houses an active homing endonuclease

Exhaustive attempts to delete Tk-RadA (TK1899) from the *T. kodakarensis* genome repeatedly failed, as did initial attempts to delete the intein-encoding sequences (aa 150-633) within Tk-RadA^{WT}. So called full-length inteins, such as the intein sequence within Tk-RadA, encode autonomous homing endonuclease (HEN) domains between conserved sequences required for protein splicing; inteins without HEN domains are often termed mini-inteins (Figure 3.1A). The HEN domain allows for intein mobility at the DNA-level by generating double-stranded DNA breaks in intein-less alleles to drive intein spreading via allelic-conversion repair with intein-containing alleles.

We reasoned that our failure to delete the sequences encoding the RadA-intein from the *T. kodakarensis* chromosome was due to an active HEN domain cleaving the intein-minus allele prior to recombination. As expected, preparations of Tk-RadA^{WT} did not cleave the TK1899^{WT} target DNA as the intein sequence was intact; however, it demonstrated robust HEN activity against a PCR amplified DNA fragment containing the intein-less allele of TK1899$^{\Delta intein}$ (Figure 3.1B) and the closely related intein-less RadA encoding-sequence from *Pyrococcus furiosus* (Pf1926) (Figure 3.2C). The active site of the putative HEN domain within Tk-RadA^{WT} was predicted based on homology with other HEN-containing inteins[39] at residues 373-381. When

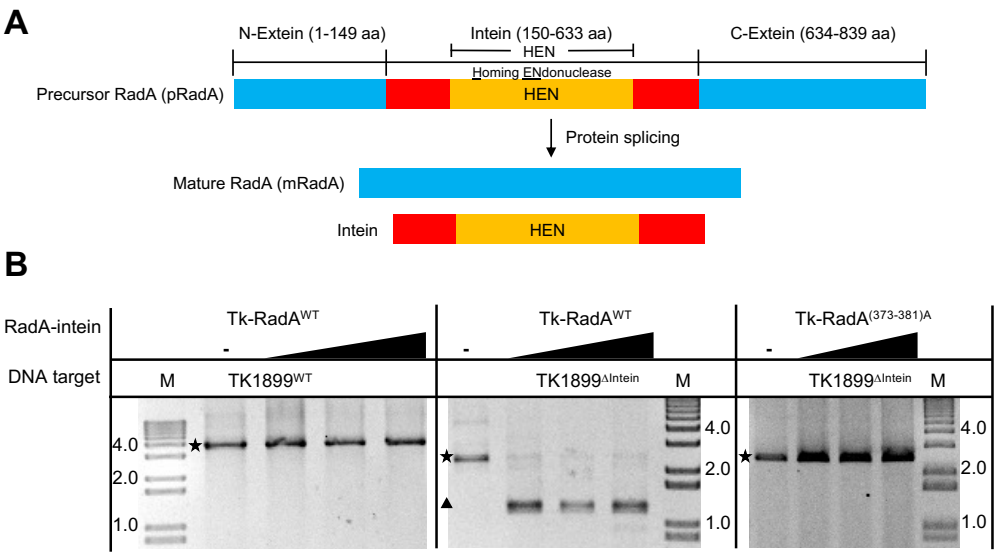


Figure 3.1. Tk-RadA encode an intein with an active HEN domain.
(A) Schematic representations of precursor RadA (pRadA, top) splicing to form both intein and mature RadA (mRadA). The precursor RadA is composed of N-extein, intein, and C-extein denoted in blue, red/orange, and blue, respectively. The region of the intein containing the homing endonuclease (HEN) domain is shown in orange. (**B)** *In vitro* RadA-mediated DNA cleavage demonstrates HEN activity targets intein-less alleles of TK1899. Tk-RadA^{WT} does not cut its own sequence (Tk1899^{WT}) (left), however can cut inteinless allele of RadA from *Thermococcus kodakarensis* (Tk1899$^{\Delta Intein\text{-}encoding\ sequences}$) (middle). HEN activities are compromised in the Tk-RadA$^{(373\text{-}381)A}$ variant wherein HEN active site residues are altered to alanines (right). The intact DNA target and resultant cleavage product(s) are represented by a star and a triangle, respectively.

the presumptive HEN active site residues were either changed to alanine (TK-RadA$^{A(373\text{-}381)}$) or deleted (TK-RadA$^{\Delta 373\text{-}381}$) all HEN activity was lost (Figure 3.1B; 3.2C). Additionally, Tk-RadA lacking the entire HEN domain was prepared (Tk-RadA$^{\Delta 286\text{-}585}$), and as expected, failed to cleave TK1899$^{\Delta Intein}$ DNA (Figure 3.2C). The exact cleavage site, sequence requirements for HEN-mediated DNA cleavage, and HEN active site residues were determined for Tk-RadA^{WT} (Figure 3.2A,B; 3.3A-F). We found that Tk-RadA intein HEN domain, recognizes and cleaves long asymmetrical DNA sequence, >23 bp, encoding for an inteinless allele of RadA producing DNA fragments with 3'-hydroxyl overhangs congruent with previous finding of similar HEN domain in *T. kodakarensis* DNA polymerase B inteins[39].

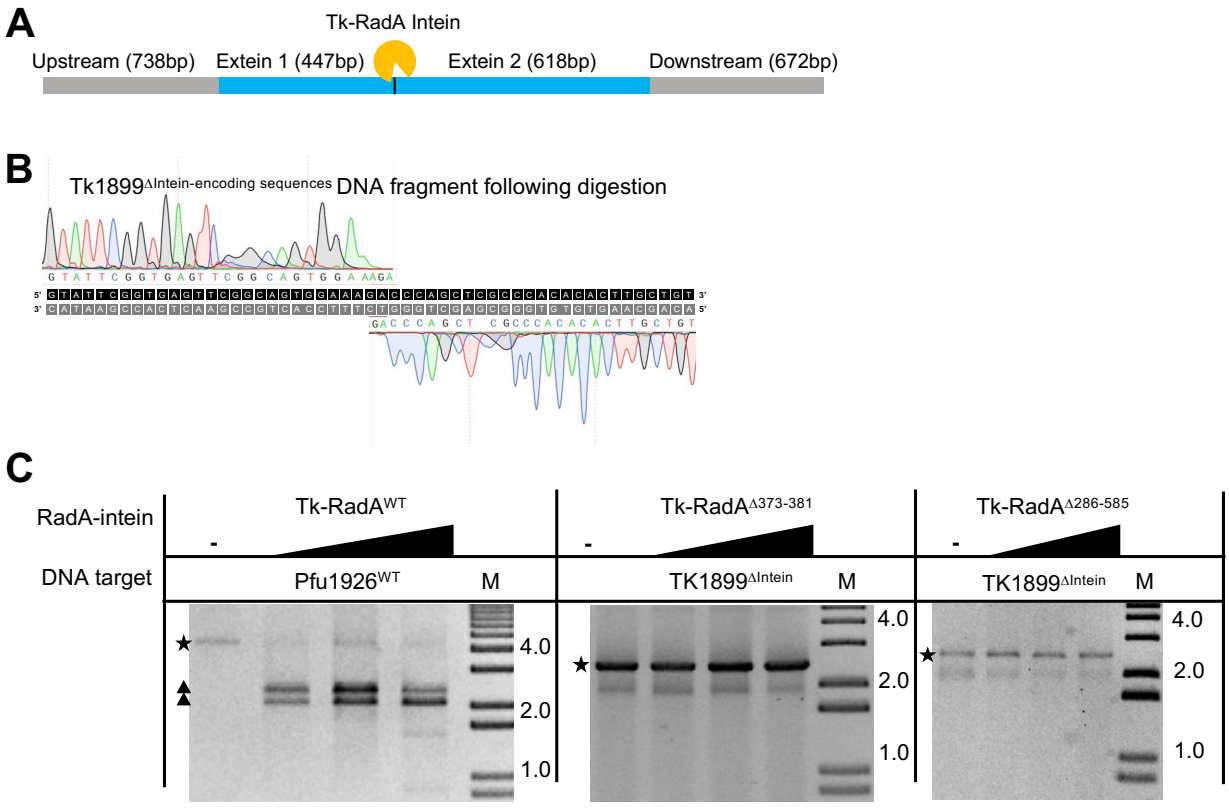
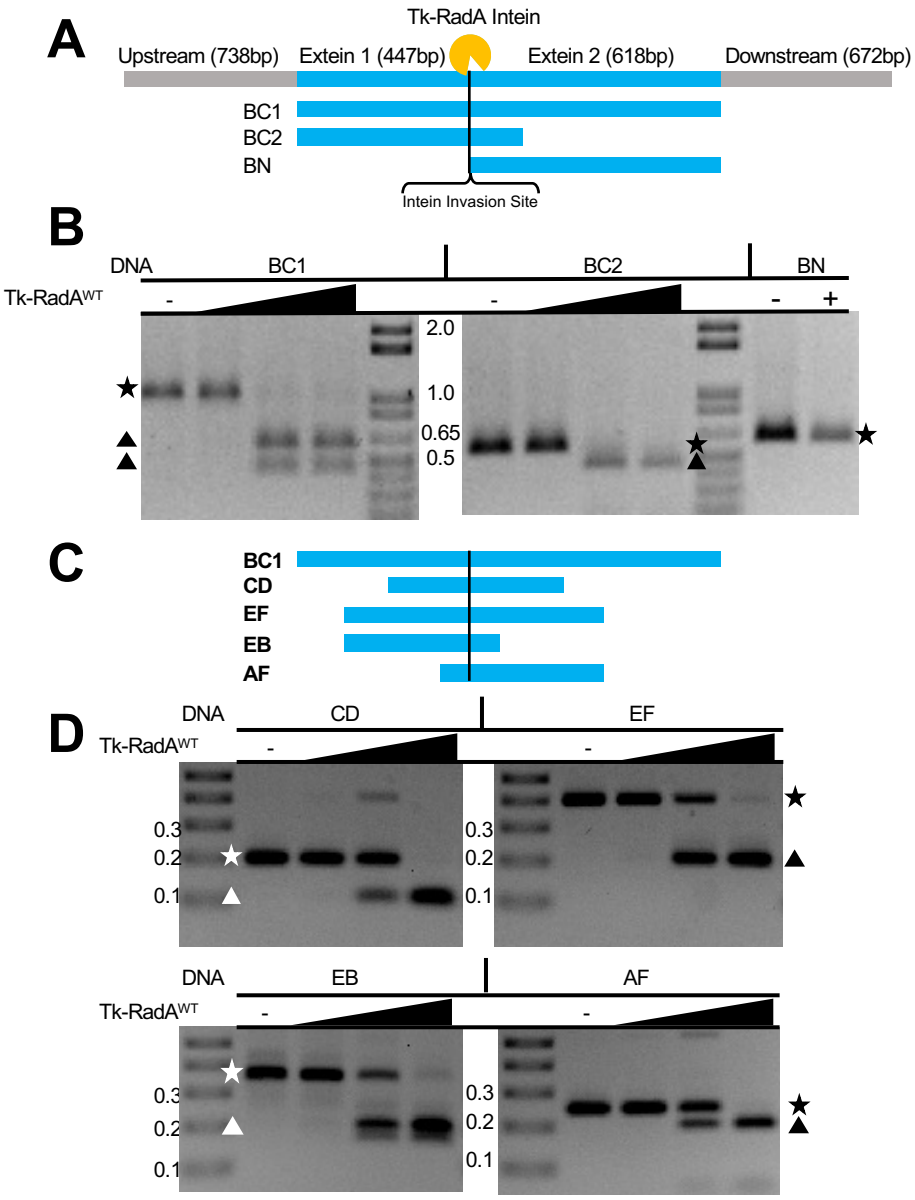


Figure 3.2. Tk-RadA^{WT} contains a self-splicing intein with active homing endonuclease activities capable of cleaving intein-less alleles.

(A) Diagrams of the DNA fragments generated and tested to confirm the cleavage site of the Tk-RadA HEN activity. Grey and blue bars denote sequences surrounding and encoding an intein-less allele of TK1899, respectively. The black line denotes the presumptinve HEN cut site between the extein-encoding sequences. (**B)** Cleavage fragments of the amplicon containing the Tk1899$^{\Delta Intein\text{-}encoding\ sequences}$ DNA sequence were Sanger sequenced to identify the exact position of cleavage on each strand. **(C)** *In vitro* HEN assays demonstrate that Tk-RadA^{WT} can efficiently cleave amplicons of the native, intein-less allele of RadA from the closely related species *Pyrococcus furiosus* (Pfu1926^{WT}; left). HEN activities are compromised in Tk-RadA variants wherein active site residues are missing or lack its HEN domain, RadA$^{\Delta 373\text{-}381}$ (middle) and RadA$^{\Delta 286\text{-}585}$ (right) respectively. Intact DNA target and cleaved DNA target(s) are denoted as a star and triangles, respectively.


A
Tk-RadA Intein
Upstream (738bp)
Extein 1 (447bp)
Extein 2 (618bp)
Downstream (672bp)
BC1
BC2
BN
Intein Invasion Site
B
DNA
BC1
BC2
BN
Tk-RadA^WT
-
-
-
+
2.0
1.0
0.65
0.5
C
BC1
CD
EF
EB
AF
D
DNA
CD
EF
Tk-RadA^WT
-
-
0.3
0.2
0.1
DNA
EB
AF
Tk-RadA^WT
-
-
0.3
0.2
0.1

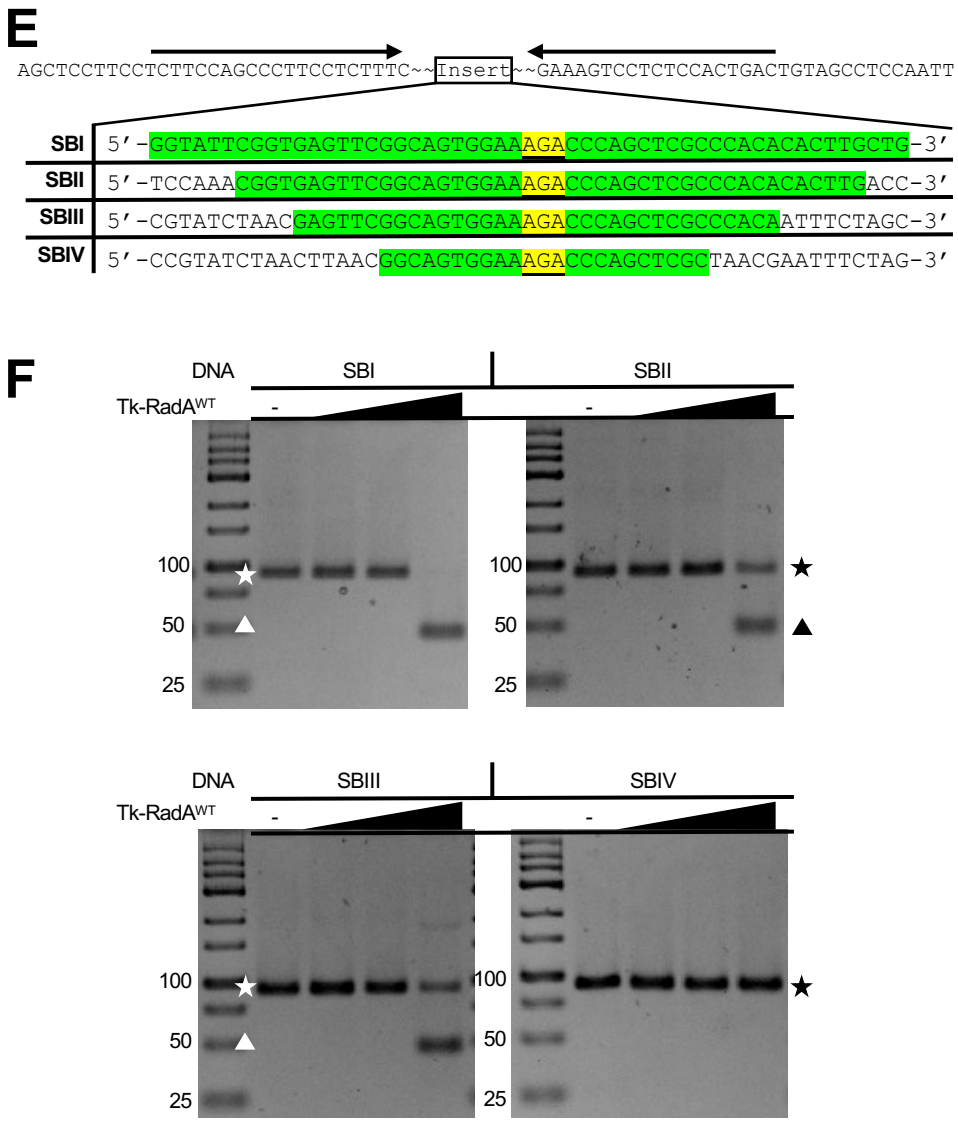


Figure 3.3. Defining the minimum recognition sequence necessary for HEN-initiated DNA cleavage of intein-less alleles.
(A, C, E) Diagrams and sequences of the DNA fragments generated and tested to narrow the recognition site of Tk-RadA HEN activity. Grey and blue bars denote surrounding and TK1899 intein-encoding sequences, respectively. The black line denotes the HEN cut site. Artificial substrates with decreasing sequences derived from TK1899 (highlighted in green) were inserted into the Blue Heron pUC MinusMCS vector, from which amplicons were generated to assay HEN activities. (**B, D, F)** HEN assays monitor the ability of purified Tk-RadA^{WT} protein to recognize amplicons containing targeting sequences and cleave the DNA fragments. HEN activities of Tk-RadA^{WT} are detected against DNA sequences containing at least 30 bp of the cut site but not detected when only 20 bp sequence surrounding the cut site are available within the amplicon. Intact DNA target and cleaved DNA target(s) are denoted as a star and triangles, respectively.

Variant intein sequences impact RadA splicing efficiency

Intein splicing accuracy and efficiency are often quantified using reporters with fully or partially substituted exteins (Figure 3.4A). Placing the Tk-RadA-intein between the artificial exteins maltose-binding-protein (N-extein) and green fluorescent protein (C-extein) generates a splicing reporter we refer to as MIG. MIG constructs provide a rapid mechanism to detail the impacts of intein composition and solution conditions on the accuracy and efficiency of protein splicing determined by the relative fluorescence of unspliced precursor to ligated exteins in total cell lysates. Precursor and ligated extein species are monitored in-gel under conditions (i.e. samples are not boiled prior to SDS-PAGE) where GFP fluorescence is maintained (Figure 3.4B)[40].

Once the Tk-RadA-intein$^{A(373\text{-}381)}$ and Tk-RadA-intein$^{\Delta 373\text{-}381}$ were placed within a MIG reporter, splicing was highly efficient and proceeded to completion during expression in *E. coli*, even at 15°C (Figure 3.4B,C). Tk-RadA-intein^{WT} with the active HEN was toxic to *E. coli* within the MIG construct and thus the relative efficiency of splicing was determined with constructs that lack HEN activity. Internal components of the Tk-RadA-intein, excluding those residues known to be directly involved in intein-mediated catalysis, can influence the efficiency of splicing[41,42]. While HEN inactivation does not fully inhibit splicing, removal of the entire HEN domain (either Δ276-585 or Δ286-585) from the Tk-RadA-intein dramatically lowers splicing efficiency from ~100% to just ~20% during expression at 15°C (Figure 3.4B,C). By comparison, the closely related *Pyrococcus horikoshii* RadA mini-intein (Ph-RadA-intein), nearly identical to the Tk-RadA-intein but naturally lacking the HEN domain (Figure 3.5A,B), splices very efficiently within the MIG splicing reporter (Figure 3.4B,C), as previously observed[22–24].

When comparing the sequences of the Ph-RadA-intein and the Tk-RadA-intein$^{\Delta 276\text{-}585}$, which are identical in length and are highly conserved (Figure 3.5A), one region of the sequence displays minimal conservation (red, Figure 3.5A). This poorly conserved region (Tk-RadA aa 270-275

and 586-592 compared to Ph-RadA aa 273-285; note that the Tk-sequence is split due to ∆276-585) where the HEN domain of Tk-RadA is found, is where Ph-RadA once presumably housed a HEN domain that was lost during evolution[22]. Interestingly, substitution of the corresponding *P.ho*. intein amino acids residues 273-285 (referred to as *P.ho.* loop) for the *T.k.* intein amino acids into an otherwise HEN-deleted Tk-RadA-intein sequence (Tk-RadA-intein$^{\Delta 270\text{-}592 + P.ho.\ loop}$) largely restored the splicing defect of Tk-RadA-intein variants with large HEN domain deletions (Figure 3.4B,C).

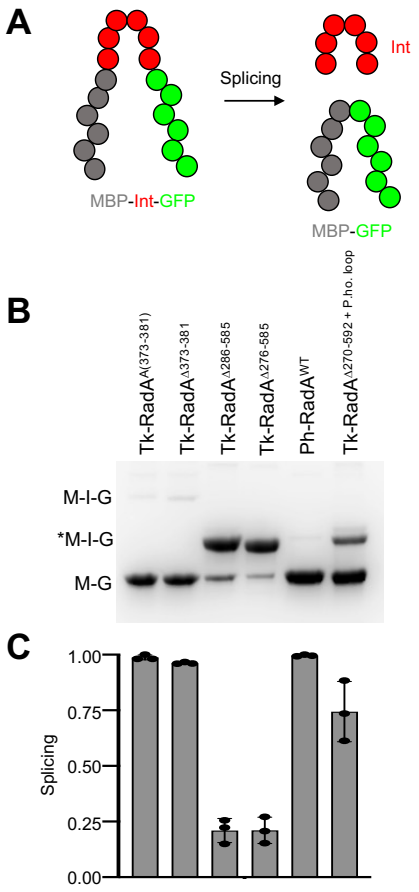


Figure 3.4. Intein-extein partnerships and intein composition impact the efficiency of intein splicing.
(A) Schematic of the MBP-Int-GFP (MIG) reporter constructs wherein intein-containing fluorescent precursor proteins must excise the intein (Int) and ligate the exteins to generate a matured, smaller fluorescent product (MBP-GFP). (**B-C)** Native PAGE and quantification of fluorescent signals reveals the impact of intein composition and the presence of the HEN domain on the efficiency of intein excision.

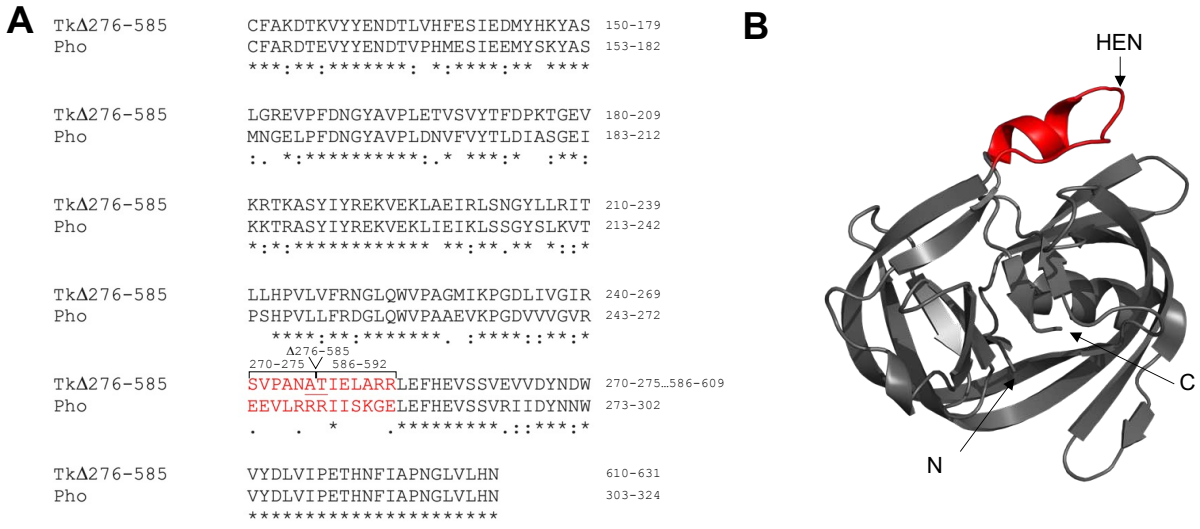


Figure 3.5. Loop residues in the *P. horikoshii* RadA intein are important for splicing. **(A, B)** Sequence alignment of the Tk-RadA$^{\Delta 276\text{-}585}$ and the Ph-RadA inteins reveals substantial congruence except for a 14 aa (highlighted in red) patch where the HEN domain of Tk-RadA is normally inserted. The divergent residues are known to form a short helix and loop within the atomic structure of the *P. horikoshii* RadA-intein (right, in cartoon) that impacts the temperature-dependence of intein splicing.

Native RadA exteins block intein splicing following HEN deletion

Given that Tk-RadA is an essential protein, the construction of *T. kodakarensis* strains wherein severe intein-splicing defects would sufficiently limit the production of mature Tk-RadA (mTk-RadA) were predicted to be problematic. Inefficient splicing of Tk-RadA-intein variants lacking the HEN domain (Δ276-585 or Δ286-585) can be rescued by incubation at elevated temperature within the MIG reporter (Figure 3.6A,B). The splicing efficiencies of Tk-RadA-intein$^{\Delta 286\text{-}585}$ and Tk-RadA-intein$^{\Delta 276\text{-}585}$ increased from just ~20% to ~80% and ~100%, respectively, within MIG constructs upon incubation at just 50˚C. We carried out additional *in vitro* and MIG reporter assays to develop an intein-splicing efficiency - that might translate *in vivo* - to generate a HEN-deleted Tk-RadA-intein that would reduce, but not eliminate splicing (Figure 3.6).

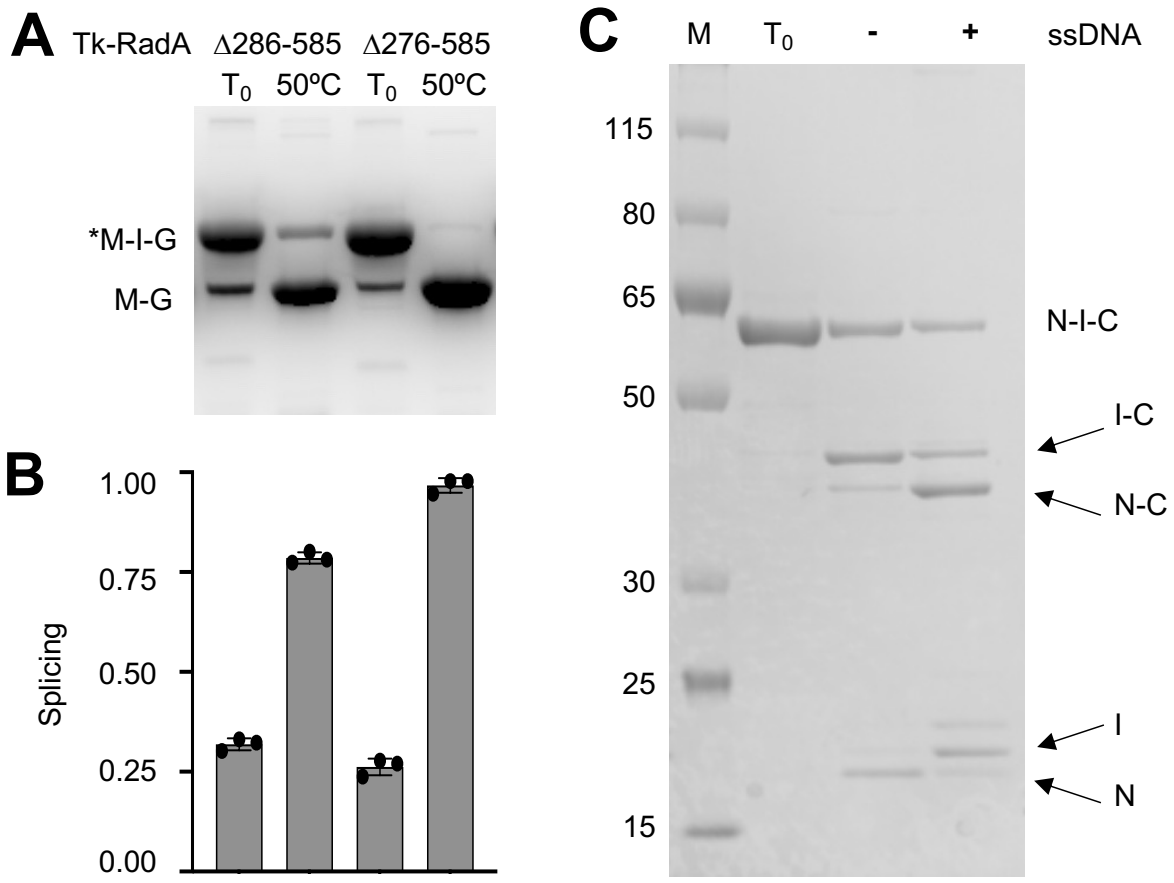


Figure 3.6. The efficiency and accuracy of pTK-RadA splicing is radically improved at high temperature or through addition of ssDNA *in vitro*.
(A-B) Native PAGE and quantification of fluorescent signals reveals the impact of temperature on the efficiency of intein excision of Tk-RadA$^{\Delta 286\text{-}858}$. **(C)** SDS-PAGE of purified pTk-RadA$^{\Delta 286\text{-}585}$ prior (T_0) to incubation without (-) and with (+) a ssDNA cofactor that dramatically increases the efficiency and accuracy of intein splicing and extein-ligation. The lane labeled M contains proteins of known molecular weights labeled in kDa.

As anticipated, we were unable to obtain large HEN-domain deletions of the Tk-RadA-intein within the natural extein context of *T. kodakarensis,* presumably due to the low splicing efficiency and thus insufficient *in vivo* levels of mTk-RadA to support growth. We were, however, successful in generating *T. kodakarensis* strains encoding large HEN-domain deletions within the Tk-RadA-intein when accompanied by the addition of the *P.ho.* loop that was demonstrated (Figure 3.4B,C) to assist splicing of HEN-deleted RadA inteins *in vitro*. The failure to recover

strains at 85˚C wherein the HEN-encoding sequences were selectively deleted adumbrates that temperature alone was unable to restore sufficient splicing of the HEN-deleted Tk-RadA-intein within *T. kodakarensis* to support growth when expressed in the native extein context, despite temperature having a large impact on splicing efficiency within the MIG reporter (Figure 3.6A); Ph-RadA-intein splicing efficiencies within the native exteins, rather than the MIG reporter, are also grossly impacted by intein-extein interactions[24]. In fact, upon expression of the Tk-RadA-intein$^{\Delta 286\text{-}585}$ in a construct containing the native exteins, no splicing is observed during expression in *E. coli* and subsequent purification (Figure 3.6C). Attempts to increase splicing of the Tk-RadA-intein lacking the HEN domain via incubation at 65°C only modestly increased splicing, and while the levels of precursor protein are reduced, splicing of this variant was inefficient (~10%) and off-pathway products resulting from cleavage of the N-extein from the intein-C-extein prior to ligation dominate (Figure 3.6C).

RadA drives recombination by forming nucleoprotein filaments on single-stranded DNA (ssDNA) as a first step in homologous recombination. As the addition of substrates (ssDNA) can rescue Ph-RadA-intein splicing within native exteins[22,23,34], we tested whether the addition of ssDNA might similarly increase the splicing efficiency of Tk-RadA-intein$^{\Delta 286\text{-}585}$ within the native extein context (Figure 3.6C). As predicted, addition of ssDNA increased the accuracy of Tk-RadA-intein$^{\Delta 286\text{-}585}$ splicing within native exteins from just ~10% to ~70% *in vitro* (Figure 3.6C). While it is speculative as to any potential role for ssDNA in Tk-RadA splicing, our results indicate that Tk-RadA$^{\Delta 286\text{-}585}$ retains at least partial ssDNA binding capacity in the precursor form. Further, these results suggest a possible means by which some mini-inteins might rely on cellular signals or substrates (e.g. ssDNA) to compensate for splicing defects following HEN domain loss.

Such dramatic splicing defects within native exteins (Figure 3.6C) likely explain our inability to obtain HEN deletions without the addition of the *P.ho.* loop *in vivo* in *T. kodakarensis*. While inviability associated with an inability to efficiently splice an essential protein such as RadA may appear an obvious result, it is the first such description within an intein-containing organism. As such, our results have implications for other intein-containing microbes and support the hypothesis that targeted approaches to impair protein splicing in human pathogens, such as *Mycobacterium tuberculosis*, represent a viable antimicrobial strategy.

Variable protein splicing within *T. kodakarensis*

Following an allelic exchange of genomic sequences to inactivate the Tk-RadA^{WT}-intein HEN active site (aa 373-381), sequences encoding the entire Tk-RadA^{WT}-intein could then be easily deleted or allelically exchanged on the *T. kodakarensis* genome (Figure 3.7A). To evaluate the impacts of variable protein splicing on the viability, fitness, and potential replicative strategies employed *in vivo* due to altered levels of mTk-RadA, we generated a series of otherwise isogenic strains of *T. kodakarensis* wherein TK1899 sequences (encoding RadA) were modified to generate alleles that encode Tk-RadA variants wherein intein-splicing efficiency was predicted to be altered (Figure 3.7A). Importantly, all strains retain the native promoter and genomic locus, and encode the same extein sequences, ensuring mTk-RadA in each strain is identical at the primary sequence level. Intein sequence variants, however, are likely to impact the efficiency of intein excision and thus steady-state mTk-RadA protein levels *in vivo*. Whole genome sequencing, at greater than 100X coverage for each strain, revealed that the newly constructed strains contained only the desired allelic modifications to TK1899 and were otherwise isogenic throughout the remainder of the > 2 Mbp genome to that of the parental strain TS559.

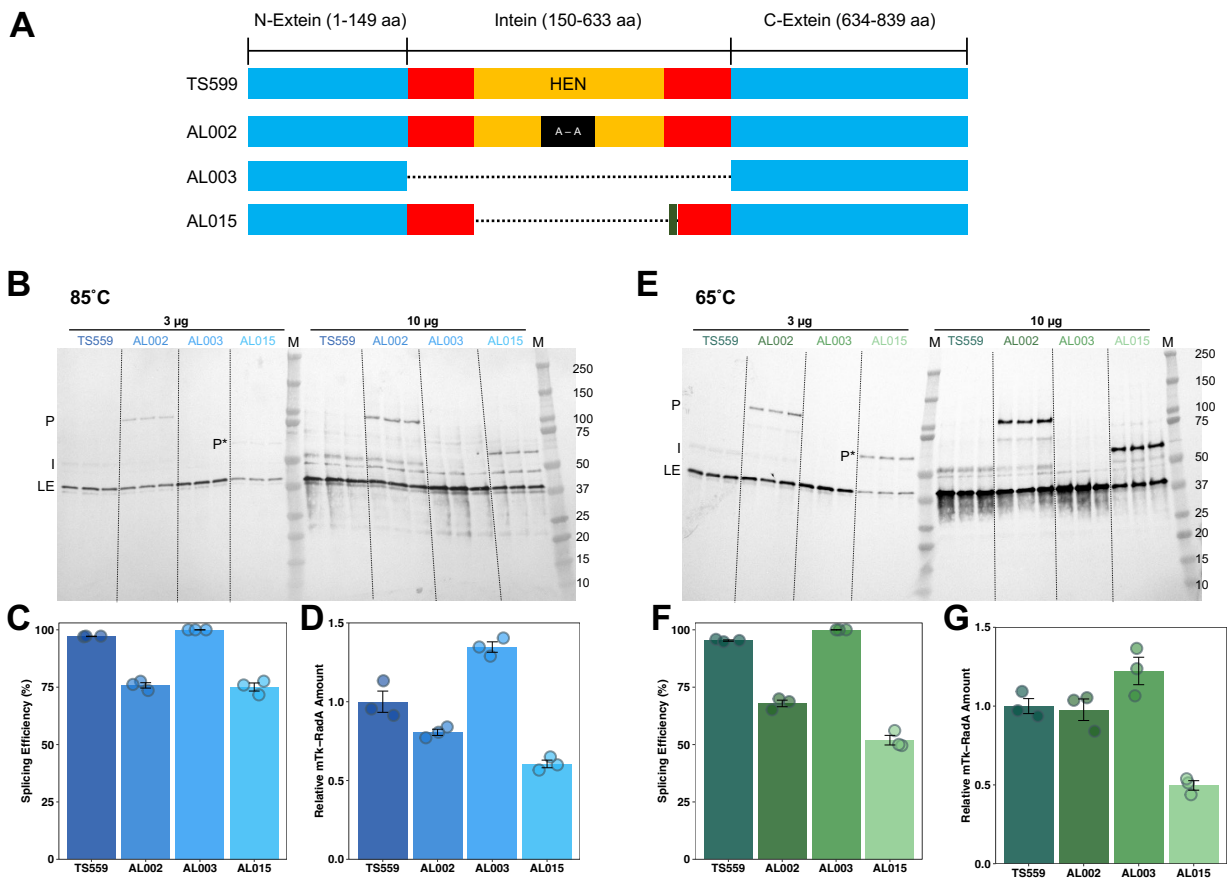


Figure 3.7. Intein splicing defects impact *in vivo* levels of mTk-RadA.
(A) Tk-RadA allelic variants introduced into the *T. kodakarensis* genome at the native locus retain native extein and promoter sequences. Strains differ solely in the sequences encoding the Tk-RadA-intein. The full-length precursor of Tk-RadA^{WT} from the parental strain TS559 contains the unspliced intein (red and orange; aa 150-633) with an active HEN domain (orange; aa 286-585) embedded within N- and C-terminal exteins (blue; aa 1-149 and 634-839, respectively). AL002 was constructed from TS559 and encodes an inactive HEN (red, orange, and black; 373-381) with alanine substitution of aa 373-381 (Tk-RadA$^{A(373-381)}$). AL003 (Tk-RadA$^{\Delta Intein}$) encodes a Tk-RadA variant lacking the entire intein sequence (aa 150-633) introduced into AL002. Reintroduction of the allele lacking the HEN domain with a small loop derived from the mini-intein of *Pyrococcus horikoshii* RadA protein into AL003 generated AL015 (Tk-RadA$^{\Delta 270-592 + P.ho.\ loop}$). (**B** and **E)** *In vivo* Tk-RadA protein and intein levels were monitored by Western blotting using polyclonal antibodies against the precursor Tk-RadA^{WT}. Strains were grown to mid-exponential phase in biological triplicate at 85˚C (**B**) and 65˚C (**E**). All Tk-RadA^{WT} isoforms were detected, including precursor Tk-RadA (P), precursor Tk-RadA lacking the HEN domain (P*), mature, ligated exteins of Tk-RadA (LE), and the spliced-intein (I). (**C** and **F**) Tk-RadA splicing efficiency at 85˚C (**C**) and 65˚C (**F**) were quantified from the ratio of LE/P or LE/P* intensities. (**D** and **G**) Relative mature Tk-RadA levels at 85˚C (**D**) and 65˚C (**G**) were quantified from the ratio of LE (variant strains)/LE (TS559). Error bars represent standard error from the mean of a minimum of three biological replicates.

The efficiency of intein-splicing *in vivo,* as well as the steady-state abundance of mTk-RadA (resultant from properly Ligated Exteins (LE)) and the Tk-Rad-intein were quantified via Western blotting employing polyclonal antibodies raised against the unspliced precursor Tk-RadA^{WT} (pTk-RadA; Figure 3.7B,E; each strain was evaluated with triplicate biological replicates). Total cell lysates derived from the *T. kodakarensis* parental strain (TS559) were compared with lysates from otherwise-isogenic strains encoding variant sequences of TK1899 that result in (i) direct production of mTk-RadA^{WT} without the need for splicing due to the removal of all intein-encoding sequences (strain AL003), (ii) a pTk-RadA$^{A(373\text{-}381)}$ variant wherein the full intein-encoding sequences are retained but modified such that the HEN active site is compromised due to replacement of residues 373-381 with alanines (strain AL002), or (iii) a pTk-RadA$^{\Delta 270\text{-}592\ +\ P.ho.\ \text{loop}}$ variant wherein the entire HEN domain was removed and a small sequence derived from the Ph-RadA-intein that improves splicing (Figure 3.4B,C) was added (strain AL015). When total cellular lysates were resolved and probed with polyclonal antibodies raised against pTk-RadA^{WT}, we were readily able to detect the anticipated dominant protein products of TK1899, namely pTk-RadA^{WT} (P), mTk-RadA^{WT} (LE), and the Tk-RadA-intein (I) (Figure 3.7B,E).

In the parental strain TS559 wherein the native extein-intein partnerships of Tk-RadA are retained, *in vivo* protein splicing is efficient (~95% of pTk-RadA is processed to mTk-RadA and Tk-RadA-intein) and accurate (no evidence of off-pathway reactions) (Figure 3.7B,C). Retention of the Tk-RadA-intein as a stand-alone protein adumbrates potential biological roles for such, including HEN activity. When the intein-encoding sequences of TK1899 are fully removed from the genome (strain AL003), mTk-RadA^{WT} is the direct product of translation and Western blotting reveals only a single band corresponding to mTk-RadA^{WT}; as anticipated, and Tk-RadA-intein signal is completely absent (Figure 3.7B,E). mTk-RadA^{WT} levels in strain AL003 increase to ~135% compared the parental strain TS559 (Figure 3.7D,G). We rationalize that this

difference in mTk-RadA^{WT} levels results from some native pTk-RadA^{WT} misfolding, improperly splicing, or degrading in TS559.

The importance of intein-sequences beyond those immediately engaged in the chemistry of protein splicing is revealed in strain AL002, where the 9 alanine substitutions in the active center of the HEN domain of the Tk-RadA-intein diminish the efficiency of splicing by ~25% and obvious levels of non-spliced pTk-RadA$^{A(373-381)}$ are retained at steady-state (Figure 3.7B,C,E,F). In strain AL015, removal of sequences encoding the entire HEN domain and addition of the corresponding sequences from the Ph-RadA-intein, generating Tk-RadA$^{\Delta 270-592\ +\ P.ho.\ loop}$, result in a smaller precursor protein (P*) that is also less efficiently spliced compared to pTk-RadA^{WT} (~75%) due to changes within intein residues not involved in the chemistry of protein-splicing (Figure 3.7B,D,E,F). The *in vivo* splicing deficiencies for pTk-RadA$^{A(373-381)}$ and pTk-RadA$^{\Delta 270-592\ +\ P.ho.\ loop}$ were significant from 85˚C grown strains (Figure 3.7B,C) and were further exacerbated when strains were grown at just 65˚C (Figure 3.7E,F), demonstrating the environmental conditions can impact splicing efficiencies and thus regulate production of intein-free proteins *in vivo*.

Impaired splicing impacts mature RadA protein levels

Changes to the efficiency of pTk-RadA splicing *in vivo* directly impact the steady-state levels of mTk-RadA available for cellular activities, including formation of nucleoprotein filaments necessary for initiation of homologous recombination. In both strains AL002 and AL015, wherein the efficiency of Tk-RadA-intein splicing was reduced, the impact on steady-state protein levels was significant, with HEN-inactivation and HEN-deletion *P. ho.* loop addition within the Tk-RadA intein resulting in ~20% and ~40% less of the identical mTk-RadA product, respectively (Figure 3.7B,D) when strains were grown at the optimal temperature of 85˚C.

Given the variable splicing efficiencies due to temperature changes observed *in vitro* and in *E. coli*, we grew *T. kodakarensis* strains at 65˚C to determine what impacts environmental conditions might play in the efficiency of Tk-RadA protein-splicing and the steady state abundance of mTk-RadA *in vivo* in the native host (Figure 3.7E,F,G). A 20˚C reduction in temperature had minimal and no impact on the splicing efficiency of Tk-RadA in strains TS559 and AL003, respectively. The noted temperature dependence for *Thermococcal* RadA proteins *in vitro*[22] is thus largely compensated for *in vivo*. In strains AL002 and AL015, where splicing efficiency was compromised at 85˚C, the reduction in growth temperature to 65˚C resulted in even greater splicing deficiencies, with protein-splicing reduced to just ~66% and ~53%, respectively (Figure 3.7E,F). The deficiency in splicing at 65˚C also significantly impacts the steady-state protein levels of mTk-RadA^{WT}. As observed at 85˚C, the RadA-intein-less strain AL003 retains a greater (~120%) amount of mTk-RadA^{WT} than observed in TS559. Surprisingly, the reduction in splicing efficiency in strain AL002 at all temperatures only very modestly reduces (~95%) the steady-state level of mTk-RadA^{WT} at 65˚C implying that a combination of splicing and degradation rates impacts protein levels in strain AL002. In line with the increased splicing deficiencies in strain AL015, the steady abundance of mTk-RadA^{WT} is halved (~50%) compared to the otherwise isogenic parental strain with the native Tk-RadA-intein sequence (Figure 3.7E,G).

Intein-splicing defects compromise growth and response to DNA damage

Deletion of the sequences encoding the RadA-intein in strain AL003 or inactivation of the HEN activities within the RadA intein in strain AL002 maintained mature RadA protein levels *in vivo* at both 65˚C and 85˚C (Figure 3.7D,G) and unsurprisingly resulted in negligible impacts on growth compared to the parental strain TS559 at both temperatures (Figure 3.8A,B). These results suggest no negative or positive fitness impact due to the absence of a functional HEN-containing intein nor the complete loss of the intein from the cytoplasm of *T. kodakarensis*.

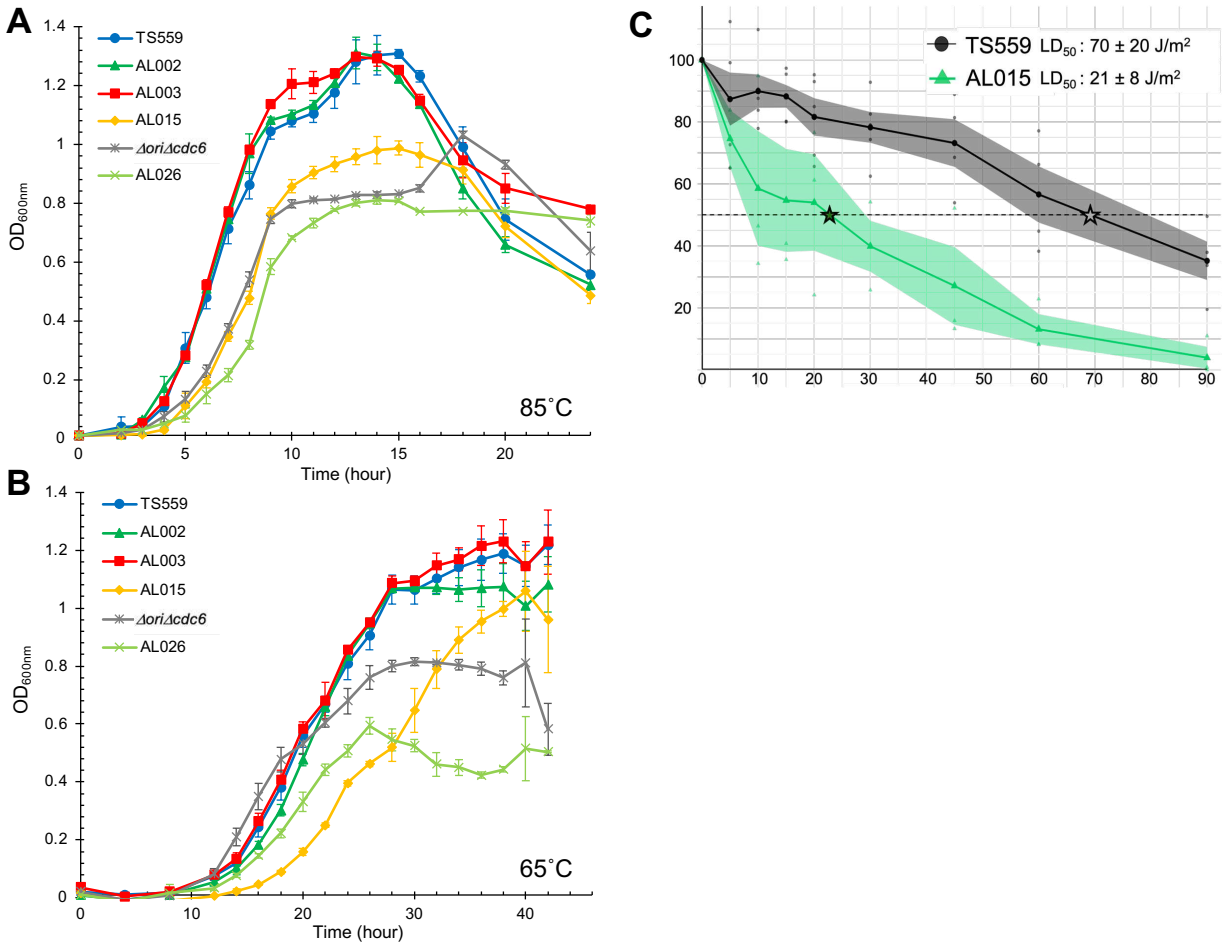


Figure 3.8. Impaired intein-excision and reduced mTk-RadA protein levels impair cellular fitness.
(A-B) Growth of triplicate biological replicates of strains of *T. kodakarensis* monitored by changes in optical density (OD_{600nm}) at 85˚C over 24 hours (**A**) and 65˚C over 42 hours (**B**) reveals that impaired intein splicing negatively impacts growth rate and final culture densities. Deletion of the TkRadA-intein (AL003) or inactivation of the HEN domain (AL002) does not obviously impact growth at either temperature. Deletion of the HEN domain in AL015 hinders growth at both temperatures, and when combined with the deletion of the *ori* and Cdc6 in AL026, synergistically reduces growth. (**C)** Reduced splicing efficiencies and reduced mTk-RadA levels in strain AL015 result in decreased survival upon UV exposure, adumbrating impacts to RadA-mediated DNA repair mechanisms *in vivo*. The median lethal dose (LD_{50}) denoted with a star was calculated for both strains based on the death curve data using Finney's probit analysis calculator.

In contrast, strain AL015 – wherein RadA-intein splicing is compromised and mRadA^{WT} levels are reduced compared to the parental strain (Figure 3.7D,G), demonstrated a consistent growth lag, and reduced culture density was observed (Figure 3.8A,B). Given the parental (TS559) and AL015 strains are otherwise isogenic, the most parsimonious cause of the growth defect is a direct association with RadA-intein excision defects. While the reduction in splicing efficiency is moderate and may not have been predicted to manifest a phenotypic consequence or fitness impact, reduced splicing efficiency and reduced mRadA^{WT} protein levels in AL015 have a demonstrated fitness cost for *T. kodakarensis* (Figure 3.8A-C).

RadA oligomerizes and forms nucleoprotein filaments that drive strand-invasion; such invasions can initiate homology-directed DNA repair, including the repair of bulky DNA lesions resultant from exposure to ultraviolet light (UV) light. To ascertain if the modest defects in RadA-intein splicing and associated approximately two-fold reductions in mRadA^{WT} steady-state protein levels impacted RadA-mediated DNA repair mechanisms, UV sensitivity assays were used to compare the ability of parental (TS559) and AL015 strains to repair bulky DNA damage. We detected a significant and substantial (~3.5-fold) UV-sensitive phenotype within AL015 cells following UV damage compared to the parental strain, showing that a ~50% reduction in mTk-RadA abundance greatly affected RadA-mediated DNA repair mechanisms and compromises cellular fitness (Figure 3.8C).

Reduced RadA-intein splicing can control the dominant mode of DNA replication

The defects in DNA repair due to reduced mRadA^{WT} levels (Figure 3.8C) suggested that other RadA-mediated processes, like recombination-dependent replication (RDR), could be similarly impacted by defects in RadA-intein excision proficiency. The dominant mode of replication in *T. kodakarensis* cells is easily determined by marker frequency analysis (MFA) (Figure 3.9)[13,43]. In rapidly growing but unsynchronized planktonic *T. kodakarensis* cultures that are using an

origin(s) of replication, DNA harvested during exponential growth should retain an overrepresentation of sequences adjacent to the origin compared to sequences located further away from the origin(s). When sequence abundances recovered from i) growing and replicating cultures and ii) non-growing, non-replicating stationary phase cultures are compared, any regional overabundance defines an origin(s) of replication and would be consistent with ODR (e.g. a peak(s) in the plotted frequency of sequence abundance defines an origin, with a smooth wave of regional overabundance extending bidirectional away from the origin providing confidence that at least some of the cells in culture are using origin-dependent replication as the dominant mode). When a relatively equal abundance of all genomic sequences is returned in MFA, all sites on the genome are equally likely to serve as replication origins and would be consistent with RDR.

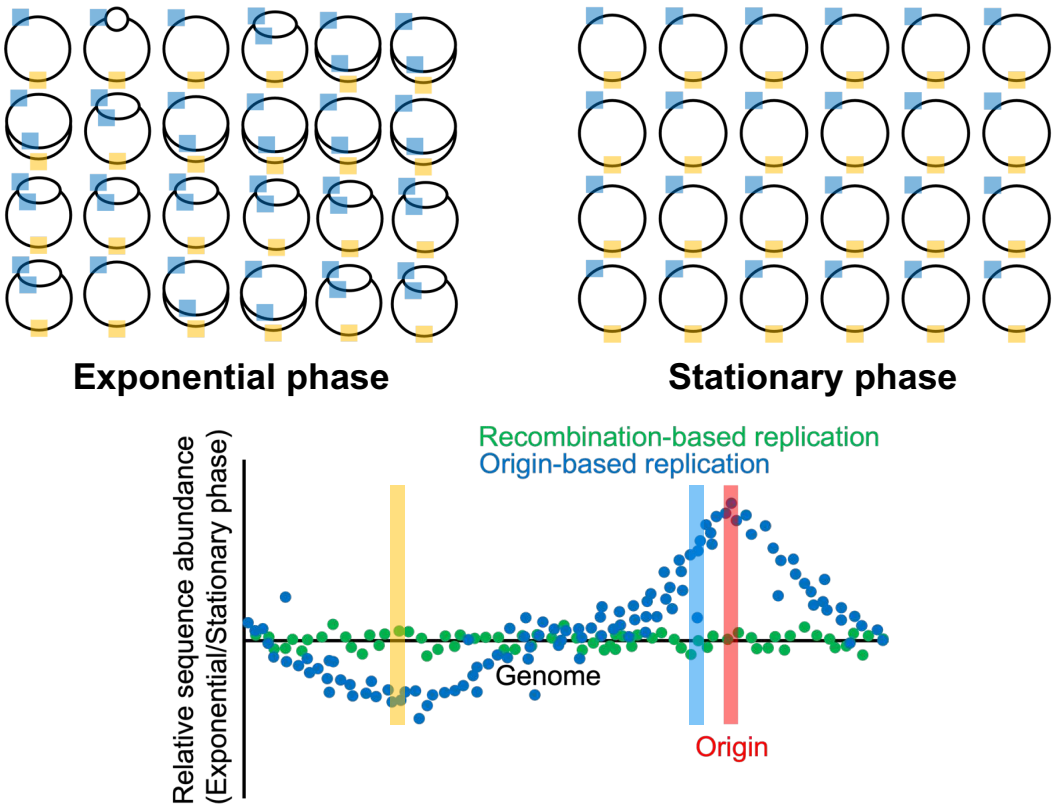


Figure 3.9. Replication strategies can be defined with marker frequency analysis. The entire *T. kodakarensis* genome was shown by a black circle, with cyan and orange boxes denoting generic regions in the genome proximal and distal, respectively, to the origin of replication (*ori*). When origin-dependent replication prevails as the dominant replicative strategy, DNA isolated from actively replicating culture during exponential phase retains an overrepresentation of the DNA regions closest to the *ori* (cyan) and an underrepresentation of regions of DNA located distantly from the *ori* (orange). When the frequencies of DNA sequences recovered and sequenced from replicating cultures are compared to the frequencies of DNA sequences recovered and sequenced from non-replicating cultures of the same strain in stationary phase are compared, a marker frequency analysis (MFA) permits identification of any areas of the genome that are overrepresented in replicating cells, indicative of a bias position of replication, with the peak of this bias defining a replication origin. The hypothetical blue scatter plot displays the anticipated wave pattern with a single apex at the predicted origin that would be anticipated from cultures wherein ODR dominants. In contrast, when recombination-dependent replication dominates within a growing culture, MFA will not reveal any sequence enrichment and would be consistent with the hypothetical green scatter plot.

The sole *T. kodakarensis* replication origin, the moniker of ODR, is located not far from the RadA (TK1899) locus, between TK1900 (predicted voltage-gated potassium channel) and TK1901 (Cdc6; the origin-recognition protein) at ~1.75 Mbp of the chromosome (Figure 3.10; red dashed line defines the location of the *ori*). Marker frequency analysis (MFA) of TS559 cultures grown at 85˚C and 65˚C reaffirmed[44] that an obvious single origin is not present, and therefore, is not necessary for normal growth of *T. kodakarensis*. These results are which is consistent with the preponderance of DNA replication initiation through RDR at every position of the genome with equal preference instead of a clear peak at the origin or anywhere else on the genome (Figure 10A). The loss of the Tk-RadA-intein in strain AL003 did not impact the use of RDR as the dominant mode of replication at either 85˚C or 65˚C (Figure 3.10B). In fact, AL003 appeared to generate a flatter MFA profile compared to TS559 (Figure 3.10A,B), suggesting a greater propensity for RDR due to the increased levels of mTk-RadA (Figure 3.7B,E). For AL002, particularly at 65˚C, MFA indicates a modest shift in the population toward ODR, although the change was subtle (Figure 3.10C). However, a pronounced switch in the mode of DNA replication from RDR to ODR is observed for AL0015, particularly at 65˚C (Figure 3.10C,D). This switch from RDR to ODR correlated with deficiencies in Tk-RadA-splicing (Figure 3.7B,C,E,F), a resultant decrease in mTk-RadA^{WT} (Figure 3.7B,D,E,G), an increase in pTk-RadA (Figure 3.7B,E), a reduced overall fitness (Figure 3.8A,B), and a UV-hypersensitivity phenotype (Figure 3.8C).

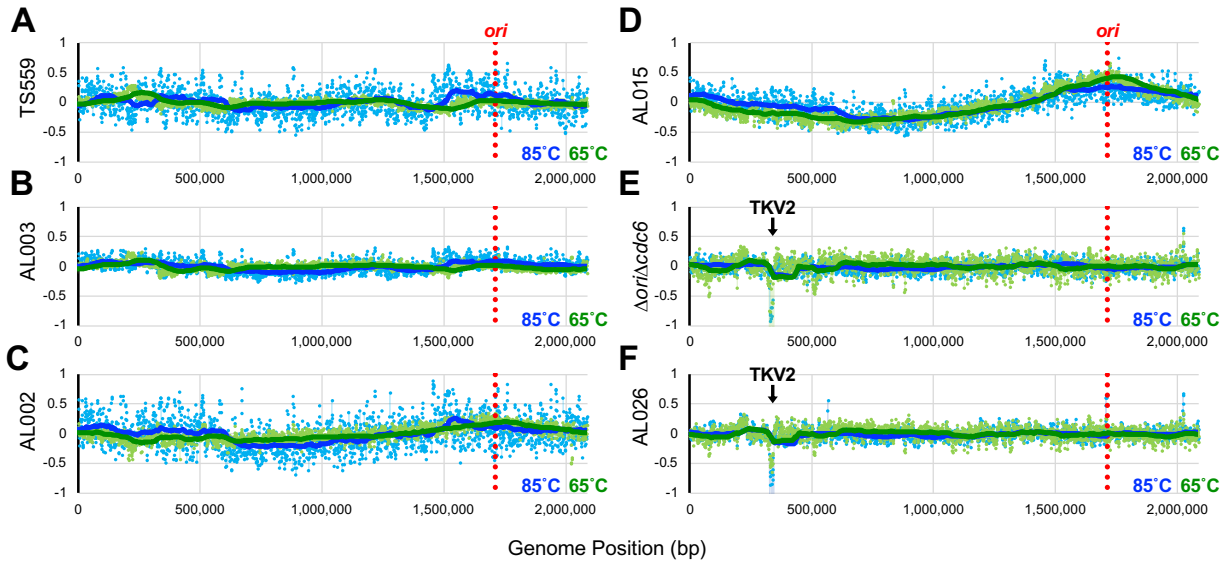


Figure 3.6. Impaired intein-excision elicits a radical shift of replication strategies. **(A-F)** Strains were grown at 85˚C (blue) and 65˚C (green) and DNAs were purified from cells harvested at both mid-exponential and stationary phase. Marker frequency analysis (MFA) reports the $\log_2$ ratio of the mapped reads from exponentially growing cells divided by reads from stationary cultures. A 100-period moving average is plotted using the thick blue/green solid lines to show the trend from the MFA. The genome of *T. kodakarensis* was segregated into 1000 bp bins. The location of the *ori-cdc6* region is indicated by the red dotted line. The downward arrow in panels **E** and **F** represents the location of the TKV2 excised region.

AL015, when compared to TS559 (parental), is completely isogenic besides changes to the Tk-RadA-intein, and it still encodes mTk-RadA^{WT}. When grown at either 65˚ or 85˚C, there was clear evidence of origin usage and thus a foundational switch in replicative strategies was manifested by a maximally two-fold change in steady-state mTk-RadA levels *in vivo*. MFA of the AL002 strain, wherein splicing efficiency is decreased to just ~80% and ~66% at 85˚C and 65˚C, respectively, displayed a more subtle transition to ODR from the RDR default; MFA provides a population average; thus some cells are likely to prefer RDR and others ODR in strain AL002. This shift correlated with accumulation of pTk-RadA rather than steady-state protein levels of mTk-RadA^{WT} (Figure 3.7C). Surprisingly, it appeared that the splicing efficiency decrease was responsible for changes in the prominence of ODR versus RDR, as total steady-state protein

levels of mTk-RadA^{WT} in AL002 were only reduced by ~20% or ~5% when compared to TS559 at 85˚C and 65˚C, respectively. Evidence for origin utilization in strain AL002 at 65˚C, wherein mTk-RadA^{WT} levels were effectively unchanged, suggested that the very act of splicing, or the retention of pTk-RadA$^{A(373-381)}$, tipped the ratio of *T. kodakarensis* cells that rely on RDR versus ODR.

***Δori Δcdc6* strains must use RDR**

The preference for RDR over ODR in *T. kodakarensis* permits deletion of the origin (*ori*) sequence and adjacently encoded origin-recognition protein Cdc6 (TK1901) without substantial growth defects[13]. *Δori Δcdc6* strains rely entirely on RDR (Figure 3.10E) and deletion of the natural origin and initiator protein does not reveal evidence of a secondary or cryptic origin of replication[13]. Construction of the *Δori Δcdc6* strain was coincident with the spontaneous deletion of TKV2, a prophage genome known to be dispensable for growth of *T. kodakarensis*[45]. Mapping of MFA results to the TS559 genome details this loss through a precise peak at ~0.32 Mbp of the genome (Figure 3.10E,F). We rationalized that introduction of the Tk-RadA$^{\Delta 270\text{-}592 + P.\ ho.\ \text{loop}}$ allele into a *Δori Δcdc6* strain incapable of ODR (generating strain AL026), would challenge both ODR and RDR mechanisms in *T. kodakarensis.* The reduced efficiency of intein splicing and reduced mTk-RadA levels favor ODR, but the deletion of the origin and initiator protein preclude use of such. As predicted, growth of AL026 strains are compromised, more drastically at 65˚C wherein intein excision is more impaired (Figure 3.8A,B), but AL026 strains are incapable of ODR as revealed by MFA (Figure 3.10F). Thus, although reduced RadA splicing (e.g. strain AL015) would normally dictate a switch in replication strategies favoring ODR when intein excision is compromised, ODR is not permitted and therefore, AL026 must use RDR. RDR is required for growth of AL026, but leads to dramatically reduced growth at 65˚C due to the sub-optimal concentrations of mTk-RadA. This represents the first demonstration of a

growth defect due to impaired splicing and the potential power of regulated intein splicing to regulate microbial physiology.

mRadA activities are unaffected by pRadA

The massive phenotypic impacts resultant from modestly decreased intein-excision and a two-fold reduction in $mRadA^{WT}$ protein levels suggest that even small changes in intein-excision efficiencies can manifest large physiological responses *in vivo*, supporting continued efforts to develop antimicrobials that control intein-splicing in essential genes. Given that RadA functions as an oligomer, the precursor RadA (pRadA) retains a response to addition of substrate (Figure 3.6), and is likely at least partially folded, we rationalized that increased steady-state levels of pRadA could impact on mRadA function. Work with the closely related *P. horikoshii* RadA demonstrated that while the precursor form of RadA cannot hydrolyze ATP, it can bind DNA[23,24], hinting that precursor forms of intein-containing proteins can perform some but not all functions. Therefore, if elevated steady-state pRadA levels permit pRadA to interrupt mRadA oligomerization and function, the true impact of deficiencies in RadA-intein excision may be reflective of both a reduced mRadA level and inhibition of mRadA-function due to unfavorable and interfering interactions between pRadA and mRadA.

RadA activity can be reconstituted and quantified through an *in vitro* recombinase assay that monitors the efficiency of purified recombinant mTk-RadA to promote strand-invasion of a ssDNA oligonucleotide into a supercoiled plasmid[21] (Figure 3.11A). Nucleoprotein filament formation was facilitated by incubating the purified recombinase with the 5'-FAM deoxyoligonucleotide (L93) prior to addition of supercoiled plasmid DNA (pUC19), which contains a sequence complementary to the L93 oligo. RadA mediates strand invasion of the L93 oligo into pUC19, forming a displacement loop (D-loop) in a RadA protein concentration and ATP-dependent manner (Figure 3.11A). The D-loop formation can subsequently be resolved

and quantified using native gel electrophoresis, revealing a shift in fluorescent signal caused by the 5'-FAM L93 oligo traveling slower with the supercoiled pUC19.

Purified mTk-RadA, as expected, functions as a recombinase and successfully catalyzes the strand invasion of the 5'FAM L93 oligo into the supercoiled pUC19, resulting in the D-loop (Figure 3.11A). The optimal temperature for RadA-mediated invasion was established to be 65˚C, comparable with prior results using RadA from *P. abysii*[21], (Figure 3.11D). Incubating the recombinase reaction with increasing amounts of pTk-RadA – in an attempt to inhibit mRk-RadA – did not impact the efficiency of D-loop formation (Figure 3.11B), suggesting that the unspliced RadA-precursors were not interacting with nor impairing the capacity of mRadA to drive D-loop formation even when present at equal molar concentrations. It should be noted that while we added pTk-RadA preparations that contained almost no-spliced proteins into the recombinase reactions, the *in vitro* recombinase assay conditions appeared to promote the splicing of pTk-RadA, as observed by SDS-PAGE (Figure 3.11C); given that the bulk of the added precursor protein does not splice, we remain confident that the excess pTk-RadA does impact mTk-RadA function *in vitro*. While we did generate a pTk-RadA variant wherein we mutated the first cysteine residue on the intein splicing junctions to completely prevent pTk-RadA from splicing (and thus eliminate any remaining concerns of splicing of pTk-RadA in the recombinase reactions), the resulting variant was unstable and repeatedly precipitated out of solution during heat treatment (data not shown). The results obtained thus argue that increased pTk-RadA levels do not negatively impact the functions of mTk-RadA and that the massive phenotypic impacts are resultant directly from reduced splicing efficiencies and reduced mTk-RadA protein levels *in vivo*.

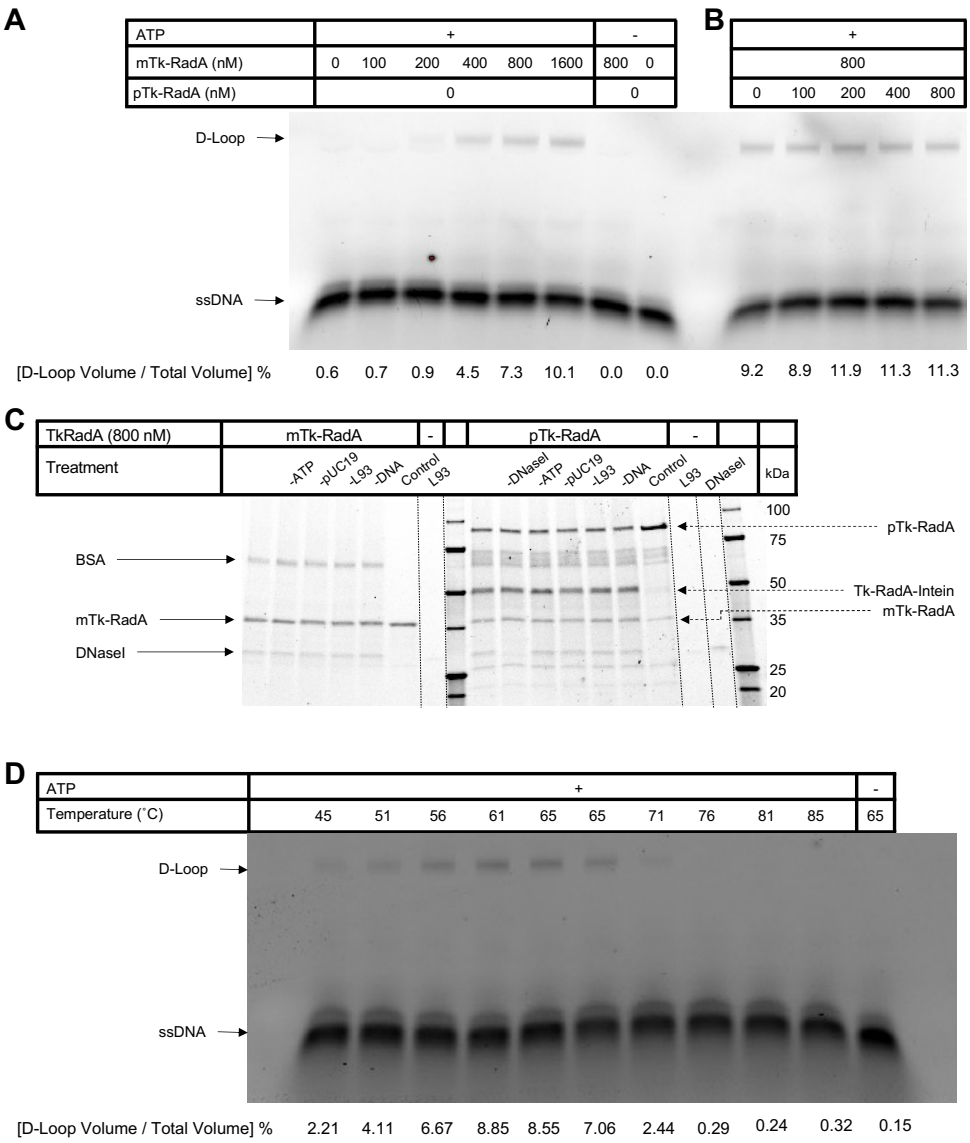


Figure 3.11. Precursor Tk-RadA does not inhibit recombinase activity of mature Tk-RadA.

Native TBE agarose gel electrophoresis visualized by fluorescence demonstrating mTk-RadA-mediated D-loop formation. **A-B.** mTk-RadA functions as a recombinase, catalyzing a protein-concentration and ATP-dependent invasion of a ssDNA into the supercoiled pUC19 plasmid forming a D-loop (**A**). Addition of pTk-RadA does not impact the efficiency of mTk-RadA recombinase activities (**B**). D-loop formation was quantified as the percentage of fluorescent oligo with retarded migration in each lane. **C.** SDS-PAGE of the purified proteins used in the D-loop formation assays reveals an elevated splicing efficiency for pTk-RadA under our D-loop assay conditions. **D.** D-loop formation assay incubated at various temperatures, 45 – 85 ˚C, with a constant mTk-RadA concentration kept at 800 nM. D-loop formation were quantified from the ratio of [D-Loop Volume/Total Volume]%.

Discussions

Inteins can be vertically inherited, transferred via endosymbiosis, move between closely related strains during mating, propagate via intragenomic transfer, or be horizontally transferred between species or even Domains, often as the result of viral mediated events[46,47]. *T. kodakarensis* encodes the most inteins relative to genome size reported, with at least 15 inteins distributed in 12 genes (Table 3.1)[36]. Remarkably, ~1% of the *T. kodakarensis* genome is composed of inteins, with these enigmatic elements representing over 20 kbp of sequence. While inteins might be traditionally viewed as a molecular parasite and fitness-negative MGEs, the last decade has yielded compelling work to demonstrate that the rate and accuracy of protein splicing for dozens of inteins is highly dependent on environment, suggesting that conditional protein splicing (CPS) may provide regulatory benefits to the intein-containing organism[26].

Table 3.1. Inteins present in the *T. kodakarensis* genome.

GENE	PROTEIN	FUNCTION	RRR	INT	INT AA	EXT AA
TK0001	PolB	DNA polymerase	Yes	2	360/536	760
TK0470	Rgy	Reverse gyrase	Yes	1	489	1222
TK0764	LHR	Large helicase-related protein	Yes	1	525	865
TK1091	TopA	DNA topoisomerase 1	Yes	1	511	718
TK1305	IF2	Initiation factor 2	No	1	546	598
TK1332	Ski2-like	DNA repair and recombination	Yes	1	403	722
TK1620	MCM-3	Replicative DNA helicase	Yes	2	140/335	682
TK1736	RNR	Ribonucleotide reductase	No	2	454/382	910
TK1853	KlbA	type II/IV secretion system ATPase	No	1	523	675
TK1899	RadA	Homologous recombinase	Yes	1	482	354
TK1903	PolD	DNA polymerase II	Yes	1	474	1324
TK2218	RFC	Clamp loader	Yes	1	540	322

The preponderance of inteins in replication-, repair-, and recombination-related (RRR) genes in *T. kodakarensis* (11/15; Table 3.1) is in line with the prevalence of inteins in RRR-related gene across both prokaryotic Domains (~65%)[36]. Given the non-random distribution and retention of

inteins in RRR-related genes, coupled with the wealth of *in vitro* work demonstrating CPS in response to environmental signals acutely relevant to the intein-containing organism[22–35], argues that some inteins have been exapted from one-time parasites to beneficial regulatory elements. However, the major criticism of this hypothesis was the complete absence of work demonstrating a connection between RRR-related protein function and intein-splicing efficiencies within an organism that naturally houses inteins. This study provides, for the first time, a compelling connection between intein activity and RRR-protein activity *in vivo*. In this case, we show the growth rates, response to fitness challenges, and the dominant mode of DNA replication can be regulated by differences the intein-splicing efficiency.

While the reduction in splicing we observe is based on intein-specific mutations in otherwise isogenic strains, we convincingly demonstrate that a reduction in mTk-RadA levels due to variable intein splicing results in *T. kodakarensis* favoring ODR over RDR (Figure 3.10), a reduction in growth rate, and an increase in UV sensitivity (Figure 3.8). Additionally, we show for the splicing-deficient AL015 strain that the switch from RDR to ODR becomes more apparent at lower temperature, which correlates with intein-splicing accuracy and efficiency both *in vivo* and *in vitro* (Figure 3.4,3.6,3.7). Although our findings rely on an artificial downregulation of intein activity by intein-specific mutations on the *T. kodakarensis* chromosome, our work suggests that CPS may be a prevalent mechanism of regulating the efficiency, rate, or mechanisms of DNA replication, recombination, and repair under distinct environmental conditions. These results represent seminal findings in an emerging field that will lay the groundwork for the investigation of CPS within the natural context of the intein.

We demonstrate that a relatively modest reduction (~50%) in the splicing of a single intein can radically impact microbial physiology. While it has been long hypothesized that splicing inhibition would result in diminished growth, this work represents the first *in vivo* demonstration and

shows the potential power of protein splicing inhibition to control microbial growth. It remains plausible that variant intein sequences within identical extein sequences can influence the folding of sequence-identical exteins, and thus impact the function of the mature protein. Regardless, knowing that differential intein-excision can radically impact total cellular physiology and fitness, combined with i) the retention of inteins in several devastating human pathogens, including *M. tuberculosis* and *Cryptococcus neoformans*, and ii) the complete absence of inteins in metazoans, lends support to the ongoing search for targeted protein splicing inhibitors as novel antimicrobials[48,49]. The ability for modest changes in the splicing of a single intein to radically impact microbial growth, combined with the retention of inteins in many pathogens, but the near absence (just 1%) or complete loss of inteins in eukaryotic genomes and metazoans, respectively, offers intriguing opportunities to target intein splicing with therapeutics[48,49].

The integration of inteins into the active sites or critical regulatory regions of RRR-related proteins may reflect a selection for intein integrations that retain the greatest likelihood of impacting function of the RRR-related protein in the precursor form. Indeed, the Tk-RadA-intein is located with the ATP-binding P-loop. It will thus be critical to determine both the efficiency of splicing and the impact on steady-state mature protein concentrations *in vivo* for inteins in different classes of genes and across each Domain to fully understand the biological role inteins play in regulating RRR mechanisms. Given even very subtle changes to *in vivo* protein concentrations and splicing efficiencies may result in dramatic changes in cellular physiology, it will be equally critical to evaluate the fitness importance of intein-splicing efficiency – an ancient selectable characteristic of inteins – to the fitness importance of regulated intein-splicing – an evolutionarily-recent exaptation of intein biology.

Materials and Methods

Microbial growth and media conditions

Parental (TS559) and newly constructed *T. kodakarensis* strains were anaerobically maintained in an artificial sea water (ASW) based medium supplemented with 5 g/l tryptone, 5 g/l yeast extract, 5 g/l pyruvate, 2 g/l elemental sulfur (S˚), a KOD1-vitamin mixture, and 1 mM agmatine sulfate at 85˚C or 65˚C[50]. Growth rates were monitored via optical density measurements at 600 nm in liquid cultures. Cultures were prepared with 1:100 inoculums from overnight cultures grown in the same medium, and the growth rates of a minimum of three independent biological replicates were monitored and reported (Figure 3.7,3.8).

T. kodakarensis strain constructions

The genomic sequences encoding Tk-RadA (TK1899) were targeted for allelic modification using standard markerless-modification protocols in *T. kodakarensis* strain TS559[50] and Δ*ori*Δ*cdc6*[13,50]. Briefly, non-replicative plasmids containing sequences homologous to the flanking regions of TK1899 along with the desired allelic change(s) were temporarily integrated, then subsequently excised from the TS559 or Δ*ori*Δ*cdc6* genome in the region surrounding TK1899 based on restoration of agmatine prototrophy and resistance to 6-methylpurine, respectively. Sanger sequencing of amplicons generated via diagnostic PCR using primers adjacent to TK1899 with genomic DNA purified from newly constructed strains confirmed the exact endpoints of deletions and any allelic modifications. Preparations of genomic DNA from strains presumed to be modified at TK1899 were sequenced at >100x coverage via MinION Nanopore sequencing (Oxford Nanopore Technologies, Oxford, UK) to finalize whole genome sequencing (WGS). Analyses of WGS results confirmed the introduction of sequences encoding allelic variants at TK1899 and the absence of any secondary mutations throughout the remainder of the genome. WGS also confirmed the loss of TKV2 in strains AL026 and

Δori Δcdc6. HEN activity of native Tk-RadA necessitated a unique path to strain constructions to ensure intein-invasion of intein-less alleles of TK1899 did not prohibit generation of desired alleles. TS559 was used to first generate AL002 (Tk-RadA$^{A(373\text{-}381)}$), the inactive HEN activity lacking variant of Tk-RadA. AL002 was used to generate AL003 (Tk-RadA$^{\Delta Intein}$), the inteinless variant of Tk-RadA. AL003 was used to generate AL015 (TkRadA$^{\Delta 270\text{-}592 + P.ho.\ loop}$), the mini-intein variant of Tk-RadA. *Δori Δcdc6* was used to generate AL026.

Cloning, expression, and purification of Tk-RadA variants

The wildtype, intein-containing, sequences of Tk-RadA (TK1899) was amplified from purified TS559 genomic DNA, incorporating sequences encoding a C-terminal 6xHis tag during amplification and cloned into the SalI site of pQE-80L via In-Fusion® Snap Assembly (Takara Bio USA, Inc.). Sequence variants, used to produce proteins variants of Tk-RadA, were introduced via Quikchange mutagenesis on the wildtype Tk-RadA expression plasmid. Sanger sequencing confirmed the full sequence of the entire insert for all constructs. Each expression plasmid was transformed into Rosetta™ 2 competent cells (Novagen). Transformants were grown at 37°C in LB media containing 100 µg/mL sodium ampicillin salt and 25 µg/mL chloramphenicol. Protein expression was induced at OD_{600nm} of ~0.4 through addition of IPTG to 0.5 mM final and protein production was permitted for an additional 4 hours of 37˚C growth. Cells were harvested via centrifugation (15,000 x *g*, 15 minutes, 4˚C), the supernatant was discarded, and cell pellets were frozen at -20˚C until protein purifications. Cell pellets were thawed and resuspended in 20 mM Tris HCl pH 8.3, 500 mM NaCl, 10% glycerol, 30 mM imidazole (Buffer A), lysed using sonication, and cellular debris were removed through centrifugation (75,000 x *g*, 4˚, 20 minutes). Clarified lysates were passed through a 5 ml HiTrap™ Chelating HP column (Cytiva) charged with nickel, washed extensively with Buffer A, then eluted using a linear, 20 column-volume gradient from 100% Buffer A to 100% Buffer B (20

mM Tris HCl pH 8.3, 500 mM NaCl, 10% glycerol, 500 mM imidazole). Fractions containing Tk-RadA were identified via SDS-PAGE, pooled, and dialyzed into 20 mM Tris HCl pH 8.3, 200 mM NaCl, 50% glycerol for long-term storage.

Homing endonuclease assays

DNA amplicons were generated via two rounds of PCR each followed with gel extraction and PCR clean-up respectively using Nucleospin® Gel and PCR Clean-up kit (MACHEREY-NAGEL, Inc.). HEN activity was evaluated at 75°C for 1 hour in 20 µL reactions containing 0 – 0.5 nM of the Tk-RadA^{WT} or Tk-RadA variant proteins and 15 nM of purified DNA substrate in 50 mM Tris HCl pH 8.3, 100 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT. The reactions were stopped by addition of 100 µL 0.6 M Tris-HCl pH 8.0, 12 mM EDTA and extracted with an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1 (v/v/v)). Precipitation of the aqueous phase was facilitated with 2.6x volumes 100% ethanol and 50 µg/mL GlycoBlue™ Coprecipitant (Invitrogen™). Purified DNAs were resolved through either 1% (Figure 3.1,3.2) or 3% (Figure 3.3) 1X TBE agarose gels run in TBE and visualized by EtBr staining. The cleaved fragments generated from Tk1899$^{\Delta Intein\text{-}encoding\ sequences}$ were Sanger sequenced by Azenta Life Sciences (Figure 3.2B) to map the exact cut sites.

D-loop formation assay

The protocol to monitor D-loop formation was adapted from Hogrel et al.[21] with a few modifications. Supercoiled pUC19 plasmids was purified using a Qiagen plasmid purification kit following a low-temperature modified protocol intended to increase isolation of supercoiled plasmids from *E. coli* cells[51]. The 5'FAM-L93 single-stranded fluorescently labeled DNA substrates (5'-[FAM]-AAA-GGC-GGT-AAT-ACG-GTT-ATC-CAC-AGA-ATC-AGG-GGA-TAA-CGC-AGG-AAA-GAA-CAT-GTG-AGC-AAA-AGG-CCA-GCA-AAA-GGC-CAG-GAA-CCG-TAA-AAA-3′) was obtained from Eurofins Genomics LLC. Briefly, 25 nM 5'FAM-L93 was mixed with

purified mTk-RadA or pTk-RadA at various concentrations from 0 – 1600 nM or 0 – 800 nM, respectively (Figure 3.11) in 20 mM Tris-HCl pH 8.0, 125 mM NaCl, 10 mM DTT, 50 µg/mL BSA, 10 mM $MgCl_2$, and 2.5 mM ATP (when indicated), followed by 10 minutes of incubation at 65 ˚C (note that incubation temperatures were varied in Figure 3.11C). Following the initial incubation, 25 nM supercoiled pUC19 was added, and the reactions were incubated again for 10 minutes at 65 ˚C (unless noted otherwise). The reactions were terminated by the addition of 50 µg/mL Proteinase K, 0.5% SDS, 40 mM EDTA followed by 15 minutes of incubation at 37˚C. Reactions were resolved through 1.2% TBE agarose gel following the addition of an equal volume 20% FICOLL. The unincorporated and D-loop-incorporated 5'FAM-L93 oligos were visualized with a Typhoon FLA 9500 (GE Healthcare). The percentage of oligo complexed with the plasmid through the activities of Tk-RadA were quantified with ImageQuant software and plotted in Excel.

Western blot analysis

Purified pTk-$RadA^{WT}$ was used as an antigen to generate polyclonal antibodies in guinea pigs (Cocalico Biologicals, Inc.); test and terminal bleeds were confirmed against purified Tk-$RadA^{WT}$ and lysates from *T. kodakarensis* for specificity. *T. kodakarensis* cultures were grown at either 85˚C or 65˚C until optical density measurements of 0.4-0.5 were achieved, representing mid-log phase growth, then rapidly chilled on ice. Chilled cells were harvested via centrifugation (15,000 x *g,* 10 min, 4˚C) and resuspended in 100 µL 25 mM Tris HCl pH 8.0, 500 mM NaCl, 10% glycerol, 2% SDS. Protein concentrations were quantified using a Qubit Protein Assay (Thermo Fisher Scientific). Proteins were resolved through 4 – 20% acrylamide gels, blotted onto PVDF membranes, blocked with 5% BSA, and probed using the primary anti-Tk-RadA antibody (1:10,000 dilution). Blots were washed, then probed with IgG-alkaline phosphatase conjugated goat anti-guinea pig secondary antibodies (1:1000 dilution) and visualized using 1-Step™

NBT/BCIP Substrate Solution (Thermo Fisher Scientific). Western blot bands were quantified with ImageQuant software and plotted in Excel.

UV Sensitivity Assay

Parental (TS559) and AL015 *T. kodakarensis* strains were grown at 85°C in rich media supplemented with 1 mM agmatine sulfate and KOD1- vitamin mixture as previously described to an OD_{600nm} of ~0.4-0.5 before being rapidly harvested via centrifugation (8,000 *x g*, 15 minutes, at 4°C), the supernatant discarded, and the cell pellets resuspended in 150 mL 1x artificial sea water (ASW). The resuspended cells were anaerobically irradiated by exposure to a UV light in 10 mL aliquots at 100 μW/cm^2 for 0, 5, 10, 15, 20, 30, 45, 60, and 90 seconds (resulting in total exposures of 0, 5, 10, 15, 20, 30, 45, 60, and 90 J/m^2, respectively) and immediately put on ice. Irradiated cultures were serially diluted tenfold from 10^{-1} through 10^{-7} and 10 μL from each dilution were plated onto rich media solidified plates supplemented with polysulfides, agmatine sulfate, and KOD1-vitamin mixture. The plates were incubated anaerobically at 85°C for 48 hours. Given that *T. kodakarensis* colonies are flat, pale, and difficult to image conventionally, colony forming unit (CFU) counts were determined after transfer of colonies to PVDF (0.2 μm) and staining of the proteins lysed from transferred cells. PVDF membranes were pre-rinsed with 100% methanol for 10 minutes and pressed onto the colony-containing plates to facilitate colony transfer. The colony-containing membranes were flash frozen with liquid nitrogen to facilitate cell lysis and protein release, then stained with Coomassie Brilliant Blue G-250 for 20 minutes with gentle rocking. The membranes were destained twice in 100% methanol for five minutes and allowed to air dry. The fraction of viable cells at each UV dose was determined by comparing the number of CFUs from countable spots at each UV dose to that of the same dilution spot from the no UV control for each strain. Assays were done in minimally triplicate for each strain.

Marker frequency analysis

Genomic DNAs were isolated at mid-exponential (0.4 – 0.5 OD_{600nm}) and late stationary phase (after the OD_{600nm} reading peaked and modestly decreased). Illumina libraries were prepared using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (New England Biolabs) as directed by the manufacturer. The quality of each library was assessed using an Agilent Bioanalyzer 2100 using an Agilent High Sensitivity Kit (Agilent Technologies). Libraries were pooled in equimolar proportion and sequenced on an Illumina Next-seq instrument, using a Nextseq 1000/2000 P2 reagent kit. Raw sequencing data was uploaded into the Galaxy platform and was processed via fastp for adaptor trimming and filtering low quality reads, HISAT2 for aligning the processed data to the TS559 reference genome, and bamCoverage to bin the TS559 reference genome into non-overlapping 1000 base pair bins for generating a bedGraph file classifying the sequencing reads into the corresponding bin normalized to RPKM. Graphs plotting the $\log_2$ (exponential/stationary) ratios were generated using Excel.

***In vitro* Splicing Assays**

All MIG splicing reporter constructs (Figure 4B,6A) containing either Tk-RadA-intein$^{A(373\text{-}381)}$, Tk-RadA-intein$^{\Delta 373\text{-}381}$, Tk-RadA-intein$^{\Delta 286\text{-}585}$, Tk-RadA-intein$^{\Delta 276\text{-}585}$ or Tk-RadA-intein$^{\Delta 270\text{-}592}$ + *P.ho.* loop were commercially synthesized and sequenced (GenScript, USA) based on previous MIG reporters in the pACYC vector backbone[24]. Construction of the Pho-RadA-intein^{WT} was previously described[24]. All inteins within the MIG reporter are flanked by 10 residues on both the N- and C-terminus from the natural RadA exteins, which are identical in all constructs.

For native extein splicing assays with Tk-RadA-intein$^{\Delta 286\text{-}585}$ (Figure 3.6C), this construct was commercially synthesized and subcloned into the pET45b(+) vector backbone in-frame with an N-terminal His-tag (GenScript, USA). This construct contains the entire natural C-extein, which

forms interactions with the intein that inhibits splicing[23,24], and a deletion in the N-extein (residues 1-112) previously shown to not influence extein-intein interactions or response to ssDNA[23].

Plasmids were transformed into *E. coli* BL21(DE3) and protein expression was induced in mid-log phase by addition of 1 mM isopropyl-b-d-1thiogalactopyranoside (GoldBio). Proteins were expressed for ~20 hours at 15°C and cells were harvested at 4000 x *g*.

For MIG splicing assays, cell pellets were resuspended in 50 mM Tris-HCl pH 8.0, 10% glycerol, and lysed by sonication. Insoluble material was removed by centrifugation and clarified lysates were either examined immediately for splicing during expression or incubated at 50°C as described in Figure 2 and 3. To measure splicing efficiencies, lysates were mixed with Laemmli sample buffer and resolved using 8-16% TGX gels (Bio-Rad). Samples were not heated in Laemmli sample buffer to maintain GFP fluorescence. GFP fluorescence was measured in-gel using an Amersham Imager 680 (GE Healthcare).

For the Tk-RadA-intein$^{\Delta 286\text{-}585}$ in native exteins (Figure 3.6), cell pellets were resuspended in 20 mM Tris-HCl pH 8.0, 500 mM NaCl, 30 mM imidazole, lysed by sonication, and insoluble material removed by centrifugation. Clarified lysates were purified using Ni-Charged MagBeads (Genscript) and dialyzed into 20 mM Tris-HCl pH 8.5, 200 mM NaCl, 10% glycerol. Purified protein was mixed with 187.5 ng/µl ssDNA (Bayou Biolabs) or Tris-EDTA buffer and incubated as described in Figure 6C. Reactions were mixed with SDS sample buffer, resolved through 8-16% Bis-Tris gels (GenScript, USA) and stained with Coomassie brilliant blue dye.

Relative levels of precursor, ligated exteins, and other species were quantified by densitometry using ImageJ (imagej.nih.gov) based on three biological replicates. Average and standard deviation are shown in bar graphs.

REFERENCES

1. Kelman, L. M. & Kelman, Z. Archaeal DNA replication. *Annu Rev Genet* **48,** 71–97 (2014).

2. Raymann, K., Forterre, P., Brochier-Armanet, C. & Gribaldo, S. Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. *Genome Biol Evol* **6,** 192–212 (2014).

3. Ausiannikava, D. & Allers, T. Diversity of DNA Replication in the Archaea. *Genes (Basel)* **8,** 56 (2017).

4. Cubonová, L., Richardson, T., Burkhart, B. W., Kelman, Z., Connolly, B. A., Reeve, J. N. & Santangelo, T. J. Archaeal DNA polymerase D but not DNA polymerase B is required for genome replication in Thermococcus kodakarensis. *J Bacteriol* **195,** 2322–8 (2013).

5. Bell, S. D. Archaeal orc1/cdc6 proteins. *Subcell Biochem* **62,** 59–69 (2012).

6. Beattie, T. R. & Bell, S. D. Molecular machines in archaeal DNA replication. *Curr Opin Chem Biol* **15,** 614–9 (2011).

7. Kunkel, T. A. & Burgers, P. M. J. Arranging eukaryotic nuclear DNA polymerases for replication: Specific interactions with accessory proteins arrange Pols α, δ, and ϵ in the replisome for leading-strand and lagging-strand DNA replication. *BioEssays* **39,** 1700070 Preprint at https://doi.org/10.1002/bies.201700070 (2017)

8. Riera, A., Barbon, M., Noguchi, Y., Reuter, L. M., Schneider, S. & Speck, C. From structure to mechanism— understanding initiation of DNA replication. *Genes Dev* **31,** 1073–1088 Preprint at https://doi.org/10.1101/gad.298232.117 (2017)

9. Bleichert, F., Botchan, M. R. & Berger, J. M. Mechanisms for initiating cellular DNA replication. *Science (1979)* **355,** eaah6317 Preprint at https://doi.org/10.1126/science.aah6317 (2017)

10. Samson, R. Y. Y., Xu, Y., Gadelha, C., Stone, T. A. A., Faqiri, J. N. N., Li, D., Qin, N., Pu, F., Liang, Y. X. X., She, Q. & Bell, S. D. D. Specificity and function of archaeal DNA replication initiator proteins. *Cell Rep* **3,** 485–96 (2013).

11. Costa, A., Hood, I. V. & Berger, J. M. Mechanisms for initiating cellular DNA replication. *Annu Rev Biochem* **82,** 25–54 Preprint at https://doi.org/10.1146/annurev-biochem-052610-094414 (2013)

12. Forterre, P. Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Mol Microbiol* **33,** 457–65 (1999).

13. Gehring, A. M., Astling, D. P., Matsumi, R., Burkhart, B. W., Kelman, Z., Reeve, J. N., Jones, K. L. & Santangelo, T. J. Genome replication in Thermococcus kodakarensis independent of Cdc6 and an origin of replication. *Front Microbiol* **8,** (2017).

14. Gaudier, M., Schuwirth, B. S., Westcott, S. L. & Wigley, D. B. Structural basis of DNA replication origin recognition by an ORC protein. *Science (1979)* **317,** 1213–1216 (2007).

15. Samson, R. Y., Abeyrathne, P. D. & Bell, S. D. Mechanism of Archaeal MCM Helicase Recruitment to DNA Replication Origins. *Mol Cell* **61,** 287–296 (2016).

16. Hawkins, M., Malla, S., Blythe, M. J., Nieduszynski, C. A. & Allers, T. Accelerated growth in the absence of DNA replication origins. *Nature* **503,** 544–547 (2013).

17. Spaans, S. K., van der Oost, J. & Kengen, S. W. M. The chromosome copy number of the hyperthermophilic archaeon Thermococcus kodakarensis KOD1. *Extremophiles* **19,** 741–750 (2015).

18. Seitz, E. M., Brockman, J. P., Sandler, S. J., Clark, A. J. & Kowalczykowski, S. C. RadA protein is an archaeal RecA protein homolog that catalyzes DNA strand exchange. *Genes Dev* **12,** 1248–1253 (1998).

19. Wardell, K., Haldenby, S., Jones, N., Liddell, S., Ngo, G. H. P. & Allers, T. RadB acts in homologous recombination in the archaeon Haloferax volcanii, consistent with a role as recombination mediator. *DNA Repair (Amst)* **55,** 7–16 (2017).

20. Makarova, K. S. & Koonin, E. V. Archaeology of eukaryotic DNA replication. *Cold Spring Harb Perspect Biol* **5,** a012963 (2013).

21. Hogrel, G., Lu, Y., Alexandre, N., Bossé, A., Dulermo, R., Ishino, S., Ishino, Y. & Flament, D. Role of RadA and DNA Polymerases in Recombination-Associated DNA Synthesis in Hyperthermophilic Archaea. *Biomolecules* **10,** 1–17 (2020).

22. Lennon, C. W., Stanger, M., Banavali, N. K. & Belfort, M. Conditional protein splicing switch in hyperthermophiles through an intein-extein partnership. *mBio* **9,** (2018).

23. Lennon, C. W., Stanger, M. & Belfort, M. Protein splicing of a recombinase intein induced by ssDNA and DNA damage. *Genes Dev* **30,** 2663–2668 (2016).

24. Topilina, N. I., Novikova, O., Stanger, M., Banavali, N. K. & Belfort, M. Post-translational environmental switch of RadA activity by extein–intein interactions in protein splicing. *Nucleic Acids Res* **43,** 6631 (2015).

25. Yalala, V. R., Lynch, A. K. & Mills, K. V. Conditional Alternative Protein Splicing Promoted by Inteins from Haloquadratum walsbyi. *Biochemistry* **61,** 294 (2022).

26. Wood, D. W., Belfort, M. & Lennon, C. W. Inteins-mechanism of protein splicing, emerging regulatory roles, and applications in protein engineering. *Front Microbiol* **14,** (2023).

27. Lennon, C. W., Wahl, D., Goetz, J. R. & Weinberger, J. Reactive Chlorine Species Reversibly Inhibit DnaB Protein Splicing in Mycobacteria. *Microbiol Spectr* **9,** (2021).

28. Green, C. M., Li, Z., Smith, A. D., Novikova, O., Bacot-Davis, V. R., Gao, F., Hu, S., Banavali, N. K., Thiele, D. J., Li, H. & Belfort, M. Spliceosomal Prp8 intein at the crossroads of protein and RNA splicing. *PLoS Biol* **17,** (2019).

29. Topilina, N. I., Green, C. M., Jayachandran, P., Kelley, D. S., Stanger, M. J., Piazza, C. L., Nayak, S. & Belfort, M. SufB intein of Mycobacterium tuberculosis as a sensor for oxidative and nitrosative stresses. *Proc Natl Acad Sci U S A* **112,** 10348–10353 (2015).

30. Mills, K. V. & Paulus, H. Reversible inhibition of protein splicing by zinc ion. *J Biol Chem* **276,** 10832–10838 (2001).

31. Reitter, J. N., Cousin, C. E., Nicastri, M. C., Jaramillo, M. V. & Mills, K. V. Salt-Dependent Conditional Protein Splicing of an Intein from Halobacterium salinarum. *Biochemistry* **55,** 1279–1282 (2016).

32. Woods, D., Vangaveti, S., Egbanum, I., Sweeney, A. M., Li, Z., Bacot-Davis, V., Lesassier, D. S., Stanger, M., Hardison, G. E., Li, H., Belfort, M. & Lennon, C. W. Conditional DnaB Protein Splicing Is Reversibly Inhibited by Zinc in Mycobacteria. *mBio* **11,** 1–14 (2020).

33. Callahan, B. P., Topilina, N. I., Stanger, M. J., Van Roey, P. & Belfort, M. Structure of catalytically competent intein caught in a redox trap with functional and evolutionary implications. *Nature Structural & Molecular Biology 2011 18:5* **18,** 630–633 (2011).

34. Lennon, C. W., Stanger, M. J. & Belfort, M. Mechanism of Single-Stranded DNA Activation of Recombinase Intein Splicing. *Biochemistry* **58,** 3335 (2019).

35. Lennon, C. W. & Belfort, M. Inteins. *Current Biology* **27,** R204–R206 (2017).

36. Novikova, O., Jayachandran, P., Kelley, D. S., Morton, Z., Merwin, S., Topilina, N. I. & Belfort, M. Intein clustering suggests functional importance in different domains of life. *Mol Biol Evol* **33,** 783–799 (2016).

37. Pavankumar, T. L. Inteins: Localized Distribution, Gene Regulation, and Protein Engineering for Biological Applications. *Microorganisms* **6,** (2018).

38. Maeder, D. L., Weiss, R. B., Dunn, D. M., Cherry, J. L., González, J. M., Diruggiero, J., Robb, F. T., Niederhausern, V., Aoyagi, A., Mahmoud, M., Hannenhalli, S., Lupas, A. N., Koretke, K. K. & Diruggiero, J. Divergence of the hyperthermophilic archaea Pyrococcus furiosus and P. horikoshii inferred from complete genomic sequences. *Genetics* **152,** 1299 (1999).

39. Nishioka, M., Fujiwara, S., Takagi, M. & Imanaka, T. Characterization of two intein homing endonucleases encoded in the DNA polymerase gene of Pyrococcus kodakaraensis strain KOD1. *Nucleic Acids Res* **26,** 4409 (1998).

40. Weinberger Ii, J. & Lennon, C. W. Monitoring Protein Splicing Using In-gel Fluorescence Immediately Following SDS-PAGE. *Bio Protoc* **11,** e4121 (2021).

41. Naor, A., Altman-Price, N., Soucy, S. M., Green, A. G., Mitiagina, Y., Turgeman-Grotta, I., Davidovich, N., Gogarten, J. P. & Gophna, U. Impact of a homing intein on recombination frequency and organismal fitness. *Proc Natl Acad Sci U S A* **113,** E4654–E4661 (2016).

42. Robinzon, S., Cawood, A. R., Ruiz, M. A., Gophna, U., Altman-Price, N. & Mills, K. V. Protein Splicing Activity of the Haloferax volcanii PolB-c Intein Is Sensitive to Homing Endonuclease Domain Mutations. *Biochemistry* **59,** 3359–3367 (2020).

43. Andersson, A. F., Pelve, E. A., Lindeberg, S., Lundgren, M., Nilsson, P. & Bernander, R. Replication-biased genome organisation in the crenarchaeon Sulfolobus. *BMC Genomics* **11,** 1–7 (2010).

44. Gehring, A. M., Astling, D. P., Matsumi, R., Burkhart, B. W., Kelman, Z., Reeve, J. N., Jones, K. L. & Santangelo, T. J. Genome replication in Thermococcus kodakarensis independent of Cdc6 and an origin of replication. *Front Microbiol* **8,** 2084 (2017).

45. Tagashira, K., Fukuda, W., Matsubara, M., Kanai, T., Atomi, H. & Imanaka, T. Genetic studies on the virus-like regions in the genome of hyperthermophilic archaeon, Thermococcus kodakarensis. *Extremophiles* **17,** 153–60 (2013).

46. Lennon, C. W. & Belfort, M. *Inteins*. *Current Biology* **27,** R204–R206 (Curr Biol, 2017).

47. Green, C. M., Novikova, O. & Belfort, M. The dynamic intein landscape of eukaryotes. *Mob DNA* **9,** 4 (2018).

48. Tharappel, A. M., Li, Z. & Li, H. Inteins as Drug Targets and Therapeutic Tools. doi:10.3389/fmolb.2022.821146

49. Wall, D. A., Tarrant, S. P., Wang, C., Mills, K. V & Lennon, C. W. Intein Inhibitors as Novel Antimicrobials: Protein Splicing in Human Pathogens, Screening Methods, and Off-Target Considerations. *Front Mol Biosci* **8,** 752824 (2021).

50. Liman, G. L. S., Stettler, M. E. & Santangelo, T. J. Transformation Techniques for the Anaerobic Hyperthermophile Thermococcus kodakarensis. *Methods Mol Biol* **2522,** 87–104 (2022).

51. Carbone, A., Fioretti, F. M., Fucci, L., Ausió, J. & Piscopo, M. High efficiency method to obtain supercoiled DNA with a commercial plasmid purification kit. *Acta Biochim Pol* **59,** 275–278 (2012).

CHAPTER 4: DUALITY IN ARCHAEAL DNA POLYMERASES

Summary

DNA polymerase (DNAP) is the central DNA replication enzyme responsible for the synthesis of nascent DNA strands, which is a necessity in all lifeforms. Through evolution, Archaea have acquired two replicative DNAPs: B-family DNAP (PolB) and D-family DNAP (PolD). Extensive research has been performed on *in vitro* characterization of the two archaeal DNAPs, but the question still persists on what the exact biological function(s) of these polymerases are. Here we genetically modify the hyperthermophilic anaerobic model archaeon, *Thermococcus kodakarensis*, to demonstrate (1) the biological significance of deleting Tk-PolB at non-optimum growth temperatures (55˚C, 65˚C, and 75˚C), (2) the peculiar synthetic lethality resulting from the reintroduction of exonuclease and steric gate mutants of Tk-PolB, and (3) the increased incorporation of genomic ribonucleotides correlated to the activity of the mutant Tk-PolB.

Introduction

Deoxyribonucleic acid (DNA) is the fundamental building blocks of all known life. The act of replicating DNA is essential to maintain the passage of DNA from the parental cells down to the daughter cells. Central to this process is activity of the DNA polymerase (DNAP), which is primarily tasked with replicating the parental template DNA strand and synthesizing the nascent DNA strand from the available deoxyribonucleotide triphosphates (dNTPs). DNAP is not only integral for DNA replication but also the maintenance of the DNA integrity through various DNA repair processes.

Although this critical activity is conserved in all living organisms, DNA replication and repair have been shown to be catalyzed by domain-specific, non-orthologous replication machineries[1]. This notion is especially true when comparing polymerases from specific organisms in Bacteria, Archaea, and Eukarya[1–3]. To date, based on their phylogeny, the three Domains and the various

categories of viruses utilize seven different families of DNAP: PolA, PolB, PolC, PolD, PolE, PolX, PolY (Figure 4.1), and the enigmatic Reverse transcriptase[4]. The main replicative polymerases in each Domain are distinct: Bacteria utilize A- and C-families DNAP; Eukarya utilize B-family DNAP; and Archaea utilize B- and/or D-families DNAP[4].

DNAPs are evaluated based on their processivity and fidelity to replicate the template DNA[5–8]. Processivity refers to the number of incorporated dNTPs within a single event of association/ dissociation[5,6]. Low processivity DNAP, on average, will bind template DNA for a shorter period before falling off, incorporate a lower amount of dNTP, and synthesize a shorter nascent DNA strand in a single binding event compared to the higher processivity DNAP. Fidelity refers to the accuracy of DNAP to incorporate the correct nucleotides during replication[7,8]. Lower fidelity DNAP have a higher propensity to incorporate the wrong dNTP without correcting the mistake compared to higher fidelity DNAP. DNAP can achieve higher fidelity by two means, which are higher nucleotide selectivity and proofreading (the ability of DNAP to detect and 3´→5´ exonuclease digest miss incorporated nucleotides from the nascent DNA)[7]. Each family of DNAP has distinct features that contribute to processivity and fidelity.

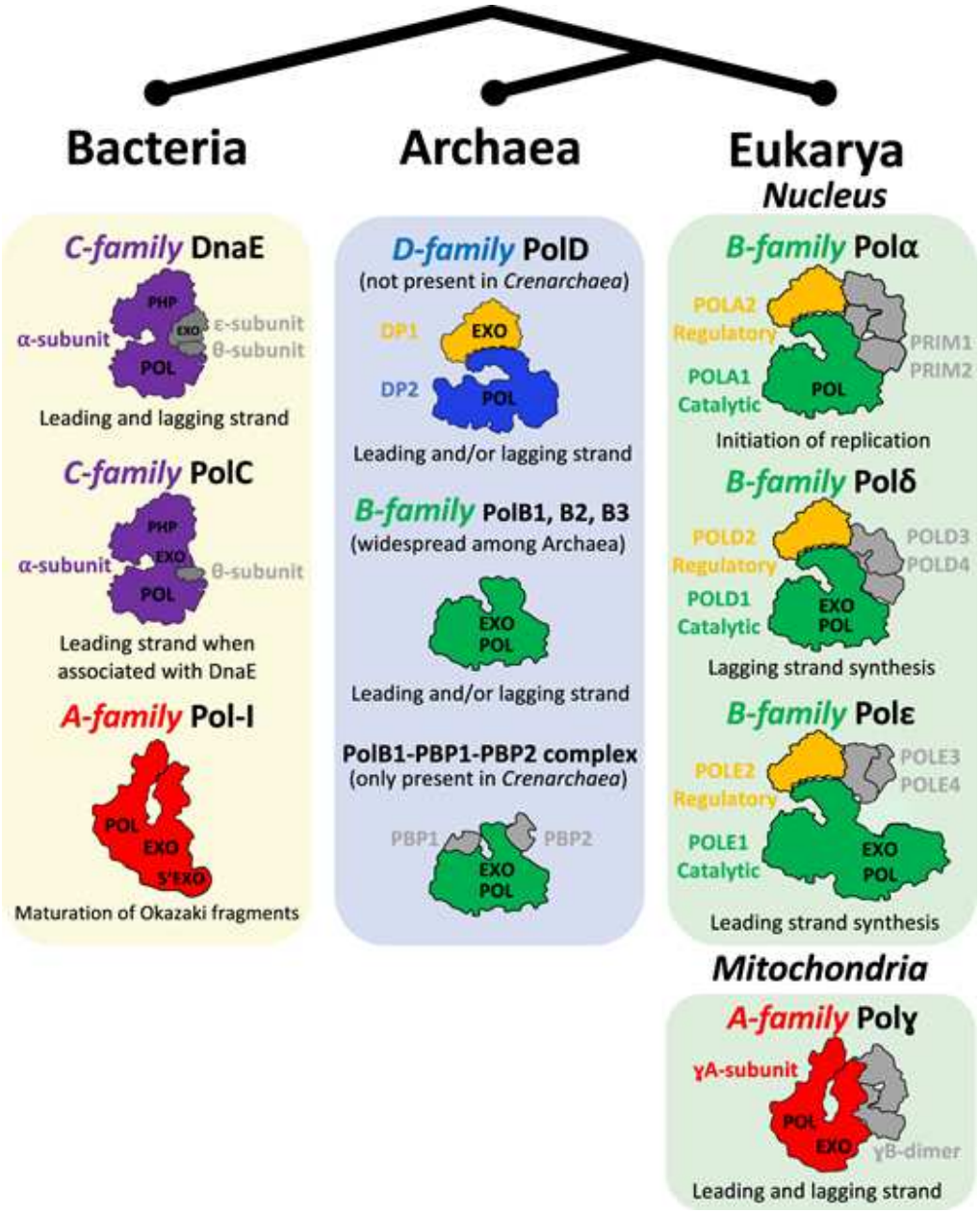


Figure 4.1. Distribution of replicative DNA polymerases in the three domains of life. Adapted from Raia et al. (2019)

Table 4.1. Comparison of the two replicative archaeal DNA polymerases. * = only in the presence of PCNA.

Properties	PolB	PolD
Subunit(s)	1	2
3´→5´ exonuclease	Yes	Yes
Bind primed DNA	Yes	Yes
Bind ssDNA	Yes	No
Elongate DNA primer	Yes	Yes
Elongate RNA primer	Yes	Yes
Displacement DNA synthesis	Yes	Yes*

The two main families of DNAP in Archaea, PolB and PolD[4], are a contrast to each other (Table 4.1). While archaeal B-family DNAP shares sequence and structural homology to the eukaryotic B-family DNAP, the D-family DNAP is unique to Archaea. Archaeal PolB is a one subunit DNAP containing both polymerase and 3´→5´ exonuclease activity in one gene[9]. In contrast, archaeal PolD is a two subunit DNAP, containing a large subunit responsible for polymerase activity and a small subunit responsible for 3´→5´ exonuclease activity[9]. Although PolD is a two subunit enzyme, it requires both subunits in order to be a functional DNAP[10]. Previous studies have evaluated the processivity and fidelity of the two archaeal DNAP *in vitro* and showed that both PolB and PolD can bind primed DNA templates and synthesize nascent DNA from synthetic DNA templates[11]. Several points to note, while both DNAP seems similar to each other in their main properties, they also showed unique properties separating the two, which are: (1) while PolD showed strong preference for binding primed DNA templates, PolB binds both primed and single-stranded DNA templates. (2) Although both DNAPs showed ability to do strand displacement DNA synthesis in the presence of PCNA, only PolB is capable to strand displacement DNA synthesis without the presence of PCNA, unlike PolD[11]. The two unique

properties separating PolB and PolD suggest a diverging function of the two polymerases *in vivo.*

The model organism *Thermococcus kodakarensis* is an anaerobic hyperthermophilic archaeon and naturally encode for the two distinct and evolutionarily unrelated DNAPs, Tk-PolB (TK0001) and Tk-PolD (TK1902 (small subunit) and TK1903 (large subunit)). Both DNAPs are co-expressed with other genes in the same operon; while Tk-PolB is in the same operon as multiple hypothetical proteins (TK2305 and TK2306), Tk-PolD subunits are in the same operon as Tk-Cdc6 (TK1901), the sole *ori* binding and DNA replication initiator protein in *T. kodakarensis* (Figure 4.2B-C). Dissecting further into the sequences of both DNAPs showed the existence of inteins mobile genetic elements invasion into Tk-PolB and Tk-PolD large subunit. Previous studies have shown Tk-PolB is dispensable from the highly polyploid *T. kodakarensis* genomes, but both subunits of Tk-PolD are essential[12,13]. Further biochemical analysis of Tk-PolB deletion strains showed that although the strains are growing at a similar rate as the parental strains, these deletion strains are highly sensitive to DNA damaging reagents, such as ultraviolet light (UV), Mitomycin C (MMC), and methyl methanesulfonate (MMS)[12,13]. These findings suggest the following: (1) Tk-PolD is the sole main replicative polymerase in *T. kodakarensis,* that can do both leading- and lagging-strand synthesis during DNA replication. (2) Tk-PolB is not essential but important for instances wherein DNA repair is necessary.

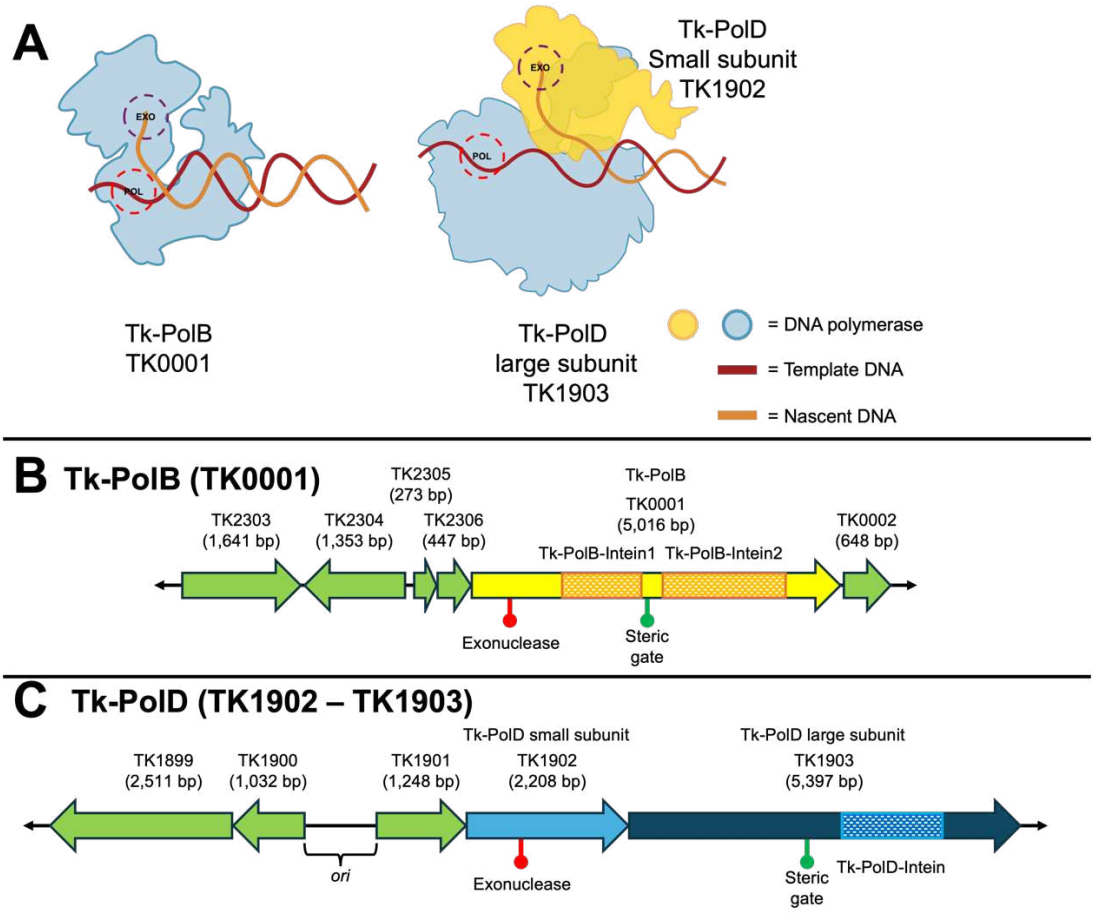


Figure 4.2. Tk-PolB and Tk-PolD genomic loci in *T. kodakarensis.* (**A**) Cartoon depictions of the one subunit Tk-PolB and the two subunit Tk-PolD. The relative exonuclease and polymerase active sites are circled. (**B**) Tk-PolB gene and its surrounding genes are depicted in yellow arrow and green arrows, respectively. Areas of interest which include the exonuclease, steric gate, and inteins are depicted in red point, green point, and orange dashed boxes, respectively. (**C**) Tk-PolD small subunit, Tk-PolD large subunit, and their surrounding genes are depicted in light blue arrow, dark blue arrow, and green arrows, respectively. Areas of interest which include the exonuclease, steric gate, intein, and *ori* are depicted in red point, green point, blue dashed box, and bracket, respectively.

The duality of DNAP in Archaea is still enigmatic and the main aims of this study will revolve around delineating the function(s) of both PolB and PolD in *T. kodakarensis.* Here we show Tk-PolB might play a bigger role *in vivo* at sub-optimal temperatures. Furthermore, through genetic approaches targeting the Tk-PolB locus, our data suggests that the gene is highly active *in vivo*, contradicting the notion that Tk-PolB is only utilized in DNA repair.

Results

ΔTk-PolB strain growth is inhibited at sub-optimal temperatures

To date, Tk-PolB has been shown to be dispensable from the *T. kodakarensis* genome, and deletion of the gene did not result in any perceivable growth defect. The impact of deleting Tk-PolB becomes apparent when the cells are subjected to DNA damaging reagents, UV, MMC, and MMS. We first recapitulate the previous finding wherein ΔTk-PolB strain growth is unimpacted by the lack of Tk-PolB in the cells at 85˚C. Surprisingly, we observed a dramatic growth defect for ΔTk-PolB strain not at the normal 85˚C optimal growth temperature of *T. kodakarensis* but at 75˚C, 65˚C, and 55˚C sub-optimal growth temperatures (Figure 4.3). This finding suggests Tk-PolB might play a bigger role at sub-optimal growth temperatures.

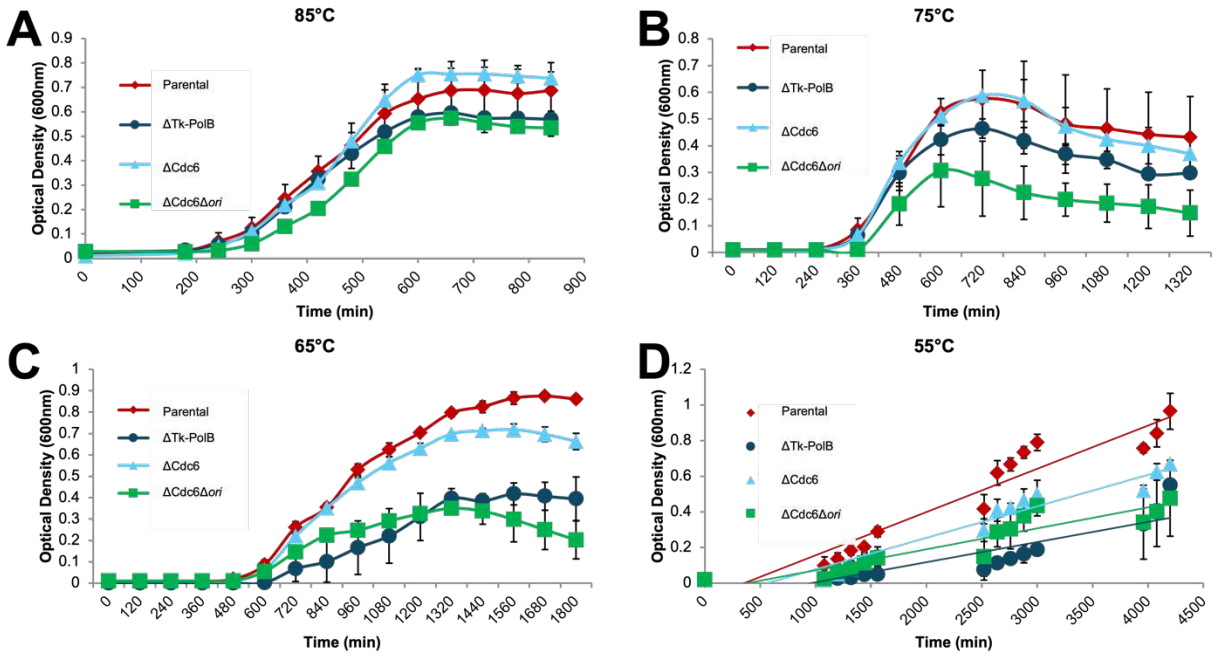


Figure 4.3. Growth defects for ΔTk-PolB at suboptimal temperatures. (**A-D**) Growth of triplicate biological replicates of strains of *T. kodakarensis*; Wildtype (red line, diamond points), ΔTk-PolB (dark blue line, circle points), ΔCdc6 (light blue line, triangle points), and ΔCdc6Δ*ori* (green line, square points), monitored by changes in optical density (OD_{600nm}) at 85˚C (**A**), 75˚C (**B**), 65˚C (**C**), and 55˚C (**D**).

We did further analysis to compare the growth rate of ΔTk-PolB to other deletion strains with similar phenotypes at optimal and lower temperatures. We found that the growth properties of ΔTk-PolB matched very closely to the ΔCdc6Δ*ori,* in which both deletion strains showed no growth defect at 85˚C but showed significant growth deficiency at lower temperatures (Figure 4.3). Tk-PolB, although dispensable, has been shown to be important for growth at sub-optimal temperatures akin to strain lacking ΔCdc6Δ*ori* and sensitivity to DNA damaging reagents.

Tracking *in vivo* activity of Tk-PolB using RADAR-seq

Tracking activities of DNAP *in vivo* using next-generation sequencing (NGS) methods have been done in both Bacteria and Eukarya. The first to utilize NGS approach to track the activity of a DNAP is done by the Kunkel lab to track the activity of the different DNAPs in Yeast[14]. The latest NGS method developed to successfully track DNAP activity was developed in the Gardner lab to track the activity of *Escherichia coli* PolI on the replication forks[15,16]. The methods, termed Hydrolytic End Sequencing (HydEn-Seq) and RAre DAmage and Repair sequencing (RADAR-seq,) both points to three main requirements to track DNAP *in vivo* in the strains: (1) ribonucleotide excision repair (RER) must be deficient, (2) Mutant DNAP is deficient in 3´→5´ exonuclease, and (3) Mutant DNAP is promiscuously incorporating ribonucleotide during nascent DNA synthesis[14–16].

Ribonucleotide triphosphates (NTPs) can sometimes get incorporated by mistake. When left unfixed, ribonucleotides (rNMPs) in the genome can induce genomic instability and replicative stress through their reactive 2'-hydroxyl group[17–19]. Life has evolved to efficiently repair this genomic incorporation of rNMPs[20–22]. Although conserved in all life is the RER pathway, the enzymes responsible in this pathway are domain-specific and non-orthologous. The steps of RER pathway involve: (1) nicking of the embedded rNMP by the RNaseH2 enzymes, (2)

extension of the nascent nick by DNAP, (3) displacement of the rNMP by DNAP's displacement DNA synthesis ability to create a flap, (4) nuclease cleavage of the rNMP containing flap, and (5) ligation of the nascent elongated nick[20–22].

The RER pathway in Archaea was first reconstituted and established in the model archaeon, *T. kodakarensis.* As shown in Yeast, the deficiency of RER in *T. kodakarensis* can also be achieved via targeted genetic deletion of the TK0805 locus (Figure 4.4A-B) encoding for Tk-RNaseH2[22,23]. Deletion of Tk-RNaseH2 (ΔTk-RNaseH2) did not show any growth deficiency compared to the parental strain, similar to the ΔTk-PolB strain at 85˚C[23]. Further characterization of the ΔTk-RNaseH2 strain is done by alkaline gel electrophoresis to qualitatively check for an increase of rNMP incorporation in genomic DNA. Indeed a single gene deletion in ΔTk-RNaseH2 showed a visible increase in degradation product under alkaline conditions, suggesting deficient RER in this strain.

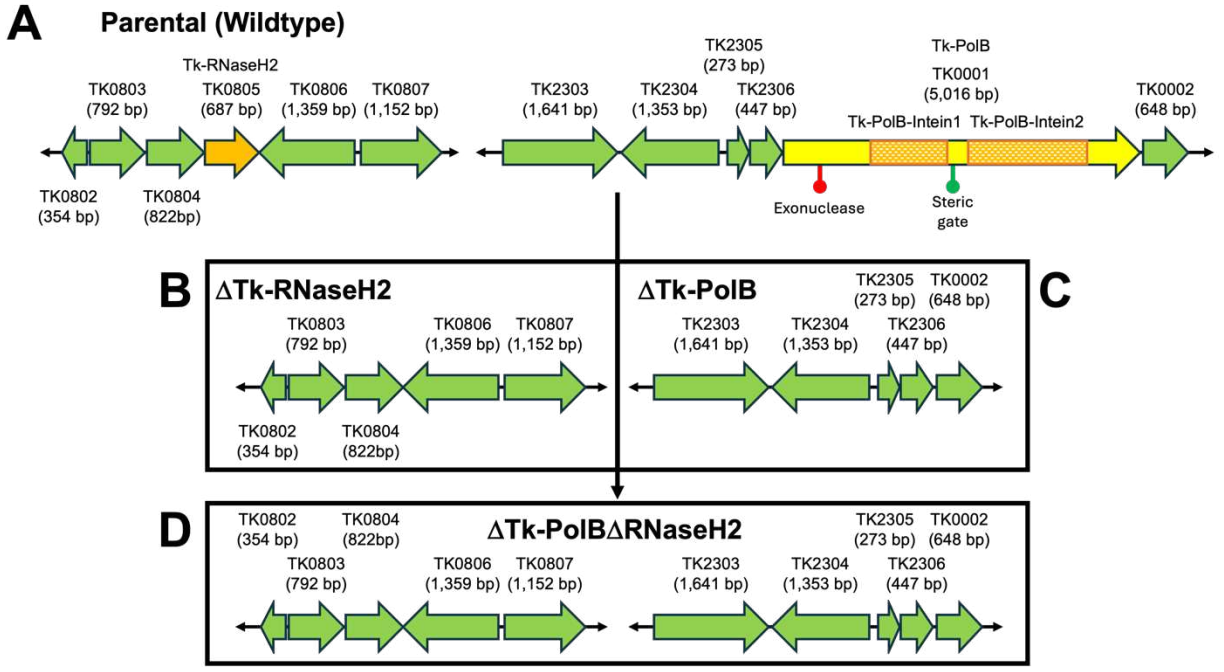


Figure 4.4. Strain construction genomic loci for RADAR-seq in *T. kodakarensis.* (**A**) Parental strain genomic loci for Tk-RNaseH2 (TK0805) and Tk-PolB (TK0001) depicted in orange arrow and yellow arrow, respectively. (**B-D**) Markerless deletion of Tk-RNaseH2 (**B**), Tk-PolB (**C**), and both (**D**) allows for specific gene targeting without addition of a selectable marker in the final strains and allows us to recycle the selectable markers.

The successful generation of the RER deficient strain of ΔTk-RNaseH2 satisfied the first requirement for tracking the *in vivo* activity of DNAP. The second and final steps require us to introduce a mutated DNAP with two main properties, which are exonuclease deficient and steric gate deficient. Although DNAP has an inherent property to exclude rNTPs from the active site by steric hindrance of the steric gate residue, eukaryotic DNAP and bacterial DNAP still incorporates 1 rNMP per 1,000 bp and 2,300 bp during replication, respectively[17,24–26]. Mutation of the steric gate residue within DNAPs in Yeast and *E. coli* has been previously shown to increase their propensity to miss incorporated rNTPs into the genome during DNA synthesis[14–17].

In *T. kodakarensis,* the presence of two inteins within Tk-PolB, wherein the steric gate residue is located 3 amino acids away from the splicing junction of Tk-PolB-intein1, becomes a major complication. Tk-PolB is expressed as an inactive precursor, and post translational protein splicing of the precursor is required to ultimately get the active mature Tk-PolB isoform. Mutation to residues near the splicing junctions has been shown to impact protein splicing. To bypass the problem, we modified our approach to start by deleting the genomic copy of Tk-PolB along with both inteins from the parental strain, and subsequently, reintroducing the mutant Tk-PolB without the two inteins, termed Tk-PolB$^{-exo/-steric/-intein}$.

Single deletion of Tk-PolB and Tk-RNaseH2 (Figure 4.4B-C) can be achieved without much impact on growth compared to the parental strain at 85˚C[12,13,22,23]. We successfully combined the two deletion constructs to generate a double deletion strain lacking both Tk-PolB and Tk-

RNaseH2, which is called ΔTk-PolBΔTk-RNaseH2 (Figure 4.4D). The successful construction of ΔTk-PolBΔTk-RNaseH2 allows us to reintroduce Tk-PolB$^{-exo/-steric/-intein}$ without any worry about compromised protein splicing. However, our attempts to reintroduce Tk-PolB$^{-exo/-steric/-intein}$ into ΔTk-PolBΔTk-RNaseH2 at the original locus of Tk-PolB results in either lethality or extra mutations within the mutant Tk-PolB$^{-exo/-steric/-intein}$ coding region rendering the gene inactive or non-functional. Overall, Tk-PolB appears to be not essential for *T. kodakarensis*, but reintroduction of Tk-PolB$^{-exo/-steric/-intein}$ appears to be synthetically lethal.

Previous approaches to introduce the mutant Tk-PolB$^{-exo/-steric/-intein}$ require homologous recombination of the mutant construct into the genomic locus previously hosting the wildtype Tk-PolB gene. We postulate additional mutations accumulating within the coding region of the mutant Tk-PolB$^{-exo/-steric/-intein}$ are the resultant of the homologous recombination process. We devised another novel approach to circumvent the current problem by reintroducing the mutant Tk-PolB$^{-exo/-steric/-intein}$ exogenously using pTS543-based plasmid, an autonomously replicating plasmid for *T. kodakarensis,* eliminating the need for homologous recombination. We were able to successfully reintroduce the plasmids carrying exogenous ectopic expression of Tk-PolB$^{-exo/-steric/-intein}$, encoding for Tk-PolB with D141A/E143A (exonuclease inactivation) and the Y409Y (silent steric gate mutant), Y409V (steric gate mutant), Y409L (steric gate mutant) (Figure 4.5).

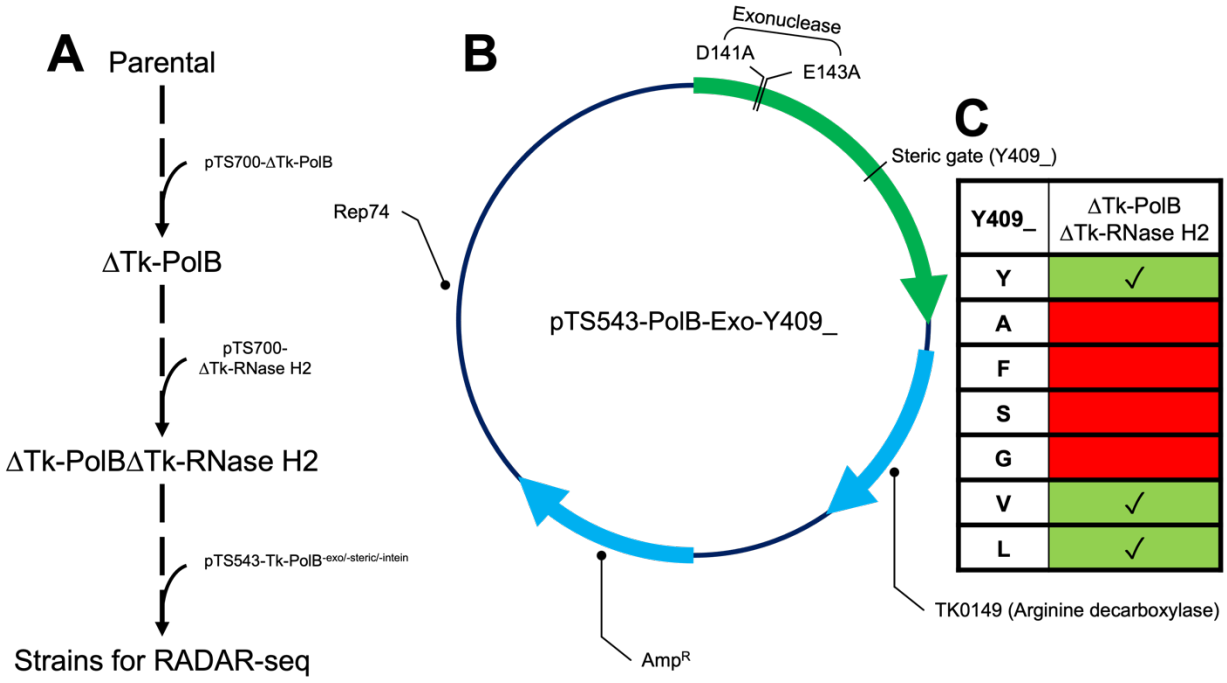


Figure 4.5. Reintroduction of the mutant Tk-PolB$^{-exo/-steric/-intein}$ using extopic expression plasmids. (**A**) The modified approach to reintroduce Tk-PolB$^{-exo/-steric/-intein}$ into the ΔTk-PolBΔTk-RNaseH2 background strain showing the incorporation of pTS543-Tk-PolB$^{-exo/-steric/-intein}$. (**B**) Detailed diagram of the pTS543-Tk-PolB$^{-exo/-steric/-intein}$ construct used in the new approach. (**C**) Table of successfully reintroduced plasmids containing the Tk-PolB$^{-exo/-steric/-intein}$ mutations. Strains confirmation were done via MinION Nanopore sequencing showing no additional mutations in the coding region of the mutant Tk-PolB$^{-exo/-steric/-intein}$.

The successful reintroduction of the mutant Tk-PolB$^{-exo/-steric/-intein}$ is predicted to increase rNMP incorporation into the genome of *T. kodakarensis.* As mentioned previously, the steric gate residues in DNAPs function through steric hindrance to exclude rNMPs from the active site. In Tk-PolB, the tyrosine at the 409th position is predicted to be the steric gate residue. Mutation of the Y409 residue in Tk-PolB into a less bulky amino acid has been shown to increase rNMPs incorporation *in vitro*[27]. We employ alkaline gel electrophoresis to qualitatively determine the increased rNMPs in the genome by the reintroduction of the mutant Tk-PolB$^{-exo/-steric/-intein}$ (Figure 6A). In short, rNMP is highly volatile, and induced nicks due to high pH, as well as degradation products from rNMP incorporation, can be visualized using alkaline gel electrophoresis. As predicted, although genomic DNA of strains lacking Tk-RNaseH2 are already showing an

increase in the degradation products under the alkaline condition, genomic DNA from strains with the mutant Tk-PolB$^{-exo/-steric/-intein}$ showed drastically more degradation product suggesting higher rNMP embedded within the genomic DNA (Figure 4.6E-F). Surprisingly, reintroduction of the Tk-PolB$^{-exo/-intein}$ with only the exonuclease deficiency alone and without the steric gate mutation is sufficient to see the increase in the degradation product from the alkaline condition (Figure 4.6D). Focusing on the Tk-PolB$^{-exo/-steric/-intein}$ variants with Y409L and Y409V mutations, subjecting the genomic DNA from these two strains showed even bigger shift with the latter showing the biggest difference in degradation products related to the alkaline condition compared to the control strains (Figure 4.6E-F). Our attempts to reintroduce the other mutant Tk-PolB$^{-exo/-steric/-intein}$ with even less bulky amino acids than valine or leucine were unsuccessful suggesting the synthetic lethality phenotype perhaps due to excess amount of rNMP incorporation.

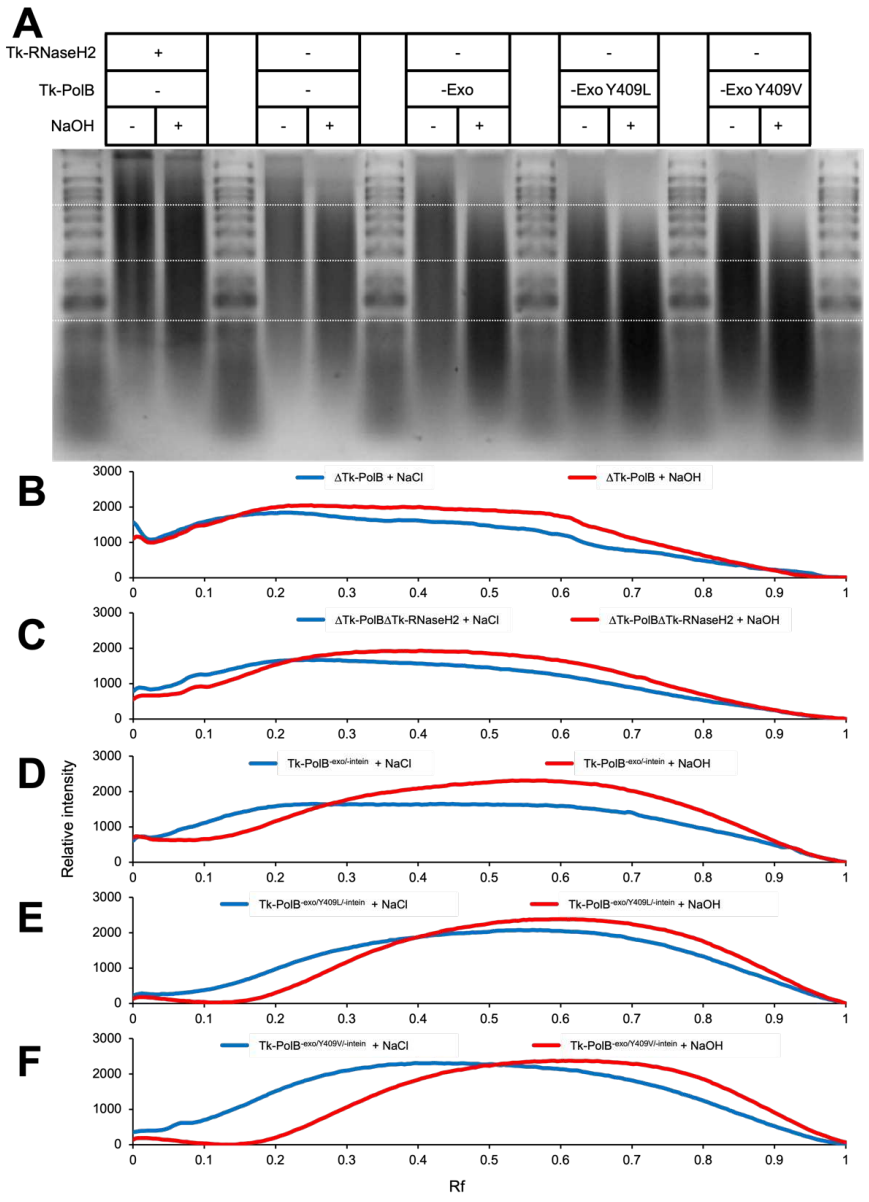


Figure 4.6. Alkaline agarose gel electrophoresis assessment showing increased rNMP incorporation into genomic DNA. (**A**) Alkaline gel electrophoresis were done on genomic DNA isolated from ΔTk-PolB, ΔTk-PolBΔTk-RNaseH2, Tk-PolB$^{-exo/-intein}$, Tk-PolB$^{-exo/Y409Y/-intein}$, and Tk-PolB$^{-exo/Y409V/-intein}$ strains, from left to right without treatment (left lane) and with NaOH treatment (right lane). (**B-F**) Quantification of the relative intensities from the alkaline agarose gel electrophoresis (**A**) for each strains were plotted against the retardation factor (Rf).
So far we have been able to generate *T. kodakarensis* strains with the following mutations: (1) RER deficiency, (2) exonuclease deficient Tk-PolB, and (3) steric gate deficient Tk-PolB. Ultimately, combining the three major mutations allows us to correlate areas of rNMP incorporations in the *T. kodakarensis* genome to the *in vivo* activity of the mutant Tk-PolB using the RADAR-seq. We are currently waiting for the RADAR-seq data from our collaborators at New England Biolabs.

Discussions and Future Directions

Although various scientific disciplines have uncovered a lot about DNAPs within the three Domains, there are still unknowns especially within the archaeal DNAPs. Most literature on DNAP for the archaeal Domain are overly focused on *in vitro* characterization of these enzymes and less about their *in vivo* significance. Our study collectively and extensively endeavors to characterize and delineate the *in vivo* function(s) of the two replicative polymerases in the hyperthermophilic anaerobic model organism, *T. kodakarensis.*

The duality of DNAPs in *T. kodakarensis* was the catalyst that piqued our interest in figuring out the biological purpose of each polymerase. Both the single subunit Tk-PolB and the two subunits Tk-PolD are active *in vitro* with relatively few differentiating features. Our study seems to point out the discrepancies of the *in vitro* data compared to the reality of *in vivo* conditions. Previous *in vitro* characterization of recombinantly purified archaeal DNAPs were are done at temperatures <85˚C, sub-optimum temperatures, perhaps due to *in vitro* parameters that prevents the assay from working properly at higher temperatures[9,27,28]. Although both Tk-PolD and Tk-PolB are functional polymerases, only Tk-PolD is shown to be essential. Most importantly, monitoring the growth rate of ΔTk-PolB not only at the optimum *T. kodakarensis* growth temperature (85˚C) but also at sub-optimum temperatures suggests functional relevancy of Tk-PolB is much more detrimental to the cells at these lower temperatures. Although *T. kodakarensis* mainly grown in laboratories at 85˚C, it also showed the ability to grow at a relatively wide range of temperatures, from 45˚C – 95˚C. We often neglected the idea of how our idealized laboratory conditions could impact our findings. These sub-optimum growth temperatures are more relevant to the temperatures selected for previously published *in vitro* assays. Based on this study, probing growth phenotypes of new strains or older strains of *T. kodakarensis* at non-optimum growth conditions might give some functional insight on dispensable genes without growth impact upon their deletion at optimum growth conditions.

Archaeal DNAPs have been extensively utilized in biotechnological purposes but their *in vivo* function(s) are still elusive. In *T. kodakarensis*, deleting the genomic copy of Tk-PolB resulted in a strain that is more sensitive to DNA damaging reagents[12,13]. Through genetic targeting we are able to achieve a couple milestones: (1) deleted Tk-PolB, (2) deleted Tk-RNaseH2, and (3) reintroduced the mutant Tk-PolBs. These mutations combined allow us to monitor the activity of Tk-PolB as a bulk through alkaline gel electrophoresis and in the near future also map the activity of Tk-PolB at the nucleotide level through RADAR-seq. Unfortunately, major hold back for progress currently focused on getting RADAR-seq to work again since 2022 until now.

Tk-PolB is the non-essential replicative DNAP in *T. kodakarensis,* whereas Tk-PolD is the only essential DNAP in *T. kodakarensis.* The biological activity of the archaeal specific family-D DNAP has only been inferred based on its essentiality and *in vitro* characterization. In theory, RADAR-seq can also be used to track the activity of PolD *in vivo.* The current approach used to obtain the Tk-PolB strain cannot be used for Tk-PolD mainly due to the essentiality of the gene itdelf and also its surrounding genes. We developed a strategy to specifically target Tk-PolD using a plasmid containing the whole operon of Tk-PolD consisting of the *ori,* Tk-Cdc6, TkPolD small subunit, and Tk-PolD large subunit (Figure 4.7). The operon of Tk-PolD is then flanked by two selectable markers allowing for a stringent selection for transformation, this plasmid is termed p190X and will be used to target the genomic copy of Tk-PolD. This approach prevents lethality induced from dysregulation of the Tk-PolD operon which also would subsequently affect the *ori.*

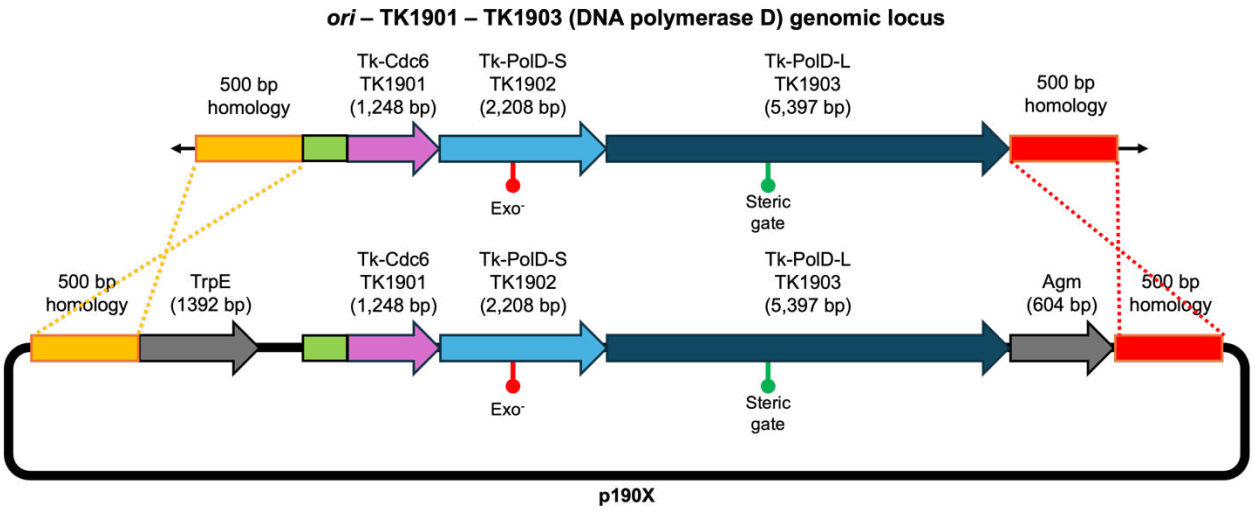


Figure 4.7. Schematic of p190X to target Tk-PolD. The operon of Tk-PolD includes the *ori,* Tk-Cdc6 (TK1901), Tk-PolD small subunit (TK1902), and Tk-PolD large subunit are denoted in green box, magenta arrow, light blue arrow, and dark blue arrow, respectively. The two selectable markers are denoted in gray arrows. The plasmid copy of the Tk-PolD operon carrying the mutations in the Tk-PolD genes will integrate into the genome through the 500 bp upstream and downstream homology denoted in orange box and red box, respectively.

The successful construction of p190X allows us to introduce mutations in both subunits of Tk-PolD at the same time without disruption of the Tk-PolD operon. We narrowed down the residues responsible for the exonuclease and steric gate capacities of Tk-PolD from previous *in vitro* studies for other PolD from a different model archaeon, *Pyrococcus abyssi* and *Thermococcus* sp. 9°N[10,29–31]. The steric gate residue of Tk-PolD is very straight forward with only one residue mapped to the histidine at position 932 in the Tk-PolD large subunit (TK1903). Mapping the exonuclease residue for Tk-PolD is tricky with multiple studies pointing at different residues to target to eliminate the exonuclease activity of Tk-PolD. We have mapped the conserved residue in Tk-PolD small subunit (TK1902) shown to be responsible for the exonuclease activity of the enzyme (Table 4.2). We are still in the process of introducing the exonuclease and steric gate mutations into the genomic copy of Tk-PolD using the p190X plasmids. Hopefully in the near future, we can also track the activity of Tk-PolD *in vivo.*

Table 4.2. List of the exonuclease residues in PolD previously targeted *in vitro.*

Thermococcus kodakarensis	*Thermococcus sp.9˚N*	*Pyrococcus abyssi*	Source
H564	H554	H451	Greenough et al. 2014
D517	D507	D404	Greenough et al. 2014
D473 / H475	-	D360 / H362	Takashima et al. 2019

Methods

T. kodakarensis strain construction

Strains used in this study are listed in Table 4.3. Deletion of Tk-PolB and Tk-RNaseH2 were done by homologous recombination based genetic targeting resulting in markerless deletions at the desired gene loci as previously described[32–34]. Newly constructed strains were confirmed via (1) diagnostic PCR screening and (2) whole genome sequencing (WGS) at >100X coverage using Oxford Nanopore MinION sequencing.

Table 4.3. Tk-PolB strains constructed for this study.

Strains	Details	Comments
TS746	∆Tk-PolB	
TS747	∆Tk-PolB∆RNaseH2	
AL006	183-Exo-747: TS747 + pAL-PolB-Exo	Exo
AL007	189-747L: TS747 + pAL-PolB-Y409L	Exo, Y409L, **extra mutations**
AL008	180-Exo-746: TS746 + pAL-PolB-Exo	Exo
AL009	185-746L: TS746 + pAL-PolB-Y409L	Exo, Y409L, **extra mutations**
AL011	185-747-A485L: TS747 + pAL-PolB-A485L	Exo, A485L, **extra mutations**
AL012	188-746-A485L: TS746 + pAL-PolB-A485L	Exo, A485L, **extra mutations**
AL027	TS747 + pTS543-PolB-Exo	Sequence confirmed
AL028	TS747 + pTS543-PolB-Exo-Y409L	Sequence confirmed
AL029	TS747 + pTS543-PolB-Exo-Y409V	Sequence confirmed

T. kodakarensis culture growth

Constructed strains of *T. kodakarensis* were anaerobically cultured inside of an anaerobic chamber maintained with an atmosphere containing 90% nitrogen and 10% hydrogen. *T. kodakarensis* produce flammable H_2 and toxic H_2S gasses as their metabolism byproduct. The toxic H_2S gas is removed using hydrogen sulfide removal column via adsorption and chemisorption by the media and the H_2 gas byproduct can react with the oxygen contaminant from the outside forming water catalyzed by the palladium catalyst. All strains were passaged in artificial sea water (ASW) media supplemented with 5 g/l tryptone, 5 g/l yeast extract, 5 g/l pyruvate, 2 g/l elemental sulfur (S˚), and a KOD1-vitamin mixture with or without 1 mM agmatine sulfate at the desired growth temperature[32–34].

Growth rates were monitored using optical density measurements at 600 nm (OD_{600nm}) using a spectrophotometer. A minimum of three biological replicates for each strain at each temperatures were plotted using Excel.

Alkaline agarose gel electrophoresis

Genomic DNA from all confirmed strains used in this study were subjected to alkaline gel electrophoresis as previously described in Heider et al. (2017) and McElhinny et al. (2010)[17,22] with some modifications. Isolation of genomic DNA were done using high-quality *T. kodakarensis* genomic DNA purification protocol. Actively growing cultures (0.3-0.5 OD_{600nm}) were lysed, digested with Proteinase K and RNaseA, and purified with both PCI extraction and isopropanol precipitation[34].

Protocol for alkaline gel electrophoresis:

Casting 1% alkaline agarose gel:

Specific steps need to be followed to prevent the NaOH from denaturing at high temperature. The following steps will prevent denaturation of the gel during casting.

1. Mix 1.5 g of agarose in 150 mL dH_2O.
2. Microwave on high for 2 minutes to dissolve agarose.
3. Let the mixture Cool to <60°C.
4. Add 0.3 ml of 0.5 M EDTA.
5. Add 3.75 ml of 2 N NaOH (see **note 1**).
6. Pour in frame with 20 wells comb.
7. Let the alkaline agarose gel set at room temperature for 15 minutes.
8. Transfer the newly casted gel into an electrophoresis box filled with the alkaline buffer containing 0.05 N NaOH and 1mM EDTA.

Loading and running the alkaline gel

1. Mix samples with 2X loading buffer (see note 2).
2. Load every other lane with sample alternating with DNA ladder (see note 3).
3. Run at 4˚C 20 volts overnight (16 hours).
4. Incubate the alkaline gel with 1 M Tris HCl pH 7.5 for 15 minutes at room temperature.
5. Stain with SYBR Gold in 1 M Tris HCl pH 7.5 at room temperature.

Notes

1. Adding the 2 N NaOH too early will turn the gel yellow. Hints of yellowish hue suggest the agarose mixture was too hot when the 2 N NaOH was added.
2. Loading dye can be different depending on which protocol you followed. The recommended loading dye is the 2X RNA loading buffer (NEB). Unfortunately, no dyes we tested were suitable to meaningfully track the progression of the alkaline agarose gel.
3. We used the 1 Kb Plus DNA Ladder (Invitrogen).

REFERENCES

1. Kelman, L. M. & Kelman, Z. Archaeal DNA replication. *Annu Rev Genet* **48,** 71–97 (2014).

2. Greci, M. D. & Bell, S. D. Archaeal DNA Replication. *Annu Rev Microbiol* **74,** 65–80 (2020).

3. Makarova, K. S. & Koonin, E. V. Archaeology of Eukaryotic DNA Replication. *Cold Spring Harb Perspect Biol* **5,** a012963 (2013).

4. Raia, P., Delarue, M. & Sauguet, L. An updated structural classification of replicative DNA polymerases. *Biochem Soc Trans* **47,** 239–249 Preprint at https://doi.org/10.1042/BST20180579 (2019)

5. Zhuang, Z. & Ai, Y. Processivity factor of DNA polymerase and its expanding role in normal and translesion DNA synthesis. *Biochim Biophys Acta* **1804,** 1081 (2010).

6. Wu, J., De Paz, A., Zamft, B. M., Marblestone, A. H., Boyden, E. S., Kording, K. P. & Tyo, K. E. J. DNA binding strength increases the processivity and activity of a Y-Family DNA polymerase. *Scientific Reports 2017 7:1* **7,** 1–12 (2017).

7. Bębenek, A. & Ziuzia-Graczyk, I. Fidelity of DNA replication—a matter of proofreading. *Curr Genet* **64,** 985 (2018).

8. Dodd, T., Botto, M., Paul, F., Fernandez-Leiro, R., Lamers, M. H. & Ivanov, I. Polymerization and editing modes of a high-fidelity DNA polymerase are linked by a well-defined path. *Nature Communications 2020 11:1* **11,** 1–11 (2020).

9. Kushida, T., Narumi, I., Ishino, S., Ishino, Y., Fujiwara, S., Imanaka, T. & Higashibata, H. Pol B, a Family B DNA Polymerase, in Thermococcus kodakarensis is Important for DNA Repair, but not DNA Replication. *Microbes Environ* **34,** 316 (2019).

10. Č uboň ová, L., Richardson, T., Burkhart, B. W., Kelman, Z., Connolly, B. A., Reeve, J. N. & Santangelo, T. J. Archaeal DNA polymerase D but not DNA polymerase B is required for genome replication in Thermococcus kodakarensis. *J Bacteriol* **195,** 2322–2328 (2013).

11. Clausen, A. R., Lujan, S. A., Burkholder, A. B., Orebaugh, C. D., Williams, J. S., Clausen, M. F., Malc, E. P., Mieczkowski, P. A., Fargo, D. C., Smith, D. J. & Kunkel, T. A. Tracking replication enzymology in vivo by genome-wide mapping of ribonucleotide incorporation. *Nature Structural & Molecular Biology 2015 22:3* **22,** 185–191 (2015).

12. Zatopek, K. M., Potapov, V., Maduzia, L. L., Alpaslan, E., Chen, L., Evans, T. C., Ong, J. L., Ettwiller, L. M. & Gardner, A. F. RADAR-seq: A RAre DAmage and Repair sequencing method for detecting DNA damage on a genome-wide scale. *DNA Repair (Amst)* **80,** 36–44 (2019).

13. McElhinny, S. A. N., Kumar, D., Clark, A. B., Watt, D. L., Watts, B. E., Lundström, E. B., Johansson, E., Chabes, A. & Kunkel, T. A. Genome instability due to ribonucleotide incorporation into DNA. *Nature Chemical Biology 2010 6:10* **6,** 774–781 (2010).

14. Lipkin, D., Talbert, P. T. & Cohn, M. The Mechanism of the Alkaline Hydrolysis of Ribonucleic Acids. *J Am Chem Soc* **76,** 2871–2872 (1954).

15. Li, Y. & Breaker, R. R. Kinetics of RNA degradation by specific base catalysis of transesterification involving the 2γ-hydroxyl group. *J Am Chem Soc* **121,** 5364–5372 (1999).

16. Vaisman, A., McDonald, J. P., Noll, S., Huston, D., Loeb, G., Goodman, M. F. & Woodgate, R. Investigating the mechanisms of ribonucleotide excision repair in Escherichia coli. *Mutat Res* **761,** 21–33 (2014).

17. Sparks, J. L., Chon, H., Cerritelli, S. M., Kunkel, T. A., Johansson, E., Crouch, R. J. & Burgers, P. M. RNase H2-initiated ribonucleotide excision repair. *Mol Cell* **47,** 980–986 (2012).

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

19. Burkhart, B. W., Cubonova, L., Heider, M. R., Kelman, Z., Reeve, J. N. & Santangelo, T. J. The GAN exonuclease or the flap endonuclease Fen1 and RNase HII are necessary for viability of Thermococcus kodakarensis. *J Bacteriol* **199,** (2017).

20. Nick McElhinny, S. A., Watts, B. E., Kumar, D., Watt, D. L., Lundström, E. B., Burgers, P. M. J., Johansson, E., Chabes, A. & Kunkel, T. A. Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases. *Proc Natl Acad Sci U S A* **107,** 4949–4954 (2010).

21. Cerritelli, S. M. & Crouch, R. J. The Balancing Act of Ribonucleotides in DNA. *Trends Biochem Sci* **41,** 434–445 (2016).

22. Yao, N. Y., Schroeder, J. W., Yurieva, O., Simmons, L. A. & O'Donnell, M. E. Cost of rNTP/dNTP pool imbalance at the replication fork. *Proc Natl Acad Sci U S A* **110,** 12942–12947 (2013).

23. McElhinny, S. A. N., Kumar, D., Clark, A. B., Watt, D. L., Watts, B. E., Lundström, E. B., Johansson, E., Chabes, A. & Kunkel, T. A. Genome instability due to ribonucleotide incorporation into DNA. *Nat Chem Biol* **6,** 774–781 (2010).

24. Gardner, A. F. & Jack, W. E. Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Res* **27,** 2545–2553 (1999).

25. Lemor, M., Kong, Z., Henry, E., Brizard, R., Laurent, S., Bossé, A. & Henneke, G. Differential Activities of DNA Polymerases in Processing Ribonucleotides during DNA Synthesis in Archaea. *J Mol Biol* **430,** 4908–4924 (2018).

26. Greenough, L., Kelman, Z. & Gardner, A. F. The roles of family B and D DNA polymerases in thermococcus species 9°N Okazaki fragment maturation. *Journal of Biological Chemistry* **290,** 12514–12522 (2015).

27. Zatopek, K. M., Alpaslan, E., Evans, T. C., Sauguet, L. & Gardner, A. F. Novel ribonucleotide discrimination in the RNA polymerase-like two-barrel catalytic core of Family D DNA polymerases. *Nucleic Acids Res* **48,** 12204 (2020).

28. Betancurt-Anzola, L., Martínez-Carranza, M., Delarue, M., Zatopek, K. M., Gardner, A. F. & Sauguet, L. Molecular basis for proofreading by the unique exonuclease domain of Family-D DNA polymerases. *Nature Communications 2023 14:1* **14,** 1–15 (2023).

29. Greenough, L., Menin, J. F., Desai, N. S., Kelman, Z. & Gardner, A. F. Characterization of Family D DNA polymerase from Thermococcus sp. 9°N. *Extremophiles* **18,** 653–664 (2014).

30. Takashima, N., Ishino, S., Oki, K., Takafuji, M., Yamagami, T., Matsuo, R., Mayanagi, K. & Ishino, Y. Elucidating functions of DP1 and DP2 subunits from the Thermococcus kodakarensis family D DNA polymerase. *Extremophiles* **23,** 161–172 (2019).

APPENDIX A: TETRAETHER ARCHAEAL LIPIDS PROMOTE LONG-TERM SURVIVAL IN EXTREME CONDITIONS

Summary

The sole unifying feature of the incredibly diverse Archaea is their isoprenoid-based ether-linked lipid membranes. Unique lipid membrane composition, including an abundance of membrane-spanning, tetraether lipids, impart resistance to extreme conditions. Many questions remain, however, regarding the synthesis and modification of tetraether lipids, and how dynamic changes to archaeal lipid membrane composition support hyperthermophily. Tetraether membranes, termed glycerol dibiphytanyl glycerol tetraethers (GDGTs), are generated by tetraether synthase (Tes) by joining the tails of two bilayer lipids known as archaeol. GDGTs are often further specialized through addition of cyclopentane rings by GDGT ring synthase (Grs). A positive correlation between relative GDGT abundance and entry into stationary phase growth has been observed, but the physiological impact of inhibiting GDGT synthesis has not previously been reported. Here, we demonstrate that the model hyperthermophile Thermococcus kodakarensis remains viable when Tes (TK2145) or Grs (TK0167) are deleted, permitting phenotypic and lipid analyses at different temperatures. The absence of cyclopentane rings in GDGTs does not impact growth in T. kodakarensis, but an overabundance of rings due to ectopic Grs expression is highly fitness negative at supra-optimal temperatures. In contrast, deletion of Tes resulted in the loss of all GDGTs, cyclization of archaeol, and loss of viability upon transition to stationary phase in this model archaea. These results demonstrate the critical roles of highly specialized, dynamic, isoprenoid-based lipid membranes for archaeal survival in high temperature.

---

[1] Most of this chapter was previously published under the following title with a few updates: Liman, G. L. S., Garcia, A. A., Fluke, K. A., Anderson, H. R., Davidson, S. C., Welander, P. V., & Santangelo, T. J. (2024). Tetraether archaeal lipids promote long-term survival in extreme conditions. Molecular microbiology, 10.1111/mmi.15240. Advance online publication. https://doi.org/10.1111/mmi.15240

Introduction

The diversity of marine and terrestrial environments containing an abundance of microbial life continues to expand and often beguiles the limits of life itself. Archaea often dominate and thrive in the extremes of temperature, salinity, pressure, and pH [1,2]. The unique, domain-specific, isoprenoid-based lipid membranes generated by all Archaea [3] are thought to assist but cannot completely resolve rationales for survival in the extremes. Beyond compositional and structural differences, the dynamic response of archaeal lipid membranes to changes in growth phase, growth rate, and temperature support adaptive responses to rapid and often dramatic stimuli in highly volatile environments [4–8].

Archaeal membranes are composed of glycerol-1-phosphate (G1P) ether-linked to two isoprenoid-based branched chains, typically 20 carbons each (archaeol) (Fig. A.1). Archaeal lipid composition contrasts with the glycerol-3-phosphate (G3P) ester-linked fatty acid-based lipids common in Bacteria and Eukarya [9,10]. Many archaea also fuse their diether bilayers (archaeols) to form tetraether monolayer membranes known as glycerol dibiphytanyl glycerol tetraethers (GDGTs) (Fig. A.1) [9,10]. GDGTs are further modified through addition of cyclopentane or cyclohexane rings within the isoprenoid chains to generate a series of cyclized GDGTs [11]. The properties, and potential biotechnological and commercial value of archaeal isoprenoid-based lipids demand an understanding of their synthesis and role in promoting fitness in extreme habitats.

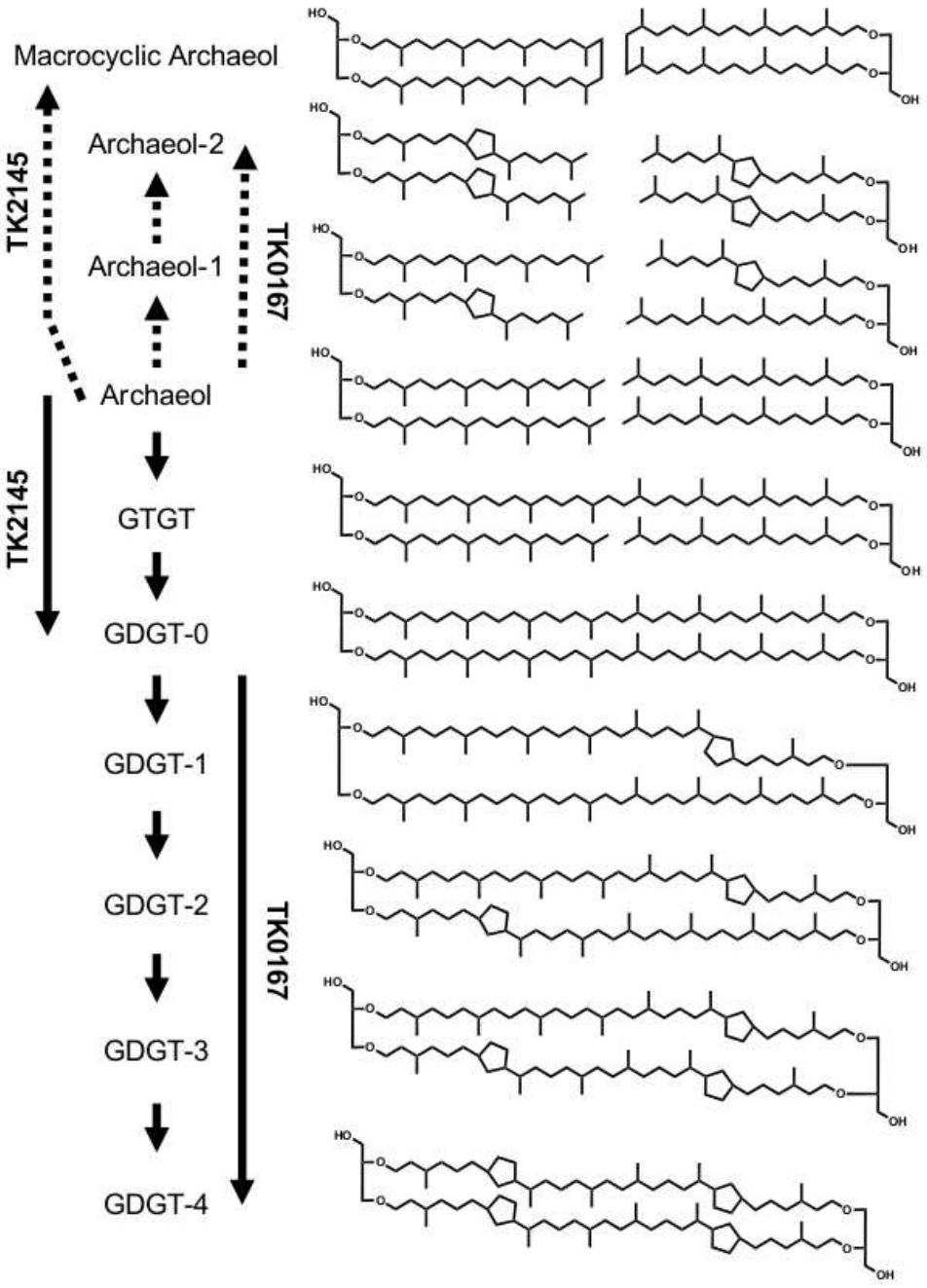


Figure A.1. Tetraether lipid biosynthetic pathway in *T. kodakarensis*. The C20 isoprenoid chains of archaeol are covalently linked to generate GDGT (also termed GDGT-0), through the activity of TK2145, Tk-Tes, with GTGT and macrocyclic archaeol (MA) as intermediate or side products, respectively. Production of cyclized-tetraether lipids (GDGT-1, 2, 3, 4), is catalyzed by TK0167, Tk-Grs. Archaeol-1 and archaeol-2 are detected only in strains ectopically expressing Tk-Grs or those with disruption of native Tk-Tes activities.

The formation of GDGT from archaeol and subsequent cyclization are thought to primarily modify cell membrane fluidity and permeability, aiding archaeal organisms to adapt to a wide range of environmental parameters [4,12,13]. The dynamic changes of archaeal lipid composition in response to environmental variation, combined with the geologically relevant stability of archaeal lipids, also provide a reliable method to infer ancient sea surface temperatures, informing paleoclimate studies [14–16]. Further, the unique properties of archaeal lipid membranes, often termed archaeosomes, offer promise for therapeutic approaches, including nano-drug delivery protocols and vaccine delivery as an alternative to the more conventional liposome-based systems [17–19]. Controlled and reliable production of specialized archaeosomes with novel compositions and novel properties is also of major biotechnological interest. However, while GDGT properties have been readily investigated in artificial systems, biological platforms wherein GDGT and cyclized GDGT synthesis can be controlled and regulated are lacking.

GDGT synthesis is achieved through the action of the radical S-adenosylmethionine (SAM) superfamily protein GDGT-macrocyclic archaeol synthase (GDGT-MAS) [20] also known as tetraether synthase (Tes) [21]. Formation of cyclopentane rings in GDGT-0 (no rings) at the C-7 and C-3 positions is catalyzed by another radical SAM superfamily protein, GDGT ring synthase (Grs), first characterized in *Sulfolobus acidocaldarius* [22]. While biochemical studies have determined the enzymes responsible for production of GDGT-0 and its cyclized derivatives, biological manipulations have been lacking due to presumed essentiality of Tes [21].

Thermococcus kodakarensis is a highly versatile, genetically tractable hyperthermophilic, anaerobic, archaeon that naturally synthesizes GDGT-0, and to a much lesser extent, its ring bearing derivatives. The relative abundance of GDGT lipids in *T. kodakarensis* under optimal growth conditions has been determined in previous studies and found to vary between ~25 – 80% based on the extraction and analysis methods used. *T. kodakarensis* undergoes

substantial diether to tetraether lipid composition changes in response to temperature and growth phase [23,24] that GDGT synthesis, which involves Tes activities, contributes to hyperthermophilic physiology. In this study, we generated Tes and Grs deletion mutants demonstrating that neither tetraether membrane formation nor membrane cyclization was essential in *T. kodakarensis*. Physiological characterization of these mutants and of strains ectopically expressing Tes and Grs demonstrated the importance of tetraether lipids for thermophily and fitness in the extremes for *T. kodakarensis* while also revealing unexpected cyclization responses.

Results

TK2145 and TK0167 encode the non-essential tetraether synthase (Tes) and GDGT-ring synthase (Grs), respectively.

Analyses of the lipidome of *T. kodakarensis* strains revealed the presence of membrane-spanning glycerol dibiphytanyl glycerol tetraethers (GDGTs) and a small percentage of cyclized GDGTs, thus predicting the presence of an active Tes and Grs (Fig. A.2C-D & A.S2A). Based on homology with recently identified Tes enzymes [21,22], TK2145 was predicted to encode the sole Tes homolog ($1e^{-170}$ e-value, 45% identity) in *T. kodakarensis* (Tk-Tes); however, no biochemical or genetic evidence supporting TK2145 as Tk-Tes was previously reported. An unbiased mutagenesis of *T. kodakarensis* [25] revealed that inactivation of TK2145 resulted in reduced hyperthermotolerance in *T. kodakarensis,* but lipid analyses were not performed to confirm loss of GDGT production. The co-purification of the product of TK2145 with TK2140, a DNA ligase [26], is the only other previous, and still unexplained, information on the role of the product of TK2145 *in vivo*. We also searched the *T. kodakarensis* genome for a potential Grs homolog and identified TK0167 as a strong candidate ($8e^{-72}$ e-value, 32% identity to GrsA). Beyond basic transcriptomics [27] or microarray data [28] demonstrating expression of TK0167, no previous information on the role(s) of TK0167 has been reported.

To establish the potential roles of TK2145 and TK0167 in lipid production and maturation in *T. kodakarensis*, we individually targeted each locus for deletion with established genetic techniques [29] (Fig. A.2A & B). While the entire open reading frame of TK2145 was deleted, a portion of TK0167 was intentionally not targeted for genomic deletion due to the presence of small RNA identified in our previous transcriptomic data that overlaps with TK0167 [27]. Deletion of TK2145 from the parental strain TS559 was successful (generating strain AL016; *Δtes*), as was the desired partial deletion of TK0167 (generating strain AL010; *Δgrs*) (Table A.1). The exact endpoints of each markerless genomic deletion were confirmed first by PCR amplification of genomic regions and Sanger sequencing of amplicons, followed by whole genome sequencing, at >100X coverage for each strain (Fig. A.S1A & B). Total genome analyses of the *Δtes and Δgrs* strains confirmed deletion of the TK2145 and TK0167 sequences, respectively, and established that neither strain acquired any secondary mutations throughout the remainder of the genome.

To confirm that deletion of putative Tes and Grs homologs abolished GDGT formation and cyclization, respectively, we carried out triplicate lipid analyses of the *T. kodakarensis* parental strain, TS559, and the Tk-Tes and Tk-Grs deletion strains grown to late-stationary phase at 85˚C (Fig. A.2C-D, A.3C-D, A.4C-D & A.S2A-B & E). Freeze-dried biomass was acid hydrolyzed to generate core lipids, which lack headgroups, for liquid chromatography-mass spectrometry (LC-MS) analyses and revealed the production of both tetraether and diether lipids (Fig. A.2C-D & A.S2A). Archaeol (m/z = 653) production was observed in all three strains (Fig. A.2C & A.S2A-B & E), while the *Δtes* strain displayed a complete absence of all tetraether lipid species (m/z = 1304-1294) (Fig. A.2C, A.S2B, & A.S3A-F); note that the trace for the *Δtes* strain is displayed at three orders of magnitude lower intensity, yet no signal indicative of GDGT-0 was observed. Complementation of the *Δtes* strain via ectopic expression of Tk-Tes restored GDGT-

0 production (Fig. A.2C & A.S2C & D) demonstrating that Tes activities are encoded at the TK2145 locus and that no other Tes activities exist within *T. kodakarensis*. The viability of the Tk-Tes mutant contrasts with the presumed essentiality of the *S. acidocaldarius* Tes [22], suggesting that tetraether membranes may be more critical in some archaeal clades than others.

Lipid analyses of the parental strain TS559 revealed small amounts (<1% of all tetraether lipids) of GDGT-1, 2, 3, and 4 whereas the Tk-Grs deletion strain displayed a complete loss of cyclized GDGTs while archaeol and GDGT-0 remained present (Fig. A.2D). Complementation via ectopic Grs (TK0167) expression restored cyclized GDGT production in the *Δgrs* strain (Fig. A.2D & A.S2F & G) confirming that the TK0167 locus encodes Grs activity. The complete loss of GDGT-1, 2, 3, and 4 in the *Δgrs* strain also adumbrates that no redundant Grs activities are present in *T. kodakarensis*.

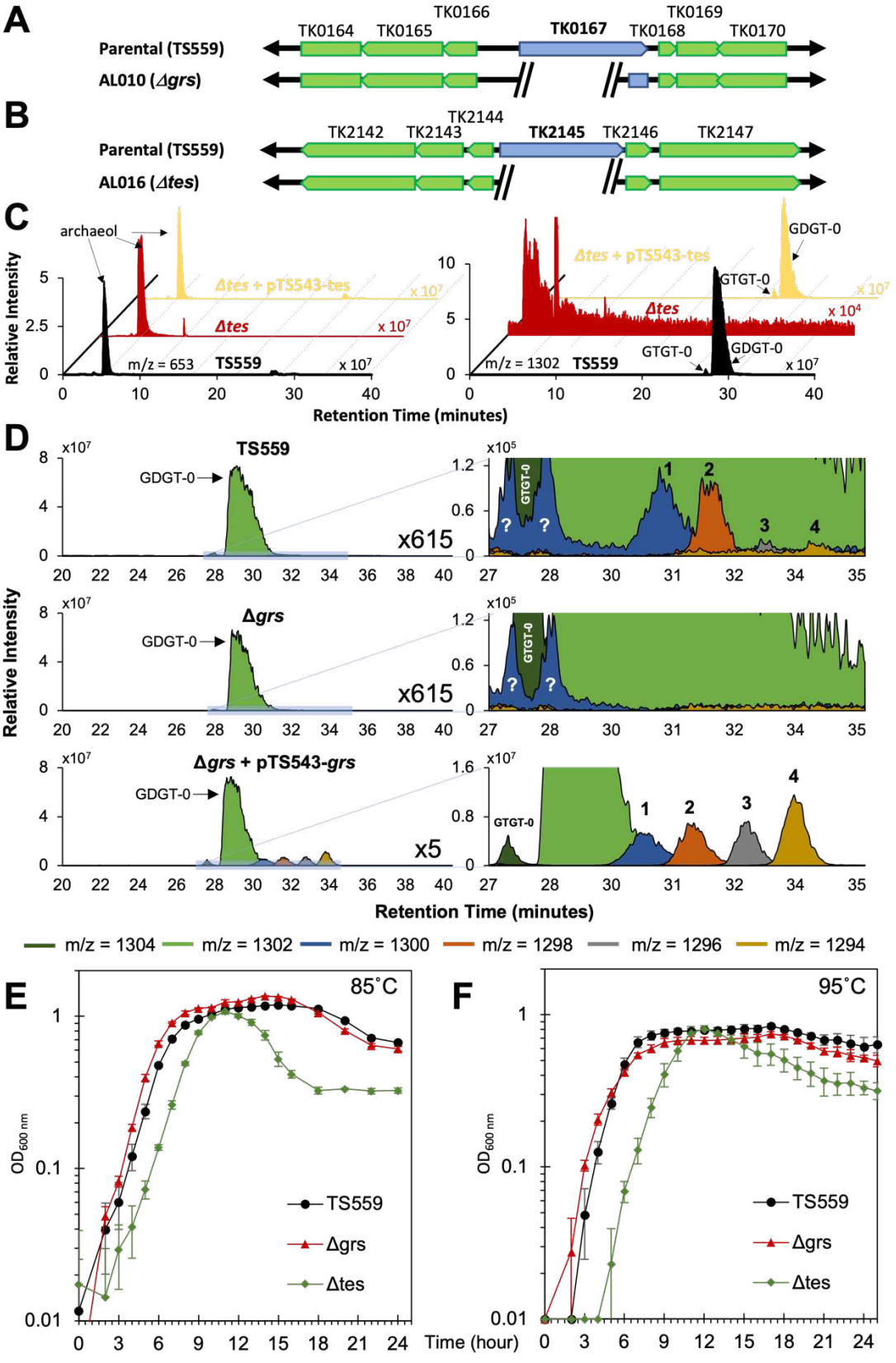

A
TK0166
TK0164
TK0165
TK0167
TK0168
TK0169
TK0170
Parental (TS559)
AL010 (Δgrs)
B
TK2144
TK2142
TK2143
TK2145
TK2146
TK2147
Parental (TS559)
AL016 (Δtes)
C
archaeol
Δtes + pTS543-tes
Δtes
m/z = 653
TS559
GDGT-0
GTGT-0
m/z = 1302
Relative Intensity
Retention Time (minutes)
D
TS559
Δgrs
Δgrs + pTS543-grs
GDGT-0
GTGT-0
x615
x5
Relative Intensity
Retention Time (minutes)
m/z = 1304
m/z = 1302
m/z = 1300
m/z = 1298
m/z = 1296
m/z = 1294
E
85˚C
F
95˚C
OD600 nm
TS559
Δgrs
Δtes
Time (hour)

Figure A.2. Long-term survival of *T. kodakarensis* demands Tes (TK2145) catalyzed tetraether lipid production but not Grs (TK0167) catalyzed cyclization. (A) The genomic locus of TK0167, predicted to encode Tk-Grs (blue), shown with flanking genes (green) in the parental strain (TS559) (Top) and in the TK0167 partial deletion strain (*Δgrs*; AL010) (Bottom). (B) The genomic locus of TK2145, predicted to encode Tk-Tes (blue), shown with flanking genes (green) in the parental strain (TS559) (Top) and in the TK2145 deletion strain (*Δtes*; AL016) (Bottom). (C) LC-MS extracted ion chromatograms of one replicate of the acid hydrolyzed lipid extracts from the parent strain TS559, the *Δtes* strain, and the *Δtes* strain complemented with TK2145. GDGT production was lost upon deletion of Tk-Tes and restored with ectopic expression of TK2145. (D) LC-MS extracted ion chromatograms of one replicate of the acid hydrolyzed lipid extracts from the parent strain TS559, the *Δgrs* strain, and the *Δgrs* strain complemented with TK0167. Deletion of Tk-Grs resulted in the loss of cyclized GDGTs and cyclization was restored with ectopic expression of TK0167. (E and F) Exponential growth rates of *T. kodakarensis* at 85˚C (E) and 95˚C (F) are not significantly impacted by deletion of TK2145 (Tk-Tes; strain AL016) or TK0167 (Tk-Grs; AL010), however, deletion of Tk-Tes (TK2145) dramatically impacts survival upon entry into stationary phase.

Table A.1. Strain names

Strain	Details	Source
TS559	Parental Strain	Santangelo et al. (2010)
AL010	TS559 ΔTK0167*	This study
AL016	TS559 ΔTK2145	This study
AL017	TS559 + pTS543	This study
AL018	TS559 + pTS543-TK0167	This study
AL019	TS559 + pTS543-TK2145	This study
AL020	AL010 + pTS543	This study
AL021	AL010 + pTS543-TK0167	This study
AL022	AL016 + pTS543	This study
AL023	AL016 + pTS543-TK2145	This study

Lack of GDGTs impacts late stationary phase survival.

The biological importance of tetraether lipids has not previously been investigated *in vivo* due to the presumed essentiality of Tes [21]. Given the small percentage (~0.1%) of tetraethers that were cyclized in the parental strain, it was not surprising that deletion of Tk-Grs (strain AL010) did not result in a significant defect in growth rate or final cell densities at 85˚C (optimal) or 95˚C (supra-optimal, but tolerable) (Fig. A.2E & F). In contrast, eliminating Tk-Tes (strain AL016) resulted in substantial impacts on survival upon transition from late-exponential phase to stationary phase (Fig. A.2E & F). Aside from the slightly elongated lag phase, the rate of *Δtes* growth largely matched that of the parental strain TS559 at both optimal and supra-optimal growth temperatures. Whereas TS559 showed only minor reductions in optical density upon reaching stationary phase, suggestive of minor impacts to viability, the optical density of *Δtes* (AL016) drops significantly upon entry into stationary phase (Fig. A.2E & F). These data imply that most (~50-75%) of the culture perishes in stationary phase due to the lack of tetraether lipids and provides *in vivo* evidence that tetraether lipids are necessary to promote viability in stationary phase *T. kodakarensis* cultures. Given the known volatile nature and ever-changing nutrient availability of hydrothermal vents, the fitness of *T. kodakarensis* strains lacking tetraether lipids pales in comparison to tetraether-containing strains in natural environments.

The rescued production of core tetraether lipids in the Tes mutant via ectopic Tk-Tes activities restored the decline in population density that was observed in the *Δtes* strain upon entry to stationary phase. Ectopic Tk-Tes expression completely restored growth phenotypes at both 85˚C and 95˚C (Fig. A.3A & B). Our findings thus not only genetically confirm the function of TK2145 as a Tes, but also confirm the stationary phase phenotype of cells lacking Tk-Tes can be rescued through complementation.

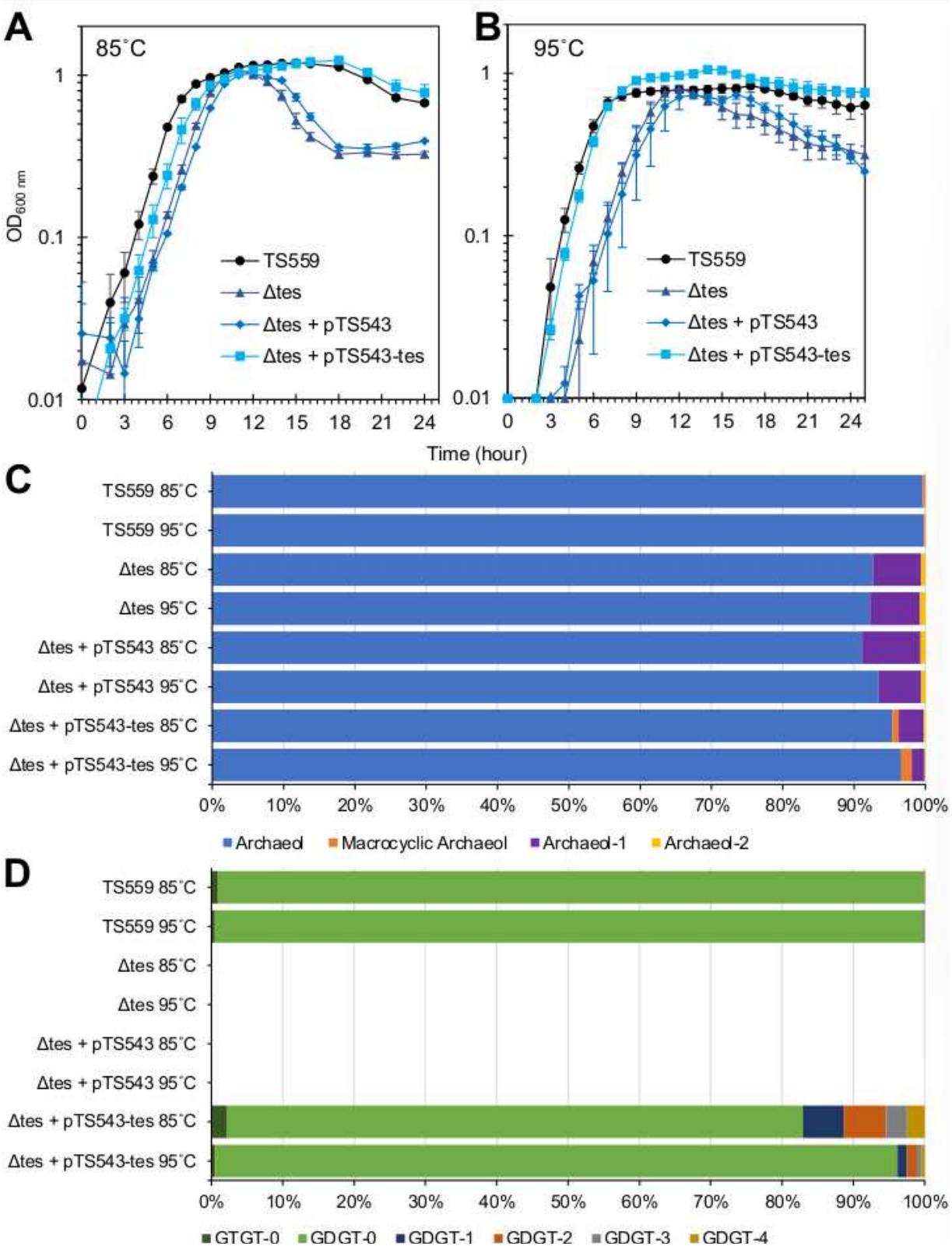

A
85˚C
B
95˚C
OD600 nm
1
0.1
0.01
0 3 6 9 12 15 18 21 24
Time (hour)
TS559
Δtes
Δtes + pTS543
Δtes + pTS543-tes
C
TS559 85˚C
TS559 95˚C
Δtes 85˚C
Δtes 95˚C
Δtes + pTS543 85˚C
Δtes + pTS543 95˚C
Δtes + pTS543-tes 85˚C
Δtes + pTS543-tes 95˚C
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Archaeol
Macrocyclic Archaeol
Archaeol-1
Archaeol-2
D
GTGT-0
GDGT-0
GDGT-1
GDGT-2
GDGT-3
GDGT-4

Figure A.3. Loss of hyperthermophily due to loss of tetraether lipid synthesis is restored via ectopic complementation of Tk-Tes activities. (A & B) The growth of triplicate biological cultures of *T. kodakarensis* strains TS559 (black), *Δtes* (dark blue), *Δtes* + pTS543 (blue), and pTS543 + pTS543-TK2145 (light blue) were monitored via changes in optical density (at 600nm) at 85˚C and 95˚C, respectively, revealing the rescue of reduced survival during transition to stationary phase due to the lack of Tk-Tes activities. (C & D) Lipid analyses of triplicate biological cultures of *T. kodakarensis* strains grown at 85˚C and 95˚C reveals that ectopic expression of Tk-Tes activities restores tetraether lipid synthesis to strains lacking TK2145 on the genome and an unanticipated increase in cyclized GDGTs. Panel C and D show diether and tetraether lipid production, respectively.

Aberrant production of cyclized tetraether lipids dramatically impairs hyperthermophilic growth.

As less than 1% of GDGTs were cyclized in the parental strain TS559, we predicted that ectopic complementation of Tk-Grs might result in an overabundance of cyclic tetraethers, permitting evaluation of the impact of increased lipid rings on microbial fitness (Fig. A.4 and A.S2A & G). We introduced an autonomously replicating plasmid [30] directing Grs expression from a strong promoter [31] into the parental strain TS559 and the *Δgrs* strain (Fig. A.4 & A.S4); strains of TS559 and *Δgrs* retaining the same plasmid (pTS543) lacking the *grs* expression cassette were also generated as controls (Table A.1).

As predicted, introduction of the Tk-Grs complementation vector (pTS543-TK0167) into the *Δgrs* strain not only restored the production of cyclized GDGTs but also dramatically increased cyclization. The Ring Index (RI) is defined as the weighted average of ring numbers in GDGT compounds [16]. We observed a massive increase in the RI from just 0.002 and 0.004 in TS559 at 85˚C and 95˚C, respectively, to 0.470 and 0.740 in the *Δgrs* strains complemented with ectopic Grs expression at 85˚C and 95˚C, respectively (Fig. A.4D & A.S5, Table A.2). Tk-Grs ectopic expression in the parental strain TS559 also dramatically shifted the lipid composition compared to the parental strain (Fig. A.S4C & D), increasing the average RI to 0.413 and 0.588 at 85˚C

and 95˚C, respectively (Fig. A.S5). Thus, ectopic expression of Tk-Grs can result in strains where ~19-28% of the tetraether lipids are cyclized, providing a route to the large-scale production of cyclized GDGTs in *T. kodakarensis* (Figs. A.4D & A.S4D). Additionally, when Tk-Grs is ectopically expressed in *Δgrs* at 60˚C, an even greater average RI of 1.67 is observed and ~53% of the tetraether lipids are cyclized, indicating that GDGT cyclization has a complex relationship with temperature (Fig. A.S5 & A.S7A).

The increased relative abundance of cyclized GDGTs in the parental and *Δgrs* strains ectopically expressing Tk-Grs is not benign and results in substantial defects at increased growth temperatures (Fig. A.4A & B, A.S4A & B). While increased Grs activities throughout the growth cycle are tolerated without significant impact at the optimal growth temperature of 85˚C, ectopic expression of Tk-Grs at 95˚C dramatically slows growth, although final cell densities matching parental strains are eventually achieved (Fig. A.4A & B, A.S4A & B). Thus, while an increase in the proportion of GDGTs that are cyclized from natural levels of ~0.1% to ~19% is well-tolerated at 85˚C, near identical increases from ~0.3% to ~28% at 95˚C results in dramatic fitness consequences (Fig. A.4B & A.S4B). It will be of interest to determine how increases in cyclized GDGTs change key parameters of membrane biology (e.g., stability, fluidity, and permeability) in *T. kodakarensis* that result in impaired growth rates at supra-optimal temperatures.

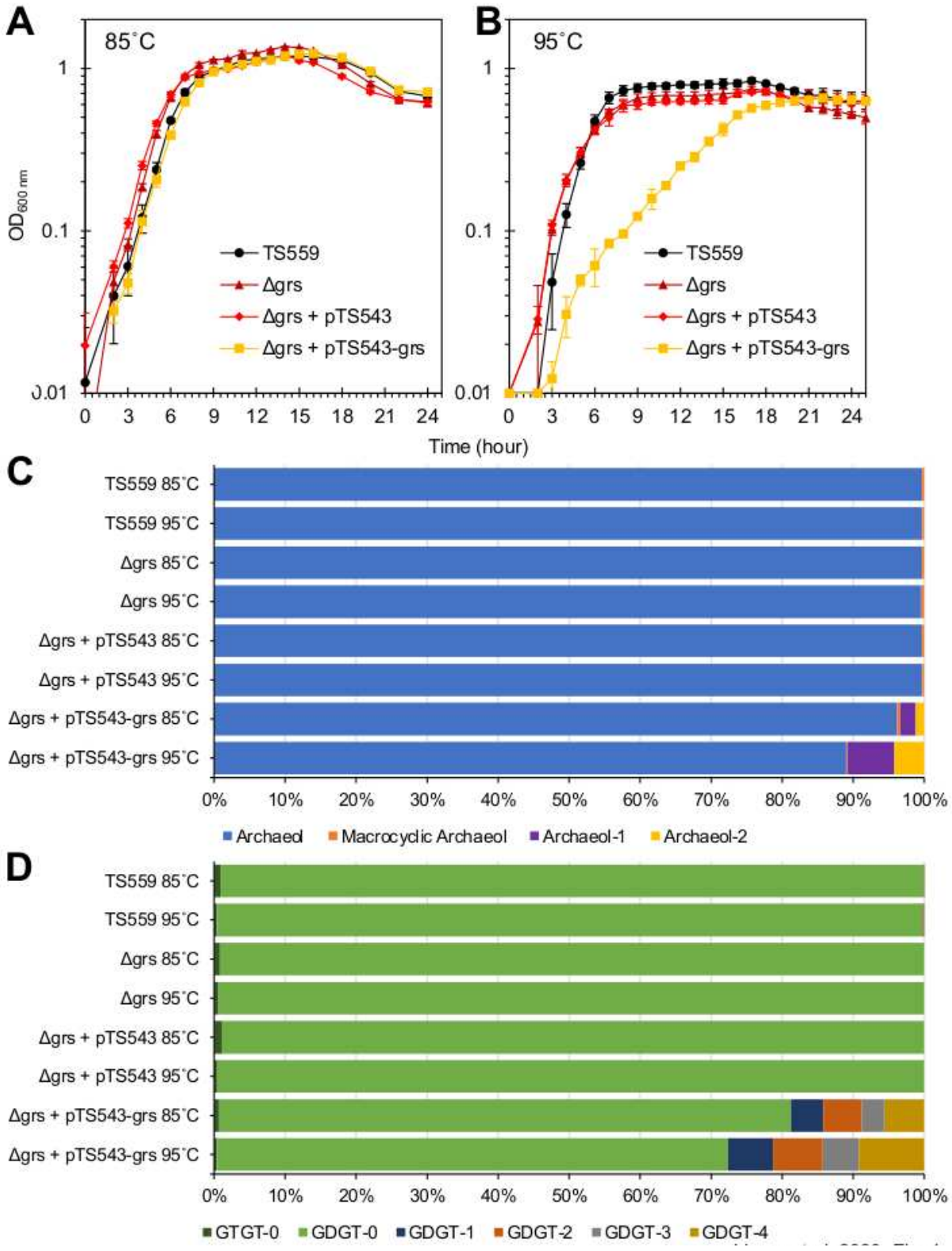

A
85˚C
B
95˚C
OD600 nm
1
0.1
0.01
0 3 6 9 12 15 18 21 24
Time (hour)
TS559
Δgrs
Δgrs + pTS543
Δgrs + pTS543-grs
C
TS559 85˚C
TS559 95˚C
Δgrs 85˚C
Δgrs 95˚C
Δgrs + pTS543 85˚C
Δgrs + pTS543 95˚C
Δgrs + pTS543-grs 85˚C
Δgrs + pTS543-grs 95˚C
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Archaeol
Macrocyclic Archaeol
Archaeol-1
Archaeol-2
D
GTGT-0
GDGT-0
GDGT-1
GDGT-2
GDGT-3
GDGT-4

Figure A.4. Dramatic increases in cyclized GDGTs from ectopic Tk-Grs expression coincide with growth defects at supra-optimal temperature. (A & B) Triplicate biological cultures of *T. kodakarensis* strains TS559 (black), *Δgrs* (dark red), *Δgrs* + pTS543 (red), and *Δgrs* + pTS543-TK0167 (orange) were monitored for changes in optical density (at 600nm) at 85˚C and 95˚C, respectively. Ectopic expression of Tk-Grs (orange) resulted in impaired growth at 95˚C. (C & D) Lipidome analyses of triplicate biological cultures of *T. kodakarensis* strains grown at 85˚C and 95˚C. Ectopic TK-Grs expression increases cyclized-GDGT levels ~50-fold (to ~20% of extracted core lipids). Panel C and D show diether and tetraether lipid production, respectively.

Disruption of native Tk-Tes or Tk-Grs activity results in unexpected lipid composition changes.

As expected, the complete loss of GDGTs in the *Δtes* strain results in a membrane composed of only diether lipids. However, unexpectedly, we found that the *Δtes* strain is capable of cyclizing these diether lipids, producing putative ring containing derivatives of archaeol, here termed archaeol-1 and archaeol-2 with one and two rings, respectively (Fig. A.1, A.3C, & A.S6A-D, Table A.2), supported by mass spectra and hydrogenation experiments described later in the text. In the absence of its presumed natural GDGT substrates, Tk-Grs uses archaeol as a substrate, resulting in the *Δtes* strain converting ~7% of its diether lipids to archaeol-1 and archaeol-2 at both 85˚C and 95˚C (Fig. A.3C and Table A.2). The significance of newly cyclized archaeols is emphasized by the fact that the cyclized diethers in *Δtes* are ~70-fold higher (~7% of core diethers) than the abundance of all cyclized tetraethers (~0.1% of core tetraethers) in the parental strain TS559 (Fig. A.3C & D).

Archaeol-1 and archaeol-2 were also detected in strains ectopically expressing Tk-Grs where ~3% and 11% of the core diether lipids were cyclized at 85˚C and 95˚C, respectively (Fig. A.4C & A.S6E & F). Growth of the *Δgrs* complementation strain (*Δgrs* + pTS543-*grs*) at 60˚C resulted in the robust production of archaeol-1 and -2, together accounting for ~51% of the total core diether lipids while archaeol-1 and -2 were completely absent in the *Δgrs* strain (Fig. A.S7B & C). The high levels of these cyclized diether lipids allowed for their characterization via gas

chromatography-mass spectrometry (GC-MS), following trimethylsilylation (TMS) derivatization. This analysis revealed the presence of two chromatographically separated archaeol-1 isomers (termed archaeol-1a and -1b) and three chromatographically separated archaeol-2 isomers (termed archaeol-2a, -2b, and -2c) with nearly identical electron impact (EI) mass spectra (Fig. A.S7D, A.S9, & A.S10). EI mass spectra of TMS-archaeol (Fig. A.S8) in these samples closely matched that in Dawson et al. [32]; however, mass spectra of both TMS-archaeol-1 isomers differed from archaeol with a single double bond (Fig. A.S9, A.S14, & A.S15 and Fig. A.2B in Dawson et al.). Further, hydrogenation of acid-hydrolyzed total lipid extracts containing archaeol-1 and -2 demonstrated that these lipids are immune to hydrogenation and thus do not contain double bonds but instead contain rings (Fig. A.S16A-C). This hydrogenation experiment, in combination with 1) the observed mass spectra, 2) the delayed elution times of archaeol-1 and -2 compared to archaeol (Fig A.S7C), typical of ring-bearing but not double bond possessing derivatives that elute earlier (Fig. A.S16B & C), and 3) the robust production of these lipids when ectopically expressing Tk-Grs (Fig. A.S7C), strongly supports that archaeol-1 and archaeol-2 contain cyclopentane rings within their C-20 isoprenoid chains (Fig. A.1).

Further, our analyses reveal that two regioisomers of archaeol-1 are possible; these are isomers that differ by which tail the ring is placed in – either the C3 glycerol bonded tail or the C2 glycerol bonded tail (Fig. A.S9). The EI mass spectra of both TMS-archaeol-1a and TMS-archaeol-1b reveals that both isomers are actually a mixture of the two regioisomers as demonstrated by the co-occurrence of the m/z = 307, 412 fragment ion pair and the m/z = 309, 410 fragment ion pair in the same mass spectrum, each pair resulting from the C2-C3 glycerol bond cleavage of a different regioisomer (Fig. A.S9, A.S14, & A.S15). However, the m/z = 307, 412 fragment ion pair is dominant over the m/z = 309, 410 pair, indicating that the ring is more commonly found in the C3 glycerol bonded tail (tail furthest from the headgroup) and suggesting that Tk-Grs prefers to cyclize at this location first (Fig. A.S9, A.S14, & A.S15).

The restoration of Tk-Tes activities in the *Δtes* complementation strain also resulted in unexpected changes to the lipidome. As expected, restoring Tk-Tes expression and activities via ectopic expression in the *Δtes* strain restored production of GDGT-0 (Fig. A.3D). Surprisingly, ectopic expression of Tk-Tes also resulted in an increased abundance of GDGT-1, 2, 3, and 4 with an average ring index of 0.358 at 85˚C. This was similar to that of strains ectopically expressing Tk-Grs at 85˚C (*Δgrs* + pTS543-*grs* RI = 0.470 and TS559 + pTS543-*grs* RI = 0.413), although this increased cyclization was much less at 95˚C with an RI of just 0.080 (Fig. A.S5). An increase in the average RI was also observed for the parent strain (TS559) overexpressing Tk-Tes. However, the RI increase is much lower at both 85˚C (RI = 0.016) and 95˚C (RI = 0.013) than in the complementation strain but still an increase as compared to 0.002 and 0.004 in the parental strain at 85˚C and 95˚C, respectively (Fig. A.S5). These findings indicate that the activities of Grs, which may be temperature-regulated, are impacted by the ectopic expression of Tes.

Discussion

Like all archaea, *T. kodakarensis* generates an isoprenoid-based lipid membrane that responds to environmental and growth phase changes. Instead of a conventional bilayer, the membranes of many archaea are dominated by membrane-spanning tetraether lipids termed GDGTs that are synthesized by the activities of a tetraether lipid synthase (Tes) [20,21]. GDGTs can be modified through the formation of ring structures within the dibiphytanyl tails by the activity of GDGT-ring synthase (Grs) [22]. Ring-containing lipids are typically only minor constituents of the total lipidome of the Thermococcales [24] but can be the dominant membrane lipids in other Archaea including Thaumarchaeota [33], Sulfolobales [34], and Thermoplasmatales [35] species. Cyclized GDGTS are thought to impart unique properties to membranes that support survival in the extremes [4,12,23,36–38] but the synthesis, modification, and impacts of such on hyperthermophily are understudied *in vivo*.

In this study, we established that *T. kodakarensis* strains lacking the tetraether synthase Tes (TK2145) are viable despite the complete lack of tetraether lipids. The hyperthermophilic growth of *T. kodakarensis* is thus not dependent on tetraether lipids, although the lack of tetraether lipid synthesis leads to a dramatic loss of viability upon entry into the stationary phase. In the absence of natural GDGT substrates in the *T. kodakarensis Δtes* deletion strain, the GDGT cyclase Grs instead modifies ~7% of archaeol to mono- or di-ring containing compounds (archaeol-1 and archaeol-2; Fig. 1) that may assist in preserving lipid membrane integrity and function in the absence of tetraether lipids. Deletion of Grs (TK0167) eliminates the production of the small (<1%) percentage of cyclic GDGTs found in natural membranes and is generally not phenotypic. While loss of ring-containing GDGTs is fitness-neutral under laboratory conditions, aberrant levels of ring-containing GDGTs are not well tolerated at supra-optimal growth temperatures upon overexpression of Grs *in vivo*. Overproduction of Grs throughout all growth phases does provide a route to abundant (~53%, 19%, and 28% of tetraether lipids at

60˚C, 85˚C, and 95˚C, respectively) ringed-GDGT production, but these high levels of derivatized GDGTs led to significant growth defects at 95˚C. Furthermore, the effects of temperature on cyclization appear to be complex, as increased levels of cyclized GDGTs are observed at both supra-optimal (95˚C) and sub-optimal (60˚C) temperatures. While increases in cyclized GDGTs in response to increased temperature have been well documented in many archaea, other physiological factors known to influence GDGT cyclization, such as growth rate, also change with temperature and thus may have confounding effects on cyclization levels. Alternatively, the turnover of cyclized GDGTs could be reduced at suboptimal growth temperatures, or Tk-Grs may simply function more efficiently under such conditions. Taken together, our findings show the importance of tetraether lipid synthesis in the long-term survival of *T. kodakarensis.* The regulation of lipid composition appears to be critical for both archaeal hyperthermophily and growth phase transitions. It is also likely that combinatorial changes to environmental variables (e.g., temperature, salinity, pressure, and pH) will also direct changes to the lipidome that will reveal other critical roles for tetraether lipids and their cyclized derivatives.

Our findings also demonstrate the production of new lipid types, specifically archaeol-1 and archaeol-2. While macrocyclic archaeol containing one and two rings has been observed in environmental lipid extracts [39], the presence of rings in archaeol and cyclized diether lipids in culture have not been previously observed. Tk-Grs appears to be promiscuous, allowing the production of cyclized diethers when ectopically expressed throughout the growth phases, especially at 60˚C where they comprise half of the diether lipids, interestingly equal to the proportion of GDGTs cyclized at this temperature. This suggests that Tk-Grs can work equally well on both diether and tetraether lipids under certain conditions.

The mass spectra of the two archaeol-1 isomers provide further biochemical insights into Tk-Grs function. Fragmentation patterns demonstrate that Tk-Grs can form the first ring in either lipid tail but that it prefers to do so on the tail furthest from the headgroup. Additionally, the presence of multiple chromatographically resolved isomers of archaeol-1 and -2 has interesting implications for our understanding of the stereochemistry of the ring moieties in archaeal lipids (Fig. A.S11A & B, A.S12A-E). While the observed isomers of archaeol-1 and -2 could in theory be the result of different ring shapes (e.g., cyclohexane) or cyclization at different positions along the phytanyl tail (e.g., at C-3), no biochemical precedence exists for such multifaceted activity of Grs. Further, we 1) do not observe LC chromatographic resolution of the isomers that would suggest varying ring shapes [40], 2) do not detect the presence of archaeol with more than 2 rings (possible if C3 and C7 cyclization occurs), and 3) do not detect fragments in the mass spectra of the archaeol-2 isomers that would indicate that the two rings can be found together on one tail. Thus, we hypothesize that the different isomers of archaeol-1 and -2 are diastereomers of one another, differing in their stereochemistry around the cyclopentane ring (cis vs. trans configuration) (Fig. A.S11A & B, A.S12A-E). Previous studies have determined that the stereochemistry of the cyclopentane rings in GDGTs from *S. acidocaldarius* [41], crenarchaeol from Thaumarchaeota [42], and environmental GDGT-derived compounds [43] appears to be exclusively trans, particularly with a C7(*S*)-C10(*S*) configuration. However, one notable exception is the potential presence of a cis-configured cyclopentane ring in the enigmatic crenarchaeol isomer [44]. Thus, it is unclear if the suggested additional stereoisomers of archaeol-1 and archaeol-2 are merely artifacts of the “unnatural” cyclization of archaeol or if they are indicative of the stereochemistry of the rings in the GDGTs of *T. kodakarensis* as well. Given the demands for biological routes for large-scale production of derivatized GDGTs [17–19], strains that express wildtype and potential-variant forms of Grs hold substantial promise.

Rescued recovery of GDGT synthesis via ectopic Tk-Tes expression in genomically-deleted Tes strains also reproducibly increases the total amount of cyclic GDGTs observed in comparison to the overexpression of Tk-Tes in the parental strain. Based on previous *T. kodakarensis* transcriptomic data, Tk-Tes (encoded by TK2145) is co-expressed with TK2146, which is annotated as a hypothetical regulatory protein. Perhaps the co-expression of Tes and TK2146 in the parental strain but not in the *Δtes* strain played a role in the regulation of cyclopentane ring production in *T. kodakarensis.* These findings ultimately warrant more investigation into the main regulatory pathway that controls the production of cyclic GDGTs.

In conclusion, the ability to manipulate lipidome composition in *T. kodakarensis* offers a powerful mechanism to study the impacts of tetraether lipid biosynthesis on archaeal physiology and survival in the extremes. Investigations into the interplay between GDGT biosynthesis, modification, and cellular viability in different environments will allow a better understanding of the roles of these tetraether lipids in the evolution of archaeal organisms. Although the properties of these new lipids still require further investigation, the expansion and control of lipid diversity in *T. kodakarensis* can be utilized for biotechnology applications in relation to drug deliveries and vaccines.

Materials and Methods

Microbial growth and media conditions

The constructed strains of *T. kodakarensis* were anaerobically cultured as previously described in artificial sea water (ASW) media supplemented with 5 g/l tryptone, 5 g/l yeast extract, 5 g/l pyruvate, 2 g/l elemental sulfur (S˚), and a KOD1-vitamin mixture with or without 1 mM agmatine [29]. Culture growth at 85˚C or 95˚C was monitored using optical density measurements at 600 nm. Growth rates of a minimum of three independent biological replicates were monitored and plotted for each strain. Biomass harvested for lipid extractions and analyses were either grown

at 85˚C or 95˚C and harvested 24 hours post-inoculum (late-stationary phase) and stored frozen (at -80˚C or on dry ice) until sample processing for lipid extraction.

T. kodakarensis strain constructions

Strains used in this study are listed in Table 1. All *T. kodakarensis* deletion strains were constructed via homologous recombination resulting in markerless deletions on the genome as previously described [29]. Deletion strain genotypes were confirmed through whole genome sequencing (WGS) at >100x coverage using Oxford Nanopore MinION sequencing and visualized using Integrative Genomics Viewer. Complemented strains were constructed as previously described [30,31]. Briefly, strains carrying the expression vectors were selected based on agmatine autotrophy and cultured without agmatine supplementation. Retention of expression plasmids were confirmed via PCR using primers flanking the insertion site of the gene of interest.

Lipid extraction and analyses

Frozen biomass samples were freeze-dried, resuspended in methanol (MeOH), and transferred to glass tubes where the solvent was evaporated under a N_2 stream. Samples were acid hydrolyzed in 2 mL 1 M HCl in MeOH for 3 hours at 90˚C before being neutralized by addition of 1 mL 2 M KOH in MeOH and diluted with 5 mL of deionized water. Samples were extracted three times with 5 mL dichloromethane (DCM) which was pooled and evaporated under a N_2 stream. Samples were then resuspended in 1 mL of 9:1 MeOH:DCM and filtered through 0.45-µm polytetrafluoroethylene filters.

Core lipids were analyzed on an Agilent 1260 Infinity II series high performance liquid chromatography (HPLC) instrument coupled to an Agilent G6125B single quadrupole mass spectrometer with the electrospray ionization (ESI) interface in positive mode. ESI-MS

conditions were as follows: drying gas temperature 300˚C, drying gas flow rate 8.0 L/min, nebulizer pressure 35 psi, capillary voltage 3500 V, and fragmentor voltage 175 V in scanning mode with a range of m/z = 600-1400.

Core lipids were separated with reverse phase chromatography on a Kinetex 1.7 µm XB-C18 100 Å LC column (150 x 2.1 mm) by a method modified from Rattray and Smittenberg [45] with mobile phase A: MeOH with 0.04% formic acid and 0.03% NH_3 and mobile phase B: isopropanol with 0.04% formic acid and 0.03% NH_3. An initial mobile phase of 60A:40B was held for 1 minute and then linearly ramped to 50A:50B over 19 minutes. This composition was held for 15 minutes and then linearly ramped back to 60A:40B over 5 minutes which was then held for 10 minutes to allow for re-equilibration. A large injection volume of 25 µL was used to detect all compounds of interest. Compounds were identified by the mass of the protonated parent ion ([M+H]+) coupled with comparison of elution times to laboratory standards or to those found in previous literature [45]. The non-response factor corrected relative intensities of compounds were calculated using the manually integrated peak areas of the [M+H]+ ions only.

For gas chromatography-mass spectrometry (GC-MS) analysis of diether lipids, acid hydrolyzed lipid extracts were trimethylsilyl (TMS) derivatized in 100 µL of a 1:1 solution of [pyridine]: [N,O-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) with 1% trimethylchlorosilane (TMCS)] for 1 hour at 70˚C. 5 µL of the resulting reaction mixture was immediately injected on the GC-MS for analysis.

Diether lipids were analyzed on an Agilent 7890B Series GC instrument coupled to an Agilent 5977A Series MSD in EI mode at 70 eV, scanning over a range of m/z = 50 - 900. Lipids were separated on two Agilent DB-17HT columns (30 m x 0.25 mm x 0.15 µm film thickness) connected in series using helium as the carrier gas with a constant flow rate of 1.1 mL/min. GC

conditions were as follows: 60˚C to 200˚C at 10 degrees/min, then 200˚C to 300˚C at 4 degrees/min, and finally held at 300˚C for 60 minutes.

Culture of *Halorubrum lacusprofundi* and Base Hydrolysis of Biomass

Liquid cultures (75 mL) of *H. lacusprofundi* DSM 5036 were grown on DSMZ Medium 372 at 10 °C, shaking for five months. *H. lacusprofundi* biomass was base hydrolyzed in 2 mL 1M KOH in MeOH for 3 hours at 70 °C. Reactions were neutralized with 1 mL 2M HCl in MeOH. The core lipids were then extracted and analyzed as described for *T. kodakarensis*.

Lipid sample hydrogenation

The lipid extract was resuspended in 2 mL of 1:1 MeOH:ethyl acetate (EtOAc). Argon was bubbled through the solution for 10 min before platinum(IV) oxide (~15 mg, 66 µmol) was added. The reaction mixture was placed in a Parr pressure vessel which was then pressurized to 60 psi with H2, and the reaction mixture was continuously stirred with magnetic stir bars. After 16 hours, the pressure was released, and argon was bubbled through the reaction for 10 min. The reaction was then filtered through celite with additional portions of EtOAc and then concentrated under vacuum.

Supplementary Figures

Fig. A.S1. Whole genome sequencing confirmation of the partial deletion strain of TK0167 (*Δgrs*) and full deletion strain of TK2145 (*Δtes*). (A & B) Integrated genome viewer panels highlight the sequencing alignments for AL010 and AL016, confirming each deletion strain with
>100x coverage.

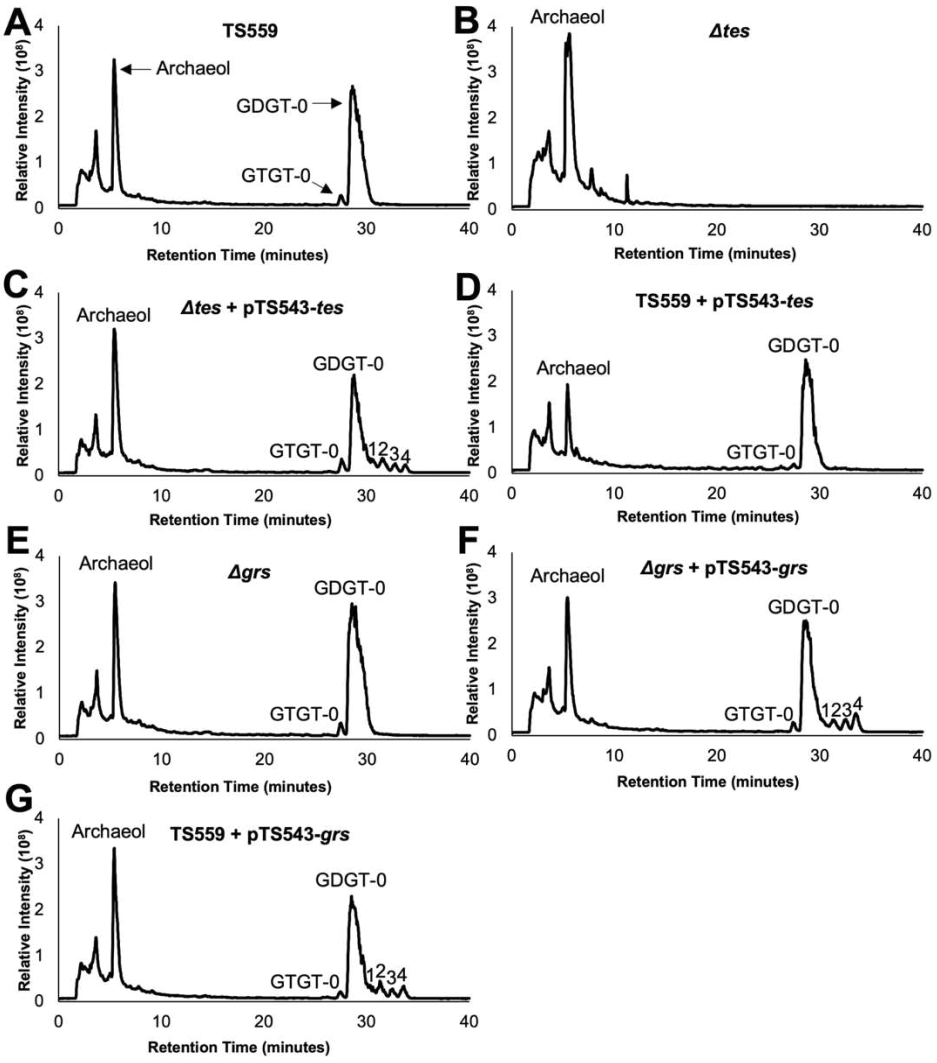


Fig. A.S2. Deletion of *tes* results in the loss of tetraether lipids and ectopic expression of *grs* results in the robust production of cyclized GDGTs. LC-MS total ion chromatograms (TICs) of acid hydrolyzed lipid extracts of *T. kodakarensis* strains grown at 85˚C to late-stationary phase. Numbers 0-4 denote the number of rings within GTGT and GDGT. (A) TIC from strain TS559 demonstrates the presence of archaeol, GDGT-0, and GTGT-0; cyclized derivatives of GDGT are not sufficiently abundant to be seen in the TIC – they are however observed in the extracted ion chromatograms of this strain. (B) TIC from the *Δtes* strain demonstrates the loss of peaks corresponding to GTGT and GDGT. (C) TIC from the *Δtes* + pTS543-*tes* strain demonstrates the restoration of GTGT and GDGT lipid peaks due to ectopic Tes expression and the unanticipated presence of abundant cyclized GDGTs. (D) TIC from the TS559 + pTS543-*tes* strain demonstrates that ectopic expression of Tes in a Tes-containing strain does not result in substantial changes to membrane lipid composition. (E) TIC from the *Δgrs* strain. (F) TIC from the *Δgrs* + pTS543-*grs* strain demonstrates the restoration of abundant cyclized GDGTs due to ectopic Grs expression. (G) TIC from the TS559 + pTS543-*grs* strain demonstrates an increase of cyclized GDGTs due to ectopic Grs expression.

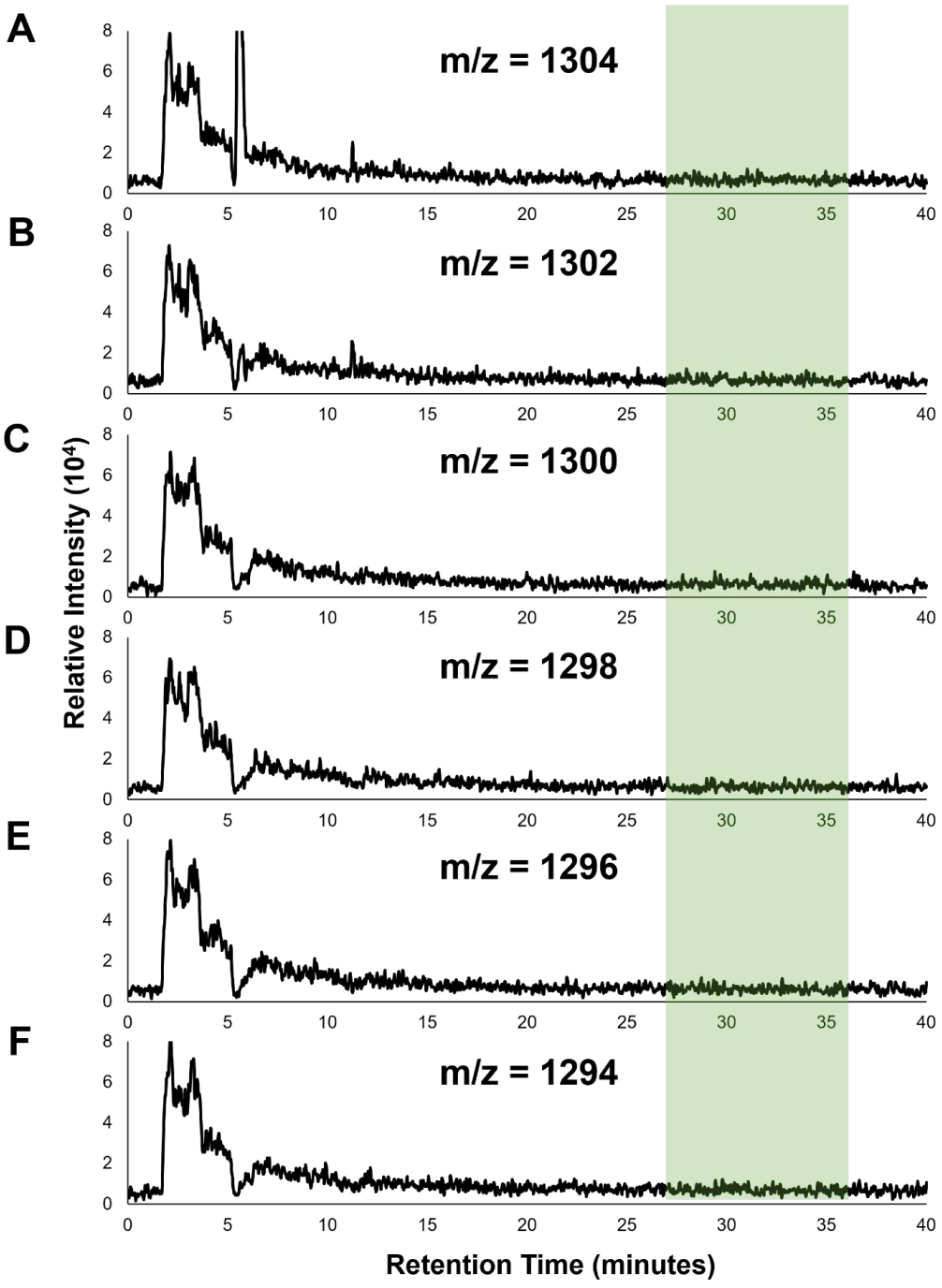


Fig. A.S3. Deletion of Tes in *T. kodakarensis* eliminates production of all tetraether lipids. Extracted ion chromatograms (EICs) of m/z = 1304 (A), 1302 (B), 1300 (C), 1298 (D), 1296 (E), and 1294 (F) of acid hydrolyzed lipids extracted from the *Δtes* strain grown at 85˚C to late-stationary phase demonstrate the complete absence, respectively, of GTGT-0, GDGT-0, GDGT-1, GDGT-2, GDGT-3, and GDGT-4 in the *Δtes* strain. Note: tetraether lipid species elute between ~27-35 minutes (denoted by the green box).

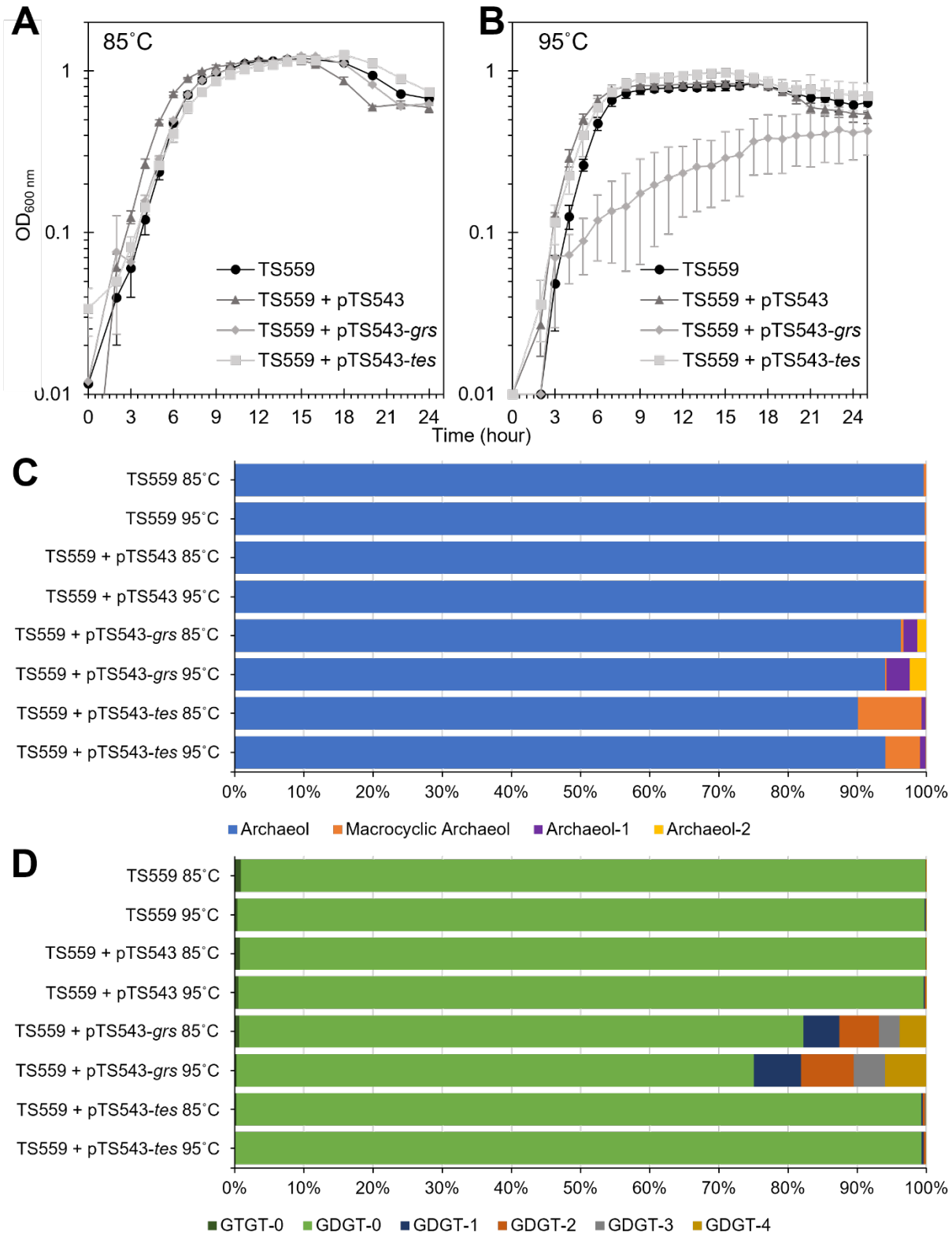

A
85˚C
B
95˚C
OD600 nm
1
0.1
0.01
0 3 6 9 12 15 18 21 24
Time (hour)
TS559
TS559 + pTS543
TS559 + pTS543-grs
TS559 + pTS543-tes
C
TS559 85˚C
TS559 95˚C
TS559 + pTS543 85˚C
TS559 + pTS543 95˚C
TS559 + pTS543-grs 85˚C
TS559 + pTS543-grs 95˚C
TS559 + pTS543-tes 85˚C
TS559 + pTS543-tes 95˚C
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Archaeol
Macrocyclic Archaeol
Archaeol-1
Archaeol-2
D
GTGT-0
GDGT-0
GDGT-1
GDGT-2
GDGT-3
GDGT-4

Fig. A.S4. Ectopic expression of Tk-Grs and Tk-Tes activities alters the lipidome of *T. kodakarensis*. (A & B) The growth of triplicate biological cultures of *T. kodakarensis* strains TS559 (black), TS559 + pTS543 (dark grey), TS559 + pTS543-TK0167 (grey), and TS559 + pTS543-TK2145 (light grey) were monitored via changes in optical density (at 600nm) at 85˚C and 95˚C, respectively, revealing that ectopic expression of Tk-Grs (TK0167) (and the accompanied increased cyclization of GDGT) resulted in a major growth defect at 95˚C but not 85˚C, congruent with the growth analysis of *Δgrs* + pTS543-TK0167 (Fig 3A & B). (C & D) Lipid analysis of the parental strain (TS559) showed that GDGT-0 was the dominant tetraether core lipid extracted and that only a small, but reproducible, amount of GDGT-1, -2, -3, and -4 were detected, whereas >100-fold increases in cyclized GDGTs were obvious upon overexpression of Tk-Grs. Panel C and D show diether and tetraether lipid production, respectively.

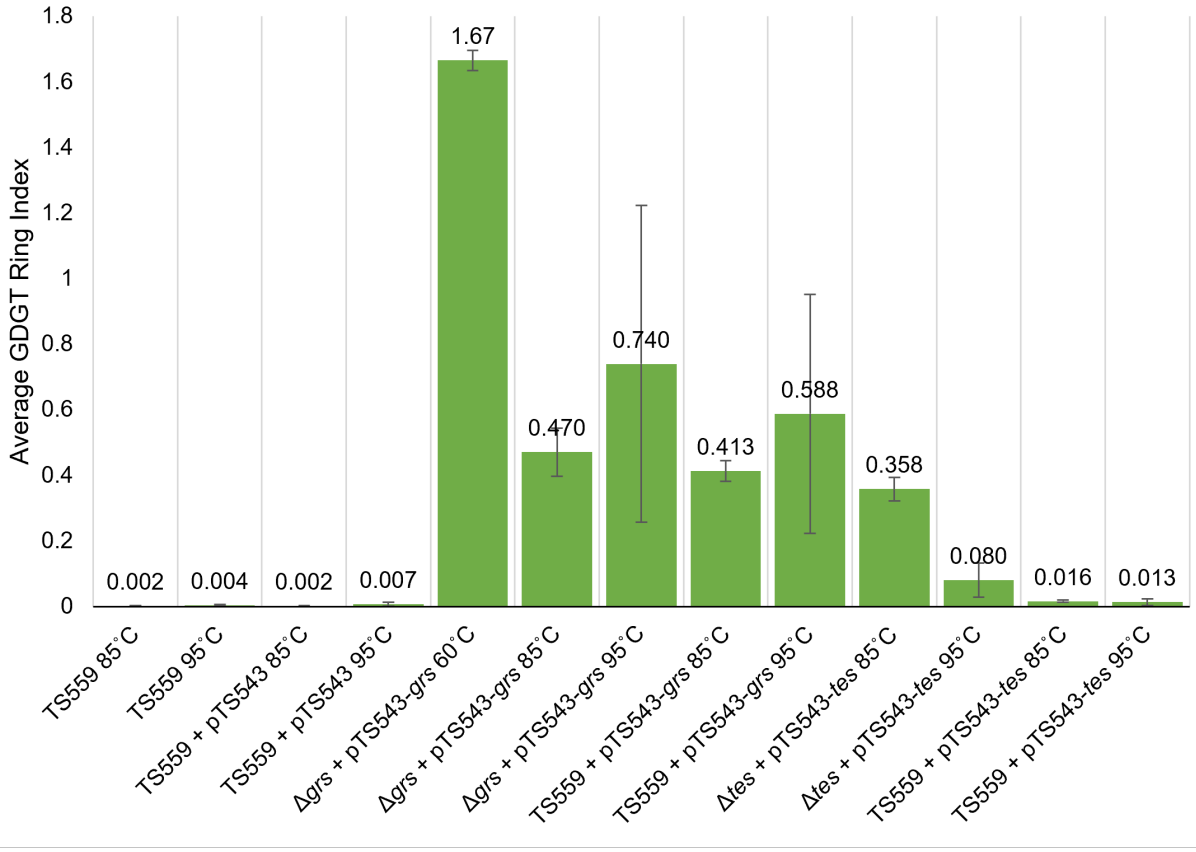


Fig. A.S5. Dramatic differences in the GDGT Ring Index (RI) are observed across temperatures and strains. Bar plots of the RI: $\frac{[(GDGT-1)+(GDGT-2\ x\ 2)+(GDGT-3\ x\ 3)+(GDGT-4\ x\ 4)]}{(GDGT-0+GDGT-1+GDGT-2+GDGT-3+GDGT-4)}$ calculated from the relative abundance of core GDGT lipids in acid hydrolyzed extracts of *T. kodakarensis* strains grown at 60˚C, 85°C, and 95°C to late stationary phase. Error bars show standard deviation from the mean of a minimum of three biological replicates.

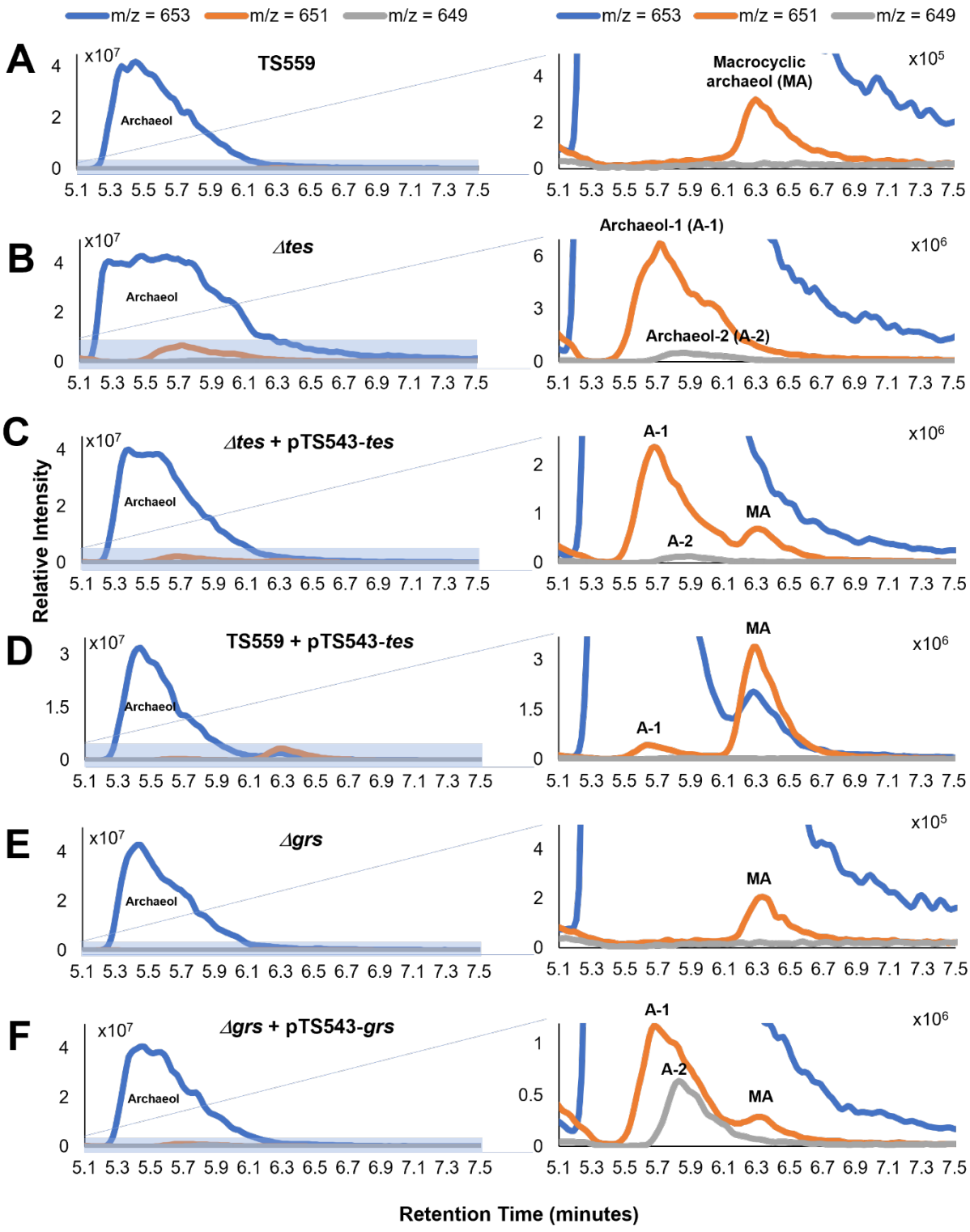

m/z = 653
m/z = 651
m/z = 649
A
TS559
Archaeol
Macrocyclic archaeol (MA)
B
Δtes
Archaeol-1 (A-1)
Archaeol-2 (A-2)
C
Δtes + pTS543-tes
A-1
MA
A-2
Relative Intensity
D
TS559 + pTS543-tes
E
Δgrs
F
Δgrs + pTS543-grs
Retention Time (minutes)

Fig. A.S6. Minor diether lipid species vary substantially amongst the *T. kodakarensis* strains used in this study. LC-MS overlayed extracted ion chromatograms (m/z = 653, 651, 649) showing the diether lipids present in the acid hydrolyzed lipid extracts of *T. kodakarensis* strains grown at 85˚C to late-stationary phase. (A) Overlayed extracted ion chromatogram (EIC) of TS559 showing the major presence of archaeol and a minor amount of macrocyclic archaeol. (B) Overlayed EIC of *Δtes* showing the presence of archaeol and the appearance of putative cyclopentane ring containing derivatives, termed archaeol-1 and archaeol-2, as well as the absence of macrocyclic archaeol upon *tes* deletion. (C) Overlayed EIC of *Δtes* + pTS543-*tes* showing the presence of archaeol and the reappearance of macrocyclic archaeol upon *tes* complementation, as well as the presence of archaeol-1 and -2, consistent with the observation that the *tes* complementation strain unexpectedly produces abundant cyclized GDGTs at 85˚C. (D) Overlayed EIC of TS559 + pTS543-*tes* showing the increase in macrocyclic archaeol abundance upon likely *tes* overexpression as well as the presence of a minor amount of archaeol-1, consistent with the observation that the abundance of cyclized GDGTs increases in this strain relative to the parental strain, TS559. (E) Overlayed EIC of *Δgrs* showing the major presence of archaeol and a minor amount of macrocyclic archaeol. (F) Overlayed EIC of *Δgrs* + pTS543-*grs* showing the appearance of archaeol-1 and archaeol-2 upon *grs* ectopic expression, as well as the presence of archaeol and macrocyclic archaeol.

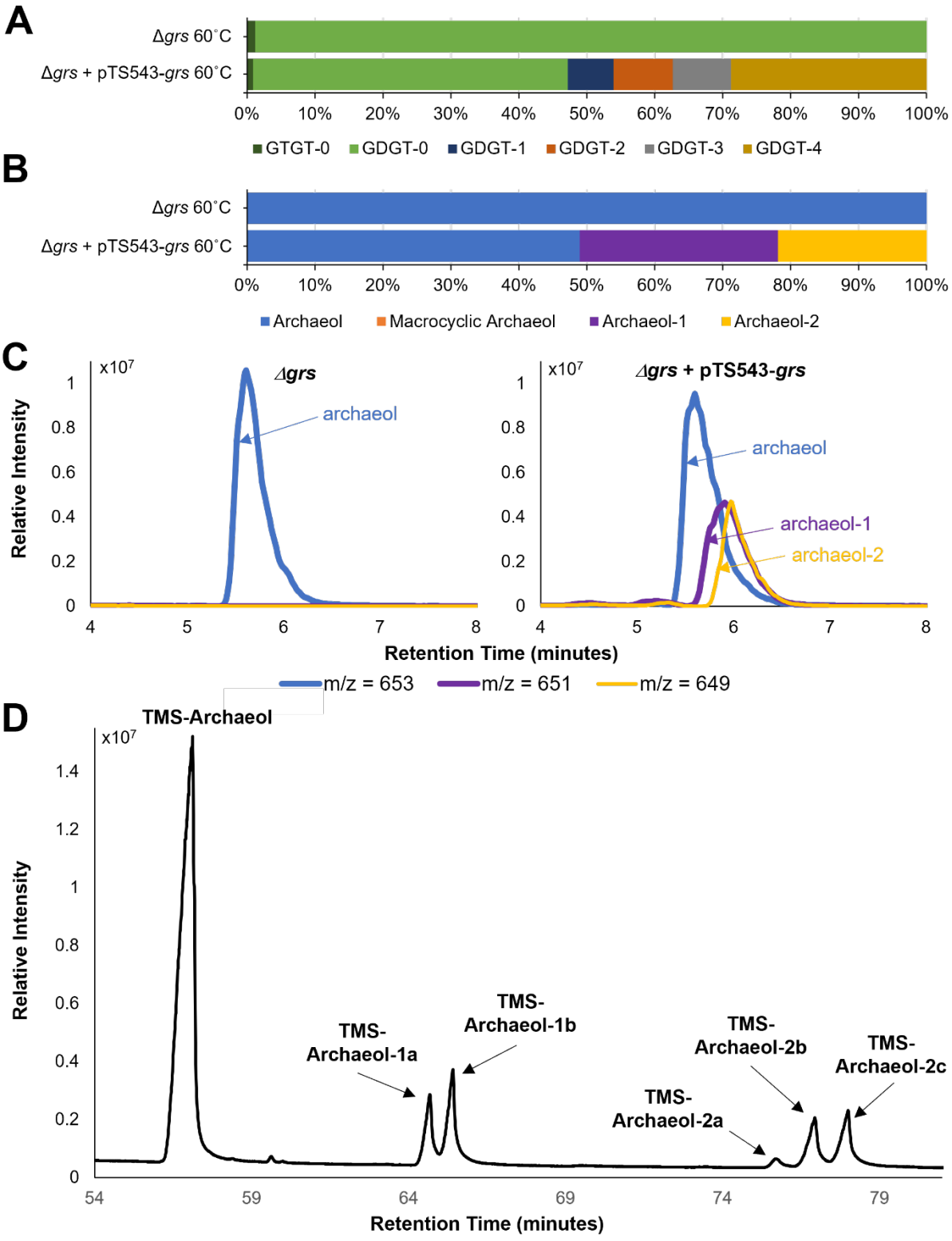

A
Δgrs 60˚C
Δgrs + pTS543-grs 60˚C
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
GTGT-0 GDGT-0 GDGT-1 GDGT-2 GDGT-3 GDGT-4
B
Δgrs 60˚C
Δgrs + pTS543-grs 60˚C
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Archaeol Macrocyclic Archaeol Archaeol-1 Archaeol-2
C
Relative Intensity
x10^7
Δgrs
archaeol
x10^7
Δgrs + pTS543-grs
archaeol
archaeol-1
archaeol-2
Retention Time (minutes)
m/z = 653 m/z = 651 m/z = 649
D
TMS-Archaeol
x10^7
Relative Intensity
TMS-Archaeol-1a
TMS-Archaeol-1b
TMS-Archaeol-2a
TMS-Archaeol-2b
TMS-Archaeol-2c
Retention Time (minutes)

Fig. A.S7. Ectopic expression of *grs* at 60°C results in the robust production of archaeol-1 and -2, each present in multiple isomeric forms. (A & B) Bar plots of the average relative abundance (n = 4) of core tetraether (A) and core diether (B) lipids present in acid hydrolyzed lipid extracts of *Δgrs* and *Δgrs* + pTS543-*grs* strains grown at 60°C to late stationary phase. The cyclized derivatives of the diether and tetraethers both comprise ~50% of the relative abundance of their respective lipid species in the strain ectopically expressing *grs*. (C) LC-MS overlayed extracted ion chromatograms (m/z = 649, 651, 653) of acid hydrolyzed lipid extracts of *Δgrs* and *Δgrs* + pTS543-*grs* strains grown at 60°C to late stationary phase. Note the sequential delayed elution times of archaeol-1 and archaeol-2 relative to archaeol. (D) GC-MS total ion chromatogram of trimethylsilyl (TMS) derivatized acid hydrolyzed lipid extract of the *Δgrs* + pTS543-*grs* strain grown at 60°C to late stationary phase showing the presence of two chromatographically resolved isomers of archaeol-1 (archaeol-1a and 1b) and three of archaeol-2 (archaeol-2a, 2b, and 2c).

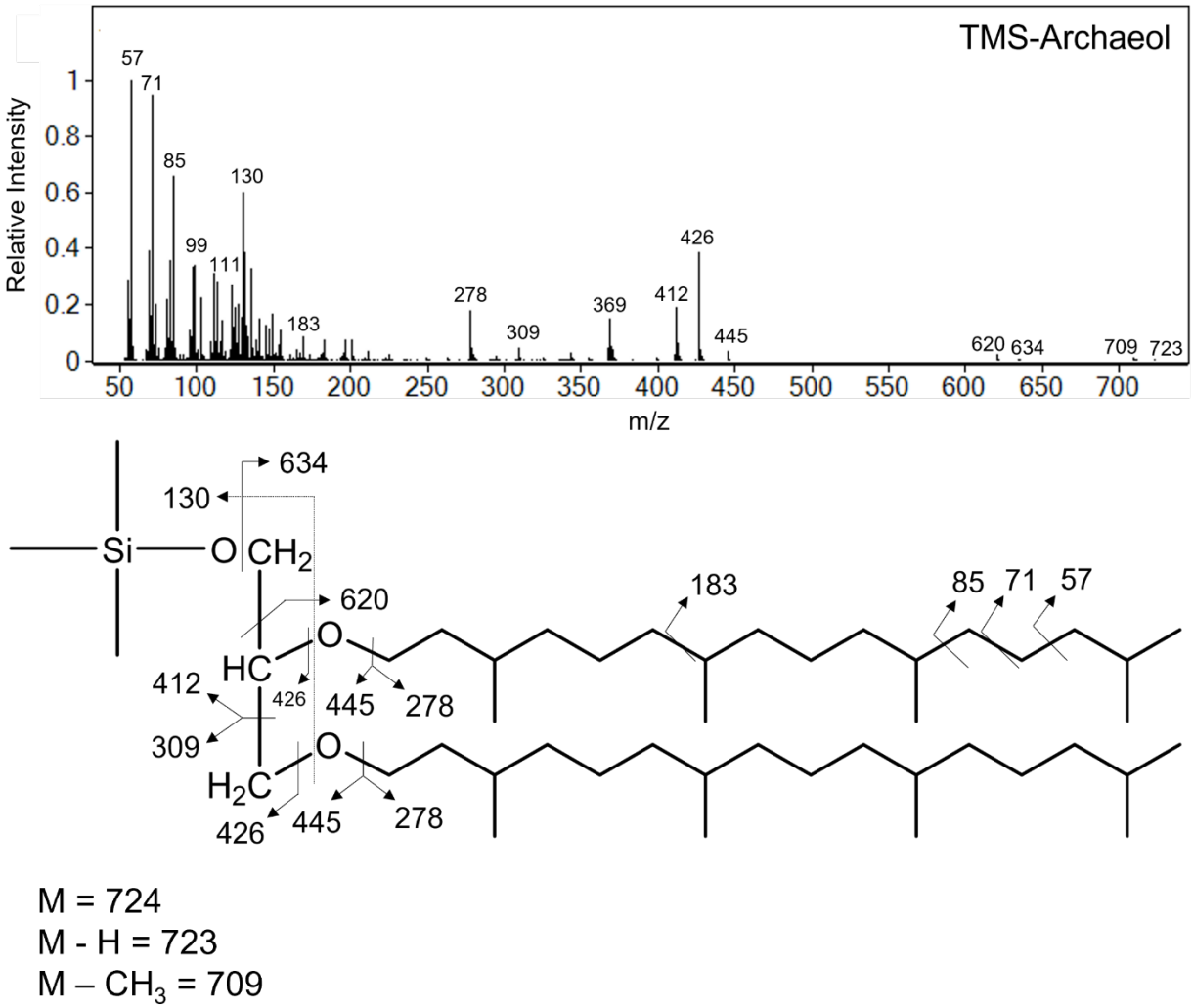


Fig. A.S8. Electron impact mass spectrum of TMS derivatized archaeol. M denotes the molecular ion. Bond cleavages producing the observed mass spectrum fragments are shown with arrows.

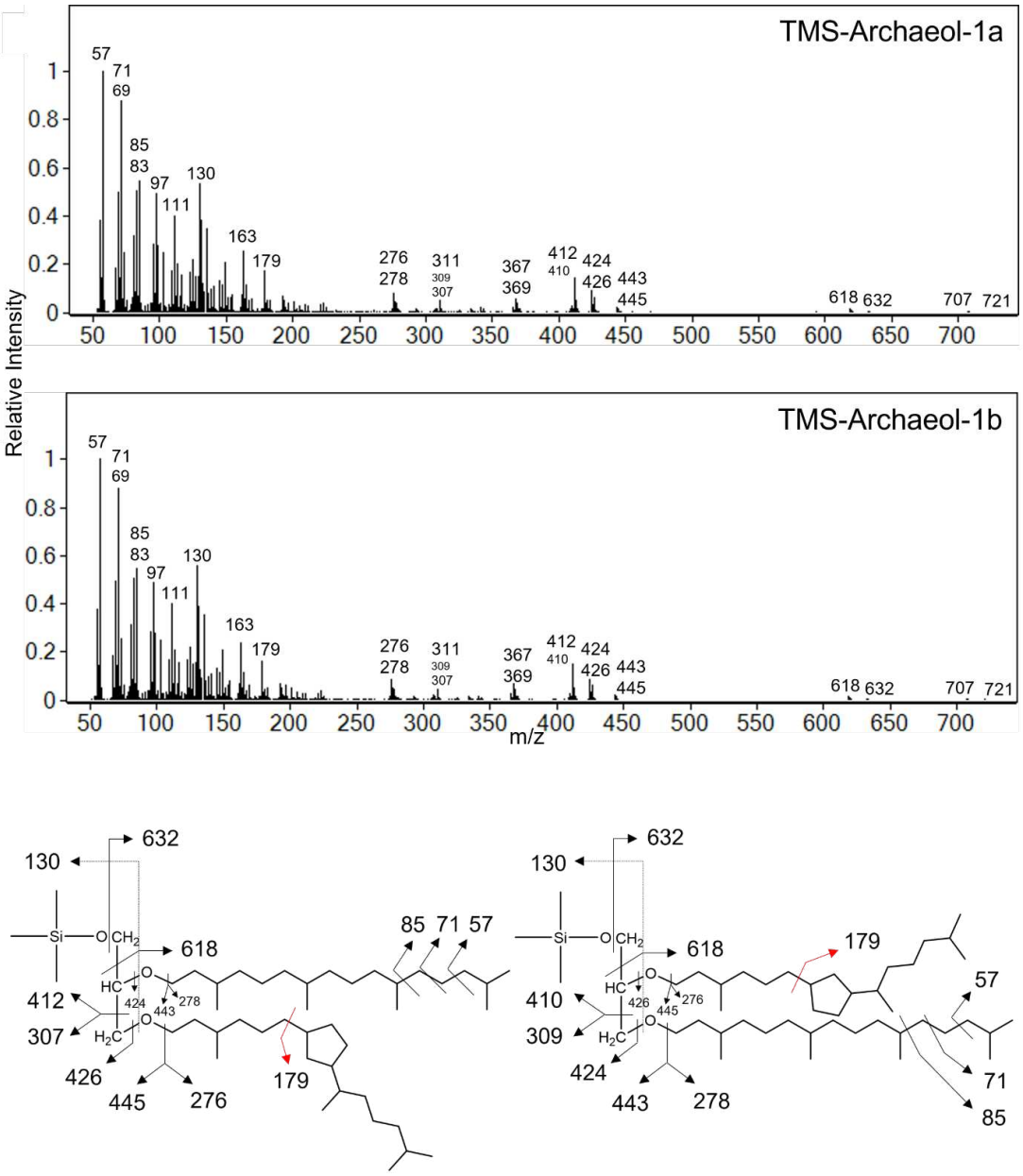
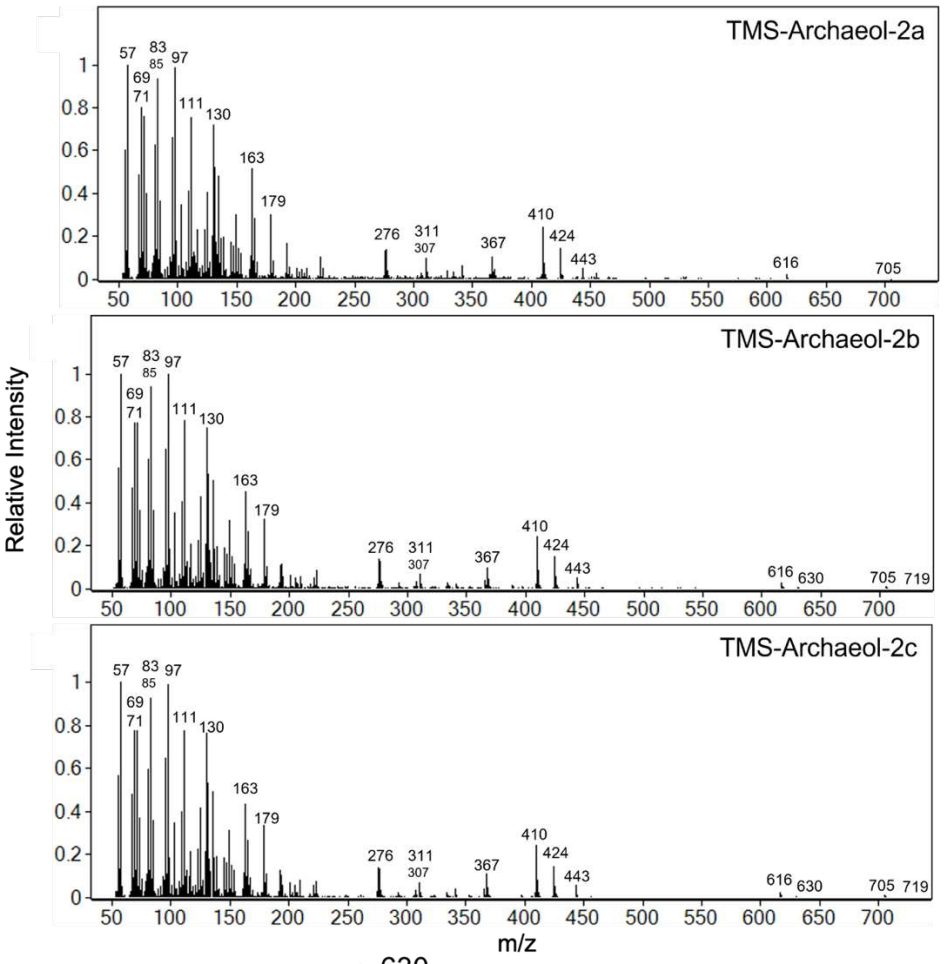
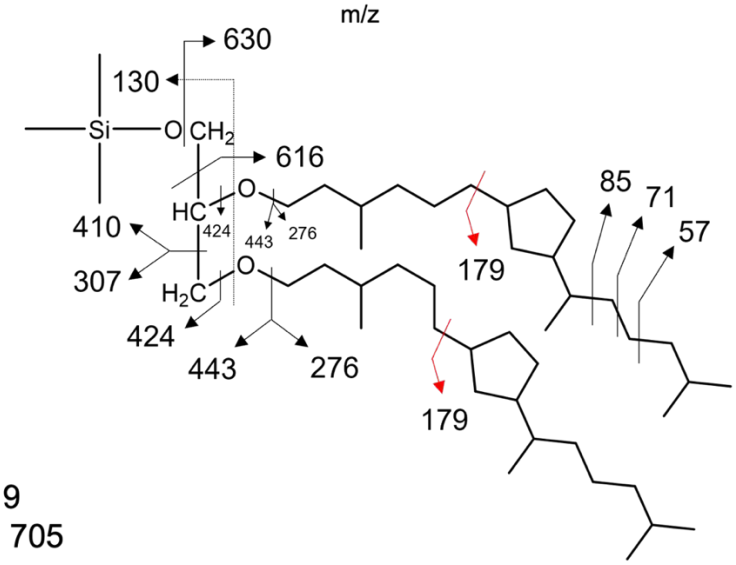


Fig. A.S9. Electron impact mass spectra of TMS derivatized archaeol-1a and archaeol-1b.

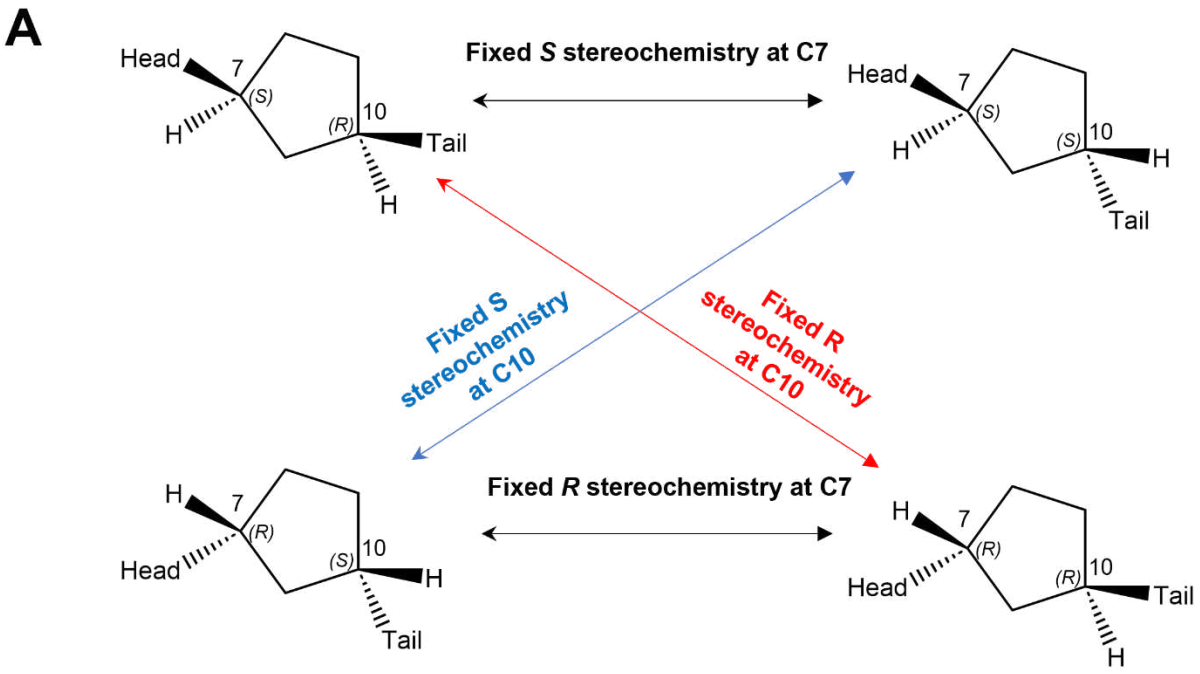
M denotes the molecular ion. Bond cleavages producing the observed mass spectrum fragments are shown with arrows. Speculative cleavage adjacent to the ring is shown in red. Note: spectra from both isomers generates a combination of m/z = 410 and 412 fragments, indicating both isomers (1a and 1b) are each a mixture of the two depicted regioisomers of archaeol-1.





M = 720
M - H = 719
M – CH_3 = 705

Fig. A.S10. Electron impact mass spectra of TMS derivatized archaeol-2a, archaeol-2b, and archaeol-2c. M denotes the molecular ion. Bond cleavages producing the observed mass spectrum fragments are shown with arrows. Speculative cleavages adjacent to the ring are shown in red.


A
Head
7
(S)
H
(R)
10
Tail
H
Fixed S stereochemistry at C7
Head
7
(S)
H
(S)
10
H
Tail
Fixed S stereochemistry at C10
Fixed R stereochemistry at C10
H
7
(R)
Head
(S)
10
H
Tail
Fixed R stereochemistry at C7
H
7
(R)
Head
(R)
10
Tail
H

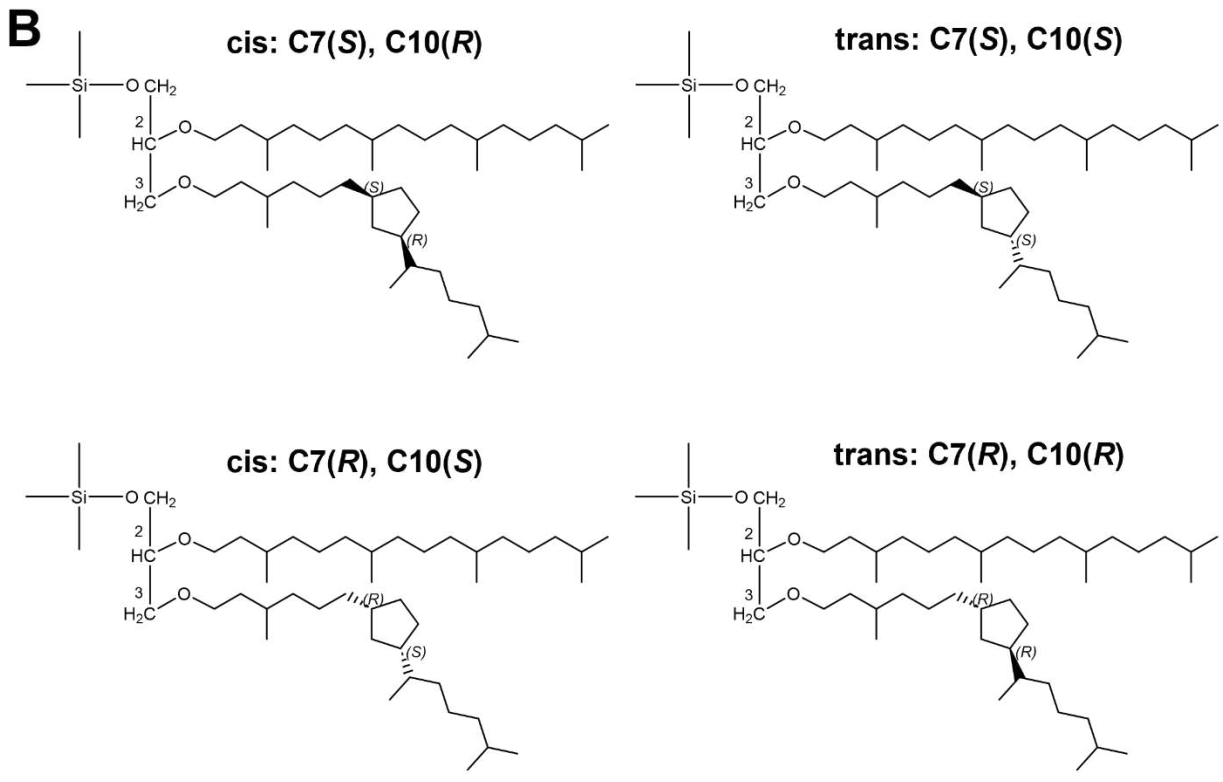

B
cis: C7(S), C10(R)
trans: C7(S), C10(S)
cis: C7(R), C10(S)
trans: C7(R), C10(R)
Si
O CH2
2
HC
O
3
H2C
O
(S)
(R)

Fig. A.S11. The cyclopentane rings in archaeal lipids possess two stereocenters. (A) Depictions of the chiral centers at C7 and C10 in the cyclopentane rings of archaeal lipids and the diastereomeric pairs that may arise if the stereochemistry at one carbon is fixed during the cyclization reaction while the other is not. “Head” refers to the alkyl chain proximal to the headgroup; “tail” refers to the alkyl chain distal to the headgroup. (B) Illustrations of the four possible diastereomers of archaeol-1 of the regioisomer bearing the ring on the tail bonded to the C3 carbon of glycerol; the other possible regioisomer, archaeol-1 bearing the ring on the other tail (that bonded to the C2 carbon of glycerol), also possess four potential diastereomers (the mass spectra of the two archaeol-1 isomer peaks indicates that both are a mix of regioisomers). Note: The two cis isomers are not enantiomers of each other, and neither are the two trans isomers – although they possess mirrored stereochemistry at the cyclopentane ring, the archaeol molecules possess multiple other chiral centers which will not change during the cyclization reaction and will thus not mirror one another, making all four stereoisomers diastereomers of one another.

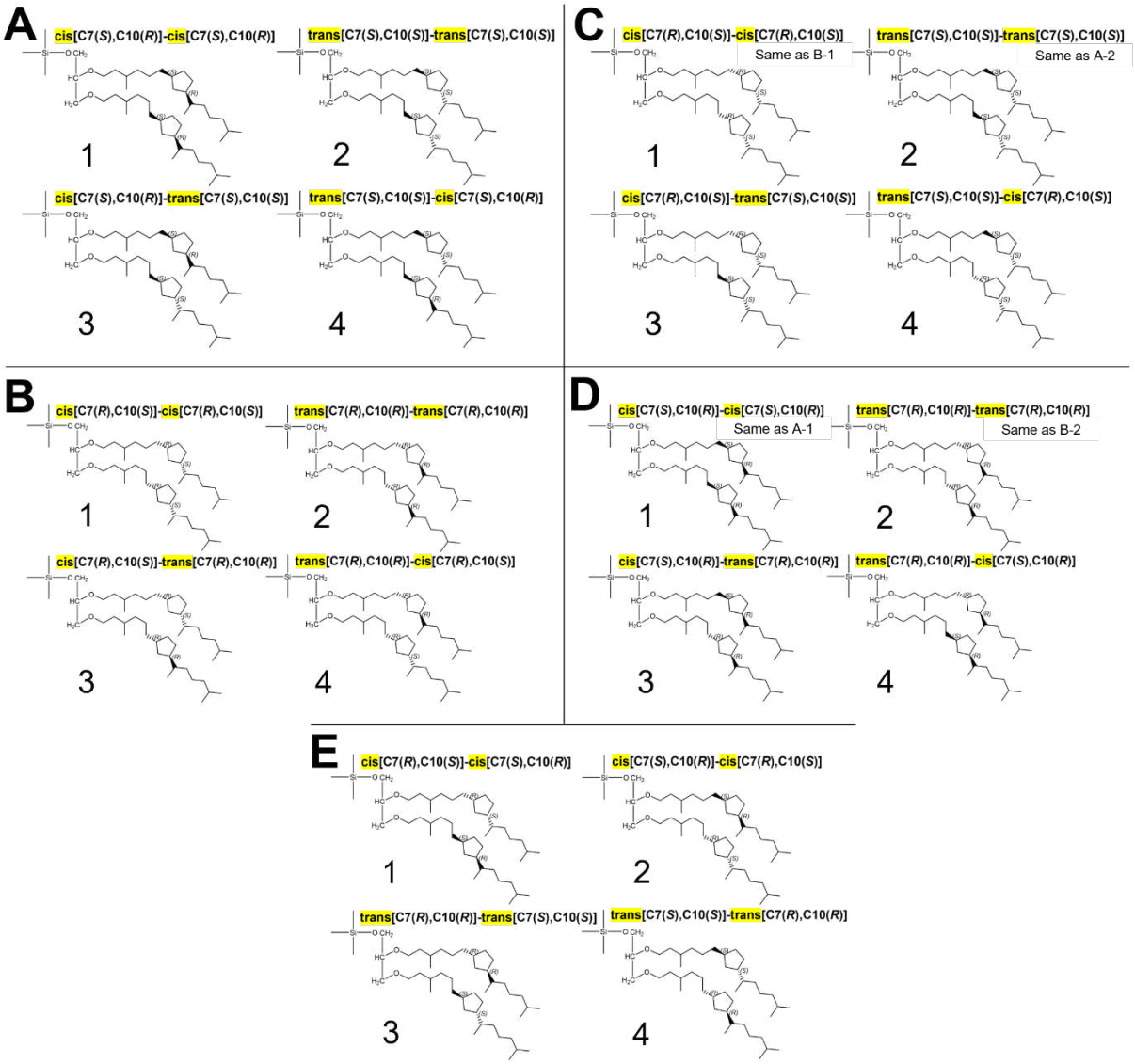


Fig. A.S12. Archaeol-2 has 16 potential diastereomers. (A) The four potential diastereomers of archaeol-2 if there is fixed *S* stereochemistry at C7. (B) The four potential diastereomers of archaeol-2 if there is fixed *R* stereochemistry at C7. (C) The four potential diastereomers of archaeol-2 if there is fixed *S* stereochemistry at C10. Note: two of the diastereomers seen here were seen previously in A and B. (D) The four potential diastereomers of archaeol-2 if there is fixed *R* stereochemistry at C10. Note: two of the diastereomers seen here were seen previously in A and B. (E) Four additional diastereomers which are only possible if the stereochemistry is not fixed at both C7 and C10.

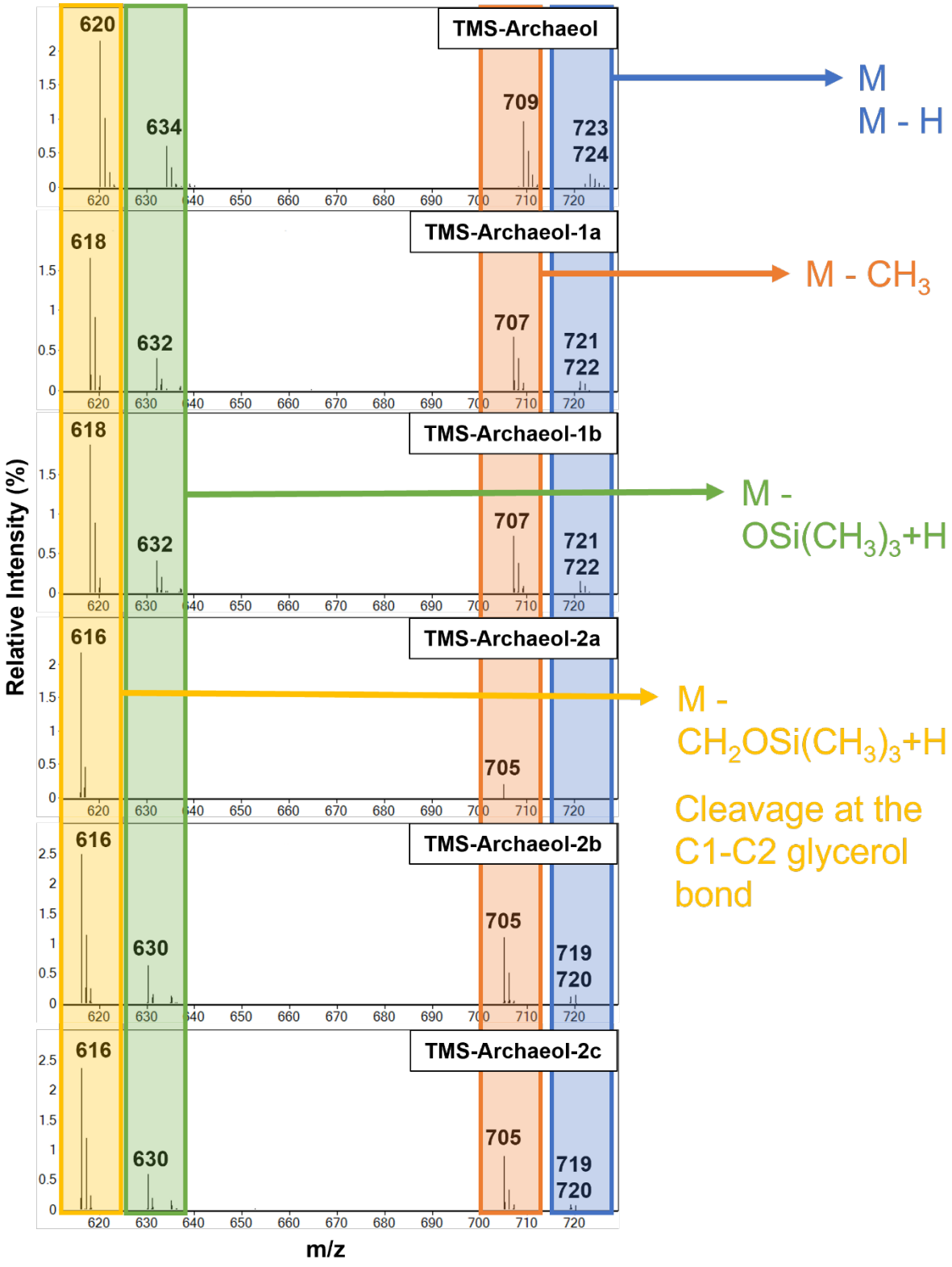

TMS-Archaeol
620
634
709
723
724
M
M - H
TMS-Archaeol-1a
618
632
707
721
722
M - CH3
TMS-Archaeol-1b
618
632
707
721
722
M - OSi(CH3)3+H
TMS-Archaeol-2a
616
705
M - CH2OSi(CH3)3+H
Cleavage at the C1-C2 glycerol bond
TMS-Archaeol-2b
616
630
705
719
720
TMS-Archaeol-2c
616
630
705
719
720
Relative Intensity (%)
m/z

Fig. A.S13. **The molecular ions and related moieties of TMS derivatized archaeol, archaeol-1, and archaeol-2.** Electron impact mass spectra of TMS derivatized archaeol, archaeol-1, and archaeol-2 between m/z = 615 – 725 reveals the molecular weight of TMS-archaeol, TMS-archaeol-1, and TMS-archaeol-2. The blue box contains the M and M-H ions. The orange box contains the M - CH_3 fragment ions. The green box contains the M - $OSi(CH_3)_3$+H fragment ions (loss of TMS and terminal oxygen). The yellow box contains the M - $OSi(CH_3)_3$+H fragment ions which results from cleavage at the C1-C2 glycerol bond. Note: the molecular ion was detected for all archaeol lipid species except for archaeol-2a which was present in much lower amounts than archaeol-2b and 2c; however, the M - CH_3 peak was observed for this isomer, indicating that it is not simply underivatized archaeol-2.

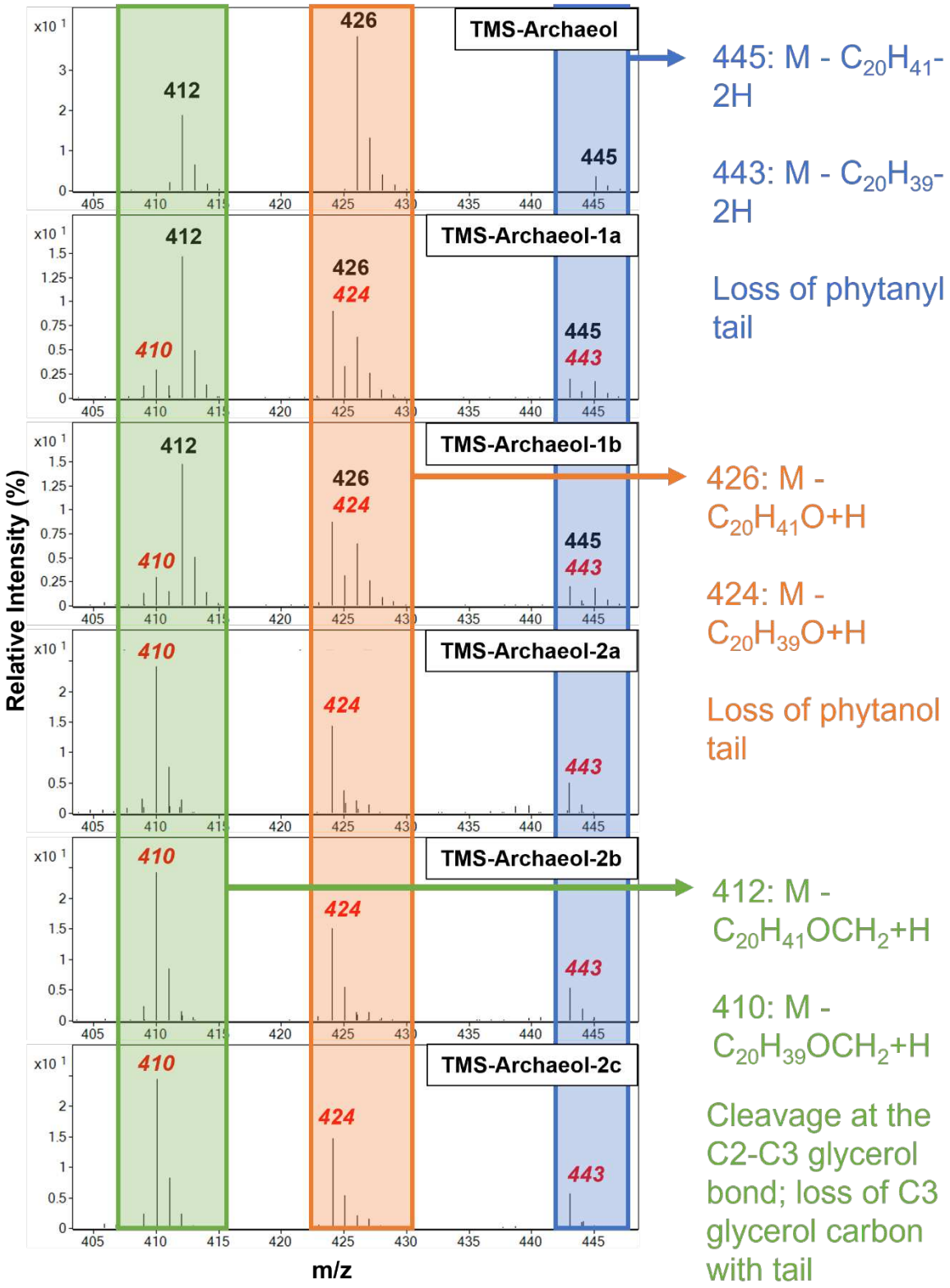
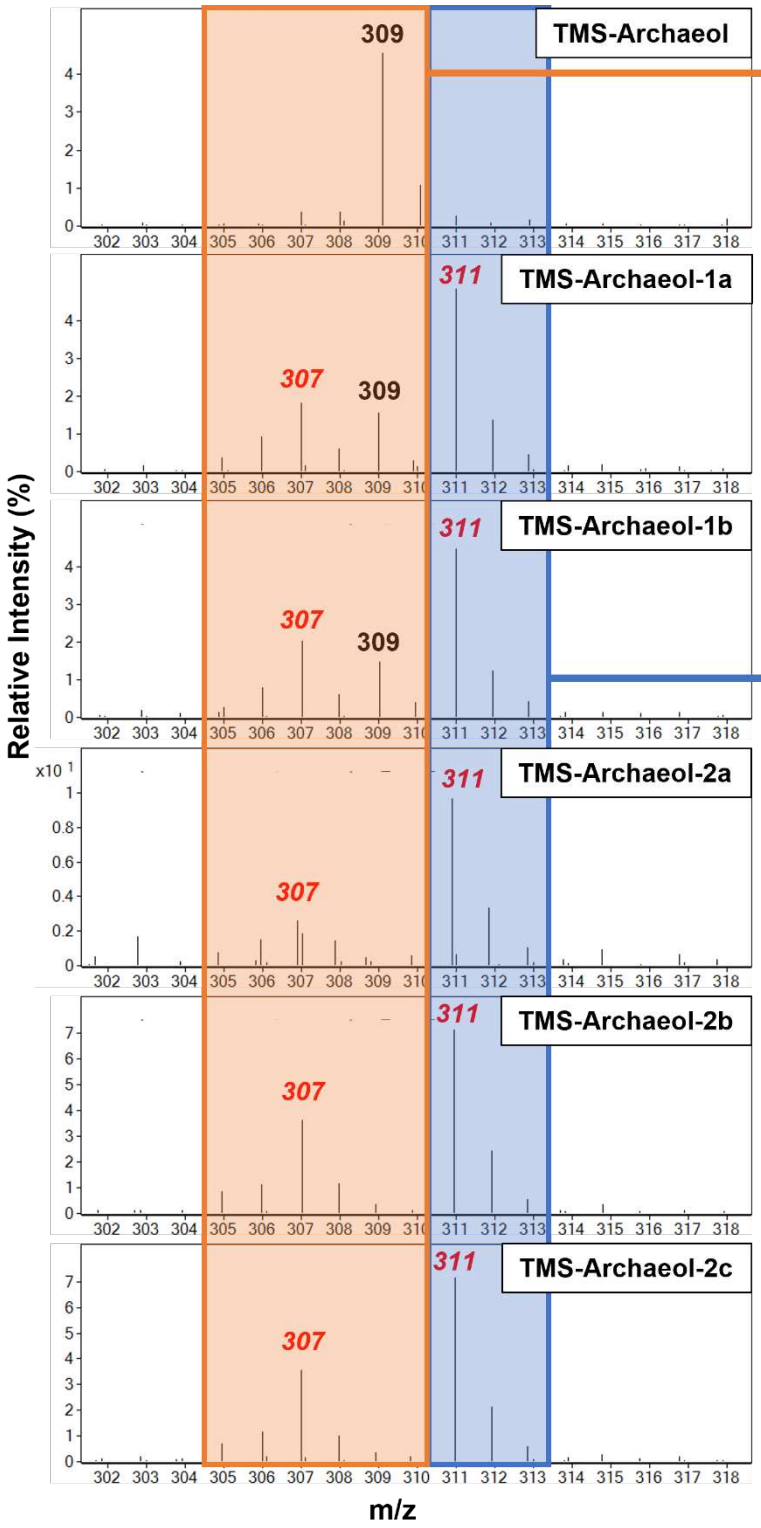

TMS-Archaeol
TMS-Archaeol-1a
TMS-Archaeol-1b
TMS-Archaeol-2a
TMS-Archaeol-2b
TMS-Archaeol-2c
412
426
445
410
424
443
Relative Intensity (%)
m/z
445: M - C20H41-2H
443: M - C20H39-2H
Loss of phytanyl tail
426: M - C20H41O+H
424: M - C20H39O+H
Loss of phytanol tail
412: M - C20H41OCH2+H
410: M - C20H39OCH2+H
Cleavage at the C2-C3 glycerol bond; loss of C3 glycerol carbon with tail

Fig. A.S14. Fragment ions arising from cleavage at the C2-C3 glycerol bond reveal archaeol-1a and archaeol-1b are both mixtures of regioisomers. Electron impact mass spectra of TMS derivatized archaeol, archaeol-1, and archaeol-2 between m/z = 400 - 450 reveal three different ways in which a tail is lost from the molecular ion. Red m/z values indicate the fragment possesses a ring. The blue box contains the M - $C_{20}H_{39and41}$-2H fragment ions resulting from the loss of a phytanyl tail. The orange box contains the M - $C_{20}H_{39and41}O$+H fragment ions resulting from the loss of a phytanol tail. The green box contains the diagnostic M - $C_{20}H_{39and41}OCH_2$+H fragment ions resulting from cleavage at the C2-C3 glycerol bond; this fragment ion loses the C3 glycerol carbon and its associated tail and thus exclusively possesses the tail bonded to the C2 glycerol carbon, providing information on the location of the ring in archaeol-1. The M - $C_{20}H_{39and41}OCH_2$+H fragments of archaeol-1a and archaeol-1b are both a mixture of ions with m/z = 410 and 412, but with a larger 412 peak, indicating that each isomer is a mixture of regioisomers with rings in the C2 and C3 glycerol carbon bonded tails but with a predominance of the ring in the C3 bonded tail.



309:
$C_{21}H_{43}O$ - 2H

307:
$C_{21}H_{41}O$ - 2H

Cleavage at the C2-C3 glycerol bond (C3-containing fragment)

$C_{21}H_{41}O$ + 2H ?

Cleavage at the C2-C3 glycerol bond (alternative C3-derived fragment for cyclized archaeol?)

Fig. A.S15. Cleavage at the C2-C3 glycerol bond appears to generate a unique fragment ion in cyclized archaeol. Electron impact mass spectra of TMS derivatized archaeol, archaeol-1, and archaeol-2 between m/z = 300 - 320. Red m/z values indicate a fragment possesses a ring. The orange box contains the “typical” C3-derived fragment ion resulting from cleavage at the C2-C3 glycerol bond; this is the C3 glycerol carbon and its associated tail which were lost from the molecular ion in the green box of the Figure S13. The blue box contains a speculative C3-derived fragment ion (m/z = 311) that is unique to cyclized archaeol. Note the m/z = 307 peak is slightly larger than the m/z = 309 peak in archaeol, which is expected if the C3 glycerol bonded tail more commonly possesses the ring; however, the difference is not as large as in the sister fragments (m/z = 410 and 412), but this difference may be accounted for if the C3 glycerol bonded tail fragment also gives rise to the m/z = 311 fragment ion.

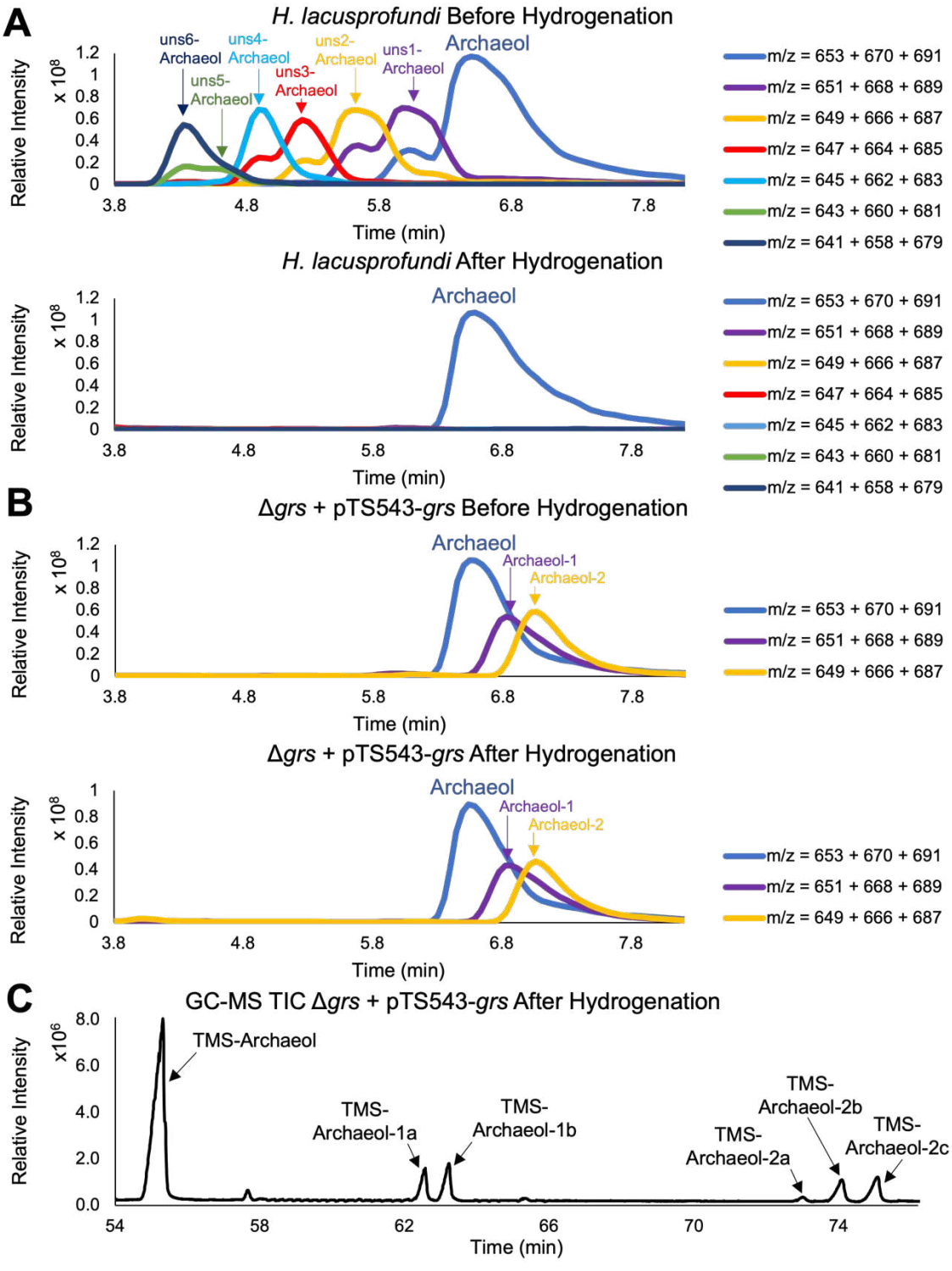

A
H. lacusprofundi Before Hydrogenation
uns6-Archaeol
uns4-Archaeol
uns2-Archaeol
uns1-Archaeol
Archaeol
uns5-Archaeol
uns3-Archaeol
Relative Intensity
x 10^8
Time (min)
m/z = 653 + 670 + 691
m/z = 651 + 668 + 689
m/z = 649 + 666 + 687
m/z = 647 + 664 + 685
m/z = 645 + 662 + 683
m/z = 643 + 660 + 681
m/z = 641 + 658 + 679
H. lacusprofundi After Hydrogenation
Archaeol
B
Δgrs + pTS543-grs Before Hydrogenation
Archaeol-1
Archaeol-2
Δgrs + pTS543-grs After Hydrogenation
C
GC-MS TIC Δgrs + pTS543-grs After Hydrogenation
x10^6
TMS-Archaeol
TMS-Archaeol-1a
TMS-Archaeol-1b
TMS-Archaeol-2a
TMS-Archaeol-2b
TMS-Archaeol-2c

Fig. A.S16. Hydrogenation of total lipid extracts containing archaeol-1 and archaeol-2 indicates they possess rings and not double bonds. (A) LC-MS overlayed extracted ion chromatograms of a base hydrolyzed lipid extract of *Halorubrum lacusprofundi* grown at 10 °C for five months, shown before and after hydrogenation. Before hydrogenation, the *H. lacusprofundi* lipid extracts possess archaeol with 0 to 6 double bonds. After hydrogenation, the peaks corresponding to the double bond containing archaeols - uns1-archaeol (1 double bond) to uns6-archaeol (6 double bonds) - are lost. For each lipid species the $[M+H]^+$, $[M+NH_4]^+$, and $[M+K]^+$ ions were extracted and summed. For example, for archaeol the $[M+H]^+$ adduct ion has a m/z = 653, the $[M+NH_4]^+$ adduct ion has a m/z = 670, and the $[M+K]^+$ adduct ion has a m/z = 691. This was done because we found that the unsaturated archaeols much more readily formed adducts with NH_4^+ than saturated archaeol did and that the unsaturated archaeols much more weakly formed adducts with H^+ so it would not be proper to compare the two with a single adduct ion. We therefore chose to sum the three most abundant adduct ions for each lipid species, which corresponded to $[M+H]^+$, $[M+NH_4]^+$, and $[M+K]^+$. (B) LC-MS overlayed extracted ion chromatograms of an acid hydrolyzed lipid extract of the Δ*grs* + pTS543-*grs* strain grown at 60 °C to late stationary phase, shown before and after hydrogenation. The archaeol-1 and archaeol-2 peaks remain after hydrogenation and are unchanged, indicating they do not possess double bonds and rather have a ring(s). (C) GC-MS total ion chromatogram of the hydrogenated trimethylsilyl (TMS) derivatized acid hydrolyzed lipid extract of the *Δgrs* + pTS543-*grs* strain grown at 60°C to late stationary phase, showing that all the isomers of archaeol-1 and -2 remain after hydrogenation, indicating none of the isomers are the result of a double bond and rather are all ring-containing.

REFERENCES

1. Teske, A., Amils, R., Ramírez, G. A. & Reysenbach, A. L. Editorial: Archaea in the Environment: Views on Archaeal Distribution, Activity, and Biogeography. *Front Microbiol* **12,** 497 (2021).

2. Merino, N., Aronson, H. S., Bojanova, D. P., Feyhl-Buska, J., Wong, M. L., Zhang, S. & Giovannelli, D. Living at the extremes: Extremophiles and the limits of life in a planetary context. *Front Microbiol* **10,** 780 (2019).

3. Koga, Y. From promiscuity to the lipid divide: On the evolution of distinct membranes in archaea and bacteria. *J Mol Evol* **78,** 234–242 (2014).

4. Zhou, A., Weber, Y., Chiu, B. K., Elling, F. J., Cobban, A. B., Pearson, A. & Leavitt, W. D. Energy flux controls tetraether lipid cyclization in Sulfolobus acidocaldarius. *Environ Microbiol* **22,** 343–353 (2020).

5. Yang, W., Chen, H., Chen, Y., Chen, A., Feng, X., Zhao, B., Zheng, F., Fang, H., Zhang, C. & Zeng, Z. Thermophilic archaeon orchestrates temporal expression of GDGT ring synthases in response to temperature and acidity stress. *Environ Microbiol* **25,** 575–587 (2023).

6. Elling, F. J., Könneke, M., Lipp, J. S., Becker, K. W., Gagen, E. J. & Hinrichs, K. U. Effects of growth phase on the membrane lipid composition of the thaumarchaeon Nitrosopumilus maritimus and their implications for archaeal lipid distributions in the marine environment. *Geochim Cosmochim Acta* **141,** 579–597 (2014).

7. Qin, W., Carlson, L. T., Armbrust, E. V., Devol, A. H., Moffett, J. W., Stahl, D. A. & Ingalls, A. E. Confounding effects of oxygen and temperature on the TEX86 signature of marine Thaumarchaeota. *Proc Natl Acad Sci U S A* **112,** 10979–10984 (2015).

8. Hurley, S. J., Elling, F. J., Könneke, M., Buchwald, C., Wankel, S. D., Santoro, A. E., Lipp, J. S., Hinrichs, K. U. & Pearson, A. Influence of ammonia oxidation rate on thaumarchaeal lipid composition and the TEX86 temperature proxy. *Proc Natl Acad Sci U S A* **113,** 7762–7767 (2016).

9. Matsumi, R., Atomi, H., Driessen, A. J. M. & van der Oost, J. Isoprenoid biosynthesis in Archaea – Biochemical and evolutionary implications. *Res Microbiol* **162,** 39–52 (2011).

10. Straub, C. T., Counts, J. A., Nguyen, D. M. N., Wu, C.-H., Zeldes, B. M., Crosby, J. R., Conway, J. M., Otten, J. K., Lipscomb, G. L., Schut, G. J., Adams, M. W. W. & Kelly, R. M. Biotechnology of extremely thermophilic archaea. *FEMS Microbiol Rev* **42,** 543–578 (2018).

11. Sinninghe Damsté, J. S., Schouten, S., Hopmans, E. C., Van Duin, A. C. T. & Geenevasen, J. A. J. Crenarchaeol. *J Lipid Res* **43,** 1641–1651 (2002).

12. Van de Vossenberg, J. L. C. M., Driessen, A. J. M. & Konings, W. N. The essence of being extremophilic: The role of the unique archaeal membrane lipids. *Extremophiles* **2,** 163–170 (1998).

13. Valentine, D. L. Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nature Reviews Microbiology 2007 5:4* **5,** 316–323 (2007).

14. Wang, J. X., Xie, W., Zhang, Y. G., Meador, T. B. & Zhang, C. L. Evaluating production of cyclopentyl tetraethers by Marine Group II Euryarchaeota in the pearl river estuary and coastal South China Sea: Potential impact on the TEX86 paleothermometer. *Front Microbiol* **8,** 2077 (2017).

15. Schouten, S., Hopmans, E. C., Schefuß, E. & Sinninghe Damsté, J. S. Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures? *Earth Planet Sci Lett* **204,** 265–274 (2002).

16. Zhang, Y. G., Pagani, M. & Wang, Z. Ring Index: A new strategy to evaluate the integrity of TEX86 paleothermometry. *Paleoceanography* **31,** 220–232 (2016).

17. Jia, Y., Agbayani, G., Chandan, V., Iqbal, U., Dudani, R., Qian, H., Jakubek, Z., Chan, K., Harrison, B., Deschatelets, L., Akache, B. & McCluskie, M. J. Evaluation of Adjuvant Activity and Bio-Distribution of Archaeosomes Prepared Using Microfluidic Technology. *Pharmaceutics 2022, Vol. 14, Page 2291* **14,** 2291 (2022).

18. Santhosh, P. B. & Genova, J. Archaeosomes: New Generation of Liposomes Based on Archaeal Lipids for Drug Delivery and Biomedical Applications. *ACS Omega* **8,** 1 (2022).

19. Lilia Romero, E. & Jose Morilla, M. Ether lipids from archaeas in nano-drug delivery and vaccination. *Int J Pharm* **634,** 122632 (2023).

20. Lloyd, C. T., Iwig, D. F., Wang, B., Cossu, M., Metcalf, W. W., Boal, A. K. & Booker, S. J. Discovery, structure and mechanism of a tetraether lipid synthase. *Nature* **609,** 197 (2022).

21. Zeng, Z., Chen, H., Yang, H., Chen, Y., Yang, W., Feng, X., Pei, H. & Welander, P. V. Identification of a protein responsible for the synthesis of archaeal membrane-spanning GDGT lipids. *Nature Communications 2022 13:1* **13,** 1–9 (2022).

22. Zeng, Z., Liu, X. L., Farley, K. R., Wei, J. H., Metcalf, W. W., Summons, R. E. & Welander, P. V. GDGT cyclization proteins identify the dominant archaeal sources of tetraether lipids in the ocean. *Proc Natl Acad Sci U S A* **116,** 22505–22511 (2019).

23. Matsuno, Y., Sugai, A., Higashibata, H., Fukuda, W., Ueda, K., Uda, I., Sato, I., Itoh, T., Imanaka, T. & Fujiwara, S. Effect of Growth Temperature and Growth Phase on the Lipid Composition of the Archaeal Membrane from Thermococcus kodakaraensis. *Biosci Biotechnol Biochem* **73,** 104–108 (2009).

24. Tourte, M., Schaeffer, P., Grossi, V. & Oger, P. M. Functionalized Membrane Domains: An Ancestral Feature of Archaea? *Front Microbiol* **11,** (2020).

25. Orita, I., Futatsuishi, R., Adachi, K., Ohira, T., Kaneko, A., Minowa, K., Suzuki, M., Tamura, T., Nakamura, S., Imanaka, T., Suzuki, T. & Fukui, T. Random mutagenesis of a hyperthermophilic archaeon identified tRNA modifications associated with cellular hyperthermotolerance. *Nucleic Acids Res* **47,** 1964 (2019).

26. Li, Z., Santangelo, T. J., Čuboňová, L., Reeve, J. N. & Kelman, Z. Affinity purification of an archaeal DNA replication protein network. *mBio* **1,** 221–231 (2010).

27. 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).

28. Čuboňová, L., Katano, M., Kanai, T., Atomi, H., Reeve, J. N. & Santangelo, T. J. An Archaeal Histone Is Required for Transformation of Thermococcus kodakarensis. *J Bacteriol* **194,** 6864 (2012).

29. Liman, G. L. S., Stettler, M. E. & Santangelo, T. J. Transformation Techniques for the Anaerobic Hyperthermophile Thermococcus kodakarensis. *Methods Mol Biol* **2522,** 87 (2022).

30. Santangelo, T. J., Čuboňová, L. & Reeve, J. N. Thermococcus kodakarensis Genetics: Tk1827-encoded β-glycosidase, new positive-selection protocol, and targeted and repetitive deletion technology. *Appl Environ Microbiol* **76,** 1044–1052 (2010).

31. Scott, K. A., Williams, S. A. & Santangelo, T. J. Thermococcus kodakarensis provides a versatile hyperthermophilic archaeal platform for protein expression. *Methods Enzymol* **659,** 243 (2021).

32. Dawson, K. S., Freeman, K. H. & Macalady, J. L. Molecular characterization of core lipids from halophilic archaea grown under different salinity conditions. *Org Geochem* **48,** 1–8 (2012).

33. Pitcher, A., Rychlik, N., Hopmans, E. C., Spieck, E., Rijpstra, W. I. C., Ossebaar, J., Schouten, S., Wagner, M. & Damsté, J. S. S. Crenarchaeol dominates the membrane lipids of Candidatus Nitrososphaera gargensis, a thermophilic Group I.1b Archaeon. *The ISME Journal 2009 4:4* **4,** 542–552 (2009).

34. Jensen, S. M., Neesgaard, V. L., Skjoldbjerg, S. L. N., Brandl, M., Ejsing, C. S. & Treusch, A. H. The effects of temperature and growth phase on the lipidomes of Sulfolobus islandicus and Sulfolobus tokodaii. *Life* **5,** (2015).

35. Uda, I., Sugai, A., Itoh, Y. H. & Itoh, T. Variation in Molecular Species of Core Lipids from the Order Thermoplasmales Strains Depends on the Growth Temperature. *J Oleo Sci* **53,** (2004).

36. Van de Vossenberg, J. L. C. M., Driessen, A. J. M. & Konings, W. N. The essence of being extremophilic: The role of the unique archaeal membrane lipids. *Extremophiles* **2,** 163–170 (1998).

37. Tierney, J. E. & Tingley, M. P. A TEX86 surface sediment database and extended Bayesian calibration. *Scientific Data 2015 2:1* **2,** 1–10 (2015).

38. Meador, T. B., Gagen, E. J., Loscar, M. E., Goldhammer, T., Yoshinaga, M. Y., Wendt, J., Thomm, M. & Hinrichs, K. U. Thermococcus kodakarensis modulates its polar membrane lipids and elemental composition according to growth stage and phosphate availability. *Front Microbiol* **5,** (2014).

39. Stadnitskaia, A., Baas, M., Ivanov, M. K., Van Weering, T. C. E. & Sinninghe Damsté, J. S. Novel archaeal macrocyclic diether core membrane lipids in a methane-derived carbonate crust from a mud volcano in the Sorokin Trough, NE Black Sea. *Archaea* **1,** 165 (2003).

40. Liu, X. L., De Santiago Torio, A., Bosak, T. & Summons, R. E. Novel archaeal tetraether lipids with a cyclohexyl ring identified in Fayetteville Green Lake, NY, and other sulfidic lacustrine settings. *Rapid Commun Mass Spectrom* **30,** 1197–1205 (2016).

41. Montenegro, E., Gabler, B., Paradies, G., Seemann, M. & Helmchen, G. Determination of the configuration of an Archaea membrane lipid containing cyclopentane rings by total synthesis. *Angewandte Chemie - International Edition* **42,** (2003).

42. Holzheimer, M., Sinninghe Damsté, J. S., Schouten, S., Havenith, R. W. A., Cunha, A. V. & Minnaard, A. J. Total Synthesis of the Alleged Structure of Crenarchaeol Enables Structure Revision**. *Angewandte Chemie - International Edition* **60,** (2021).

43. Lutnaes, B. F., Brandal, Ø., Sjöblom, J. & Krane, J. Archaeal C80 isoprenoid tetraacids responsible for naphthenate deposition in crude oil processing. *Org Biomol Chem* **4,** (2006).

44. Sinninghe Damsté, J. S., Rijpstra, W. I. C., Hopmans, E. C., den Uijl, M. J., Weijers, J. W. H. & Schouten, S. The enigmatic structure of the crenarchaeol isomer. *Org Geochem* **124,** (2018).

45. Rattray, J. E. & Smittenberg, R. H. Separation of Branched and Isoprenoid Glycerol Dialkyl Glycerol Tetraether (GDGT) Isomers in Peat Soils and Marine Sediments Using Reverse Phase Chromatography. *Front Mar Sci* **7,** (2020).

APPENDIX B: DYNAMIC RNA ACETYLATION REVEALED BY QUANTITATIVE CROSS-EVOLUTIONARY MAPPING

Summary

N4-acetylcytidine (ac4C) is an ancient and highly conserved RNA modification that is present on tRNA and rRNA and has recently been investigated in eukaryotic mRNA[1–3]. However, the distribution, dynamics and functions of cytidine acetylation have yet to be fully elucidated. Here we report ac4C-seq, a chemical genomic method for the transcriptome-wide quantitative mapping of ac4C at single-nucleotide resolution. In human and yeast mRNAs, ac4C sites are not detected but can be induced—at a conserved sequence motif—via the ectopic overexpression of eukaryotic acetyltransferase complexes. By contrast, cross-evolutionary profiling revealed unprecedented levels of ac4C across hundreds of residues in rRNA, tRNA, non-coding RNA and mRNA from hyperthermophilic archaea. Ac4C is markedly induced in response to increases in temperature, and acetyltransferase-deficient archaeal strains exhibit temperature-dependent growth defects. Visualization of wild-type and acetyltransferase-deficient archaeal ribosomes by cryo-electron microscopy provided structural insights into the temperature-dependent distribution of ac4C and its potential thermoadaptive role. Our studies quantitatively define the ac4C landscape, providing a technical and conceptual foundation for elucidating the role of this modification in biology and disease[4–6].

Main

Acetylation is an ancient mechanism that regulates biomolecular function. Perhaps the most well conserved of these mechanisms is the enzymatic modification of RNA to form the

---

[1] Most of this chapter was previously published under the following title with a few updates: Sas-Chen, A., Thomas, J. M., Matzov, D., Taoka, M., Nance, K. D., Nir, R., Bryson, K. M., Shachar, R., Liman, G. L. S., Burkhart, B. W., Gamage, S. T., Nobe, Y., Briney, C. A., Levy, M. J., Fuchs, R. T., Robb, G. B., Hartmann, J., Sharma, S., Lin, Q., Florens, L., … Schwartz, S. (2020). Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. Nature, 583(7817), 638–643. https://doi.org/10.1038/s41586-020-2418-2

acetylated nucleobase ac^{4}C. Ac^{4}C occurs in all domains of life, and its formation is catalysed by the acetyltransferases NAT10 in humans and Kre33 in yeast[1–3]. NAT10 and Kre33 are essential in humans and yeast, respectively, and the four target sites of these enzymes in rRNA and tRNA are also conserved between these two distant eukaryotes[1–3]. The deposition of ac^{4}C at its two tRNA targets (tRNA-Ser and tRNA-Leu) requires an additional adaptor protein—THUMPD1 in humans and Tan1 in yeast[1]—and has been implicated in tRNA stability[7,8]. Conversely, NAT10 is guided towards its two target sites in rRNA by specialized small nucleolar RNAs[9]. Recently, antibody-based mapping suggested the existence of additional NAT10-regulated ac^{4}C sites in human mRNAs[10]; however, the lack of base-resolution quantification of any single ac^{4}C site precluded orthogonal validation and functional prioritization on the basis of modification stoichiometries. Thus, the quantitative distribution of ac^{4}C among rRNA, tRNA and mRNA remains to be comparatively defined in any organism.

Nucleotide-resolution ac4C sequencing

To quantitatively study cytidine acetylation in the transcriptome, we developed a chemical method to enable the sensitive detection of ac^{4}C at single-nucleotide resolution. Building on previous work[11], we found that the reaction of ac^{4}C with sodium cyanoborohydride ($NaCNBH_3$) under acidic conditions forms the reduced nucleobase *N*^{4}-acetyltetrahydrocytidine. The altered structure of this reduced nucleobase compared with ac^{4}C causes the incorporation of non-cognate deoxynucleotide triphosphates (dNTPs) upon reverse transcription[11], which can be detected via cDNA sequencing. Compared with previous chemistries, this reaction shows faster kinetics and causes increased misincorporation at known ac^{4}C sites in rRNA (Extended Data Fig. B.1, Supplementary Note 1). Critically, ac^{4}C-dependent mutations are not observed when the modification is hydrolysed (chemically deacetylated) using mild alkali before analysis[12] (Fig. B.1a). Integrating these chemistries with next-generation sequencing led to the development of ac^{4}C-seq, a method that enables the transcriptome-wide, quantitative analysis of ac^{4}C at single-

nucleotide resolution (Fig. B.1b, Methods). Inspection of sequencing data revealed that $NaCNBH_3$ treatment caused C>T misincorporation at acetylated sites, which were reduced upon alkali-induced deacetylation (Fig. B.1c). This guided the development of an analytical pipeline for ac^4C detection, based on the following observations: C>T misincorporation upon treatment with acid and $NaCNBH_3$; the reduction in C>T misincorporation upon pre-treatment with alkali; and the absence of C>T misincorporation in mock-treated RNA. These three requirements were formalized as two statistical tests, comparing misincorporations in $NaCNBH_3$-treated samples with those in alkali- or mock-treated controls. In practice, excellent signal-to noise ratios could be obtained on the basis of the latter comparison, enabling the former to be used as an optional filter to increase confidence in identified sites (Fig. B.1d). To evaluate our ability to quantitatively measure acetylation levels, we applied ac^4C-seq to four synthetic RNAs, each harbouring a single ac^4C site. In these synthetic RNAs, ac^4C was embedded within several sequence contexts, and spiked into complex RNA samples at varying stoichiometries. We observed excellent absolute agreement between the synthesized ac^4C stoichiometries and the experimentally measured misincorporation levels (Pearson's $R = 0.99$) across the entire range of stoichiometries (Fig. B.1e). Thus, given sufficient read-depth, ac^4C-seq is able to detect and quantify even low-stoichiometry (4%) modifications with excellent accuracy and precision.

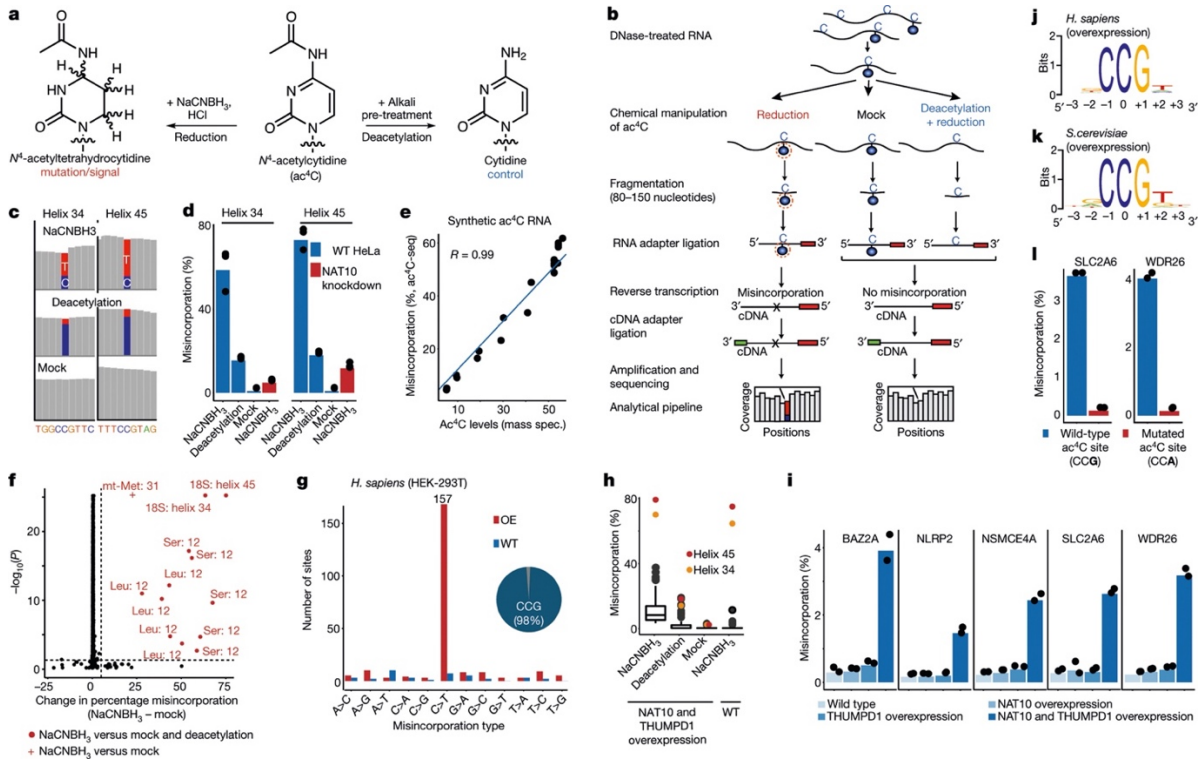

a
N^4-acetyltetrahydrocytidine
mutation/signal
+ NaCNBH3, HCl
Reduction
N^4-acetylcytidine (ac^4C)
+ Alkali pre-treatment
Deacetylation
Cytidine
control
b
DNase-treated RNA
Chemical manipulation of ac^4C
Reduction
Mock
Deacetylation + reduction
Fragmentation (80–150 nucleotides)
RNA adapter ligation
Reverse transcription
Misincorporation
No misincorporation
cDNA
cDNA adapter ligation
Amplification and sequencing
Analytical pipeline
Coverage
Positions
c
Helix 34
Helix 45
NaCNBH3
Deacetylation
Mock
TGGCCGTTC TTTCCGTAG
d
Helix 34
Helix 45
WT HeLa
NAT10 knockdown
Misincorporation (%)
NaCNBH3
Deacetylation
Mock
e
Synthetic ac^4C RNA
R = 0.99
Misincorporation (%, ac^4C-seq)
Ac^4C levels (mass spec.)
f
mt-Met: 31
18S: helix 45
18S: helix 34
Ser: 12
Leu: 12
$-\log_{10}(P)$
Change in percentage misincorporation (NaCNBH3 – mock)
NaCNBH3 versus mock and deacetylation
NaCNBH3 versus mock
g
H. sapiens (HEK-293T)
157
OE
WT
CCG (98%)
Number of sites
A>C
A>G
A>T
C>A
C>G
C>T
G>A
G>C
G>T
T>A
T>C
T>G
Misincorporation type
h
Misincorporation (%)
Helix 45
Helix 34
NaCNBH3
Deacetylation
Mock
NAT10 and THUMPD1 overexpression
WT
i
BAZ2A
NLRP2
NSMCE4A
SLC2A6
WDR26
Misincorporation (%)
Wild type
THUMPD1 overexpression
NAT10 overexpression
NAT10 and THUMPD1 overexpression
j
H. sapiens
(overexpression)
Bits
5′ −3 −2 −1 0 +1 +2 +3 3′
k
S.cerevisiae
(overexpression)
Bits
5′ −3 −2 −1 0 +1 +2 +3 3′
l
SLC2A6
WDR26
Misincorporation (%)
Wild-type ac^4C site (CCG)
Mutated ac^4C site (CCA)

Figure B.1. Development and application of ac^{4}C-seq in human and yeast. a, Reaction schemes showing the reduction and the deacetylation of ac^{4}C. **b**, Schematic showing the ac^{4}C-seq procedure: RNA is deacetylated in the pre-treatment step (or mock-pretreated), followed by treatment with $NaCNBH_3$ (or mock treatment). After library preparation as illustrated, ac^{4}C is detected by the analysis of C>T misincorporation. **c**, Misincorporation rates in total RNA from HeLa cells are shown for known sites in 18S (blue, cytidine; red, thymidine). **d**, Misincorporation rates in 18S sites in wild-type and NAT10-depleted cells (bars, mean of 3 biological samples; dots, individual measurements). **e**, Misincorporation rates of 4 synthetic spikes measured by ac^{4}C-seq (*y* axis) plotted against ac^{4}C stoichiometry as measured by mass spectrometry (*x* axis). Pearson's *R*, *n* = 1 experiment. **f**, Statistical significance plotted against the difference in misincorporation rates between $NaCNBH_3$ and mock-treated total RNA from HeLa cells. Vertical dashed line, 5%; horizontal dashed line, *P* = 0.05 (χ^2 test). *n* = 3 biological samples. **g**, Frequency of the 12 possible misincorporation patterns (*y* axis) for sites found in poly(A)-enriched RNA from wild-type (WT) HEK-293T cells and from HEK-293T cells overexpressing NAT10 and THUMPD1 (OE). The pie chart shows the proportion of sites harbouring C>T misincorporations within a C**C**G motif. **h**, Misincorporation rate at ac^{4}C sites within CCG motifs identified in **g** in wild-type cells and in cells overexpressing NAT10 and THUMPD1, shown for RNA treated with $NaCNBH_3$ and indicated controls (*n* = 2 biological samples for overexpression $NaCNBH_3$- or mock-treated and 1 sample for the rest). For the box plots, the centre line indicates the median, the box boundaries mark the 25th and 75th percentiles, the whiskers indicate ±1.5× the interquartile range (IQR) and outliers are shown as individual dots. **i**, Misincorporation level (obtained from ac^{4}C-seq) at amplicons spanning ac^{4}C sites in HEK-293T cells, depicted as in **d**. *n* = 2 biological samples. **j**, **k**, Sequence motif surrounding the ac^{4}C sites identified in indicated organisms. **l**, Misincorporation rate at two wild-type and mutated ac^{4}C sites in HEK-293T cells overexpressing NAT10/THUMPD1, quantified via targeted ac^{4}C-sequencing, depicted as in **d**. *n* = 2 biological samples.

Ac^{4}C in eukaryotic RNA

We next explored the properties of ac^{4}C in eukaryotic RNA. To strengthen this study, we used a cross-evolutionary approach, analysing two human cell lines and the budding yeast, *Saccharomyces cerevisiae*. Applying ac^{4}C-seq to total RNA from these organisms recapitulated both known sites of ac^{4}C modification in 18S rRNA, as well as the two known sites of ac^{4}C on tRNA: tRNA-Ser and tRNA-Leu[1–3] (Fig. B.1f, Extended Data Fig. B.2a–c). No additional rRNA or tRNA sites met detection thresholds for ac^{4}C. Acetylation of rRNA and tRNA sites were reduced after the disruption of human NAT10, and eliminated after the mutation of yeast Kre33 (Fig. B.1d, Extended Data Fig. B.2d). These results suggest that eukaryotic rRNA

and tRNA ac^4C is well annotated, and that in these abundant RNAs, ac^4C-seq demonstrates very good sensitivity and specificity.

Next we explored the properties of ac^4C in eukaryotic mRNA[10,13]. Applying ac^4C-seq to poly(A)-enriched mRNA from HEK-293T cells readily identified the known sites on rRNA (Extended Data Fig. B.2e). However, only four additional C>T misincorporations passed detection thresholds (Fig. B.1g)—a number consistent with the anticipated false discovery rate (Fig. B.1g). To address the possibility that the absence of detectable ac^4C in mRNA is unique to HEK-293T cells, we applied ac^4C-seq to poly(A)-RNA isolated from HeLa cells and from *S. cerevisiae*, in which ac^4C has been previously suggested to be present using other approaches[1,10,13]. In both models we detected the known rRNA ac^4C sites (Extended Data Fig. B.2a–c). However, no additional sites passed detection thresholds in HeLa cells, and in yeast, three additional sites were identified in mRNA, but they were not eliminated after the mutation of yeast Kre33, and no enrichment was observed for C>T misincorporations—suggesting that they do not represent ac^4C sites. Although these observations do not rule out the existence of rare or low-stoichiometry acetylation sites (Extended Data Fig. B.2f–h), we find no confirmatory evidence for the presence of ac^4C in eukaryotic mRNA.

To understand the potential for eukaryotic cytidine acetyltransferases to modify mRNA, we co-overexpressed NAT10 and THUMPD1 in HEK-293T cells, and their orthologues Kre33 and Tan1 in yeast (Extended Data Fig. B.3a–d). Notably, overexpression of these complexes led to the identification of 146 and 66 putative novel ac^4C sites in human and yeast mRNA, respectively (Fig. B.1g, Extended Data Fig. B.3e). Misincorporation levels within mRNA remained modest (median 7.7% and 4.9% in human and yeast, respectively) even when NAT10 and THUMPD1 were co-overexpressed at very high levels (Fig. B.1h, Extended Data Fig. B.3a, d)). Targeted deep sequencing of five of these sites (median 120,000 reads per site) recapitulated acetylation upon dual overexpression of NAT10 and THUMPD1 (approximately 3–

4% misincorporation), whereas misincorporation rates in RNA from cells in which only one protein was overexpressed were on the order of 0.2%, identical to the wild type (Fig. B.1i, Extended Data Fig. B.3e, Supplementary Note 2). To characterize substrates of the NAT10–THUMPD1 complex and explore factors that direct its specificity, we performed additional analysis of induced eukaryotic ac^{4}C sites. We found that 154 out of 157 (98%) sites in human mRNA and 73 out of 74 (98.6%) sites in yeast mRNA occurred at a CCG motif, with the central cytidine being acetylated (Fig. B.1g,j,k, Extended Data Fig. B.3e). It is noteworthy that all four ac^{4}C sites that were previously identified in eukaryotic rRNA and tRNA occur within precisely this motif (Extended Data Fig. B.3f). Induced ac^{4}C sites were randomly distributed across genes and displayed no preference for a particular position in a codon (Extended Data Fig. B.3g,h)). The obligate nature of the CCG motif was validated by plasmid-based reconstitution of an inducible ac^{4}C site, the acetylation of which—dependent on NAT10–THUMPD1—was abolished by mutation of the guanosine immediately downstream of the acetylated site (Fig. B.1l). Systematic mutagenesis experiments further indicate that base-paired structural elements may have a role in ac^{4}C deposition; this suggests that 'CCG' is required, but is not sufficient, for induced acetylation (Extended Data Fig. B.4). Overall, our studies define rRNA and tRNA as the predominant sites of ac^{4}C in eukaryotes, suggest that ac^{4}C is absent or present at very low levels in endogenous eukaryotic mRNA, and demonstrate that RNA acetylation can be induced at hundreds of sites via dual overexpression of NAT10–THUMPD1, invariably within a C**C**G motif.

Unprecedented ac^{4}C levels in archaeal RNA

A cross-evolutionary analysis of total RNA by liquid chromatography coupled to mass spectrometry (LC–MS) revealed high concentrations of ac^{4}C in the archaeal hyperthermophile *Thermococcus kodakarensis*[14] (Fig. B.2a). Motivated by this, we applied ac^{4}C-seq to quantitatively map cytidine acetylation in *T. kodakarensis* cultured at its optimal growth

temperature of 85 °C. We found an unprecedented number (404) of ac^4C sites spread across rRNA, tRNA, non-coding (nc)RNAs and mRNA (Extended Data Fig. B.5a). Of these sites, 99% occurred within CCG motifs and were highly enriched for C>T misincorporation signatures (Fig. B.2b). To validate these identifications, we performed quantitative tandem LC–MS analysis of purified and partially digested *T. kodakarensis* rRNA[15]. This revealed 25 uniquely mapped ac^4C sites, which fully overlapped with positions identified using ac^4C-seq (Fig. B.2c, Supplementary Data B.1). Estimates of modification stoichiometry based on LC–MS analysis agreed very well (Pearson's $R = 0.97$) with those from ac^4C-seq (Fig. B.2c). Deletion of the *NAT10* homologue *TK0754* (hereafter 'Tk*NAT10*'; recently reported to acetylate *T. kodakarensis* tRNA[16]), but not of the *THUMPD1* homologue *TK2097* ('Tk*THUMPD1*'), caused complete loss of ac^4C in all RNA substrates (Fig. B.2d); this result was confirmed by ac^4C-specific northern blotting and mass spectrometric analysis (Extended Data Fig. B.5b–g). To understand whether pervasive RNA acetylation is a common feature of archaeal extremophiles, we used ac^4C-seq to profile *Pyrococcus furiosus* and *Thermococcus* sp. AM4—close euryarchaeal relatives of *T. kodakarensis* within the order Thermococcales—and the more phylogenetically distant species *Methanocaldococcus jannaschii* (a Methanococcale from the Euryarchaeota phylum) and *Saccharolobus solfataricus* (a Sulfolobale from the Crenarchaeota phylum), for evolutionary breadth. This revealed that ac^4C is widespread within each of the Thermococcales species, occurring at hundreds of sites across diverse RNA types (Fig. B.2e), almost exclusively within CCG consensus motifs. In *T.* sp. AM4 and *P. furiosus*, ac^4C was not only widely present, but the precise sites and stoichiometry of ac^4C were also highly conserved (Fig. B.2f,g, Extended Data Fig. B.5h). By contrast, ac^4C detected in *S. solfataricus* was confined to 41 CCG sites mostly in tRNAs (Fig. B.2e), whereas *M. jannaschii* lacked ac^4C entirely, consistent with the absence of an apparent *NAT10* homologue in this organism[17]. These studies establish the existence and regulation of prevalent RNA acetylation in the archaeal order Thermococcales.

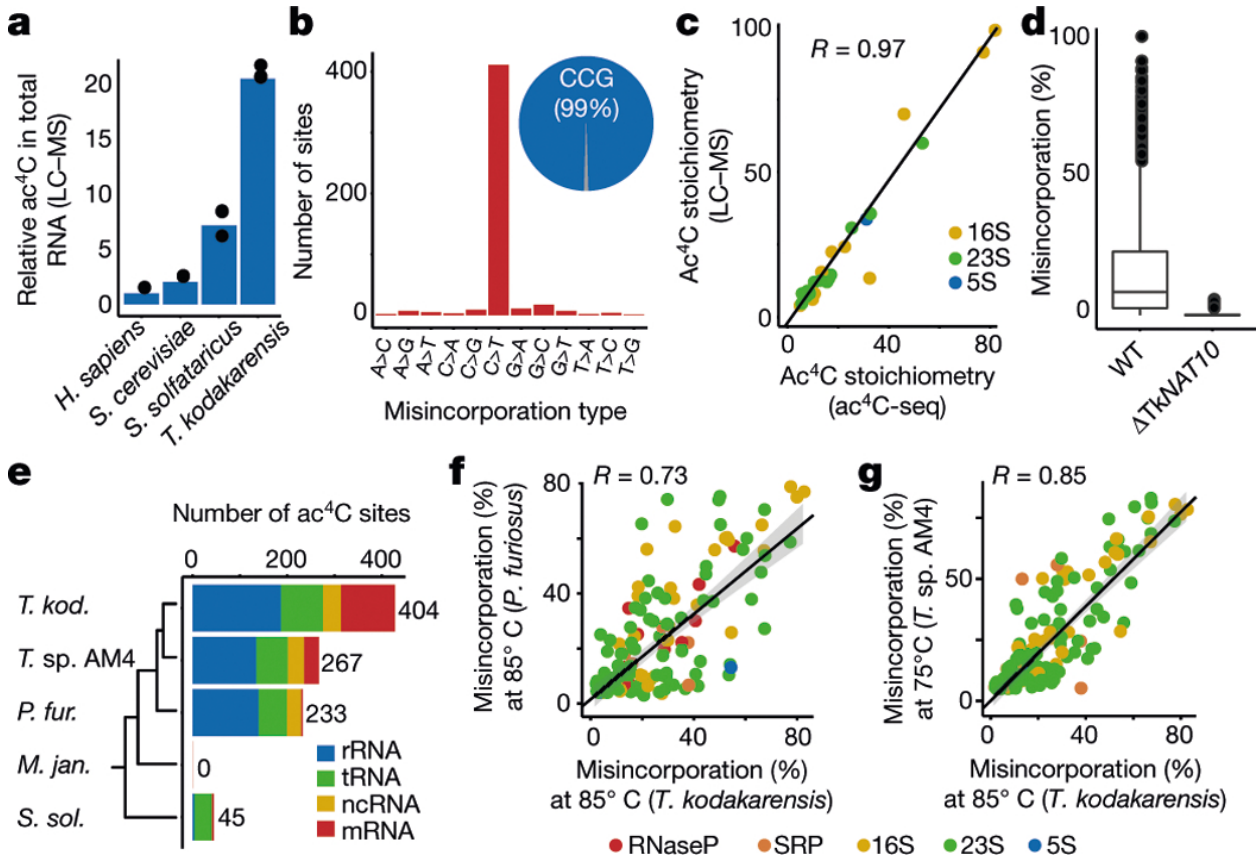


Figure B.2. Ac^{4}C is present at unprecedented levels across diverse RNA species in archaea. a, Relative quantification of ac^{4}C in total RNA isolated from *H. sapiens*, *S. cerevisiae*, *S. solfataricus* and *T. kodakarensis*. Mean of *n* = 3 technical replicates. *H. sapiens* total RNA was isolated from HeLa cells. **b**, Distribution of misincorporations (as in Fig. 1g) across all identified sites in *T. kodakarensis*. Of the C>T misincorporation sites, 99% are embedded within a C**C**G motif. **c**, Correlation (Pearson's *R*) between ac^{4}C levels as measured by ac^{4}C-seq and those measured by LC–MS, shown for 25 sites that were quantified by both methodologies. *n* = 2 and 1 independent samples for LC–MS and ac^{4}C-seq experiments, respectively. **d**, Ac^{4}C-seq quantification of sites identified in wild-type and ΔTk*NAT10* strains. Box plot parameters are as in Fig. 1h. *n* = 4 and 2 independent biological samples for wild-type and ΔTk *NAT10*, respectively. **e**, The number of identified ac^{4}C sites in the different RNA types as found in total RNA of different archaeal species. Note that for *T. kodakarensis*—but not for the others—ac^{4}C-seq was applied also to rRNA-depleted RNA. Non-coding RNAs (ncRNAs) reflect sites in RNaseP RNA, signal-recognition-particle (SRP) RNA and small nucleolar RNA (snRNA), the latter being present only in *P. furiosus*. The phylogenetic tree represents evolutionary distance between the species. **f**, **g**, Correlation between misincorporation levels in ncRNA of *T. kodakarensis* and *P. furiosus* (**f**) and *T.* sp. AM4 (**g**), identified by ac^{4}C-seq. Pearson's *R*, *n* = 4 and 1 independent biological samples for *T. kodakarensis* and other archaea, respectively. Shading indicates 95% confidence interval for predictions from a linear model.

Dynamic acetylation of archaeal RNA

To investigate how ac^4C responds to environmental cues, we applied ac^4C-seq to RNA from *T. kodakarensis* cultures grown at 55–95 °C, spanning the range of temperatures at which this organism can be cultivated. These experiments revealed that ac^4C across all classes of RNA increases markedly with temperature (Fig. B.3a), a finding that was validated by northern blotting and LC–MS analysis (Fig. B.3b, Extended Data Fig. B.6a). Proteomic analysis of the subsequent gene products indicates that the expression of *Tk*NAT10 is increased at high temperatures (Extended Data Fig. B.6b,c), which is consistent with increased ac^4C. These temperature-dependent patterns of ac^4C modification in rRNA, tRNA, ncRNA and mRNA are described in further detail in Fig. B.3c, Extended Data Fig. B.6d–h and Supplementary Note 3. Notably, the Tk*NAT10*-knockout strain (denoted ΔTk*NAT10*) showed a temperature-dependent growth lag in comparison to the wild-type strain, beginning at 75 °C and reaching a maximum at 95 °C (Fig. B.3d). The reduced fitness of ΔTk*NAT10* strains at higher temperatures parallels the increased prevalence of ac^4C in wild-type strains under these conditions, suggesting that ac^4C is required in particular for growth at high temperatures. If cytidine acetylation is a response to thermal stress, we might expect closely related organisms to also use this mechanism. Indeed, induced acetylations at higher temperatures were also conserved in *P. furiosus* and *T*. sp. AM4—two species closely related to *T. kodakarensis* (Fig. B.3e, Extended Data Fig. B.6i). Moreover, the precise sites and stoichiometries at which ac^4C was induced were also highly conserved in these organisms (Extended Data Fig. B.6j). These studies suggest that temperature-dependent cytidine acetylation is a unique adaptive survival strategy and is used by the archaeal order Thermococcales.

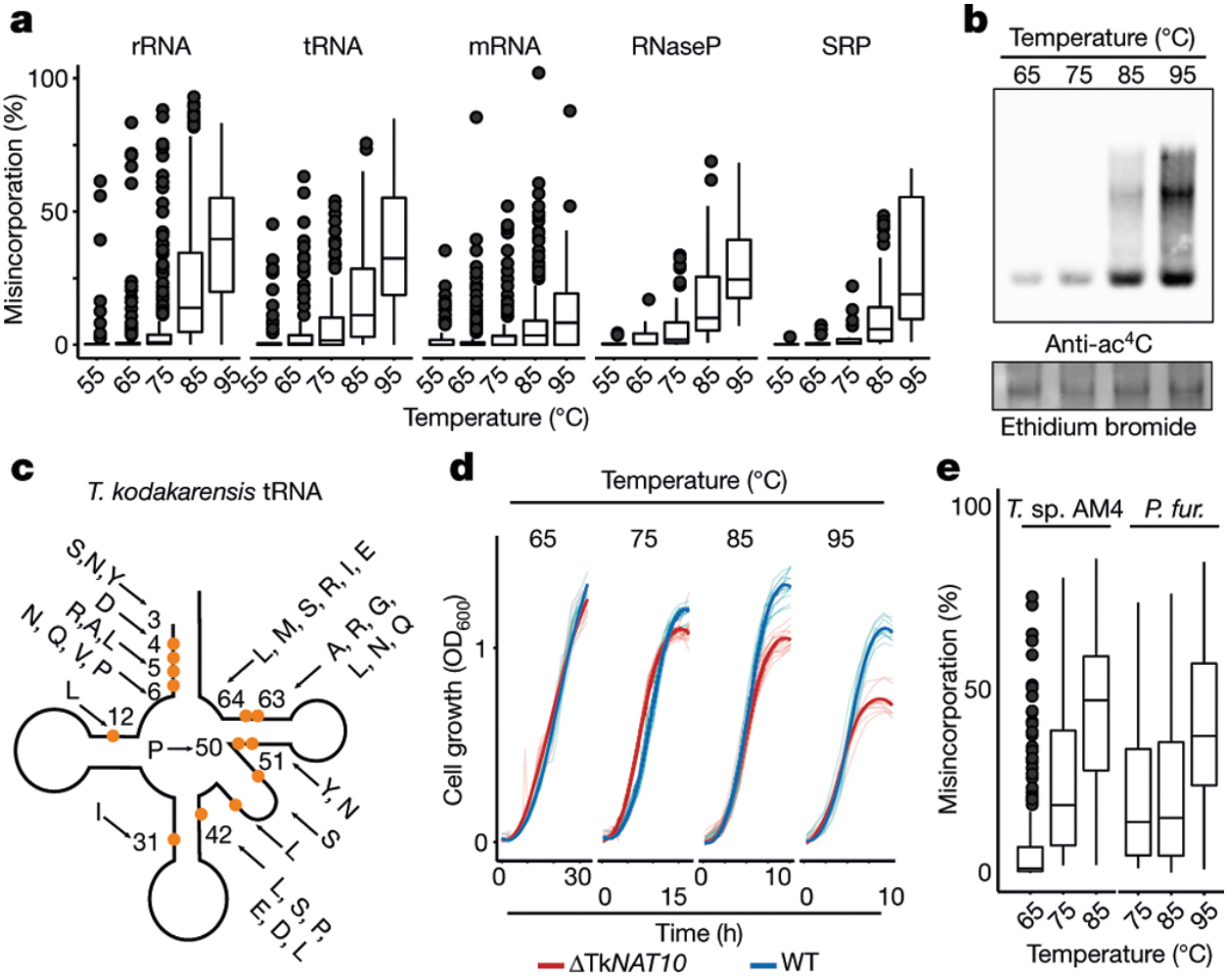


Figure B.3. Ac^{4}C accumulates in a temperature-dependent manner across all RNA species in archaea and is required for growth at higher temperatures. a, Distributions of misincorporation level at ac^{4}C sites across temperatures ranging from 55 to 95 °C. Box plot parameters are as in Fig. 1h. *n* = 4 biologically independent samples for 85 °C, *n* = 2 for 65 °C and 75 °C and *n* = 1 for 55 and 95 °C. **b**, Immuno-northern blot for the analysis of ac^{4}C in *T. kodakarensis* total RNA as a function of temperature. Ethidium bromide staining is used to visualize total RNA. Results are representative of two biological replicates. For gel source data, see Supplementary Data 3. **c**, Schematic representation of a tRNA molecule. A total of 77 ac^{4}C sites found within 19 tRNA species (indicated by the one-letter code of the relevant amino acid) were distributed across 13 distinct positions within the tRNA molecule. Each modified position is indicated by an orange circle. Numbers indicate position within the tRNA. Note that positions in the variable region are not numbered. **d**, Wild-type *T. kodakarensis* and ΔTk*NAT10* cells were grown at diverse temperatures (65–95 °C), and the optical density at 600 nm (OD_{600}) was measured hourly. The average curve of each replicate is shown by the thick line (*n* = 11 for 95 °C and *n* = 12 for each of 65–85 °C), and individual replicates are shown by thin lines. **e**, Quantification by ac^{4}C-seq of total RNA collected from cells grown at a range of temperatures. Shown are misincorporation levels for ac^{4}C sites identified in *P. furiosus* and *T*. sp. AM4. Box plot visualization parameters are as in Fig. 1h. *n* = 1 biological sample per condition.

Profiling ac^{4}C in an archaeal ribosome

The dynamics of ac^{4}C on the *T. kodakarensis* ribosome are to our knowledge unprecedented, with both the number of sites and their stoichiometry of modification increasing substantially with temperature (Fig. B.3a). In comparison, characterized eukaryotic ribosomes have at most two ac^{4}C sites[18] whereas their bacterial counterparts have none[18–20]. To visualize the distribution of ac^{4}C in *T. kodakarensis* rRNA, we obtained cryo-electron microscopy (cryo-EM) structures of ribosomes derived from wild-type and ΔTk*NAT10* strains with nominal resolutions of 2.95 Å and 2.65 Å, respectively (Extended Data Figs. B.7 & B.8). This resolution enabled full delineation of the architecture of the *T. kodakarensis* 70S ribosome—including assignment of the three RNA constituents, associated core proteins, and visualization of modified nucleotides (Fig. B4a, b, Extended Data Fig. B.8b). Comparing the structures of ribosomes from the wild-type and ΔTk*NAT10* strains, we found that the density associated with ac^{4}C was exclusively observed in wild-type ribosomes (Fig. B.4b, c, Extended Data Fig. B.8b). Cryo-EM maps directly supported the presence of 69 ac^{4}C sites in the *T. kodakarensis* ribosome grown at 85 °C (Fig. B.4a). The ability to visualize these residues using cryo-EM was consistent with the high stoichiometry estimated at these sites on the basis of the ac^{4}C-seq measurements (Extended Data Fig. B.9a). The cryo-EM analysis enhanced the information available from ac^{4}C-seq by also identifying six locations for the doubly modified nucleoside ac^{4}Cm (Extended Data Fig. B.9b–e), which is both acetylated at N4 and methylated at the 2′O sugar and has been previously suggested to have a role in thermostability[21,22]. To explore the dynamics of ac^{4}C using cryo-EM, we also determined the structure of ribosomes derived from wild-type *T. kodakarensis* grown at 65 °C (2.55 Å resolution) (Extended Data Figs. B.7,8). Consistent with the results of ac^{4}C-seq, the strain grown at 65 °C exhibited substantially lower ac^{4}C levels than that grown at 85 °C, with only five cytidine residues showing a clear density for acetylation (Extended Data Fig. B.8b).

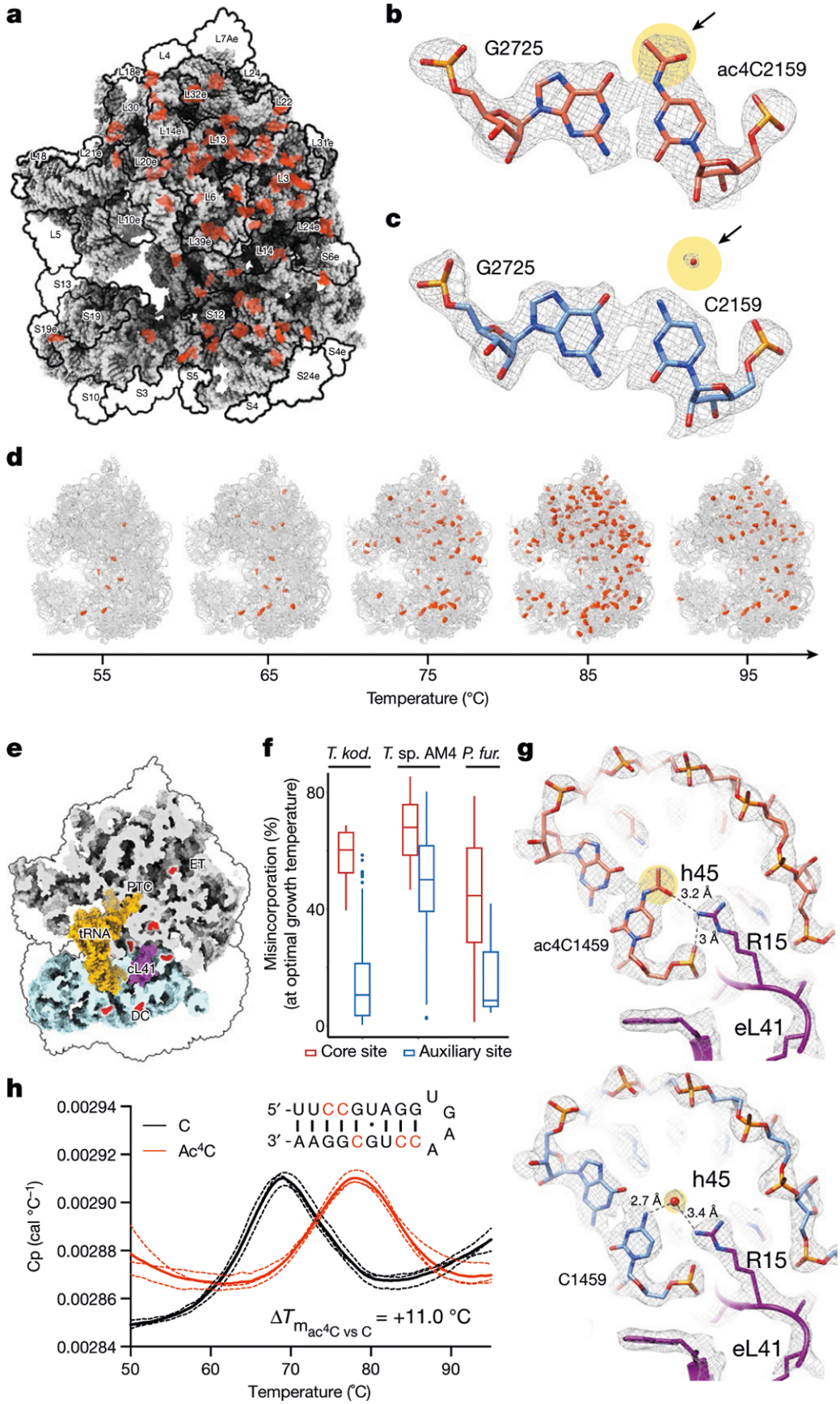

a
L7Ae
L4
L18e
L24
L32e
L30
L22
L14e
L13
L31e
L18
L21e
L20e
L3
L6
L10e
L5
L24e
L39e
L14
S6e
S13
S19
S12
S19e
S4e
S5
S24e
S10
S3
S4
b
G2725
ac4C2159
c
G2725
C2159
d
55
65
75
85
95
Temperature (°C)
e
ET
PTC
tRNA
cL41
DC
f
T. kod.
T. sp. AM4
P. fur.
Misincorporation (%)
(at optimal growth temperature)
80
40
0
Core site
Auxiliary site
g
h45
3.2 Å
ac4C1459
3 Å
R15
eL41
h45
2.7 Å
3.4 Å
C1459
R15
eL41
h
0.00294
0.00292
0.00290
0.00288
0.00286
0.00284
C
Ac4C
5′ -UUCCGUAGG U G
3′ -AAGGCGUCC A A
Cp (cal °C⁻¹)
ΔTmac4C vs C = +11.0 °C
50
60
70
80
90
Temperature (°C)

Figure B.4. Cryo-EM structure of wild-type and ac^{4}C-deficient *T. kodakarensis* ribosomes. a, Ac^{4}C distribution as observed by a cryo-EM image of wild-type *T. kodakarensis* grown at 85 °C. Modified residues are highlighted in orange, rRNA in grey and r-proteins are contoured in black. **b**, **c**, Ac^{4}Cs participate in Watson–Crick pairing with guanine residues. **b**, An example of ac^{4}C density shown in mesh. Residues correspond to ac^{4}C2159 and G2725 of LSU. Acetate is highlighted yellow and is indicated by an arrow. **c**, The same position in the ΔTk*NAT10* strain indicates that, in the mutant, the acetyl moiety is replaced by a structured solvent molecule. **d**, Ac^{4}C in *T. kodakarensis* ribosomes derived from archaea grown at different temperatures, identified by ac^{4}C-seq and LC–MS. **e**, 'Core' ac^{4}Cs (shown in red) present at high stoichiometries across temperatures are enriched in the intersubunit interface and are in proximity to eL41 and to the ribosomal substrates. The functional ribosome regions indicated are the decoding centre (DC), the peptidyl-transferase centre (PTC) and the protein exit tunnel (ET). tRNA and mRNA are highlighted yellow, eL41 is shown in purple. The tRNA and mRNA coordinates are from PDB structure 4V5D. **f**, Misincorporation at core and auxiliary sites from *T. kodakarensis* and their conserved counterparts in *P. furiosus* and *T.* sp. AM4, grown at optimal growth temperatures (85 °C for *T. kodakarensis* and *T.* sp. AM4 and 95 °C for *P. furiosus*). n = 4 and 1 independent biological samples for *T. kodakarensis* and other archaea, respectively. Box plot visualization parameters are as in Fig. 1h. **g**, A representative example of the electrostatic interaction between ac^{4}C and ribosomal proteins is shown between O(7) of ac^{4}C1459 at h45 of small-subunit (SSU) and R15 of eL41 (top). The same position in the ΔTk*NAT10* strain (bottom) implicates a solvent molecule that serves to mediate the same interaction network in the absence of an acetyl group. **h**, Thermal melting curves of synthetic RNA hairpin containing C (black) or ac^{4}C (red) obtained by differential scanning calorimetry (DSC). Cp, heat capacity; T_m, melting temperature. Data are mean ± s.d. of n = 3 independent experiments.

A notable feature of ac^{4}C in the *T. kodakarensis* ribosome is that acetylation seems to be spread across core and surface residues in both subunits (Fig. B.4a). This contrasts starkly with rRNA base modifications in eukaryotes and bacteria, which are enriched at functional regions near the ribosome core (Extended Data Fig. B.9f). Nonetheless, inspection of modification level as a function of temperature revealed a clear pattern of ac^{4}C in archaeal rRNA (Fig. B.4d). The seven ac^{4}C residues detected at low temperatures (herein termed 'core' sites) were found to concentrate at the interface between the two ribosomal subunits, making direct interactions with the ribosomal substrates (Fig. B.4e). Six of these sites envelop an inter-subunit bridge comprising the large-subunit (LSU) ribosomal protein eL41, whereas an additional site is localized at the ribosome exit tunnel (Fig. B.4e). Of note, the eukaryotic homologue of eL41 (RPL41) also localizes in an environment that is enriched in modified nucleosides[18]. Core sites were acetylated at very high levels across all temperatures (median of 77% misincorporation at

85 °C; Extended Data Fig. B.9g, h), and were also modified at high levels in *T*. sp. AM4 and *P. furiosus* (Fig. B.4f), emphasizing a potential role in ribosome function. By contrast, ac^4C sites detected only at higher temperatures were modified at lower levels (median 18% at 85 °C) and distributed widely across the ribosome, suggestive of a non-catalytic 'auxiliary' role (Extended Data Fig. B.9g, h). Considering physical mechanisms that are affected by ac^4C, we noted that in the vast majority of sites visualized by cryo-EM (64/70, 91%), the N4-acetyl group present in wild-type ribosomes is replaced by an ordered solvent molecule in the deletion strain ΔTk*NAT10* (Fig. B.4b, c, Extended Data Fig. B.8b). Similar replacement was observed in unmodified positions from the strain grown at 65 °C (Extended Data Fig. B.8b). Ordered solvent molecules are often visualized in near-atomic-resolution structures and can contribute to the structural integrity of protein and RNA architecture; it is tempting to speculate that ac^4C may have evolved as a covalent installation to replace tightly bound solvent molecules that might otherwise undergo displacement at high temperatures. Concomitantly, we identified a small subset of positions in which cytidine acetylation created the potential for unique RNA–protein interactions. Representative examples include the interaction of O(7) of ac^4C1459—a core site located in helix 45 of the *T. kodakarensis* small subunit—with Arg15 of eL41 (Fig. B.4g) and ac^4C1434 of LSU with OP2 of A1786 (Extended Data Fig. B.9i). In these examples, the ordered solvent molecule bridges the interactions that are otherwise mediated by the acetyl group (Fig. B.4g). Examining the potential influence of ac^4C on RNA–RNA interactions, we found that the vast majority (68 out of 69; 99%) of modified residues lie in duplexed rRNA and engage in canonical C–G base pairing. Consistent with the potential for acetylation to strengthen these interactions, biophysical analyses of a synthetic ribosomal RNA hairpin found that its thermal stability is enhanced by the replacement of cytidine with ac^4C[23,24] (Fig. B.4h, Extended Data Fig. B.9j). Overall, our structural survey highlights several ways by which dynamic cytidine acetylation at higher temperatures may alter the catalytic properties and physical robustness of the archaeal ribosome.

Conclusion

Here we describe ac^{4}C-seq, a method for the quantitative, nucleotide-resolution profiling of RNA cytidine acetylation. This method leverages acid-catalysed reactivity enhancement to achieve an efficient chemical reduction of ac^{4}C, which was integrated with next-generation sequencing to enable transcriptome-wide detection of ac^{4}C in diverse organisms and RNA species. Applied to eukaryotes, our studies define rRNA and tRNA as the major physiological repositories of ac^{4}C, and suggest that cytidine acetylation is absent or is present at very low levels in endogenous eukaryotic mRNA. This diverges substantially from the findings of previous experiments using antibody-based enrichment[10]. It remains to be established whether this discrepancy originates from technical differences in the methods (Supplementary Note 2a) or as a result of artefacts caused by antibody promiscuity, the latter of which has substantial precedent in the field[25–28] (Supplementary Note 2b).

The application of ac^{4}C-seq in archaea revealed pervasive programs of RNA acetylation. In the context of rRNA base modifications, ac^{4}C in Thermococcales is unprecedented in its prevalence and responsiveness to environmental cues. The dynamic and widespread distribution of ac^{4}C in the *T. kodakarensis* ribosome challenges our orthodox view of rRNA modifications, in which target sites of rRNA-modifying enzymes are classically conceptualized as being deterministic—that is, each RNA-modifying enzyme catalyses the modification of one or more highly specific sites. The high number and partial modification of 'auxiliary' sites in the *T. kodakarensis* ribosome instead raises the possibility that ac^{4}C catalysis at these positions may be statistical—that is, each site harbours a predefined probability of being targeted by the acetyltransferase, and contributes in an additive manner to overall rRNA function. It remains to be addressed whether such deposition is primarily required for the function of mature ribosomes or to facilitate rRNA folding and processing under increased temperatures. Our results further suggest that such 'statistical' deposition of ac^{4}C is not limited to rRNA, but is also widespread in

other highly structured RNAs. Collectively, our studies define the ac^4C landscape across archaeal and eukaryotic lineages, providing a technical and conceptual foundation for elucidating the role of this modification in biology and disease[4–6].

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and no allocation to groups was made in this study. Results obtained by ac^4C-seq and LC–MS were conducted in different laboratories and were compared only after the data were analysed, making them blind to each other.

Human cell culture

Wild-type HeLa (ATCC) and NAT10-depleted HeLa cells[10] were maintained in Dulbecco's Modified Eagle's Medium (DMEM, Quality Biological, 112–013-101) supplemented with 10% fetal bovine serum (FBS, VWR, 89510–194), 25 mM d-glucose, 2 mM l-glutamine, and 1 mM sodium pyruvate. HEK-293T cells (ATCC) were maintained in Dulbecco's Modified Eagle's Medium (DMEM, Quality Biological, 112–013-101) supplemented with 10% fetal bovine serum (FBS), 25 mM d-glucose, and 2 mM l-glutamine. All cells were maintained at 37 °C in the presence of 5% CO_2, and all cell culture reagents were purchased from Invitrogen unless otherwise noted. Cells were found to be free of mycoplasma contamination and did not undergo authentication.

Microbial growth and media conditions

T. kodakarensis strains—TS559 and their derivatives thereof—were grown as previously described[29–31] in artificial seawater (ASW) medium supplemented with vitamins and trace minerals. ASW contains (per litre) 20 g NaCl, 3 g $MgCl_2{\cdot}6H_2O$, 6 g $MgSO_4{\cdot}7H_2O$, 1 g $(NH_4)_2SO_4$,

200 mg $NaHCO_3$, 300 mg $CaCl_2 \cdot 2H_2O$, 0.5 g KCl, 420 mg KH_2PO_4, 50 mg NaBr, 20 mg $SrCl_2 \cdot 6H_2O$ and 10 mg $Fe(NH_4)_2(SO_4)2 \cdot 6H_2O$. The trace mineral solution (1,000× per litre) contains 0.5 g $MnSO_4 \cdot H_2O$, 0.1 g $CoCl_2 \cdot 6H_2O$, 0.1 g $ZnSO_4 \cdot 7H_2O$, 0.01 g $CuSO_4 \cdot 5H_2O$, 0.01 g $AlK(SO_4)_2 \cdot 12H_2O$, 0.01 g H_3BO_3 and 0.01 g $Na_2MoO_4 \cdot 2H_2O$. The vitamin mixture (200× per litre) contains 0.2 g niacin, 0.08 g biotin, 0.2 g pantothenate, 0.2 g lipoic acid, 0.08 g folic acid, 0.2 g *p*-aminobenzoic acid, 0.2 g thiamine, 0.2 g riboflavin, 0.2 g pyridoxine and 0.2 g cobalamin. 5 g/l yeast extract (Y), 5 g/l tryptone (T), 5 g/l pyruvate (Pyr) and 2 g/l elemental sulfur (S˚) were added to rich medium (ASW-YT-Pyr-S˚). ASW-S˚ mixture supplemented with a combination of 20 amino acids formed minimal medium (ASW-aa-S˚). The amino acid mixture contains (20× per litre) 5 g cysteine, 5 g glutamic acid, 5 g glycine, 2.5 g arginine, 2.5 g proline, 2 g asparagine, 2 g histidine, 2 g isoleucine, 2 g leucine, 2 g lysine, 2 g threonine, 2 g tyrosine, 1.5 g alanine, 1.5 g methionine, 1.5 g phenylalanine, 1.5 g serine, 1.5 g tryptophan, 1 g aspartic acid, 1 g glutamine and 1 g valine.

All *T. kodakarensis* cultures were grown at 55–95 °C under strict anaerobic conditions in sealed vessels with a headspace gas composition of 95% N_2/5% H_2 at 1 atmosphere at 22 °C; 1 mM agmatine was provided when necessary. Solid medium was prepared by the addition of 1% gelzan, with polysulfides substituting for S˚ (refs. [31,32]). Polysulfides were prepared (500×, per 15 ml) by dissolving 10 g $Na_2S \cdot 9H_2O$ and 3 g S˚ with heat to a deep red mixture. Colonies formed on solid medium were observed by lifting cells to polyvinylidene difluoride membranes that were then flash-frozen in liquid N_2 before being thawed and stained with Coomassie Brilliant Blue. *P. furiosus* strain COM1 was cultured at 75–95 °C in an artificial-seawater-based medium supplemented with cellobiose, maltose, yeast extract, S˚, trace minerals, cysteine and sodium tungstate as previously described[33]. *Thermococcus sp*. AM4[34] was cultured under identical conditions to those for *T. kodakarensis*.

Yeast growth and media conditions

S. cerevisiae strains were grown at 30 °C in standard YEP medium (1% yeast extract, 2% Bacto Peptone) supplemented with 2% dextrose (YPD). For induction of *Tan1* by galactose, cells were washed twice with water, resuspended in YEP medium (1% yeast extract, 2% Bacto Peptone) supplemented with 2% galactose (YPG) and grown at 30 °C for 21 h before collection.

Construction of *T. kodakarensis* strains markerlessly-deleted for *TK0754* or *TK2097*

Plasmids used to direct the markerless deletion of genomic sequences from the parental strain TS559 were each individually constructed from the parental plasmid pTS700[30] and contain approximately 700 bp sequences complementary to both upstream and downstream regions of the respective locus under study[29]. Each vector also encodes expression cassettes for *TK0149* (provides agmatine autotrophy) and *TK0664* (provides sensitivity to 6-methylpurine). Strains were constructed as previously described[29,30,35]. In brief, plasmids incapable of autonomous replication in *T. kodakarensis* were individually transformed into *T. kodakarensis* TS559 (Δ*TK0149*; Δ*TK0664*; Δ*TK0254*::*TK2276*; Δ*TK2276*)[29,30,32,35]. Plasmid integration at the desired locus was confirmed by several diagnostic PCR amplicons generated from genomic DNA purified from intermediate strains. Overnight growth in the presence of 1 mM agmatine permitted spontaneous plasmid excision, and colonies were selected on solid media containing 20 amino acids, 6-methylpurine and agmatine. DNA was extracted from 1 ml ASW-YT-Pyr-S°-agmatine cultures grown from individual 6-MP resistant colonies for use in diagnostic PCRs to confirm the deletion of the desired locus. Final confirmation of each strain included whole-genome sequencing[29] to confirm deletion endpoints and to ensure no unanticipated modifications were introduced into the genome at remote locations.

Plasmids for NAT10, Tan1 and THUMPD1 overexpression

Tan1 was synthesized and cloned into pD1201 and pD1231 by ATUM. The remaining plasmids were constructed using Gateway recombination cloning (Thermo Fisher) as follows: NAT10 was amplified from a cDNA plasmid (Dharmacon, accession number BC035558) by PCR and cloned into pDonr-255 with BP Clonase. The insert was sequence-verified and subcloned with LR Clonase into a neomycin-resistant mammalian transfection backbone with CMV promoter and N-terminal 3×Flag–eGFP fusion. The same strategy, NAT10 entry clone, and expression vector were used to generate 3×Flag–NAT10. THUMPD1 was amplified from a cDNA plasmid (Dharmacon, accession number BC000448) by PCR and the entry clone was generated and verified in a similar fashion. This entry clone was then subcloned with LR Clonase into a neomycin-resistant mammalian transfection backbone with CMV promoter, and N-terminal myc tag. Transfection-quality plasmid DNAs were prepared using ZymoPURE II Plasmid Maxiprep Kit (Zymo Research)

Overexpression of eGFP–NAT10 in HEK-293T cells

HEK-293T cells were plated in a 10 cm dish (2.5×10^6 cells per dish in 10 ml DMEM medium) and allowed to adhere and grow for 24 h. eGFP-tagged NAT10 was overexpressed using FuGENE 6 transfection reagent (Promega, E2691). Before transfection, 600 µl of OPTI-MEM (Gibco, 31985062) was incubated with 18 µl FuGENE 6 for 5 min at room temperature before adding 6 µg of eGFP–NAT10 plasmid and incubating for an additional 30 min. Transfection mixture was carefully added to the cell monolayer without changing the medium. Overexpression was carried out by incubating the cells for 24 h at 37 °C under a 5% CO_2 atmosphere, after which cells were imaged using an EVOS FL fluorescence microscope at 10× and 40× magnifications.

Co-overexpression of NAT10 and THUMPD1 in HEK-293T cells

HEK-293T cells were seeded into twenty 10 cm dishes (2.5×10^6 cells per dish in 10 ml DMEM medium) and allowed to adhere and grow for 24 h. 3×Flag-tagged NAT10 and myc-tagged THUMPD1 were overexpressed using FuGENE 6 transfection reagent (Promega, E2691). For each 10 cm dish, 600 µl of Opti-MEM I Reduced Serum Medium (Gibco, 31985062) was incubated with 18 µl FuGENE 6 for 5 min at room temperature before adding 3 µg each of NAT10 and THUMPD1 plasmid and incubating for an additional 30 min. Transfection mixtures were carefully added to the cell monolayer without changing the medium. Overexpression was carried out by incubating the cells for 24 h at 37 °C under 5% CO_2 atmosphere, after which 19 plates were collected by trypsinization and snap-frozen for total RNA extraction. The remaining plate was collected using ice-cold PBS and pelleted for western blot analysis of overexpression. The cell pellet was resuspended in 500 µl of ice-cold PBS containing protease inhibitor cocktail (1X, EDTA-free, Cell Signaling Technology, 5871S). Samples were then lysed by sonication using a 100 W QSonica XL2000 sonicator (3 × 1 s pulse, amplitude 1, 60 s resting on ice between pulses). The lysate was pelleted by centrifugation (20,817 rcf × 30 min, 4 °C) and quantified using the Qubit 4.0 Fluorometer and Qubit Protein Assay Kit. Protein was run on SDS–PAGE alongside non-transfected control and immunoblotted with anti-Flag-tag (Cell Signaling, 2044), anti-NAT10 (Bethyl Laboratories, A304–385A), and anti-myc-tag (Cell Signaling, 5605) antibodies. For immunoblotting, SDS–PAGE gels were transferred to nitrocellulose membranes (Novex, Life Technologies, LC2001) by electroblotting at 30 V for 1 h using a XCell II Blot Module (Novex). Membranes were blocked using StartingBlock (PBS) Blocking Buffer (Thermo Scientific) for 30 min and incubated overnight at 4 °C in primary antibody. The membranes were washed with TBST buffer and incubated with secondary HRP-conjugated antibody (Cell Signaling, 7074) for 1 h at room temperature. The membranes were again washed with TBST and treated with chemiluminescence reagents (Western Blot Detection System, Cell Signaling) for 1 min, and imaged for chemiluminescent signal using an ImageQuant Las4010 Digital Imaging System (GE Healthcare).

For targeted ac^{4}C-sequencing in cells overexpressing either NAT10, THUMPD1, neither, or both, HEK-293T cells were seeded in replicates in wells of a 6-well plate (0.5×10^6 cells per well in 2 ml DMEM media) and allowed to adhere and grow for 24 h. Cells were transfected using PolyJet (SignaGen Laboratories) according to the manufacturer's protocol, either with 0.5 µg NAT10, or with 0.5 µg THUMPD1, neither or both. In all samples a total of 50 ng GFP plasmid was used to monitor transfection efficiency. Cells were grown for 24 h before collecting for RNA purification.

Growth analysis of *T. kodakarensis*

Parental strain TS559 and Tk*NAT10*-deleted *T. kodakarensis* cells were grown as described above at 65–95 °C (11–12 replicates from each temperature). Growth of liquid cultures was monitored by measurements of optical density at 600 nm at hourly intervals for a total of 33 h. Measurements were used to model cell growth using the 'locally estimated scatterplot smoothing' (loess) method[36].

Total RNA isolation from yeast, human and archaea

Total RNA from human cells was extracted using TRIzol according to the manufacturer's protocol. 1 ml TRIzol was used per 1×10^7 cells. The RNA pellet was resuspended by briefly heating at 50 °C in 1.0 ml 1X TE buffer pH 8.0. Samples were quantified by UV absorbance and stored at −80 °C. Typical extractions were carried out with 4×10^7 cells and yielded 400 µg of total RNA.

For targeted ac^{4}C-sequencing, RNA was extracted using Nucleozol (Macherey Nagel) according to the manufacturer's instructions.

Total RNA was isolated from yeast using hot acidic phenol. In brief, a frozen yeast (*S. cerevisiae*) pellet was suspended in 1.0 ml AES buffer (50 mM sodium acetate, 10 mM EDTA pH 8.0, 1% SDS) per 0.5 ml pellet volume. To the suspended pellet, 1.0 ml acid-buffered phenol per ml of AES buffer used was added. The sample was mixed by vortexing and incubated in a 65 °C water bath for 30 min, vortexing every 2 min to mix. Samples were put on ice for 10 min and 1.0 ml chloroform:isoamyl alcohol (24:1) was added for each 1.0 ml phenol used. The sample was vortexed to mix and centrifuged at 5,000 rcf for 15 min. The aqueous layer (top) was transferred to a clean tube and extracted three times with an equal volume of acid-buffered phenol:chloroform:isoamyl alcohol (24:23:1). After each extraction the sample was centrifuged at 5,000 rcf for 10 min and the aqueous layer was transferred to a new tube. A final extraction with chloroform:isoamyl alcohol was carried out to remove residual phenol. The aqueous layer was transferred to a clean tube and RNA was precipitated by the addition of an equal volume of 100% isopropanol and 1/9th volume of 3 M sodium acetate. Samples were incubated at −20 °C for 30 min and centrifuged at 12,000 rcf at 4 °C for 15 min. The supernatant was decanted and the pellet was washed with 4 ml ice-cold 70% ethanol. The RNA pellet was resuspended by briefly heating at 50 °C in 1.0 ml 1X TE buffer at pH 8.0. Samples were quantified by UV absorbance and stored at −80 °C. Typical extractions were carried out with cell pellets of 1.0 ml volume and yielded 20 mg of total RNA. Total RNA was isolated from archaeal samples using TRIzol according to the manufacturer's protocol.

Poly(A) RNA isolation from yeast and human cells

Poly(A) RNA from yeast and human total RNA was isolated by two rounds of purification using the GenElute mRNA miniprep kit (Sigma) according to the manufacturer's protocol. 500 µg total RNA was used per purification column. A typical yield after two rounds of isolation was 1.2%. For targeted ac^{4}C-sequencing, poly(A) RNA was isolated from total RNA of HEK-293T cells by two rounds of purification using Dynabeads mRNA DIRECT Kit (Invitrogen), according to the

manufacturer's protocol. 75 µg total RNA was taken from each sample, using 150 µl oligo dT beads. Typical yield after two rounds of isolation was 1.6%.

Ribosome purification

Purification of *T. kodakarensis* ribosomes of the wild-type and the Tk*NAT10* deletion strains were conducted similarly to previously documented procedures[37]. In brief, cell lysis was obtained through sonication in buffer A (20 mM HEPES, pH 7.5, 10.5 mM magnesium acetate, 100 mM ammonium acetate, 0.5 mM EDTA and 6 mM β-mercaptoethanol). Cell debris was discarded by centrifugation at 30,000*g* for 20 min at 4 °C, and the cytoplasmic fraction was loaded onto a 1.1 M sucrose cushion in buffer B (20 mM HEPES, pH 7.5, 10 mM magnesium acetate, 150 mM potassium acetate 6 mM β-mercaptoethanol). The ribosome-enriched pellet was obtained by overnight centrifugation at 220,000*g* at 4 °C. The pellet was resuspended in buffer B and ribosome particles were purified on a 10–40% sucrose gradient using a SW-28 rotor, at 43,000*g* for 17 h at 4 °C. Fractions containing 70S ribosomes were collected, combined and centrifuged at 230,000*g* overnight at 4 °C. The pellet was resuspended in buffer B and an additional centrifugation step at 200,000*g* for 1.5 h at 4 °C was designed to remove sucrose traces. The ribosomal pellet was resuspended in buffer C (20 mM HEPES pH 7.5, 10 mM magnesium acetate, 100 mM potassium acetate, 100 mM ammonium acetate and 1 mM DTT), diluted to a concentration of 1 mg/ml aliquoted and stored at −80 °C until further use.

rRNA depletion from total RNA of *T. kodakarensis*

To deplete abundant *T. kodakarensis* rRNAs before RNA-seq, we adapted a method originally reported previously[38] using reagents provided in the NEBNext rRNA Depletion Kit (NEB, E6310). The protocol in the manual for the kit was followed with the following changes. The NEBNext rRNA Depletion Solution provided in the kit was substituted for an equimolar mixture of 85 oligonucleotides complementary to *T. kodakarensis* rRNA sequences (Supplementary

Table 1b). The concentration of the oligo mix was 85 µM, such that each individual oligo was at 1 µM in the mix. All volumes for the probe hybridization, RNase H treatment and DNase I treatment sections of the protocol were scaled up twofold and 24 µl of 62.5 ng/µl *T. kodakarensis* RNA was used as the starting material. Instead of bead purification as indicated in the manual, samples were purified using the Monarch RNA Cleanup Kit (NEB, T2030) using the standard protocol. Sixteen depletion reactions were performed as described above for each *T. kodakarensis* total RNA sample and these were then concentrated into a single depleted RNA sample by pooling them and performing a second round of purification with the Monarch RNA Cleanup Kit. The yield of RNA after depletion was measured using the Qubit RNA BR Assay Kit (Thermo Fisher).

UV spectroscopic analysis of ac^{4}C reduction rates

Model reactions to assess the rate of reduction of ac^{4}C by $NaBH_4$ and $NaCNBH_3$ were performed using free N^4-acetylcytidine nucleoside. For $NaBH_4$ reductions, stock solutions of $NaBH_4$ (100 mM) and N^4-acetylcytidine (2 mM) were prepared fresh daily in water. Reactions (25 µl) consisted of N^4-acetylcytidine (100 µM), $NaBH_4$ (20 mM) and reaction buffer (water, 100 mM sodium acetate (pH 4.5), or 100 mM potassium phosphate (pH 7.5)). At the indicated time point, reactions were adjusted to 50 µl using 100 mM HCl. To normalize pH, a further aliquot of 50 µl 100 mM sodium phosphate (pH 7.2) was added and reactions were transferred to Greiner-UV Star 96-well half-area microplates (655801) for analysis. For $NaCNBH_3$ reductions, stock solutions of $NaCNBH_3$ (1 M) and N^4-acetylcytidine (2.5 mM) were prepared fresh daily in water. Reactions (100 µl) consisted of N^4-acetylcytidine (100 µM), $NaCNBH_3$ (100 mM) and HCl (100 mM). At the indicated time point, reactions were quenched with 30 µl of 1 M Tris-HCl (pH 8.0), and added to Greiner-UV Star 96-well microplates for analysis. Reduction of N^4-acetylcytidine was analysed on a Biotek Synergy plate reader by monitoring the absorbance of N^4-acetylcytidine (λ_{max} = 300 nm) and cytidine (λ_{max} = 270 nm). For N^4-acetylcytidine reactions, the

percentage decrease in N^4-acetylcytidine was calculated from absorbance (A) values at 300 nm using the formula: Percentage decrease = $(A_{ac4C(start)} - A_{ac4C(end)})/(A_{ac4C(untreated)} - A_{water(blank)}) \times 100$.

UV spectroscopic analysis of ac^{4}C deacetylation

Model reactions to assess the rate of acid- and base-induced deacetylation of ac^{4}C were performed using free N^4-acetylcytidine nucleoside. Stock solutions of N^4-acetylcytidine (2.5 mM) were prepared fresh daily in water. Reactions (50 µl) consisted of N^4-acetylcytidine (250 µM) and reaction buffer (KCl/HCl buffer (pH 1) or $NaHCO_3$ buffer (pH 10)) added to a Greiner-UV Star 96-well half-area microplate. Control reactions were set up similarly with cytidine (250 µM). Deacetylation of N^4-acetylcytidine was analysed on a Biotek Synergy plate reader by monitoring the absorbance of N^4-acetylcytidine (pH 1 λ_{max} = 310 nm; pH 10 λ_{max} = 300 nm) over 18 h. For N^4-acetylcytidine reactions, the percentage decrease in N^4-acetylcytidine was calculated from λ_{max} absorbance values using the formula: percentage decrease = $(A_{ac4C(start)} - A_{ac4C(end)})/(A_{ac4C(untreated)}) \times 100$.

In vitro transcription of synthetic ac^{4}C-containing RNAs as spike-in controls

In vitro transcription was performed with the HiScribe T7 Kit (New England Biolabs), according to the manufacturer's instructions using DNA templates containing a T7 promoter upstream of a template sequence harbouring a single cytidine within an ACA, GCA, ACG or GCG sequence context (Supplementary Table 1a). For ac^{4}C-containing transcripts, CTP was replaced in the reaction mixture with ac^{4}CTP (10 mM) as described previously[12]. In vitro transcription reactions were analysed by denaturing polyacrylamide gel electrophoresis on 10% TBE-urea gels and visualized using SYBR Gold staining. Synthetic RNA products were used in ac^{4}C-seq, LC–MS quantification, and reverse transcription stop experiments, the latter of which were performed as previously described[11].

Mass spectrometry analysis of ac^{4}C in synthetic spike-in controls

Mass spectrometry analysis of ac^{4}C reduction in RNA probes was assessed after nuclease digest as described previously[12]. In brief, in vitro transcribed ac^{4}C RNA was treated with nuclease P1 (2U/10 µg RNA, N8630, Sigma) in 50 µl of buffer containing 100 mM ammonium acetate (pH 5.5), 2.5 mM NaCl and 0.25 mM $ZnCl_2$ for 2 h at 37 °C. Sample volumes were adjusted to 60 µl by adding 3.5 µl of H_2O, 6 µl of 10× Antarctic Phosphatase buffer (B0289S, NEB) and 0.5 µl of Antarctic Phosphatase (1 U/10 µg RNA, M0289S, NEB). Samples were further incubated at 37 °C for 2 h, adjusted to 150 µl with RNase-free water and filtered via centrifugation to remove enzymatic constituents (Amicon Ultra 3K, UFC500396). After lyophilization, samples were reconstituted in 10 µl RNase-free water and analysed via LC–MS/MS using reverse phase chromatography (Shimadzu LC-20AD) coupled to a triple-quadrupole mass spectrometer (Thermo TSQ-ultra) operated in positive electrospray ionization mode. Quantification was accomplished by monitoring nucleoside-to-base ion transitions and generating standard curves for each nucleoside using the stable isotope dilution internal standardization method.

Primer extension and reverse transcription stop analysis of ac^{4}C RNAs

Primer extension assays were performed using PAGE-purified model RNAs containing a single site of either ac^{4}C or cytidine produced by in vitro transcription (sequence provided above). For each reaction, RNA (2 µg) was treated in a final reaction volume of 100 µl. For $NaBH_4$-treated samples: 1 M $NaBH_4$ was added to 2 µg RNA in nuclease-free H_2O to a final concentration of 100 mM and samples were incubated for 60 min at 37 °C, $NaBH_4$ was quenched with 1 M HCl (15 ml), and neutralized by the addition of 1 M Tris-Cl (pH 8.0) buffer (15 ml). For $NaCNBH_3$ treated samples: 1 M $NaCNBH_3$ was added to 2 µg RNA in nuclease free H_2O to a final concentration of 100 mM. Reactions were initiated by the addition of 1 M HCl to a final concentration of 100 mM and samples were incubated 20 min at room temperature (20 °C). The

reaction was stopped by neutralizing the pH by the addition of 30 µl 1 M Tris-HCl pH 8.0. For untreated control samples: 1 M HCl was added to 2 µg RNA in nuclease-free water to a final concentration of 100 mM and samples were incubated for 20 min at room temperature (20 °C). Reactions were stopped by neutralizing the pH by the addition of 30 µl 1 M Tris-HCl pH 8.0. Reactions were adjusted to 200 ml with H_2O, precipitated with ethanol, desalted with 70% ice-cold ethanol, briefly dried on Speedvac, resuspended in H_2O, and quantified by absorbance using a Nanodrop 2000 spectrophotometer. RNA from individual reactions (5 pmol) was incubated with 5′-Cy5 IVT primer (5′-/Cy5/ACTCATCACTTTTCTCCCTCTACACAATC-3′; 3.5 pmol) in a final volume of 50 µl. Individual reactions were heated to 65 °C for 5 min and cooled at a rate of 5 °C per min to a final temperature of 4 °C to facilitate annealing, with the following buffer conditions used for specific RTs: AMV: 1X AMV reaction buffer (NEB), 1.0 mM dNTPs; Superscript III: 500 mM dNTPs; TGIRT: 1X TGIRT reaction buffer (Ingex), 5 mM $MgCl_2$. After annealing, reverse transcriptions were performed as follows: 1) AMV reactions: 100 units RNaseOUT (Invitrogen), 25 U AMV RT, incubate 60 min, 48 °C; 2) Superscript III: 1x SSIII reaction buffer (from 10x stock; Thermo Fisher), 5 mM $MgCl_2$, 10 mM DTT, 100 U RNaseOUT, 500 U Superscript III, incubate 60 min, 48 °C; 3) TGIRT reactions: first add 5 mM DTT, 500 U TGIRT RT, incubate 20 min room temperature, then add 500 mM dNTPs, incubate 1 h, 57 °C. After the indicated incubation time, reactions were adjusted to 200 µl with H_2O, extracted with phenol:chloroform, precipitated with ethanol, desalted with 70% ice-cold ethanol, briefly dried on Speedvac, and resuspended in 20 ml of 1X RNA denaturing RNA loading buffer. Samples were heated at 95 °C for 4 min, cooled on ice, loaded onto a 10% denaturing polyacrylamide gel and run at 400 V (20 V/cm) for 5 h. Gels were fluorescently visualized using an ImageQuant Las4010 (GE Healthcare) with red LED excitation (λ_{max} = 630 nm) and a R670 filter, with band intensities quantified by densitometry using Imagequant software. To calculate the product/stop ratio, the fluorescence intensity of the bands observed at the ac^4C site (−1, 0 or +1) were

divided the total fluorescence intensity of all other primer extension products observed in each gel lane.

Reverse transcription and misincorporation analysis of RNAs by Sanger sequencing

For each reaction, RNA (1 µg) was incubated with either $NaCNBH_3$ (100 mM in H_2O + 100 mM HCl) or untreated 'mock' control (H_2O + 100 mM HCl) in a final reaction volume of 100 ul. Samples were incubated for 20 min at 20 °C. Reactions were stopped by neutralization of pH by the addition of 30 µl 1 M Tris-HCl pH 8.0. Reactions were adjusted to 200 µl with H_2O, precipitated with ethanol, desalted with 70% ice-cold ethanol, briefly dried on Speedvac, resuspended in H_2O, and quantified by absorbance using a Nanodrop 2000 spectrophotometer. RNA from individual reactions (200 pg) was incubated with 4.0 pmol RT primer in a final volume of 20 µl. Individual reactions were heated to 65 °C for 5 min and transferred to ice for 3 min to facilitate annealing in 1× TGIRT reaction buffer (Ingex), 5 mM $MgCl_2$. After annealing, reverse transcriptions were performed as follows using TGIRT-III; DTT was added to 5 mM along with 100 U TGIRT RT and 25 U RNasin Plus (Promega). The reaction was incubated for 20 min at room temperature. The reverse transcription reaction was initiated by addition of dATP, dTTP and dCTP to 500 mM and dGTP to 250 mM. Reactions were incubated for 1 h at 57 °C. cDNA (2 µl) was used as template in 50 ml PCR reaction with Phusion Hot start flex (NEB). Reaction conditions: 1X supplied HF buffer, 2.5 pmol each forward and reverse primer, 200 mM each dNTPs, 2 U Phusion hot start enzyme, 2 µl template and the following specific conditions:

In vitro transcribed 'single ac^4C': Primers: IVT rev (PCR primer), IVT forward (PCR primer). Thermocycling conditions: 71 °C annealing, 34 cycles.

Human 18S rRNA, helix 45 ac^4C site: Primers: human 18S helix 45 fwd, human 18S helix 45 rev. Thermocycling conditions: 67.4 °C annealing, 34 cycles.

PCR products were run on a 2% agarose gel, stained with SYBR safe and visualized on UV transilluminator at 302 nm. Bands of the desired size were excised from the gel and DNA extracted using QIA-quick gel extraction kit from Qiagen and submitted for Sanger sequencing (GeneWiz) using the forward PCR primer for 18S sites and reverse PCR primer for IVT 'single ac^4C'. Processed sequencing traces were viewed using 4Peaks software. The peak height for each base was measured and the percentage misincorporation was determined using the equation: Percentage misincorporation = (Sum of non-cognate base peaks intensities)/(sum of total base peaks) × 100%.

Ac^4C-seq library preparation

Strand-specific ac^4C-seq libraries were generated on the basis of previously described protocols[39,40]. In brief, RNA was first subjected to FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific), followed by a 3′ ligation of an RNA adaptor using T4 ligase (NEB). Ligated RNA was reverse transcribed using TGIRT-III (InGex), and the cDNA was subjected to a 3′ ligation with a second adaptor using T4 ligase. The single-stranded cDNA product was then amplified for 9–12 cycles in a PCR reaction. Libraries were sequenced on Illumina NextSeq 500 or NovaSeq 6000 platforms generating short paired-end reads, ranging from 25 to 55 bp from each end.

Samples used in ac^4C-seq analysis

Human: Three experiments were conducted. In the first experiment, total RNA from wild-type HeLa cells or cells with reduced expression of NAT10[10] were treated with $NaCNBH_3$ (with and without alkali pre-treatment) or mock-treated in three biological replicates. In the second, a set of 5 poly(A)-enriched HeLa samples (3 and 2 biological replicates for wild-type and NAT10 knock-down, respectively) were treated with $NaCNBH_3$ or mock-treated. For the third

experiment, poly(A)-enriched HEK-293T cells co-overexpressing NAT10 and THUMPD1 (2 biological replicates) and a sample of wild-type cells were treated with $NaCNBH_3$ (with and without alkali pre-treatment) or mock-treated.

Yeast: Two experiments were conducted. In the first, biological duplicates of wild-type *S. cerevisiae* cells and cells expressing a catalytic mutant of Kre33[1] were treated with $NaCNBH_3$ (with and without alkali pre-treatment) or mock-treated. In the second, cells co-overexpressing Kre33 and Tan1 in a Kre33-catalytic mutant strain were analysed in comparison to wild-type *S. cerevisiae* cells. One replicate of the co-overexpression cells expressed Tan1 under a constitutive GPD promoter, the other under a GAL1-inducible promoter. These cells were grown in YPD and YPG, respectively, along with a matching wild-type sample grown under the same conditions. These four samples were treated with $NaCNBH_3$ or were mock-treated. All libraries of yeast were prepared from poly(A)-enriched RNA.

***T. kodakarensis*:** A total of 17 samples were analysed, representing 25 treatment conditions. For all samples total RNA was analysed from a single biological sample, unless stated otherwise. TS559 cells grown at 55, 65, 75, 85 and 95 °C were treated with $NaCNBH_3$ or mock-treated. For the 85 °C condition, four biological replicates were assessed, and one of them also underwent alkali pre-treatment. For 65 and 75 °C two biological replicates were assessed. Biological duplicates of cells in which Tk*NAT10* (*TK0754*) or Tk*THUMPD1* (*TK2097*) were deleted were treated with $NaCNBH_3$. ΔTk*NAT10* samples were also mock-treated. rRNA-depleted RNA from TS559 cells grown at 85 and 95 °C were treated with $NaCNBH_3$ or were treated with $NaCNBH_3$ and mock-treated, respectively. Purified ribosomes from TS559 cells grown at 85 °C were treated with $NaCNBH_3$.

***T. sp*. AM4:** total RNA from cells grown at 65, 75 and 85 °C was treated with $NaCNBH_3$ or mock-treated.

P. furiosus: total RNA from cells grown at 75, 85 and 95 °C was treated with $NaCNBH_3$ or mock-treated.

S. solfataricus: total RNA from cells grown at 85 °C was treated with $NaCNBH_3$ (with and without alkali pre-treatment) or mock-treated. A total of three samples were used, representing a single biological sample.

M. jannaschii: a single sample was treated with $NaCNBH_3$ (with and without alkali pre-treatment) or mock-treated.

Identification of putative ac^{4}C sites

Reference genomes were generated on the basis of the following genome assemblies: ASM996v1 for *T. kodakarensis*, ASM27560v1 for *P. furiosus*, ASM15120v2 was used for *T. sp*. AM4, ASM700v1 for *S. solfataricus* and ASM9166v1 for *M. jannaschii*. For human poly(A)-enriched samples we used the GRCh37/hg19 with UCSC Genes annotations, supplemented with tRNA, rRNA and snRNAs sequences, obtained from the Modomics database[41]. Samples from total RNA of human cells were aligned to a subset of the full reference containing only the tRNA, rRNA and snRNA sequences. For *S. cerevisiae* samples the sacCer3 assembly was used in experiments designed to detect modification in mRNA, whereas a limited reference containing only rRNAs and tRNAs (filtered to only retain non-redundant sequences) was used in experiments designed to detect only sites in these non-coding transcripts.

Samples were aligned to the genome using STAR aligner[42]. For archaeal and *S. cerevisiae* samples intron size was limited to 500 bases ('alignIntronMax = 500'). For poly(A)-enriched samples (applicable to some of the human and yeast samples, as indicated in the main text) duplicated reads and chimeric pairs were filtered out by the dedup function of UMI-tools[43] (using '–chimaeric-pairs = discard') followed by removal of overlapping reads by the clipOverlap function of bamUtil[44]. For human and yeast samples aligned to a limited reference containing only the ncRNA sequences mentioned above, multiple mapping was allowed ('multiMapping = 200').

Single nucleotide variants were detected using the JACUSA software in pileup mode[45], which outputs a tabular format summarizing the abundance of each nucleotide (with minimal coverage of 5 reads) at each position. A custom script was used to extract the misincorporation rate at each position as well as to identify the most abundant nucleotide appearing instead of the wild-type nucleotide (the 'predominant base conversion').

For a position to be considered as putatively modified, it had to meet two sets of requirements. First, at the level of an individual $NaCNBH_3$-treated sample compared to a suitable control (whereby the control is in most cases a mock-treated sample, but in some cases is a chemically deacetylated sample or a NAT10-deficient genetic control) the fundamental requirement it had to meet was that the *P* value obtained from the χ^2 test comparing the misincorporation rates in the treated versus control samples was lower than 0.05. In experiments with multiple replicates, the χ^2 test was conducted on 'pooled samples' combining misincorporation information from all replicates. Second, to reduce the computational load, we applied this statistical framework only to sites matching the minimal criteria below: (1) At least three reads with misincorporations in the $NaCNBH_3$-treated sample (or wild-type sample, when comparing to NAT10-deficient). (2) A misincorporation rate > MIN_RATE in the $NaCNBH_3$-treated sample (for archaea we used a

MIN_RATE_TREAT = 2%, for human and yeast with larger genomes and consequently slightly reduced signal:noise ratios we used 3%). (3) A misincorporation rate lower than MAX_RATE_CONT in the control sample (MAX_RATE_CONT = 5% in archaea, 1% in human and *S. cerevisiae*). (4) Misincorporation rates in the $NaCNBH_3$-treated sample were at least 2% higher than in their control counterparts. (5) The predominant base conversion at the site in the $NaCNBH_3$-treated sample was from cytidine to thymidine (C>T). To eliminate redundancies, positions harbouring identical sequences in a 21-bp window (10 bp upstream + 10 bp downstream) surrounding the putative site were filtered to retain only one. Furthermore, when possible on the basis of the experimental design, we demanded that such a site be reproducibly identified across at least two distinct comparisons. The distinct experimental design for the different organisms (in some cases we monitored distinct temperatures, in others distinct genetic backgrounds, in others we obtained static snapshots under one condition) was taken into consideration, and the precise set of comparisons performed for each organism is detailed in Supplementary Table 2. This set of comparison was used to create a final 'catalogue of ac^4C sites' for each organism, which was used in downstream analyses. All catalogues, segregated by organism.

Motif analysis

For each species, we extracted the 20 nt flanking the ac^4C positions in its catalogue of 'significantly modified' sites. These 21-nt long sequences were used to generate sequence logos using the WebLogo software (available at https://weblogo.berkeley.edu/logo.cgi)[46], in which the height of each stack indicates the information content at that position (measured in bits), whereas the height of letters within the stack reflects the relative frequency of the corresponding nucleic acid at that position.

Targeted ac^4C-sequencing

mRNA samples treated with $NaCNBH_3$ were incubated with Turbo DNase (Invitrogen) for 30 min at 37 °C. 400 ng of the DNase-treated mRNA was reverse transcribed using TGIRT-III (InGex), with random primers (Applied Biosystems). After cleanup of cDNA using Dynabeads MyOne SILANE beads (Life Technologies), 10 cycles of PCR were carried out using Kapa HiFi HotStart Readymix PCR kit (Kapa Biosystems), and pairs of primers described in Supplementary Table 1a. 1 µl of the PCR reaction was used as template for a second PCR reaction (Kapa HiFi, 25 µl reaction volume, 20 cycles), in which barcoded Illumina adaptors were added. Amplicons were analysed on 2% E-gel EX agarose gels (Invitrogen), and cleaned using two rounds of AMPure XP beads (Beckman Coulter). For targeted ac^4C-sequencing of overexpressed sequences, total RNA samples were treated with $NaCNBH_3$ and incubated with Turbo DNase (Invitrogen) for 30 min at 37 °C. 600 ng of the DNase-treated total RNA was reverse transcribed using TGIRT-III (InGex), with random primers (Applied Biosystems). After cleanup of cDNA using Dynabeads MyOne SILANE beads (Life Technologies), 20 cycles of PCR were carried out using Kapa HiFi HotStart Readymix PCR kit (Kapa Biosystems), adding the barcoded Illumina adaptors.

Construction of plasmids for overexpression of wild-type (CCG) and mutated (CCA) ac^4C sites

The sequences described in Supplementary Table 1a were cloned using FastDigest SgsI (AscI) and BcuI (SpeI) restriction enzymes (Thermo Scientific) into pZDonor FC plasmid, as a 3′ UTR of a reporter gene[47].

Targeted ac^4C -sequencing of a pool of sequence variants of BAZ2A mRNA

Pool design: A 91-base-long sequence surrounding the ac4C site identified in *BAZ2A* mRNA was used as a wild-type control fragment. Variants of the wild-type *BAZ2A* fragment were made by introducing a single point mutation at each base of the wild-type sequence, by replacing it with all possible bases. *BAZ2A* fragments were preceded by an 8-base barcode, allowing each

variant to be uniquely mapped, and flanked by SpeI and AscI restriction sites to facilitate cloning, Illumina adaptor sequences to allow sequencing, and primer sequences to allow amplification of the entire construct in the cloning stage.

Cloning of the oligonucleotide pool: The pool of sequences was cloned as 3′ UTR downstream of a reporter gene in the pZDonor FC plasmid, essentially as described previously[48]. Specifically, the library was amplified in 5 different PCR reactions, each using 50 pg as a template and 14 cycles. The reactions were combined, cleaned using an QIAquick PCR purification kit (Qiagen), and a total of 540 ng was cut by SgsI (AscI) and BcuI (SpeI) restriction enzymes (FastDigest, Thermo Scientific). After electro-elution from a gel using Midi GeBAflex tubes (GeBA, Kfar Hanagid, Israel), the library was ligated (in 1:1 ratio) to pZDonor FC plasmid digested by SgsI and BcuI, using CloneDirect Rapid Ligation kit (Lucigen Corporation) and transformed into E. cloni 10G electrocompetent cells (Lucigen) in a single cuvette. The bacteria were grown on four 14-cm plates, reaching on average about 1,500 colonies per each sequence variant. Plasmids were purified directly from collected bacterial colonies.

Transfection, treatment and library preparation: The plasmids pool was transfected to 10-cm plates of HEK-293T cells in replicates using PolyJet reagent (SignaGen Laboratories), either by itself (2 µg) or together with both NAT10 and THUMPD1 (1.5 µg each). For targeted ac^4C-sequencing of the library variants, total RNA samples were treated with $NaCNBH_3$ and incubated with Turbo DNase (Invitrogen) for 30 min at 37 °C. 1 µg of the DNase-treated total RNA was reverse transcribed using TGIRT-III (InGex), with random primers (Applied Biosystems). After cleanup of cDNA using Dynabeads MyOne SILANE beads (Life Technologies), half of the cleaned cDNA was used in a 25-cycle PCR reaction, using Kapa HiFi HotStart Readymix PCR kit (Kapa Biosystems), and Illumina adaptors as primers.

Analysis: SAMtools mpileup was used to assess misincorporation rates at the ac^{4}C site of BAZ2A variants.

mRNA expression analysis

To estimate expression levels, reads were aligned against the human, yeast or *T. kodakarensis* genome using RSEM (version 1.2.31) in paired-end and strand-specific mode with default parameters[49]. For robust comparison between different samples, we used trimmed mean of M values (TMM) normalization[50] of the RSEM read counts as implemented by the NOISeq package[51] in R.

Analysis of codon enrichment and distribution across transcript body

Our analysis identified 146 and 119 putative ac^{4}C sites in mRNA of human and *T. kodakarensis*, respectively. For each site its relative position within the codon was identified on the basis of the genome annotation. As a control, the distribution of all remaining cytidines embedded in CCG sequences in the examined mRNAs was calculated. For *T. kodakarensis*, we further calculated the distribution of the putative ac^{4}C sites and the control cytidines between specific codons encoding the different amino acids. For human sites we mapped the location of each ac^{4}C site and control cytidines (as described above) within the transcript body (that is, 5′ UTR, CDS or 3′ UTR) and calculated the distribution across transcript regions.

Multiple alignment of tRNAs

All *T. kodakarensis* tRNA sequences were multiply aligned against each other using MAFFT v7.402 with default parameters[52]. Manual inspection of aligned sequences facilitated assignment of ac^{4}C sites into distinct regions within the tRNA structure and into specific positions within a canonical model of a tRNA.

Conservation analysis between archaea

Sequences of 16S, 23S, 5S, RNaseP RNA and SRP RNA were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) from genome references NC_006624.1, NC_018092.1 and NC_016051.1 for *T. kodakarensis*, *P. furiosus* and *T. sp*. AM4, respectively. Multiple sequence alignment was conducted across all three archaea for each gene separately using the Clustal Omega software with default parameters (https://www.ebi.ac.uk/Tools/msa/clustalo/)[53]. A custom script was used to detect ac^4C at positions conserved between at least two species and assign it with the relevant misincorporation rates as calculated using ac^4C-seq across all samples. This dataset was used for archaea conservation-related analysis presented in the main text.

Phylogenetic tree

A phylogenetic tree for the archaea analysed by ac^4C-seq was generated using the default parameters of phyloT tree generator (https://phylot.biobyte.de) based on the following NCBI taxonomy IDs: *T. kodakarensis*, 69014; *T*. sp. AM4, 246969; *P. furiosus*, 1185654; *S. solfataricus*, 555311 and *M. jannaschii*, 2190.

Comparison between ac^4C sites in *T. kodakarensis* rRNAs as measured by ac^4C-seq and LC–MS

A total of 172 ac^4C sites at CCG motifs were identified in *T. kodakarensis* rRNA under the full set of comparisons detailed in Supplementary Table 2. Although LC–MS identified a total of 146 potential ac^4C sites, only 25 of these could be uniquely assigned to specific positions within the ribosome, owing to redundancies in the oligo sequences identified in the LC–MS. All comparisons of ac^4C between the methods were therefore conducted on a subset of these 25 sites.

Northern blot analysis of ac^{4}C in archaeal total RNA

Immuno-northern blots were performed using Ambion NorthernMax reagents (Thermo Fisher Scientific). The amount of RNA used was dependent upon sample type, with 15 µg used for analysis of human and yeast total RNA, and 3 µg used for hyperthermophilic archaea. Equal amounts of RNA were mixed together with 1 vol of NorthernMax-Gly Sample Loading Dye (Thermo Fisher Scientific), incubated at 65 °C for 30 min, and separated on a 1% agarose-1X Glyoxal Gel prepared using 10X NorthernMax-Gly Gel Prep/Running Buffer (Thermo Fisher Scientific). Gels were run at 75 V for approximately 70 min, or until the dye front had migrated about 3 inches (7.3 cm). Loading controls were analysed by UV-imaging of ethidium bromide before transfer. RNA was transferred onto Amersham Hybond-N+ membranes (GE Healthcare) using a downward capillary method. After transfer, membranes were crosslinked three times at 150 mJ/cm^2 in a UV254nm Stratalinker 2400 (Stratagene). Membranes were then blocked in a solution of blocking buffer (5% non-fat milk in 0.1% TBST) for 1 h at room temperature and washed 3 times at 5 min each in 0.1% TBST. Membranes were then incubated overnight at 4 °C with the anti-ac^{4}C antibody (1:10,000 dilution, Abcam) in blocking buffer. Membranes were washed 3 × 5 min in 0.1% TBST and then incubated with HRP-conjugated secondary anti-rabbit IgG in 5% non-fat milk in 0.1% TBST at room temperature for 2 h. Membranes were washed 3 times at 10 min each in 0.1% TBST. SuperSignal ELISA Femto Maximum Sensitivity Substrate reagent (Thermo Fisher Scientific) was added directly to the membrane and signal was detected via chemiluminescent imaging. Typical exposure times ranged from 2 to 20 min depending on the concentration of individual RNA samples. We found that for hyperthermophilic archaea a 2-min exposure time was optimal, but yeast and human RNA required a 15- to 20-min exposure time to yield optimal results.

LC–MS analysis of ac^{4}C in total RNA

For assessment of cellular ac^{4}C levels by LC–MS, total RNA was analysed using a similar method as previously described[54]. In brief, before UHPLC–MS analysis, 2,000 ng of each oligonucleotide was treated with 0.5 pg/µl of internal standard (IS), isotopically labelled guanosine, [^{13}C] [^{15}N]G (Cambridge Isotope Laboratories). The enzymatic digestion was carried out using Nucleoside Digestion Mix (NEB) according to the manufacturer's instructions. Finally, the digested samples were lyophilized and reconstituted in 100 µl of RNase-free water, 0.01% formic acid before UHPLC–MS/MS analysis. The UHPLC–MS analysis was accomplished on a Waters XEVO TQ-STM (Waters Corporation) triple quadruple mass spectrometer equipped with an electrospray source (ESI) source maintained at 150 °C and a capillary voltage of 1 kV. Nitrogen was used as the nebulizer gas, which was maintained at a pressure of 7 bar, a flow rate of 500 l/h and a temperature of 500 °C. UHPLC–MS/MS analysis was performed in ESI positive-ion mode using multiple-reaction monitoring (MRM) from ion transitions (*m*/*z* 286.16 > 154.07 and *m*/*z* 286.16 > 112.06) previously determined for ac^{4}C[54]. A Waters ACQUITY UPLCTM HSS T3 guard column, 2.1 × 5 mm, 1.8 µm, attached to a HSS T3 column, 2.1 × 50 mm, 1.7 µm were used for the separation. Mobile phases included RNase-free water (18 MΩ/cm) containing 0.01% formic acid (Buffer A) and 50:50 acetonitrile in Buffer A (Buffer B). The digested nucleotides were eluted at a flow rate of 0.5 ml/min with a gradient as follows: 0–2 min, 0–10%B; 2–3 min, 10–15% B; 3–4 min, 15–100% B; 4–4.5 min, 100% B. The total run time was 7 min. The column oven temperature was kept at 35 °C and the sample injection volume was 10 µl. Three injections were performed for each sample. Data acquisition and analysis were performed with Waters software MassLynx V4.1 and TargetLynx. Calibration curves were plotted using linear regression with a weight factor of 1/*x*.

Preparation of RNase digests and direct nanoflow LC–MS and tandem MS of rRNA fragments

rRNAs were extracted from purified *T. kodakarensis* 70S ribosomes. An aliquot of the sample (100 µl, 1 mg/ml) was mixed with 800 µl of ISOGEN reagent (Nippon Gene) and passed 100 times through a 23–gauge needle. The sheared sample was mixed with 200 µl of chloroform and centrifuged at 10,000*g* for 15 min at 4 °C. The resulting upper phase (around 500 µl) was mixed with a glycogen solution (0.5 µl, 20 mg/ml) and isopropanol (500 µl) and centrifuged to yield rRNAs as a precipitate. The precipitate was dissolved in RNase free water and stored at −80 °C until further use. The three rRNA classes (5S, 16S and 23S) were separated by reversed-phase LC through a PLRP-S 4,000 Å column (4.6 × 150 mm, 10 µm, Agilent Technologies). After applying around 10 µg total RNA to the column, the rRNAs were eluted with a 60-min linear gradient of 12–14% (v/v) acetonitrile in 100 mM TEAA, pH 7.0, 0.1 mM diammonium phosphate at a flow rate of 200 µl/min at 60 °C while monitoring the absorbance of the eluate at 260 nm[55].

RNA (around 50 ng) was digested with RNase T1 (20 ng) in 100 mM triethylammonium acetate buffer (pH 7.0) at 37 °C for 1 h. The RNA fragments were separated using a direct nanoflow LC–MS system as described[56,57]. In brief, the digests were injected onto a reversed-phase Develosil C30-UG tip column (150 µm i.d. × 120 mm, 3-µm particle size; Nomura Chemical Co.) equilibrated with solvent A (10 mM TEAA, pH 7, in water:methanol, 9:1). Samples were eluted at 100 nl/min with a 60-min 0–24.5% linear gradient of solvent B (10 mM TEAA, pH 7:acetonitrile, 60:40). The column was subsequently washed with 70% B for 10 min and re-equilibrated with A.

Each LC eluate was sprayed online at −1.4 kV with the aid of a spray-assisting device[57] into a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific) operating in the negative ion mode and in the data-dependent mode to automatically switch between MS and tandem MS acquisition. Full-scan mass spectra (m/z = 480–1,980) were acquired at a mass resolution of

350,000. At most, the five most intense peaks, (>100,000 counts per second with a 60-ms maximum injection time), were isolated within a 3-*m*/*z* window for fragmentation. Precursors were fragmented by switching to a higher energy collision-induced dissociation mode with a normalized collision energy of 20 or 50%. To retain mass resolution and to increase spectral quality, three tandem mass spectral micro-scans were acquired for each sample. A fixed starting value of *m*/*z* = 100 was set for each tandem mass spectrum.

Interpretation of the tandem mass spectra and quantification of modifications

Ariadne[58] (http://ariadne.riken.jp/) was used for assignment of the tandem mass spectral peaks in conjunction with the sequence of rRNAs of *T. kodakarensis* (Gene ID: 3253116, 3253120 and 3253121). The Ariadne search parameters were: the maximum number of missed cleavages was one; two methylations per RNA fragment at any residue position were allowed; an RNA mass tolerance of ±20 ppm and a tandem spectral tolerance of ±50 ppm were allowed.
The quantification of post-transcriptional modification (PTM) was performed by the peak-area-based method. The target oligonucleotide peaks were obtained from extracted-ion chromatograms with their theoretical mass values (±5 ppm). Each peak area was measured using the Xcalibur software (Thermo Fisher Scientific) including a manual determination of the start-end of the peak. The stoichiometry of PTM was calculated from the peak areas obtained by MS with the following equation, in which *P* and *N* refer to peak areas of the oligonucleotide with or without PTM, respectively. Stoichiometry (%) = 100 × *P*/(*P* + *N*).

Proteomic analysis of *T. kodakarensis*

Proteins isolated from cultures of *T. kodakarensis* were precipitated by trichloroacetic acid (TCA) and washed twice with cold acetone before digestion for MS analysis. In brief, TCA-precipitated proteins were resuspended in 100 mM Tris pH 8.5 containing 8 M urea. Cysteine residues were reduced by 5 mM TCEP for 30 min at room temperature and further modified by

2-chloroacetamide for 30 min in the dark at room temperature. Proteins were first digested by recombinant Lys-C (Promega) overnight at 37 °C with shaking. The urea was diluted to 2 M before additional digestion overnight at 37 °C by the addition of trypsin at a ratio of 1:100 enzyme to substrate (Promega). The digestion reaction was quenched with the addition of formic acid to 5% final concentration. Peptides were quantified by the Pierce Colorimetric Peptide Assay (Thermo Scientific) and diluted in buffer A (5% acetonitrile (ACN), 0.1% formic acid (FA)) such that 1 µg was analysed per technical replicate. Each sample was trapped on an Acclaim PepMap 100 C18 column (5 µm particles 0.3 mm × 5 mm, Thermo Scientific) using the Ultimate 3000 autosampler (Dionex). Using chromatography conditions previously optimized[59], peptides were separated on an in-house-packed reverse phase chromatography column (1.9 µm particles (ReproSil, Dr. Maish), 75 µm × 20 cm), directly interfaced to a QExactive Plus (QE+) mass spectrometer (Thermo Scientific). Peptides were eluted over a quick gradient from 2–7% buffer B (80% ACN, 0.1% formic acid) in 10 min before the gradient was gradually increased to 40% buffer B over 6 h before ramping to 95% B in 15 min. The flow was kept at 95% B for 15 min before 20 min of re-equilibration at 2% B before the next injection. Flow rate was 180 nl/min. The application of a 2.5 kV distal voltage electrosprayed the eluting peptides directly into the QE+ mass spectrometer equipped with the Nanospray Flex source (Thermo Scientific). Full MS spectra were recorded on the eluting peptides at a resolving power of 70,000 over a 400 to 1,600 *m*/*z* range, followed by higher energy dissociation fragmentation at 30% normalized collision energy on the 15 most intense ions selected from the full MS spectrum. MS^2 spectra were collected in the Orbitrap at a resolving power of 17,500. Dynamic exclusion was enabled for 30 seconds[60]. Mass spectrometer scan functions and HPLC solvent gradients were controlled by the XCalibur data system (Thermo Scientific).

RAW files were extracted into .ms2 file format[61,62] using RawDistiller v. 1.0, in-house developed software[61]. RawDistiller D(g, 6) settings were used to abstract MS1 scan profiles by Gaussian

fitting and to implement dynamic offline lock mass using six background polydimethylcyclosiloxane ions as internal calibrants[61]. MS/MS spectra were first searched using ProLuCID[63] with a 10 ppm mass tolerance for peptide and 25 ppm tolerance for fragment ions. Trypsin specificity was imposed on both ends of candidate peptides during the search against a protein database containing 2,301 *T. kodakarensis* proteins (NCBI 2018–11-09 release), as well as 386 usual contaminants such as human keratins, IgGs and proteolytic enzymes. To estimate false discovery rates (FDR), each protein sequence was randomized (keeping the same amino acid composition and length) and the resulting 'shuffled' sequences were added to the database, for a total search space of 5,440 amino acid sequences. A mass of 15.9949 Da was differentially added to methionine residues.

DTASelect v.1.9[64] was used to select and sort peptide/spectrum matches (PSMs) passing the following criteria set: PSMs were only retained if they had a DeltCn of at least 0.08; minimum XCorr values of 1.0 for singly-, 1.4 for doubly-, and 2.1 for triply-charged spectra; peptides had to be at least 7 amino acids long. Results from each sample were merged and compared using CONTRAST[64]. Combining all replicate injections, proteins had to be detected by at least 2 peptides and/or 2 spectral counts. Proteins that were subsets of others were removed using the parsimony option in DTASelect on the proteins detected after merging all runs. Proteins that were identified by the same set of peptides (including at least one peptide unique to such protein group to distinguish between isoforms) were grouped together, and one accession number was arbitrarily considered as representative of each protein group.

NSAF7[65] was used to create the final reports on all detected peptides and non-redundant proteins identified across the different runs. Spectral and protein level FDRs were, on average, 0.17 ± 0.05% and 0.18 ± 0.05%, respectively.

Cryo-EM data acquisition and analysis

3.5 µl of 70S ribosome sample (0.25 mg/ml for the wild-type strains grown at 85 °C and 65 °C, and 0.4 mg/ml for the ac^{4}C-deficient strains) was applied on glow-discharged holey carbon grids (Quantifoil R2/2) coated with a thin layer of continuous carbon film. Grids were blotted (3 s) and plunge-frozen using a Vitrobot Mark IV (FEI, Thermo Fisher Scientific). Micrographs were recorded at liquid nitrogen temperature on a Titan Krios electron microscope (FEI, Thermo Fisher Scientific) operating at 300 kV and equipped with a Falcon 3 direct electron detector (FEI, Thermo Fisher Scientific). Nominal magnification used was 96K and corresponded to a calibrated pixel size of 0.85 Å per pixel, with a dose rate of approximately 1.16 $e^-/Å^2/s$ and defocus values ranging from −0.5 to −1.5 µm. Automatic data acquisition was performed using EPU (FEI, Thermo Fisher Scientific) and yielded a total of 2,509 micrographs for the WT85 (wild-type grown at 85 °C), 3,115 for the WT65 and 4,211 for the mutant. Micrographs were processed using MotionCor2[66] to correct for patched frame motion and dose-weighting and contrast transfer function parameters were estimated by CTFFIND 4.1[67,68]. Particle picking, extraction and classifications were performed using Relion 3.0[69]. The 60-Å low-pass-filtered cryo-EM map of the *P. furiosus* ribosome (EMD-2009) was used as an initial reference and has been used for further particle classification in 3D. Final maps reconstructed from 53,737, 283,424 and 116,586 particles for the WT85, WT65 and mutant strains, respectively, were obtained through multibody refinement with the LSU, the SSU body and SSU head masked individually as demonstrated in Extended Data Fig. 7a[69]. Density maps were corrected for the modulation transfer function of the detector, and then sharpened by applying a negative B-factor that was estimated using automated procedures in Relion3[70]. Averaged map resolutions were 2.95 Å, 2.55 Å and 2.65 Å for the WT85, WT65 and Tk*NAT10*, respectively and were determined using the gold-standard FSC = 0.143 criterion as implemented in Relion3 and M-triage as implemented in Phenix[71] (Extended Data Fig. 7b–d). Local resolutions were estimated using Resmap[72] (Extended Data Fig. 8a).

Model building and refinement

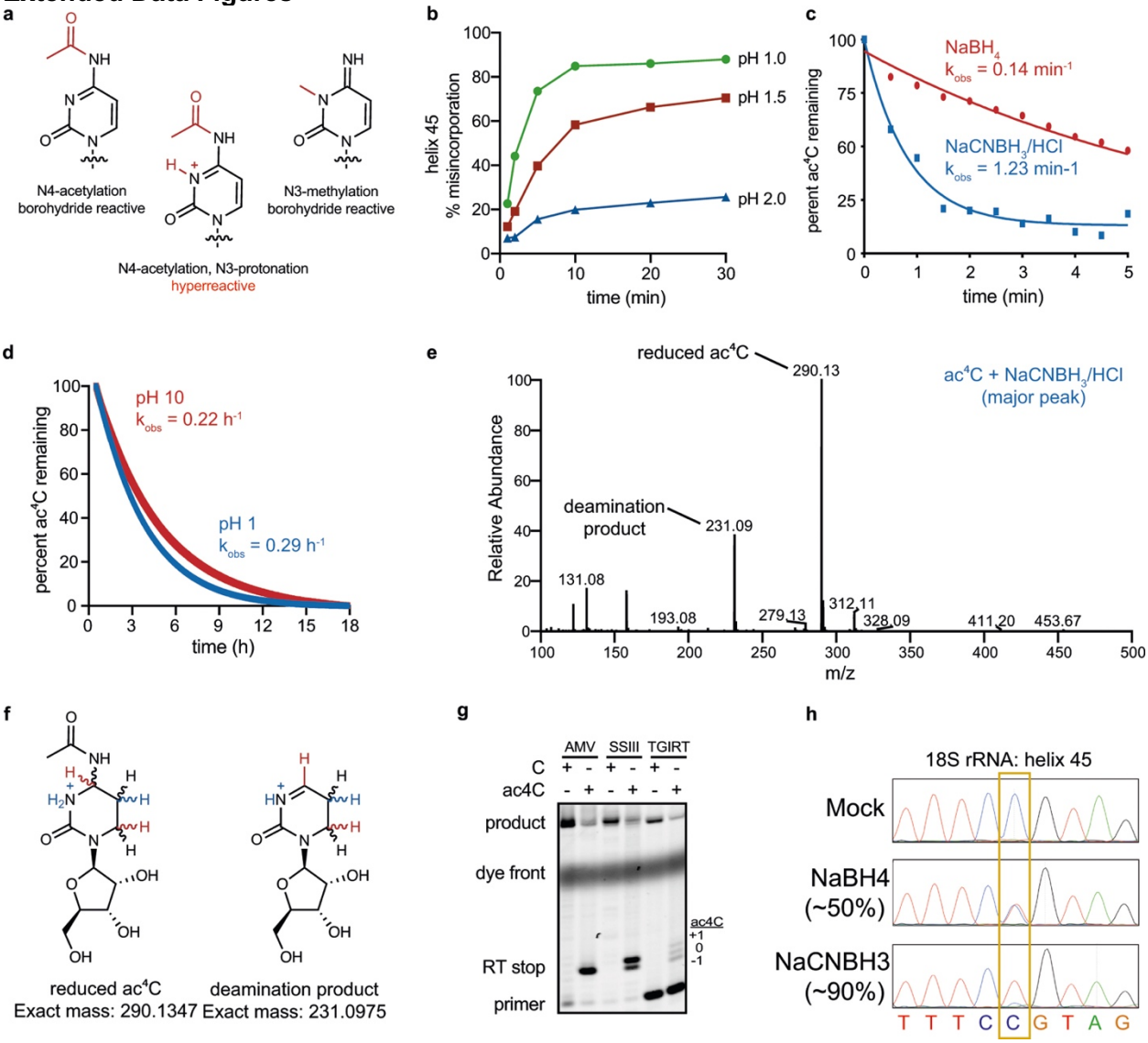
The initial template of *T. kodakarensis* ribosome was derived from the cryo-EM model of *P. furiosus* ribosome (PDB 4V6U). The model was docked into the electron microscopy density maps using UCSF Chimera[73], followed by iterative manual building in Coot[74]. Coordinates and library files for the modified residues were generated through phenix.elbow[75] and were manually docked into the relevant positions using Coot followed by real-space refinement. The final model was subjected to global refinement and minimization in real space using phenix.real_space_refine in Phenix[71]. MolProbity[76] was used to evaluate model geometry. The final refinement parameters are provided in Supplementary Table 6, and map versus model diagrams are in Extended Data Fig. 7e–g. Examples of the ac^4C model in density view are shown in Extended Data Fig. 8b.

Biophysical characterization of ac^4C containing hairpins

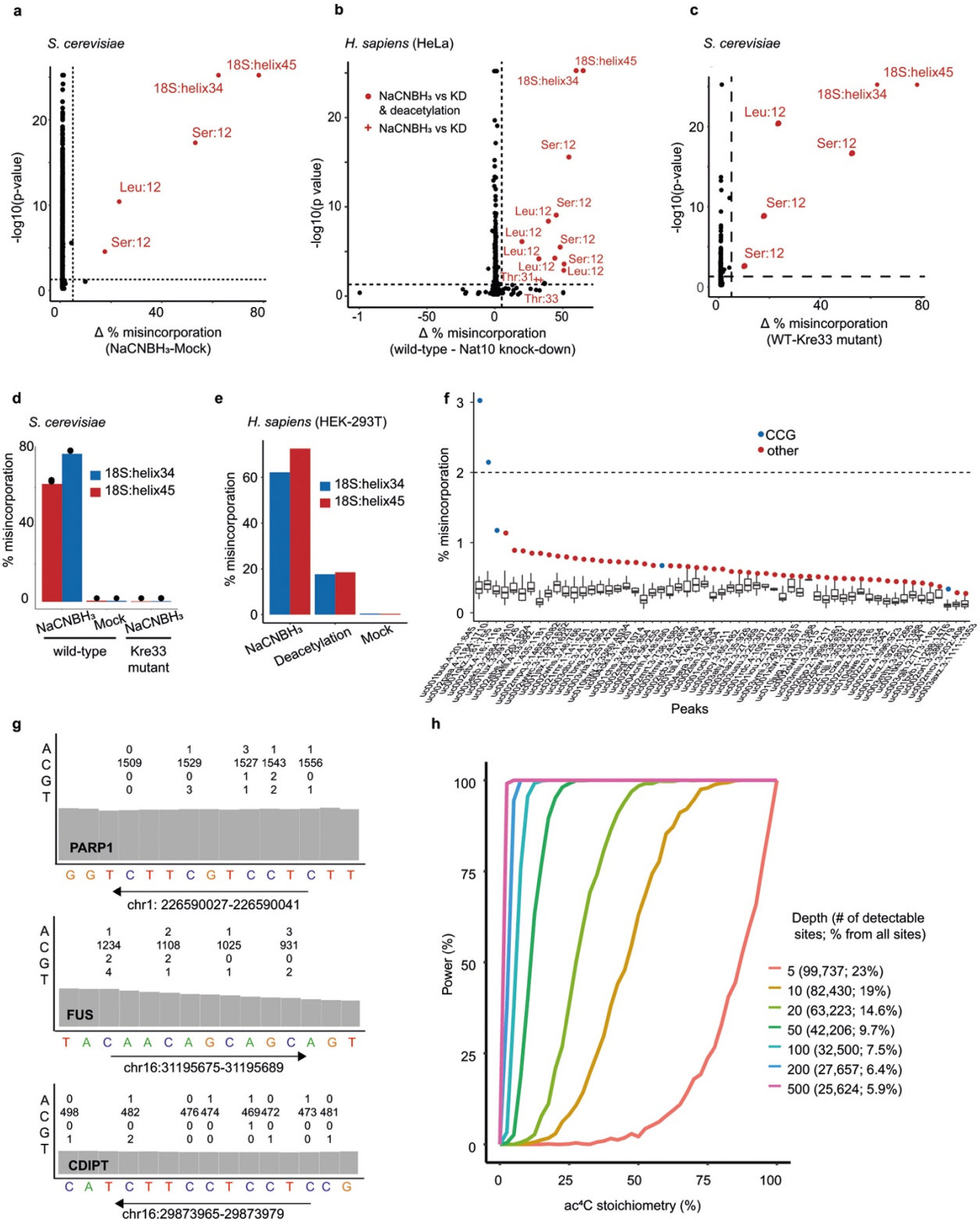
In vitro transcription was performed using the NEB Highscribe T7 highyield RNA synthesis kit according to the manufacturer's instructions using DNA templates containing a T7 promoter upstream of a template sequence (Supplementary Table 1a). For ac^4C-containing transcripts, CTP was replaced in the reaction mixture with ac^4CTP (50 mM). Crude in vitro transcription reactions were purified by denaturing polyacrylamide gel electrophoresis (PAGE). Full length product bands were visualized by UV shadowing and excised with a razor blade. RNA was extracted by crushing the gel slices and shaking in 500 mM ammonium acetate with 0.2 mM EDTA pH 8.0. RNA was desalted by four sequential rounds of dilution and concentration in a 1K MWCO centrifugal ultrafiltration device. Before use in DSC and circular dichroism experiments, purified RNAs were analysed for purity by denaturing PAGE and visualized using SYBR Gold Nucleic Acid Gel Stain from Invitrogen. DSC experiments were carried out on a VP-DSC instrument (Microcal). Desalted PAGE purified helix-45 oligos were diluted to 18 μM in 1X Oligo

DSC buffer (10 mM phosphate buffer, 50 mM NaCl) and folded by heating to 95 °C for 10 min and rapidly cooled by placing on ice for 10 min. Samples were vacuum degassed with stirring for 8 min at 35 °C. DSC was equilibrated with 550 µl freshly degassed 1x Oligo DSC buffer in sample and reference cells through multiple scan cycles until a stable and flat differential heat flow curve was established. During downscanning, the sample cell was emptied, and 550 µl freshly degassed helix-45 hairpins were loaded between 40 °C and 35 °C. Samples were equilibrated at 35 °C for 15 min and calorimetric data was collected from 35 °C to 120 °C at a scan rate of 1 °C/min. Raw DSC data from each scan was processed by linear baseline subtraction and the absolute value of each baseline was adjusted to allow curves to be observed on a single plot. Melting temperatures were calculated as the mean value of the local maxima of the major transition on each scan (n = 3) and errors were calculated as the standard deviation. CD analyses were performed on a JASCO J-1500 CD Spectrometer using a 1 mm pathlength quartz cuvette. In brief, desalted helix-45 oligos were diluted to 5 µM in 1X melting buffer (100 mM NaCl, 1.97 mM KCl, 0.1 mM EDTA and 8.7 mM sodium phosphate (pH 7.4)) and folding by fast cooling. Denaturation curves were recorded by monitoring the change in ellipticity at 260 nm while the temperature was increased from 30 °C to 95 °C at a rate of 2 °C/min. The minimum points in the first-derivative curves of CD melting spectra were recorded (n = 3) and errors were calculated as standard deviation.

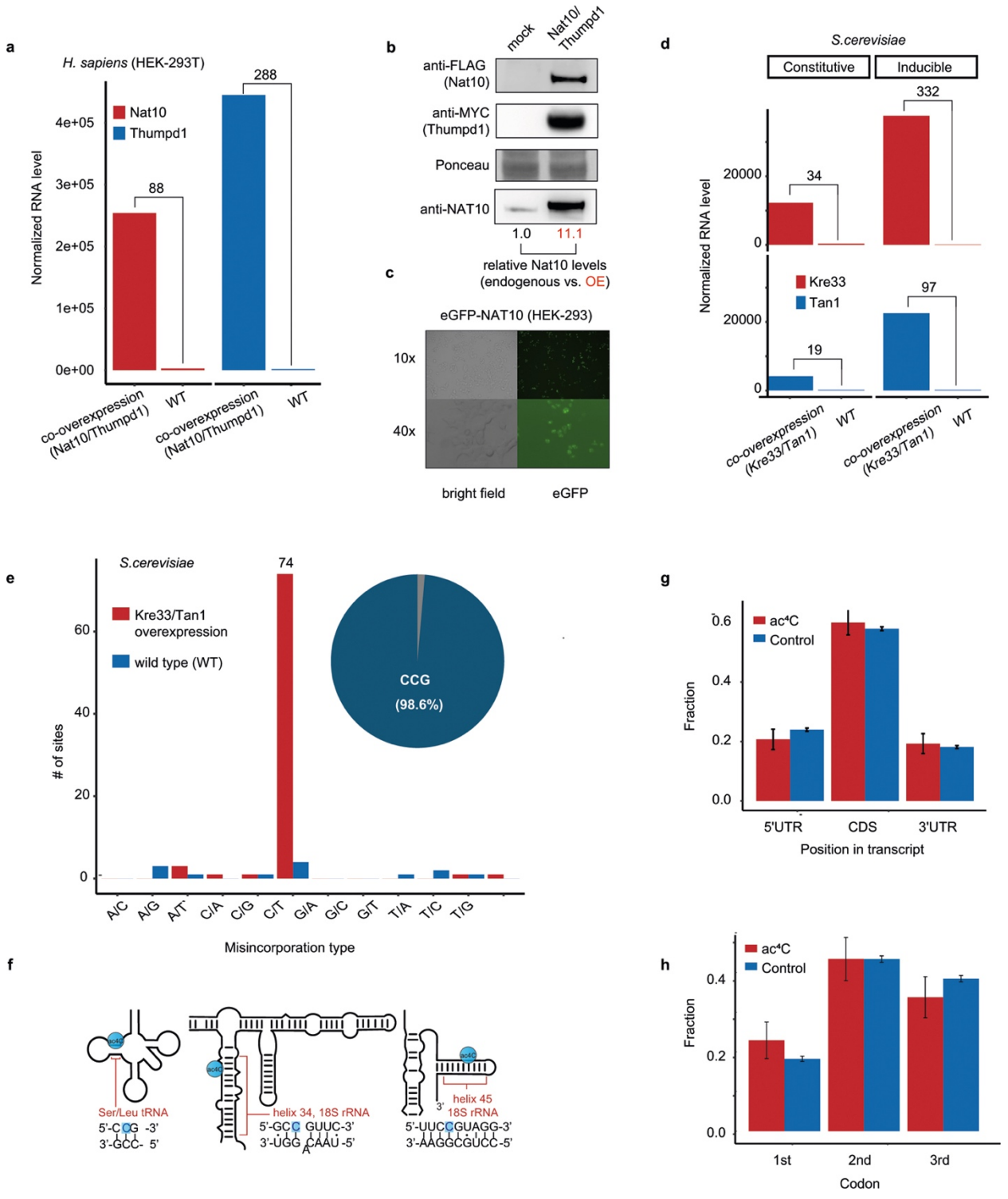
Extended Data Figures



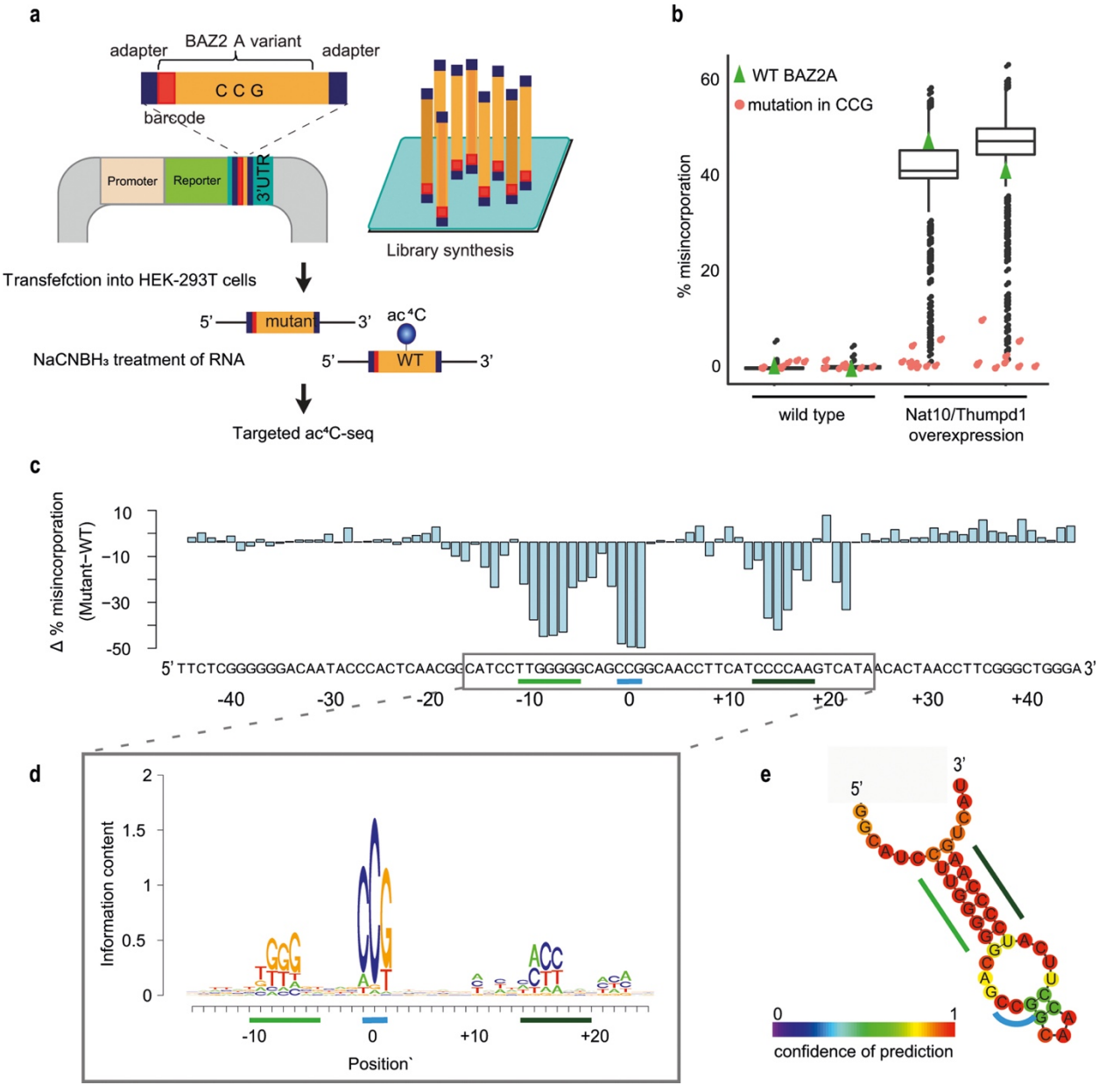
Extended Data Fig. B.1 An optimized reaction for sequencing of N^4-acetylcytidine in RNA. a, Protonation under acidic conditions hyperactivates ac^4C, increasing its reactivity with $NaCNBH_3$. Efficient reduction manifests as quantitative misincorporation of deoxynucleotide triphosphates at ac^4C upon reverse transcription. **b**, $NaCNBH_3$-dependent misincorporation at the known ac^4C site in human helix 45 is increased at more acidic pH. The percentage misincorporation at ac^4C sites after chemical reduction, reverse transcription and PCR was quantified from Sanger sequencing data. One independent experiment. **c**, Kinetic analysis of ac^4C reduction. Reaction progress was assessed by monitoring the disappearance of ac^4C absorbance at 300 nm in the presence of first and second-generation hydride donors. Reaction conditions: ac^4C (0.1 mM, free nucleoside), reductant (20 mM), H_2O. $NaBH_4$ reactions were carried out at pH 10, whereas $NaCNBH_3$ reactions were adjusted to pH 1 using HCl before initiation. Representative of 3 independent experiments. **d**, Kinetic analysis of the hydrolysis of ac^4C at pH values used in $NaBH_4$ (pH 10) and $NaCNBH_3$ (pH 1) reduction reactions. Reaction progress was assessed by monitoring the disappearance of ac^4C absorbance at 300 nm. Acid- and base-catalysed hydrolysis occurred at similar rates, and were slow compared to ac^4C reduction by $NaBH_4$ and $NaCNBH_3$. Representative of 3 independent experiments. **e**, LC–MS/MS analysis confirms reduction of ac^4C to reduced ac^4C in the presence of $NaCNBH_3$. Reaction conditions: ac^4C (0.1 mM, free nucleoside), $NaCNBH_3$ (20 mM), HCl pH 1. Representative of 2 independent experiments. **f**, Exact mass of reduced ac^4C and deamination product observed in LC–MS/MS experiments. **g**, Primer extension analysis of ac^4C-containing RNAs after $NaCNBH_3$ treatment (100 mM, pH 1, 37 °C, 1 h). **h**, Sanger sequence traces of a known ac^4C site in helix 45 of human HAP1 cells. C>T misincorporation is exclusively observed at the ac^4C site in reduced ($NaBH_4$ and $NaCNBH_3$) but not in mock-treated samples. ac^4C sites are highlighted in yellow.


a
S. cerevisiae
18S:helix45
18S:helix34
Ser:12
Leu:12
Ser:12
-log10(p-value)
Δ % misincorporation
(NaCNBH₃-Mock)
b
H. sapiens (HeLa)
NaCNBH₃ vs KD
& deacetylation
NaCNBH₃ vs KD
-log10(p value)
Δ % misincorporation
(wild-type - Nat10 knock-down)
c
S. cerevisiae
Leu:12
-log10(p-value)
Δ % misincorporation
(WT-Kre33 mutant)
d
S. cerevisiae
% misincorporation
18S:helix34
18S:helix45
NaCNBH₃
Mock
wild-type
Kre33
mutant
e
H. sapiens (HEK-293T)
% misincorporation
Deacetylation
f
CCG
other
% misincorporation
Peaks
g
PARP1
chr1: 226590027-226590041
FUS
chr16:31195675-31195689
CDIPT
chr16:29873965-29873979
h
Power (%)
ac⁴C stoichiometry (%)
Depth (# of detectable
sites; % from all sites)
5 (99,737; 23%)
10 (82,430; 19%)
20 (63,223; 14.6%)
50 (42,206; 9.7%)
100 (32,500; 7.5%)
200 (27,657; 6.4%)
500 (25,624; 5.9%)

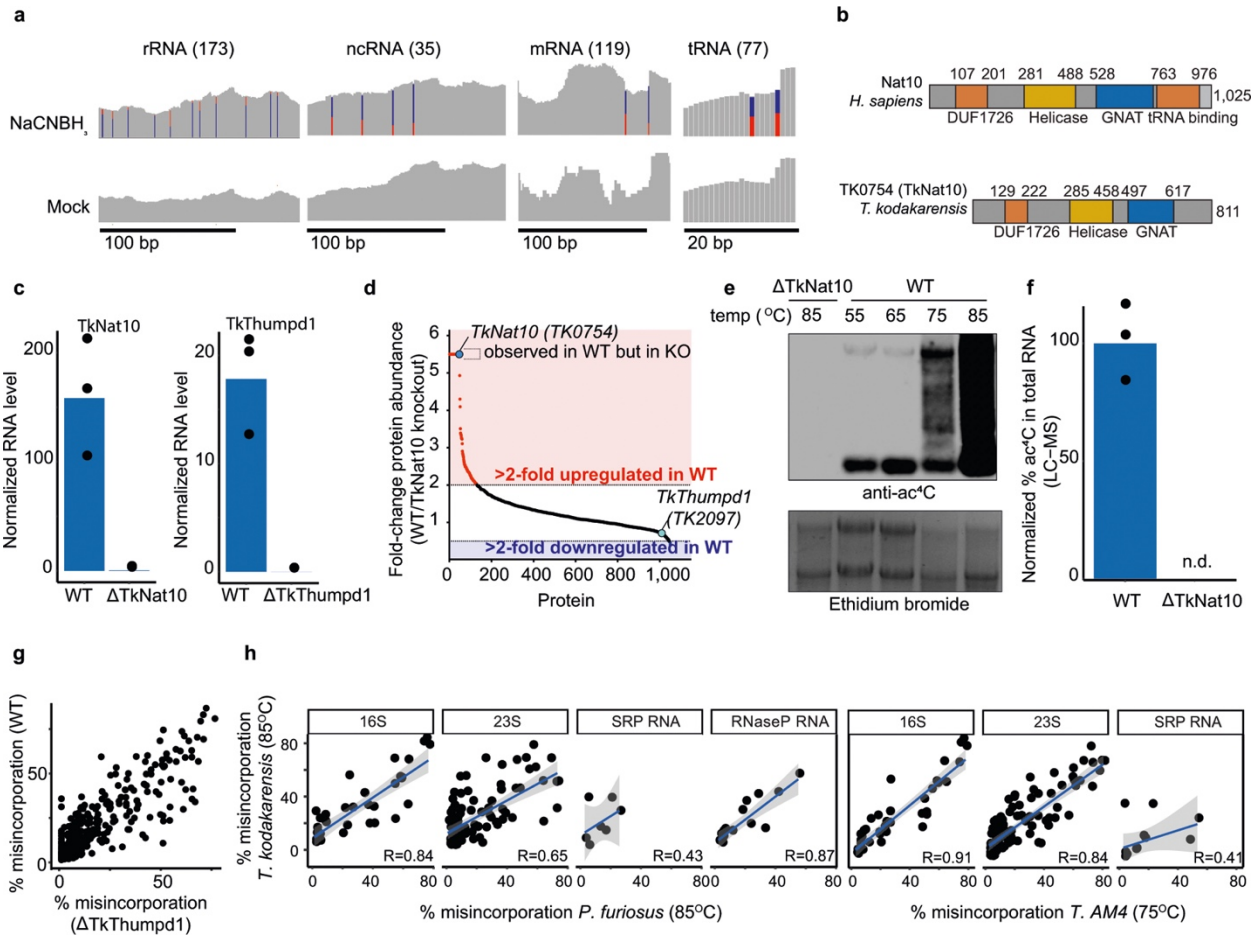
Extended Data Fig. B.2 Ac^4C in eukaryotic cells with wild-type NAT10 expression. **a–c**, ac^4C-seq was conducted on RNA from *S. cerevisiae* (**a, c**) and HeLa cells (**b**). Statistical significance from the χ^2 test is plotted against the difference in misincorporation rates (corresponding to ac^4C levels) between $NaCNBH_3$-treated and mock-treated RNA from *S. cerevisiae* (**a**), RNA from wild-type and NAT10-depleted cells (**b**) or from wild-type *S. cerevisiae* cells and a strain expressing a catalytic mutant of Kre33 (**c**), treated with $NaCNBH_3$. Sites with a differential misincorporation level >5% and a *P* value <0.05 are labelled and marked in red. For HeLa cells (**b**) an additional comparison between $NaCNBH_3$ and deacetylation pre-treatment was conducted. Sites that do not pass significance under these conditions are marked with a plus sign (shown only for sites found significant between $NaCNBH_3$ and mock treatment). Significant sites are labelled with the identity of the molecule and the relative position (or helix) of ac^4C. *n* = 3 biologically independent samples for all but NAT10-depleted HeLa cells, in which case *n* = 2 biologically independent samples. **d, e,** Misincorporation level in the two known sites in 18S (helix 34 and helix 45), compared with controls in poly(A)-enriched RNA from wild-type *S. cerevisiae* cells and *S. cerevisiae* cells expressing a catalytic mutant of Kre33 (**d**) and from wild-type HEK-293T cells and HEK-293T cells overexpressing NAT10 and THUMPD1 (**e**). **f, g,** ac^4C-seq data from poly(A)-enriched RNA from HEK-293T cells overexpressing NAT10 and THUMPD1 on 'ac^4C peaks' that have been identified previously[10] as harbouring ac4C. **f,** Distribution of misincorporation across each of 57 'ac^4C peaks' that had a coverage of more than 400 reads in more than 80% of the cytosines in the peak. For each peak the cytosine harbouring the highest misincorporation rate is indicated in colour, presented in blue if it harbours a CCG motif and red otherwise. **g**, Traces from the Integrative Genomic Viewer (IGV) browser of three such genes, with highest coverage in the ac^4C-seq data. For each gene the 15 bases motif identified in ref. [10] is presented. The numbers above each cytidine indicate the number of bases (A, C, G and T) observed in our data at that position. **h,** Power analysis for ac^4C detection, as a function of sequencing depths and stoichiometries. Each data point in each curve is based on 1,000 simulations. For each sampled depth, numbers in the legend indicate the sequencing depth, which was kept identical for treatment and control samples. In addition, the legend indicates the number of CCG sites found in wild-type HEK-293T samples that have such a minimal depth and the percentage of these detectable CCG sites from all CCG sites in the transcriptome.


a
H. sapiens (HEK-293T)
Nat10
Thumpd1
Normalized RNA level
4e+05
3e+05
2e+05
1e+05
0e+00
88
288
co-overexpression (Nat10/Thumpd1)
WT
co-overexpression (Nat10/Thumpd1)
WT
b
mock
Nat10/ Thumpd1
anti-FLAG (Nat10)
anti-MYC (Thumpd1)
Ponceau
anti-NAT10
1.0
11.1
relative Nat10 levels (endogenous vs. OE)
c
eGFP-NAT10 (HEK-293)
10x
40x
bright field
eGFP
d
S.cerevisiae
Constitutive
Inducible
Normalized RNA level
20000
0
34
332
Kre33
Tan1
20000
0
19
97
co-overexpression (Kre33/Tan1)
WT
co-overexpression (Kre33/Tan1)
WT
e
S.cerevisiae
Kre33/Tan1 overexpression
wild type (WT)
# of sites
60
40
20
0
74
CCG
(98.6%)
A/C
A/G
A/T
C/A
C/G
C/T
G/A
G/C
G/T
T/A
T/C
T/G
Misincorporation type
f
Ser/Leu tRNA
5'-CCG -3'
3'-GCC- 5'
helix 34, 18S rRNA
5'-GCC GUUC-3'
3'-UGG CAAU -5'
A
helix 45
18S rRNA
3'
5'-UUCCGUAGG-3'
3'-AAGGCGUCC-5'
g
ac4C
Control
Fraction
0.6
0.4
0.2
0.0
5'UTR
CDS
3'UTR
Position in transcript
h
ac4C
Control
Fraction
0.4
0.2
0.0
1st
2nd
3rd
Codon

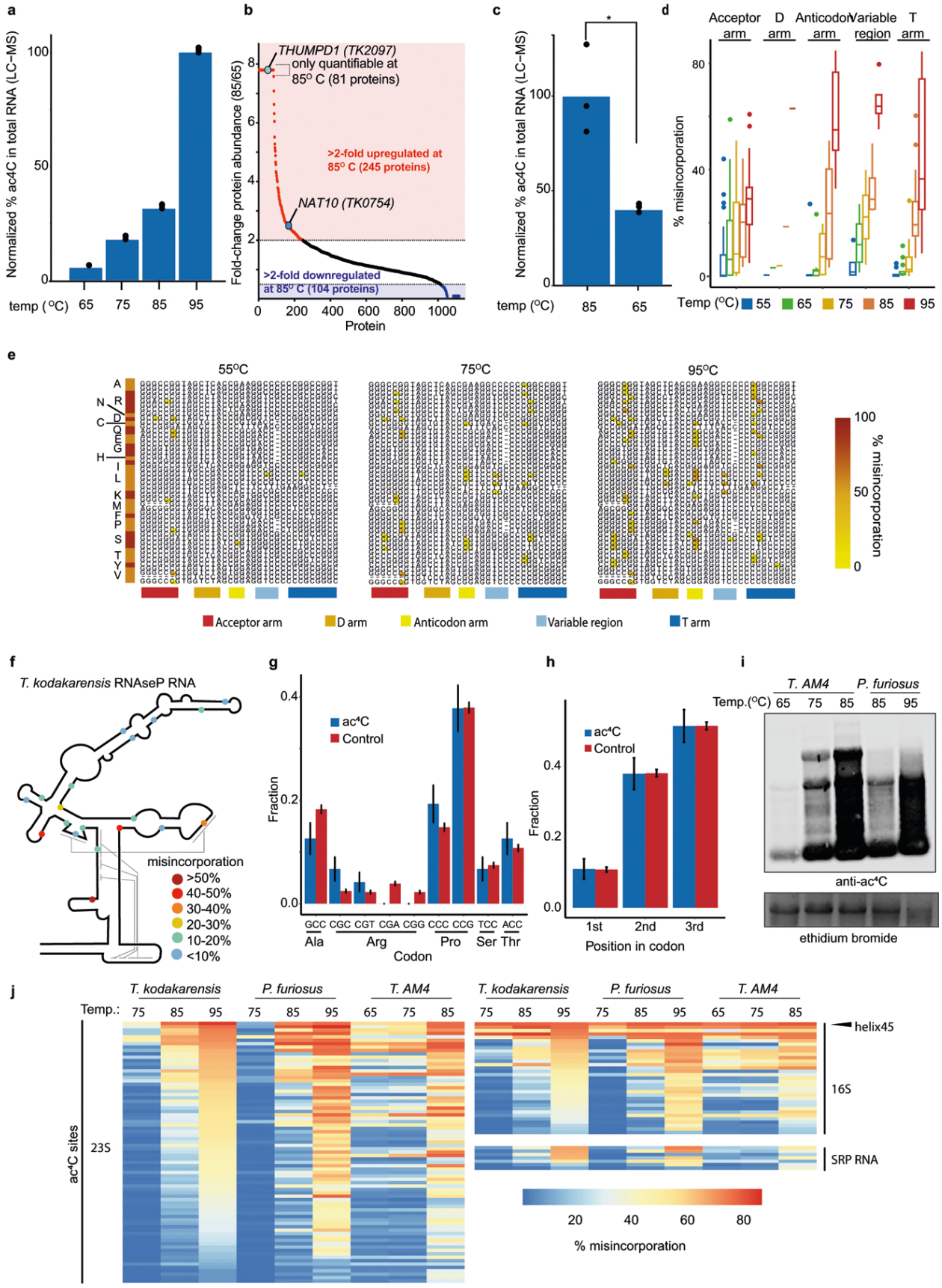
Extended Data Fig. B.3 Ac^4C in eukaryotic cells with manipulated NAT10 expression. **a**, RNA expression of *NAT10* and *THUMPD1* in HEK-293T cells overexpressing both genes compared to wild-type cells. Shown are TMM-normalized read counts. The numbers above the bars indicate fold increase compared with the wild type. **b**, Immunoblotting analysis of NAT10 and THUMPD1 overexpression in HEK-293T cells. Representative of 3 independent experiments with similar results. For gel source data, see Supplementary Data B.3. **c**, Microscopy images of the eGFP–NAT10 construct, confirming nuclear and nucleolar localization of ectopically expressed N-terminally tagged protein. Representative of 3 independent experiments with similar results. **d**, RNA expression of *Kre33* and *Tan1* in wild-type yeast cells and in cells stably overexpressing *Kre33* and either stably or inducibly overexpressing *Tan1*. The numbers above the bars indicate fold increase from the wild type. **e**, The number of sites displaying each of the 12 possible misincorporation patterns are displayed (bar plot, *y* axis) for sites found in poly(A)-enriched RNA from both wild-type *S. cerevisiae* cells and *S. cerevisiae* cells overexpressing both *Kre33* and *Tan1*. The pie chart displays the proportion of sites harbouring C>T misincorporations that were embedded within a C**C**G motif (73 out of 74, 98.6%). **f**, Schematic of the known ac^4C sites in human tRNAs (Leu and Ser) and in helix 34 (C1337) and helix 45 (C1842) of human 18S rRNA. The acetylated cytidine residue (highlighted in blue) is embedded within a C**C**G motif in all known sites. **g**, Fraction of ac^4C sites found within the 5′ UTR, CDS and 3′ UTR (CDS, coding sequence; UTR, untranslated region). Results are shown for the set of ac^4C sites in mRNA of HEK-293T cells overexpressing NAT10 and THUMPD1 (red bars, n = 139), and—as controls—for all CCG motifs present within all genes within which any ac^4C was found (blue bars, n = 6,129). Error bars representing standard distribution of the binomial distribution. Data are based on 2 biologically independent samples. **h**, Fraction of ac^4C sites at the first, second and third position of each codon, shown for ac^4C sites and controls as in **g**. Data are mean ±s.d. of the binomial distribution and are based on two biologically independent samples.


a
adapter
BAZ2 A variant
adapter
C C G
barcode
Promoter
Reporter
3'UTR
Library synthesis
Transfefction into HEK-293T cells
5'
mutan
3'
ac4C
NaCNBH3 treatment of RNA
5'
WT
3'
Targeted ac4C-seq
b
WT BAZ2A
mutation in CCG
% misincorporation
60
40
20
0
wild type
Nat10/Thumpd1
overexpression
c
Δ % misincorporation
(Mutant–WT)
10
−10
−30
-50
5' TTCTCGGGGGGACAATACCCACTCAACGGCATCCTTGGGGGCAGCCGGCAACCTTCATCCCCAAGTCATAACACTAACCTTCGGGCTGGGA 3'
-40
-30
-20
-10
0
+10
+20
+30
+40
d
Information content
2
1.5
1
0.5
0
-10
0
+10
+20
Position`
e
3'
5'
0
1
confidence of prediction

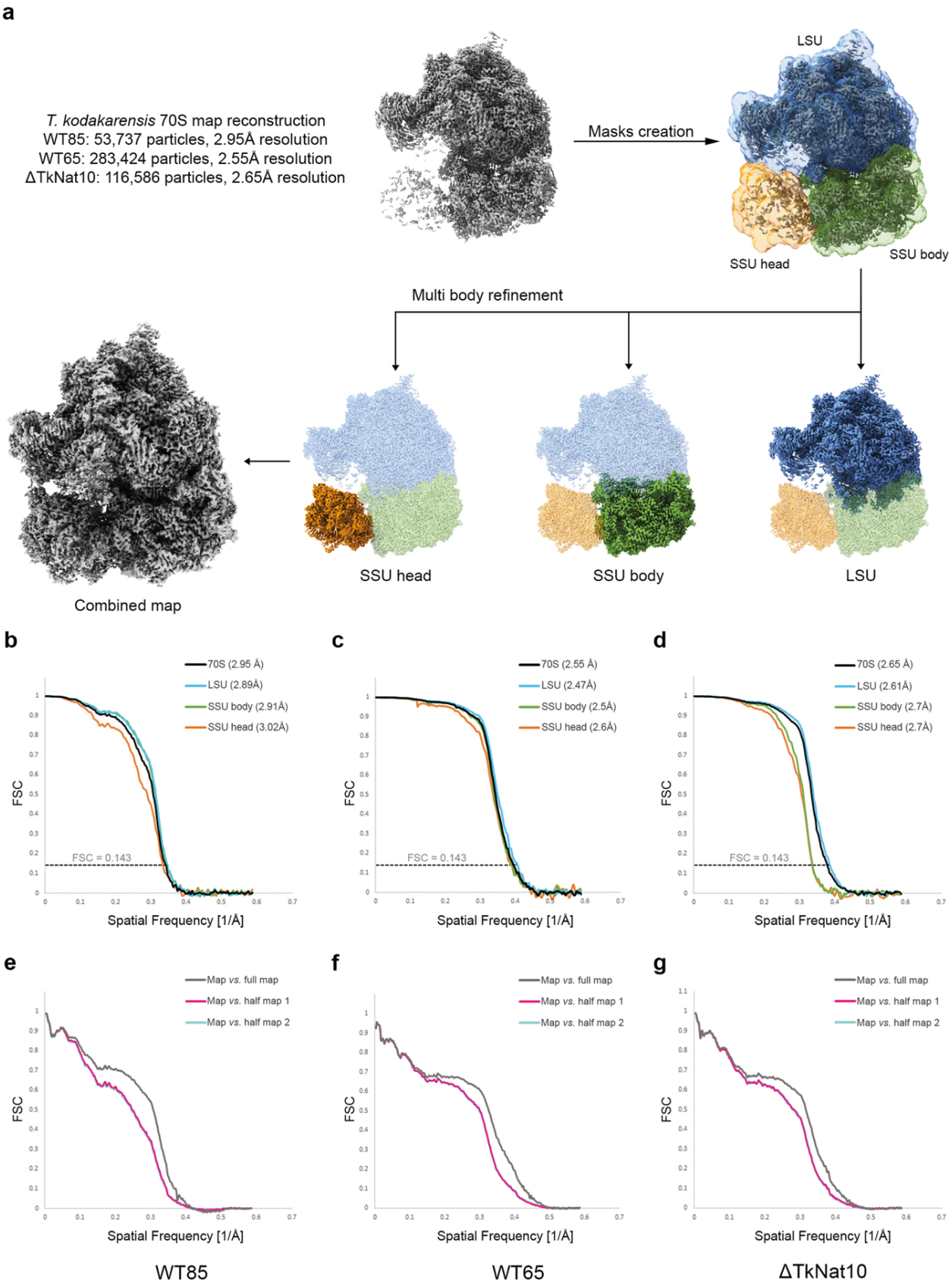
Extended Data Fig. B.4 Sequence and structure requirements for deposition of ac^4C. **a**, Oligonucleotides representing the wild-type sequence surrounding the acetylated site in *BAZ2A* mRNA, or variants with single mutations across the wild-type sequence, were synthesized as a pool and cloned into the 3′UTR of a reporter gene. The pool of plasmids was transfected into wild-type HEK-293T cells or cells transiently overexpressing NAT10 and THUMPD1. RNA extracted from cells was subjected to targeted ac^4C-seq and ac^4C levels were estimated on the basis of misincorporation rates. **b**, Misincorporation rate of oligonucleotides described in **a**, harbouring the wild-type sequence of *BAZ2A* (green triangles) or a sequence mutated at the CCG motif and at its surrounding bases (red and black, respectively). Box plot visualization parameters are as in Fig. 1h. $n = 2$ biologically independent samples. **c**, The difference in misincorporation rate of oligonucleotides with a single base mutation compared with the wild-type oligonucleotide is shown across all positions of the construct. **d**, De novo construction of the motifs surrounding the modified cytidine were built on the basis of the contribution of single-base mutations in the *BAZ2A* sequence to the reduction in misincorporation rate compared to wild-type *BAZ2A* sequence. **e**, Secondary structure of the *BAZ2A* mRNA fragment as predicted by RNAfold. Bases are colour-coded according to confidence level of the prediction. Regions highlighted by a blue and green line in **c**–**e** represent the CCG motif and a stem structure surrounding the modified cytidine, respectively.


a
rRNA (173)
ncRNA (35)
mRNA (119)
tRNA (77)
NaCNBH3
Mock
100 bp
100 bp
100 bp
20 bp
b
Nat10
H. sapiens
107 201 281 488 528 763 976
1,025
DUF1726 Helicase GNAT tRNA binding
TK0754 (TkNat10)
T. kodakarensis
129 222 285 458 497 617
811
DUF1726 Helicase GNAT
c
TkNat10
TkThumpd1
Normalized RNA level
WT ΔTkNat10
WT ΔTkThumpd1
d
TkNat10 (TK0754)
observed in WT but in KO
>2-fold upregulated in WT
TkThumpd1
(TK2097)
>2-fold downregulated in WT
Fold-change protein abundance
(WT/TkNat10 knockout)
Protein
e
ΔTkNat10
WT
temp (°C) 85 55 65 75 85
anti-ac4C
Ethidium bromide
f
Normalized % ac4C in total RNA
(LC–MS)
n.d.
WT ΔTkNat10
g
% misincorporation (WT)
% misincorporation
(ΔTkThumpd1)
h
% misincorporation
T. kodakarensis (85°C)
16S
23S
SRP RNA
RNaseP RNA
R=0.84
R=0.65
R=0.43
R=0.87
% misincorporation P. furiosus (85°C)
16S
23S
SRP RNA
R=0.91
R=0.84
R=0.41
% misincorporation T. AM4 (75°C)

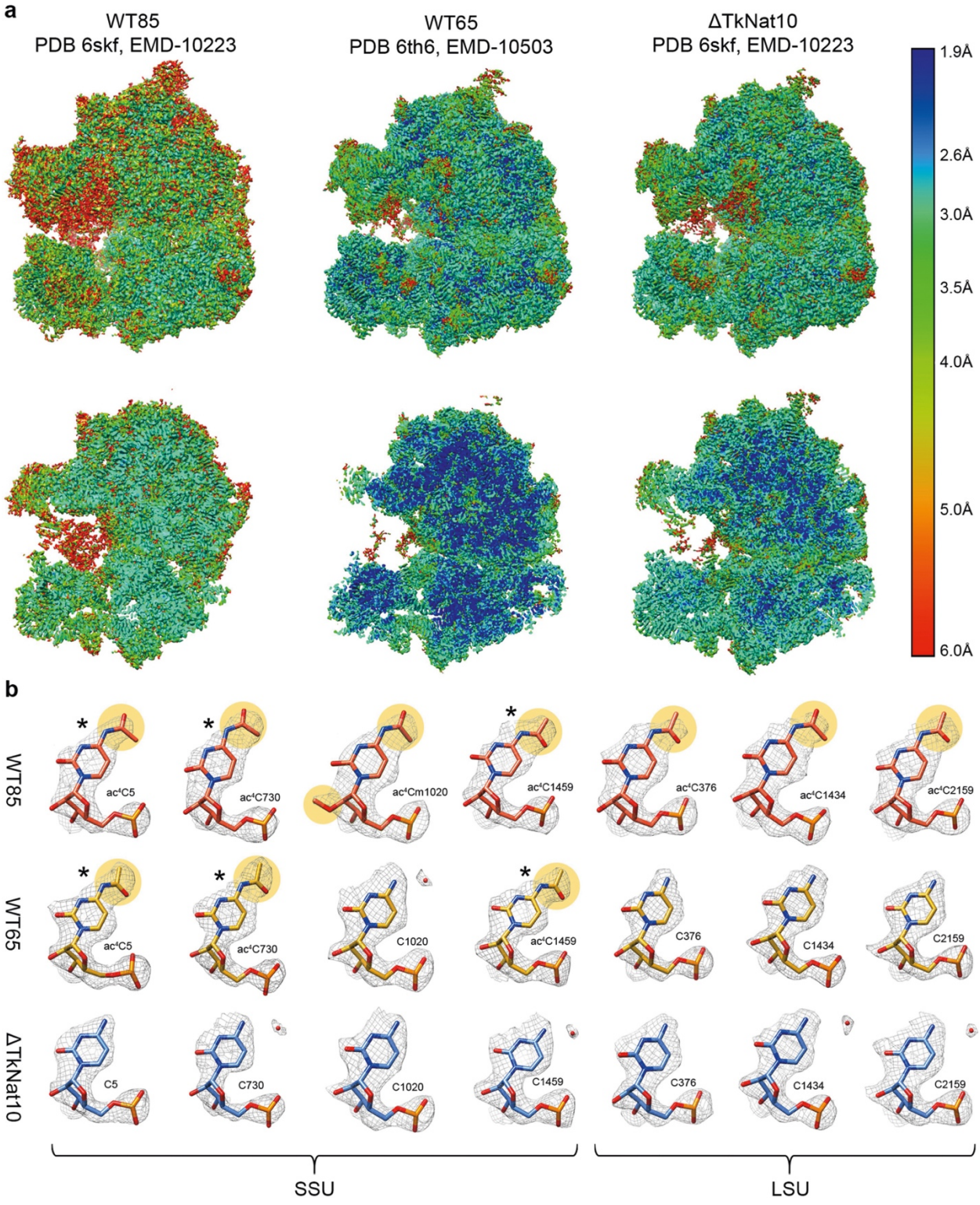
Extended Data Fig. B.5 Deletion of Tk*NAT10* and Tk*THUMPD1* in *T. kodakarensis*. **a**, Total RNA from *T. kodakarensis* was analysed via ac^{4}C-seq. IGV browser traces display representative ac^{4}C sites in rRNA, ncRNA, mRNA and tRNA of *T. kodakarensis*, visualized as in Fig. B.1c. The number in parentheses indicates the number of sites identified for each class of molecules. **b**, Conserved domain architecture of human *NAT10* and its homologue in *T. kodakarensis*, *TK0754* (referred to as Tk*NAT10* in the text). **c**, Expression of Tk*NAT10* and Tk*THUMPD1* (*TK2097*) in wild-type *T. kodakarensis* and the indicated deletion strains was quantified from ac^{4}C-seq data. Shown are mean TMM-normalized values (n = 3 and 2 biological replicates in wild-type and deletion strains, respectively). **d**, Quantitative LC–MS/MS proteomics analysis of wild-type and ΔTk*NAT10 T. kodakarensis*. Fold-change in protein abundance was based on comparison of distributed normalized spectral abundance factor for individual proteins. Fold-change for proteins detectable exclusively in the wild-type or the knockout (KO) condition (fold-change = ∞) are graphed at 5.5 and 0.1, respectively, which represents the maximum and minimum of measured values. n = 3 LC–MS/MS runs for each condition. **e**, Anti-ac^{4}C immuno-northern blot in *T. kodakarensis* total RNA. Ethidium bromide staining is used to visualize total RNA. Results are representative of two biological replicates. For gel source data, see Supplementary Data B.3. **f**, Relative quantification of ac^{4}C in total RNA isolated from wild-type and ΔTk*NAT10 T. kodakarensis* strains as measured by LC–MS. Mean of n = 3 technical replicates. n.d., not detectable. **g**, Scatter-plot depicting misincorporation rate of ac^{4}C sites in wild-type *T. kodakarensis* is compared with the Tk*THUMPD1*-deletion strain, showing no effect of the deletion of the gene on the ac^{4}C status. **h**, Correlation between misincorporation rates in *T. kodakarensis* compared to *P. furiosus* and *T*. sp. AM4 for the different types of ncRNAs identified by ac^{4}C-seq. The Pearson's correlation coefficient is indicated at the bottom of each plot. n = 4 and 1 independent biological samples for *T. kodakarensis* and other archaea, respectively. Shading indicates 95% confidence intervals for predictions from a linear model.


a
Normalized % ac4C in total RNA (LC–MS)
temp (°C) 65 75 85 95
b
Fold-change protein abundance (85/65)
THUMPD1 (TK2097)
only quantifiable at
85° C (81 proteins)
>2-fold upregulated at
85° C (245 proteins)
NAT10 (TK0754)
>2-fold downregulated
at 85° C (104 proteins)
Protein
c
Normalized % ac4C in total RNA (LC–MS)
temp (°C) 85 65
d
Acceptor arm
D arm
Anticodon arm
Variable region
T arm
% misincorporation
Temp (°C) 55 65 75 85 95
e
55°C
75°C
95°C
% misincorporation
Acceptor arm
D arm
Anticodon arm
Variable region
T arm
f
T. kodakarensis RNAseP RNA
misincorporation
>50%
40-50%
30-40%
20-30%
10-20%
<10%
g
ac4C
Control
Fraction
GCC CGC CGT CGA CGG CCC CCG TCC ACC
Ala Arg Pro Ser Thr
Codon
h
ac4C
Control
Fraction
1st 2nd 3rd
Position in codon
i
T. AM4
P. furiosus
Temp.(°C) 65 75 85 85 95
anti-ac4C
ethidium bromide
j
T. kodakarensis
P. furiosus
T. AM4
Temp.: 75 85 95 75 85 95 65 75 85
ac4C sites
23S
helix45
16S
SRP RNA
20 40 60 80
% misincorporation

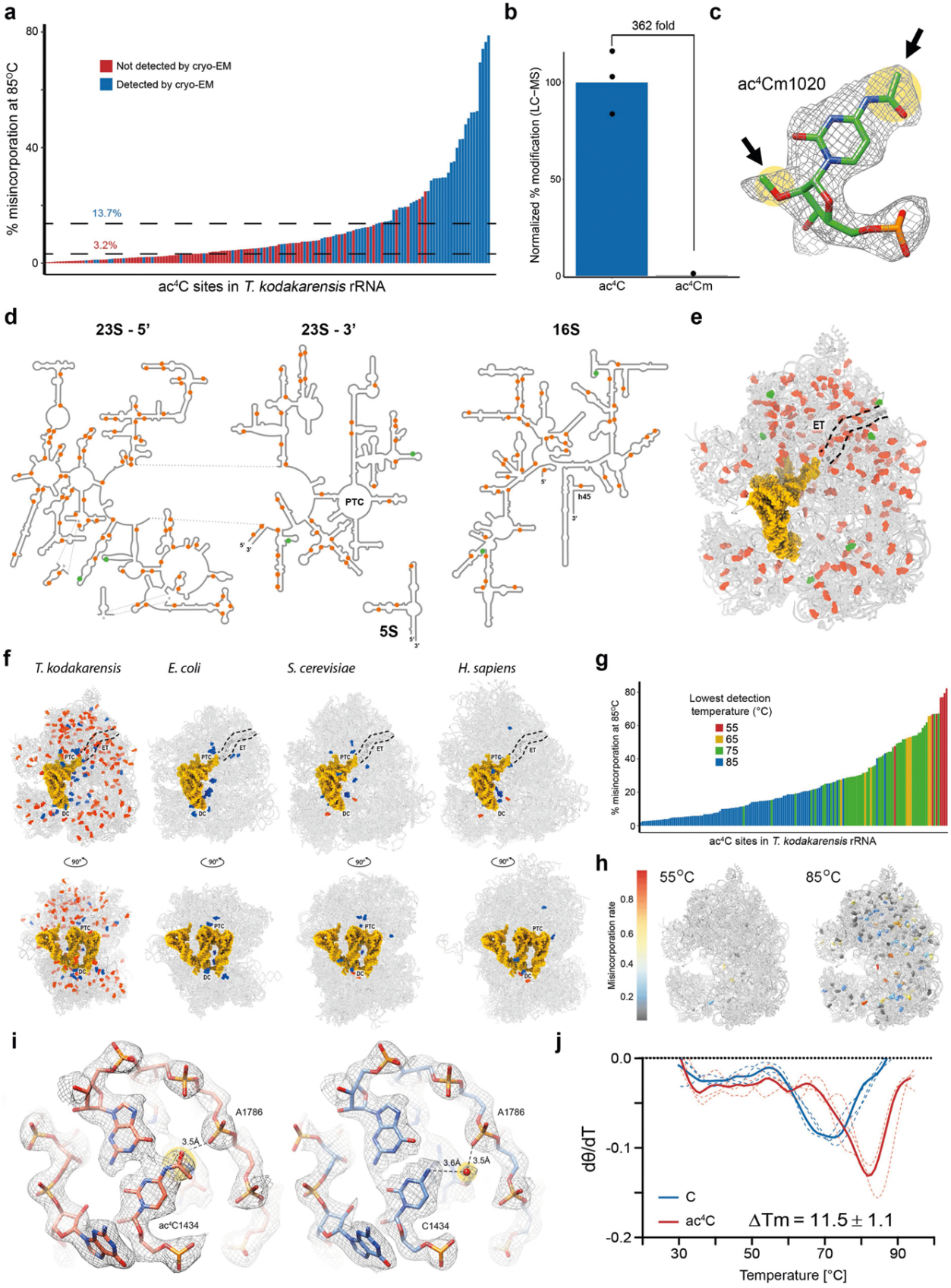
Extended Data Fig. B.6 Ac^{4}C accumulates in a temperature-dependent manner across all RNA species in archaea. a, Relative quantification of ac^{4}C in total RNA isolated from wild-type *T. kodakarensis* as a function of temperature as measured by LC–MS. Mean is shown along with individual data points. *n* = 3 technical replicates. For 65, 75, 85 °C: representative of 2 independent experiments, for 95 °C: 1 experiment. **b**, **c**, Quantitative LC–MS/MS proteomics analysis of *T. kodakarensis* temperature-dependent protein expression. Fold-change in *T. kodakarensis* protein abundance between growth conditions at 85 °C and 65 °C was based on comparison of distributed normalized spectral abundance factor for individual proteins. Fold-change for proteins detectable exclusively in the 85 °C or 65 °C condition (fold-change = ∞) was set at 7.8 and 0.1, respectively, which represents the maximum and minimum of measured values. *n* = 3 LC–MS/MS runs for each condition. Student's *t*-test, paired, two tailed *P* = 0.012. **d**, Misincorporation rates of ac^{4}C sites at distinct regions of *T. kodakarensis* tRNAs as a function of growth temperature (55–95 °C), segregated into distinct regions within the tRNA molecule. Only sites with a minimal stoichiometry of 5% in any sample are shown. Box plot visualization parameters are as in Fig. B.1h. *n* = 4 biologically independent samples for 85 °C, *n* = 2 for 65 °C and 75 °C and *n* = 1 for 55 °C and 95 °C. **e**, Multiple alignment of 37 tRNA molecules, representing 19 distinct tRNAs in *T. kodakarensis*, plotted across three distinct temperatures. ac^{4}C sites are coloured on the basis of misincorporation rate (see colour bar). The red–orange bar on the left segregates the aligned sequences into distinct tRNA molecules, identified by the single-letter abbreviation of their amino acid. Selected regions from the multiple alignment, where ac^{4}C is particularly abundant, are shown and colour-coded according to the bottom colour bar. **f**, Schematic representation of RNaseP RNA in *T. kodakarensis*. ac^{4}C sites (all in C**C**G) are marked with circles colour-coded by misincorporation rate measured in cells grown at 85 °C. Fine grey lines indicate regions that base pair in the folded structure of the molecule, according to the model in ref. [79]. **g**, **h**, Distribution of 119 acetylated cytidine residues (in 86 mRNAs) in *T. kodakarensis* across different codons (**g**) and at specific position within codons (**h**) are shown, and compared to that of 2,245 control non-acetylated cytidines, found at CCG motifs of the same mRNAs. The *y* axis presents the fraction of cytidines in each position. *n* = 1 set of sites (comprising 119 ac^{4}Cs and 2,245 Cs) with error bars representing standard deviation of the binomial distribution. **i**, Anti-ac^{4}C immuno-northern blot in *P. furiosus* and *T.* sp. AM4 total RNA as a function of temperature. Ethidium bromide staining was used to visualize total RNA. Results are representative of two biological replicates. For gel source data, see Supplementary Data B.3. **j**, A heat map showing misincorporation rates at conserved ac^{4}C sites in 5S, 16S, 23S, RNase P RNA and SRP RNA of *T. kodakarensis, P. furiosus* and *T.* sp. AM4 grown at various temperatures. Rows are ordered according to misincorporation rates quantified in *T. kodakarensis* grown at 95 °C. The arrowhead indicates the conserved ac^{4}C site at helix 45 (top site in the heat map).


a
T. kodakarensis 70S map reconstruction
WT85: 53,737 particles, 2.95Å resolution
WT65: 283,424 particles, 2.55Å resolution
ΔTkNat10: 116,586 particles, 2.65Å resolution
Masks creation
LSU
SSU head
SSU body
Multi body refinement
Combined map
SSU head
SSU body
LSU
b
70S (2.95 Å)
LSU (2.89Å)
SSU body (2.91Å)
SSU head (3.02Å)
FSC
FSC = 0.143
Spatial Frequency [1/Å]
c
70S (2.55 Å)
LSU (2.47Å)
SSU body (2.5Å)
SSU head (2.6Å)
FSC
FSC = 0.143
Spatial Frequency [1/Å]
d
70S (2.65 Å)
LSU (2.61Å)
SSU body (2.7Å)
SSU head (2.7Å)
FSC
FSC = 0.143
Spatial Frequency [1/Å]
e
Map vs. full map
Map vs. half map 1
Map vs. half map 2
FSC
Spatial Frequency [1/Å]
WT85
f
Map vs. full map
Map vs. half map 1
Map vs. half map 2
FSC
Spatial Frequency [1/Å]
WT65
g
Map vs. full map
Map vs. half map 1
Map vs. half map 2
FSC
Spatial Frequency [1/Å]
ΔTkNat10

Extended Data Fig. B.7 Cryo-EM data processing and map reconstruction. a, Schematic representation of electron microscopy data processing for the *T. kodakarensis* ribosomes. Data processing was performed in Relion 3 and included motion correction, contrast transfer function correction, particle picking and classification. Initial map reconstruction and post processing was performed by the 3D refinement algorithm implemented in Relion on the complete 70S particle, indicating high residual mobility of the SSU head domain (top left, grey). Further implementation of multibody refinement with individual masks prepared for the LSU (blue), SSU body (green) and SSU head (orange) resulted in the complete reconstruction of the 70S particle. The final map consisting of all three ribosomal domains for the wild-type ribosome derived from cells grown at 85 °C is presented at the bottom left. Fourier shell correlation curves indicating overall (black) and per domain (colour coded according to the relevant masks) resolutions are presented in **b** for the wild-type strain grown at 85 °C (WT85), **c** for the wild-type strain grown at 65 °C (WT65) and **d** for the ΔTk*NAT10* ribosomes (mutant). FSC comparisons of full (grey) and half-maps (pink/cyan) to the final refined model are presented in **e**, **f** and **g** for WT85, WT65 and mutant strains, respectively. The excellent agreement between the cyan and pink curves indicates the lack of overfitting.


a
WT85
PDB 6skf, EMD-10223
WT65
PDB 6th6, EMD-10503
ΔTkNat10
PDB 6skf, EMD-10223
1.9Å
2.6Å
3.0Å
3.5Å
4.0Å
5.0Å
6.0Å
b
WT85
WT65
ΔTkNat10
ac4C5
ac4C730
ac4Cm1020
ac4C1459
ac4C376
ac4C1434
ac4C2159
C5
C730
C1020
C1459
C376
C1434
C2159
SSU
LSU

Extended Data Fig. B.8 Cryo-EM data quality and ac^{4}C visualization. a, Surface (top) and cross-section (bottom) representations of the cryo-EM density maps coloured according to local resolution distribution. Growth conditions and *T. kodakarensis* strains used in the study along with PDB and EMDB accession codes are indicated. Resolution values are colour-coded according to the right index and are presented in Å. **b**, Model in density for multiple ac^{4}C positions in wild-type *T. kodakaresis* grown at 85 °C (orange) and 65 °C (yellow) compared to an ac^{4}C ΔTk*NAT10* strain (mutant, blue) indicating the absence of acetate density (highlighted in light orange) in the mutant and in multiple positions of the strain grown at 65 °C. Positions highlighted with an asterisk are also acetylated in the 65 °C strain whereas positions that are unmarked are acetylated only in the archaea grown at 85 °C. These data are in good agreement with both the genomic sequencing and MS approaches described in this manuscript, that similarly indicate that ac^{4}C distribution is highly dependent on growth temperature. For a 2D map with ac^{4}C distribution, see Extended Data Fig. B.9d.


a
% misincorporation at 85°C
80
40
0
Not detected by cryo-EM
Detected by cryo-EM
13.7%
3.2%
ac4C sites in T. kodakarensis rRNA
b
362 fold
Normalized % modification (LC–MS)
100
50
0
ac4C
ac4Cm
c
ac4Cm1020
d
23S - 5'
23S - 3'
16S
PTC
h45
5S
e
ET
f
T. kodakarensis
E. coli
S. cerevisiae
H. sapiens
ET
PTC
DC
90°
g
Lowest detection temperature (°C)
55
65
75
85
% misincorporation at 85°C
80
60
40
20
0
ac4C sites in T. kodakarensis rRNA
h
55°C
85°C
Misincorporation rate
0.8
0.6
0.4
0.2
i
A1786
3.5Å
ac4C1434
3.6Å
C1434
j
dθ/dT
0.0
-0.1
-0.2
C
ac4C
ΔTm = 11.5 ± 1.1
30
50
70
90
Temperature [°C]

Extended Data Fig. B.9 RNA modifications of *T. kodakarensis* ribosome and thermostability. a, Misincorporation level as quantified by ac^{4}C-seq across all ac^{4}C sites identified in ribosomes of *T. kodakarensis* at 85 °C. Blue and red bars indicate sites that were and were not detected by cryo-EM, respectively. Dashed lines indicate median misincorporation of cryo-EM detected (upper, 13.7%) and not-detected (lower, 3.2%) sites. Acetylation detected by ac^{4}C-seq and also observed in the cryo-EM were generally of medium to high stoichiometry whereas the majority of acetylation sites detected by ac^{4}C-seq but not observed in the cryo-EM map density were of relatively low stoichiometry, rendering them invisible in the ensemble cryo-EM structure, which averages thousands of individual particles for map reconstruction. **b**–**e**, Combined cryo-EM and mass spectrometric analysis indicated six ac^{4}C residues that are also methylated at their 2′-O position. Relative quantification of ac^{4}C and ac^{4}Cm detection in *T. kodakarensis* RNA via LC–MS is presented in **b**. Mean and individual data points are shown. *n* = 3 technical replicates. An example of ac^{4}Cm in density is shown in **c** with acetate and methyl installations indicated by black arrows. 2D (**d**) and 3D (**e**) visualization of ac^{4}C and ac^{4}Cm distribution in the *T. kodakarensis* ribosome with ac^{4}C highlighted orange and ac^{4}Cm green. Data are presented for the *T. kodakarensis* grown at optimal growth temperature (85 °C). Ac^{4}C positions highlighted in orange include genomic, mass spectrometry and electron microscopy data. Ac^{4}Cm positions are a combination between cryo-EM and mass spectrometry data. In **e**, RNA and proteins are presented as grey ribbons, modified residues are highlighted as spheres. The protein exit tunnel (ET) is highlighted with a dashed black line, and tRNA is in yellow. The tRNA and mRNA coordinates are from PDB 4V5D. **f**, A comparative view of RNA modification distribution in *E. coli*, yeast (*S. cerevisiae*), human (*H. sapiens*) and *T. kodakarensis*. Base modifications are coloured blue, ac^{4}Cs in red, tRNA and mRNAs in yellow. Ribosome functional regions are designated in black with decoding centre (DC), the peptidyl transferase centre (PTC) and the protein exit tunnel (ET) highlighted by a dashed black line. PDB codes for the structures used for comparison are 5AFI, 4V88 and 4UGO, for the *E. coli*, *S. cerevisiae* and human ribosome, respectively. **g**, Misincorporation rate as quantified by ac^{4}C-seq for all ac^{4}C sites in the *T. kodakarensis* ribosome. The bar colour indicates the lowest growth temperature at which the site was detected. **h**, 3D representation of the *T. kodakarensis* ribosome with ac^{4}C sites detected at 55 °C and 85 °C shown and colour-coded according to misincorporation rate in each temperature. **i**, Ac^{4}Cs were shown to stabilize the *T. kodakarensis* ribosome via direct interactions with protein and RNA residues. An example of stabilization through RNA– protein interactions is presented in Fig. 4g. RNA–RNA interactions correspond to interactions of ac^{4}C1434 with OP2 of A1786 of LSU. **j**, Temperature-dependent circular dichroism spectra of synthetic RNAs containing cytidine (blue) or ac^{4}C (red). Solid and dashed lines represent mean and individual measurements, respectively. *n* = 3 independent experiments. *θ*, ellipticity at 260 nm.

REFERENCES

1. Sharma S et al. Yeast Kre33 and human NAT10 are conserved 18S rRNA cytosine acetyltransferases that modify tRNAs assisted by the adaptor Tan1/THUMPD1. *Nucleic Acids Res* 43, 2242–2258 (2015).

2. Ito S et al. A single acetylation of 18S rRNA is essential for biogenesis of the small ribosomal subunit in Saccharomyces cerevisiae. *J. Biol. Chem* 289, 26201–26212 (2014).
3. Ito S et al. Human NAT10 is an ATP-dependent RNA acetyltransferase responsible for N^4-acetylcytidine formation in 18 S ribosomal RNA (rRNA). *J. Biol. Chem* 289, 35724–35730 (2014).
4. Larrieu D, Britton S, Demir M, Rodriguez R & Jackson SP Chemical inhibition of NAT10 corrects defects of laminopathic cells. *Science* 344, 527–532 (2014).
5. Tschida BR et al. Sleeping Beauty insertional mutagenesis in mice identifies drivers of steatosis-associated hepatic tumors. *Cancer Res* 77, 6576–6588 (2017).
6. Zhang H et al. GSK-3β-regulated N-acetyltransferase 10 is involved in colorectal cancer invasion. *Clin. Cancer Res* 20, 4717–4729 (2014).
7. Kotelawala L, Grayhack EJ & Phizicky EM Identification of yeast tRNA Um44 2′-O-methyltransferase (Trm44) and demonstration of a Trm44 role in sustaining levels of specific $tRNA^{Ser}$ species. *RNA* 14, 158–169 (2008).
8. Dewe JM, Whipple JM, Chernyakov I, Jaramillo LN & Phizicky EM The yeast rapid tRNA decay pathway competes with elongation factor 1A for substrate tRNAs and acts on tRNAs lacking one or more of several modifications. *RNA* 18, 1886–1896 (2012).
9. Sharma S et al. Specialized box C/D snoRNPs act as antisense guides to target RNA base acetylation. *PLoS Genet* 13, e1006804 (2017).
10. Arango D et al. Acetylation of cytidine in mRNA promotes translation efficiency. *Cell* 175, 1872–1886.e24 (2018).
11. Thomas JM et al. A chemical signature for cytidine acetylation in RNA. *J. Am. Chem. Soc* 140, 12667–12670 (2018).
12. Sinclair WR et al. Profiling cytidine acetylation with specific affinity and reactivity. *ACS Chem. Biol* 12, 2922–2926 (2017).

13. Tardu M, Jones JD, Kennedy RT, Lin Q & Koutmou KS Identification and quantification of modified nucleosides in Saccharomyces cerevisiae mRNAs. *ACS Chem. Biol* 14, 1403–1409 (2019).
14. Kowalak JA, Dalluge JJ, McCloskey JA & Stetter KO The role of posttranscriptional modification in stabilization of transfer RNA from hyperthermophiles. *Biochemistry* 33, 7869–7876 (1994).
15. Taoka M et al. Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic Acids Res* 46, 9289–9298 (2018).
16. Orita I et al. Random mutagenesis of a hyperthermophilic archaeon identified tRNA modifications associated with cellular hyperthermotolerance. *Nucleic Acids Res* 47, 1964–1976 (2019).
17. Yu N et al. tRNA modification profiles and codon-decoding strategies in Methanocaldococcus jannaschii. *J. Bacteriol* 201, e00690–18 (2019).
18. Sharma S & Lafontaine DLJ 'View from a bridge': a new perspective on eukaryotic rRNA base modification. *Trends Biochem. Sci* 40, 560–575 (2015).
19. Fischer N et al. Structure of the E. coli ribosome–EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. *Nature* 520, 567–570 (2015).
20. Polikanov YS, Melnikov SV, Söll D & Steitz TA Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly. *Nat. Struct. Mol. Biol* 22, 342–344 (2015).
21. Kawai G et al. Conformational rigidity of N4-acetyl-2′-O-methylcytidine found in tRNA of extremely thermophilic Archaebacteria (Archaea). *Nucleosides Nucleotides* 11, 759–771 (1992).
22. Bruenger E et al. 5S rRNA modification in the hyperthermophilic archaea Sulfolobus solfataricus and Pyrodictium occultum. *FASEB J* 7, 196–200 (1993).

23. Kumbhar BV, Kamble AD & Sonawane KD Conformational preferences of modified nucleoside N^4-acetylcytidine, ac^4C occur at “wobble” 34th position in the anticodon loop of tRNA. *Cell Biochem. Biophys* 66, 797–816 (2013).
24. Parthasarathy R, Ginell SL, De NC & Chheda GB Conformation of N^4-acetylcytidine, a modified nucleoside of tRNA, and stereochemistry of codon–anticodon interaction. *Biochem. Biophys. Res. Commun* 83, 657–663 (1978).
25. Safra M et al. The m^1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature* 551, 251–255 (2017).
26. Li X et al. Base-resolution mapping reveals distinct m^1a methylome in nuclear- and mitochondrial-encoded transcripts. *Mol. Cell* 68, 993–1005.e9 (2017).
27. Grozhik AV et al. Antibody cross-reactivity accounts for widespread appearance of m1A in 5′ UTRs. *Nat. Commun* 10, 5126 (2019).
28. Helm M, Lyko F & Motorin Y Limited antibody specificity compromises epitranscriptomic analyses. *Nat. Commun* 10, 5669 (2019).
29. Gehring AM, Sanders TJ & Santangelo TJ Markerless gene editing in the hyperthermophilic archaeon Thermococcus kodakarensis. *Bio Protoc* 7, e2604 (2017).
30. Hileman TH & Santangelo TJ Genetics techniques for Thermococcus kodakarensis. *Front. Microbiol* 3, 195 (2012).
31. Santangelo TJ, Cubonov L, James CL & Reeve JN TFB1 or TFB2 is sufficient for Thermococcus kodakaraensis viability and for basal transcription in vitro. *J. Mol. Biol* 367, 344–357 (2007).
32. Santangelo TJ & Reeve JN Deletion of switch 3 results in an archaeal RNA polymerase that is defective in transcript elongation. *J. Biol. Chem* 285, 23908–23915 (2010).
33. Lipscomb GL et al. Natural competence in the hyperthermophilic archaeon Pyrococcus furiosus facilitates genetic manipulation: construction of markerless deletions of genes

encoding the two cytoplasmic hydrogenases. *Appl. Environ. Microbiol* 77, 2232–2238 (2011).

34. Oger P et al. Complete genome sequence of the hyperthermophilic archaeon Thermococcus sp. strain AM4, capable of organotrophic growth and growth at the expense of hydrogenogenic or sulfidogenic oxidation of carbon monoxide. *J. Bacteriol* 193, 7019–7020 (2011).
35. Farkas JA, Picking JW & Santangelo TJ Genetic techniques for the archaea. *Annu. Rev. Genet* 47, 539–561 (2013).
36. Wickham H *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
37. Matzov D et al. The cryo-EM structure of hibernating 100S ribosome dimer from pathogenic Staphylococcus aureus. *Nat. Commun* 8, 723 (2017).
38. Morlan JD, Qu K & Sinicropi DV Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS ONE* 7, e42882 (2012).
39. Shishkin AA et al. Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat. Methods* 12, 323–325 (2015).
40. Engreitz JM et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* 341, 1237973 (2013).
41. Machnicka MA et al. MODOMICS: a database of RNA modification pathways-2013 update. *Nucleic Acids Res* 41, D262–D267 (2013).
42. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).
43. Smith T, Heger A & Sudbery I UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* 27, 491–499 (2017).

44. Jun G, Wing MK, Abecasis GR & Kang HM An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* 25, 918–925 (2015).
45. Piechotta M, Wyler E, Ohler U, Landthaler M & Dieterich C JACUSA: site-specific identification of RNA editing events from replicate sequencing data. *BMC Bioinformatics* 18, 7 (2017).
46. Crooks GE, Hon G, Chandonia J-M & Brenner SE WebLogo: a sequence logo generator. *Genome Res* 14, 1188–1190 (2004).
47. Vainberg Slutskin I, Weingarten-Gabbay S, Nir R, Weinberger A & Segal E Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. *Nat. Commun* 9, 529 (2018).
48. Weingarten-Gabbay S et al. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* 351, aad4939 (2016).
49. Li B & Dewey CN RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323 (2011).
50. Robinson MD & Oshlack A A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11, R25 (2010).
51. Tarazona S, Garca F, Ferrer A, Dopazo J & Conesa A NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.journal* 17, 18–19 (2012).
52. Katoh K, Misawa K, Kuma K & Miyata T MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30, 3059–3066 (2002).
53. Sievers F et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol* 7, 539 (2011).

54. Basanta-Sanchez M, Temple S, Ansari SA, D'Amico A & Agris PF Attomole quantification and global profile of RNA modifications: Epitranscriptome of human neural stem cells. *Nucleic Acids Res* 44, e26 (2016).
55. Yamauchi Y et al. Denaturing reversed phase liquid chromatographic separation of non-coding ribonucleic acids on macro-porous polystyrene-divinylbenzene resins. *J. Chromatogr. A* 1312, 87–92 (2013).
56. Taoka M et al. An analytical platform for mass spectrometry-based identification and chemical analysis of RNA in ribonucleoprotein complexes. *Nucleic Acids Res* 37, e140 (2009).
57. Nakayama H, Yamauchi Y, Taoka M & Isobe T Direct identification of human cellular microRNAs by nanoflow liquid chromatography-high-resolution tandem mass spectrometry and database searching. *Anal. Chem* 87, 2884–2891 (2015).
58. Nakayama H et al. Ariadne: a database search engine for identification and chemical analysis of RNA using tandem mass spectrometry data. *Nucleic Acids Res* 37, e47 (2009).
59. Zhang Y, Wen Z, Washburn MP & Florens L Evaluating chromatographic approaches for the quantitative analysis of a human proteome on Orbitrap-based mass spectrometry systems. *J. Proteome Res* 18, 1857–1869 (2019).
60. Zhang Y, Wen Z, Washburn MP & Florens L Effect of dynamic exclusion duration on spectral count based quantitative proteomics. *Anal. Chem* 81, 6317–6326 (2009).
61. Zhang Y, Wen Z, Washburn MP & Florens L Improving proteomics mass accuracy by dynamic offline lock mass. *Anal. Chem* 83, 9344–9351 (2011).
62. McDonald WH et al. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom* 18, 2162–2168 (2004).

63. Xu T et al. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *J. Proteomics* 129, 16–24 (2015).

64. Tabb DL, McDonald WH & Yates JR III. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res* 1, 21–26 (2002).

65. Zhang Y, Wen Z, Washburn MP & Florens L Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal. Chem* 82, 2272–2281 (2010).

66. Zheng SQ et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* 14, 331–332 (2017).

67. Rohou A & Grigorieff N CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol* 192, 216–221 (2015).

68. Mindell JA & Grigorieff N Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol* 142, 334–347 (2003).

69. Zivanov J et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* 7, e42166(2018).

70. Rosenthal PB & Henderson R Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol* 333, 721–745 (2003).

71. Afonine PV et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* 74, 531–544 (2018).

72. Kucukelbir A, Sigworth FJ & Tagare HD Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* 11, 63–65 (2014).

73. Pettersen EF, Goddard TD & Huang CC UCSF Chimera-a visualization system for exploratory research and analysis. *J. Comput. Chem* 25, 1605–1612 (2004).

74. Emsley P, Lohkamp B, Scott WG & Cowtan K Features and development of Coot. *Acta Crystallogr. D* 66, 486–501 (2010).
75. Moriarty NW, Grosse-Kunstleve RW & Adams PD electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr. D* 65, 1074–1080 (2009).
76. Williams CJ et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci* 27, 293–315 (2018).
77. Deutsch EW et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* 45, D1100–D1106 (2017).
78. Perez-Riverol Y et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 47, D442–D450 (2019).
79. Ueda T et al. Mutation of the gene encoding the ribonuclease P RNA in the hyperthermophilic archaeon Thermococcus kodakarensis causes decreased growth rate and impaired processing of tRNA precursors. *Biochem. Biophys. Res. Commun* 468, 660–665 (2015).