DISSERTATION

THE EPITRANSCRIPTOME IN HEAT-LOVING ARCHAEA ENHANCES THERMOPHILY

Submitted by

Kristin Alison Fluke

Graduate Degree Program in Cell & Molecular Biology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2023

Doctoral Committee:

Advisor: Thomas J. Santangelo

Carol Wilusz
Daniel Sloan
Zaid Abdo

ABSTRACT

THE EPITRANSCRIPTOME IN HEAT-LOVING ARCHAEA ENHANCES THERMOPHILY

>170 RNA modifications are known to decorate the transcriptome across all three Domains of life. The totality of RNA modifications in a cell is called the epitranscriptome. Modifications expand the form and function of RNA, often invoking new structures, activities, and interactions. The molecular consequences, fitness impacts, transcriptome-wide distribution, and genesis of the vast majority of modifications are largely unknown, but more > 100 human diseases are linked to mutations in the genes that encode RNA modifying enzymes. It is therefore critical to elucidate the generation and impact of RNA modifications on fitness and function.

5-methylcytidine (m^5C) is one of the most abundant and conserved modifications across Domains and is generated through the post-transcriptional activities of several RNA m^5C methyltransferases (R5CMTs). RNA modifications, especially m^5C, have largely been studied in the context of abundant rRNA and tRNAs while research into the impact of mRNA modifications is lacking due to their low abundance in the cell. Archaeal model organisms have been shown to incorporate a higher abundance of select modifications compared to Eukarya, proving a new avenue to resolve fundamental questions regarding the phenotypic consequences of epitranscriptomic changes.

In the model hyperthermophilic archaeon, *Thermococcus kodakarensis*, I comprehensively mapped m^5C to the transcriptome. I identified at least five R5CMTs that site-specifically generate m^5C and showed an unprecedented level of m^5C incorporation that includes 10% of unique transcripts, mainly in mRNA. I demonstrated that R5CMTs target

mRNAs for modification with both sequence and structural specificity. Cells lacking m^5C exhibit a severe temperature dependent growth defect, indicating the m^5C epitranscriptome is critical for cellular fitness under heat stress. The extensive m^5C epitranscriptome coupled with the large collection of R5CMTs indicate that *T. kodakarensis* is the ideal model system to pursue fundamental questions regarding the epitranscriptome.

Efforts to identify RNA methyltransferases that install m^5C led to the discovery of a novel modification, N4,N4-dimethylcytidine (m^4_2C) and the enzyme responsible for its *in vivo* and *in vitro* installation. I showed that m^4_2C is robustly resistant to bisulfite-driven deamination, potentially indicating that all bisulfite-sequencing datasets may be falsely reporting m^5C sites that are instead occupied by m^4_2C. I mapped a single m^4_2C residue to the ribosomal decoding center in the 16S rRNA and showed that cells lacking m^4_2C exhibit a severe growth defect at higher temperatures. Structural studies of the enzyme that generates m^4_2C, tentatively named m^4_2C synthase, demonstrate it adopts a canonical class I Rossman fold at the C-terminal lobe and a unique N-terminal lobe. I showed that m^4_2C synthase methylates assembled ribosomes and defined the catalytic amino acid residue. Taken together, I report a novel writer enzyme and show that both m^5C and m^4_2C promote hyperthermophilic growth. The dense and chemically diverse epitranscriptome argues that *Thermococcus* provides an excellent model system for further epitranscriptomic studies that probe the impact of both ubiquitous and rare modifications on core biological processes.

ACKNOWLEDGMENTS

I must express my deepest gratitude to all members of the Santangelo lab, past and present, for their relentless support. Completing a PhD has been the most challenging experience of my life, and without the support of my lab mates, I surely would have failed. I have learned a great deal, on a personal and professional level, from each and every one of you.

Tom – Because of your guidance, I am not only a competent scientist, but a better person. You showed me how to be a fearless leader, how to be dignified in front of an audience, and to always strive for greatness. When no one else gave me a chance, you saw my potential and welcomed me into the lab. Your patients with me during our first few years working together was …impressive, to say the least. You are the first person to truly invest in my growth. For these reasons and many more, you will be a life-long friend. As such, I still expect approximately 11 lb of fudge annually, an invite to the annual Santangelo lab party, and any future *thermococcus*-themed swag.

Victoria, **Liam**, **Jackson** – As Undergraduate Research Assistants, you have all had a direct hand in the work presented here and provided a vital life-line to my research goals and achievements. Your contributions were and still are critical to my success and the progression of the field as a whole. Your efforts have pushed the boundaries of science and yielded valuable contributions to the epitranscriptomics community. A sincere ‘Thank you’ is earned, and I hope to continue to support you all in your future plans.

The Epitranscriptomics Consortium – It truly takes a village to make great science happen. I would be remiss if not to acknowledge and express my deepest gratitude to all members of the consortium, and especially the folks at New England Biolabs, for their essential contributions to my graduate research and professional development.

DEDICATION

I have had many invaluable experiences as a PhD student that have forced my growth personally and professionally. Many of these experiences were anticipated, however, I did not expect to develop such high value emotional investments through mentorship. My growth as a leader and mentor is immeasurable, and by a wide margin, brought the most significant and worthwhile lessons to my educational endeavors. **I dedicate my work to Victoria**, **Liam, and Jackson.** You have all challenged me in unique ways technically and personally, and I have watched you all grow into competent researchers who can think critically and execute experiments with integrity and rigor. Your efforts have contributed immensely to my accomplishments and your success as aspiring scientists has become my number-one goal. It seems fitting to dedicate my work to those who had the most significant impact on my growth.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGMENTS iv
DEDICATION v
CHAPTER 1: INTRODUCTION 1
RNA is modified in all domains of life 1
Methods for and limitations of detecting RNA modifications 4
Writers and erasers of the epitranscriptome 21
Select RNA modifications are abundant in heat-loving archaea 29
REFERENCES 32
CHAPTER 2: THE EXTENSIVE AND DYNAMIC M5C EPITRANSCRIPTOME OF THERMOCOCCUS KODAKARENSIS IS GENERATED BY A SUITE OF RNA METHYLTRANSFERASES THAT SUPPORT LIFE IN THE EXTREMES 48
Summary 48
Introduction 49
Results 51
Discussion 72
Materials and Methods 76
REFERENCES 92
CHAPTER 3: A NOVEL N4,N4-DIMETHYLCYTIDINE STABILIZES THE ARCHAEAL RIBOSOME 101
Summary 101
Introduction 101
Results 104
Discussion 119
Materials and Methods 121
CHAPTER 4: THE IN VIVO CELLULAR AND MOLECULAR IMPACTS OF THE M5C EPITRANSCRIPTOME 138
Summary 138
Background 138
Results 140
Discussion 149
Materials and Methods 150
REFERENCES 154
APPENDIX 1: THERMOCOCCUS KODAKARENSIS PROVIDES A VERSATILE HYPERTHERMOPHILIC ARCHAEAL PLATFORM FOR PROTEIN EXPRESSION 156
Summary 156
Introduction 157
2. Genetic systems for Thermococcus kodakarensis 159
3. Purifying proteins and identifying protein interactions 172

4. Reagents, recipes and equipment..... 174
5. Protocols..... 180
6. Advantages and future perspectives..... 184
REFERENCES..... 188
APPENDIX 2: CHAPTER 2 SUPPLEMENTARY ITEMS..... 194
APPENDIX 3: CHAPTER 3 SUPPLEMENTARY ITEMS..... 209

CHAPTER 1: INTRODUCTION

RNA is modified in all domains of life

Incredibly diverse life forms have emerged throughout earth's history that thrive across vastly different environments, and each have earned a designation on the tree of life within one of the three Domains; Eukarya, Bacteria, and Archaea. Although Bacteria and Archaea both lack membrane bound organelles, recent discoveries of novel archaeal species have challenged the three-domain theory as the standard markers of phylogenetic relatedness have led to blurred lines between the Archaeal and Eukaryotic Domains.[1] Despite the vast differences observed across the spectrum of life, a perplexingly simple genetic code (DNA) and the message of that code (RNA) are composed of a minimal set of five nucleosides (A, T, U, C, and G) that shape all organisms. Perhaps unsurprisingly, organisms have adapted strategies to expand their genetic capacity by chemically modifying nucleosides to take on new identities and functions (Figure 1.1).[2] Species from all Domains, and even viruses, chemically modify their genetic material, but the details surrounding how most modifications are generated and their impact on fitness and function are lacking.

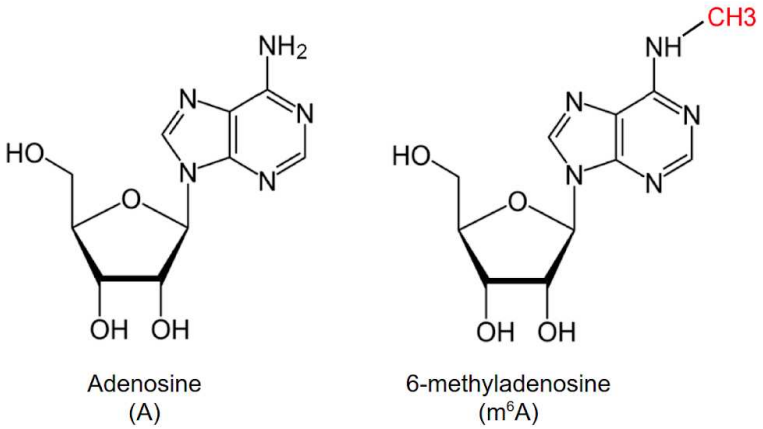


Figure 1.1 RNA modifications expand the nucleotide code. As a single example, adenosine can be methylated at position N6, generating a distinct nucleotide, 6-methyladenosine (m6A).

The interest in DNA and RNA modifications has led to the emergence of the epigenomics and epitranscriptomics fields of study, respectively. DNA modifications can expand sequence composition by at least 17 nucleotides while upwards of 170 unique RNA modifications have been identified.[3] DNA modifications generally do not impact Watson-Crick base pairing and therefore do not directly alter the coding potential of the resulting codons, but these modified residues do alter protein-DNA interactions and have critical impacts on normal and adaptive cellular processes.[4] These interactions have been repeatedly shown to control gene expression (Figure 1.2), transcription elongation, DNA replication, DNA damage response, and allow self-recognition of restriction-modification systems that protect from invading DNA.[4–7] Inosine (the deaminated analog of adenosine) and 8-oxoguanine (the oxidized analog of guanine) are exceptions that lead to altered base-pairing but are widely regarded as lesions that elicit a DNA damage response that is quickly repaired through base excision repair pathways.[8,9] DNA modifications participate in core biological processes, so it is unsurprising that aberrantly modified DNA can be detrimental to human health, leading to severe disease, impaired development, and aging.[4] The epigenome is relatively uncomplicated in model species, and although there is still much to learn about how the epigenome is dynamically regulated, the the function of common DNA modifications is well studied.

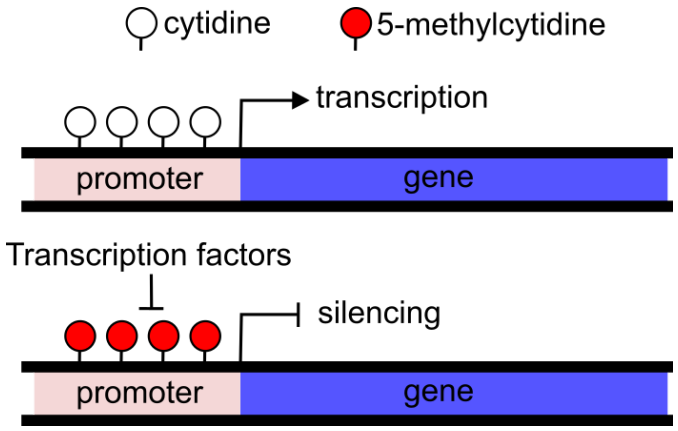


Figure 1.2 Hypermethylation of cytidine at promoter regions leads to transcriptional silencing.

The unbridled complexity, low abundance, and low stoichiometry of most RNA modifications in model species has resulted in the field of epitranscriptomics lagging behind and the impacts of the vast majority of RNA modifications having yet to be explored.[10] The discovery of new RNA modifications has outpaced the development of methodologies to detect them with adequate resolution necessary for targeted studies.[10] Mass spectrometry approaches have been critical for the identification of nucleoside analogs from total RNA. These studies have revealed that some RNA modifications are unique to one or two Domains of life, but many are conserved across Domains.[11] Stark differences in the modification profile of stable RNAs and messenger RNAs (mRNAs) has revealed that ribosomal- and transfer-RNAs (rRNA and tRNAS) are densely modified compared to mRNAs. Stable RNAs are the major components of protein synthesis machinery and therefore central to cellular function.[2] Many RNA modifying enzymes have well-defined structural and sequence motifs, leading to the success of bioinformatics approaches followed by biochemical experimentation to identify some of the enzymes that generate the epitranscriptome. For these reasons, many epitranscriptomics studies have focused on the modification profile of rRNA and tRNA where the impacts and molecular roles of modifications and modifying enzymes have been pursued in a shared context. Even still, only a select few RNA modifications, and mainly those abundant in stable RNAs, have been the focus of sufficient study to define their individual and cellular impacts.

RNA modifications are installed post-transcriptionally by enzymes that site-specifically target individual residues for conversion. Spontaneous modifications, such as deamination events, namely that of cytidine and adenosine, or the oxidation of guanosine can also result in the non-enzymatic production of nucleotide analogs. Cytosine deamination results in its conversion to uracil while deamination of adenosine results in the distinct nucleoside, Inosine. Although, dedicated enzymes called adenosine deaminases acting on RNA (ADARs), site-specifically generate inosine in diverse species.[12] Spontaneous deamination events are rare, non-targeted, difficult to measure *in vivo*, and their biological relevance is unclear.

Enzymatic production of modified RNAs is the predominant mechanism for the establishment and maintenance of the epitranscriptome, and loss or malfunction of RNA modifying enzymes has shed light on the role of the epitranscriptome to impact health and fitness.[13]

The extraordinary quantity of known modifications and their ubiquity leads to the reasonable assumption that the epitranscriptome provides tangible benefits to cellular fitness and overall health. Indeed, over 100 distinct human diseases have been linked to the epitranscriptome, and mutations in approximately half of known RNA modifying enzymes result in maladies such as cancer, cardiovascular diseases, birth defects, metabolic diseases, neurological disorders, infertility, and mitochondrial defects.[13,14] Although associated diseases are wide-ranging, neurological diseases are largely overrepresented, likely due to the observed enrichment of several types of RNA modifications across neuronal tissues.[13] Emerging knowledge as a result of advances in RNA-sequencing has revealed that RNA modifications are widespread and involved in major biological processes including cell differentiation, sex determination, stress responses, maintaining proteostasis, cellular signaling, translational control, and DNA damage response.[15] RNA modifications support core biological processes, but large gaps in our understanding remain regarding the molecular and cellular impacts of most RNA modifications and the fate of RNAs once modified.

Methods for and limitations of detecting RNA modifications

Several methods to detect modifications to RNAs have been developed that differentiate nucleotides based on their unique chemical properties. Liquid chromatography coupled mass spectrometry (LC-MS) is a universal approach that allows the detection of modified nucleosides based on their known retention times and shifts in their mass-to-charge ratio compared to unmodified nucleosides. Purification of total cellular RNA, selective degradation of rRNAs, enrichment of mRNAs, size selection of small RNAs or tRNAs, or gel purification of individual RNAs provides a means to analyze distinct RNA fractions. RNA fractions can be digested to

single nucleosides or small oligonucleotides and subjected to LC-MS where each unique nucleoside can be quantitatively identified. Such methods have demonstrated highly disparate modification patterns across tRNAs[16], the exclusion or ubiquity of RNA modifications across the 23S rRNA, 16S rRNA, long and short RNAs, mRNAs, or tRNA [17], and the abundance of some modifications in one mRNA isoform over another[18]. Although LC-MS is extremely sensitive and provides accurate quantitation within each sample, this method lacks high throughput site-specificity. Moreover, abundant rRNA and tRNAs are difficult to remove from RNA preparations and often lead to artifacts in fractions thought to be free of these RNAs.[19]

Thin-layer chromatography (TLC) has also provided a means to differentiate nucleotide identity through separation of hydrolyzed nucleotides according to their polar properties.[19,20] RNA samples are first enzymatically digested before each nucleotide is radiolabeled and spotted onto a TLC plate and solvent mobilized, driving their separation. Optionally, the nucleotides can be further resolved by separation in a second direction using a different solvent (2D-TLC). The TLC plate is imaged using phosphorimaging and the positions in which each nucleotide migrates compared to a standard is indicative of nucleotide identity (Figure 1.3). This method is relatively simple and has been used to define the composition and role of RNA modifications, such as the impact of tRNA modifications on translation rate[20,21] and how long-non coding RNA modifications influence DNA repair pathways[22].

Blotting methods also provide a means to detect whether a modification is present in a pool of RNA. Blotting methods entail spotting an RNA sample on a membrane (dot blot) or separation by gel electrophoresis before transferring to a membrane. An antibody against the targeted modification is applied to the membrane and can be quantitatively resolved (Figure 1.3). Unlike TLC and LC-MS, blotting methods require oligomerized RNA of length > ~15 nt to allow the RNA to bind to the membrane with adequate strength to remain bound through multiple washes. Blotting methods have been heavily utilized to detect the presence of specific modifications in total or fractionated RNA.[22–24] Antibody-based methods are quantitative, highly

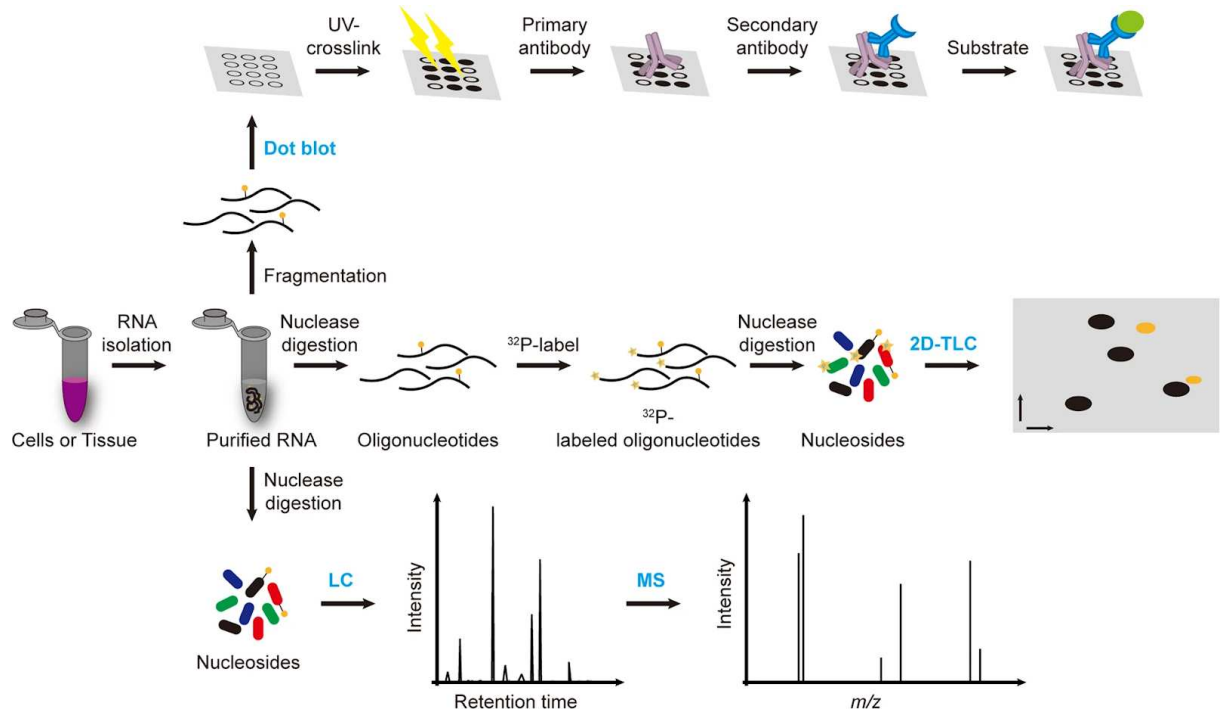


Figure 1.3 Technical flowchart of dot blot, 2D-TLC and LC–MS. Adapted from Zhang et. al., 2022.

sensitive, and also have the added benefit of tracking the size of the modified RNA. These methods have all been instrumental in detangling the complexities of the epitranscriptome and all carry the advantages of being quantitative and comprehensive to many different modifications, but all are limited in throughput and site-specific, transcriptome-wide identification. These methods are often best-suited for targeted studies where analysis of purified transcripts or RNA fractions are sufficient to address the research goals.

Transcriptome-wide approaches are ideal for measuring the global distribution and biological impact of modified residues. Approximately 10 unique RNA modifications have been reliably mapped to the transcriptome using next-generation sequencing (NGS) approaches.[24] Immunoprecipitation of RNAs using an antibody raised against modified nucleotides followed by RNA-sequencing has allowed detection and detailed studies of individual RNAs that harbor such modifications. Although immunoprecipitation-based methods lack quantitative power and site-specific resolution, this method has been used to generate a detailed map of m6A and

elucidate its role in RNA processing, nuclear export, and alternative splicing.[25,26] Although RNA-immunoprecipitation followed by sequencing has been instrumental in understanding the transcriptome-wide distribution and biological impacts of several modifications, antibody generation against nucleosides is challenging due to modified nucleosides being naturally present in host species, necessitating the elicitation of an autoimmune response. And, because nucleosides are small molecules, legitimate concerns regarding antibody specificity to unique modifications have led researchers to seek alternative methods moving forwards.

Another method aimed at mapping the precise location of modified nucleotides relies on the inability of reverse transcriptase to incorporate nucleotides at the site of modification, leading to a truncated cDNA.[27] Some modifications are known to stall specific reverse transcriptases, and sequenced reads will appear to end abruptly at the site of modification. This method provides a means to site-specifically map modifications with semi-quantitative power, yet may not be specific for the intended modification, as multiple types of modifications may stall the same reverse transcriptase. Also, additional sites of modification directly downstream of where reverse transcriptase stalled may escape detection.

Methods for quantitatively mapping modifications with single-nucleotide resolution have been developed for m5C[28], ac4C[23], and most recently m6A[29] that rely on misincorporation of nucleotides during cDNA synthesis. Chemical or enzymatic deamination of modified vs unmodified nucleosides results in a mismatched base relative to the reference genome. Some base analogs such as Inosine and 8-oxoG inherently result in misincorporation during cDNA synthesis, so are also detectable by NGS without a specially tailored protocol. These misincorporated nucleotides can be precisely mapped to a reference genome, allowing single nucleotide resolution of the modification.[27] Depending on the conversion chemistry, these methods can be highly quantitative in determining the stoichiometry of the modification at each position within the transcriptome. Misincorporation-based approaches are widely accepted as the current gold standard for epitranscriptomics studies[30] and have led to detailed investigations

in a wide range of species.

So called “third-generation sequencing” methods, such as those introduced by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), are quickly making their way into the spotlight. PacBio instruments are able to sequence long reads, and while doing so, measure the interpulse distance of nucleotide incorporation by Polymerase.[31] Some modified nucleotides elicit distinct interpulse distances relative to their unmodified counterparts. However, the RNA library preparation process is arduous and modification detection is limited to a small set of modifications. Single Molecule Minion sequencing is a promising development that has the potential to distinguish nucleotide identity based on the unique voltage of each nucleotide as it passes through a nanopore.[32] This method has yet to mature to acceptable standards for quantitatively detecting a variety of RNA modifications, but has the potential to provide quantitative detection of multiple types of modifications with single-nucleotide resolution and within a single experiment. Direct sequencing methods would also have the advantage of being robust to reverse transcription and PCR bias while also being less arduous to perform.[24,33] Advances in NGS approaches have permitted the targeted study of the function and enzymatic generation of select RNA modifications at the molecular level, and continue to revolutionize the field of epitranscriptomics. Pending the maturation of data analysis methodology, third-generation sequencing will likely be the next big development in this arena.

Modifications expand the function of RNA

Systematic efforts to generate quantified epitranscriptomic maps across diverse model species have revealed that some modifications are reversible, dynamically regulated, and abundantly present in a variety of RNAs inclusive of mRNAs, small RNAs, intergenic RNAs, and antisense RNAs.[15,23,25,28,34,35] The Modomics database is a useful resource concerning the structure, biosynthetic pathway, localization, and interacting enzymes related to the epitranscriptome.[3] Although modifications are distributed throughout the transcriptome, rRNA and tRNA are known to be densely modified with a plethora of different modifications, adding

new layers of complexity to otherwise highly conserved sequences and structures. Many unique modifications have only been described in the context of rRNA and tRNA, which are the most abundant and stable RNAs in a cell. As such, rRNA and tRNAs have historically served as the frontier for the discovery and characterization of RNA modifications.[15]

tRNA modifications

tRNAs are transcribed from the genome to initially produce RNAs that can reach up to several hundred nucleotides in length before undergoing a maturation process that commonly reduces the RNA to ~76 nucleotides.[36] During the maturation process, tRNAs undergo diverse chemical modifications that stabilize the secondary and tertiary structures as well as elicit a wide variety of translation outcomes.[37] Some modifications, like isopentenyl adenosine and wybutosine (yW) occur rarely, while other modifications, such as dihydrouridine (D) or pseudouridine (Ψ) are virtually ubiquitous across all tRNAs.[15] While more than 100 unique types of RNA modifications have been mapped to tRNAs, individual tRNAs typically harbor ~11-13 distinct modification sites, each invoking a particular function.[11] Although 20 canonical amino acids are used to synthesize proteins, there are ~300 expressed tRNAs encoded in the human genome that give rise to isoacceptors and isodecoders. The modification status of tRNAs has been noted to differ depending on its organellar localization.[38] Namely, modifications to mitochondrial tRNAs sometimes differ from that of a cytoplasmic tRNA. Differential modification profiles observed within the vast pool of tRNAs only deepen the complexity of these molecules and provide a model to explore the chemical and functional diversity of RNA.

Nearly all tRNAs adopt a clover-leaf secondary structure (Figure 1.4A) and a tertiary structure resembling an up-side-down L-shape (Figure 1.4B) that are crucial for maintaining translation outputs. The secondary structure is represented with four arms and a variable loop. The tertiary structure is largely achieved due to the interaction between the D- and T-loops.[39] The V-loop contains a variable number of nucleotides and that bulges outward to avoid steric hindrance.[39] This particular structure is maintained, in part, by modifications that are often

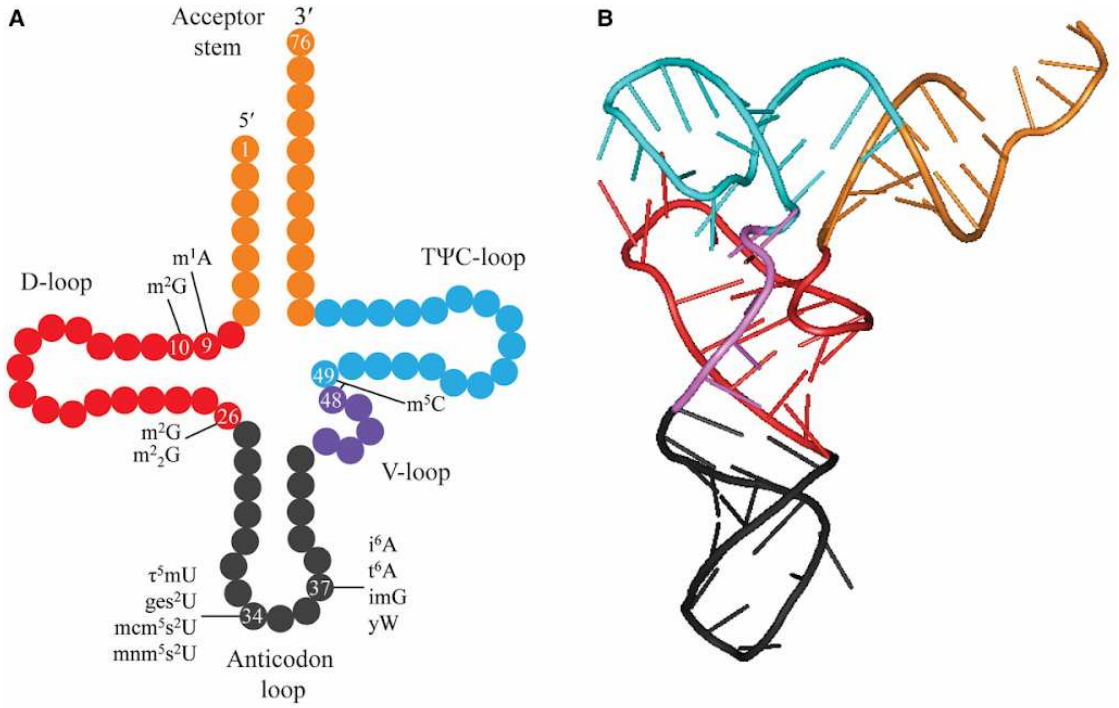


Figure 1.4 tRNA secondary and tertiary structure. Adapted from Jordan Ontiveros et. al., 2019. (A) Two-dimensional clover-leaf structure and the five domains that make up a tRNA. TΨC-loop and T-loop are synonymous. Specific residues whose modifications are discussed here. (B) Tertiary, L-shaped structure of tRNA. The color code of the secondary structure matches with that of the tertiary structure.

essential for maintaining proteostasis.[37] 1-methyladenosine (m1A) at the junction between the D-loop and acceptor stem inhibits base-pairing leading to the proper folding of the clover-leaf and is critical for the maturation of some tRNAs.[40] A variety of methylated guanosine analogs are often found clustered at the arm-junctions around the D-loop.[36] These modification positions are often conserved across all life as they promote the formation of the internal loop which keeps the core of the tRNA open enough to allow the L-shaped tertiary structure. 2-methylguanosine (m2G) is energetically identical to guanosine and does not impair base-pairing with cytidine, while N2,N2-dimethylguanosine (m2,2G) base pairs instead with adenosine.[41] There are known m2,2G-A interactions in some tRNAs between G26 and A44 that help to maintain the clover-leaf.[42]

Ψ, an isomer of uridine, has also been shown to induce drastic alterations in the

structure and function of tRNAs, as this nucleotide base has increased rotational freedom around the ribose and can form non-canonical base-pairing.[15] Ψ increases the thermodynamic stability of tRNAs and greatly increases the stacking ability of bases.[43] Although Ψ is often not essential to cellular viability, it is critical for the maintenance of the anticodon loop and proper binding of the tRNA to ribosome.[44] The most conserved tRNA modification, Dihydrouridine (D), simply replaces the C5-C6 double bond with hydrogen bonds in a uracil base. In the D-loop, this modification promotes secondary and tertiary structure by preventing base-stacking, and has been shown to destabilize helices and reduce the melting temperature necessary for proper D-loop folding.[45]

5-methylcytidine (m5C) has not been detected in *E. coli* tRNAs but is present in archaeal and eukaryotic tRNAs at positions 48 and 49 at the junction between the T- and the V-loops.[11,15] At these positions, m5C is known to enhance tRNA stability and protein synthesis. While not participating in non-canonical base-pairing per se, m5C at position 48 engages in Levitt pairing with the nucleotide at position 15, assuming a reverse Watson-Crick base pairing geometry that joins the D- and V-loops.[46] m5C has been shown to increase the melting temperature similar to Ψ,[47] but the direct contribution of m5C to tRNA function is largely unknown.

In addition to playing a major role in tRNA structural integrity, modifications to the anticodon stem and other regions have profound impacts on gene expression by modulating the rate of amino acid incorporation and frameshifting.[48,49] The structures elicity by Ψ and m2,2G as previously described also have indirect impacts on gene expression. Ψ in the anticodon loop can help maintaining homeostatic translation levels by strengthening codon-anticodon recognition and facilitating the rejection of incorrect tRNAs by reducing the rate of protein synthesis.[50] Interestingly, a study correlated an increase of m2,2G at position 26 in the D-loop with a decrease in RNA polymerase III in yeast and humans, perhaps due to a decrease in the rate of RNA polymerase III synthesis by m2,2G-containing tRNAs.[51] Modifications to positions 34 and 37 in the anticodon loop are the most varied between tRNAs and have been implicated

in tRNA maturation, stability, ribose conformation, and decoding potential.[15] In addition to modulation of gene expression levels, modifications also have the potential to directly impact protein sequence, directly altering the coding capacity of individual tRNAs.

Modifications at position 34, the wobble position, can expand or restrict the decoding range of a given codon, and a majority of tRNAs harbor modifications at this position. The codon-anticodon base pairing between mRNAs and tRNAs can accommodate certain mismatches at the 3rd codon position, although most modifications to the wobble position do not impact Watson-Crick base pairing[3,15] and do not directly affect the anticodon-codon interface. However, there are several examples of modifications at the wobble position that expand the wobble recognition repertoire and can result in non-cognate amino acid incorporation. 5-taurinomethylurdine (τ5mU) in mitochondrial $tRNA^{Leu}_{UAA}$ and bacterial 2-geranylatedthiouridine have been shown to stabilize hydrogen bonding and base stacking of adjacent nucleosides and contribute to the wobble pairing of U and G.[52] Modifications at position 37, directly 3' of the anticodon, often induce steric hindrance or hydrophobicity that support the anticodon stem-loop structure and provide stability to the codon-anticodon interaction.[15] The nucleoside composition at position 37 has variable impacts on translation efficiency and accuracy. Translational control through wobble-base pair interactions and structural maintenance may act to fine-tune the coding potential of individual codons.

rRNA modifications

Ribosomes are composed of approximately 3-4 rRNA fragments and upwards of 80 ribosomal proteins along with various metal cofactors such as magnesium and zinc. These complex molecules harbor highly conserved sequences and structures that are critical for translation. At the decoding center, mRNAs and tRNAs are bound together through a codon-anticodon interaction during translation elongation. The mRNA-tRNA complex is translocated to the P-site where polypeptide bond formation is catalyzed by the rRNA before the tRNA is ejected through the E-site. Hundreds of modified residues have been mapped to rRNA

that are mainly concentrated near the decoding center.[50,53] Although no single rRNA modification is essential to cell survival, modified residues confer fitness advantages, especially under certain growth or stress conditions.[23,54–57] There has been speculation that modifications may play a role in catalytic activity based on their mapped positions near the P-site[58,59], but ultimately, evidence to support the particular role of individual modified residues to impact peptidyl-transfer is largely missing.

One well recognized role of some rRNA modification is their ability to confer antibiotic resistance in bacterial species. In some species, methylated residues are positioned within the ribosome to prevent the binding of antibiotics that would otherwise prohibit protein synthesis, thus killing the cell.[60–62] A notable example, adenosine methylation in the bacterial P-site simultaneously blocks 8 different classes of antibiotics.[63] Multidrug resistant infections are a growing health concern, and clinical research efforts include developing antibiotics that would target unmodified residues. On the other hand, researchers have taken advantage of this adaptation by using antibiotic resistance as selectable markers for retention of genetic material, such as plasmids. However, the vast majority of rRNA modifications have not been shown to correlate with adaptive responses to environmental insults.

Like tRNA maturation, ribosomal maturation is an intricate process of trimming and folding different rRNA fragments into their final form. Modifications are installed during all stages of ribosomal biogenesis, and some even play a role in the maturation process.[64,65] Several hypermodified Ψ and 2'O methylated residues distributed throughout the rRNA have been shown to increase the efficiency of RNA cleavage.[66] The nucleolytic cleavages and structures of pre-rRNA are often dependent on select modifications, and it is therefore hypothesized that modifications facilitate proper ribosome formation.

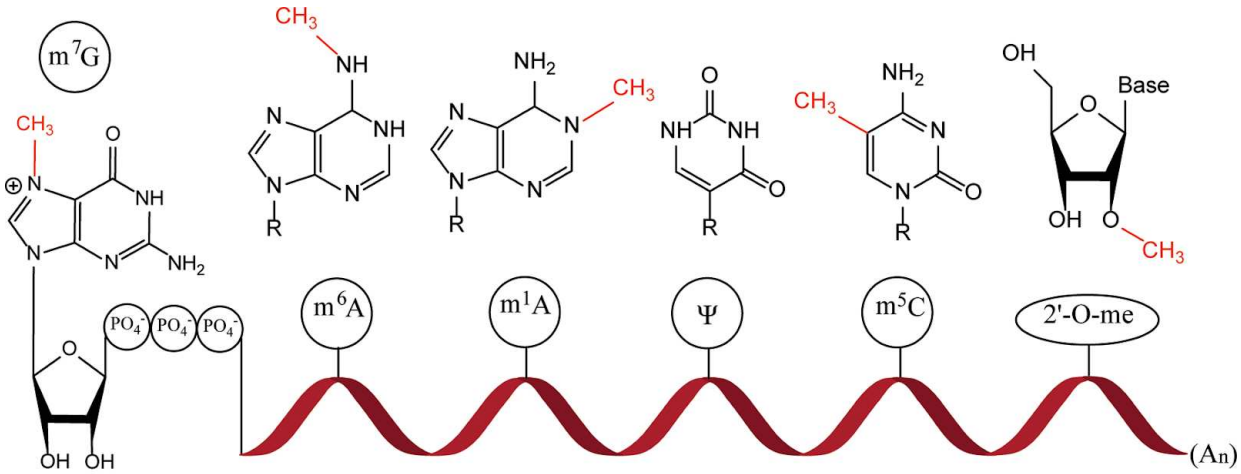
The highly structured domains found in rRNA are often supported by various stabilizing modifications that increase the rigidity or flexibility of certain nucleotides. More than 100 sites of 2'O methylation can be found in the eukaryotic ribosome, and this modification is known to favor

a particular conformation of the sugar base while protecting the nucleotide from hydrolysis of the phosphate backbone.[15,67] 2'O-methylation is perhaps the predominant modification found in rRNA, and likely plays a significant role in shaping its canonical secondary and tertiary structures.

The P-site is arguably the most functionally important ribosomal region as it is the location of nascent peptide bond formation. Maintaining the structural and functional integrity of this region is critical, and it is therefore unlikely a coincidence that the P-site is the most heavily modified region in the ribosome. There are approximately 90 Ψ residues mapped to the ribosome, most of which are clustered in the P-site.[67] Studies have shown that Ψ sites are collectively essential to translation outputs and healthy cellular growth.[67] Mainly, it is speculated that Ψ facilitates secondary and tertiary structures required for ribosome biogenesis and protein-RNA interactions.[66] Despite the abundance of dozens of unique modifications and disparate modification profiles found in rRNA across Domains of life, relatively little is known regarding the collective or individual role of most known modifications.

mRNA modifications

Transcriptome-wide mapping of modifications has renewed interest in uncovering the role of mRNA modifications, beyond traditional capping, polyadenylation, and splicing variants. In mammalian systems, m6A, m1A, Ψ, 2'O-methylation, and m5C in mRNAs and their associated enzymes (Figure 1.5) has sparked recent interest. But, detailed studies have mainly focused on elucidating the role of m6A, as it is highly abundant in mammalian mRNA. Studies have repeatedly shown that modifications are dynamically regulated in response to environmental cues.[13,68,69] Proteins that install modifications, termed *writers*, are counterbalanced by *eraser* enzymes that remove modifications. Proteins that bind to modified residues are termed *readers*, and some mRNA modifications have been shown to recruit reader proteins that alter RNA activity.[70,71] Writers and erasers work together to balance regulatory signals introduced by modifications that impact a wide-range of cellular processes such as



Writer(s)	METTL3/14 WTAP KIAA1429	TRMT6/61	PUS1-4 PUS6 PUS7 PUS9	NOL1/Nsun2 DNMT2	Unknown
Eraser(s)	ALKBH5 FTO	ALKBH3	Unknown	TET1-3	Unknown
Reader(s)	YTHDF1-3 YTHDC1/2 eIF3 hnRPNC IGF2BP1-3	Unknown	Unknown	ALYREF	Unknown

Figure 1.5 mRNA modifications and their associated reader, writer, and eraser proteins in humans. Adapted from Jordan Ontiveros et. al., 2019. The 5′ cap structure consists of m7G, ribose sugar, and the inverse triphosphate linkage. Five internal mRNA modifications and their structures and their associated proteins are diagrammed. An = poly(A) tail.

translation, half-life, structure, cellular localization, and intramolecular interactions.[2,68,72] Modifications of low abundance mRNAs and low modification stoichiometries lead some researchers to speculate whether mRNA modifications are resultant of promiscuous activity of enzymes with true substrate specificity for rRNAs or tRNA. In many cases, it is unclear if mRNA modifications are biologically relevant and drive cellular fitness.

Virtually all eukaryotic and some viral mRNAs are capped with a 5’ inverted, triphosphorylated guanosine that is methylated at the 7th position. It is common to also observe adjacent nucleosides be methylated at the 2’OH position (Figure 1.6), where higher eukaryotes tend to harbor more extensively methylated caps.[73] Cap structures enhance mRNA stability by

protecting against degradation by exonucleases, promote nuclear export, stimulate translation initiation through interactions with initiation factors, and increase alternative splicing accuracy and efficiency, possibly by facilitating interactions with cap-binding proteins.[73] At the 3' terminal, mRNAs are polyadenylated with hundreds of unmodified adenosines which also impart protective features and assist in nuclear export and translation initiation. Prokaryotic species do not contain capped or polyadenylated mRNAs, only simply a 5' triphosphorylated nucleoside.[74] Because the extent of cap methylation becomes more complex as organisms evolve, mRNA caps and their level of methylation are one of the characteristics that distinguish eukaryotes from prokaryotes and higher eukaryotes from lower eukaryotes.[73] Accordingly, the degree of methylation on viral mRNA caps mirrors that of the host. Modified 5' caps are fundamental to eukaryotic biology, and systemic failure to cap mRNAs is lethal.

The most prevalent modification mapped within eukaryotic mRNAs is m6A where 1-3 sites are predicted to occur per mRNA.[15] Several studies have found that m6A affects the geometry of the adenine base depending on whether the nucleoside is single or double stranded, indicating this modification induces local structural switches.[75,76] Researchers would later show the single stranded geometry of m6A facilitates the binding of reader proteins.[75,77] m6A localizes at splicing junctions within intronic regions, and several m6A writers and erasers have been shown to impact splicing outcomes.[78–80] Deletion of m6A writers results in higher levels of intron retention and exon skipping while depletion of m6A erasers increases exon exclusion. An m6A reader protein has also been shown to promote exon inclusion. Taken together, strong evidence is presented that m6A is dynamically regulated and directly facilitates alternative splicing of eukaryotic mRNAs. In addition to pre-mRNA maturation, researchers have shown that m6A also impacts a wider-range of cellular processes. Studies have shown that m6A promotes nuclear export of mRNAs, enhances the recruitment of translation factors, and destabilizes m6A-containing transcripts as a result of interactions with various reader proteins.[81] The direct impacts of m6A in these contexts are not entirely understood, but the evidence thus

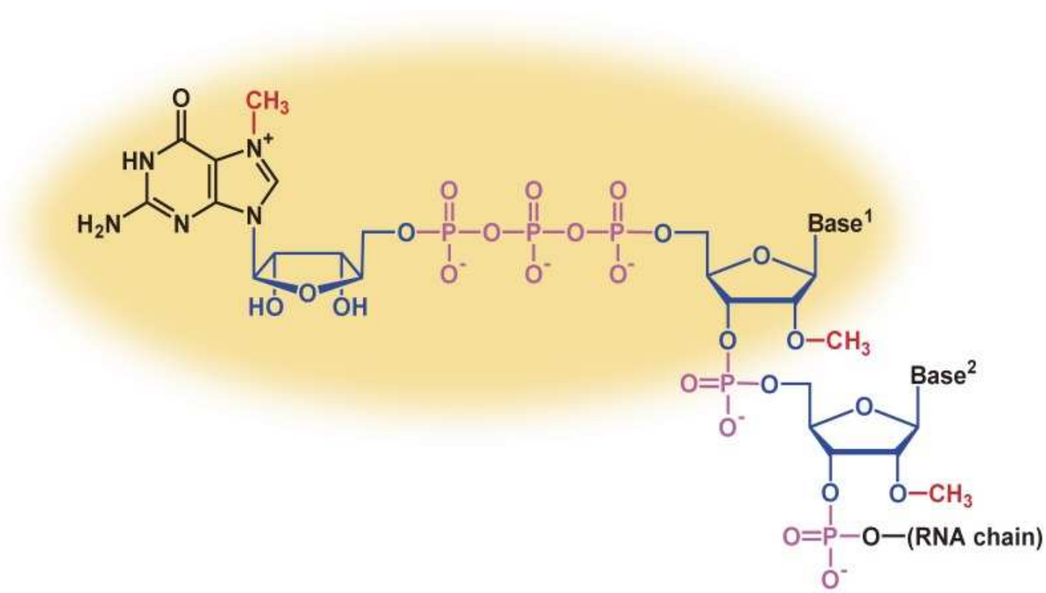


Figure 1.6 mRNA cap structure. Adapted from Furuichi, 2015. Inverted, triphosphorylated Guanosine is appended to the 5’ end of mRNAs a methyl moiety is appended. Two 2’O methylated nucleosides are incorporated adjacent to the 5’ guanosine.

far suggests m6A allows for short bursts of increased translation by dynamically regulating RNA levels and isoforms before being quickly degraded.[15] This modification dominates eukaryotic mRNAs and appears to elicit complex levels of gene expression regulation that may be critical for responding to different physiological events.

A far less abundant modification, m1A is hardly understood due to its rarity in eukaryotic mRNA. This modified nucleotide is positively charged and mostly occurs in rRNA and tRNAs, but also appears to be enriched at translation start sites and the first splice junction in mRNAs.[82] Little else is known about m1A other than that writer and eraser proteins have been identified, indicating m1A is dynamically regulated and therefore likely important for cellular function.

Ψ has been mapped with single-nucleotide resolution in mammalian and yeast mRNA and is dynamically regulated in response to stress, cellular growth phase, and metabolic conditions.[83,84] Due to its ability to impact RNA structure, Ψ may enhance translation initiation, RNA localization, and RNA interference, but strong evidence is lacking.[83] A recent study has demonstrated that mRNA Ψ in humans is enriched at splicing regions and has measurable

impacts on splicing efficiency.[85] The ubiquity and abundance of Ψ warrants further study into its molecular and cellular impacts.

Studies have identified m5C in diverse RNAs, and have established the ubiquity of m5C across life. The distribution of m5C across model species has placed m5C in eukaryotic and archaeal mRNAs, tRNAs, and rRNAs, while m5C in Bacteria seems to be exclusive to rRNA.[11,28] Hundreds, and in some studies thousands, of m5C sites have been mapped to human and mouse mRNAs where differential m5C-profiles are observed across cell and tissue types, however these studies often lack congruency.[35,86,87] To date, only one archaeal species, *Sulfolobus solfataricus*, has been assessed for m5C in mRNA where 14 sites are detected.[28]

We are only just beginning to uncover the levels and distribution of mRNA-m5C, and little is known regarding its functions. Studies have repeatedly correlated m5C with heat stress across Domains[88,89], potentially indicating its role in mitigating thermal insults. m5C does not impact Watson-Crick base pairing, but it does increase hydrophobicity of the RNA, and therefore may affect base stacking and overall RNA structure.[15] Gene ontology enrichment analysis of m5C-containing mRNAs tends to correlate m5C with basic cellular and metabolic pathways.[2,35] Studies have correlated m5C sites with binding sites for RNA regulatory proteins, such as Argonaute, splicing, and decay factors, suggesting m5C may impact protein activities.[90] Correlations have also been drawn between m5C-mRNAs and nuclear export[70] and decreased translation levels of m5C-containing reporter mRNAs[88]. The precise mechanisms by which m5C impacts molecular processes are currently unknown.

The distribution of m5C in mRNA depends on the species. In mammalian species, m5C sites in mRNA tend to be enriched near where the 5’ and 3’ untranslated regions (UTRs) meet the coding region.[91] In *Arabidopsis*, the evidence is conflicting with several studies placing m^5C in 3’ UTRs, 5’ UTRs or reporting a slight enrichment within coding sequences.[88,92,93] In zebrafish and rice, an enrichment has been observed within coding sequences.[89,94] Positional enrichment was not detected in *S. solfataricus* mRNAs.[28] These studies would suggest that the position of

m5C within the mRNA has differential effects on mRNA function. Although few researchers dispute the presence of m5C in mRNAs, these disparate results bring into question the reliability of different approaches and highlight the lack of standard data analysis practices geared towards eliminating false-positive m5C sites. There is still much to learn about the transcriptome-wide distribution of this highly conserved and ubiquitous modification.

Other non-coding RNA modifications

A subset of small noncoding RNAs (sncRNAs) have been identified in eukaryotes with important functions in RNA silencing pathways.[95] sncRNAs are typically < ~200 nt in length, and in higher eukaryotes, traditionally include microRNAs (miRNAs), short interfering RNAs (siRNAs), or P-element-induced wimpy testis interacting RNA (piRNAs).[96] These sncRNAs form complexes with and act as guide RNAs for Argonaute family proteins, and provide specificity to mRNA degradation and translation inhibition pathways.[34] The 3' ends of these sncRNAs are highly heterogeneous, but are sometimes 2'O methylated, which prevents poly-uridylation, a signal that would otherwise elicit the degradation of these sncRNAs.[97,98] In addition, 2'O methylation likely impacts the stability of sncRNAs by modulating their thermodynamic properties by increasing base stacking and structural rigidity.[99] Some miRNAs have been shown to harbor inosine in their seed sequence, which is generally thought to module target specificity[100], as inosine is an adenosine analog that instead base pairs with cytosine. Inosine-containing miRNAs provide tissue specific silencing of enzymes involved in uric acid production.[100] m6A is found in mature miRNA and is essential for miRNA processing, but its exact role is unclear.[101] Taken together, modifications to this subset of sncRNAs have profound impacts on gene expression across tissue and cell types *via* RNA silencing pathways.

Small nuclear RNAs (snRNAs) are typically ~150 nt and are major components of the spliceosome. Some snRNAs are heavily modified, mainly incorporating Ψ and 2'O methylation, and nearly all snRNAs contain a 2,2,7-trimethylguanosine cap.[102] In vertebrates, snRNA modification identity and position are highly conserved[102], demonstrating these modifications

increase cellular fitness. Moreover, most internal snRNA modifications are clustered in functionally important regions relevant to splicing,[103,104] suggesting modified nucleotides play a direct role in the splicing process. Modifications in these RNAs impact RNA-RNA and RNA-protein interactions in the spliceosome, and directly modulate catalytic splicing reactions.[105] snRNA modifications therefore impact pre-mRNA processing and mRNA isoform generation, and in-turn, gene expression.

There are many more non-coding RNAs that harbor modified residues, such as long non-coding RNAs (lncRNAs) and tRNA-derived small RNAs (tsRNAs). Although the function and distribution of modifications in these RNAs are largely unknown. Accumulation of tsRNAs in combination with m5C-deficient tRNAs impacts protein expression by exacerbating translation inaccuracy.[106,107] lncRNAs regulate transcription and chromatin remodeling, and multiple m6A sites have been mapped to lncRNAs.[108] The role of m6A in these transcripts remains unknown.

Thus far, we have only discussed a select few modifications, but the chemical diversity of nucleotides that constitute many types of RNA in all Domains is staggeringly complex. Nucleotide analogs result in altered chemical properties that modulate structure, stability, and intermolecular interactions of RNAs. In-turn, these often subtle alterations have profound impacts on biological processes such as cell cycle progression, stress response, DNA damage response, metabolism, viral-host interactions, apoptosis, cell differentiation, embryonic development, carcinogenesis and tumorigenesis, and inflammation.[2,14] RNA modifications are often essential, dynamic, and reversible, indicating their importance as a regulatory switch in response to physiological events. These post-transcriptional modifications are site-specifically generated by writer enzymes, the malfunction of which has been linked to severe human disease. Mass spectrometry, RNA-sequencing, and biochemical assays have revolutionized our understanding of RNA modifications, but there are still many unknowns regarding how the epitranscriptome is generated and the fate of RNAs once modified.

Writers and erasers of the epitranscriptome

Modifications are not distributed randomly, rather enzymes target specific nucleotides post-transcriptionally for modification. Eraser enzymes remove the initial modification installed by the writer enzyme. Together, writers and erasers dynamically maintain the epitranscriptome and are actively pursued targets for therapeutic intervention.[14,109] Several such enzymes have been identified, but the genesis of the significant majority of modifications, even those that have been the focus of intense study, are still unknown. Moreover, the strategies that target many RNA modifying enzymes to a specific nucleotide are elusive.

Methylation is the most common modification to decorate the transcriptome and is usually installed by a single enzymatic reaction.[3] RNA methyltransferases (RMTases) can install methyl moieties at many different positions on any of the four canonical nucleotides. RMTases appear to universally prefer S-adenosyl-L-methionine (SAM or AdoMet), likely due to favorable energetics.[110] The methyl group is transferred to the substrate by a RMTases resulting in a covalently methylated substrate and the byproduct, S-adenosyl-L-homocysteine (SAH). RMTases can be multi-domain and multi-subunit proteins but are minimally composed of the catalytic site(s), SAM-binding site(s), and RNA recognition site(s). The catalytic and SAM-binding sites are easily identified using computational approaches, but substrate recognition sites are far more difficult to discern.[111–113] DNA MTases recognize distinct sequence motifs as the primary strategy for site-specificity, but most RMTases have not been shown to target distinct RNA sequence motifs complex enough for site-specific modification, suggesting additional factors (i.e. RNA structure, protein-interactions, etc) must contribute to RNA targeting strategies. Current efforts to identify druggable RMTases often focus on structural characteristics related to substrate recognition[109] and provide opportunities to develop novel therapeutic routes to treat diseases of the epitranscriptome.

Methyltransferases fall into one of five structural classes (I-V), but most RMTases are

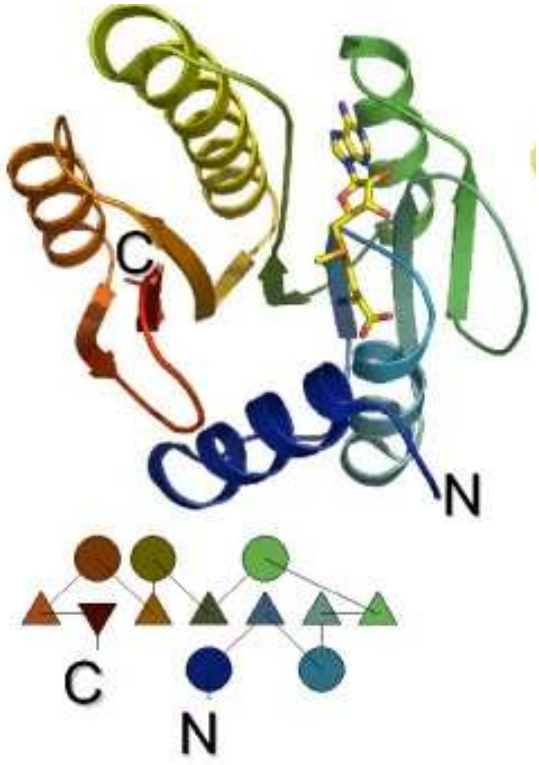


Figure 1.7 The basic structure of class I Rossmann fold, SAM-dependent methyltransferases. Adapted from Kozbial and Mushegian, 2005.

characterized by a class I Rossmann fold at its core. Enzymes from class I are comprised of seven β-strands with a topological switch-point at the center and a reversed β hairpin at the C-terminal end of the sheet (Figure 1.7).[114–116] Both sides of the β sheet are typically flanked by 2-3 α-helices, and the overall structure forms an αβα sandwich.[114] Although some class I MTases contain variants in this base-structure, most share this highly conserved fold. The other structural classes are less common, but can be described in a similar light with a β-sheet flanked or surrounded by α-helices, or some variation thereof. SAM-cofactor and RNA-substrate binding sites generally constitute two separate cavities that are linked by a short, narrow channel.[109] Dozens of RMTase structures have been experimentally solved and made available on the Protein Data Bank,[117] and although each contain highly conserved structural domains, their substrate recognition sites are not readily apparent or predictable.

Humans encode a staggering 57 known RMTases, 43 of which (~70%) adopt the class I

Rossmann fold.[109] m6A is predominantly installed by the METTL family of RMTases, where homologs have been identified in mammals, yeast, flies, and plants. METTL RMTases are known to form complexes with other proteins that regulate cellular levels of m6A.[118] The METTL3-METTL14 dimeric complex is responsible for the installation of the majority of the m6A residues in human mRNA.[109] METTL3-METTL14 methylates a DRACH nucleotide motif (D = A, G or T; R = A or G; H = A, C or U), but only a fraction of RNAs that contain this sequence are methylated by the METTL3-METTL14 complex.[81] The UAC**A**GAGAA consensus motif wasfound to be enriched in substrates targeted by METTL16[118], and this protein has also been shown to methylate a particular stem loop structure where the consensus sequence is disjoined (UAC**A**G, AGAA)[109]. These data indicate additional factors beyond sequence contribute to the substrate targeting strategies employed by these enzymes. FTO and ALKBH5 are m6A erasers. Both are α-ketoglutarate-dependent oxidases that mediate the demethylation of m6A. First, N6-hydroxy-methyl-adenosine is generated as an intermediate product before further oxidation results in N6-formyl-adenosine.[119] FTO demethylates m6A and m6Am on mRNA and snRNAs and m1A on tRNA and does not appear to target a strong sequence motif, but rather prefers single-stranded RNA.[68] ALKBH5 has been reported to demethylate m6,2A in rRNA within a consensus sequence.[68] These eraser enzymes are enriched in certain tissues and cell types, and given their essentiality,[68] it would not be surprising if additional cell- or tissue-specific m6A demethylases are identified.

m5C is installed by a highly conserved family of enzymes across all three Domains. In humans, m5C is mainly installed by the NOL1/NOP2/sun (NSUN) family and DNA methyltransferase 2 (DNMT2). DNMT2 was originally thought to install m5C in DNA based on homology to known DNA methyltransferases, but later studies showed it actually methylates tRNA.[120] m5C removal is mediated by ALKBH1 and TET resulting in the oxidative intermediate, C5-formyl-cytidine.[86] In many species, m5C is enriched in G-C rich regions, but a defined sequence motif has not been identified and strong sequence motifs have not been associated

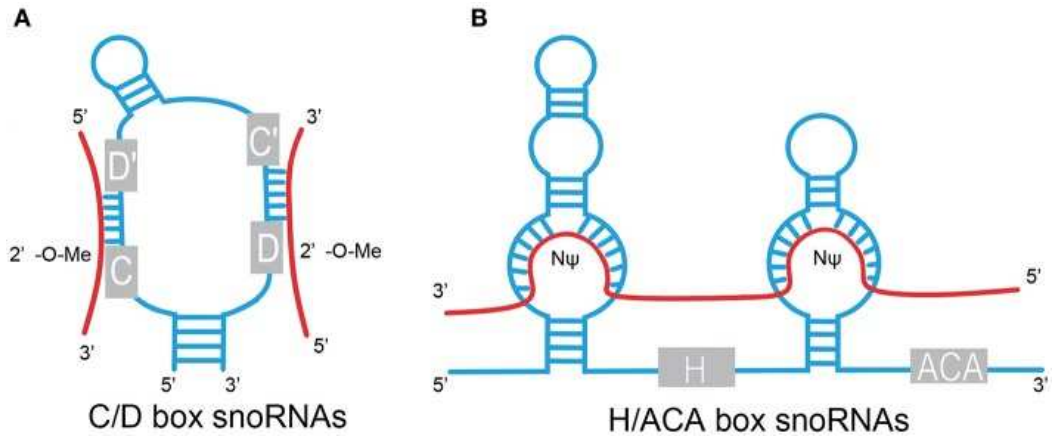


Figure 1.8 C/D box and H/ACA box snoRNAs. Adapted from Liang et al., 2019. (A) snoRNAs contain two conserved sequence elements, named box C, and box D. The elements upstream of the box D/D′ motifs are complementary to substrate RNAs and catalyze site-specific 2'-O-methylation. **(B)** H/ACA box snoRNAs contain box H, box ACA, and two pseudouridylation pockets. Pseudouridylation pockets are complementary to the substrate RNAs and responsible for pseudouridylation (NΨ). Red line = substrate RNAs.

with m5C writers.[35] As an example, NSUN2 has been shown to target 5'GGGN**C** at the apex of a stem-loop structure in mammalian mRNAs.[35] As a 3-4 nucleotide consensus sequence does not encode enough complexity to target enzymes to specific cytidines to the exclusion of all other non-targeted sites in the transcriptome, it is likely a consensus structure may also contribute to substrate recognition. Since m5C is abundant in tRNAs, some suggest that the consensus structure of m5C writers may resemble that of a stem-loop found in tRNAs.[121] Many m5C writers have been identified and studied, but little else is known about how these enzymes selectively methylate substrate RNAs nor the molecular functions and cellular impacts of individual m5C sites.

A much more defined strategy of modification targeting are those guided by small nucleolar RNAs (snoRNAs) in Eukarya and Archaea, but not Bacteria. snoRNAs are typically ~60-170 nt in length and associate with proteins to form the snoRNA-ribonucleoprotein complex (snoRNP). snoRNAs include H/ACA and C/D box RNAs (Figure 1.8) and act as a guide system[122] to target writer proteins mainly to the ribosome for installation of Ψ and 2'O methylation, respectively.[123] Archaea do not encapsulate a nucleus and so sRNAs are

equivalent to the eukaryotic snoRNAs. C/D box snoRNAs interact with fibrillarin, Nop56, Nop58, and 15.5kD.[123] Fibrillarin performs the methylation while Nop56, Nop58, and 15.5kD contribute to maturation, stability, and localization of the snoRNA.[123] Similarly, a set of proteins assemble with H/ACA box snoRNAs including NHP2, NOP10, GAR1, and pseudouridine synthase dyskerin. In both cases, the substrate targeting lies in sequence homology between the snoRNA and the substrate RNA targeted for methylation.[123] snoRNAs have been implicated in a wide range of cellular processes and aberrant expression of snoRNAs has been correlated with many different cancers.[124] Research has found that their expression can be detected in bodily fluids indicating these snoRNAs may potentially be used as cancer biomarkers.[124,125]

The epitranscriptome reflects the diversity of life

We now know that the intricate and expanding network of readers, writers, and erasers establish and interpret with the epitranscriptome and confer a dynamic layer of complexity in regulating cellular homeostasis. Improvements in mass spectrometry and RNA-sequencing methodologies continue to yield new discoveries and exciting hypotheses regarding the biological role of different modifications. The exact consequences of individual modifications are largely unsolved, but the epitranscriptomic profile across organelles, species, and Domains present clues about the molecular and fitness impacts of modifications in different contexts.

The impacts of diverse RNA modification profiles is perhaps best understood through the lens of highly conserved tRNAs. The tRNAmodviz web server[126] is an excellent tool to interactively visualize tRNA modifications. Although nearly all tRNAs adopt a canonical structure and are similar in sequence, the tRNA modification profile can differ significantly in composition and density. Eukaryotic model systems typically modify their tRNAs to a higher extent compared to bacteria, and mitochondrial and plastid tRNA are among the lowest modified tRNAs.[126] Although many tRNA modifications are conserved across life, some modifications are unique to one or two Domains; about half of tRNA modifications are unique to one Domain, about 1/5 are conserved in all three Domains, and another 1/5 are conserved in two Domains (Figure 1.9).[11]

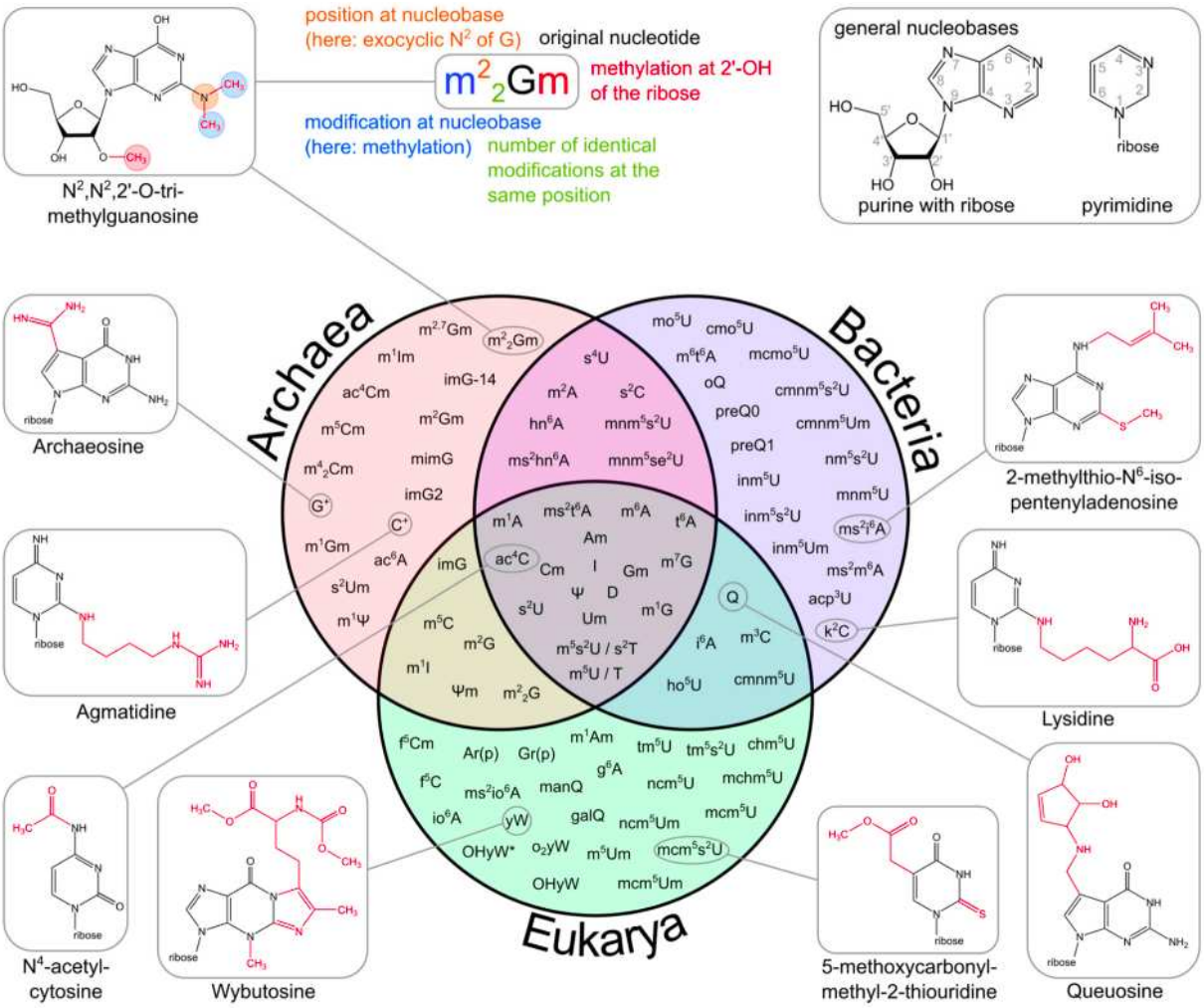


Figure 1.9 tRNA modifications across Domains of life. Adapted from Lorenz et. al., 2017.

Maintaining tRNA integrity is essential. It is tempting to speculate that different modifications impart features that perfectly adapt tRNAs to serve their respective environmental context. This hypothesis provides a rationale for the evolutionary acquisition of distinct tRNA modification profiles across organelles, species, and Domains. In many cases, different epitranscriptomic marks are readily apparent. As previously stated, m5C is excluded from bacterial tRNAs but highly abundant and often essential in eukaryotic and archaeal tRNAs, where m5C promotes structural stability, aminoacylation, and codon-anticodon recognition. Archaeosine (G+) is found in the D-arm and is entirely unique to archaeal tRNAs.[127] tRNA nucleotides methylated at the 2'

OH position are often substrates for additional modification in Archaea but not in Eukarya or Bacteria.[11] m1A is a common mitochondrial tRNA modification at position 9, and mitochondrial tRNAs are A- and U-rich, often leading to alternative folds that yield non-functional tRNAs.[128] tRNA-m1A9 strongly contributes to the cloverleaf structure and it is thought to help these tRNAs avoid misfolded states.[128] Mostly in nematodes but also in humans and yeast, mitochondrial tRNAs can adopt bizarre forms that lack D- or T-arms but are still functional.[129–131] It is unclear how these tRNAs retain a sufficient tertiary structure, but strong evidence suggests modifications play a critical role in forcing an L-shaped structure.[40] These are a small few examples that highlight the presence and function of differential epitranscriptomic profiles that have emerged to serve in unique contexts. It remains to be discovered how most modifications holistically impact cellular fitness in diverse RNAs and harsh environments.

Organisms thrive in virtually every environmental niche, even those previously thought to be inhospitable. Such organisms are classified as extremophiles due to their unique ability to thrive in environments that are hypersaline, hypotonic, extremely hot or cold, extremely acidic or basic, or those exposed to high levels of radiation. Archaea are ubiquitous across environmental niches, but almost all extremophiles are from the archaeal lineage.[132] Although relatively little is known about these organisms, unique RNA modification profiles are apparent in the select few archaea that have been studied. The mechanisms that allow most extremophiles to thrive in such harsh conditions are enigmatic, but likely multifaceted. Regardless, these organisms must adapt specific mechanisms to protect the integrity of their biomolecules. Modified nucleosides expand the function of RNA, and given the evidence thus far, it is reasonable to hypothesize that the epitranscriptome in extremophiles has adapted to accommodate extreme environments that would otherwise jeopardize critical RNAs.

Characterization of the role of specific modifications is complicated by their low abundance in Eukarya and Bacteria, while studies have shown higher overall abundance of select modifications in heat-loving archaea.[23] Hyperthermophiles, organisms that thrive at

extremely high temperatures, require adaptations of their biomolecules to prevent thermal denaturation. Several RNA modifications have already been shown to provide structural stability and increase the thermal tolerance across Domains of life.[89,127,133,134] Our preliminary data demonstrates that RNA modifications are relatively more abundant in hyperthermophiles compared to their mesophilic counterparts, indicating a role for RNA modifications to enhance thermophily in these organisms.[23] Archaea are the progenitors to Eukarya and share much of the same information processing machinery involved in DNA- and RNA-driven processes, often to the exclusion of bacteria.[55] Studies of archaeal RNA modification systems have a proven record of identifying eukaryotic factors and processes that underscore snoRNA, rRNA and tRNA modifications.[23,122] For these reasons, Archaea provide the ideal platform to study how the epitranscriptome is generated, dynamically maintained, and supports life in the extremes.

Select RNA modifications are abundant in heat-loving archaea

In most other organisms, high temperatures result in denaturation of critical biomolecules such as DNA, RNA, and proteins. Yet, many hyperthermophiles grow rapidly at physiological temperatures, indicating these organisms have acquired thermal adaptation to essential biomolecules. A higher abundance of salt-bridges, disulfide bridges, and hydrophobic interactions in heat-living organisms likely contribute to protein thermal stability.[135] Reverse gyrase leads to positive supercoiling of DNA and is universally conserved in thermophiles and hyperthermophiles, leading many to hypothesize this protein's involvement in stabilizing DNA against thermal insults.[136] However, this gene is non-essential, albeit, cells lacking reverse gyrase grow poorly. Similarly, distinct epitranscriptomic signatures have proven to promote thermophily, but the exact rules of thermophily are poorly defined. ***I hypothesize that select RNA modifications are abundant in T. kodakarensis and provide a strategy for hyperthermophilic growth by protecting RNA from thermal denaturation.***

The model hyperthermophilic archaeon, *Thermococcus kodakarensis*, has historically

served as the flagship platform for empirically defining the rules of thermophily. This is because *T. kodakarensis* is genetically tractable and grows at a wide range of temperatures, allowing for the temperature-dependent impacts of individual genes on cellular growth to be studied.[23,137] As previously discussed, certain tRNA modifications are subject to phylogenetic segregation, and the tRNA modification profile in hyperthermophilic archaea has been shown to directly contribute to hyperthermophilic growth. ***T. kodakarensis is therefore the ideal model species to address questions concerning the relationship between the epitranscriptome and thermophily.***

A tRNA modification unique to Archaea, G+, contains a large and complex chemical moiety, and its installation is a multistep, multienzyme pathway.[127] Eukarya and Bacteria partially encode this pathway which produces a related modification, queuosine (Q), in the anticodon loop of tRNAs. Its position suggests a role in translation regulation.[138] In Archaea, the position in the D-stem and bulky structure of G+ suggests a role in stabilizing tRNA structure *via* electrostatic interactions between the positively charged modification and the negatively charged RNA backbone.[139] Computational analysis indicates a stronger Levitt base pairing in G+-containing tRNAs, which would theoretically stabilize the tertiary L-structure.[140] When the G+ biosynthetic pathway is disrupted resulting in the absence of G+, *T. kodakarensis* experiences defective growth with increasing severity at higher temperatures.[127] These data suggest G+ provides much needed structural stability to tRNAs in otherwise strongly denaturing environments. Hypermodified nucleotides, those modified at multiple positions, are common in archaeal tRNAs, and several studies also indicate their temperature-dependent impact on cell growth across thermophilic and hyperthermophilic model organisms.[11]

Growth at cold temperatures equally challenges our understanding of biological systems. Low temperatures increase the rigidity of molecules, and it is common to see adaptations that improve the flexibility of proteins and membrane lipids in psychrophiles.[141] There are no noted changes in GC content or sequence diversity in most psychrophile tRNAs compared to those

encoded in mesophiles. Non-aromatic dihydrouridine are shown to increase RNA flexibility, and an increase in this particular modification is observed in psychrophile tRNAs while thermophilic organisms contain very low levels of non-aromatic dihydrouridine.[11] An intriguing study correlated the cold response with a marked increase in RNA A-to-I editing in octopus, indicating the recoding of mRNAs as a mechanism of cold adaptation.[142] The epitranscriptome of cold-loving organisms is poorly understood, and although limited, the noted epitranscriptomic differences between cold-loving and heat-loving organisms add credence to the proposed role for RNA modifications in temperature adaptation.

The epitranscriptomic composition in other types of RNAs is much less defined in extremophiles. In thermophiles and hyperthermophiles, there is a recorded increase in the number of 2'O methylated and ac4C residues in the ribosome[55,90], and my research also indicates a substantial enrichment in m5C residues in the *T. kodakarensis*. My research has also led to the identification of a novel modification, N4,N4-dimethylcytidine (m4,2C) in the ribosomal decoding center. In the helix 31 structure of 16S rRNA, m4,2C sits adjacent to m1Ψ and 2'OmG. Not only is this a unique epitranscriptomic marker, but also a hypermodified region compared to Bacteria or Eukarya where only 1 or 2 modified residues are typically present.[143] ***I have shown that a single m4,2C residue at C918 in the 16S rRNA and the m5C epitranscriptome holistically are essential to hyperthermophilic growth, adding a substantial body of evidence supporting the hypothesis that modifications stabilize RNA against thermal denaturation.***

The thermal stability of tRNAs and rRNAs is undoubtedly supported by the epitranscriptome, but thermal stability of mRNAs is also likely critical. The mRNA epitranscriptome in archaea is nearly entirely unmapped. Until recently, ac4C was the only modification mapped to *T. kodakarensis* mRNAs.[23] A single previous study mapped 14 m5C sites to the entire *Sulfolobus* epitranscriptome[28], a hyperthermophilic and acidophilic archaeon, suggesting m5C is likely rare within crenarchaeota. ***I reported for the first time m5C in T.***

kodakarensis mRNA, and m5C is abundantly present in coding and non-coding RNAs. Although m6A is abundant in eukaryotic mRNAs, I did not detect m6A in *T. kodakarensis* mRNA, rather m6A is exclusive to rRNA and/or tRNA. All ac4C sites are generated by Nat10, whereas I have discovered that the installation of m5C is dependent on at least 6 enzymes that potentially share partial redundancy. Enzymes that install m5C are dubbed RNA m5C methyltransferases (R5CMT). Cells deleted for Nat10[23] or one of the five R5CMT, and as previously mentioned, the enzyme that installed m4,2C, experience severe growth defects at increasing temperatures (unpublished). ***These data would strongly indicate that RNA modifications are not universally more abundant in hyperthermophilic models, rather, RNA modifications are selectively employed in hyperthermophiles and support life in the extremes.***

REFERENCES

1. Liu, Y., Makarova, K. S., Huang, W.-C., Wolf, Y. I., Nikolskaya, A. N., Zhang, X., Cai, M., Zhang, C.-J., Xu, W., Luo, Z., Cheng, L., Koonin, E. V. & Li, M. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593,** 553–557 (2021).
2. Wilkinson, E., Cui, Y.-H. & He, Y.-Y. Roles of RNA Modifications in Diverse Cellular Functions. *Front Cell Dev Biol* **10,** 828683 (2022).
3. Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., Kurkowska, M., Shirvanizadeh, N., Destefanis, E., Groza, P., Avşar, G., Romitelli, A., Pir, P., Dassi, E., Conticello, S. G., Aguilo, F. & Bujnicki, J. M. MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Res.* **50,** D231–D235 (2021).
4. Liyanage, V. R. B., Jarmasz, J. S., Murugeshan, N., Del Bigio, M. R., Rastegar, M. & Davie, J. R. DNA modifications: function and applications in normal and disease States. *Biology* **3,** 670–723 (2014).
5. Zatopek, K. M., Burkhart, B. W., Morgan, R. D., Gehring, A. M., Scott, K. A., Santangelo, T. J. & Gardner, A. F. The Hyperthermophilic Restriction-Modification Systems of Thermococcus kodakarensis Protect Genome Integrity. *Front. Microbiol.* **12,** 657356 (2021).
6. Reisenauer, A., Kahng, L. S., McCollum, S. & Shapiro, L. Bacterial DNA methylation: a cell cycle regulator? *J. Bacteriol.* **181,** 5135–5139 (1999).
7. Payelleville, A. & Brillard, J. Novel Identification of Bacterial Epigenetic Regulations Would Benefit From a Better Exploitation of Methylomic Data. *Front. Microbiol.* **12,** 685670 (2021).
8. Alseth, I., Dalhus, B. & Bjørås, M. Inosine in DNA and RNA. *Curr. Opin. Genet. Dev.* **26,** 116–123 (2014).
9. Hahm, J. Y., Park, J., Jang, E.-S. & Chi, S. W. 8-Oxoguanine: from oxidative damage to epigenetic and epitranscriptional modification. *Exp. Mol. Med.* **54,** 1626–1642 (2022).

10. Schaefer, M., Kapoor, U. & Jantsch, M. F. Understanding RNA modifications: the promises and technological bottlenecks of the ‘epitranscriptome’. *Open Biol.* **7,** 170077 (2017).
11. Lorenz, C., Lünse, C. E. & Mörl, M. tRNA Modifications: Impact on Structure and Thermal Adaptation. *Biomolecules* **7,** (2017).
12. Heraud-Farlow, J. E. & Walkley, C. R. What do editors do? Understanding the physiological functions of A-to-I RNA editing by adenosine deaminase acting on RNAs. *Open Biol.* **10,** 200085 (2020).
13. Jonkhout, N., Tran, J., Smith, M. A., Schonrock, N., Mattick, J. S. & Novoa, E. M. The RNA modification landscape in human disease. *RNA* **23,** 1754–1769 (2017).
14. Cayir, A. RNA modifications as emerging therapeutic targets. *Wiley Interdiscip. Rev. RNA* **13,** e1702 (2022).
15. Ontiveros, R. J., Stoute, J. & Liu, K. F. The chemical diversity of RNA modifications. *Biochem. J* **476,** 1227–1245 (2019).
16. Yan, T.-M., Pan, Y., Yu, M.-L., Hu, K., Cao, K.-Y. & Jiang, Z.-H. Full-Range Profiling of tRNA Modifications Using LC–MS/MS at Single-Base Resolution through a Site-Specific Cleavage Strategy. *Anal. Chem.* **93,** 1423–1432 (2021).
17. Petrov, D. P., Kaiser, S., Kaiser, S. & Jung, K. Opportunities and Challenges to Profile mRNA Modifications in Escherichia coli. *Chembiochem* **23,** e202200270 (2022).
18. Estevez, M., Li, R., Paul, B., Daneshvar, K., Mullen, A. C., Romerio, F. & Addepalli, B. Identification and mapping of post-transcriptional modifications on the HIV-1 antisense transcript Ast in human cells. *RNA* rna.079043.121 (2022).
19. Richter, F., Plehn, J. E., Bessler, L., Hertler, J., Jörg, M., Cirzi, C., Tuorto, F., Friedland, K. & Helm, M. RNA marker modifications reveal the necessity for rigorous preparation protocols to avoid artifacts in epitranscriptomic analysis. *Nucleic Acids Res.* **50,** 4201–4215 (2021).
20. Grosjean, H., Droogmans, L., Roovers, M. & Keith, G. Detection of enzymatic activity of transfer RNA modification enzymes using radiolabeled tRNA substrates. *Methods Enzymol.*

425, 55–101 (2007).

21. Roovers, M., Wouters, J., Bujnicki, J. M., Tricot, C., Stalon, V., Grosjean, H. & Droogmans, L. A primordial RNA modification enzyme: the case of tRNA (m1A) methyltransferase. *Nucleic Acids Res.* **32,** 465–476 (2004).
22. Chen, H., Yang, H., Zhu, X., Yadav, T., Ouyang, J., Truesdell, S. S., Tan, J., Wang, Y., Duan, M., Wei, L., Zou, L., Levine, A. S., Vasudevan, S. & Lan, L. m5C modification of mRNA serves a DNA damage code to promote homologous recombination. *Nat. Commun.* **11,** 2834 (2020).
23. 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., Washburn, M. P., Isobe, T., Santangelo, T. J., Shalev-Benami, M., Meier, J. L. & Schwartz, S. Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. *Nature* **583,** 638–643 (2020).
24. Zhang, Y., Lu, L. & Li, X. Detection technologies for RNA modifications. *Exp. Mol. Med.* **54,** 1601–1616 (2022).
25. Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., Sorek, R. & Rechavi, G. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485,** 201–206 (2012).
26. Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E. & Jaffrey, S. R. Comprehensive analysis of mRNA methylation reveals enrichment in 3’ UTRs and near stop codons. *Cell* **149,** 1635–1646 (2012).
27. Liu, N. & Pan, T. Probing RNA Modification Status at Single-Nucleotide Resolution in Total RNA. *Methods Enzymol.* **560,** 149–159 (2015).
28. Edelheit, S., Schwartz, S., Mumbach, M. R., Wurtzel, O. & Sorek, R. Transcriptome-wide

mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. *PLoS Genet.* **9,** e1003602 (2013).

29. Liu, C., Sun, H., Yi, Y., Shen, W., Li, K., Xiao, Y., Li, F., Li, Y., Hou, Y., Lu, B., Liu, W., Meng, H., Peng, J., Yi, C. & Wang, J. Absolute quantification of single-base m6A methylation in the mammalian transcriptome using GLORI. *Nat. Biotechnol.* **41,** 355–366 (2023).
30. Bao, Z., Li, T. & Liu, J. Determining RNA Natural Modifications and Nucleoside Analog-Labeled Sites by a Chemical/Enzyme-Induced Base Mutation Principle. *Molecules* **28,** (2023).
31. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13,** 278–289 (2015).
32. Zhao, X., Zhang, Y., Hang, D., Meng, J. & Wei, Z. Detecting RNA modification using direct RNA sequencing: A systematic review. *Comput. Struct. Biotechnol. J.* **20,** 5740–5749 (2022).
33. Chen, Y., Davidson, N. M., Wan, Y. K., Patel, H., Yao, F., Low, H. M., Hendra, C., Watten, L., Sim, A., Sawyer, C., Iakovleva, V., Lee, P. L., Xin, L., Ng, H. E. V., Loo, J. M., Ong, X., Ng, H. Q. A., Wang, J., Koh, W. Q. C., Poon, S. Y. P., Stanojevic, D., Tran, H.-D., Lim, K. H. E., Toh, S. Y., Ewels, P. A., Ng, H.-H., Gopalakrishna Iyer, N., Thiery, A., Chng, W. J., Chen, L., DasGupta, R., Sikic, M., Chan, Y.-S., Tan, B. O. P., Wan, Y., Tam, W. L., Yu, Q., Khor, C. C., Wüstefeld, T., Pratanwanich, P. N., Love, M. I., Goh, W. S. S., Ng, S. B., Oshlack, A., Göke, J. & SG-NEx consortium. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. *bioRxiv* 2021.04.21.440736 (2021). doi:10.1101/2021.04.21.440736
34. Kim, Y.-K., Heo, I. & Kim, V. N. Modifications of small RNAs and their associated proteins. *Cell* **143,** 703–709 (2010).
35. Huang, T., Chen, W., Liu, J., Gu, N. & Zhang, R. Genome-wide identification of mRNA 5-methylcytosine in mammals. *Nat. Struct. Mol. Biol.* **26,** 380–388 (2019).

36. Phizicky, E. M. & Hopper, A. K. tRNA biology charges to the front. *Genes Dev.* **24,** 1832–1860 (2010).
37. Pereira, M., Francisco, S., Varanda, A. S., Santos, M., Santos, M. A. S. & Soares, A. R. Impact of tRNA Modifications and tRNA-Modifying Enzymes on Proteostasis and Human Disease. *Int. J. Mol. Sci.* **19,** (2018).
38. Salinas-Giegé, T., Giegé, R. & Giegé, P. tRNA biology in mitochondria. *Int. J. Mol. Sci.* **16,** 4518–4559 (2015).
39. Rich, A. & RajBhandary, U. L. Transfer RNA: molecular structure, sequence, and properties. *Annu. Rev. Biochem.* **45,** 805–860 (1976).
40. Sakurai, M., Ohtsuki, T. & Watanabe, K. Modification at position 9 with 1-methyladenosine is crucial for structure and function of nematode mitochondrial tRNAs lacking the entire T-arm. *Nucleic Acids Res.* **33,** 1653–1661 (2005).
41. Rife, J. P., Cheng, C. S., Moore, P. B. & Strobel, S. A. N 2-methylguanosine is iso-energetic with guanosine in RNA duplexes and GNRA tetraloops. *Nucleic Acids Res.* **26,** 3640–3644 (1998).
42. Edqvist, J., Grosjean, H. & Stråby, K. B. Identity elements for N2-dimethylation of guanosine-26 in yeast tRNAs. *Nucleic Acids Res.* **20,** 6575–6581 (1992).
43. Davis, D. R. Stabilization of RNA stacking by pseudouridine. *Nucleic Acids Res.* **23,** 5020–5026 (1995).
44. Davis, D. R., Veltri, C. A. & Nielsen, L. An RNA model system for investigation of pseudouridine stabilization of the codon-anticodon interaction in tRNALys, tRNAHis and tRNATyr. *J. Biomol. Struct. Dyn.* **15,** 1121–1132 (1998).
45. Finet, O., Yague-Sanz, C., Marchand, F. & Hermand, D. The Dihydrouridine landscape from tRNA to mRNA: a perspective on synthesis, structural impact and function. *RNA Biol.* **19,** 735–750 (2022).
46. Bohnsack, K. E., Höbartner, C. & Bohnsack, M. T. Eukaryotic 5-methylcytosine (m5C) RNA

Methyltransferases: Mechanisms, Cellular Functions, and Links to Disease. *Genes* **10,** 102 (2019).

47. Agris, P., Guenther, R., Sochacka, E., Newman, W., Czerwińska, G., Liu, G., Ye, W. & Malkiewicz, A. Thermodynamic contribution of nucleoside modifications to yeast tRNA\^ Phe anticodon stem loop analogs. *Acta Biochim. Pol.* **46,** 163–172 (1999).
48. Björk, G. R., Durand, J. M., Hagervall, T. G., Leipuviene, R., Lundgren, H. K., Nilsson, K., Chen, P., Qian, Q. & Urbonavicius, J. Transfer RNA modification: influence on translational frameshifting and metabolism. *FEBS Lett.* **452,** 47–51 (1999).
49. Phelps, S. S., Malkiewicz, A., Agris, P. F. & Joseph, S. Modified Nucleotides in tRNALys and tRNAVal are Important for Translocation. *J. Mol. Biol.* **338,** 439–444 (2004).
50. Liang, X.-H., Liu, Q. & Fournier, M. J. Loss of rRNA modifications in the decoding center of the ribosome impairs translation and strongly delays pre-rRNA processing. *RNA* **15,** 1716–1728 (2009).
51. Arimbasseri, A. G., Blewett, N. H., Iben, J. R., Lamichhane, T. N., Cherkasova, V., Hafner, M. & Maraia, R. J. RNA Polymerase III Output Is Functionally Linked to tRNA Dimethyl-G26 Modification. *PLoS Genet.* **11,** e1005671 (2015).
52. Kurata, S., Weixlbaumer, A., Ohtsuki, T., Shimazaki, T., Wada, T., Kirino, Y., Takai, K., Watanabe, K., Ramakrishnan, V. & Suzuki, T. Modified uridines with C5-methylene substituents at the first position of the tRNA anticodon stabilize U\textperiodcentered G wobble pairing during decoding. *J. Biol. Chem.* **283,** 18801–18811 (2008).
53. Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M. & Yusupova, G. New structural insights into the decoding mechanism: Translation infidelity via a G·U pair with Watson–Crick geometry. *FEBS Lett.* **587,** 1848–1857 (2013).
54. Saraiya, A. A., Lamichhane, T. N., Chow, C. S., SantaLucia, J., Jr & Cunningham, P. R. Identification and role of functionally important motifs in the 970 loop of Escherichia coli 16S ribosomal RNA. *J. Mol. Biol.* **376,** 645–657 (2008).

55. Dennis, P. P., Tripp, V., Lui, L., Lowe, T. & Randau, L. C/D box sRNA-guided 2'-O-methylation patterns of archaeal rRNA molecules. *BMC Genomics* **16,** 632 (2015).
56. Burakovsky, D. E., Prokhorova, I. V., Sergiev, P. V., Milón, P., Sergeeva, O. V., Bogdanov, A. A., Rodnina, M. V. & Dontsova, O. A. Impact of methylations of m2G966/m5C967 in 16S rRNA on bacterial fitness and translation initiation. *Nucleic Acids Res.* **40,** 7885–7895 (2012).
57. Andersen, N. M. & Douthwaite, S. YebU is a m5C methyltransferase specific for 16 S rRNA nucleotide 1407. *J. Mol. Biol.* **359,** 777–786 (2006).
58. Chow, C. S., Lamichhane, T. N. & Mahto, S. K. Expanding the nucleotide repertoire of the ribosome with post-transcriptional modifications. *ACS Chem. Biol.* **2,** 610–619 (2007).
59. Lane, B. G., Ofengand, J. & Gray, M. W. Pseudouridine in the large-subunit (23 S-like) ribosomal RNA. The site of peptidyl transfer in the ribosome? *FEBS Lett.* **302,** 1–4 (1992).
60. Schaenzer, A. J. & Wright, G. D. Antibiotic Resistance by Enzymatic Modification of Antibiotic Targets. *Trends Mol. Med.* **26,** 768–782 (2020).
61. Toh, S.-M. & Mankin, A. S. An indigenous posttranscriptional modification in the ribosomal peptidyl transferase center confers resistance to an array of protein synthesis inhibitors. *J. Mol. Biol.* **380,** 593–597 (2008).
62. Sergeeva, O. V., Bogdanov, A. A. & Sergiev, P. V. What do we know about ribosomal RNA methylation in Escherichia coli? *Biochimie* **117,** 110–118 (2015).
63. Tsai, K., Stojković, V., Noda-Garcia, L., Young, I. D., Myasnikov, A. G., Kleinman, J., Palla, A., Floor, S. N., Frost, A., Fraser, J. S., Tawfik, D. S. & Fujimori, D. G. Directed evolution of the rRNA methylating enzyme Cfr reveals molecular basis of antibiotic resistance. *Elife* **11,** e70017 (2022).
64. Liao, H., Gaur, A., McConie, H., Shekar, A., Wang, K., Chang, J. T., Breton, G. & Denicourt, C. Human NOP2/NSUN1 regulates ribosome biogenesis through non-catalytic complex formation with box C/D snoRNPs. *Nucleic Acids Res.* **50,** 10695–10716 (2022).

65. Van Haute, L., Hendrick, A. G., D'Souza, A. R., Powell, C. A., Rebelo-Guiomar, P., Harbour, M. E., Ding, S., Fearnley, I. M., Andrews, B. & Minczuk, M. METTL15 introduces N4-methylcytidine into human mitochondrial 12S rRNA and is required for mitoribosome biogenesis. *Nucleic Acids Res.* **47,** 10267–10281 (2019).

66. Sloan, K. E., Warda, A. S., Sharma, S., Entian, K.-D., Lafontaine, D. L. J. & Bohnsack, M. T. Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biol.* **14,** 1138–1152 (2017).

67. Piekna-Przybylska, D., Przybylski, P., Baudin-Baillieu, A., Rousset, J.-P. & Fournier, M. J. Ribosome Performance Is Enhanced by a Rich Cluster of Pseudouridines in the A-site Finger Region of the Large Subunit*. *J. Biol. Chem.* **283,** 26026–26036 (2008).

68. Yang, Y., Hsu, P. J., Chen, Y.-S. & Yang, Y.-G. Dynamic transcriptomic m6A decoration: writers, erasers, readers and functions in RNA metabolism. *Cell Res.* **28,** 616–624 (2018).

69. Trixl, L. & Lusser, A. The dynamic RNA modification 5-methylcytosine and its emerging role as an epitranscriptomic mark. *Wiley Interdiscip. Rev. RNA* **10,** e1510 (2019).

70. Yang, X., Yang, Y., Sun, B.-F., Chen, Y.-S., Xu, J.-W., Lai, W.-Y., Li, A., Wang, X., Bhattarai, D. P., Xiao, W., Sun, H.-Y., Zhu, Q., Ma, H.-L., Adhikari, S., Sun, M., Hao, Y.-J., Zhang, B., Huang, C.-M., Huang, N., Jiang, G.-B., Zhao, Y.-L., Wang, H.-L., Sun, Y.-P. & Yang, Y.-G. 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m5C reader. *Cell Res.* **27,** 606–625 (2017).

71. Li, A., Chen, Y.-S., Ping, X.-L., Yang, X., Xiao, W., Yang, Y., Sun, H.-Y., Zhu, Q., Baidya, P., Wang, X., Bhattarai, D. P., Zhao, Y.-L., Sun, B.-F. & Yang, Y.-G. Cytoplasmic m6A reader YTHDF3 promotes mRNA translation. *Cell Res.* **27,** 444–447 (2017).

72. Zaccara, S., Ries, R. J. & Jaffrey, S. R. Reading, writing and erasing mRNA methylation. *Nat. Rev. Mol. Cell Biol.* **20,** 608–624 (2019).

73. Furuichi, Y. Discovery of m(7)G-cap in eukaryotic mRNAs. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **91,** 394–409 (2015).

74. Bail, S. & Kiledjian, M. Tri- to be mono- for bacterial mRNA decay. *Structure* **17,** 317–319 (2009).

75. Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M. & Pan, T. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* **518,** 560–564 (2015).

76. Roost, C., Lynch, S. R., Batista, P. J., Qu, K., Chang, H. Y. & Kool, E. T. Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification. *J. Am. Chem. Soc.* **137,** 2107–2115 (2015).

77. Alarcón, C. R., Goodarzi, H., Lee, H., Liu, X., Tavazoie, S. & Tavazoie, S. F. HNRNPA2B1 Is a Mediator of m(6)A-Dependent Nuclear RNA Processing Events. *Cell* **162,** 1299–1308 (2015).

78. Adhikari, S., Xiao, W., Zhao, Y.-L. & Yang, Y.-G. m(6)A: Signaling for mRNA splicing. *RNA Biol.* **13,** 756–759 (2016).

79. Geula, S., Moshitch-Moshkovitz, S., Dominissini, D., Mansour, A. A., Kol, N., Salmon-Divon, M., Hershkovitz, V., Peer, E., Mor, N., Manor, Y. S., Ben-Haim, M. S., Eyal, E., Yunger, S., Pinto, Y., Jaitin, D. A., Viukov, S., Rais, Y., Krupalnik, V., Chomsky, E., Zerbib, M., Maza, I., Rechavi, Y., Massarwa, R., Hanna, S., Amit, I., Levanon, E. Y., Amariglio, N., Stern-Ginossar, N., Novershtern, N., Rechavi, G. & Hanna, J. H. Stem cells. m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* **347,** 1002–1006 (2015).

80. Zhao, X., Yang, Y., Sun, B.-F., Shi, Y., Yang, X., Xiao, W., Hao, Y.-J., Ping, X.-L., Chen, Y.-S., Wang, W.-J., Jin, K.-X., Wang, X., Huang, C.-M., Fu, Y., Ge, X.-M., Song, S.-H., Jeong, H. S., Yanagisawa, H., Niu, Y., Jia, G.-F., Wu, W., Tong, W.-M., Okamoto, A., He, C., Rendtlew Danielsen, J. M., Wang, X.-J. & Yang, Y.-G. FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Res.* **24,** 1403–1419 (2014).

81. He, P. C. & He, C. m6 A RNA methylation: from mechanisms to therapeutic potential.

EMBO J. **40,** e105977 (2021).

82. Dominissini, D., Nachtergaele, S., Moshitch-Moshkovitz, S., Peer, E., Kol, N., Ben-Haim, M. S., Dai, Q., Di Segni, A., Salmon-Divon, M., Clark, W. C., Zheng, G., Pan, T., Solomon, O., Eyal, E., Hershkovitz, V., Han, D., Doré, L. C., Amariglio, N., Rechavi, G. & He, C. The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA. *Nature* **530,** 441–446 (2016).
83. Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M. & Gilbert, W. V. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* **515,** 143–146 (2014).
84. Li, X., Zhu, P., Ma, S., Song, J., Bai, J., Sun, F. & Yi, C. Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat. Chem. Biol.* **11,** 592–597 (2015).
85. Martinez, N. M., Su, A., Burns, M. C., Nussbacher, J. K., Schaening, C., Sathe, S., Yeo, G. W. & Gilbert, W. V. Pseudouridine synthases modify human pre-mRNA co-transcriptionally and affect pre-mRNA processing. *Mol. Cell* **82,** 645–659.e9 (2022).
86. Li, M., Tao, Z., Zhao, Y., Li, L., Zheng, J., Li, Z. & Chen, X. 5-methylcytosine RNA methyltransferases and their potential roles in cancer. *J. Transl. Med.* **20,** 214 (2022).
87. Heissenberger, C., Liendl, L., Nagelreiter, F., Gonskikh, Y., Yang, G., Stelzer, E. M., Krammer, T. L., Micutkova, L., Vogt, S., Kreil, D. P., Sekot, G., Siena, E., Poser, I., Harreither, E., Linder, A., Ehret, V., Helbich, T. H., Grillari-Voglauer, R., Jansen-Dürr, P., Koš, M., Polacek, N., Grillari, J. & Schosserer, M. Loss of the ribosomal RNA methyltransferase NSUN5 impairs global protein synthesis and normal growth. *Nucleic Acids Res.* **47,** 11807–11825 (2019).
88. Yang, L., Perrera, V., Saplaoura, E., Apelt, F., Bahin, M., Kramdi, A., Olas, J., Mueller-Roeber, B., Sokolowska, E., Zhang, W., Li, R., Pitzalis, N., Heinlein, M., Zhang, S., Genovesio, A., Colot, V. & Kragler, F. m5C Methylation Guides Systemic Transport of Messenger RNA over Graft Junctions in Plants. *Curr. Biol.* **29,** 2465–2476.e5 (2019).

89. Tang, Y., Gao, C.-C., Gao, Y., Yang, Y., Shi, B., Yu, J.-L., Lyu, C., Sun, B.-F., Wang, H.-L., Xu, Y., Yang, Y.-G. & Chong, K. OsNSUN2-Mediated 5-Methylcytosine mRNA Modification Enhances Rice Adaptation to High Temperature. *Dev. Cell* **53,** 272–286.e7 (2020).
90. Boo, S. H. & Kim, Y. K. The emerging role of RNA modifications in the regulation of mRNA stability. *Exp. Mol. Med.* **52,** 400–408 (2020).
91. Sibbritt, T., Patel, H. R. & Preiss, T. Mapping and significance of the mRNA methylome. *Wiley Interdiscip. Rev. RNA* **4,** 397–422 (2013).
92. Cui, X., Liang, Z., Shen, L., Zhang, Q., Bao, S., Geng, Y., Zhang, B., Leo, V., Vardy, L. A., Lu, T., Gu, X. & Yu, H. 5-Methylcytosine RNA Methylation in Arabidopsis Thaliana. *Mol. Plant* **10,** 1387–1399 (2017).
93. David, R., Burgess, A., Parker, B., Li, J., Pulsford, K., Sibbritt, T., Preiss, T. & Searle, I. R. Transcriptome-Wide Mapping of RNA 5-Methylcytosine in Arabidopsis mRNAs and Noncoding RNAs. *Plant Cell* **29,** 445–460 (2017).
94. Xue, S., Xu, H., Sun, Z., Shen, H., Chen, S., Ouyang, J., Zhou, Q., Hu, X. & Cui, H. Depletion of TRDMT1 affects 5-methylcytosine modification of mRNA and inhibits HEK293 cell proliferation and migration. *Biochem. Biophys. Res. Commun.* **520,** 60–66 (2019).
95. Xiong, Q. & Zhang, Y. Small RNA modifications: regulatory molecules and potential applications. *J. Hematol. Oncol.* **16,** 64 (2023).
96. Zhang, Z., Zhang, J., Diao, L. & Han, L. Small non-coding RNAs in human cancer: function, clinical utility, and characterization. *Oncogene* **40,** 1570–1577 (2021).
97. Ren, G., Xie, M., Zhang, S., Vinovskis, C., Chen, X. & Yu, B. Methylation protects microRNAs from an AGO1-associated activity that uridylates 5' RNA fragments generated by AGO1 cleavage. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 6365–6370 (2014).
98. Ji, L. & Chen, X. Regulation of small RNA stability: methylation and beyond. *Cell Res.* **22,** 624–636 (2012).
99. Kumar, S., Mapa, K. & Maiti, S. Understanding the effect of locked nucleic acid and

2’-O-methyl modification on the hybridization thermodynamics of a miRNA-mRNA pair in the presence and absence of AfPiwi protein. *Biochemistry* **53,** 1607–1615 (2014).

100. Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A. G. & Nishikura, K. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315,** 1137–1140 (2007).

101. Konno, M., Koseki, J., Asai, A., Yamagata, A., Shimamura, T., Motooka, D., Okuzaki, D., Kawamoto, K., Mizushima, T., Eguchi, H., Takiguchi, S., Satoh, T., Mimori, K., Ochiya, T., Doki, Y., Ofusa, K., Mori, M. & Ishii, H. Distinct methylation levels of mature microRNAs in gastrointestinal cancers. *Nat. Commun.* **10,** 1–7 (2019).

102. Morais, P., Adachi, H. & Yu, Y.-T. Spliceosomal snRNA Epitranscriptomics. *Front. Genet.* **12,** 652129 (2021).

103. Aoyama, T., Yamashita, S. & Tomita, K. Mechanistic insights into m6A modification of U6 snRNA by human METTL16. *Nucleic Acids Res.* **48,** 5157–5168 (2020).

104. Adachi, H. & Yu, Y.-T. Insight into the mechanisms and functions of spliceosomal snRNA pseudouridylation. *World J. Biol. Chem.* **5,** 398–408 (2014).

105. Karijolich, J. & Yu, Y.-T. Spliceosomal snRNA modifications and their function. *RNA Biol.* **7,** 192–204 (2010).

106. Tuorto, F., Liebers, R., Musch, T., Schaefer, M., Hofmann, S., Kellner, S., Frye, M., Helm, M., Stoecklin, G. & Lyko, F. RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat. Struct. Mol. Biol.* **19,** 900–905 (2012).

107. Blanco, S., Dietmann, S., Flores, J. V., Hussain, S., Kutter, C., Humphreys, P., Lukk, M., Lombard, P., Treps, L., Popis, M., Kellner, S., Hölter, S. M., Garrett, L., Wurst, W., Becker, L., Klopstock, T., Fuchs, H., Gailus-Durner, V., Hrabě de Angelis, M., Káradóttir, R. T., Helm, M., Ule, J., Gleeson, J. G., Odom, D. T. & Frye, M. Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. *EMBO J.* **33,** 2020–2039 (2014).

108. Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X., Dai,

Q., Chen, W. & He, C. A METTL3–METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.* **10,** 93–95 (2013).

109. Schapira, M. Structural Chemistry of Human RNA Methyltransferases. *ACS Chem. Biol.* **11,** 575–582 (2016).
110. Bartke, T. & Schneider, R. You are what you eat - How nutrition and metabolism shape the genome through epigenetics. *Mol Metab* **38,** 100987 (2020).
111. Bujnicki, J. M. & Rychlewski, L. RNA:(guanine-N2) methyltransferases RsmC/RsmD and their homologs revisited--bioinformatic analysis and prediction of the active site based on the uncharacterized Mj0882 protein structure. *BMC Bioinformatics* **3,** 10 (2002).
112. Tkaczuk, K. L., Dunin-Horkawicz, S., Purta, E. & Bujnicki, J. M. Structural and evolutionary bioinformatics of the SPOUT superfamily of methyltransferases. *BMC Bioinformatics* **8,** 73 (2007).
113. Petrossian, T. & Clarke, S. Bioinformatic Identification of Novel Methyltransferases. *Epigenomics* **1,** 163–175 (2009).
114. Schubert, H. L., Blumenthal, R. M. & Cheng, X. Many paths to methyltransfer: a chronicle of convergence. *Trends Biochem. Sci.* **28,** 329–335 (2003).
115. Martin, J. L. & McMillan, F. M. SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold. *Curr. Opin. Struct. Biol.* **12,** 783–793 (2002).
116. Kozbial, P. Z. & Mushegian, A. R. Natural history of S-adenosylmethionine-binding proteins. *BMC Struct. Biol.* **5,** 19 (2005).
117. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
118. Oerum, S., Meynier, V., Catala, M. & Tisné, C. A comprehensive review of m6A/m6Am RNA methyltransferase structures. *Nucleic Acids Res.* **49,** 7239–7255 (2021).
119. Fu, Y., Jia, G., Pang, X., Wang, R. N., Wang, X., Li, C. J., Smemo, S., Dai, Q., Bailey, K. A.,

Nobrega, M. A., Han, K.-L., Cui, Q. & He, C. FTO-mediated formation of N6-hydroxymethyladenosine and N6-formyladenosine in mammalian RNA. *Nat. Commun.* **4,** 1798 (2013).

120. Jeltsch, A., Ehrenhofer-Murray, A., Jurkowski, T. P., Lyko, F., Reuter, G., Ankri, S., Nellen, W., Schaefer, M. & Helm, M. Mechanism and biological role of Dnmt2 in Nucleic Acid Methylation. *RNA Biol.* **14,** 1108–1123 (2017).

121. Liu, R.-J., Long, T., Li, J., Li, H. & Wang, E.-D. Structural basis for substrate binding and catalytic mechanism of a human RNA:m5C methyltransferase NSun6. *Nucleic Acids Res.* **45,** 6684–6697 (2017).

122. Lui, L. & Lowe, T. Small nucleolar RNAs and RNA-guided post-transcriptional modification. *Essays Biochem.* **54,** 53–77 (2013).

123. Liang, J., Wen, J., Huang, Z., Chen, X.-P., Zhang, B.-X. & Chu, L. Small Nucleolar RNAs: Insight Into Their Function in Cancer. *Front. Oncol.* **9,** 587 (2019).

124. Thorenoor, N. & Slaby, O. Small nucleolar RNAs functioning and potential roles in cancer. *Tumour Biol.* **36,** 41–53 (2015).

125. Dimitrova, D. G., Teysset, L. & Carré, C. RNA 2'-O-Methylation (Nm) Modification in Human Diseases. *Genes* **10,** (2019).

126. Machnicka, M. A., Olchowik, A., Grosjean, H. & Bujnicki, J. M. Distribution and frequencies of post-transcriptional modifications in tRNAs. *RNA Biol.* **11,** 1619–1629 (2014).

127. Turner, B., Burkhart, B. W., Weidenbach, K., Ross, R., Limbach, P. A., Schmitz, R. A., de Crécy-Lagard, V., Stedman, K. M., Santangelo, T. J. & Iwata-Reuyl, D. Archaeosine Modification of Archaeal tRNA: Role in Structural Stabilization. *J. Bacteriol.* **202,** (2020).

128. Helm, M., Giegé, R. & Florentz, C. A Watson-Crick base-pair-disrupting methyl group (m1A9) is sufficient for cloverleaf folding of human mitochondrial tRNALys. *Biochemistry* **38,** 13338–13346 (1999).

129. Wende, S., Platzer, E. G., Jühling, F., Pütz, J., Florentz, C., Stadler, P. F. & Mörl, M.

Biological evidence for the world's smallest tRNAs. *Biochimie* **100,** 151–158 (2014).

130. Wolstenholme, D. R., Macfarlane, J. L., Okimoto, R., Clary, D. O. & Wahleithner, J. A. Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 1324–1328 (1987).

131. Messmer, M., Pütz, J., Suzuki, T., Suzuki, T., Sauter, C., Sissler, M. & Catherine, F. Tertiary network in mammalian mitochondrial tRNAAsp revealed by solution probing and phylogeny. *Nucleic Acids Res.* **37,** 6881–6895 (2009).

132. Rampelotto, P. H. Extremophiles and extreme environments. *Life* **3,** 482–485 (2013).

133. Navarro, I. C., Tuorto, F., Jordan, D., Legrand, C., Price, J., Braukmann, F., Hendrick, A. G., Akay, A., Kotter, A., Helm, M., Lyko, F. & Miska, E. A. Translational adaptation to heat stress is mediated by RNA 5-methylcytosine in Caenorhabditis elegans. *EMBO J.* **40,** e105496 (2021).

134. Lui, L. M., Uzilov, A. V., Bernick, D. L., Corredor, A., Lowe, T. M. & Dennis, P. P. Methylation guide RNA evolution in archaea: structure, function and genomic organization of 110 C/D box sRNA families across six Pyrobaculum species. *Nucleic Acids Res.* **46,** 5678–5691 (2018).

135. Wu, H., Chen, Q., Zhang, W. & Mu, W. Overview of strategies for developing high thermostability industrial enzymes: Discovery, mechanism, modification and challenges. *Crit. Rev. Food Sci. Nutr.* **63,** 2057–2073 (2023).

136. Heine, M. & Chandra, S. B. C. The linkage between reverse gyrase and hyperthermophiles: A review of their invariable association. *The Journal of Microbiology* **47,** 229–234 (2009).

137. Gehring, A. M., Sanders, T. J. & Santangelo, T. J. Markerless Gene Editing in the Hyperthermophilic Archaeon Thermococcus kodakarensis. *Bio Protoc* **7,** (2017).

138. Rakovich, T., Boland, C., Bernstein, I., Chikwana, V. M., Iwata-Reuyl, D. & Kelly, V. P. Queuosine deficiency in eukaryotes compromises tyrosine production through increased tetrahydrobiopterin oxidation. *J. Biol. Chem.* **286,** 19354–19363 (2011).

139. Gregson, J. M., Crain, P. F., Edmonds, C. G., Gupta, R., Hashizume, T., Phillipson, D. W. & McCloskey, J. A. Structure of the archaeal transfer RNA nucleoside G*-15 (2-amino-4,7-dihydro- 4-oxo-7-beta-D-ribofuranosyl-1H-pyrrolo[2,3-d]pyrimidine-5-carboximi dam ide (archaeosine)). *J. Biol. Chem.* **268,** 10076–10086 (1993).

140. Oliva, R., Tramontano, A. & Cavallo, L. Mg2+ binding and archaeosine modification stabilize the G15 C48 Levitt base pair in tRNAs. *RNA* **13,** 1427–1436 (2007).

141. Collins, T. & Margesin, R. Psychrophilic lifestyles: mechanisms of adaptation and biotechnological tools. *Appl. Microbiol. Biotechnol.* **103,** 2857–2871 (2019).

142. Birk, M. A., Liscovitch-Brauer, N., Dominguez, M. J., McNeme, S., Yue, Y., Hoff, J. D., Twersky, I., Verhey, K. J., Sutton, R. B., Eisenberg, E. & Rosenthal, J. J. C. Temperature-dependent RNA editing in octopus extensively recodes the neural proteome. *Cell* **186,** 2544–2555.e13 (2023).

143. Lamichhane, T. N., Abeydeera, N. D., Duc, A.-C. E., Cunningham, P. R. & Chow, C. S. Selection of peptides targeting helix 31 of bacterial 16S ribosomal RNA by screening M13 phage-display libraries. *Molecules* **16,** 1211–1239 (2011).

CHAPTER 2: THE EXTENSIVE AND DYNAMIC M^5C EPITRANSCRIPTOME OF *THERMOCOCCUS KODAKARENSIS* IS GENERATED BY A SUITE OF RNA METHYLTRANSFERASES THAT SUPPORT LIFE IN THE EXTREMES[1]

Summary

RNAs are synthesized using the four canonical nucleotides yet are often modified to invoke new structures, activities, and interactions. While many modifications are limited in frequency, restricted to non-coding RNAs, or present only in select organisms, 5-methylcytidine (m^5C) is abundant across diverse RNAs and fitness-relevant across Domains of life, but the synthesis and impacts of m^5C have yet to be fully investigated. Here, we mapped, then assessed the impacts of m^5C in the model hyperthermophile, *Thermococcus kodakarensis*. We demonstrate that m^5C is ~25x more abundant in *T. kodakarensis* than human cells and the m^5C epitranscriptome includes ~10% of unique transcripts. *T. kodakarensis* rRNAs harbor tenfold more m^5C compared to most Eukarya or Bacteria, and m^5C levels dynamically respond to environmental cues. We identified at least five RNA m^5C methyltransferases (R5CMTs) that modify RNAs *in vivo* and *in vitro*. Strains deleted for individual R5CMTs lack site-specific m^5C modifications that limit hyperthermophilic growth. We demonstrate that both RNA sequence and structure play a role in site-specific methylation and m^5C is likely generated through partial redundancy in target sites among R5CMTs. The complexity of, and requirements for the extensive m^5C epitranscriptome in *T. kodakarensis* argues that m^5C supports biological function in the extremes.

[1]This chapter has been submitted for publication.

This work was completed by the following contributors: Kristin A. Fluke, Ryan T. Fuchs, Yueh-Lin Tsai, Victoria Talbott, Liam Elkins, Hallie P. Febvre, Nan Dai, Eric J. Wolf, Jackson Schiltz, G. Brett Robb, Ivan R. Corrêa Jr., Thomas J. Santangelo. (2023)

Introduction

Studies of the bacterial and eukaryotic epitranscriptomes have shown that RNA chemical modifications are often abundant and dynamically modulated in response to physiological or environmental stimuli[1–5], parasitic-[6], pathogen- [7,8] or viral-infection[9,10] or are introduced within specific cells and lineages[11–13]. Modifications to even individual residues in long RNAs are known to elicit dramatic impacts on structure[14–16], stability[17–20], protein-interactions[21], cellular localization[22–24], translation efficiency[18,25–27], half-life[28], nuclear transport[29], mitochondrial[30] and plastid transport[31]. Specialized nucleotides within rRNAs and tRNAs often provide essential stability and importantly, these modified RNAs must be passed to daughter cells, providing a heritable role for epitranscriptomic modifications. The conserved modification profiles defined for rRNA and tRNA demonstrate that the distribution of modifications are not random, but rather that modifications are strategically positioned to impact function at specific nucleotides.[32] Evidence suggests this regulatory paradigm extends to mRNAs and across Domains, with RNA modifications being essential for all life.[33,34]

mRNAs have been increasingly identified as modification targets. Site-specific, and often sub-stoichiometric, modification of mRNA is linked to core biological functions and responses to environmental changes.[33,35] Critical questions remain regarding how the epitranscriptome is generated, how modifications impact mRNA function, and how mRNAs are selected for modification. Despite increasingly sophisticated methods to site-specifically identify and quantify a variety of RNA modifications, the low abundance of mRNA transcripts and their low modification frequency hinders investigations into the regulation imposed by the epitranscriptome.[36] Specific RNA modifications, including archaeosine[16] and 4-acetylcytidine[2] (ac^4C) are critical for life at high temperatures, suggesting that hyperthermophilic species might harbor greater densities of critical RNA modifications. Many archaea thrive in conditions inhospitable to most extant life wherein increased RNA stability is essential.[2,5,16] Maintaining

RNA structure at high temperature, or at extremes of salinity, pressure, or pH is facilitated , in part, by chemical modifications, and loss of epitranscriptomic modifications is often lethal.[2,3,16,37,38] Multiple RNA modifications are known to be more prevalent in Archaea than in Eukarya; many RNA modifications have been identified and quantified in the model hyperthermophilic species *Thermococcus kodakarensis*, representing an order of magnitude higher absolute abundance and several orders of magnitude higher frequency than found in model eukaryotes.[2,38,39] The abundance of RNA modifications in *T. kodakarensis*, its tractable genetic system, and the range of growth temperatures permitting robust growth provide an ideal platform to resolve fundamental open questions regarding the phenotypic consequences of epitranscriptomic changes, including RNA targeting and RNA modifying enzymes.

m^5C is one of the most abundant and conserved modifications that decorates the transcriptome across Domains.[40,41] m^5C in most species is predicted to be generated through the coordinated, post-transcriptional activities of several RNA m^5C methyltransferases (R5CMTs)[34]. Here we established the genesis, phenotypic impact, redundancy, and targeting mechanisms of the m^5C epitranscriptome within *T. kodakarensis*. *T. kodakarensis* retains a ~25 fold higher abundance of m^5C over human cell lines. With ultra-deep bisulfite sequencing, we comprehensively mapped m^5C sites across the *T. kodakarensis* transcriptome with single-nucleotide resolution under different biological conditions. We identified at least 232 unique m^5C sites in a diverse set of coding and non-coding RNAs, established that ~10% of all unique transcripts contain m^5C, and that mRNA represents the largest fraction of m^5C-modified RNAs. Moreover, we identified that at least five R5CMTs are responsible for installing individual m^5C sites. Through bisulfite-sequencing of RNA recovered from 17 unique strains, each deleted for a non-essential, putative RNA methyltransferase (RMTases), we paired specific sites of m^5C with unique R5CMTs, providing clues for targeting preferences of each R5CMT. R5CMT activity was further validated orthogonally through *in vitro* methylation assays and mass spectrometry.

Dynamic changes in the m^5C epitranscriptome were observed across distinct biological conditions. Strains deleted for individual (or pairs of) R5CMTs exhibited impaired growth at increasing temperatures, demonstrating that m^5C modifications support life at high temperatures. Our results suggest overlap in R5CMT function, with many m^5C sites being targeted for modification by at least two R5CMTs *in vivo*, indicating a partially redundant network of enzymes that maintain the m^5C epitranscriptome. The combined single-nucleotide, quantitative mapping of m^5C, phenotypic analyses of strains with abrogated m^5C-profiles, *in vitro* biochemistry with multiple recombinant R5CMTs, and homology of *T. kodakarensis* R5CMTs with RNA methyltransferases in each Domain provides a wealth of predictive power for understanding the impacts of the m^5C epitranscriptome in extant life.

Results

m^5C is abundant in T. kodakarensis coding and non-coding RNA

The nucleoside pools derived from total and rRNA-depleted RNAs isolated from *T. kodakarensis* strain TS559 were quantified and compared to that of 10 human cell lines (referred to here as the universal human reference (UHR)) by LC-MS/MS (supplementary Figure A2.1, see Appendix 2 for supplementary items). Total RNA preparations were dominated (~94%) by ncRNA sequenced reads, while fractions enriched for mRNA through selective degradation of rRNA and size selection to remove RNAs < 200 nt (see Materials and Methods) generated nearly ncRNA-free (~99.9%) mRNA sequencing results (supplementary Figure A2.1A). Within total RNA from the UHR, m^5C constitutes ~0.1% (~1:1,000) of cytidines (supplementary Figure A2.1B). The bulk of total m^5C in the UHR is likely present in stable, abundant rRNAs and tRNAs, and the order of magnitude reduction in m^5C levels upon mRNA selection confirms such. In stark contrast, total RNA preparations from *T. kodakarensis* revealed an unprecedented ~2.5% of total cytidines (~1:40) were replaced by m^5C (supplementary Figure

A2.1B). While degradation of *T. kodakarensis* rRNAs and size exclusion of tRNAs similarly reduced the levels of m^5C nearly-ten fold, the resulting levels of m^5C in mRNAs were still > 6.5-times greater than m^5C levels in the UHR mRNA pools. Analysis of small RNA fractions (< 200 nt) indicated that small RNAs are also rich in m^5C (supplementary Figure A2.1C).

The extremely high levels of modified cytidines in total, size fractionated, and mRNA preparations from *T. kodakarensis* were confirmed to be nearly exclusively m^5C, and not 5-hydroxymethylcytidine, 3-methyl cytidine, or 4-methyl cytidine, by comparing retentions times and mass transitions of corresponding modified nucleosides (data not shown). While m^5C levels are much higher in *T. kodakarensis* compared to the UHR, RNA modifications are not universally more prevalent in *T. kodakarensis*. Analysis of total and rRNA-repleted samples revealed that m^6A is more abundant in the UHR than m^5C and that m^6A is exceptionally low in *T. kodakarensis* total RNA and undetectable in mRNA (supplementary Figure A2.1B). The selective employment of m^5C over m^6A in RNA is likely important for the biological functions of the *T. kodakarensis* epitranscriptome.

Bisulfite sequencing reveals abundant, high-confidence, and reproducible m^5C modifications in the T. kodakarensis transcriptome

Sodium bisulfite treatment of RNA results in cytidine deamination, converting non-modified cytidines to uridines. Cytidines modified at the C-5 position, however, are largely resistant to bisulfite-driven deamination and retain their cytidine identity. Bisulfite treatment followed by cDNA synthesis, sequencing, and analysis (supplementary Figure A2.2A) thus permits quantitative assessment of m^5C modification frequencies at single-nucleotide resolution. The cytidine conversion rates of six bisulfite-sequencing libraries, three from exponential (e1-3) and three from stationary (s1-3) cultures of strain TS559, ranged from ~ 99.8% - 99.9% (supplementary Figure A2.2B), confirming near complete cytidine deamination which allows

accurate identification of m^5C sites.

RNA sequencing libraries were prepared for total RNA and mRNA-enriched fractions from *T. kodakarensis* TS559 cultures grown to early exponential or stationary growth phase. Sequenced reads were subjected to rigorous quality control pre- and post-alignment to the reference genome, and objective parameters were applied to call individual m^5C sites (supplementary Figure A2.2). Initial examination of candidate methylation sites revealed obvious false-positives likely due to incomplete bisulfite conversion of some reads (i.e., sequencing reads wherein > 3% of cytidines were not converted), and these reads were removed from further analysis. To assign a site as modified, we defined parameters for minimum overall sequencing coverage, minimum m^5C coverage (supplementary Figure A2.2C), minimum m^5C modification frequency (supplementary Figure A2.2D), and reproducibility in two of three biological replicates (supplementary Figure A2.3A). m^5C modification frequencies below 10% approached background levels in all libraries (supplementary Figure A2.2D), and therefore a 10% minimum modification frequency was applied as a requirement. Owing to the depth of high-quality sequencing, such as that of sites with ≥ 1000x coverage, the acceptable minimum m^5C frequency was lowered to 5%. Total sequencing depth at high-confidence m^5C sites typically ranged from ~100-1000x (supplementary Figure A2.2E), indicating that total coverage requirements were met for most m^5C sites. Low abundance RNAs did not provide significant statistical representation to accurately define sites with m^5C modification; the m^5C profile reported here likely underrepresents the totality of the m^5C epitranscriptome in *T. kodakarensis*. However, the extraordinarily high sequencing depth achieved as a result of the small genome (~2 Mbp) and high transcription levels typical for *T. kodakarensis* improves our ability to detect low stoichiometric modifications with greater accuracy and comparative power.

The position and frequency of m^5C at each site proved to be highly reproducible between replicates, suggesting minimal false positives within the final data. Linear regression of

m^5C sites between biological replicates revealed Pearson's correlation coefficients average at ~0.90, indicating high reproducibility of m^5C frequencies across the epitranscriptome (supplementary Figure A2.3A). Although we detected a minor proportion of m^5C sites that appear as outliers among replicates (when the m^5C site was present in only two of three replicates), the vast majority of our reads met our high-confidence thresholds within each sample; these outliers form the small number of data points seen next to the *x*- or *y*-axis. Sites that were high-confidence in all 3 replicates produced Pearson's correlation coefficients >0.90 (supplementary Figure A2.3B). To account for the discrepancy in reproducibility of some sites, we present a parallel analysis of datasets that include i) m^5C sites of high-confidence in at least two replicates (Figure 2.1) and ii) m^5C sites of high-confidence in all three replicates (supplementary Figure A2.4).

m^5C is highly abundant in a diverse set of RNAs with positional bias

Quantitative analysis of the m^5C epitranscriptome revealed a dense modification profile where shifts in modification sites and frequencies were observed between two biological conditions: exponentially growing and stationary phase cells (Figure 2.1, supplementary Figure A2.4). *T. kodakarensis* encodes just 2,306 open reading frames and an impressive ~1900 unique transcripts (~82%) were expressed at or above our detection threshold. Of those genes with sufficient expression to meet our statistical thresholds, ~10% (n=188 unique transcripts) contain at least one m^5C site (Figure 2.1A); we identified a total of 232 unique positions that satisfied our minimum criteria for high-confidence and reproducibility, mapping to 188 unique transcripts (Figure 2.1B). An identical analysis for sites reproducible in three of three replicates (supplementary Figure A2.4A), revealed 77 of ~1250 RNAs (~6%) retains at least one site of m^5C modification. The total number of m^5C sites, m^5C frequencies, and modified transcripts changed between exponentially growing and stationary phase cells (Figure 2.1B and supplementary Figure A2.4B), suggesting a dynamic m^5C-profile that shifts in response to for

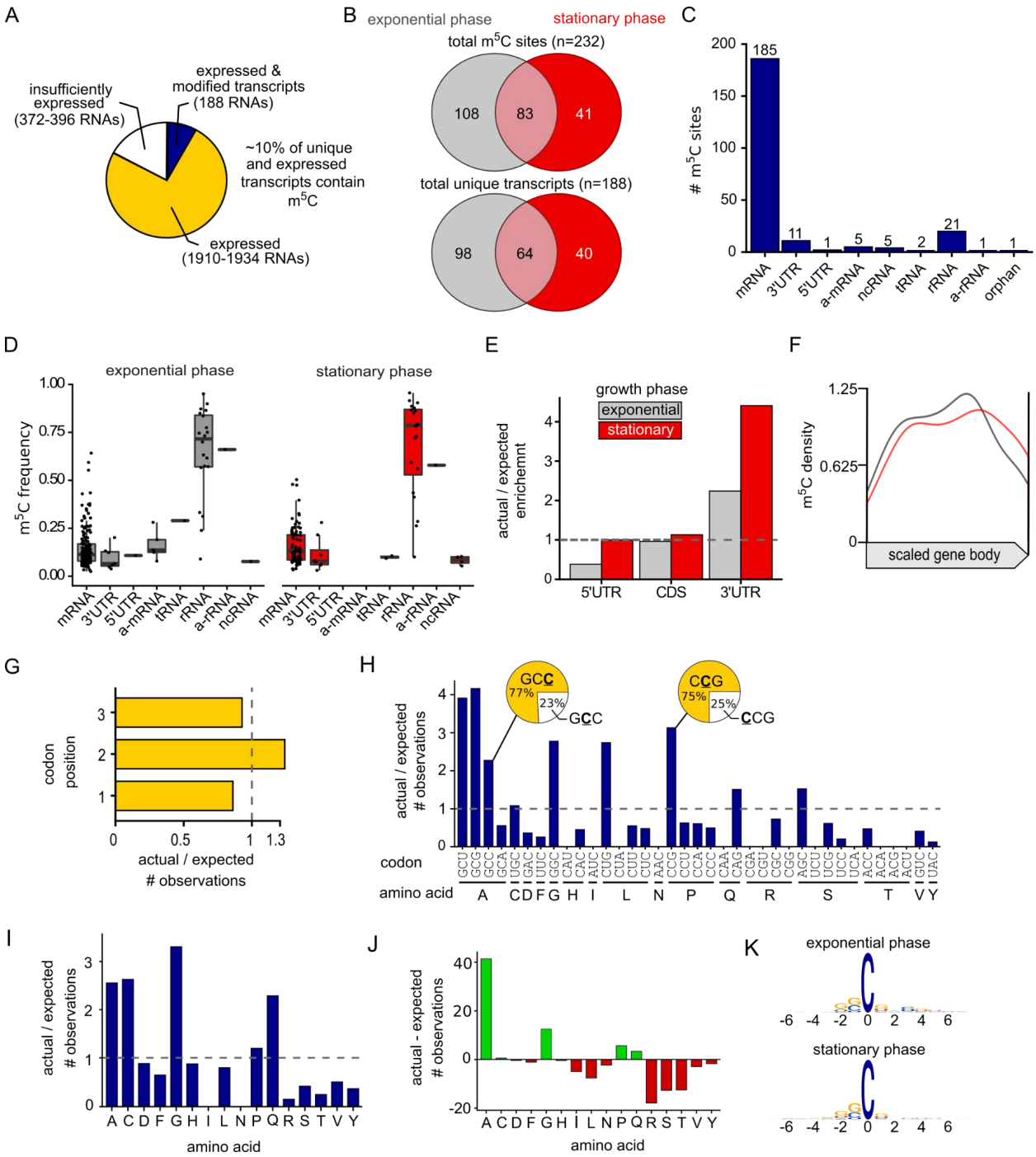


Figure 2.1 *T. kodakarensis* has an extensive and dynamic m^{5}C epitranscriptome in a diverse set of RNAs. (A) Empirical analysis of the *T. kodakarensis* m^{5}C epitranscriptome. Out of 1910-1934 expressed RNAs between two biological conditions, 188 contained at least one m^{5}C site. **(B)** The number of distinct m^{5}C sites and modified transcripts varied between exponential and stationary growth phase cells. **(C)** Number of m^{5}C sites mapping to diverse RNAs, including some antisense (a-) RNAs. **(D)** Modification frequencies across RNA types. **(E)** Comparison of the observed and expected number of m^{5}C sites in different regions of an mRNA in exponential and stationary growth phase cells. The ratio (actual/expected) of m^{5}C sites

equals to 1 if the m^5C sites are distributed randomly within the region. **(F)** Positional bias of m^5C sites when mapped within coding sequences. **(G)** Positional bias of m^5C sites at each codon position. **(H)** The number of m^5C sites mapped to each codon compared to the number expected by random chance. Obvious deviations from expectation (actual/expected = 1) were observed. There were also apparent biases in codons that include more than one cytidine (see pie chart insets). **(I)** The actual/expected and **(J)** difference between actual and expected (actual - expected = 0 where no bias is detected) number of m^5C sites mapping to particular amino acid codons. **(K)** Logo sequence analysis of nucleotides adjacent to m^5C sites.

minimally two biologically reproducible sites (Figure 2.1C); 77 of 99, ~78% for three biologically environmental cues. Most m^5C sites were detected within coding sequences (185 of 232, ~80% reproducible sites (supplementary Figure A2.4C)). The 16S and 23S ribosomal RNA fragments are densely modified, with 21 or 19 m^5C sites detected in mature rRNA, in two or three biological replicates, respectively; no m^5C sites were detected in either copy of the 5S rRNA. We observed only sparse modification of untranslated regions (UTRs), transcripts that mapped antisense (a-) to mRNAs, rRNAs, and other non-coding RNAs. Although tRNAs were size-excluded during RNA library preparation, one or two m^5C sites were detected in tRNAs with 3/3 or 2/3 replicates, respectively.

Modification frequencies in mRNAs ranged from ~5-65% while averaging at ~15-20%, whereas modification frequencies in rRNA averaged at ~75% (Figure 2.1D and supplementary Figure A2.4D). mRNA modifications in some model species are typically enriched in UTRs or near UTR-ORF boundaries.[48,49] We observed more m^5C sites in 3'UTRs than expected if distributed randomly throughout an mRNA (Figure 2.1E), but as no 3'UTR sites were reproducibly identified in all three replicates (supplementary Figure A2.4E), it is invalid to definitively suggest a 3'UTR bias. The distribution of modifications that mapped to coding sequences did not have an obvious positional bias over the open reading frame (Figure 2.1F and supplementary Figure A2.4F).

As m^5C sites accumulate within the coding sequence, it is possible that mRNA modifications may impact translation dynamics by statistically over impacting specific codons or

codon positions. We did not observe a significant bias at any codon position relative to what would be expected if m^5C sites were distributed at random (actual / expected = 1, Figure 2.1G and supplementary Figure A2.4G). The absence of m^5C enrichment at the third codon position likely rules out sole impacts on wobble base-pairing and therefore non-cognate amino acid incorporation. However, m^5C is enriched in select codon sequences (Figure 2.1H and supplementary Figure A2.4H). We mapped m^5C within GCG codons (alanine, n=18) to an extent that is fourfold higher than expected if by random chance. Modifications to codons GCU (alanine, n=13), GGC (glycine, n=18), CCG (proline, n=24), and CUG (leucine, n=16) were similarly enriched. When codons include multiple cytidines, there is a strong bias for which cytidine is selected for modification. Out of 35 modified GCC codons, 77% of the modifications were to the third cytidine. (Figure 2.1H, pie chart inset). When considering sites reproducible in all three replicates, 100% of modifications to this codon occur at the third cytidine (supplementary Figure A2.4H, pie chart inset). Modifications mapping to leucine and proline codons indicate a strong bias for one of the four possible codons (CUG and CCG, respectively) whereas modifications are highly underrepresented at the other 3 codon possibilities. These data suggest select codons are targeted for modification within mRNAs and may impact the translatability of those codons.

A similar representation is observed at the amino acid level. m^5C occurrences in particular amino acid sequences indicate 41 more m^5C sites in alanine encoded codons than expected (actual - expected) (Figure 2.1I and J). Codons encoding cysteine and glutamine are overrepresented with high-confidence m^5C sites by two to sixfold (Figure 2.1I). We expected to see 0.38 m^5C sites in cysteine codons, but we detected 1. Even though this indicates a significant fold change (actual / expected) in the number of observations, less than 1 additional m^5C site was detected compared to expected (actual – expected) (Figure 2.1I and J). Therefore, overrepresentation of modified cysteine codons is not likely biologically relevant. Near identical

conclusions can be drawn in the analysis of m^5C sites reproducible in all 3 replicates (supplementary Figure A2.4G-J). Taken together, these data indicate a strong positional bias in m^5C sites in particular codon and amino acid contexts, likely indicating m^5C impacts the translatability of these codons. However, there does not appear to be a congruent nucleotide sequence context that describes all m^5C sites (Figure 2.1K). This observation is consistent with previous studies that failed to identify a single sequence motif associated with m^5C modifications.[25,50,51] It is plausible that the *T. kodakarensis* m^5C epitranscriptome is installed by a collection of RNA modifying enzymes, each targeting a unique sequence or RNA structure. In this case, a single nucleotide context would not be anticipated, rather, individual enzymes would be expected to target unique sequence motifs and RNA structures for the installation of m^5C.

T. kodakarensis encodes a suite of RNA modifying enzymes

Publically available annotations of the *T. kodakarensis* genome revealed 16 ORFs likely to encode RNA methyltransferases (Table 2.1). Six of the sixteen putative RNA methyltransferase (RMTase) enzymes were predicted to install m^5C based on domain comparisons with known RMTases in the Modomics Database but none have previously been determined to be *bona fide* RNA methyltransferases. Each of the six enzymes predicted to install m^5C were also identified by pBLAST to have domain homology to the human NSUN family of proteins that are known to install the m^5C epitranscriptome in mammals. Upon the deletion of any individual RMTases, any differences in the m^5C epitranscriptome compared to the parental strain (TS559) could be rationalized to require the activities of the deleted enzyme. Using established techniques[44], we thus attempted to individually, markerlessly delete sequences encoding the entirety of the ORF for each putative RMTase from the TS559 genome.

We were successful in generating 13 novel strains, each lacking a single putative RMTase. The sequences surrounding the targeted locus were first confirmed to lack the

Table 2.1 Putative RNA methyltransferases in T. kodakarensis

Gene ID[a]	Description[b]	Predicted modification[c]	essential
TK0224	class I SAM-dependent Mtase, UbiE/COQ5 family	mcm5s2U, mcm5U	no
TK0234	Mtase domain-containing protein	mmpN	no
TK0360	class I SAM-dependent RMtase, RsmB/NOL1/NOP2/Sun family	**m5C**	no
TK0704	class I SAM-dependent Mtase, UbiE/COQ5 family	mcm5s2U, mcm5U	no
TK0729	class I SAM-dependent Mtase, UbiE/COQ5 family	mcm5s2U, mcm5U	no
TK0872	class I SAM-dependent RMtase, RsmB/NOL1/NOP2/Sun family	**m5C**	no
TK1273	class I SAM-dependent Mtase, UbiE/COQ5 family	cmo5U, mcmo5U, m3C, m1G	no
TK1784	MTase domain-containing protein, METTL16/RlmF/DUF890 family	m6A	no
TK1917	class I SAM-dependent Mtase	cmo5U, cmo5U, m3C, m1G	no
TK1935	class I SAM-dependent RNA Mtase, RsmB/NOL1/NOP2/Sun family	**m5C**	no
TK2122	class I SAM-dependent RNA Mtase, RsmB/NOL1/NOP2/Sun family	**m5C**	no
TK2241	class I SAM-dependent MTase, UbiE/COQ5 family	cmo5U, mcmo5U, m3C, m1G	no
TK2304	class I SAM-dependent RMtase fused to NusB regulator domain, RsmB/NOL1/NOP2/Sun family	**m5C**	no
TK0008	Predicted MTase, UPF0020 family	m2G	yes
TK1785	class I SAM-dependent RMtase,RsmB/NOL1/NOP2/Sun family	**m5C**	yes
TK1933	Predicted MTase, METTL5 family protein family	m6A, m2G	yes

[a] Two predicted methyltransferases (rows highlighted in orange and yellow) are essential and cannot be deleted from the cell, while 14 are non-essential. It is unclear if TK0008 is essential.

[b] Gene descriptions are summarizes of information gathered from NCBI, Uniprot, KEGG and InterPro.

[c] Modification predictions are based on sequence and structural similarities to known RMTases from Modomics

RMTase encoding sequences by several diagnostic PCRs (Figure 2.2A). Then, we sequenced the entire genome of each deletion strain (typically at > 15x coverage; Figure 2.2B and supplementary Figure A2.5) to ensure the lack of second-site modifications. Sequencing of bisulfite-treated RNA (Figure 2.2C and supplementary Figure A2.5) at ~100x to ~1600x coverage revealed, in all strains, a lack of transcripts from the loci targeted for deletion. By repeating our markerless genomic modification procedures, we were further able to construct double deletion strains wherein two putative RMTases were simultaneously deleted. Despite exhaustive efforts, two of the targeted putative RMTases (TK1785 and TK1933) could not be deleted. Attempts to delete gene TK0008 yielded mixed results; *T. kodakarensis* regularly maintains many copies of its genome[52], and while we were able to generate strains that appeared to lack TK0008 *via* diagnostic PCR and whole genome sequencing, some genomes containing TK0008 must have been retained within the population that led to repetitive failures to yield strains that completely lacked TK0008 sequences after two culture passages. Thus, TK0008 was excluded from further analysis.

We collected and bisulfite sequenced RNA from each single or double deletion strain grown to early exponential or late stationary phase in duplicate. High-confidence and reproducible m^5C sites were identified in each deletion strain (see linear regression analysis in supplementary Figure A2.6) following the pipeline described in supplementary Figure A2.2 and compared with m^5C sites established in the parental strain, TS559. Differential m^5C frequencies demanded minimally a greater than twofold change and a p-value cutoff of 0.01 to be deemed statistically relevant. The number of sites identified in parental and deletion strains where m^5C frequencies were completely lost (Table 2.2) or differed sufficiently to meet our stringent criteria (supplementary Table A2.2) are reported. We identified multiple sites where an absolute loss in m^5C modification frequency (≤ 2%) was observed in one of five deletion strains; ΔTK0360, ΔTK0872, ΔTK1935, ΔTK2122, and ΔTK2304 (Table 2.1), indicating these genes encode

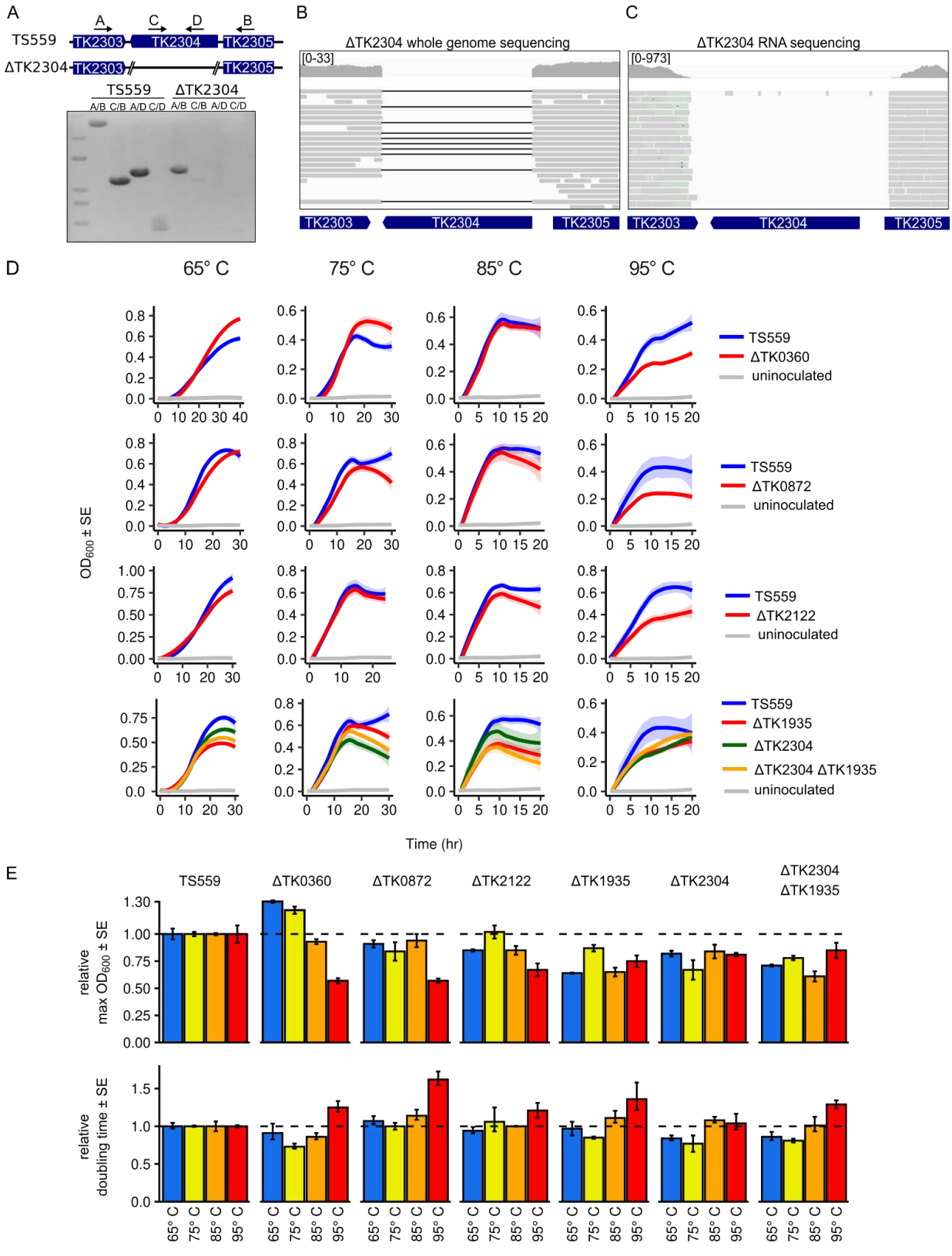

A
TS559
ΔTK2304
TK2303
TK2304
TK2305
A/B C/B A/D C/D
B
ΔTK2304 whole genome sequencing
[0-33]
C
ΔTK2304 RNA sequencing
[0-973]
D
65° C
75° C
85° C
95° C
TS559
ΔTK0360
ΔTK0872
ΔTK2122
ΔTK1935
ΔTK2304
ΔTK2304 ΔTK1935
uninoculated
OD600 ± SE
Time (hr)
E
relative max OD600 ± SE
relative doubling time ± SE

Figure 2.2 The m^5C epitranscriptome supports hyperthermophilic growth. (A) The deletion of gene TK2304 was initially confirmed by PCR using 4 sets of primers. External primers (A/B) amplify across the deleted locus and result in a full length amplicon in parent strain TS559 and an amplicon reduced by the size of gene TK2304 in the ΔTK2304 strain. Internal primers (C/D) or a combination of internal and external primers (A/D, C/B) result in amplification in strain TS559 but not ΔTK2304. Final confirmation was performed using **(B)** Minion whole genome sequencing (DNA-seq) and **(C)** Illumina RNA bisulfite sequencing (RNA-seq). Visual inspection of DNA and RNA sequences aligned to the *T. kodakarensis* reference genome using Integrative Genomics Viewer show the absence of reads from the deleted loci. In the DNA-seq windows, black lines connect contiguous reads and indicate a gapped alignment. The coverage range for each window is listed in brackets. **(D)** Head-to-head growth competitions were performed at 65°, 75°, 85°, and 95° C for each strain deleted for an R5CMT (ΔTK0360, ΔTK0872, ΔTK2122, ΔTK1935, ΔTK2304), the double deletion (ΔTK1935 ΔTK2304), and parent strain TS559. **(E)** The maximum optical density (Max OD_{600}) and the rate of growth (doubling time) for each culture is illustrated relative to parent strain TS559.

Table 2.2 Absolute losses and gains in m^5C sites in strains deleted for individual R5CMTs.

		exponential phase[b]		stationary phase[c]		
Gene ID	Description[a]	abs. gains	abs. losses	abs. gains	abs. losses	# losses/ gains
TK0360	class I SAM-dependent RMTase, RsmB/NOL1/NOP2/Sun family	3	21	12	1	0
TK0872	class I SAM-dependent RMTase, RsmB/NOL1/NOP2/Sun family	1	7	0	4	5
TK1935	class I SAM-dependent RMTase, RsmB/NOL1/NOP2/Sun family	0	15	1	11	10
TK2122	class I SAM-dependent RMTase, RsmB/NOL1/NOP2/Sun family	0	11	0	4	20
TK2304	class I SAM-dependent RMTase fused to NusB, RsmB/NOL1/NOP2/Sun family	2	6	0	6	30
TK1935; TK2304	double deletion	4	39	5	17	40

[a] The NCBI description and protein family (Uniprot and InterPro predictions) are provided.

[b,c] BS-seq of individual or double deletion strains identified absolute losses and gains in m5C sites between exponential and stationary growth phase cells.

enzymes that likely directly target RNA for m^5C installation. Bisulfite-sequencing of RNAs from the remaining ten single deletion strains did not reveal any absolute losses within the m^5C epitranscriptome (supplementary Table A2.2), suggesting that these enzymes either do not install m^5C or were not active under our experimental conditions; note that RNA transcripts for each putative RMTase gene targeted for deletion were abundantly present in the TS559 parent stain, indicating these genes are expressed under our experimental conditions. Our results are thus consistent with the predicted modification potentials listed in Table 2.1 and pBLAST homology searches for R5CMTs. To our surprise, certain strains deleted for putative m^5C-specific and alternative RMTase enzymes showed gains in m^5C sites or increased modification frequencies. Sites that show new modification status in deletion strains compared to the parental strain met our stringent statistical rigors for reproducibility. The m^5C gains imply that the loss of selected m^5C sites due to deletion of distinct RMTases from *T. kodakarensis* results in compensatory modifications throughout the epitranscriptome. This level of complexity and intertanglement of enzyme activities reinforces that the epitranscriptome is dynamic and responsive to changes in gene expression and environmental cues.

A redundant network of R5CMTs likely generates the m^5C epitranscriptome.

We detected 232 high-confidence and reproducible m^5C sites (Figure 2.1A) and correlated the absolute loss of ~46 of these sites with the loss of one of five R5CMTs (Table 2.2). Out of 21 sites mapping to mature rRNA, we only identified enzymes likely to install 9 m^5C sites (Figure 2.3). It is possible that essential enzymes from Table 2.1, or enzymes missed by our annotations of the entire genome, are responsible for the installation of the remaining m^5C sites detected here. Alternatively, multiple R5CMTs may target identical sites for installation of m^5C. To explore the latter possibility, we deleted both TK1935 and TK2304 within the same cell. Bisulfite sequencing of the double deletion strain (ΔTK1935 ΔTK2304) during exponential or stationary growth phases revealed a synergistic absolute 39 losses in m^5C sites compared to

the 17 and 6 losses identified in the ΔTK1935 and ΔTK2304 single deletion strains, respectively (Table 2.2 and Figure 2.4I). This suggests that *T. kodakarensis*, and perhaps other species, may generate epitranscriptomes with overlapping and complex activities for distinct RNA modifications.

The m^5C epitranscriptome enhances cellular viability at increasing temperatures

T. kodakarensis grows over a wide temperature range (~50 to ~98˚C), with a growth optimum of 85˚C resulting in a ~40-minute doubling time under ideal conditions. When constructing the single and double deletion R5CMT strains, we noted an obvious growth impairment of some deletion strains. Growth of the parental and deletion strains were then monitored at 65˚, 75˚, 85˚, and 95˚C to present strains with different environments (Figure 2.2D). At 65˚ and 75˚C, the growth of strains deleted for one of the five non-essential R5CMTs is largely comparable to the parent strain TS559 (Figure 2.2D), but deletion strains fared more poorly at increasing culture temperatures. Notable at 85˚C and more obviously at 95˚C, each R5CMT deletion strain displayed significant growth defects compared to the parental strain, and in many cases exhibit slower growth rates and reduced end-point culture densities (Figure 2.2E). The strain deleted for both TK1935 and TK2304 (ΔTK2304 ΔTK1935) experienced a compounding, growth defect compared to either individual deletion strain, suggesting the m^5C epitranscriptome holistically supports cell viability and survivability at high temperatures.

Growth defects were also observed in strains deleted for the other 10 putative RMTases. The growth rates for most strains were generally within expectations at 65˚, 75˚, and 85˚C, but longer doubling times were observed at 95°C in many strains (supplementary Figure A2.7). Many deletion strains also displayed a lower end-point culture density compared to the parent strain, and this defect becomes increasingly more apparent at elevated temperatures (supplementary Figure A2.7). Unsurprisingly, strains deleted for multiple RMTases experienced

a compounding effect compared to their single deletion counterparts.

The T. kodakarensis ribosome is hypermodified at mRNA decoding regions

Ribosomes across bacterial and eukaryotic model species typically harbor 2-3 m^5C sites. In *E. coli*, each of the three m^5C sites is installed by a unique Rsm MTase[53], whereas human cells rely on the NSUN family of MTases to install both ribosomal m^5C sites[54]. Surprisingly, we detected an order of magnitude increase (n=23) in m^5C sites mapping to the single *T. kodakarensis* 16S-$tRNA^{Ala}$-23S rRNA operon (Figure 2.3A and supplementary Figure A2.7). An additional m^5C site was mapped antisense to the 16S-rRNA, two m^5C sites map to the region between the 16S and $tRNA^{Ala}$, and 21 m^5C sites are incorporated into the final 16S (n=12) or 23S (n=9) sequences. There are no noticeable differences in m^5C sites or frequencies between exponential or stationary growth phases, and no m^5C residues are detectable in either 5S fragment (supplementary Figure A2.7A and B). A previous study demonstrated that *T. kodakarensis* ribosomes are hypermodified with ac^4C, implying that additional modifications support rRNA function and perhaps ribosome performance at high temperatures.[2] Our data provides new evidence that *T. kodakarensis* ribosome is also hypermodified with m^5C, implying that RNA modifications may support ribosomal function and performance at high temperatures

We were able to correlate 9 m^5C sites in the 16S and 23S rRNA fragments with specific R5CMT activities, based on the absolute losses in m^5C sites due to deletion of specific enzymes. Four of the five identified R5CMTs appear to target rRNAs (Figure 2.3A), to the exclusion of TK0360. Although tRNA and other small RNAs were largely removed from the sequencing libraries, TKt41 ($tRNA^{Trp}$) persisted through sequencing, and TK0360 was identified to methylate this transcript. Otherwise, TK0360 mainly targets mRNA transcripts. The m^5C sites mapping to rRNA are generally distributed throughout the 16S and 23S rRNA fragments with a linear clustering of sites in the 23S fragment. However, when viewed in the atomic structure of

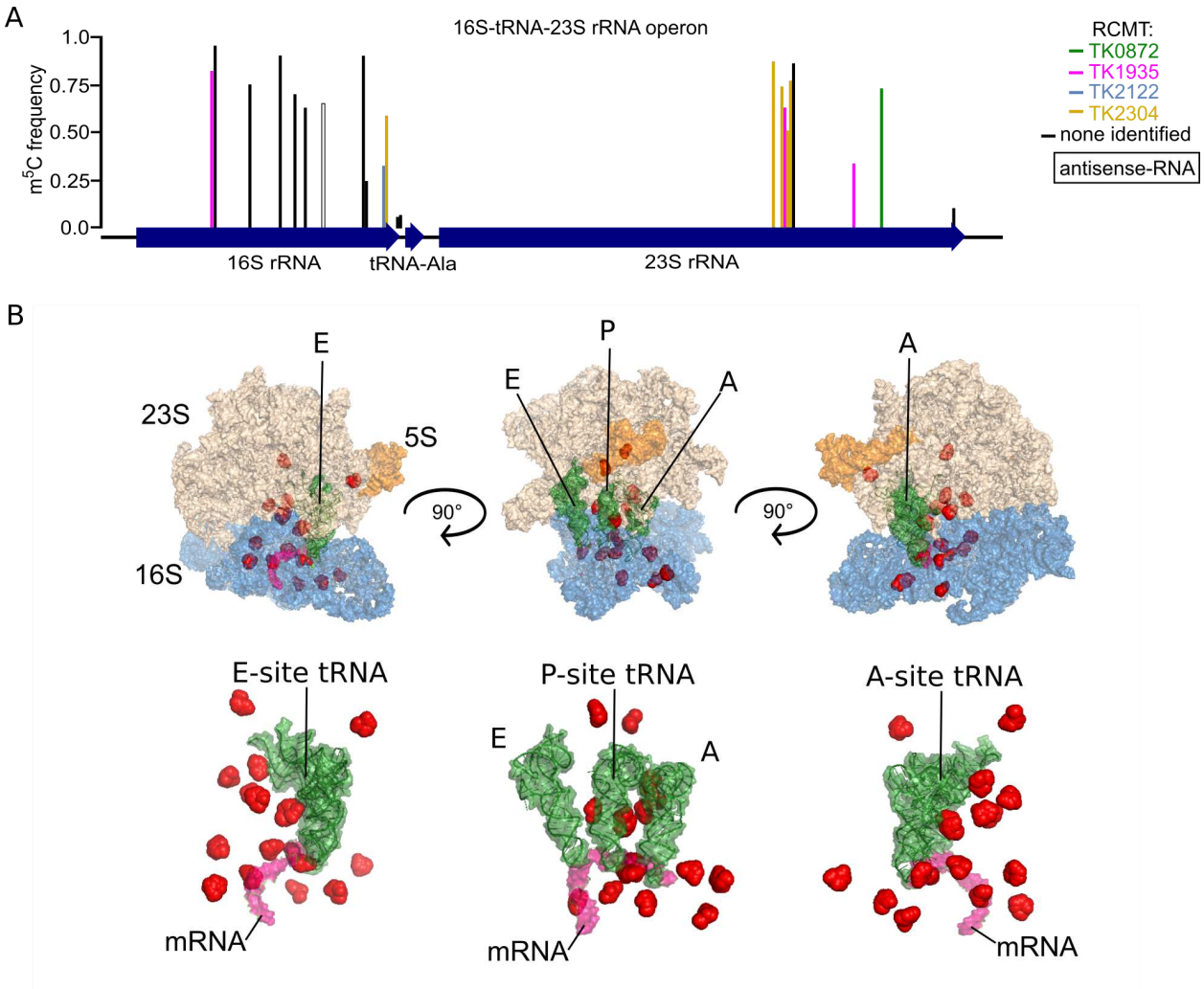


Figure 2.3 The *T. kodakarensis* ribosome is densely modified at key functional regions involved in translation. (A) Each bar represents an m^5C site mapped to the 16S-tRNAAla-23S operon; bar height represents modification frequency and its color corresponds to the enzyme responsible for its installation. One m^5C site was mapped antisense to the 16S rRNA (black outline). **(B, C)** m^5C sites (red spheres) are not randomly distributed. 16S subunit (blue), 23S subunit (nude), 5S subunit (orange), A, P, and E site tRNAs (green), and the mRNA (pink) are shown at 3 angles rotated at 90°.

the *T. kodakarensis* 70S ribosome (PDB: 6TH6), m^5C sites obviously crowd the A, P, and E sites and are positioned along the path where the mRNA is fed into the decoding center (Figure 2.3B). It is therefore clear that m^5C sites in fully assembled ribosomes are not randomly distributed but enriched in highly conserved and functional regions. Further studies are necessary to detail the impacts of m^5C modifications to these key regions on translation.

R5CMTs display sequence and structural specificity for substrate RNAs

The mechanisms that target MTases to distinct positions in the transcriptome are unclear. To elucidate the targeting strategies of R5CMTs identified here, we analyzed the sites where differential m^5C modification between deletion and parental strain were detected. A representative full analysis is shown for the deletion of TK2304 (Figure 2.4) while analysis for deletions of TK0360, TK0872, TK1935, and TK2122 are recorded in supplementary Figures A2.8 and A2.9. Deletion of TK2304 resulted in 6 absolute losses and 2 gains in m^5C sites. These sites were mapped mainly to rRNAs, but one m^5C site was detected in a single mRNA (Figure 2.4A and C). Logos sequence analysis indicates some conservation of sequence contexts for gains and losses, but the totality of sequence context alone is insufficiently complex to completely explain targeting of these sites. Three definitive nucleotides (**C<u>C</u>**nnn**U**) constitute a partial RNA sequence motif targeted by TK2304, while the two gains in m^5C sites follow an eight nucleotide motif (Figure 2.4B). Gains and losses found within distinct sequence contexts were identified for all other R5CMTs (supplementary Figure A2.9A-D.II), but in almost all cases, no sequence context alone would suffice as the sole targeting mechanism for a single R5CMT. The sole exception, interestingly, identifies more complex sequence motifs targeted by TK1935, but, surprisingly, the motif varied depending on whether the m^5C occurs in rRNA or mRNA (supplementary Figure A2.8B.II). Our analyses suggest that R5CMTs partially rely, to varying extents, on primary sequence for site-specific methylation. However, three to six nucleotides are not unique enough to encode the specificity required to target an enzyme without likely producing thousands of off-target methylation events throughout the transcriptome, even as small as that of *T. kodakarensis*.

Modification frequencies in mRNAs are typically low (~15-20%, Figure 2.1D) and questions remain regarding whether any biological relevance of substoichiometric mRNA

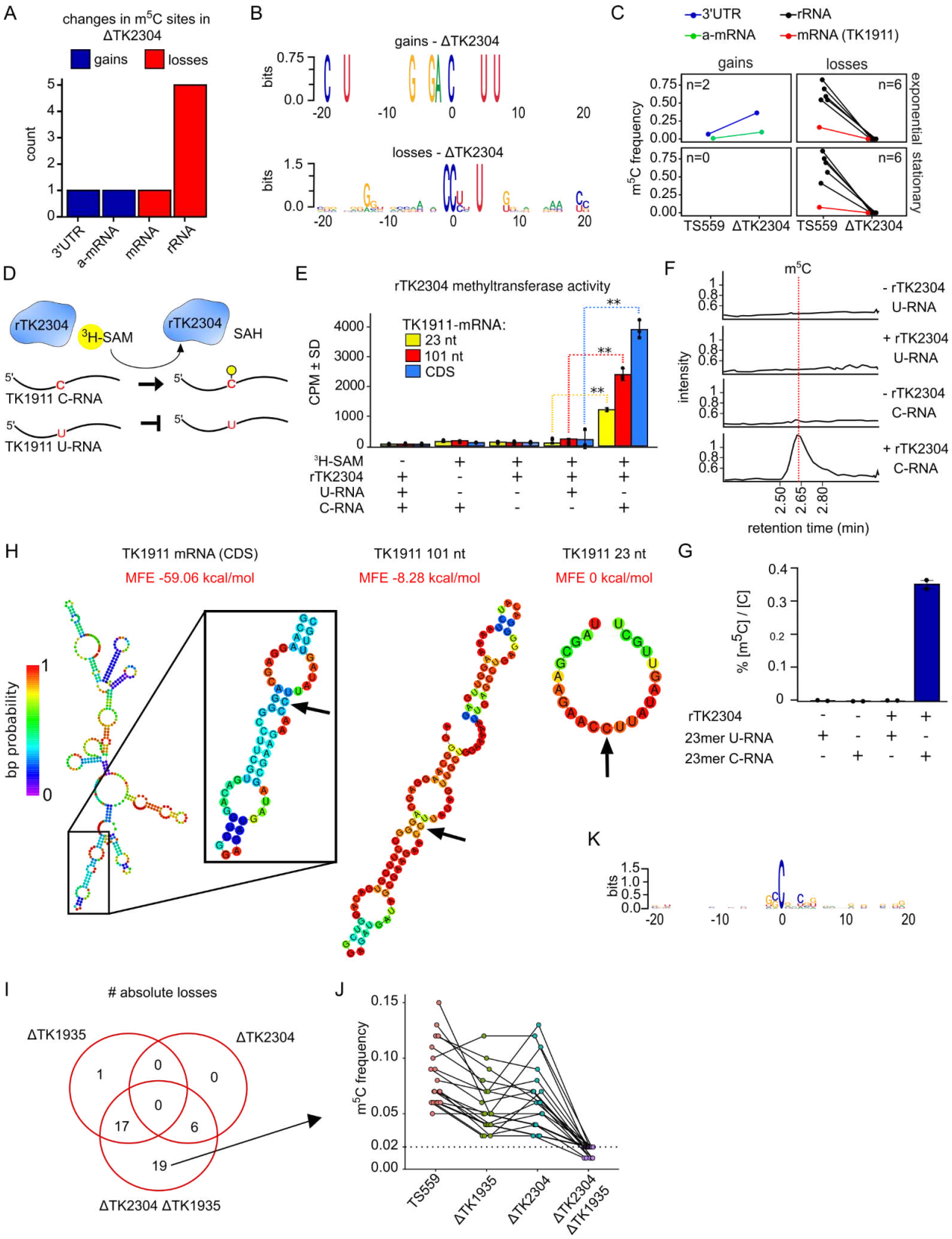

A
changes in m5C sites in ΔTK2304
gains
losses
count
3'UTR
a-mRNA
mRNA
rRNA
B
gains - ΔTK2304
losses - ΔTK2304
bits
C
3'UTR
a-mRNA
rRNA
mRNA (TK1911)
gains
losses
m5C frequency
n=2
n=6
n=0
n=6
exponential
stationary
TS559
ΔTK2304
D
rTK2304
3H-SAM
SAH
TK1911 C-RNA
TK1911 U-RNA
E
rTK2304 methyltransferase activity
TK1911-mRNA:
23 nt
101 nt
CDS
CPM ± SD
3H-SAM
rTK2304
U-RNA
C-RNA
F
m5C
- rTK2304
U-RNA
+ rTK2304
U-RNA
- rTK2304
C-RNA
+ rTK2304
C-RNA
intensity
retention time (min)
H
TK1911 mRNA (CDS)
MFE -59.06 kcal/mol
TK1911 101 nt
MFE -8.28 kcal/mol
TK1911 23 nt
MFE 0 kcal/mol
bp probability
G
% [m5C] / [C]
rTK2304
23mer U-RNA
23mer C-RNA
K
bits
I
absolute losses
ΔTK1935
ΔTK2304
ΔTK2304 ΔTK1935
J
m5C frequency
TS559
ΔTK1935
ΔTK2304
ΔTK2304 ΔTK1935

Figure 2.4 TK2304 encodes a *bona fide* R5CMT with apparent RNA sequence and structural specificity. (A) Bisulfite sequencing of strains deleted for the gene TK2304 reveals losses and gains in m^5C sites. **(B)** Logo sequence analysis of adjacent nucleotides surrounding sites where m^5C was lost or gained. **(C)** The change in modification frequency between parent strain TS559 and ΔTK2304 is faceted by growth phase and direction of regulation. The number of sites (n) is listed within each window and the color corresponds to RNA type. **(D)** The methyltransferase assay schematic is shown. Recombinant enzyme (rTK2304), radiolabeled methyl co-factor (^{3}H-SAM), and RNA substrate are incubated together. The enzyme will transfer the labeled methyl group to the cytidine target in the C-RNA but not the U-RNA where the cytidine is replaced with a uridine. **(E)** The results of the methyltransferase assays are displayed in counts per minute (CPM) which measures the β-decay of the radiolabeled methyl group covalently bound to the RNA substrate. **(F, G)** Mass spectrometry analysis of RNAs methylated by rTK2304. The 101 nt oligonucleotide was digested to single nucleosides and the m^5C levels were quantified relative to cytidine across controls. **(H)** The secondary structures of the 23 nt, 101 nt, and full coding sequence (CDS) RNAs encoding the gene TK1911 were predicted using the Vienna RNAFold webserver[71]. The arrow points to the cytidine targeted for methylation. The color of each base corresponds to the base pair (bp) probability, and the minimum free energy (MFE) for each structure is listed. **(I)** The number of unique and absolute losses in m^5C sites between single and double deletion strains ΔTK2304 and ΔTK1935. **(J)** The change in modification frequencies between strains of the 19 m^5C sites completely lost (m^5C frequency ≤ 2%) in ΔTK2304 ΔTK1935. **(K)** Logo sequence analysis of the 19 sites uniquely and completely lost in ΔTK2304 ΔTK1935. * $p < 0.05$, ** $p < 0.01$.

modifications are merited given the low modification frequencies. mRNAs may be intentional targets of MTases or be occasionally mistargeted by MTases with primary roles in rRNA or tRNA. To add credence to specific mRNA targeting, we identified a high-confidence, but low modification frequency m^5C within the TK1911 mRNA that was completely lost due to deletion of TK2304. During exponential growth, the TK1911 mRNA is modified at a frequency of ~15-18%, while in stationary phase, modification frequency of the same site drops to just ~8-13% (Figure 2.4C), thus representing a site that minimally but convincingly falls within our stringent criteria for m^5C modification calls. To directly correlate the *in vivo* loss of m^5C sites with enzymatic activity, we designed an *in vitro* assay where a recombinant TK2304 enzyme (rTK2304) was challenged to site-specifically methylate the mRNA encoding TK1911 (Figure 2.4D). When rTK2304 was provided with a radio-labeled, methyl-donor S-adenosyl-L-methionine (^{3}H-SAM) and an RNA substrate identical in sequence to TK1911 (C-RNA), an obvious transfer of the radiolabeled methyl group from SAM to the RNA was detected (Figure 2.4E). Reactions lacking

SAM, rTK2304, or an RNA substrate displayed background levels of radioactivity recovered on RNA substrates. As an idealized negative control, we also generated an mRNA with the sequence of TK1911 that contained a single nucleotide substitution, replacing the C targeted for m^5C modification *in vivo* with a U (U-RNA) (Figure 2.4D); note that all other cytidines within the U-RNA were retained. Replacing the native TK1911 RNA (C-RNA) with an otherwise identical RNA with a single U substitution (U-RNA) resulted in only background levels of methyl transfer, equivalent to that of no SAM, no enzyme, or no RNA controls (Figure 2.4E).

To elucidate the necessary and sufficient sequences targeting the TK1911 mRNA for modification by rTK2304, we generated RNA substrates of decreasing lengths (444 nt full length mRNA, 101 and 23 nt truncated RNAs; Figure 2.4E) with both a C or U at the site of modification identified *in vivo*. While rTK2304 is unable to methylate any of the U-containing RNAs above background levels, rTK2304 methylates all C-RNAs with increasing activity on longer RNAs. As the sequence immediately surrounding the target cytidine is identical between substrates, other sequence and structural elements within the RNA must elicit differential activity of rTK2304. Structural comparisons between the three RNA substrates (Figure 2.4H) adumbrate that the secondary and tertiary structures between these RNAs may play a role in the substrate recognition or catalytic activity of rTK2304 *in vitro*. Our results demonstrate that rTK2304 activity *in vitro* is improved by supplying more RNA sequences and structures to guide site-specific methylation of the TK1911 mRNA and are in contrast to expectations of spurious and non-biologically relevant recognition and methylation of the TK1911 mRNA by TK2304.

To confirm modification of the TK1911 mRNA at the site predicted based on the loss of *in vivo* modification due to deletion of TK2304, the 101 nt substrate containing a C or U at the central position was mixed with unlabeled SAM and rTK2304 *in vitro*, purified, then digested to single nucleosides for LC-MS/MS analysis. We observed 0.37% of m^5C/C ratio (or 5% oligo modification frequency) on the C-containing 101-nt RNA (Figure 2.4F and G). The U-containing

101 nt RNA did not reveal any mass or m^5C/C ratio changes following incubation with SAM and rTK2304. Intact oligonucleotide mass spectrometry analysis of RNase T1 digested 101 nt substrate also indicated the presence of a methyl group at the predicted cytidine target site (supplementary Figure A2.8A and B).

To better understand substrate recognition, we analyzed the activity levels of recombinant R5CMTs on several RNAs. Like rTK2304, the other R5CMTs were challenged to methylate different 23 nt RNAs corresponding to sequences identified as likely *in vivo* targets based on differential m^5C frequencies in parental and deletion strains. Statistically significant, and often highly disparate activity levels were observed for rTK2122 and rTK1935 on C- versus U-containing substrates despite all substrates likely lacking base-paired structures (supplementary Figure A2.9A-D.IV). rTK1935 displayed robust activity on RNAs encoding both identified sequence motifs (supplementary Figure A2.9B.II and IV). When each 23 nt was aligned (supplementary Figure A2.9A-D.IV insets), greater sequence consensus was observed with gapped alignments, which was not captured in the logos sequence analysis. These data indicate nucleotide sequence does play a role in specificity, and analysis of non-contiguous sequence contexts may be more appropriate for identifying motifs most ideal for enzymatic targeting.

In vitro experimentation employing recombinant forms of five R5CMTs with C- and U-containing RNA substrates demonstrated specific MTase activity for three enzymes. While methyltransferase activities for rTK2304, rTK1935, and rTK2122 were robust, specific, and validated by LC-MS/MS (Figure 2.4 and supplementary Figure A2.8), rTK0872 and rTK0360 displayed low and non-specific *in vitro* methyltransferase activities on a host of substrates (supplementary Figure A2.9). It is likely that TK0872 and TK0360 activities are regulated *in vivo* by additional factors to ensure faithful modifications within the epitranscriptome.

Analysis of m^5C sites differentially modified in the double deletion strain, ΔTK2304 ΔTK1935, indicate that all but one site in ΔTK1935 and all sites in ΔTK2304 were likewise lost in ΔTK2304 ΔTK1935 (Figure 2.4I). We identified 74 sites with reduced m^5C frequencies and 12 sites with increased m^5C frequencies (greater than twofold change). Thirty-nine of these sites exhibited an absolute loss in modification frequency, and 19 of these sites were uniquely lost in ΔTK2304 ΔTK1935; these 19 sites displayed robust methylation frequencies in TS559 and either single deletion strain, but a complete loss in m^5C signal above background is observed when both enzymes are deleted (Figure 2.4J). All overlapping sites map to coding sequences and little-to-no logos sequence could be discerned (Figure 2.4K) nor does a sequence alignment produce a recognition motif. This would indicate the targeting strategies for these overlapping sites rely on other factors such as RNA structure. Taken together, we have identified a suite of R5CMTs with partial redundancy that install the m^5C epitranscriptome.

Discussion

New and sensitive techniques that combine bioinformatics, biochemistry, next-generation sequencing, genetic manipulations, mass spectrometry, and evolutionary comparisons continue to add to our understanding of temporal changes to the epitranscriptome and how these changes impart phenotype. How m^5C residues impact the fate and functions of mRNAs at the individual transcript level has been historically challenging to address due to low abundance RNAs and substoichiometric modification frequencies in conventional model organisms. It remains contested whether mRNAs are specifically targeted for modification and whether m^5C in coding regions are functionally relevant. Here, we achieved exceedingly deep coverage of the m^5C epitranscriptome and have detected substoichiometric m^5C sites with high-confidence and reproducibility in low abundance RNAs, as well as discovered the enzymes responsible for modification at many of these sites. As such, we have identified several ideal candidate m^5C

sites for future mechanistic studies.

Using a combination of RNA-bisulfite sequencing and mass spectrometry, we showed that *T. kodakarensis* has a densely modified m^5C epitranscriptome that includes diverse RNAs. In bacterial model species, m^5C is largely exclusive to rRNA[40,55], and unlike eukaryotic models, we have shown that mRNAs dominate the pool of m^5C-containing RNAs in *T. kodakarensis*. Although mass spectrometry analysis indicated that tRNAs are rich in m^5C, our sequencing approach excluded tRNAs and other small RNAs, and since no m^5C sites were detected in the 5S rRNA, many additional m^5C sites are likely present in the epitranscriptome than what is presented here. In *T. kodakarensis*, differential m^5C sites were detected across exponential or stationary growing cells where media nutrients are plentiful or depleted, respectively. These data would indicate that m^5C is likely dynamically responding to metabolic cues. Likely due to approximately half of *T. kodakarensis* genes having no annotated function, we were unable to observe enrichment of any particular gene ontology. However, gene ontology enrichment analysis of m^5C-containing mRNAs in human cell lines tend to correlate m^5C with metabolic pathways[50,56]. Numerous studies have observed shifting m^5C-profiles dependent of stress response[57,58], viral infection[59], cell types and tissues[22,48,60–62], and even organellar localization[48,63], repeatedly indicating that the m^5C epitranscriptome is highly responsive to external and internal stimuli.

The regional position of m^5C within an mRNA likely has differential effects on its function. In mammalian species, m^5C in mRNA tends to be enriched in 5’ or 3’ UTRs.[49] In *Arabidopsis*, the evidence is conflicting with several studies placing m^5C in 3’ UTRs, 5’ UTRs or reporting a slight enrichment within coding sequences.[22,57,64] In zebrafish and rice, an enrichment has been observed within coding sequences.[65,66] In *T. kodakarensis*, we observed a slight enrichment in 3’ UTRs. However, this enrichment was only apparent when our strict high-confidence thresholds were met in at least 2/3 replicates and not when datasets were limited to only include sites

detected in all 3 replicates. Otherwise, m^5C appears to be distributed evenly throughout mRNAs. We also did not observe a strong bias in codon position, likely eliminating the possibility of m^5C in mRNA having substantial impacts on wobble base pairing. We did observe an m^5C enrichment in select codon sequences, leading us to speculate whether m^5C residues regulate the translatability of select codons. In mice, NSUN2-mediated methylation of mRNA encoding interleukin-17a had no observable impact on RNA half-life, but led to increased translation output, indicating a direct role of m^5C to regulate the translatability of interleukin-17a.[67] While the mechanism of m^5C-dependent translatability is unclear, further research is needed to discern whether the positional distribution of m^5C within an mRNA leads to unique regulatory effects at specific codons or within specific regions.

Previous work has indicated that some m^5C sites in eukaryal models increase the stability of mRNAs. The exact nature of these stabilizing effects are ill-defined, but studies have demonstrated that 'reader' proteins recognize some m^5C sites, leading to specific regulatory outcomes. It has been shown that these intermolecular partnerships can protect mRNAs from nucleolytic cleavage[68], prolong half-life[17,69], relocate mRNAs within a cell[29], and/or increase protein yield[25]. Apart from acting as a substrate for RNA-binding proteins, m^5C has inherent properties distinct from a naked cytidine. m^5C does not impact Watson-Crick base pairing, but it does increase hydrophobicity and the melting temperature of CG hydrogen bonds[32], and therefore may affect base stacking and overall RNA structure. It is possible that m^5C residues strengthen secondary structures which may impact the function and stability of RNAs. High temperatures are a constant insult to *T. kodakarensis*, and it is possible if not likely that the dense m^5C epitranscriptome provides much needed stability to RNAs.

Several reports have demonstrated that deletion of genes encoding R5CMTs leads to hypersensitivity to heat stress in rice and worms.[65,63] *T. kodakarensis* is a heat-loving archaeon, but the rules of thermophily are poorly understood. The selective employment of m^5C over m^6A,

and its unprecedented abundance would lead us to speculate that m^5C may increase the thermal stability of RNAs. *T. kodakarensis* encodes at least 5 R5CMTs, and we showed that deletion of genes encoding one of 5 R5CMTs results in a temperature-dependent growth defect with increasing severity at 95°C compared to lower temperatures. When multiple R5CMTs were deleted from the same cell, a compounding growth defect was apparent. These data indicate that the m^5C epitranscriptome supports life at extreme temperatures.

The strategies that grant RNA modifying enzymes specificity for distinct positions in the transcriptome have yet to be entirely defined. Previous work by us and others have shown some RNA modifications to have defined sequence contexts surrounding the site of modification. Nat10 in *T. kodakarensis* was shown to target G**C**C motifs for acetylation[2], and the R5CMTs identified here show small ~3-6 nucleotide motifs. Previous studies have also shown that most m^5C sites where the R5CMT has been identified are present in a sequence context of no more than ~4 nucleotides.[22,57,58,70] These short motifs are perplexing, as these sequences do not encode enough complexity to target an enzyme to a select cytidine in the transcriptome without producing thousands of off target modifications. Yet, we see highly precise targeting of RNA modifications *in vivo*. It is possible that non-contiguous sequences may serve as a better benchmark for identifying sequence motifs because RNA is flexible to the extent as to allow nucleotides to bulge out in a way that would permit non-contiguous nucleotides to contact the enzyme. However, this does not explain why we observed differential enzymatic activity levels on the same RNA substrate of varying lengths (Figure 2.4E). Further studies should probe the extent to which structural elements present in RNAs contribute to the specificity of methyltransferases.

Here we have shown that the extensive m^5C epitranscriptome is established and maintained by a suite of enzymes. We first showed that the individual loss of these 5 enzymes leads to the *in vivo* loss of site-specific m^5C *via* RNA BS-seq. Then, we demonstrated the *in vitro*

activity of 3 R5CMTs to install m^5C using radiolabelling assays and mass spectrometry. Of the > 230 m^5C sites detected in the parent strain, and since we were only able to correlate the installation of a quarter of these sites with an R5CMT, there is likely at least one more R5CMT encoded in *T. kodakarensis*. Two MTases we screened, TK1785 and TK1933, are essential and could not be deleted. Annotations would strongly suggest TK1785 encodes a R5CMT, and its essentiality would likewise indicate the essentiality of the m^5C epitranscriptome. It is also possible that cytidines may be redundantly targeted for modification and deletion of all R5CMT would be necessary to remove all m^5C sites. It has long been speculated but sparse evidence would indicate whether R5CMTs share partial redundancy in target sites. To test this, we generated a strain deleted for both TK1935 and TK2304, two genes that encode *bona fide* R5CMTs (Figure 2.4 and supplementary Figure A2.8). The *in vivo* data shows vast changes in the m^5C-profile and many additional losses in m^5C sites not detected in either individual deletion strain, indicating a partially redundant network of R5CMTs with overlapping methylation targets.

We observed hundreds of m^5C sites in *T. kodakarensis* in diverse RNAs and detected changes in m^5C sites in strains deleted for putative RMTases as well as differential m^5C-profiles between growth phases. The dynamic nature of the m^5C profile reinforces that the epitranscriptome is highly responsive to physiological changes. Deletion of R5CMTs result in temperature-dependent growth defects with increasing severity at higher temperatures, consistent with the hypothesis that RNA m^5C supports hyperthermophilic growth. Taken together these data suggest that a suite of R5CMTs with partial redundancy maintain the m^5C epitranscriptome and support life at extreme temperatures.

Materials and Methods

Cell growth

All *T. kodakarensis* strains were grown anaerobically at 85°C in artificial sea-water with

yeast, tryptone (ASW-YT) and supplemented with agmatine.[42] Cultures grown for total ribonucleoside analysis by LC-MS/MS were grown to mid exponential growth phase (OD_{600} = ~0.3) in duplicate before harvested by centrifugation. Culture prepared for bisulfite sequencing were harvested during early exponential growth (OD_{600} 0.1-0.2) and then left to continue growing before harvesting the remaining culture after two hours after reaching its maximum optical density (0.6-0.8, stationary growth phase); 300 ml of culture were harvested by centrifugation during early exponential growth phase and 200 ml of culture were harvested during stationary phase. Cell pellets were stored at -20°C.

Strain construction

For each putative RMTase, the gene and ~700 bp up and downstream were PCR amplified from genomic DNA and gel purified (Qiagen, cat# 28706X4). Primers were modified to include 13 bp terminal extensions with homology to pTS700 at SwaI. All primer sequences can be made available upon request. pTS700 was linearized with SwaI (NEB, cat# R0604S) before the insert was cloned into pTS700 by ligation independent cloning. The coding sequence of the target gene was then removed from the plasmid using QuickChange site directed mutagenesis Agilent cat# 200516). Regions where the target gene overlapped with other genomic elements (i.e. other genes or known promoter sequences) were retained in the plasmid. Deletion-plasmid sequences were confirmed using sanger sequencing. Procedures for generating deletion plasmids are previously reported in Hileman *et al.* (43).

Procedures for generating deletion strains are published in Gehring *et al.* (44). For each RMTase gene targeted for deletion, the corresponding deletion-plasmid was transformed into *T. kodakarensis* strain TS559 and cells were plated on agmatine-free, rich medium. After 2-5 days of growth at 85°C, transformants were picked into rich liquid media lacking agmatine and grown overnight. For each isolate, genomic DNA was extracted from 1 ml of fully grown culture by

phenol/chloroform/isoamyl alcohol (25/24/1; v/v/v) followed by alcohol precipitation in an equal volume of ethanol. Plasmid integration into the genome was confirmed by PCR using two primer pairs, with one primer from each pair having homology to the genome and the other primer to the plasmid. This ensures that the amplicon originates from a genomically integrated sequence.

Isolates where the plasmid had integrated at the predicted loci were plated on minimal medium with the counter selectable marker, 6-methyl purine, and grown anaerobically for 2-5 days at 85°C. Colonies were then picked into rich liquid media and grown overnight. Genomic DNA was purified from each strain and screened for deletion of the target loci. Candidate deletion strains were identified *via* PCR using four sets of primers. First, PCR using primer pairs that flanked the deletion loci (A/B primers) resulted in a reduced amplicon size equivalent to the size of the deleted region compared to the parent strain. Primer sets both with homology to a region internal to the deletion loci (C/D primers) resulted in PCR amplification in the parent strain but not the deletion strain. Two combinations of external and internal primers (A/D and C/B primer pairs), which should not produce an amplicon in the deletion strain were used to ensure the deletion loci was absent in the deleted strain. Candidate deletion strains where PCR indicated a successful deletion were confirmed *via* whole genome sequencing.

To ensure the region targeted for deletion was correctly excised and to ensure no off-target genome modifications were present, each candidate strain was screened by whole genome sequencing on a MinIon platform. Each strain was grown anaerobically overnight in 5 ml of rich medium with agmatine at 85°C. Cultures were pelleted and genomic DNA purified using the Monarch Genomic DNA Purification kit (NEB, cat# T3010S). Library preparation for each strain was completed using the Rapid Barcoding kit 96 (ONT, cat# SQK-RBK110.96). Sequencing was done on the MinIon Mk1C with the R9.4.1 flow cell. Fast5 files were converted to Fastq files using Guppy default settings. Fastq files were aligned to the TS559 reference genome using MiniMap2 and alignment files were compared to the reference genome using

Medaka Variant Calling Pipeline via Neural Networks. Visual inspection of the deletion loci on Integrated Genomics Viewer confirmed deletion of the loci. As a third confirmatory measure, the bisulfite sequencing of RNA (described below) was checked for RNA that aligned to the deleted loci. For all libraries, we observed virtually zero coverage at the deleted loci, fully indicating the gene had been cleanly deleted.

RNA preparation

The universal human reference (UHR) RNA (Agilent, cat# 740000) was generated by pooling 10 cell lines to reduce cell type bias. *T. kodakarensis* cells were resuspended in 1 mL TRIzol and 200 µL chloroform with a 5-10 minute incubation at room temperature proceeding each reagent. Cellular debris were removed *via* centrifugation at 15K RPM for 15 minutes. The aqueous phase was alcohol precipitated in 2.7x the volume of 100% ethanol and incubated at -80°C for 30 minutes. RNA pellets were collected *via* centrifugation at 15K RPM for 30 minutes at 4°C. RNA was resuspended in nuclease free water and treated with DNaseI (NEB, cat# M0303) for 30 minutes in 1X DNase buffer at 37°C. RNA from each sample was purified using the Monarch RNA Clean Up kit (NEB, cat# T2040S or T2030S) or the Zymo RNA Clean and Concentrator kit (Zymo, cat# R1017). A fraction of total RNA was depleted for tRNAs using the Zymo RNA Clean and Concentrator Kit following the manufacturer's protocol for selective recovery of RNA > 200 nt.

rRNA was depleted from a fraction of the RNA samples using reagents provided in the NEBNext rRNA Depletion Kit (NEB, cat# E6310) and custom DNA oligos.[45] The manufacturer's protocol was followed with the following changes: The NEBNext rRNA Depletion Solution provided in the kit was substituted for a mixture of 85 oligonucleotides at 1 µM concentration for each oligo whose sequences were complementary to *T. kodakarensis* rRNA. All volumes for the probe hybridization, RNase H treatment and DNase I treatment sections of the protocol were

scaled up 2-fold and 24 µl of 62.5 ng/µl *T. kodakarensis* RNA was used as the starting material.

LC-MS/MS

Cellular total RNA, rRNA-depleted samples and RNA substrates for the *in vitro* methyltransferase assays (see below) were digested to nucleosides at 37°C overnight using Nucleoside Digestion Mix (NEB, cat# M0649S). Tandem liquid chromatography-mass spectrometry (LC-MS/MS) analysis was performed by injecting digested RNAs on an Agilent 1290 Infinity II UHPLC equipped with a G7117A diode array detector and a 6495C triple quadrupole mass detector operating in the positive electrospray ionization mode (+ESI). UHPLC was carried out on a Waters XSelect HSS T3 XP column (2.1×100 mm, 2.5 µm) with a gradient mobile phase consisting of methanol and 10 mM aqueous ammonium acetate (pH 4.5). MS data acquisition was performed in the dynamic multiple reaction monitoring (DMRM) mode. Each nucleoside was identified in the extracted chromatogram associated with its specific MS/MS transition (m^5C precursor ion *m/z*: 258.1; product ion *m/z*: 126.1). To resolve m^4C and m^5C positional isomers, 0.1% formic acid was used as aqueous mobile phase. The abundance of each nucleoside derived from the digested samples were quantified based on chromatogram peak integration standard curves. Relative abundance of each modified nucleoside was determined by further dividing the peak integrations of modified and unmodified nucleosides (for example: m^5C/C).

For site-specific analysis of the 101 nt rTK2304 substrate, the fragment was instead digested with RNase T1 (Thermo, cat# EN0542) at 37°C for 1 hour and subject to an Eclipse Fusion Orbitrap mass spectrometer (Thermo Fisher Scientific) coupled with a Vanquish UHPLC (Thermo Fisher Scientific). Digested oligonucleotides were separated in a mobile phase gradient consists of buffer A (1% hexafluoroisopropanol (HFIP), 0.1% *N,N*-diisopropylethylamine (DIEA), 1 µM EDTA) and buffer B (90% Methanol, 10% water, 0.075% HFIP, 0.0375% DIEA, 1

μM EDTA) on a C18 column (Waters ACQUITY Premier Oligonucleotide, 1.7 μm, 2.1 x 100 mm). The obtained chromatogram peaks were deconvoluted by a ProMass HR software package (Novatia LLC, USA).

RNA library preparation and bisulfite sequencing

Frozen cell pellets were resuspended in a total volume of 7.5 ml TRI reagent RT (MRC Inc. Cincinnati, OH), vortexed and incubated at room temperature for 5-10 min prior to cooling on ice. 375 μl of 4-bromoanisole (MRC Inc.) was added and tubes were mixed by inversion prior to centrifugation at 12,000 g for 15 min at 4°C. The aqueous layer (~4.5 ml) was removed, 6.75 ml of 100% isopropanol was added, tubes were mixed by inversion and the centrifugation step was repeated for 30 min. Liquid was removed from the tubes, each pellet was washed with 75% ethanol and the centrifugation was repeated for 10 min. The 75% ethanol was removed and pellets were air dried for 5 min or less. Pellets were resuspended in 450 μl nuclease free water (ThermoFisher), 50 μl DNase I buffer, 5 μl DNase I (NEB, cat# M0303) and incubated at 37°C for 30 min. An equal volume of acid-phenol:chloroform, pH4.5 with IAA 125:24:1 (ThermoFisher) was added, tubes were mixed by inversion, centrifuged and aqueous layer removed and precipitated similarly as above. After the 75% ethanol wash step, each pellet was dissolved in nuclease free water and stored at -80°C.

The rRNAs- and tRNA-depletion was performed as described above. Bisulfite treatment of RNA recovered from rRNA depletion or 400-750 ng of total RNA was carried out using an EZ RNA Methylation Kit (Zymo Research Irvine, CA) with a 65°C for 120 min incubation in place of the 54°C for 45 min incubation in the protocol. Eluted RNA was either immediately used for construction of sequencing libraries or stored at -80°C. Libraries for sequencing were prepared using the NEBNext® Ultra™ or Ultra™ II Directional RNA Library Prep Kit for Illumina® (NEB, cat# E7420 or E7760) using the “protocol for use with purified mRNA or rRNA depleted RNA”.

The recommended fragmentation conditions for partially degraded RNA were used in the fragmentation step. Bisulfite-treated libraries were sequenced on one of three Illumina platforms, Replicate 2 of strain TS559 and replicate 1 of all single deletion strains were sequenced on the Next-Seq, with the exception that both replicate 1 and 2 for single deletion strains ΔTK0360 and ΔTK1784 were all sequenced on the Nova-seq. Replicate 3 of strain TS559, replicate 2 for single deletion strains and all double deletion strains were sequenced on the Nova-seq.

Bisulfite sequencing fastq files have been deposited into NCBI Sequence Read Archive (SRA) under BioProject PRJNA937301.

Raw data processing

Fastq reads were subjected to adapter trimming and quality control using Trimmomatic v0.39 using the following command: *PE <sampleID>_R1.fastq.gz strain_R2.fastq.gz ILLUMINACLIP:adapters.fa:5:15:10 SLIDINGWINDOW:1:20 MINLEN:20*. Reads where any base was sequenced with a PHRED score < 20 were removed from further analysis. Reads less than 15 nt were also removed. Using the custom python3 script, BSfilter.py, fastq reads with > 3% cytosine retention (# cytosines / read length) were removed. Fastq files for each pair end were independently mapped to our custom reference genome (*Thermococcus kodakarensis* strain TS559) using BSseeker2 v2.1.8. Fastq reads for pair R1 were reverse complemented prior to mapping using the following command: *python2 /BSseeker2/Antisense.py -i strain_R1_C.fastq -o <sampleID>_R1_CA.fastq*. Mapping was completed using the following command: *python2 bs_seeker2-align.py -i <sampleID>_R1_CA.fastq -g TS559_reference_genome.fasta --temp_dir=/tmp -m 2 --XS=0.03,3 --bt2--mm --bt-p 10 --aligner=bowtie2 -p /bowtie2/.* For mapping, bowtie2 v2.2.14 was used to support BSseeker2. Two mismatches per alignment were allowed (excluding C-to-T mismatches). Bam files for each

pair end were merged using samtools v1.17. Using samtools, alignments were removed from bam files where MAPQ score was < 20 and when detected as a PCR duplicate using the following series of commands: *samtools view -h -b -q 20 <sampleID>_unsorted.bam* | *samtools sort -n -o <sampleID>_mapq_nsorted.bam; samtools fixmate -rm strain_mapq_nsorted.bam; <sampleID>_fixmate.bam; samtools sort <sampleID>_fixmate.bam > <sampleID>_fixmate_csorted.bam; samtools markdup -r <sampleID>_fixmate_csorted.bam <sampleID>_rmdup.bam*. To reduce file size and retain as much data as possible prior to generating CGmaps, each bam file was pseudoreplicated (split) into 10 bam files using the custom python3 script, pseudoreplicate_paired_samfile.py and the following command: *python3 pseudoreplicate_paired_alignment_file.py -r 10 -b -s <strainID> -i <sampleID>_rmdup.bam*. For each strain, the seed (-s) corresponds to the 4 digit TK gene ID (i.e. TS559, 2304, etc). CGmaptools v0.1.2 was used to obtain CGmaps where C/T coverage and C coverage are calculated at each genomically encoded cytidine where total coverage > 0. The following command was used to generate CGmaps for each pseudoreplicate: cgmaptools convert bam2cgmap -b <*sampleID*>_rmdup_split10.bam -g TS559_genome.fasta --rmOverlap -o <*sampleID*>_rmdup_split10. Pseudoreplciate CGmaps were then merged into a single CGmap using the following command: cgmaptools mergelist tosingle -i <*sampleID*>_rmdup_split1.CGmap.gz,<*sampleID*>_rmdup_split2.CGmap.gz…<*sampleID*>_rmdup_split10.CGmap.gz -o <*sampleID*>.CGmap.gz.

R software

CGmaps were then analyzed using R v4.2.2 software. Jupyter notebooks are provided for analysis and figure generation. The tidyverse v2.0.0 collection of packages include dplyr v1.1.0, readr v2.1.4, forcats v1.0.0, stringr v1.5.0, ggplot2 v3.4.1, tibble v3.1.8, lubridate v1.9.2, tidyr v1.3.0, and purrr v1.0.1.

Custom R and python scripts and associated files are provided at: https://github.com/kscott94/The_m5C-epitranscriptome_of_Thermococcus_kodakarensis.

Power calculations for coverage requirements

Statistical power was calculated using the R software and the pwr v1.3.0 library (one-sample t test with effect size (d) = 0.5). Effects sizes were estimated with a significance level of 0.01. For comparison of m^5C frequencies where a 2X fold change is observed between parent and deletion strains, a total coverage of 47X is needed to achieve an 80% power. We therefore required all candidate m^5C sites to have a total coverage of greater than or equal to 47 reads.

Cytosine deamination ratio

The data analysis pipeline for detection of high-confidence and reproducible m^5C sites was written using R software and the Tidyverse collection of packages. CGmaps provide coverage information at all genomically encoded cytidines and were the initial input into the pipeline. We first calculated total cytidine conversion across all sequenced cytosines. We queried the total T or C coverage at all cytidine positions in the reference genome and divided T coverage by C+T coverage (T/(C+T)) at all genomically encoded cytidines, which gave us conversion rates ≥ ~99.8% for all libraries (Figure 2.2B). These metrics indicate that cytidines were adequately deaminated.

High-confidence m^5C detection

Three metrics were implemented to call high-confidence m^5C sites; total coverage, m^5C coverage and m^5C frequency. Based on a power calculation, a total coverage of 47X at individual sites of modification were necessary for adequate m^5C frequency comparisons between strains. In addition to total coverage, m^5C coverage (cytidine coverage) was

parameterized based on m^5C conversion rates at each genomically encoded cytosine and is therefore different for each library. Cytidine coverage at each reference cytosine was to calculate the 99% percentile for cytosine coverage (visually represented as a histogram in Figure 2.2C), and as most sites have very few remaining cytidine due to bisulfite conversion to uridine, we required that for a true modification site, m^5C coverage must be ≥ the 99% percentile. Across 6 parent strain libraries, the minimum m^5C coverage was calculated to be between 5X and 18X, depending on the library. A minimum m^5C frequency was also applied to capture high-confidence and biologically relevant modifications. Modification frequency was calculated by dividing cytidine coverage by cytidine + thymidine coverage (C/(C+T)). From modification sites that met both total and m^5C coverage minima, m^5C frequencies were visualized via histogram (Figure 2.2D). Visual inspection indicates that modification frequencies level-off at ~10%, indicating many modifications with a frequency below 10% may be false-positives. We determined that a high-confidence m^5C site must reach a 10% modification frequency. For sites with exceedingly high total coverage ≥ 1000X, we lowered the modification frequency minima to 5%. These three criteria define high-confidence m^5C sites and were applied to all libraries.

Reproducible m^5C detection

The reproducibility of m^5C sites was also considered. Our main analysis includes sites that met our high-confidence thresholds in at least 2 replicates. Parallel analysis was performed for sites that were determined to be high-confidence in all 3 replicates.

Linear regression of m^5C frequencies

We performed linear regression on modification frequencies between replicates where sites were high-confidence in at least 2 or 3 replicates (supplementary Figure A2.2A and B), using R software; the function geom_smooth(method = “lm”) was used to visualize modification

frequencies across replicates. Pearson's correlation coefficients were calculated in R using cor(x,y method=c("person")) where x and y are modification frequencies between replicates.

Gene expression threshold

To determine the proportion of unique transcripts with and without modification as represented in Figure 2.3A, we calculated the number of genes with expression levels at or above the expression threshold required for total m^5C coverage. As such, genes with an average expression of ≥ 47x at each base were considered to be expressed at or above our thresholds. The expression threshold was applied for at least 2 replicates for m^5C sites reproducible in ≥ 2 replicates (Figure 2.3A) or 3 replicates for m^5C sites reproducible in all 3 replicates (supplementary Figure A2.4A).

m^5C site annotation

The complete annotation of each m^5C site was performed using the two custom python scripts; annotation_pipeline.py and helpers.py. In addition, the reference genome for *T. kodakarensis* strain TS559 and its corresponding gtf annotation file as well as the wildtype KOD1 reference genome were queried. All sites where a m^5C modification was detected were recorded in a text file, each separated by a new line. This line separated file was inputted into annotation_pipeline.py where the complete annotation for each site was generated. All additional analysis and figure generation were performed using R software.

Amino acid and codon bias

The transcriptome-wide occurrence of each codon sequence and amino acid identity were queried using a custom python script, amino_acid_probability.py. Probability of each codon or amino acid were calculated from known or suspected open reading frames for strain TS559. These calculations are recorded in supplementary file 2.

m^{5}C mRNA distribution

Modifications sites that mapped to mRNAs were assigned to a region: CDS, 5' UTR[46] or 3' UTR (unpublished, or within 20 nt of the gene end when not indicated). To determine the enrichment of m^{5}C sites in each region, the number of m^{5}C sites in each region was divided by the average length of each region; 5'UTRs average 17 nt, 3'UTRs average 33 nt, and coding sequences average 1000 nt. When a m^{5}C site maps within a coding sequence, the percent position within the transcript was determined by calculating distance from the transcription start site and divided by the length of the gene. The percent position in the transcript was graphed *via* histogram in R using geom_density().

Logos sequence analysis

A 41 nt region surrounding each m^{5}C site was queried using the annotation pipeline described above. The RNA sequences were analyzed in R using geom_logo() and theme_logo() from the third party package, ggseqlogo v0.1. Sequence alignments as illustrated in supplementary Figure A2.8 were performed using the Clustal Omega Webserver hosted by EMBL-EBI.

Gene ontology enrichment analysis

Gene ontology (GO) analysis of m^{5}C-containing mRNAs was performed using the PANTHER bioinformatics database[47]. GO terms with p value of less than 0.05 were considered as statistically significant. No enrichment was detected.

Phenotypic growth analysis

Individual cultures were passages at least twice before inoculated into 10 mL rich medium with agmatine and sulfur in triplicate. Cultures were grown at 65°, 75°, 85°, or 95° C in a water or oil bath. The optical density at 600 nm (OD_{600}) was measured for each culture every

60 (85°/95°) or 120 (65°/75°) minutes using UV-Vis spectroscopy. The OD_{600} and the standard error was plotted in R using geom_smooth(method = 'loess', SE=TRUE). The maximum OD_{600} of each of the three replicates was divided by that of the average maximum OD_{600} of the control TS559 strain to graph the relative proportion to the control. The doubling time of each replicate was calculated using the following equation: (Δtime * ln(2)) / ln(final OD_{600} / initial OD_{600}). Generally, the initial OD_{600} was taken where the OD first doubled (~ 0.05) and the final OD_{600} was taken where the OD_{600} was that of half the maximum OD_{600} (~0.3 - 0.4). This ensures calculation of the quickest doubling time of the culture. The doubling time of each culture was averaged across 3 replicates, divided by that of the control TS559 strain, and graphed as a relative proportion to the control.

Recombinant protein expression and purification

The gene sequences for each RNA MTase were PCR amplified from the *T. kodakarensis* genome using primers with 17 bp terminal extensions homologous to the expression vector. The coding sequences along with an *E. coli* ribosome entry site were cloned into a pQE80 expression vectors at the EcoRI restriction site using Infusion cloning (Takara Bio., Cat # 638910). For genes where the start codon is GTG in *T. kodakarensis*, we exchanged the start codon for ATG for expression in *E. coli*. Gene sequences were C-terminally tagged with 6 histidines (6xHis) to aid in affinity purification. Transcriptional expression is controlled by the *lac* promoter, and the plasmid confers ampicillin resistance. Expression vectors were transformed into the Nico21 *E. coli* cell line (NEB, cat# C2529H). We additionally transformed the pRARE plasmid (confers chloramphenicol resistance) into the Nico21 cell line to overcome the codon bias between *T. kodakarensis* and *E. coli*. Cells were grown on ampicillin (100 µg/mL)- and chloramphenicol (25 µg/mL)-containing solid LB medium. A single colony was picked into liquid broth containing the necessary antibiotics for plasmid retention and grown overnight. Overnight cultures were used to inoculate larger cultures in a 1:100 ratio and grown at 37°C with shaking.

At an OD_{600} of 0.3-0.5, cultures were spiked with isopropyl β-d-1-thiogalactopyranoside (IPTG, 200-400 μM final concentration) to induce protein expression and D-sorbitol (2-3% w/v final concentration) to improve protein solubility. Cultures were shaken at 37°C for an additional 3 hours before they were harvested *via* centrifugation. Cell pellets were stored at -20°C.

Cell pellets were thawed and resuspended in buffer A (25 mM Tris-HCl pH 8.0, 500 mM NaCl, 10% glycerol) and sonicated on ice for 30 minutes to lyse the cells. Cellular debris were removed by centrifugation at 15K RPM for 15 minutes. The supernatant was heated to 65°C for 15 minutes to remove most of the *E. coli* proteins before centrifugation at 15K RPM for 15 minutes. The 6xHis-tagged protein in the heat-treated supernatant was purified on an AKTA system using a HiTrap Chelating HP column (Cytiva Life Sciences, cat# 17040901). The column was charged with Ni^{2+} before passing through the head-treated cell lysate. After the protein bound to the column, the column was washed with buffer A until the UV spectra returned to base-line, indicating the removal of unbound *E. coli* proteins. The column was then washed in 20-40 mM imidazole to further remove non-specific proteins. The protein was eluted from the nickel at increasing concentrations of buffer B (25 mM Tris-HCl pH 8.0, 100 mM NaCl, 10% glycerol, 200 mM imidazole). Protein elution typically peaked at 165 mM imidazole. Elution fractions were analyzed by SDS-PAGE (BioRad, cat# 5678095) under denaturing conditions followed by western blot with antibodies against the 6xHis tag. Fractions where the target protein was cleanly eluted were pooled and dialyzed into storage buffer (25 mM Tris-HCl pH 8.0, 100 mM NaCl, 50% glycerol v/v), and diluted to a 10 μM concentration in storage buffer.

Generation of RNA substrates for in vitro methyltransferase assays

RNA substrates used in methyltransferase activity assays were selected where the RNA was modified *in vivo*, but the m^5C was lost in a strain deleted for an R5CMT. All 23 nt RNA substrates were ordered from IDT and resuspended in nuclease-free water (ThermoFisher

Scientific, cat# AM9932). Full length TK1911 mRNA (TK1911 coding sequence without UTRs) was amplified from the *T. kodakarensis* genome by PCR with primers 5'-atttaggtgacactatagATGACTGACAAGAAGAAGC and 5'-TTATCCCTCACTGCCCCTAAG. The equivalent U-RNA (where the target cytidine is substituted for a uracil) was purchased from Twist Biosciences and PCR amplified using the same primers. The forward primer is tailed with the Sp6 Polymerase promoter sequence. The PCR reaction was run on a 1% TBE agarose gel and gel purified (Quiagen, cat# 28706X4). 500 ng of DNA amplicon was used in an *in vitro* transcription reaction according to the manufacturer's protocol for the HiScribe Sp6 RNA synthesis kit (NEB, cat# E2070S). Reactions were quenched with 2x volume of TRizol reagent (ThermFisher Scientific, cat# 15596026) and a 1:5 ratio of chloroform (ThermoFisher Scientific, Inc). Reactions were mixed by pulse vortex and separated by centrifugation at 15K RPM for 1 minute. The aqueous layer was taken into 1 µL GlycoBlue (ThermoFisher Scientific, cat# AM9515) and 3x the volume of 100% ethanol. RNA samples were incubated at -80°C for 30 minutes then pelleted by centrifugation at 15K RPM for 30 minutes. Reactions were aspirated completely, resuspended in nuclease free water (ThermoFisher Scientific, cat# AM9932) and quantified using the Qubit RNA broad range kit (Thermofisher Scientific, cat# Q10210). RNA samples were concentrated by alcohol precipitation or diluted to 10 µM in nuclease free water.

Methyltransferase activity assay

To confirm the suspected methyltransferase activity of R5CMTs, we performed enzymatic reactions with the recombinantly expressed and purified enzyme and RNA substrates generated *in vitro* or purchased from IDT. Complete reactions were done in a 20 µL volume with 1 µM enzyme in storage buffer, 1 µM RNA, ~1 µM [^{3}H-methyl]-SAM (PerkinElmer, cat# NET155V001MC) and 1x MTase buffer (25 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM DTT). Negative control reactions where 1 reaction component was excluded in place of water or the U-RNA was substituted in place of the C-substrate were performed in parallel. Reactions were

incubated at 95°C for 30 seconds to denature the RNA and then 65°C for 15 minutes to facilitate enzymatic activity. All reactions were completed in biological triplicate.

Reaction temperatures were lowered to 4°C before 15 µL of each reaction was spotted onto cut 1”x1” squares of Hybond-N+ paper (Cytiva, cat # RPN2020B), where the RNA strongly binds to the membrane. The squares were left to dry for 10 minutes before submerged in 200 mL 5% w/v Trichloroacetic acid (TCA, ThermoFisher Scientific, cat# 421455000) for 5-10 minutes with rocking. The TCA was exchanged for fresh TCA 5 times to wash away reaction components while ^{3}H-methylated RNA remained bound to the membrane. After 5 TCA washes, the squares were air dried for 20 minutes, then placed in a scintillation vial containing 5 mL of liquid EcoScint scintillation fluid (FisherScientific, cat#50-899-90184). The β-decay of each reaction was measured in counts per minute (CPM) using the Liquid Scintillation counter TRI-CARB 2900TR (Pickard). CPMs between reactions with C- and U-RNA were compared and p-values were calculated using a two-sample unpaired t-test. Comparisons resulting in a p-value <0.05 earned one star (*) while p-values <0.01 earned two stars (**).

RNA structural analysis

RNA secondary structures and minimum free energies were computed using Vienna RNAfold 2.4.18 webserver or the Vienna RNAfold package v2.4.14. RNA fold predictions for *in vitro* RNAs were done at 65°C, while *in vivo* RNA folds were predicted at 85°C. For local RNA folds, a 41 nt region surrounding the target cytidine was queried from the full length fold. Nucleotides that are base paired are represented by a | symbol while single stranded nucleotides are represented by periods in Supplementary File 1.

REFERENCES

1. Thomas, J.M., Batista, P.J., and Meier, J.L. (2019). Metabolic Regulation of the Epitranscriptome. ACS Chem. Biol. *14*, 316–324.

2. 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., et al. (2020). Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. Nature *583*, 638–643.

3. Birk, M.A., Liscovitch-Brauer, N., Dominguez, M.J., McNeme, S., Yue, Y., Hoff, J.D., Twersky, I., Verhey, K.J., Sutton, R.B., Eisenberg, E., et al. (2023). Temperature-dependent RNA editing in octopus extensively recodes the neural proteome. Cell *186*, 2544–2555.e13.

4. Zhang, C., and Jia, G. (2018). Reversible RNA Modification N1-methyladenosine (m1A) in mRNA and tRNA. Genomics Proteomics Bioinformatics *16*, 155–161.

5. Ohira, T., Minowa, K., Sugiyama, K., Yamashita, S., Sakaguchi, Y., Miyauchi, K., Noguchi, R., Kaneko, A., Orita, I., Fukui, T., et al. (2022). Reversible RNA phosphorylation stabilizes tRNA for cellular thermotolerance. Nature *605*, 372–379.

6. Chikne, V., Doniger, T., Rajan, K.S., Bartok, O., Eliaz, D., Cohen-Chalamish, S., Tschudi, C., Unger, R., Hashem, Y., Kadener, S., et al. (2016). A pseudouridylation switch in rRNA is implicated in ribosome function during the life cycle of Trypanosoma brucei. Sci. Rep. *6*, 25296.

7. He, Y., Zhang, H., Yin, F., Guo, P., Wang, S., Wu, Y., Xin, S., Han, Y., and Zhang, J. (2021). Novel insights into the role of 5-Methylcytosine RNA methylation in human abdominal aortic aneurysm. Front. Biosci. *26*, 1147–1165.

8. Shen, Q., Zhang, Q., Shi, Y., Shi, Q., Jiang, Y., Gu, Y., Li, Z., Li, X., Zhao, K., Wang, C., et al. (2018). Tet2 promotes pathogen infection-induced myelopoiesis through mRNA oxidation. Nature *554*, 123–127.

9. Gokhale, N.S., McIntyre, A.B.R., Mattocks, M.D., Holley, C.L., Lazear, H.M., Mason, C.E., and Horner, S.M. (2020). Altered m6A Modification of Specific Cellular Transcripts Affects Flaviviridae Infection. Mol. Cell *77*, 542–555.e8.

10. Gokhale, N.S., McIntyre, A.B.R., McFadden, M.J., Roder, A.E., Kennedy, E.M., Gandara, J.A., Hopcraft, S.E., Quicke, K.M., Vazquez, C., Willer, J., et al. (2016). N6-Methyladenosine in Flaviviridae Viral RNA Genomes Regulates Infection. Cell Host Microbe *20*, 654–665.

11. Șelaru, A., Costache, M., and Dinescu, S. (2021). Epitranscriptomic signatures in stem cell differentiation to the neuronal lineage. RNA Biol. *18*, 51–60.

12. Heck, A.M., and Wilusz, C.J. (2019). Small changes, big implications: The impact of m6A RNA methylation on gene expression in pluripotency and development. Biochim. Biophys. Acta Gene Regul. Mech. *1862*, 194402.

13. Zhang, M., Zhang, Y., Ma, J., Guo, F., Cao, Q., Zhang, Y., Zhou, B., Chai, J., Zhao, W., and Zhao, R. (2015). The Demethylase Activity of FTO (Fat Mass and Obesity Associated Protein) Is Required for Preadipocyte Differentiation. PLoS One *10*, e0133788.

14. Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. Nature *518*, 560–564.

15. Helm, M., Giegé, R., and Florentz, C. (1999). A Watson-Crick base-pair-disrupting methyl group (m1A9) is sufficient for cloverleaf folding of human mitochondrial tRNALys.

Biochemistry *38*, 13338–13346.

16. Turner, B., Burkhart, B.W., Weidenbach, K., Ross, R., Limbach, P.A., Schmitz, R.A., de Crécy-Lagard, V., Stedman, K.M., Santangelo, T.J., and Iwata-Reuyl, D. (2020). Archaeosine Modification of Archaeal tRNA: Role in Structural Stabilization. J. Bacteriol. *202*. 10.1128/JB.00748-19.

17. Yang, Y., Wang, L., Han, X., Yang, W.-L., Zhang, M., Ma, H.-L., Sun, B.-F., Li, A., Xia, J., Chen, J., et al. (2019). RNA 5-Methylcytosine Facilitates the Maternal-to-Zygotic Transition by Preventing Maternal mRNA Decay. Mol. Cell *75*, 1188–1202.e11.

18. Arango, D., Sturgill, D., Alhusaini, N., Dillman, A.A., Sweet, T.J., Hanson, G., Hosogane, M., Sinclair, W.R., Nanan, K.K., Mandler, M.D., et al. (2018). Acetylation of Cytidine in mRNA Promotes Translation Efficiency. Cell *175*, 1872–1886.e24.

19. Du, H., Zhao, Y., He, J., Zhang, Y., Xi, H., Liu, M., Ma, J., and Wu, L. (2016). YTHDF2 destabilizes m(6)A-containing RNA through direct recruitment of the CCR4-NOT deadenylase complex. Nat. Commun. *7*, 12626.

20. Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120.

21. Yang, Y., Hsu, P.J., Chen, Y.-S., and Yang, Y.-G. (2018). Dynamic transcriptomic m6A decoration: writers, erasers, readers and functions in RNA metabolism. Cell Res. *28*, 616–624.

22. Yang, L., Perrera, V., Saplaoura, E., Apelt, F., Bahin, M., Kramdi, A., Olas, J., Mueller-Roeber, B., Sokolowska, E., Zhang, W., et al. (2019). m5C Methylation Guides Systemic Transport of Messenger RNA over Graft Junctions in Plants. Curr. Biol. *29*,

2465–2476.e5.

23. Xiao, W., Adhikari, S., Dahal, U., Chen, Y.-S., Hao, Y.-J., Sun, B.-F., Sun, H.-Y., Li, A., Ping, X.-L., Lai, W.-Y., et al. (2016). Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing. Mol. Cell *61*, 507–519.

24. Vallecillo-Viejo, I.C., Liscovitch-Brauer, N., Diaz Quiroz, J.F., Montiel-Gonzalez, M.F., Nemes, S.E., Rangan, K.J., Levinson, S.R., Eisenberg, E., and Rosenthal, J.J.C. (2020). Spatially regulated editing of genetic information within a neuron. Nucleic Acids Res. *48*, 3999–4012.

25. Schumann, U., Zhang, H.-N., Sibbritt, T., Pan, A., Horvath, A., Gross, S., Clark, S.J., Yang, L., and Preiss, T. (2020). Multiple links between 5-methylcytosine content of mRNA and translation. BMC Biol. *18*, 40.

26. Li, A., Chen, Y.-S., Ping, X.-L., Yang, X., Xiao, W., Yang, Y., Sun, H.-Y., Zhu, Q., Baidya, P., Wang, X., et al. (2017). Cytoplasmic m6A reader YTHDF3 promotes mRNA translation. Cell Res. *27*, 444–447.

27. Shi, H., Wang, X., Lu, Z., Zhao, B.S., Ma, H., Hsu, P.J., Liu, C., and He, C. (2017). YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA. Cell Res. *27*, 315–328.

28. Boo, S.H., and Kim, Y.K. (2020). The emerging role of RNA modifications in the regulation of mRNA stability. Exp. Mol. Med. *52*, 400–408.

29. Dominissini, D., and Rechavi, G. (2017). 5-methylcytosine mediates nuclear export of mRNA. Cell Res. *27*, 717–719.

30. Bohnsack, M.T., and Sloan, K.E. (2018). The mitochondrial epitranscriptome: the roles of

RNA modifications in mitochondrial translation and human disease. Cell. Mol. Life Sci. *75*, 241–260.

31. Larkin, R.M. (2019). RNA editing implicated in chloroplast-to-nucleus communication. Proceedings of the National Academy of Sciences *116*, 9701–9703.

32. Ontiveros, R.J., Stoute, J., and Liu, K.F. (2019). The chemical diversity of RNA modifications. Biochem. J *476*, 1227–1245.

33. Höfer, K., and Jäschke, A. (2018). Epitranscriptomics: RNA Modifications in Bacteria and Archaea. Microbiol Spectr *6*. 10.1128/microbiolspec.RWR-0015-2017.

34. Jonkhout, N., Tran, J., Smith, M.A., Schonrock, N., Mattick, J.S., and Novoa, E.M. (2017). The RNA modification landscape in human disease. RNA *23*, 1754–1769.

35. Wilkinson, E., Cui, Y.-H., and He, Y.-Y. (2022). Roles of RNA Modifications in Diverse Cellular Functions. Front Cell Dev Biol *10*, 828683.

36. Schaefer, M., Kapoor, U., and Jantsch, M.F. (2017). Understanding RNA modifications: the promises and technological bottlenecks of the “epitranscriptome.” Open Biol. *7*, 170077.

37. Orita, I., Futatsuishi, R., Adachi, K., Ohira, T., Kaneko, A., Minowa, K., Suzuki, M., Tamura, T., Nakamura, S., Imanaka, T., et al. (2019). Random mutagenesis of a hyperthermophilic archaeon identified tRNA modifications associated with cellular hyperthermotolerance. Nucleic Acids Res. *47*, 1964–1976.

38. Dennis, P.P., Tripp, V., Lui, L., Lowe, T., and Randau, L. (2015). C/D box sRNA-guided 2’-O-methylation patterns of archaeal rRNA molecules. BMC Genomics *16*, 632.

39. Legrand, C., Tuorto, F., Hartmann, M., Liebers, R., Jacob, D., Helm, M., and Lyko, F. (2017). Statistically robust methylation calling for whole-transcriptome bisulfite sequencing

reveals distinct methylation patterns for mouse RNAs. Genome Res. *27*, 1589–1596.

40. Lorenz, C., Lünse, C.E., and Mörl, M. (2017). tRNA Modifications: Impact on Structure and Thermal Adaptation. Biomolecules *7*. 10.3390/biom7020035.

41. Trixl, L., and Lusser, A. (2019). The dynamic RNA modification 5-methylcytosine and its emerging role as an epitranscriptomic mark. Wiley Interdiscip. Rev. RNA *10*, e1510.

42. Scott, K.A., Williams, S.A., and Santangelo, T.J. (2021). Thermococcus kodakarensis provides a versatile hyperthermophilic archaeal platform for protein expression. Methods Enzymol. *659*, 243–273.

43. Hileman, T.H., and Santangelo, T.J. (2012). Genetics Techniques for Thermococcus kodakarensis. Front. Microbiol. *3*, 195.

44. Gehring, A.M., Sanders, T.J., and Santangelo, T.J. (2017). Markerless Gene Editing in the Hyperthermophilic Archaeon Thermococcus kodakarensis. Bio Protoc *7*. 10.21769/BioProtoc.2604.

45. Morlan, J.D., Qu, K., and Sinicropi, D.V. (2012). Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. PLoS One *7*, e42882.

46. Jäger, D., Förstner, K.U., Sharma, C.M., Santangelo, T.J., and Reeve, J.N. (2014). Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. BMC Genomics *15*, 684.

47. Mi, H., and Thomas, P. (2009). PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol. Biol. *563*, 123–140.

48. Amort, T., Rieder, D., Wille, A., Khokhlova-Cubberley, D., Riml, C., Trixl, L., Jia, X.-Y., Micura, R., and Lusser, A. (2017). Distinct 5-methylcytosine profiles in poly(A) RNA from

mouse embryonic stem cells and brain. Genome Biol. *18*, 1.

49. Sibbritt, T., Patel, H.R., and Preiss, T. (2013). Mapping and significance of the mRNA methylome. Wiley Interdiscip. Rev. RNA *4*, 397–422.

50. Yang, X., Yang, Y., Sun, B.-F., Chen, Y.-S., Xu, J.-W., Lai, W.-Y., Li, A., Wang, X., Bhattarai, D.P., Xiao, W., et al. (2017). 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m5C reader. Cell Res. *27*, 606–625.

51. Liu, M., Guo, G., Qian, P., Mu, J., Lu, B., He, X., Fan, Y., Shang, X., Yang, G., Shen, S., et al. (2022). 5-methylcytosine modification by Plasmodium NSUN2 stabilizes mRNA and mediates the development of gametocytes. Proc. Natl. Acad. Sci. U. S. A. *119*. 10.1073/pnas.2110713119.

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

53. Sergeeva, O.V., Bogdanov, A.A., and Sergiev, P.V. (2015). What do we know about ribosomal RNA methylation in Escherichia coli? Biochimie *117*, 110–118.

54. Sharma, S., and Entian, K.-D. (2022). Chemical Modifications of Ribosomal RNA. Methods Mol. Biol. *2533*, 149–166.

55. Edelheit, S., Schwartz, S., Mumbach, M.R., Wurtzel, O., and Sorek, R. (2013). Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. PLoS Genet. *9*, e1003602.

56. Song, Y., Jiang, Y., Shi, L., He, C., Zhang, W., Xu, Z., Yang, M., and Xu, Y. (2022). Comprehensive analysis of key m5C modification-related genes in type 2 diabetes. Front.

Genet. *13*, 1015879.

57. Cui, X., Liang, Z., Shen, L., Zhang, Q., Bao, S., Geng, Y., Zhang, B., Leo, V., Vardy, L.A., Lu, T., et al. (2017). 5-Methylcytosine RNA Methylation in Arabidopsis Thaliana. Mol. Plant *10*, 1387–1399.

58. Jian, H., Zhang, C., Qi, Z., Li, X., Lou, Y., Kang, Y., Deng, W., Lv, Y., Wang, C., Wang, W., et al. (2021). Alteration of mRNA 5-Methylcytosine Modification in Neurons After OGD/R and Potential Roles in Cell Stress Response and Apoptosis. Front. Genet. *12*, 633681.

59. Wnuk, M., Slipek, P., Dziedzic, M., and Lewinska, A. (2020). The Roles of Host 5-Methylcytosine RNA Methyltransferases during Viral Infections. Int. J. Mol. Sci. *21*. 10.3390/ijms21218176.

60. Lin, Y., Zhao, Z., Nie, W., Huang, M., Cai, J., Wang, Y., Wang, H., Huang, Y., and Bai, Y. (2023). Overview of distinct 5-methylcytosine profiles of messenger RNA in normal and knock-down NSUN2 colorectal cancer cells. Front. Genet. *14*, 1121063.

61. He, Y., Zhang, Q., Zheng, Q., Yu, X., and Guo, W. (2020). Distinct 5-methylcytosine profiles of circular RNA in human hepatocellular carcinoma. Am. J. Transl. Res. *12*, 5719–5729.

62. Bataglia, L., Simões, Z.L.P., and Nunes, F.M.F. (2022). Transcriptional expression of m6A and m5C RNA methyltransferase genes in the brain and fat body of honey bee adult workers. Front Cell Dev Biol *10*, 921503.

63. Navarro, I.C., Tuorto, F., Jordan, D., Legrand, C., Price, J., Braukmann, F., Hendrick, A.G., Akay, A., Kotter, A., Helm, M., et al. (2021). Translational adaptation to heat stress is mediated by RNA 5-methylcytosine in Caenorhabditis elegans. EMBO J. *40*, e105496.

64. David, R., Burgess, A., Parker, B., Li, J., Pulsford, K., Sibbritt, T., Preiss, T., and Searle, I.R.

(2017). Transcriptome-Wide Mapping of RNA 5-Methylcytosine in Arabidopsis mRNAs and Noncoding RNAs. Plant Cell *29*, 445–460.

65. Tang, Y., Gao, C.-C., Gao, Y., Yang, Y., Shi, B., Yu, J.-L., Lyu, C., Sun, B.-F., Wang, H.-L., Xu, Y., et al. (2020). OsNSUN2-Mediated 5-Methylcytosine mRNA Modification Enhances Rice Adaptation to High Temperature. Dev. Cell *53*, 272–286.e7.

66. Xue, S., Xu, H., Sun, Z., Shen, H., Chen, S., Ouyang, J., Zhou, Q., Hu, X., and Cui, H. (2019). Depletion of TRDMT1 affects 5-methylcytosine modification of mRNA and inhibits HEK293 cell proliferation and migration. Biochem. Biophys. Res. Commun. *520*, 60–66.

67. Wang, N., Tang, H., Wang, X., Wang, W., and Feng, J. (2017). Homocysteine upregulates interleukin-17A expression via NSun2-mediated RNA methylation in T lymphocytes. Biochem. Biophys. Res. Commun. *493*, 94–99.

68. Blanco, S., Bandiera, R., Popis, M., Hussain, S., Lombard, P., Aleksic, J., Sajini, A., Tanna, H., Cortés-Garrido, R., Gkatza, N., et al. (2016). Stem cell function and stress response are controlled by protein synthesis. Nature *534*, 335–340.

69. Chen, X., Li, A., Sun, B.-F., Yang, Y., Han, Y.-N., Yuan, X., Chen, R.-X., Wei, W.-S., Liu, Y., Gao, C.-C., et al. (2019). 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. Nat. Cell Biol. *21*, 978–990.

70. Huang, T., Chen, W., Liu, J., Gu, N., and Zhang, R. (2019). Genome-wide identification of mRNA 5-methylcytosine in mammals. Nat. Struct. Mol. Biol. *26*, 380–388.

71. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. (2008). The Vienna RNA websuite. Nucleic Acids Res. *36*, W70–W74.

CHAPTER 3: A NOVEL N4,N4-DIMETHYLCYTIDINE STABILIZES THE ARCHAEAL RIBOSOME[2]

Summary

Thermal denaturation of translation machinery is lethal in all extant life, but how heat-loving organisms protect ribosome integrity at extreme temperatures is poorly defined. rRNA is among the most densely modified RNAs in the cell, and distinct modification profiles reported in hyperthermophilic ribosomes are known to promote thermophily. Here, we described a unique epitranscriptomic marker in the 16S rRNA helix 31 that is hypermodified and critical for hyperthermophilic growth in *Thermococcus kodakarensis*. In the H31 loop, we mapped N4,N4-dimethylcytidine (m^4_2C) with single nucleotide resolution using bisulfite-sequencing (BS-seq) and LC-MS/MS. Bisulfite-driven deamination of cytidine was previously thought to only be resistive by modifications at C5, however we report that m^4_2C is robustly resistant to deamination. We identified the *bona fide* writer (m^4_2C synthase) that generates m^4_2C *in vivo* and *in vitro*. Structural studies demonstrate m^4_2C synthase adopts a Rossman fold and a unique N-terminal domain. We then demonstrated that m^4_2C synthase methylates assembled ribosomes but not naked or endogenously modified 16S rRNA and defined the catalytic motif. Taken together, we report a novel writer enzyme and epitranscriptomic marker in the ribosomal decoding center that enhances thermophily.

Introduction

In addition to the five canonical bases (A, U, T, C, G), nucleotides can be enzymatically modified and take on new forms and functions. Ribosomal (rRNA) and transfer (tRNA) RNAs

[2]This work was completed by the following contributors: Kristin A. Fluke, Ryan T. Fuchs, Yueh-Lin Tsai, Victoria Talbott, Liam Elkins, Hallie P. Febvre, Nan Dai, Eric J. Wolf, Ryan Czarny, Jackson Schiltz, G. Brett Robb, P. Shing Ho Ivan R. Corrêa Jr., Thomas J. Santangelo. (2023)

are among the most densely and diversely modified RNAs across Domains where modifications have wide-ranging impacts on translation outcomes.[1,2] The function of the vast majority of RNA modifications and the writer enzymes that generate modified nucleotides have yet to be discovered, but recent investigations have shown that some hyperthermophilic archaea modify RNA at unprecedented levels compared to Eukarya and Bacteria.[3,4] These studies have demonstrated that archaeal ribosomes are abundantly modified in some hyperthermophilic archaea, and the epitranscriptome is essential to hyperthermophilic growth, suggesting a role for RNA modifications in cellular thermophily.

Ribosomal RNA is decorated with a plethora of modifications concentrated in structurally conserved regions and are known to impact the structure, function, and molecular interactions of rRNA.[5,6] At the decoding center, tRNAs bind to and decode mRNAs, resulting in the incorporation of amino acids into a nascent polypeptide in the P-site. Amino acid misincorporation occurs at a basal rate of 10^{-3}-10^{-6}, where a steep decline in translational accuracy is apparent at increasing temperatures.[7,8] The physiological consequences of sustained mis-translation are often devastating.[6,9,10] It is therefore unsurprising that the decoding center is a hot-spot for modified residues that likely contribute to the stability and overall integrity of translation machinery.

The strategies employed by hyperthermophiles to maintain optimal translation accuracy under harsh denaturing conditions is unclear, but likely multifaceted. It has been shown that the model hyperthermophile, *T. kodakarensis*, is enriched for 4-acetylcytidine (ac^4C), 5-methylcytidine (m^5C), and 2'-O methylation, and an obvious increase in ac^4C sites dispersed throughout the ribosome is apparent at increasing culture temperatures.[3,4] In Chapter 2, I showed that the decoding center and A, P, and E sites are specifically enriched for m^5C. The loss of either ac^4C or m^5C results in a temperature-dependent growth defect with increasing severity at rising temperatures. It is likely that rRNA modifications support translation under

thermal insults by fortifying ribosome structure and function. How disparate modification profiles are generated and contribute to ribosome function in diverse species and thermophily in heat-loving archaea are only just beginning to be uncovered.

The ribosome has a highly conserved sequence, structure and function across all life, but different species have adopted distinct rRNA modification profiles.[3,11–13] Essential nucleotides in the helix 31 (H31) stem-loop of the small subunit rRNA jut into the P-site. H31 directly contacts the P-site tRNA, translation initiation factors, and ribosomal associated proteins, and structural and biochemical studies have suggested the importance of H31 nucleosides in cellular viability and translational fidelity.[14] There is an incredible diversity in the modification profile of this conserved region across model species. In eukaryotes, a highly modified uridine analog, such as 1-methyl-3-α-amino-α-carboxyl-propylated pseudouridine ($m^1acp^3\Psi$) in humans, is typically present at the apex of the H31 loop. The loss of $m^1acp^3\Psi$ in H31 has a strong correlation with more than 22 distinct cancers, including colon cancer.[15] The extent and complexity to which this uridine is modified varies depending on species. In contrast, all Bacteria bear an entirely conserved 2-methylguanosine (m^2G) and m^5C in the H31 loop.[11] In *Haloferax volcanii,* a moderate thermophilic archaeon with a growth optimum of 45°C, $acp^3\Psi$ sits in the H31 loop.[12] Little else has been reported regarding the modification profile of this region in other archaea, nor have there been reports of this nature in hyperthermophiles. The modification profile of H31 has apparent benefits to translation, but the advantages conferred by select modifications are hitherto unaddressed. Further research is required to elucidate the factors that drove the evolution of variable modification profiles in this otherwise highly conserved region in the ribosomal decoding center.

Here, we report for the first time the modification status of the H31 stem-loop in the hyperthermophilic archaea, *T. kodakarensis*. We have demonstrated a unique modification signature in the ribosomal recoding center. The H31 loop harbors two adjacent nucleosides,

C1-methylpseudouridine ($m^1\Psi$) and N4,N4-dimethylcytidine (m^4_2C). To our surprise, we found that m^4_2C is largely resistant to chemical deamination by sodium bisulfite, and we have mapped a single m^4_2C residue to H31 using bisulfite-sequencing (BS-seq). The identity of this modified nucleotide and surrounding nucleosides was validated using liquid chromatography coupled tandem mass spectrometry (LC-MS/MS). We demonstrated the *in vivo* and *in vitro* activity and solved the atomic structure of m^4_2C synthase, the writer enzyme that generates m^4_2C. To our knowledge, this is the first study to map any m^4_2C residue with single nucleotide resolution or identify the *bona fide* enzyme responsible for its installation. Strains lacking m^4_2C exhibit a growth defect with increasing severity at higher temperatures, suggesting the single m^4_2C residue enhances thermophily and likely impacts translation outcomes under thermal insults.

Results

A unique m^4_2C residue sits in the H31 loop

In an effort to investigate the thermophilic attributes of m^5C, we bisulfite sequenced RNA from *T. kodakarensis* collected during exponential or stationary growth phase. Cytidines modified at the C5 position are resistant to bisulfite-driven deamination and retain their cytidine identity while unmodified cytidines or those modified at different positions in the nucleotide are converted to uridine and sequenced as thymine. cDNA synthesis followed by next-generation sequencing allows comprehensive and quantitative mapping of m^5C sites with single nucleotide resolution where cytidine is retained. We additionally sequenced RNA from strains uniquely deleted for a putative RNA methyltransferase in an attempt to correlate a loss in site-specific modifications with the loss of the enzyme responsible for its installation. In parallel, we sequenced the parent strain TS559 and a strain deleted for gene TK2045, a candidate RNA m^5C methyltransferase. The deletion of TK2045 was confirmed using whole genome Minion

sequencing (Figure 3.1a, DNA-seq). The deletion locus was also largely void of bisulfite sequenced RNA reads (Figure 3.1a, RNA-seq).

In Chapter 2, I showed that bisulfite sequencing revealed more than 230 sites of cytidine retention, and presumably sites of m^5C in diverse RNAs. Here, I show that stains deleted for gene TK2045 exhibited a single loss in cytidine retention that maps to C918 in the 16S rRNA fragment corresponding to an essential nucleotide in the H31 loop (Figure 3.1b). In our parent strain, the modification appears to occur in ~60% of the 16S rRNAs where a sequencing depth of >10,000 was achieved in three biological replicates and across two growth conditions (Figure 3.1c). In ΔTK2045 strains, the modification frequency drops to zero regardless of growth phase, suggesting the enzyme encoded by gene TK2045 is responsible for modifying C918. No other losses in modified cytidines were detected in the transcriptome. This data would suggest that TK2045 is a writer enzyme that modifies C918 in the *T. kodakarensis* 16S rRNA, and this modified cytidine is resistant to bisulfite-driven deamination.

To further investigate the modification status of H31, we collected total RNA and excised the fragment of rRNA that includes H31 (Figure 3.1d). We applied DNA oligonucleotides with complementarity to the sequences flanking the H31 region and RNaseH digestion the hybridized RNA. We excluded the DNA oligos complementary to the H31 region, leaving intact a 60 nt fragment. The H31 fragment was then hybridized to a complementary, biotinylated DNA oligonucleotide and purified using streptavidin-coated magnetic beads (Figure 3.1e). The H31 fragment was further fragmented using RNase T1 or RNase 4, and the resulting subfragments were analyzed by LC-MS/MS (Figure 3.1f and h). In the fragment purified from the parent strain, we did not detect m^5C, rather we detected a mass shift consistent with N4,N4-dimethylcytidine (m^4_2C) (Figure 3.1g). No modifications were detected at C918 in H31 subfragment derived from ΔTK2045 or an unmodified *in vitro* transcribed RNA, indicating a complete loss of m^4_2C in strains deleted for gene TK2045. Interestingly, we did not detect any traces of unmodified

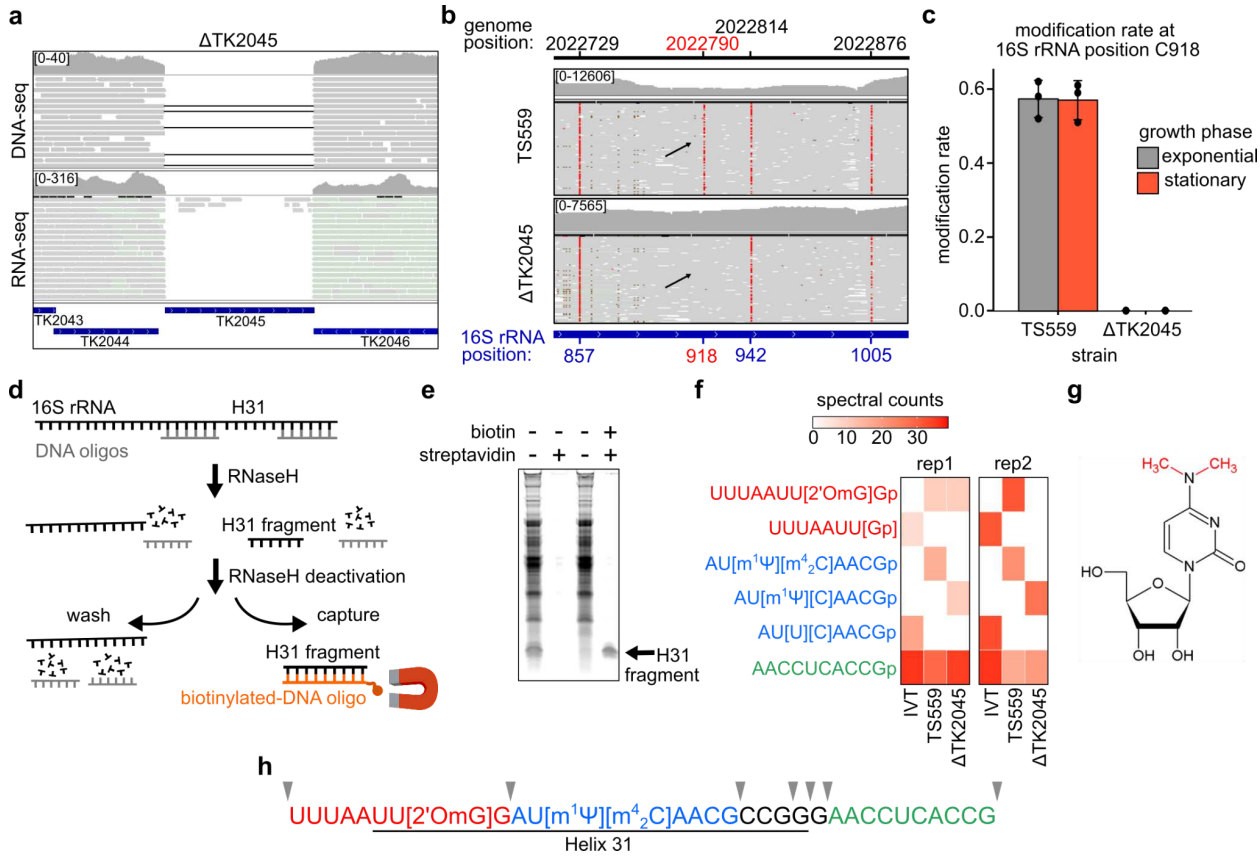


Figure 3.1 Bisulfite-sequencing reveals a single m^4_2C residue in the 16S rRNA helix 31 loop and the novel enzyme responsible for its installation. a, Genomic deletion of gene TK2045 was confirmed using whole genome sequencing (DNA-seq) and RNA-bisulfite sequencing (RNA-seq). Black lines that connect gray bars represent single reads with a gapped alignment. **b,** Bisulfite sequenced RNA collected from TS559 or ΔTK2045 was aligned to a reference genome where all cytidines were replaced with thymines, and visualized using Integrative Genomics Viewer. Sites of modification are indicated where a cytidine is retained (red) within a sequenced read. **c,** The frequency of cytidine retention (modification frequency) in TS559 and ΔTK2045 at position C918 in the 16S rRNA. **d,** Purification schematic to isolate the H31-containing fragment from ribosomal or total RNA. **e,** The supernatant (wash) and streptavidin capture fractions were analyzed on an agarose gel. Only when the biotinylated DNA oligo was included was the H31-containing fragment purified (arrow). **f,** LC-MS/MS analysis of the H31 fragment and unmodified *in vitro* transcribed (IVT) RNA. **g,** The nucleoside structure of m^4_2C. **h,** The modification status of the H31 region. The colors of subfragments are consistent with those illustrated in panel f. The H31 sequence is underlined.

cytidine at C918 in TS559, indicating a virtually completely modified cytidine. This contrasts with the ~60% modification frequency measured by BS-seq. Where m^5C is virtually completely resistant to bisulfite deamination, m^4_2C appears to be partially resistant to the deamination

chemistry. Taken together, these data would strongly indicate that the $m^4{}_2C$ is robustly detectable by BS-sequencing and gene TK2045 encodes an $m^4{}_2C$ writer, which we have tentatively named $m^4{}_2C$ synthase.

The H31 region in model species is highly conserved in sequence, structure and function, but its modification status is divergent between Domains and species.[12] We detected several other modifications in the *T. kodakarensis* H31 fragment (Figure 3.1h); directly adjacent to $m^4{}_2C$, we detected $m^1\Psi$ in the loop, and 2'O-methylguanosine in the stem. The atomic structure of the *T. kodakarensis* 70S ribosome was recently released (PDB: 6TH6) and the nucleotide modifications in the structure appear to be largely consistent with the modifications we identified by LC-MS/MS. In eukaryotic models and *H. volcanii*, a single uridine nucleoside in H31 is modified while in bacteria, two modified nucleosides are present. Not only did we detect a unique modification signature in an essential stem-loop structure, it is unusual to observe 3 modifications in this region, indicating helix 31 is hypermodified in *T. kodakarensis* compared to eukaryotic and bacterial model species.

$m^4{}_2C$ synthase methylates assembled ribosomes

The *in vivo*, site-specific loss of $m^4{}_2C$ upon deletion of TK2045 strongly suggests the enzyme encoded by TK2045 is directly responsible for modifying C918. Although $m^4{}_2C$ synthase is bioinformatically predicted to encode a methyltransferase domain, it is possible that the loss of $m^4{}_2C$ is the outcome of downstream reguatory impacts resulting from the deletion of TK2045. Although no potential intermediate modifications were detected by LC-MS/MS at this site, it is also possible that $m^4{}_2C$ is installed through the cooperation of multiple enzymes, and $m^4{}_2C$ synthase could be installing one of the two methyl moieties or a different modification that is later replaced with $m^4{}_2C$.

To better understand the role of $m^4{}_2C$ synthase in the generation of $m^4{}_2C$, we first sought

to demonstrate its *bona fide* activity *in vitro*. rRNA fragments of varying lengths ranging from 201 nt to 966 nt were synthesized using *in vitro* transcription, resulting in unmodified RNA fragments (supplementary Table A2.1). We also purchased unmodified 23 nt and modified 46 nt fragments where the adjacent modifications we detected in H31 (Figure 3.1h) are present. Recombinant $m^4{}_2C$ synthase (rTK2045) (Figure 3.2a) was challenged to transfer a radiolabeled methyl group from the common methyl cofactor, S-adenosyl-L-methionine (^{3}H-SAM), to the RNA substrate. We did not detect SAM-dependent methylation activity on modified or unmodified substrates (data not shown). These data may indicate that the strategies for substrate recognition or enzymatic activity extend beyond RNA sequence and secondary structure.

Since methyltransferase activity of $m^4{}_2C$ synthase is not apparent on *in vitro* synthesized RNA, we speculated that substrate recognition may rely on additional features only present in fully assembled ribosomes or just the small 30S subunit. We purified fully assembled ribosomes from parent strain TS559 and ΔTK2045 (Figure 3.2b) and challenged $m^4{}_2C$ synthase to methylate either ribosome species under low magnesium conditions where the large and small ribosomal subunits largely dissociate. Ribosomes derived from TS559 contain $m^4{}_2C$ and cannot be modified further at C918, whereas ribosomes derived from ΔTK2045 will have an unoccupied C918 available for $m^4{}_2C$ installation (Figure 3.2c). Site specific activity of $m^4{}_2C$ synthase should therefore only be apparent in substrates derived from ΔTK2045. We also purified rRNA from these ribosomes, generating substrates that are fully modified excluding $m^4{}_2C918$. We detected robust SAM-dependent methyltransferase activity of $m^4{}_2C$ synthase on ribosomes derived from ΔTK2045 but not TS559 or rRNA regardless of its source (Figure 3.2d). Reactions lacking ^{3}H-SAM or enzyme displayed background levels of radioactivity recovered on RNA substrates. To ensure that $m^4{}_2C$ synthase is methylating RNA rather than ribosomal proteins, we performed a series of reactions where each reaction was quenched with ProteinaseK or RNaseA to remove all proteins or RNA from the reaction. We demonstrate that reactions treated with

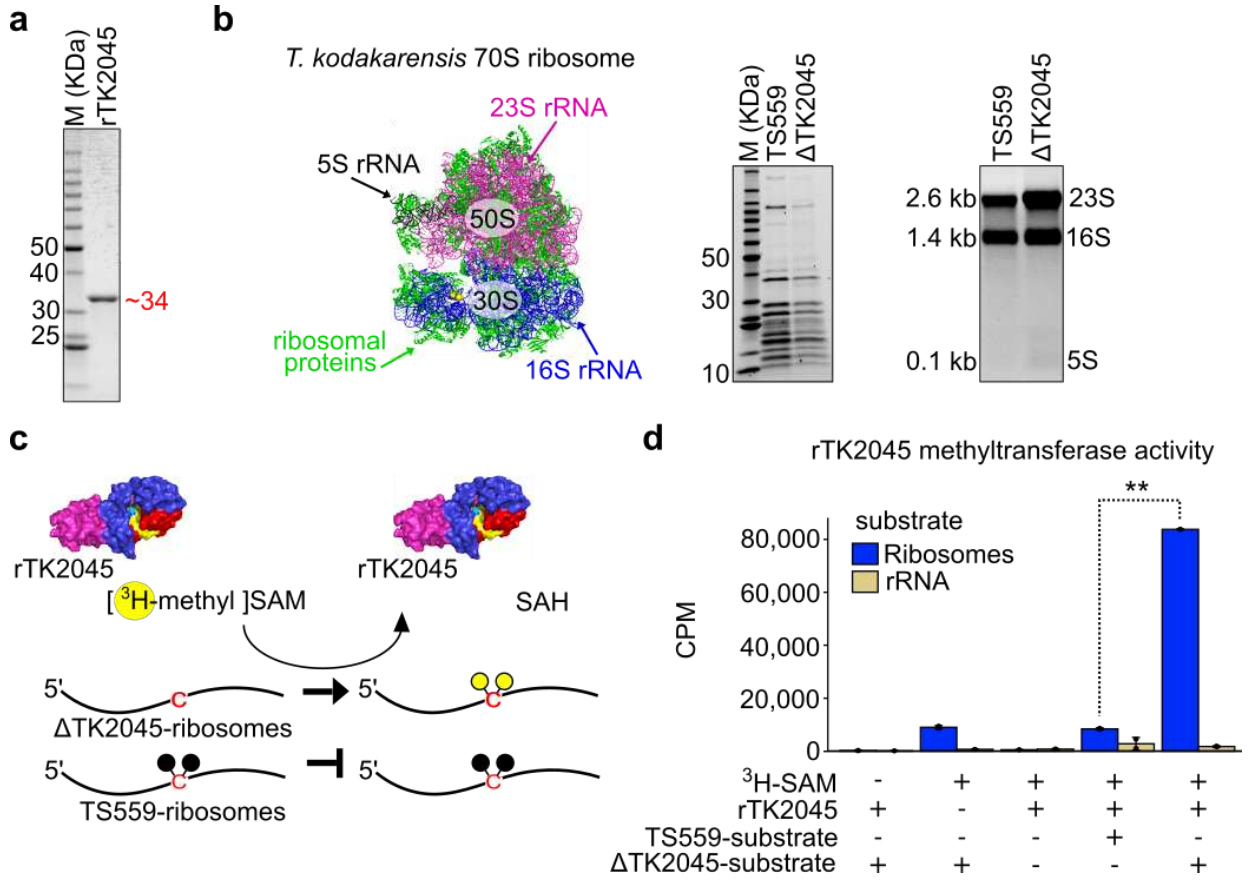


Figure 3.2 Mature ribosomes are the primary substrate of m^4_2C synthase. a, SDS-PAGE of recombinant m^4_2C synthase (rTK2045, ~34 KDa). **b,** SDS-PAGE and agarose gel electrophoresis analysis of ribosomes purified from parent strain TS559 or ΔTK2045. M=marker. **c,** *In vitro* methyltransferase activity assay schematic. **d,** SAM-dependent methyltransferase activity of rTK2045 on ribosomes and rRNA purified from TS559 or ΔTK2045. **p=0.001-4

ProteinaseK retain full activity recovered on rRNA while no activity was recovered from reactions treated with RNaseA (supplementary Figure A3.1), indicating that methylation is to the RNA. Taken together, these data show that m^4_2C synthase is directly involved in the methylation of C918, and ribosomes are the preferred substrate.

H31 is hypermodified in T. kodakarensis

The sequence, structure, and function of the ribosome is conserved across all life, and

although many positions of modification are often conserved, the frequency and identity of modified residues can be highly divergent.[3,12] Sequence comparisons of the H31 region across common model species demonstrate that this region has high sequence similarity within and across Domains (Figure 3.3a). We compared 16 nucleotides in H31 and an additional 10 nucleotides up and downstream of H31. We show that 22 of the 36 nucleotides (~61%) share sequence identity (Figure 3.3a, stars). Decades of structural studies have also demonstrated that H31 adopts a helical, stem-loop structure.[11] The sequence and structural conservation of H31 across evolutionary lineages and its essentiality are highly suggestive that H31 is biologically important, but the exact role of modified nucleotides in H31 is unclear.

Despite high sequence and structural conservation, there is an incredible diversity in the modification profile of H31 across Domains and species (Figure 3.3b). For the first time, we report the modification status of H31 nucleotides in a hyperthermophilic archaea. We detected 3 modified residues in *T. kodakarensis* H31 (Figure 3.1h) including $m^1\Psi$ and $m^4{}_2C$ at U917 and C918, respectively. The modification status of H31 in eukaryotic model species is diverse, but typically includes a uracil analog at the position equivalent to *T. kodakarensis* U917. The only report in archaea indicates that *H. volcanii* includes acp^3U at 910 in the H31 loop. In contrast, all bacterial species bear a highly conserved m^2G and m^5C (at G966 and C967, respectively) in the H31 loop. These data would argue that *T. kodakarensis* hypermodifies H31 compared to what is currently known for model species.

$m^4{}_2C918$ enhances thermophilic growth

Previous reports indicate that the H31 loop directly contacts the P-site tRNA.[11] When viewed in the *T. kodakarensis* 70S ribosome structure (PDB 6TH6, unpublished) the electron density of H31 nucleotides is largely consistent with the modifications reported here (Figure 3.3c), further confirming the identity of these nucleotides. Although the atomic structure depicts

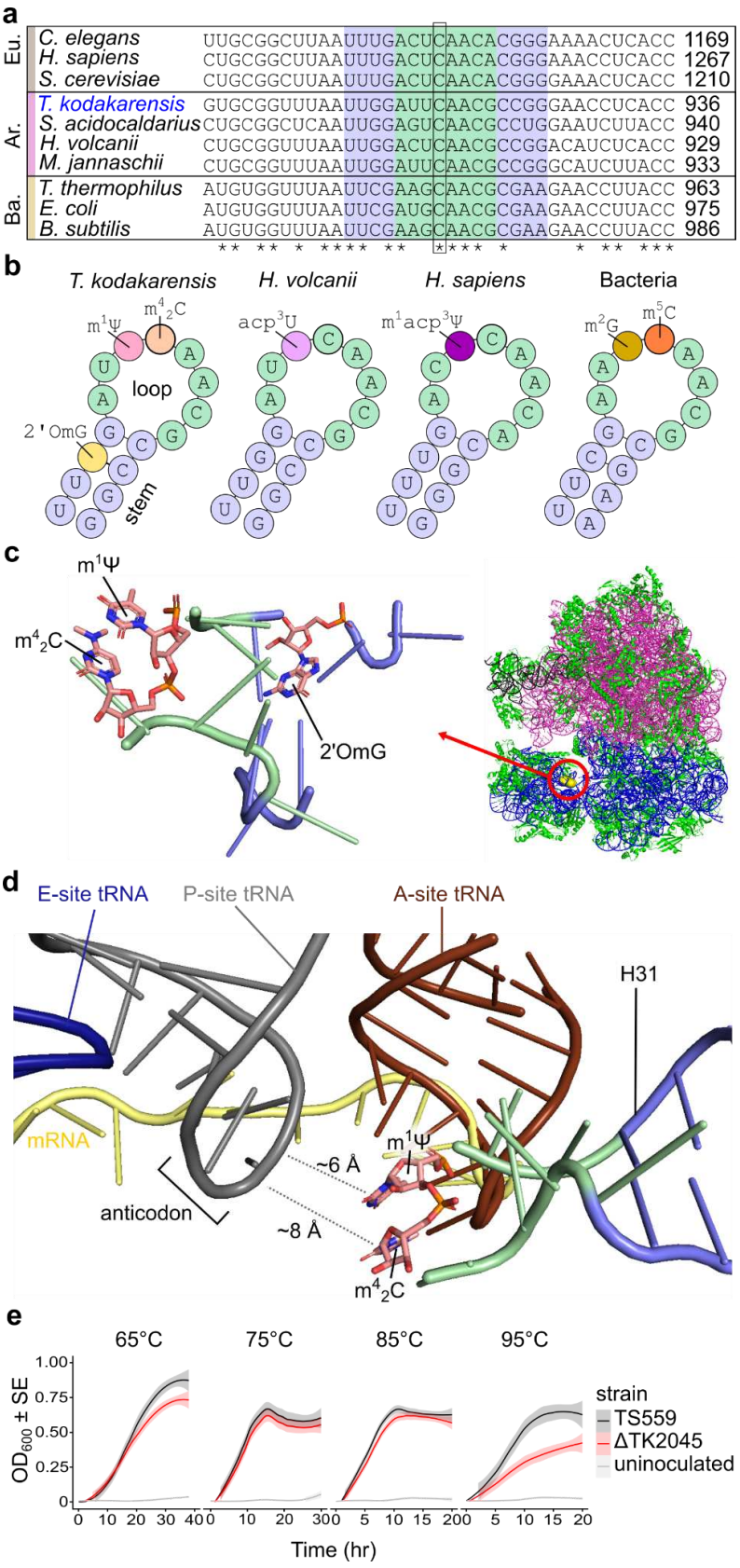

a
Eu.
Ar.
Ba.
C. elegans UUGCGGCUUAAUUUGACUCAACACGGGAAAACUCACC 1169
H. sapiens CUGCGGCUUAAUUUGACUCAACACGGGAAACCUCACC 1267
S. cerevisiae CUGCGGCUUAAUUUGACUCAACACGGGGAAACUCACC 1210
T. kodakarensis GUGCGGUUUAAUUGGAUUCAACGCCGGGAACCUCACC 936
S. acidocaldarius CUGCGGCUCAAUUGGAGUCAACGCCUGGAAUCUUACC 940
H. volcanii CUGCGGUUUAAUUGGACUCAACGCCGGACAUCUCACC 929
M. jannaschii CUGCGGUUUAAUUGGAUUCAACGCCGGGCAUCUUACC 933
T. thermophilus AUGUGGUUUAAUUCGAAGCAACGCGAAGAACCUUACC 963
E. coli AUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACC 975
B. subtilis AUGUGGUUUAAUUCGAAGCAACGCGAAGAACCUUACC 986
b
T. kodakarensis
H. volcanii
H. sapiens
Bacteria
loop
stem
2'OmG
c
m1Ψ
m42C
2'OmG
d
E-site tRNA
P-site tRNA
A-site tRNA
H31
mRNA
anticodon
~6 Å
~8 Å
m1Ψ
m42C
e
65°C
75°C
85°C
95°C
OD600 ± SE
Time (hr)
strain
TS559
ΔTK2045
uninoculated

Figure 3.3 Chemical modifications decorate H31 across Domains and are fitness-relevant in *T. kodakarensis*. a, Sequence alignment of the H31 region across Domains (Eu.=Eukarya, Ar.=Archaea, Ba.=Bacteria). Green and blue highlighted nucleotides correspond to H31. The position of the last nucleotide depicted (3'-most nucleotide) in the small subunit rRNA is listed. Stars indicate conserved nucleotides within the alignment. **b,** Modifications to H31 nucleotides across Domains. **c,** The structure and composition of H31 in the atomic structure of the T. kodakarensis 70S ribosome (PDB 6TH6). **d,** The approximate distance measured between H31 nucleotides and the anticodon loop of the P-site tRNA. **e,** Head-to-head growth competition between the parent strain TS559 and ΔTK2045. Confidence intervals in the regression model are represented by ± 1 standard error (SE).

m^5U at U917, it is likely that cryoelectron microscopy often lacks sufficient resolution to distinguish between the isomerization of the uracil base. Our mass spectrometry results suggest this nucleotide is $m^1\Psi$, not m^5U. When viewed with respects to the A, P, and E site tRNAs, $m^1\Psi$917 and $m^4{}_2C$918 interface with the anticodon loop of the P-site tRNA (Figure 3.3d). The interface between H31 modifications and the P-site tRNA likely indicates a direct impact on how tRNAs dock and translocate through the ribosome.

The position of $m^4{}_2C$ in the ribosome could indicate this modification is important for regulating translation dynamics and cellular fitness.[14] To better understand the phenotypic impact of $m^4{}_2C$, we performed head-to-head growth comparisons between parent strain TS559 and ΔTK2045. *T. kodakarensis* grows over a wide temperature range (~50 to ~98˚C), with a growth optimum of 85˚C. We tracked the growth of ΔTK2045 at 65, 75, 85, and 95˚C over 20-40 hours. A slight growth impairment was noted at 65, 75, 85˚C and a severe growth defect at 95˚C (Figure 3.3e). These data indicate that this single $m^4{}_2C$ residue imparts massive fitness benefits under heat stress, and is critical for hyperthermophilic growth.

$m^4{}_2C$ synthase adopts a canonical Rossman fold

Since $m^4{}_2C$ synthase is a newly discovered enzyme, we first used computational approaches to predict important residues. We then sought to solve its atomic structure by X-ray crystallography to better understand the enzymological strategies of catalysis, SAM binding, and

substrate recognition. The atomic structure was solved at ~2 Å and revealed two distinct domains (Figure 3.4a and b). The C-terminal lobe (red, yellow, and blue) adopts a canonical class I Rossman fold, and bioinformatic predictions indicate a highly conserved domain that contains the methyltransferase and SAM binding residues (red and yellow). The structure nor sequence of the N-terminal lobe is suggestive of its purpose. The surface structure (Figure 3.4c) clearly depicts a hollow core within the C-terminal lobe where cytidine and SAM binding residues are predicted. The electrostatic potential (Figure 3.4d) shows a highly negatively charged pocket where SAM binds and several positively charged residues near the cytidine binding site. The outer surface is highly positively charged on one side opposite of the SAM and cytidine binding pocket.

TGS1 (PDB: 3GHD), a human methyltransferase that dimethylates guanosine at N2, and shares significant structural homology to $m^4{}_2C$ synthase. The atomic structure of TGS1 was solved while bound to guanosine and SAH. Alignment of these TGS1 and $m^4{}_2C$ synthase allowed us to dock the approximate position of SAM and cytidine. These data indicate that SAM and cytidine sit within the hollow core of $m^4{}_2C$ synthase, where the SAM binding residues are predicted to occupy (Figure 3.4e).

RNA m^5C methyltransferases typically encode a DAPC motif, where aspartic acid acts as the catalytic residue.[16] While exocyclic amino methylating enzymes typically adopt a DPPY motif, we instead identified a DPPR motif common to many RNA m^5U methyltransferases.[16] Extensive research conducted over recent decades leads us to speculate that the aspartic acid (D) shares two hydrogen bonds with the cytidine at N4 and N3, and the carboxyl group linking the two prolines (P) hydrogen bond with cytidine at N4 (Figure 3.4f). Since any two amino acids would produce a carboxyl group available to hydrogen bond with cytidine, the importance of either proline is unclear, as is that of the 4th arginine. The first three amino acids of this motif (DPP) are entirely conserved within the 100 closest homologs (Figure 3.5), indicating these

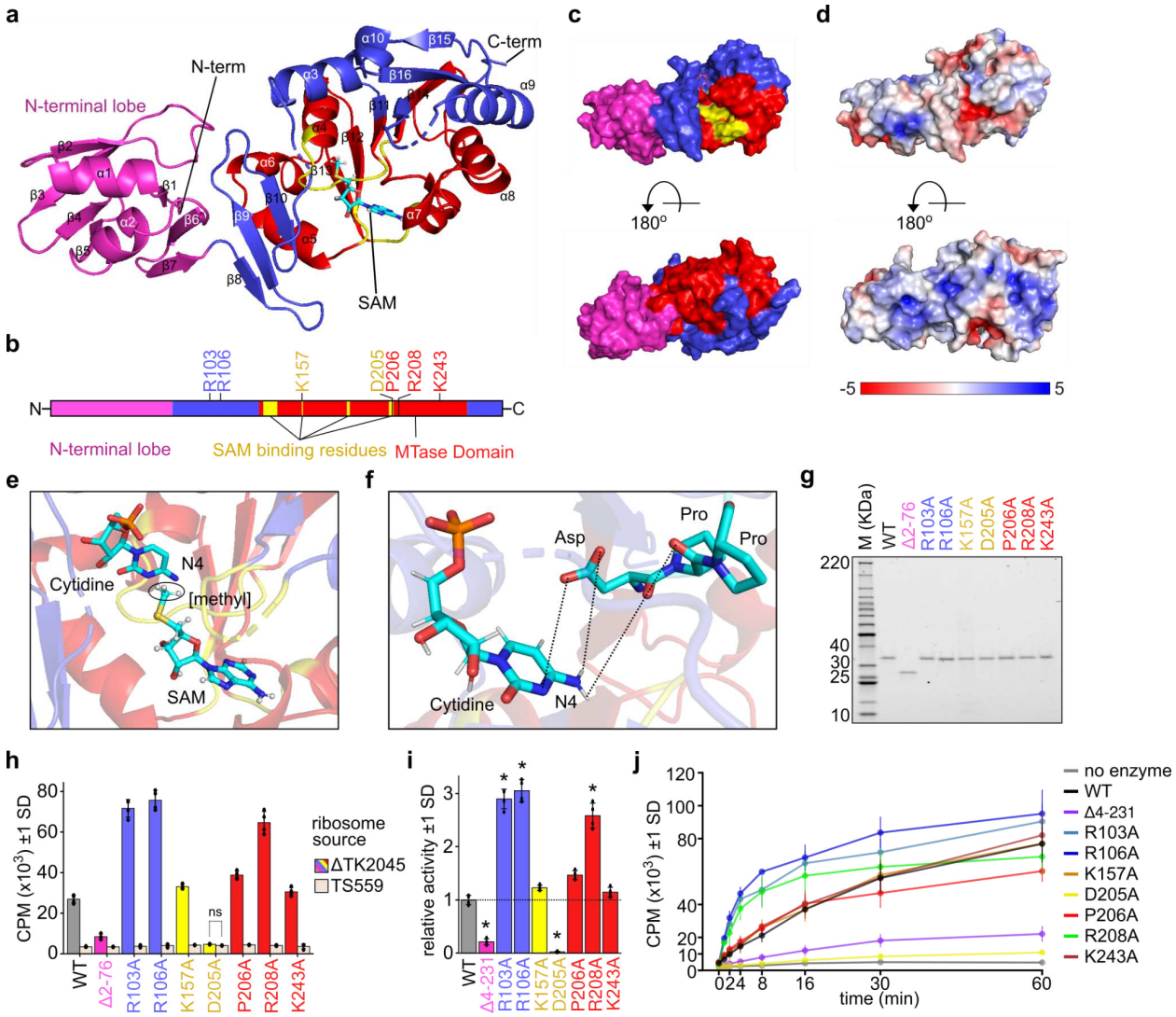


Figure 3.4 Structural and enzymological characterization of $m^4{}_2C$ synthase. a, Atomic structure of $m^4{}_2C$ synthase at 2 Å. The N- and C-termini, α-helices, β-sheets, and SAM are labeled. **b,** Linear representation of m4,2C synthase. The N-terminal lobe (magenta), C-terminal lobe (blue, red, yellow), methyltransferase domain (yellow, red), and SAM binding residues (yellow) are labeled. Amino acid residues of interest are annotated. **c,** Surface structure and **d,** electrostatic potential of $m^4{}_2C$ synthase. **e,** Cytidine and SAM docked into the active site. **f,** proposed interaction between cytidine and the DPPR motif. **g,** Approximate distance between D205 and cytidine. **h,** SDS-PAGE analysis of recombinant $m^4{}_2C$ synthase wild type and variant proteins. **i,** Methyltransferase activity of each enzyme variant relative to wild type protein. **j,** Time course methyltransferase assay.

residues are likely important for catalytic function and are a shared signature of enzymes that generate $m^4{}_2C$. To better gauge the impact of individual amino acid residues to the overall activity of $m^4{}_2C$ synthase, we recombinantly expressed and purified 8 $m^4{}_2C$ synthase variants (Figure 3.4h). The function of the N-terminal lobe is entirely elusive, and so a mutant lacking amino acids 2-76 was generated (the start codon was retained). Additional mutants included single alanine substitutions at residues of the DPPR motif, as well as positively charged residues that may interact with negatively charged RNA substrates. All mutant and wild type enzymes showed little-to-no activity on TS559-derived ribosomes, indicating none of these mutations lead to substrate promiscuity (Figure 3.4h). The D205A mutant (suspected catalytic residue within the DPPR motif) did not show activity above background, strongly indicating it is in fact the catalytic residue (Figure 3.4h-j). The enzyme lacking the N-terminal lobe retains approximately 20% activity compared to wild type enzyme (Figure 3.4h-j), indicating this domain is not required but highly important for methylation. To our surprise, the amino acid identity of P206 and R208 residues within the DPPR motif do not appear to be essential for catalysis, and the R208A substitution actually results in slightly higher midpoint activity levels compared to wild type enzyme (Figure 3.4i). Two additional mutants, R103A and R106A, were predicted to bind to the RNA due to their positive charges, however, both of these mutants showed ~3-fold higher midpoint activity levels. Time-course assays demonstrate that R103A, R106A, and R208A results in markedly faster activity rather than higher endpoint activity (Figure 3.4j). It is possible that single amino acid substitutions that reduce binding affinity to the substrate or SAM cofactor lead to faster release of the methylated product and therefore faster reaction times while retaining tight substrate specificity.

The N-terminal lobe is essential for in vivo $m^4{}_2C$ installation

Although the C-terminal lobe of $m^4{}_2C$ synthase is highly conserved, the N-terminal lobe lacks homology to any protein with a known function. Unique N- or C-terminal lobes of unknown

function are not uncommon in RNA methylating enzymes, and researchers speculate that these unique domains likely participate in substrate recognition. We demonstrated *in vitro* that the N-terminal lobe is not required for site-specific methylation and enzymes lacking the N-terminal lobe retain ~20% activity levels (Figure 3.4j). To further probe the impact of this domain, we measured the *in vivo* modification frequency at C918 in strains lacking nucleotides Δ4-231 of TK2045 (supplementary Figure A3.2a). Bisulfite sequencing of RNA isolated from this strain indicates a complete loss in modification at C918 (supplementary Figure A3.2b). If this mutant protein retains 20% activity, we would expect to see a ~12% modification frequency when accounting for a ~60% deamination resistance. However, we were unable to detect any retained cytidines at this position at over 13000X coverage, indicating the N-terminal lobe is essential for *in vivo* installation of $m^4{}_2C$.

$m^4{}_2C$ is not restricted to Archaea

To better understand the distribution of $m^4{}_2C$ across Domains, we performed total nucleoside analysis of RNA isolated from a variety of species. In the archaea, we detected a robust signal for the presence of $m^4{}_2C$ in two hyperthermophiles, *Methanocaldococcus jannaschii* and *Thermococcus sp.* AM4 (data not shown). BlastP searches through KEGG reveal obvious $m^4{}_2C$ synthase homologs in these species encoded by genes MJ_1233 and TAM4_1357, respectively. We did not detect $m^4{}_2C$ in the hyperthermophile *Sulfolobus islandicus*, the thermophile *Halobacterium salinarum,* or the mesophile *Methanococcus maripaludis* (data not shown). Homologs were likewise not detected within these species or their respective genera. Previous studies have indicated the potential presence of $m^4{}_2C$ in eukaryotes, bacteria, and some human viruses, however we failed to detect $m^4{}_2C$ in human cell lines, yeast, or *E. coli*, nor did we detect homologs in these species (data not shown). A BlastP search for the top 100 KEGG homologs reveal close homologs spanning all domains of life (Figure 3.5a). Although not all heat-loving species generate $m^4{}_2C$, all close, archaeal homologs

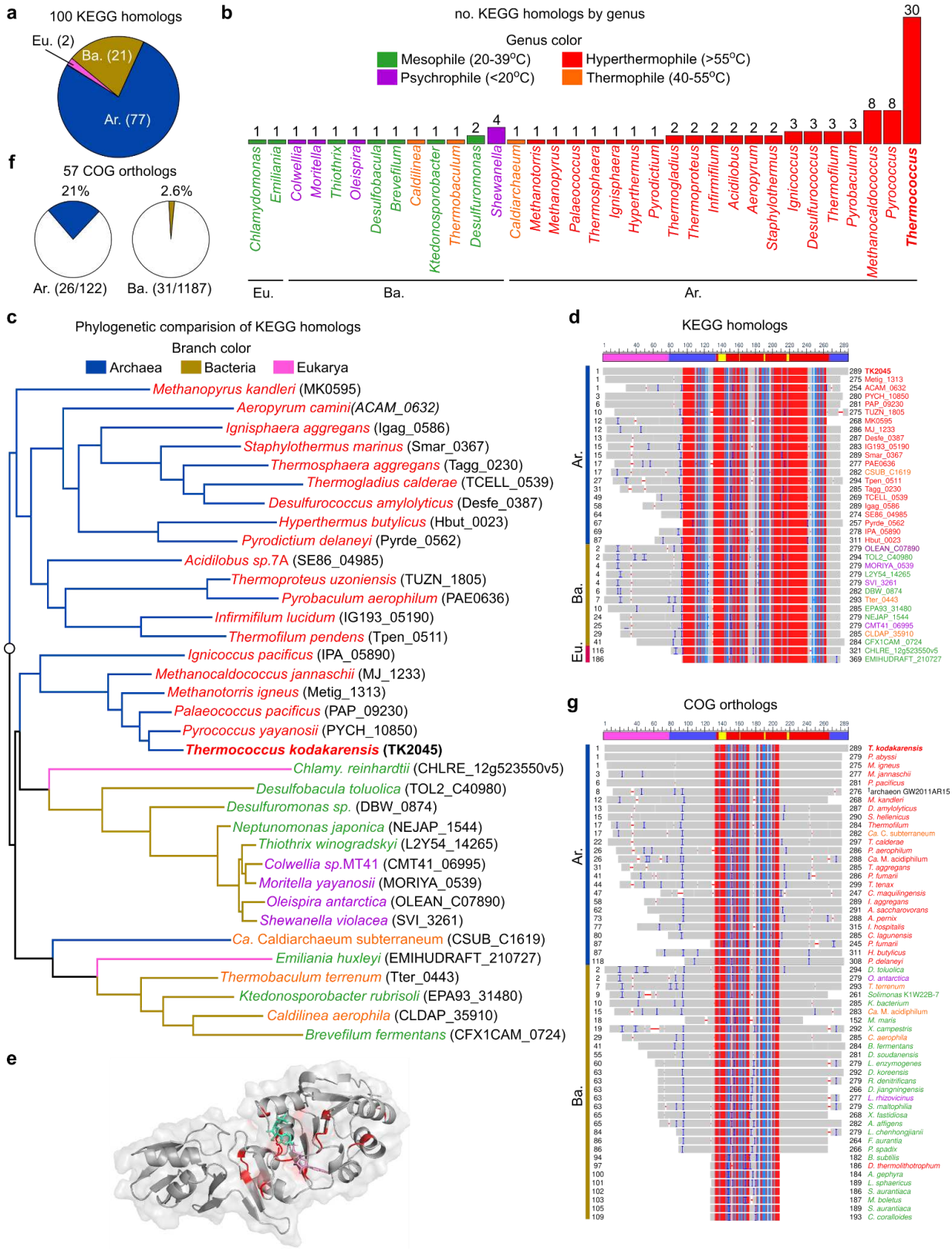

a
100 KEGG homologs
Eu. (2)
Ba. (21)
Ar. (77)
f
57 COG orthologs
21%
2.6%
Ar. (26/122)
Ba. (31/1187)
b
no. KEGG homologs by genus
Genus color
Mesophile (20-39°C)
Psychrophile (<20°C)
Hyperthermophile (>55°C)
Thermophile (40-55°C)
Chlamydomonas
Emiliania
Colwellia
Moritella
Thiothrix
Oleispira
Desulfobacula
Brevefilum
Caldilinea
Ktedonosporobacter
Thermobaculum
Desulfuromonas
Shewanella
Caldiarchaeum
Methanotorris
Methanopyrus
Palaeococcus
Thermosphaera
Ignisphaera
Hyperthermus
Pyrodictium
Thermogladius
Thermoproteus
Infirmifilum
Acidilobus
Aeropyrum
Staphylothermus
Ignicoccus
Desulfurococcus
Thermofilum
Pyrobaculum
Methanocaldococcus
Pyrococcus
Thermococcus
Eu.
Ba.
Ar.
c
Phylogenetic comparision of KEGG homologs
Branch color
Archaea
Bacteria
Eukarya
Methanopyrus kandleri (MK0595)
Aeropyrum camini(ACAM_0632)
Ignisphaera aggregans (Igag_0586)
Staphylothermus marinus (Smar_0367)
Thermosphaera aggregans (Tagg_0230)
Thermogladius calderae (TCELL_0539)
Desulfurococcus amylolyticus (Desfe_0387)
Hyperthermus butylicus (Hbut_0023)
Pyrodictium delaneyi (Pyrde_0562)
Acidilobus sp.7A (SE86_04985)
Thermoproteus uzoniensis (TUZN_1805)
Pyrobaculum aerophilum (PAE0636)
Infirmifilum lucidum (IG193_05190)
Thermofilum pendens (Tpen_0511)
Ignicoccus pacificus (IPA_05890)
Methanocaldococcus jannaschii (MJ_1233)
Methanotorris igneus (Metig_1313)
Palaeococcus pacificus (PAP_09230)
Pyrococcus yayanosii (PYCH_10850)
Thermococcus kodakarensis (TK2045)
Chlamy. reinhardtii (CHLRE_12g523550v5)
Desulfobacula toluolica (TOL2_C40980)
Desulfuromonas sp. (DBW_0874)
Neptunomonas japonica (NEJAP_1544)
Thiothrix winogradskyi (L2Y54_14265)
Colwellia sp.MT41 (CMT41_06995)
Moritella yayanosii (MORIYA_0539)
Oleispira antarctica (OLEAN_C07890)
Shewanella violacea (SVI_3261)
Ca. Caldiarchaeum subterraneum (CSUB_C1619)
Emiliania huxleyi (EMIHUDRAFT_210727)
Thermobaculum terrenum (Tter_0443)
Ktedonosporobacter rubrisoli (EPA93_31480)
Caldilinea aerophila (CLDAP_35910)
Brevefilum fermentans (CFX1CAM_0724)
d
KEGG homologs
Ar.
Ba.
Eu.
TK2045
Metig_1313
ACAM_0632
PYCH_10850
PAP_09230
TUZN_1805
MK0595
MJ_1233
Desfe_0387
IG193_05190
Smar_0367
PAE0636
CSUB_C1619
Tpen_0511
Tagg_0230
TCELL_0539
Igag_0586
SE86_04985
Pyrde_0562
IPA_05890
Hbut_0023
OLEAN_C07890
TOL2_C40980
MORIYA_0539
L2Y54_14265
SVI_3261
DBW_0874
Tter_0443
EPA93_31480
NEJAP_1544
CMT41_06995
CLDAP_35910
CFX1CAM_0724
CHLRE_12g523550v5
EMIHUDRAFT_210727
e
g
COG orthologs
Ar.
Ba.
T. kodakarensis
P. abyssi
M. igneus
M. jannaschii
P. pacificus
archaeon GW2011AR15
M. kandleri
D. amylolyticus
S. hellenicus
Thermofilum
Ca. C. subterraneum
T. calderae
P. aerophilum
Ca. M. acidiphilum
T. aggregans
P. fumarii
T. tenax
C. maquilingensis
I. aggregans
A. saccharovorans
A. pernix
I. hospitalis
C. lagunensis
P. fumarii
H. butylicus
P. delaneyi
D. toluolica
O. antarctica
T. terrenum
Solimonas K1W22B-7
K. bacterium
Ca. M. acidiphilum
M. maris
X. campestris
C. aerophila
B. fermentans
D. soudanensis
L. enzymogenes
D. koreensis
R. denitrificans
D. jiangmingensis
L. rhizovicinus
S. maltophilia
X. fastidiosa
A. affigens
L. chenhongjianii
F. aurantia
P. spadix
B. subtilis
D. thermolithotrophum
A. gephyra
L. sphaericus
S. aurantiaca
M. boletus
S. aurantiaca
C. coralloides

Figure 3.5 Predicted domain and species distribution of $m^4{}_2C$ synthase. a, Distribution of the top 100 KEGG homologs by domain. **b,** Distribution of the top 100 KEGG homologs by genus. Genera are colored by growth temperature. **c,** Phylogenetic comparison of KEGG homologs where one representative species from each genus is included. The gene ID corresponding to each $m^4{}_2C$ synthase homolog is provided. Branches are colored by Domain and organism by growth temperature. **d,** Sequence alignment of KEGG homologs. Redundant genera were removed. Red residues represent highly conserved amino acid positions and blue residues have lower conservation. A linear representation of the protein (from fig.4b) is included for reference. †growth optimum unknown. **e,** Conserved residues within non-redundant KEGG homologs (red) mapped to the atomic structure of $m^4{}_2C$ synthase, including cytidine (green) and SAM (pink). **f,** Percent of archaeal (Ar.) and bacterial (Ba.) species in the COG database that encode $m^4{}_2C$ synthase orthologs. **g,** Sequence alignment of COG orthologs. Redundant genera were removed.

identified here are in hyperthermophilic or thermophilic species (Figure 3.5b). Seventy-seven KEGG homologs are from Archaea (Figure 3.5a), but this protein is not solely restricted to the archaeal domain, nor heat-loving species. Two close homologs are from the mesophilic eukaryotes *Chlamydomonas reinhardtii* (algea) and *Emiliania huxleyi* (protist). The 21 homologs from Bacteria are divided among psychrophilic, mesophilic, thermophilic, and hyperthermophilic species and are from both gram-positive and gram-negative organisms. $m^4{}_2C$ synthase is highly conserved across the *Thermococcus*, *Pyrococcus*, and *Mathanocaldococcus* genera, yet absent from the *Sulfolobus* and *Methanococcus* genera (Figure 3.5b). Phylogenetic comparison of a representative homolog from each genera largely demonstrates that archaeal homologs share more similarity to each other than other Domains, with the expectation of a single homolog from a *Thaumarchaeota*, *Ca*. Caldiarchaeum subterraneum, which clusters near the eukaryotic homolog in *E. huxleyi* (Figure 3.5c). Sequence alignment of representative homologs from each genera indicates that the amino acid positions within the C-terminal lobe are highly conserved across all homologs. The N-terminal lobe is largely conserved within most other homologs, but distinctly missing from many bacterial and both eukaryotic homologs (Figure 3.5d). Within non-redundant genera, 21 amino acids are entirely consvered within the KEGG homologs (Figure 3.5e). These positions are located within the core of the C-terminal lobe and include DPP of the DPPR motif, where the aspartic acid (D) is likely the catalytic residue.

The Clusters of Orthologous Genes (COG) database, which currently include 1187 and 122 bacterial and archaeal genomes, respectively, indicate that 4.3% (57/1309) of these organisms encode an ortholog of $m^4{}_2C$ synthase. Eukaryotic genomes have not been cataloged in COG and therefore the distribution of orthologs within eukaryotes can not be known. Gene TK2045 belongs to COG2521 and arCOG00054, and is annotated as a 'predicted archaeal methyltransferase'. Within the COG catalog, $m^4{}_2C$ synthase orthologs appear in 21% (26/122) of archaeal organisms and 2.6% (31/1187) of bacterial organisms (Figure 3.5f). Similar to our analysis by sequence homology, predicted orthologs appear to be specific to heat-loving archaea, but in bacteria there is no growth temperature bias. Bacterial orthologs are rather divergent and most lack homology to the N-terminal lobes present in archaeal orthologs (Figure 3.5g). We did not detect orthologs in the *Sulfolobus* and *Methanococcus* genera, nor in any halophiles. Although most extremophiles are of archaeal lineage, archaea thrive across all environmental niches, and there is a clear temperature bias within the archaea that include $m^4{}_2C$ synthase. Taken together, this analysis indicates that $m^4{}_2C$ synthase homologs are encoded within all Domains of life, but it is relatively rare and has largely been retained by heat-loving archaea.

Discussion

Investigations into archaeal epitranscriptomes have lagged compared to those in bacterial and eukaryotic model organisms. Several recent reports have probed the epitranscriptomic makeup of hyperthermophilic archaea and have shown that *T. kodakarensis* abundantly and extensively modifies coding and non-coding RNAs at unprecedented levels. *Thermococcus* ribosomes are particularly hypermodified with $ac^4C^{(3)}$ and m^5C (Chapter 2), so it is not surprising that H31 is likewise hypermodified. We report here a unique epitranscriptomic marker in the H31 loop. The exact molecular function of modified nucleotides in H31 is unknown, but it has been consistently demonstrated that the *Thermococcus* epitranscriptome

confers a fitness advantage under heat stress. We show that $m^4{}_2C$ is no exception. The dense and chemically diverse epitranscriptome argues that *Thermococcus* provides an excellent model system for further epitranscriptomic studies that probe the impact of both ubiquitous and rare modifications on core biological processes.

BS-seq is currently the gold standard for quantitative and comprehensive mapping m^5C in DNA and RNA. Bisulfite-driven deamination of cytosine begins with sulfonation at C2 followed by hydrolysis of the exocyclic N4 amino group, resulting in the generation of uracil sulfonate and ultimately uracil. While cytosines are readily deaminated, the deamination of C5-modified cytosines is very slow, taking approximately 12-16 hours.[17] C5-modified cytosines are resistant to the initial conversion step of sulfonation, but m^4C is highly susceptible to bisulfite-driven deamination.[18] We showed that $m^4{}_2C$ is ~60% resistant to deamination. It is unclear why a doubly methylated but not singly methylate cytosine at N4 would be resistant to deamination. It is possible that $m^4{}_2C$ is readily sulfonated while being largly resistant to hydrolysis at N4, resulting in the retention of cytidine. Hundreds of BS-seq datasets that purport m^5C have been published. It is unclear how widespread $m^4{}_2C$ is, but its robust resistance to deamination potentially indicates that existing datasets may be falsely reporting m^5C sites that are actually occupied by $m^4{}_2C$. Future studies that rely on bisulfite deamination should dedicate efforts to orthogonally validating nucleoside identity when reporting the modification status of individual nucleotides.

Several known modifications include the installation of two methyl moieties onto a single nitrogen, but the mechanism by which an enzyme can produce a doubly-methylated product is unclear. A similar enzyme, TGS1, has been shown to deposit two methyl groups into the N2 position of the m^7G mRNA-cap resulting in a hypermodified G-cap.[19] The atomic structure of TGS1 bound to SAH and *in vitro* investigations indicate a single site for SAM binding within TGS1 and results in the installation of both methyl moieties.[20] In the ΔTK2045 strain, we did not

detect m^4C *in vivo* or any other potential intermediates indicative of a multienzyme pathway. Although this class of enzymes is not well studied, previous studies and bioinformatic predictions would lead us to speculate that $m^4{}_2C$ synthase first installs m^4C, releases SAH before binding a new SAM molecule and installing the second methyl group to N4. Only after $m^4{}_2C$ is generated does $m^4{}_2C$ synthase dissociates from the ribosome.

Methylation is the most common chemical modification to RNAs[21], but the mechanisms that dictate the chemistry of methyl transfer are only well understood for a small subset of methylation events. Endocyclic modifications (modifications within the nucleotide ring) are common, and are often catalyzed by 1 or 2 cysteine residues within well defined RNA binding and catalytic motifs.[16,22] RNA m^5C methyltransferases typically require two cysteine residues, one configured in a DAPC motif and another in a TCS motif, to coordinate methyl transfer. Similarly, RNA m^5U methyltransferases usually encode DPPR and SCN motifs for catalysis. However, some exocyclic amino methyltransferases have been shown to encode a DPPY motif while lacking catalytic cysteine residue(s). Here we show that $m^4{}_2C$ synthase encodes a DPPR catalytic motif, reminiscent of RNA m^5U methyltransferases, and we failed to detect a cysteine-containing motif seen in other methyltransferases. It is tempting to speculate that $m^4{}_2C$ methyltransferases are evolutionarily related to m^5U methyltransferases where the DPPR motif was exapted. The loose conservation of the fourth arginine may adumbrate that it is an evolutionary remnant rather than an essential residue.

Materials and Methods

Cell growth

All *T. kodakarensis* strains were grown anaerobically at 85°C in artificial sea-water with yeast, tryptone (ASW-YT) and supplemented with agmatine.[23] Cultures grown for total ribonucleoside analysis by LC-MS/MS were grown to mid exponential growth phase (OD_{600} =

~0.3) in duplicate before harvested by centrifugation. Culture prepared for bisulfite sequencing were harvested during early exponential growth (OD_{600} 0.1-0.2) and then left to continue growing before harvesting the remaining culture after two hours after reaching its maximum optical density (0.6-0.8, stationary growth phase); 300 ml of culture were harvested by centrifugation during early exponential growth phase and 200 ml of culture were harvested during stationary phase. Cell pellets were stored at -20°C.

Strain construction

Procedures for generating deletion plasmids are previously reported in Hileman *et al.* (24). Briefly, the TK2045 genomic loci and ~700 bp up and downstream were PCR amplified from genomic DNA and gel purified (Qiagen, cat# 28706X4). Primers were modified to include 13 bp terminal extensions with homology to pTS700 at SwaI. All primer sequences can be made available upon request. pTS700 was linearized with SwaI (NEB, cat# R0604S) before the insert was cloned into pTS700 by ligation independent cloning. The coding sequence of the target gene or just the N-terminal sequence (nucleotides 4-231) were then removed from the plasmid using QuickChange site directed mutagenesis Agilent cat# 200516). In parallel, HA-6xHis tags were inserted into the plasmid at the C-terminal end of either the TK2045 or TK2045-NΔ plasmids. Regions where the target gene overlapped with other genomic elements (i.e. other genes or known promoter sequences) were retained in the plasmid. Deletion-plasmid sequences were confirmed using sanger sequencing.

Procedures for generating deletion strains are published in Gehring *et al.* (25). The deletion-plasmid was transformed into *T. kodakarensis* strain TS559 and cells were plated on agmatine-free, rich medium. After 2-5 days of growth at 85°C, transformants were picked into rich liquid media lacking agmatine and grown overnight. For each isolate, genomic DNA was extracted from 1 ml of fully grown culture by phenol/chloroform/isoamyl alcohol (25/24/1; v/v/v)

followed by alcohol precipitation in an equal volume of ethanol. Plasmid integration into the genome was confirmed by PCR using two primer pairs, with one primer from each pair having homology to the genome and the other primer to the plasmid. This ensures that the amplicon originates from a genomically integrated sequence. The C-tagged plasmids were transformed into ΔTK2045 strains following the same protocol.

Isolates where the plasmid had integrated at the predicted loci were plated on minimal medium with the counter selectable marker, 6-methyl purine, and grown anaerobically for 2-5 days at 85°C. Colonies were then picked into rich liquid media and grown overnight. Genomic DNA was purified from each strain and screened for deletion of the target loci. Candidate deletion strains were identified *via* PCR using four sets of primers. First, PCR using primer pairs that flanked the deletion loci (A/B primers) resulted in a reduced amplicon size equivalent to the size of the deleted region compared to the parent strain. Primer sets both with homology to a region internal to the deletion loci (C/D primers) resulted in PCR amplification in the parent strain but not the deletion strain. Two combinations of external and internal primers (A/D and C/B primer pairs), which should not produce an amplicon in the deletion strain were used to ensure the deletion loci was absent in the deleted strain. Candidate deletion strains where PCR indicated a successful deletion were next confirmed *via* whole genome sequencing.

To ensure the region targeted for deletion was correctly excised and to ensure no off-target genome modifications were present, each candidate strain was screened by whole genome sequencing on a MinIon platform. Each strain was grown anaerobically overnight in 5 ml of rich medium with agmatine at 85°C. Cultures were pelleted and genomic DNA purified using the Monarch Genomic DNA Purification kit (NEB, cat# T3010S). Library preparation for each strain was completed using the Rapid Barcoding kit 96 (ONT, cat# SQK-RBK110.96). Sequencing was done on the MinIon Mk1C with the R9.4.1 flow cell. Fast5 files were converted to Fastq files using Guppy default settings. Fastq files were aligned to the TS559 reference

genome using MiniMap2 and alignment files were compared to the reference genome using Medaka Variant Calling Pipeline via Neural Networks. Visual inspection of the deletion loci on Integrated Genomics Viewer confirmed deletion of the loci. As a third confirmatory measure, the bisulfite sequencing of RNA (described below) was checked for RNA that aligned to the deleted loci. For all libraries, we observed virtually zero coverage at the deleted loci, fully indicating the gene had been cleanly deleted.

RNA preparation for MS

The universal human reference (UHR) RNA (Agilent, cat# 740000) was generated by pooling 10 cell lines to reduce cell type bias. *T. kodakarensis* cells were resuspended in 1 mL TRIzol and 200 µL chloroform with a 5-10 minute incubation at room temperature proceeding each reagent. Cellular debris were removed *via* centrifugation at 15K RPM for 15 minutes. The aqueous phase was alcohol precipitated in 2.7x the volume of 100% ethanol and incubated at -80°C for 30 minutes. RNA pellets were collected *via* centrifugation at 15K RPM for 30 minutes at 4°C. RNA was resuspended in nuclease free water and treated with DNaseI (NEB, cat# M0303) for 30 minutes in 1X DNase buffer at 37°C. RNA from each sample was purified using the Monarch RNA Clean Up kit (NEB, cat# T2040S or T2030S) or the Zymo RNA Clean and Concentrator kit (Zymo, cat# R1017). A fraction of total RNA was depleted for tRNAs using the Zymo RNA Clean and Concentrator Kit following the manufacturer's protocol for selective recovery of RNA > 200 nt.

rRNA depletion

rRNA was depleted from a fraction of the RNA samples using reagents provided in the NEBNext rRNA Depletion Kit (NEB, cat# E6310) and custom DNA oligos.[26] The manufacturer's protocol was followed with the following changes: The NEBNext rRNA Depletion Solution provided in the kit was substituted for a mixture of 85 oligonucleotides at 1 µM concentration for

each oligo whose sequences were complementary to *T. kodakarensis* rRNA. All volumes for the probe hybridization, RNase H treatment and DNase I treatment sections of the protocol were scaled up 2-fold and 24 µl of 62.5 ng/µl *T. kodakarensis* RNA was used as the starting material.

RNA library preparation and bisulfite sequencing

Frozen cell pellets were resuspended in a total volume of 7.5 ml TRI reagent RT (MRC Inc. Cincinnati, OH), vortexed and incubated at room temperature for 5-10 min prior to cooling on ice. 375 µl of 4-bromoanisole (MRC Inc.) was added and tubes were mixed by inversion prior to centrifugation at 12,000 g for 15 min at 4°C. The aqueous layer (~4.5 ml) was removed, 6.75 ml of 100% isopropanol was added, tubes were mixed by inversion and the centrifugation step was repeated for 30 min. Liquid was removed from the tubes, each pellet was washed with 75% ethanol and the centrifugation was repeated for 10 min. The 75% ethanol was removed and pellets were air dried for 5 min or less. Pellets were resuspended in 450 µl nuclease free water (ThermoFisher), 50 µl DNase I buffer, 5 µl DNase I (NEB, cat# M0303) and incubated at 37°C for 30 min. An equal volume of acid-phenol:chloroform, pH4.5 with IAA 125:24:1 (ThermoFisher) was added, tubes were mixed by inversion, centrifuged and aqueous layer removed and precipitated similarly as above. After the 75% ethanol wash step, each pellet was dissolved in nuclease free water and stored at -80°C.

The rRNAs- and tRNA-depletion was performed as described above. Bisulfite treatment of RNA recovered from rRNA depletion or 400-750 ng of total RNA was carried out using an EZ RNA Methylation Kit (Zymo Research Irvine, CA) with a 65°C for 120 min incubation in place of the 54°C for 45 min incubation in the protocol. Eluted RNA was either immediately used for construction of sequencing libraries or stored at -80°C. Libraries for sequencing were prepared using the NEBNext® Ultra™ or Ultra™ II Directional RNA Library Prep Kit for Illumina® (NEB, cat# E7420 or E7760) using the “protocol for use with purified mRNA or rRNA depleted RNA”.

The recommended fragmentation conditions for partially degraded RNA were used in the fragmentation step. Bisulfite-treated libraries were sequenced on one of three Illumina platforms; replicate 1 and 2 of strain TS559 and replicate 1 of ΔTK2045 were sequenced on the Next-Seq. Replicate 3 of strain TS559, replicate 2 of ΔTK2045 and both replicates of TK2045-NΔ were sequenced on the Nova-seq.

Raw data processing

Fastq reads were subjected to adapter trimming and quality control using Trimmomatic. Reads where any base was sequenced with a PHRED score < 20 were removed from further analysis. Reads less than 15 nt were also removed. Reads with > 3% cytosine retention (# cytosines / read length) were removed using the custom python script, BSfilter.py. Fastq files were then mapped to the reference genome (*Thermococcus kodakarensis* strain TS559) using BSseeker2. Two mismatches per alignment were allowed (excluding C-to-T mismatches). Using samtools, alignments were removed where MAPQ score was < 20 and when detected as a PCR duplicate. CGmaptools was used to obtain CGmaps where C/T coverage and C coverage are calculated at each genomically encoded cytosine where total coverage > 0. Methods for calling high confidence and reproducible modification sites are reported in Chapter 2.

Cytosine modification rate

The data analysis pipeline for detection of high confidence and reproducible m^5C sites was performed using R software and the Tidyverse collection of packages. CGmaps provide coverage information at all genomically encoded cytidines and were the initial input into the data analysis pipeline. We first calculated total cytidine conversion across all sequenced cytosines. We queried the total T or C coverage at all cytidines in the reference genome and divided T coverage by C+T coverage (T/(C+T)) at all genomically encoded cytidines, which gave us conversion rates ≥ 99.8% for all libraries. These metrics indicate that cytidines were adequately

deaminated.

Excision and purification of the H31 fragment

The H31 fragment was purified from total RNA or ribosomes methylated *in vitro* (see below) using the rRNA-depletion method described above but only included DNA oligos flanking the H31 fragment.

LC-MS/MS

Cellular total RNA and RNA substrates for the *in vitro* methyltransferase assays (see below) were digested to nucleosides at 37°C overnight using Nucleoside Digestion Mix (NEB, cat# M0649S). Tandem liquid chromatography-mass spectrometry (LC-MS/MS) analysis was performed by injecting digested RNAs on an Agilent 1290 Infinity II UHPLC equipped with a G7117A diode array detector and a 6495C triple quadrupole mass detector operating in the positive electrospray ionization mode (+ESI). UHPLC was carried out on a Waters XSelect HSS T3 XP column (2.1×100 mm, 2.5 µm) with a gradient mobile phase consisting of methanol and 10 mM aqueous ammonium acetate (pH 4.5). MS data acquisition was performed in the dynamic multiple reaction monitoring (DMRM) mode. Each nucleoside was identified in the extracted chromatogram associated with its specific MS/MS transition. The abundance of each nucleoside derived from the digested samples were quantified based on chromatogram peak integration standard curves. Relative abundance of each modified nucleoside was determined by further dividing the peak integrations of modified and unmodified nucleosides (for example: m^5C/C).

For site-specific analysis of the 60 nt H31 fragment, the fragment was digested with RNase T1 (Thermo, cat# EN0542) or RNase4 (NEB, Inc.) at 37°C for 1 hour and subject to an Eclipse Fusion Orbitrap mass spectrometer (Thermo Fisher Scientific) coupled with a Vanquish UHPLC (Thermo Fisher Scientific). Digested oligonucleotides were separated in a mobile phase

gradient consists of buffer A (1% hexafluoroisopropanol (HFIP), 0.1% *N,N*-diisopropylethylamine (DIEA), 1 µM EDTA) and buffer B (90% Methanol, 10% water, 0.075% HFIP, 0.0375% DIEA, 1 µM EDTA) on a C18 column (Waters ACQUITY Premier Oligonucleotide, 1.7 µm, 2.1 x 100 mm). The obtained chromatogram peaks were deconvoluted by a ProMass HR software package (Novatia LLC, USA).

Phenotypic growth analysis

Individual cultures were passages at least twice before inoculated into 10 mL rich medium with agmatine and sulfur in triplicate. Cultures were grown at 65°, 75°, 85°, or 95° C in a water or oil bath. The optical density at 600 nm (OD_{600}) was measured for each culture every 60 or 120 minutes using UV-Vis spectroscopy. The OD_{600} and the standard error was plotted in R using geom_smooth(method = 'loess', SE=TRUE).

Recombinant protein expression and purification

The gene sequences for TK2045 and variant sequences were designs such that each end included 17 nt of homology to the expression vector. The 5' end also included an *E. coli* ribosome entry site (AGGAGATAATTAA) before the start codon, and the 3' end included 6x Histidine tag (6xHis, CACCATCACCATCACCAT) to aid in affinity purification. Each insert was cloned into a pQE80 expression vector at the EcoRI restriction site using Infusion cloning (Takara Bio., Cat # 638910). Expression vectors confer ampicillin resistance and were transformed into the Rosetta2 *E. coli* cell line (..., cat # …). We note here that the pRARE plasmid in Rosetta2 cells, which overcomes codon bias, is required for the expression of TK2045 and confers resistance to chloramphenicol. Cells were grown on ampicillin (100 ug/mL)- and chloramphenicol (25 ug/mL)-containing solid LB medium. A single colony was picked into liquid broth containing the necessary antibiotics and grown overnight. Overnight cultures were used to inoculate larger cultures in a 1:100 ratio and grown at 37°C with shaking. At an OD_{600} of

~0.3, cultures were spiked with isopropyl β-d-1-thiogalactopyranoside (IPTG, 400 uM final concentration) to induce protein expression and D-sorbitol (3% w/v final concentration) to improve protein solubility. Cultures were shaken at 37°C for an additional 2 hours before cells were harvested *via* centrifugation at 8K RPM for 10 minutes. Cell pellets were stored at -20°C.

Cell pellets were thawed in buffer A (25 mM Tris-HCl pH 8.0, 500 mM NaCl, 10% glycerol), completely resuspended, and sonicated on-and-off on ice for 30 minutes to lyse the cells. Cellular debris were removed by centrifugation at 4K RPM for 30 minutes. The supernatant was heated to 65°C for 15 minutes to denature most of the *E. coli* proteins before centrifugation at 4K RPM for 30 minutes. The 6xHis-tagged protein in the heat-treated supernatant was purified on an AKTA system using a 1 mL HiTrap Chelating HP column (Cytiva Life Sciences, cat# 17040901) charged with Ni^{2+}. The column was washed with 50 column volumes of 80% buffer A and 20% buffer B (25 mM Tris-HCl pH 8.0, 100 mM NaCl, 10% glycerol, 200 mM imidazole) to remove non-specific proteins retained in the column. The protein was eluted from at increasing concentrations of buffer B. Protein elution typically peaked at 165 mM imidazole. Elution fractions were analyzed by SDS-PAGE (BioRad, cat# 5678095) under denaturing conditions followed by western blot with antibodies against the 6xHis tag. Fractions where the target protein was cleanly eluted were pooled and dialyzed into protein storage buffer (25 mM Tris-HCl pH 7.5, 100 mM NaCl, 50% glycerol v/v) and stored at -20°C.

In vitro Generation of RNA substrates

Unmodified 23 nt RNA substrates were purchased from IDT and modified 23 nt substrates were purchased from TriLink. Longer substrates (200 nt, 300 nt, 400, and 966 nt) were generated using *in vitro* transcription using an Sp6 system (NEB cat# E2070S). Reactions were DNaseI treated according to the manufacturer's protocol, purified in 2x the volume of TRIzol reagent, and precipitated in 3x the volume of 100% ethanol. RNA pellets were

suspended in nuclease-free water and stored at -20°C. Sequences for each RNA substrate are provided in supplementary Table A3.1.

Purification of ribosomes and rRNA

Ribosomes were purified from TS559 of ΔTK2045. Cultures were grown in 1L artificial sea water with yeast, tryptone, and pyruvate (ASW-YTP), supplemented with agmatine and sulfur and grown anaerobically to an OD_{600} of ~0.5. ΔTK2045 was also grown up in a 40 L batch in media lacking sulfur using a fermenter. Tightly coupled ribosomes were purified following the protocol outlined in Mehta et al. 2012 ([27]). Briefly, cells were resuspended in ribosome buffer 1 (10 mM $MgCl_2$, 20 mM TrisHCl pH7.5, 300 mM NH_4Cl, 0.5 mM EDTA, 6 mM β-mercaptoethanol) and stored at -20°C. The cell resuspension was subjected to 3 freeze-thaw cycles at -80°C and 50°C to lyse cells in the presence of DNase I. Cellular debris were removed by centrifugation at 15K RPM for 15 minutes. The supernatant was then centrifuged at 40K RPM for 1 hour to pellet the crude ribosome fraction. Crude ribosomes pellets were gently washed and resuspended in ribosome buffer 1 and then layered onto a 30% sucrose cushion (ribosome buffer 1 with 30% (w/v) sucrose). The sample was centrifuged at 40K RPM for 18 hours and the supernatant was poured off. The ribosome pellet was resuspended in ribosome buffer 2 (1 mM $MgCl_2$, 20 mM TrisHCl pH7.5, 300 mM NH_4Cl, 0.5 mM EDTA, 6 mM β-mercaptoethanol) and stored at -20°C.

A portion of each ribosome preparation was subjected to organic extraction using phenol/chloroform/isoamylalcohol (PCI, 25/24/1). 50% volume of PCI was added to the ribosome preparation and mixed vigorously. Samples were centrifuged at 15K RPM for 5 minutes, and the aqueous phase taken into 2.5x the volume of 100% ethanol. Samples were mixed well by shaking and incubated at -20°C for a minimum of 30 minutes before centrifugation at 15K RPM for 30 minutes. rRNA pellets were decanted, washed with 100% ethanol, aspirated completely, and air dried for no more than 5 minutes before resuspended in nuclease free water.

rRNA preparations were stored at -20°C.

Methyltransferase activity assay

Activity of recombinant m^4_2C synthase (rTK2045) to methylate *in vitro* generated RNA, rRNA or ribosomes was measured *in vitro*. Data represented in Figure 3.2d and supplementary Figure A3.2 performed such that reactions in 20 uL volume include 6 or 3 µg RNA, respectively. Also included were 1 uM enzyme, ~1 µM [^{3}H-methyl]-SAM (PerkinElmer, cat# NET155V001MC) and 1x MTase buffer (25 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM $MgCl_2$, 1 mM DTT). Negative control reactions where one reaction component was excluded in place of water were performed in parallel. Reactions were incubated at 65°C for 45 minutes in biological triplicate before 15 uL of each reaction was transferred to a 1 in^2 positively charged nylon membrane(Cytiva, cat # RPN2020B). Midpoint and time course assays (Figure 3.4) were performed in quadruplicate, and a 1:20 enzyme-to-substrate ratio was achieved using 50 nM enzyme and 1 µM substrate. Time course assays were performed in 80 uL of total volume and 10 uL of each reaction were drawn at each time point.

Each membrane was air dried and washed in 300 mL 5% trichloroacetic acid (TCA) for 20 minutes to wash away free ^{3}H-SAM while ^{3}H-methylated RNA remained bound to the membrane. The wash was performed 5 times, each time replacing the TCA. Membranes were air dried and transferred to individual vials containing 3-4 mL Ecoscint scintillation cocktail fluid. The β-decay of each reaction was measured in counts per minute (CPM) using the Liquid Scintillation counter TRI-CARB 2900TR (Pickard). CPMs between reactions were compared and p-values were calculated using a two-sample unpaired t-test. Comparisons resulting in a p-value <0.05 earned one star (*) while p-values <0.01 earned two stars (**). Comparisons where the p-values were greater than 0.05, are denoted “ns”.

X-ray Crystallography

Recombinant protein was dialyzed into crystallography buffer (25 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM DTT), concentrated to 2 mg/mL and screened at Hauptman-Woodward Medical Research Institute Crystallization screening center. The mother liquor that resulted in crystal growth contains 0.5 M AMPD pH 9.3 and 15% (v/v) PEG 3350. Crystal growth was recapitulated using the hanging drop vapor diffusion method in a 24-well plate where AMPD and PEG concentrations were varied. Crystals were collected from wells with 0.5 M AMPD pH ranging from 9.2 to 9.3 and PEG 3350 ranging from 10% to 20% (v/v). Crystals were stored in liquid nitrogen and diffraction data collected at the HHMI Beamline at the ALS Berkeley.

The atomic structure was solved using molecular replacement and the AlphaFold predicted structure. The side chains were removed from the AlphaFold structure and only included the backbone. Side chains were later manually replaced in the solved structure. The final atomic structure was solved at ~2 Å.

Structure analysis

The m^4_2C synthase protein structure in pdb format was analyzed using pymol. The function of amino acid residues and domains were annotated using EMBL-EBI InterPro webserver ([28]). The docked position of cytidine and SAM were predicted based on homologous positions of GTP and SAH in TGS1 (PDB: 3GHD), using the “align” function in pymol. Electrostatic potential was visualized using the pymol plugin, APBS Electrostatic.

Homology and orthology

The amino acid sequence corresponding to TK2045 was subject to a sequence similarity search using the BLASTP function through the Kyoto Encyclopedia of Genes and Genomes (KEGG) web database. The top 100 queries were cataloged as close homologs and met an e value of ≥ 1e-47. A phylogenetic tree was generated on KEGG of the closest match from each unique genera. Within the Clusters of Orthologous Genes (COG) and arCOG databases, TK2045 corresponds to COG2521 and arCOG00054, respectively. Each ortholog was cataloged. Alignments of non-redundant homologs and orthologs were performed using

BLASTP multiple sequence alignment through NCBI and visualized on NCBI MSA viewer. Amino acids are colored according to the “conservation” method which highlights highly conserved and less conserved amino acid positions based on the relative entropy threshold of the residue. Only alignment positions with no gaps are colored. Red indicates highly conserved positions and blue indicates lower conservation.

REFERENCES

1. Jansson, M. D., Häfner, S. J., Altinel, K., Tehler, D., Krogh, N., Jakobsen, E., Andersen, J. V., Andersen, K. L., Schoof, E. M., Ménard, P., Nielsen, H. & Lund, A. H. Regulation of translation by site-specific ribosomal RNA methylation. *Nat. Struct. Mol. Biol.* **28,** 889–899 (2021).
2. Baudin-Baillieu, A. & Namy, O. Saccharomyces cerevisiae, a Powerful Model for Studying rRNA Modifications and Their Effects on Translation Fidelity. *Int. J. Mol. Sci.* **22,** (2021).
3. 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., Washburn, M. P., Isobe, T., Santangelo, T. J., Shalev-Benami, M., Meier, J. L. & Schwartz, S. Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. *Nature* **583,** 638–643 (2020).
4. Lui, L. M., Uzilov, A. V., Bernick, D. L., Corredor, A., Lowe, T. M. & Dennis, P. P. Methylation guide RNA evolution in archaea: structure, function and genomic organization of 110 C/D box sRNA families across six Pyrobaculum species. *Nucleic Acids Res.* **46,** 5678–5691 (2018).
5. Sharma, S. & Lafontaine, D. L. J. 'View From A Bridge': A New Perspective on Eukaryotic rRNA Base Modification. *Trends Biochem. Sci.* **40,** 560–575 (2015).
6. Liang, X.-H., Liu, Q. & Fournier, M. J. Loss of rRNA modifications in the decoding center of the ribosome impairs translation and strongly delays pre-rRNA processing. *RNA* **15,** 1716–1728 (2009).
7. Martinez-Miguel, V. E., Lujan, C., Espie-Caullet, T., Martinez-Martinez, D., Moore, S., Backes, C., Gonzalez, S., Galimov, E. R., Brown, A. E. X., Halic, M., Tomita, K., Rallis, C.,

von der Haar, T., Cabreiro, F. & Bjedov, I. Increased fidelity of protein synthesis extends lifespan. *Cell Metab.* **33,** 2288–2300.e12 (2021).

8. Evans, C. R., Fan, Y. & Ling, J. Increased mistranslation protects E. coli from protein misfolding stress due to activation of a RpoS-dependent heat shock response. *FEBS Lett.* **593,** 3220–3227 (2019).
9. Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M. & Yusupova, G. New structural insights into the decoding mechanism: Translation infidelity via a G·U pair with Watson–Crick geometry. *FEBS Lett.* **587,** 1848–1857 (2013).
10. Pavlov, M. Y. & Ehrenberg, M. Substrate-Induced Formation of Ribosomal Decoding Center for Accurate and Rapid Genetic Code Translation. *Annu. Rev. Biophys.* **47,** 525–548 (2018).
11. Lamichhane, T. N., Abeydeera, N. D., Duc, A.-C. E., Cunningham, P. R. & Chow, C. S. Selection of peptides targeting helix 31 of bacterial 16S ribosomal RNA by screening M13 phage-display libraries. *Molecules* **16,** 1211–1239 (2011).
12. Kowalak, J. A., Bruenger, E., Crain, P. F. & McCloskey, J. A. Identities and phylogenetic comparisons of posttranscriptional modifications in 16 S ribosomal RNA from Haloferax volcanii. *J. Biol. Chem.* **275,** 24484–24489 (2000).
13. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169,** 1187–1200 (2017).
14. Saraiya, A. A., Lamichhane, T. N., Chow, C. S., SantaLucia, J., Jr & Cunningham, P. R. Identification and role of functionally important motifs in the 970 loop of Escherichia coli 16S ribosomal RNA. *J. Mol. Biol.* **376,** 645–657 (2008).
15. Babaian, A., Rothe, K., Girodat, D., Minia, I., Djondovic, S., Milek, M., Spencer Miko, S. E., Wieden, H.-J., Landthaler, M., Morin, G. B. & Mager, D. L. Loss of m1acp3Ψ Ribosomal RNA Modification Is a Major Feature of Cancer. *Cell Rep.* **31,** 107611 (2020).
16. Bujnicki, J. M., Feder, M., Ayres, C. L. & Redman, K. L. Sequence-structure-function

studies of tRNA:m5C methyltransferase Trm4p and its relationship to DNA:m5C and RNA:m5U methyltransferases. *Nucleic Acids Res.* **32,** 2453–2463 (2004).

17. Hayatsu, H., Negishi, K. & Shiraishi, M. DNA methylation analysis: speedup of bisulfite-mediated deamination of cytosine in the genomic sequencing procedure. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **80,** 189–194 (2004).
18. Huo, W., Adams, H. M., Zhang, M. Q. & Palmer, K. L. Genome Modification in Enterococcus faecalis OG1RF Assessed by Bisulfite Sequencing and Single-Molecule Real-Time Sequencing. *J. Bacteriol.* **197,** 1939–1951 (2015).
19. Hausmann, S. & Shuman, S. Specificity and mechanism of RNA cap guanine-N2 methyltransferase (Tgs1). *J. Biol. Chem.* **280,** 4021–4024 (2005).
20. Monecke, T., Dickmanns, A. & Ficner, R. Structural basis for m7G-cap hypermethylation of small nuclear, small nucleolar and telomerase RNA by the dimethyltransferase TGS1. *Nucleic Acids Res.* **37,** 3865–3877 (2009).
21. Sibbritt, T., Patel, H. R. & Preiss, T. Mapping and significance of the mRNA methylome. *Wiley Interdiscip. Rev. RNA* **4,** 397–422 (2013).
22. Bheemanaik, S., Reddy, Y. V. R. & Rao, D. N. Structure, function and mechanism of exocyclic DNA methyltransferases. *Biochem. J* **399,** 177–190 (2006).
23. Scott, K. A., Williams, S. A. & Santangelo, T. J. Thermococcus kodakarensis provides a versatile hyperthermophilic archaeal platform for protein expression. *Methods Enzymol.* **659,** 243–273 (2021).
24. Hileman, T. H. & Santangelo, T. J. Genetics Techniques for Thermococcus kodakarensis. *Front. Microbiol.* **3,** 195 (2012).
25. Gehring, A. M., Sanders, T. J. & Santangelo, T. J. Markerless Gene Editing in the Hyperthermophilic Archaeon Thermococcus kodakarensis. *Bio Protoc* **7,** (2017).
26. Morlan, J. D., Qu, K. & Sinicropi, D. V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS One* **7,** e42882 (2012).

27. Mehta, P., Woo, P., Venkataraman, K. & Karzai, A. W. Ribosome purification approaches for studying interactions of regulatory proteins and RNAs with the ribosome. *Methods Mol. Biol.* **905,** 273–289 (2012).
28. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., Sigrist, C. J. A., Sillitoe, I., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H. & Bateman, A. InterPro in 2022. *Nucleic Acids Res.* **51,** D418–D427 (2023).

CHAPTER 4: THE *IN VIVO* CELLULAR AND MOLECULAR IMPACTS OF THE M5C EPITRANSCRIPTOME[3]

Summary

In chapter 2, I demonstrated that RNA m^5C is highly abundant in *T. kodakarensis* compared to human cells and showed that the m^5C epitranscriptome enhances thermophily. Most RNA modifications are known to stabilize RNA, and our preliminary data support that the structural and functional maintenance of RNA in extreme temperatures is facilitated, in part, by dynamic regulation of the epitranscriptome. We detected > 230 high confidence and reproducible m^5C sites throughout the *T. kodakarensis* transcriptome. The distribution of m^5C places most modified residues within mRNA, but the molecular details surrounding the fate of these RNAs once modified is unknown. It is possible that m^5C stabilizes RNA in extreme heats and prolongs its degradation. It is also possible that m^5C sites may impact the translatability of mRNA, leading to more favorable translation outcomes. The densely modified epitranscriptome coupled with the known collection of writer enzymes provide an ideal platform to investigate the impacts of m^5C on RNA driven processes. **We aim to determine the *in vivo* impact of individual m^5C sites on mRNA and protein levels, and the holistic impact of the m^5C epitranscriptome on hyperthermophilic growth and cellular fitness.** The results generated from this work will provide fundamental insights into the *in vivo* role of the m^5C epitranscriptome.

Background

>170 distinct RNA modifications are known and result in RNAs with unique biochemical properties.[1,2] Although individual RNA modifications might be predicted to have small overall

[3] This work was progressed by the following contributors: Kristin A. Fluke, Victoria Talbott, Liam Elkins, Jackson Schiltz, and Thomas J. Santangelo. (2023)

impacts, defects in generating select RNA modifications result in human diseases.[3] rRNAs and tRNAs are heavily modified, and increasing evidence suggests that modifications to mRNAs are critical for stability and translatability.[4] Major gaps remain regarding how individual modified residues impact the fate of RNA once modified.

5-methylcytidine (m^5C) is one of the most abundant modifications to decorate the transcriptome in Eukarya and Archaea.[5] We demonstrated that m^5C is ~25x more abundant in *T. kodakarensis* compared to human cell lines and the m^5C epitranscriptome includes ~10% of unique transcripts. Using ultradeep bisulfite-sequencing (BS-seq), we mapped > 230 high confidence and reproducible m^5C sites with single nucleotide resolution in our parent strain and strains individually deleted for a unique, putative RNA methyltransferase (RMTase). We identified five RNA m^5C methyltransferases (R5CMTs) that modify coding and non-coding RNAs *in vivo* and three R5CMTs that modify RNA *in vitro*. Strains lacking individual or multiple R5CMTs lack site-specific m^5C modifications that restrict hyperthermophilic growth, demonstrating a requirement for m^5C for hyperthermophily. But, the molecular mechanism underlying the conferral of thermophily by m^5C has yet to be discovered.

Characterization of the role of specific m^5C sites is complicated by low abundance in Eukarya, whereas RNA modifications are typically more abundant in some model archaea.[6,7] Many archaea thrive in conditions inhospitable to most extant life wherein increased RNA stability and fidelity of core biological processes are essential. We have found hundreds of modifications in total RNA from *T. kodakarensis,* representing an order of magnitude higher abundance and frequency than observed in model eukaryotes.[6,8] The abundance of RNA modifications and their multiplicity of potential roles to regulate RNA in environmental extremes indicates that archaea provide the ideal model to pursue fundamental questions regarding epitranscriptomic changes, the role of m^5C, and to generate a suite of thermostable enzymes that provide site- and modification-specific writing activities for manipulation of RNAs.

In this chapter, I will describe the future direction of the m^5C study presented in Chapter 2. We aim to investigate the holistic impact of the m^5C epitranscriptome on hyperthermophilic growth and cellular fitness. I will additionally propose a preliminary study to explore the molecular consequences of individual m^5C residues to impact mRNA and protein levels *in vivo*. These results will elucidate the extent to which m^5C confers thermophily and will elucidate how m^5C residues alter the molecular behavior of RNA.

Results

The contribution of the m5C epitranscriptome to cellular fitness and thermophily

Our preliminary data suggests that cells lacking select m^5C sites experience growth defects that become increasingly more pronounced with rising temperatures. We now know that a collection of at least 5 writer enzymes generate the m^5C epitranscriptome, but we have yet to measure the impact on cells completely lacking RNA m^5C. To better understand the holistic impact of the m^5C epitranscriptome on hyperthermophilic growth, we will generate a collection of strains deleted for multiple and all known R5CMTs and perform phenotypic analysis. Genes that encode R5CMTs include TK0360, TK0872, TK2122, TK1935, and TK2304. The cellular fitness of strains lacking multiple or all known R5CMTs will be assessed using head-to-head growth comparisons at a range of temperatures (65-95°C) and compared to a control strain wherein all m^5C writers are present. I anticipate that cells lacking multiple R5CMTs will exhibit a severe growth defect at increasing temperatures while strains lacking all or nearly all R5CMTs will not be viable. These results will indicate the extent to which the m^5C epitranscriptome promotes survival under heat stress and its essentiality. Strain construction progress in pursuit of this aim are detailed in Table 4.1. Final confirmation of precise genome modifications include whole genome sequencing (Figure 4.1), but most of these strains have not yet been sequenced. Current efforts include continued strain construction, whole genome sequencing of most strains, and phenotypic analysis when all strains have been successfully generated.

Table 4.1 Strain construction progress.

Strain genotype	Trans.[a]	Interm.[b]	XYZ PCR[c]	Finals [d]	I/E PCR[e]	Clean[f]	WGS[g]
ΔTK2304 Δ1935	x	x	x	x	x	x	x
ΔTK2304 Δ1935 Δ0872	x	x	x	x	x	x	
ΔTK2304 Δ1935 Δ2122	x	x	x	x	x	x	
ΔTK2304 Δ1935, p0360B	x	x	x	x	x		
ΔTK2304 Δ1935 Δ0872, p2122B	x	x	x	x			
ΔTK2304 Δ1935 Δ0872, p0360B							
ΔTK2304 Δ1935 Δ2122, p0360B	x	x	x	x			
ΔTK2304 Δ1935 Δ2122, p0872B	x	x	x	x	x		

[a]. Trans. = B-plasmid (deletion plasmid) has been transformed into the parent strain.
[b]. Interim. = Intermediate colonies where the plasmid has “pop’ed in” to the genome at the target loci are cultured.
[c]. XYZ PCR = PCR confirmation of the intermediate genome.
[d]. Finals = Candidates that have “pop’ed out” the plasmid are cultured.
[e]. I/E PCR = PCR confirmation of the final genome.
[f]. Clean = Final strains where the target loci has been deleted are monocultured and wildtype genomes are not present.
[g]. WGS = Whole genome sequencing confirms the correct genotype.

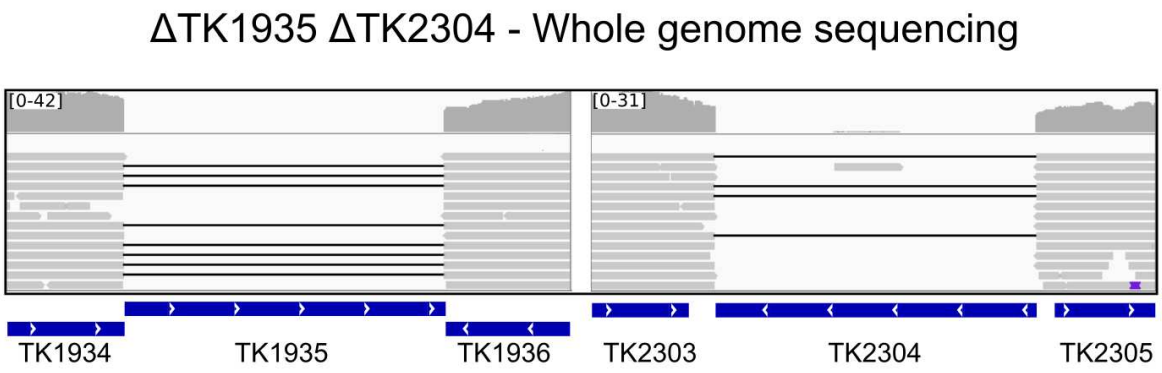


Figure 4.1 Whole genome sequencing confirms the double deletion of TK1935 and TK2304. DNA sequences aligned to the *T. kodakarensis* reference genome were visualized using Integrative Genomics Viewer. In each window, the coverage and read tracks are stacked. Reads absent from the deleted loci indicate markerless removal of the loci encoding the target gene. Black lines connect contiguous reads and indicate a gapped alignment, further confirming the target DNA region is removed. The boundaries of each gene and their TK gene IDs are represented by blue bars, and white arrows illustrate the direction in which the gene is transcribed. The coverage range for each window is listed in brackets.

The in vivo role of m^{5}C to impact stability and translation of mRNAs

We have introduced precision modification to the *T. kodakarensis* genome to alter coding sequences such that the resultant mRNA lacks the cytidine targeted for m^{5}C modification (C→T mutation) and contains a C-terminal epitope- and affinity- tag (C-tag) (Figure 4.2). These transcripts will not be modified as the target cytidine is no longer present. Strains of this nature are referred to as K-strain and are paired with C-strains that encode the target cytidine and a C-tag (Table 4.2). m^{5}C-containing mRNAs selected for *in vivo* studies are based on the criteria that the C->T mutation at the target cytidine would introduce a silent mutation and therefore not alter the protein sequence, the mRNA is expressed as sufficient levels to produce protein at measurable levels, and the m^{5}C modification frequency is at or above 20% *in vivo*. The molecular impact of the m^{5}C on RNA and protein levels will be compared to a control strain where the m^{5}C site is retained (Figure 4.3A, B). I hypothesize that modified RNAs will be more abundant than unmodified RNAs, indicating a role for m^{5}C in RNA stability. Alternatively, it is also possible that modified RNAs will be translated more efficiently, resulting in higher protein levels compared to translation of RNAs lacking m^{5}C.

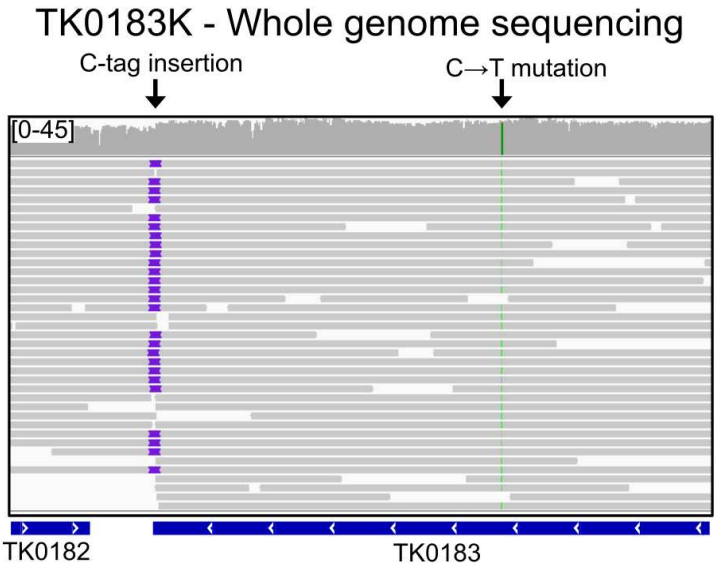


Figure 4.2 Whole genome sequencing of a representative K strain, TK0183K. TK0183 (encoded on the non-reference strand) was tagged at the C terminus resulting in a 41 nt insertion (purple) corresponding to the 6xHis and HA sequences. A C→T mutation (green) is apparent at the cytidine known to be methylated. In each window, the coverage and read tracks are stacked. The boundaries of each gene and their TK gene IDs are represented by blue bars, and white arrows illustrate the direction in which the gene is transcribed. The coverage range is listed in brackets.

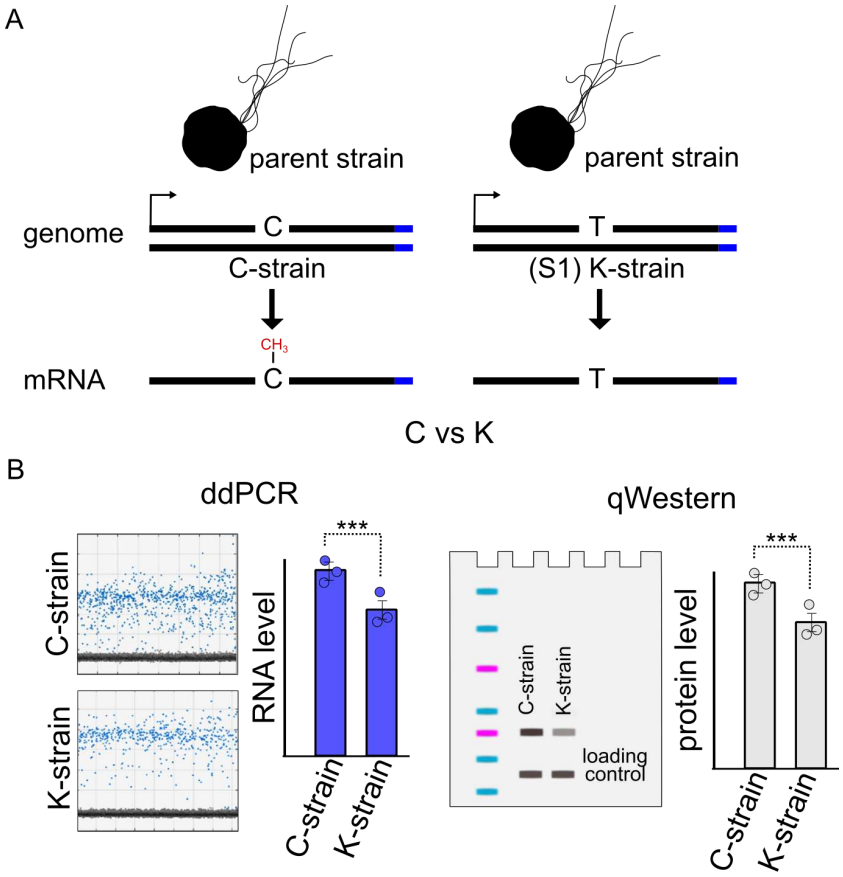


Figure 4.3 (A) Genomic C- and K-strain constructs (in series 1) and the resultant RNA. Blue sequence represents an epitope- and affinity-tag. Genome modifications are in a parent/control strain background. **(B)** Data shown is theoretical. Total RNA and protein sourced from C and K strains will be comparatively assessed for abundance of the target RNA or protein by digital droplet PCR (ddPCR) and quantitative western blotting (qWestern), respectively.

For many sequence confirmed strains listed in Table 4.2, total protein (Table 4.3) and total RNA (Table 4.4) has been purified from cultures in triplicate. Efforts to visualize each tagged protein with the ideal signal demonstrate that 1-10 µg of total protein load is sufficient to detect the protein by western blotting (Table 4.3). Initial efforts have focused on quantifying protein levels encoded by gene TK0183. Strains TK0183C and TK0183K were grown to mid exponential phase at 85°C and total protein collected. In technical and biological replicate, 10 ug of total protein as well as 0-1000 ng of the HA-standard (TK2304-6xHis-HA) were analyzed by SDS-PAGE (Figure 4.4A). When the intensity of the HA standard protein is graphed by linear regression, the R^2 value is 0.97, but linearity is lost at higher loads (Figure 4.4B). Repeated

efforts have demonstrated that optimal linearity is achieved at protein loads up to 600 ng (R^2 = 0.99, Figure 4.4C). Quantification of the TK0183 C- and K-proteins across biological triplicates indicate that methodology should be improved to achieve consistency between replicates (Figure 4.4D). These data suggest that on average more TK0183 protein is present in the C-strain compared to the K-strain, indicating the loss of m^5C results in lower protein yield (Figure 4.4E). It is possible that the loss of m^5C results in reduced levels of mRNA, leading to lower protein levels. Future efforts will address whether RNA levels differ between the C- and K-strains.

Table 4.2 Series 1 K-strain constructs and progress.

Strains	code	Candidate Interm.[a]	Confirmed Interm.	interim. push final	Candidate Finals	WGS[b]
0183C	A1	160, 161, 162	160, 161, 162	160	180-183	183
0183K	B1	160, 161, 162	160, 161, 162	160	180-183	182
0186C	C1	160, 161, 162	160, 161, 162	160	180-183	183
0186K1	D1	160, 161, 162	160, 161, 162	160		
0186K2	E1	160, 161, 163	160, 161, 163	160	680-694	694
0186K1:2	F1	160, 161, 162	160, 161, 162	160	380-394	380
0494C	G1	160, 161, 162	160, 161, 162	160	180-183	180
0494K	H1	160, 161	160, 161	160	384386	384, 386
1046C	H2	260, 261, 262, 263	260, 261, 262, 263	260		
1046K1	A2	160, 161, 162	160, 161, 162	160	182	182
1046K2	B2	160, 161, 162	160, 161, 162	160	180, 280, 281	280
1046K1:2	C2	160, 161, 162	160, 161, 162	160	280, 281	280
1736C	F2	160, 161, 162	160, 161, 162	160	180-183	180
1736K	G2	160, 161, 162	160, 161, 162	160	280-283	283
1984C		☑	☑	☑	☑	☑
1984K	E2	160, 161, 163	160, 161, 163	160	280	280
0895C		160	160	160		

0895D	
0895K1	
0895K2	
0895K1:2	

[a] Interim. = Intermediate colonies where the plasmid has “pop’ed in” to the genome at the target loci are cultured.
[b] WGS. =Whole genome sequencing.

Table 4.3 Optical density of series 1 K-strain cultures at time of harvest and optimal total protein load

Strain	replicate	OD@75°C	OD@85°C	OD@95°C	recommended protein load
0183C	1	0.393	0.554	0.415	10 ug
	2	0.427	0.524	0.328	
	3	0.367	0.551	0.354	
0183K	1	0.327	0.373	0.359	10 ug
	2	0.311	0.355	0.189	
	3	0.306	0.426	0.262	
0186C	1	0.362	0.531	0.338	10 ug
	2	0.47	0.479	0.283	
	3	0.453	0.495	0.286	
0186K2	1	0.472	0.47	0.387	10 ug
	2	0.452	0.523	0.334	
	3	0.384	0.51	0.32	
0186K1:2	1	0.439	0.563	0.489	
	2	0.479	0.672	0.505	
	3	0.39	0.523	0.4	
0494C	1	0.414	0.557	0.476	
0494K	1	0.385	0.502	0.391	
1046K1	1	0.341	0.516	0.525	
1046K2	1	0.477	0.604	0.509	
1046K1:2	1	0.281	0.513	0.442	
1736C	1	0.157	0.428	0.258	1 ug
	2	0.17	0.466	0.309	
	3	0.181	0.435	0.25	
1736K	1	0.265	0.457	0.461	1 ug

	2	0.295		0.368	
	3	0.283	0.391	0.193	
1984C	1	0.253	0.441	0.622	1 ug
	2	0.233	0.566	0.519	
	3	0.266	0.428	0.505	
1984K	1		0.534	0.312	1 ug
	2	0.358	0.492	0.343	
	3	0.322		0.319	

Table 4.4 Series 1 K-strain RNA purification and ddPCR progress

Strain	RNA	DNA-free RNA	cDNA synthesis	2nd strand synthesis	ddPCR
0183C	x	x			
0183K	x	x			
0186C	x	x			
0186K1	x	x			
0186K2	x	x			
0186K1:2	x	x			
0494C	x	x			
0494K	x	x			
1046C					
1046K1	x	x			
1046K2	x	x			
1046K1:2	x	x			
1736C	x	x			
1736K	x	x			
1984C	x	x	x		
1984K	x	x	x		

Previously described were series 1 K-strains. In series 1 K-strains, the enzymes that install the selected m^5C sites are unknown, so all mutations are introduced into a parent background. Series 2 K strains include introducing a C-tag and C→ T mutation into the parent strain and a strain deleted for the R5CMT (Figure 4.5, Table 4.5). The C-tag is appended to the open reading frame where the resultant wildtype RNA is destined to be methylated with a known R5CMT. In C→T mutants (T-mutants), the RNA will not be methylated due to the loss of the

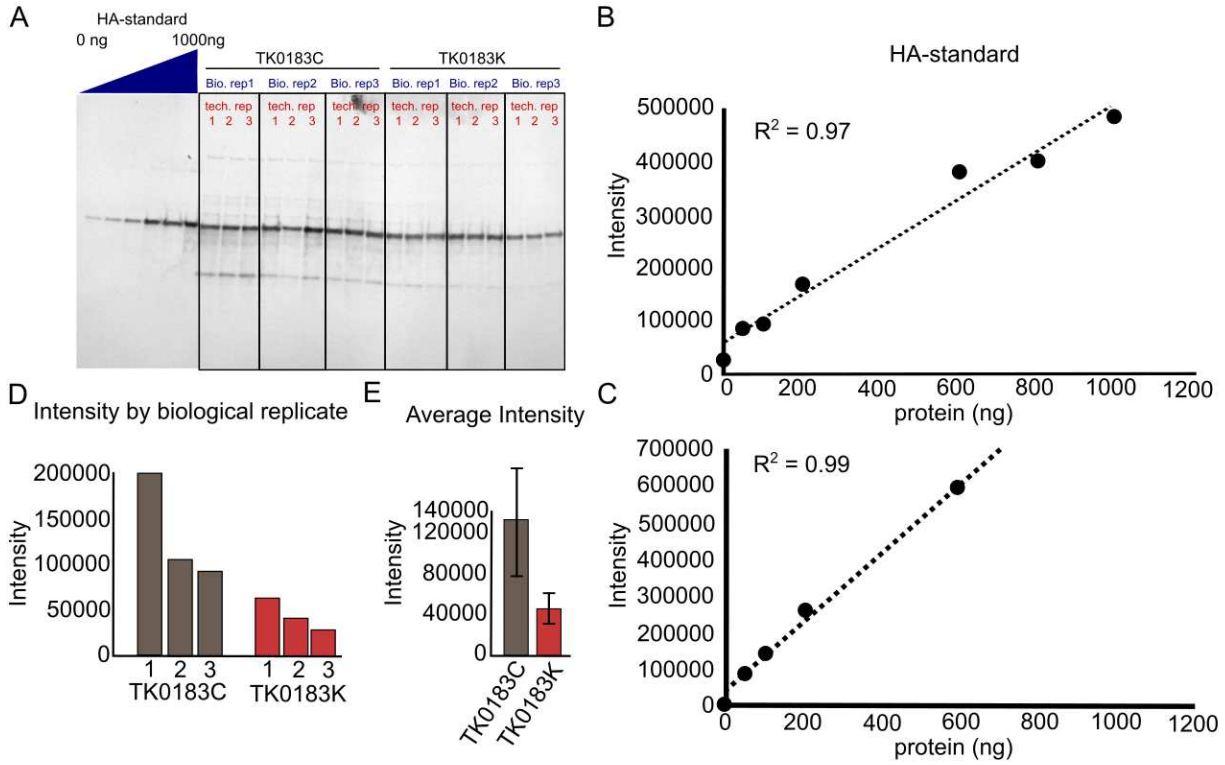


Figure 4.4 m^{5}C impacts translation outcomes. (A) Western blot against the HA epitope of total protein from TK0183 C- and K-strains in technical and biological triplicate. **(B, C)** Band intensity of the HA-standard at increasing protein load. **(D)** Quantification of protein intensity within each biological replicate and **(E)** the average intensity across replicates (± SD).

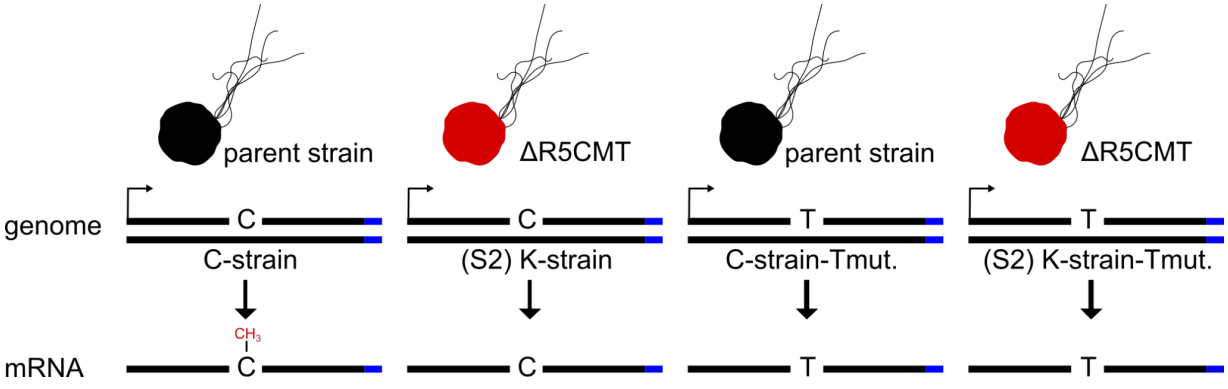


Figure 4.5 Second series C- and K-strain constructs and the resultant mRNA.

target cytidine. It is important to note that in strains deleted for R5CMTs, ribosomes and tRNAs will also have reduced levels of m^5C modification, and this is likely to have substantial impacts on translation outcomes regardless of the modification to the target mRNA. RNA and protein levels will be compared between all strains. A comparison between the C-tagged gene that is either a wildtype or T-mutant sequence in the R5CMT deletion strain will demonstrate whether the modification results in differential RNA or protein levels.

Table 4.5 Series 2 K-strain constructs and progress

Strain name	parent	genotype	Conf. Interm.	Conf. Finals	WGS
TK0136C	TS559	TK0136C			
ΔTK0872:TK0136C	ΔTK0872	ΔTK0872; TK0136C			
TK0136C-Tmut	TS559	TK0136C-Tmut			
ΔTK0872:TK0136C-Tmut	ΔTK0872	ΔTK0872; TK0136C-Tmut			
TK1034C	TS559	TK1034C			
ΔTK0872:TK1034C	ΔTK0872	ΔTK0872; TK1034C			
TK1034C-Tmut	TS559	TK1034C-Tmut			
ΔTK0872:TK1034C-Tmut	ΔTK0872	ΔTK0872; TK1034C-Tmut			
TK0112C	TS559	TK0112C			
ΔTK1935:TK0112C	ΔTK1935	ΔTK1935; TK0112C			
TK0112C-Tmut	TS559	TK0112C-Tmut			
ΔTK1935:TK0112C-Tmut	ΔTK1935	ΔTK1935; TK0112C-Tmut			
TK1152C	TS559	TK1152C			
ΔTK1935:TK1152C	ΔTK1935	ΔTK1935; TK1152C			
TK1152C-Tmut	TS559	TK1152C-Tmut			
ΔTK1935:TK1152C-Tmut	ΔTK1935	ΔTK1935; TK1152C-Tmut			
TK1911C	TS559	TK1911C			
ΔTK2304:TK1911C	ΔTK2304	ΔTK2304; TK1911C			
TK1911C-Tmut	TS559	TK1911C-Tmut			
ΔTK2304:TK1911C-Tmut	ΔTK2304	ΔTK2304; TK1911C-Tmut			
TK1259C	TS559	TK1259C			
ΔTK2122:TK1259C	ΔTK2122	ΔTK2122; TK1259C			
TK1259C-Tmut	TS559	TK1259C-Tmut			
ΔTK2122:TK1259C-Tmut	ΔTK2122	ΔTK2122; TK1259C-Tmut			

TK1170D	TS559	TK1170D
ΔTK2122:TK1170D	ΔTK2122	ΔTK2122; TK1170D
TK1170D-Tmut	TS559	TK1170D-Tmut
ΔTK2122:TK1170D-Tmut	ΔTK2122	ΔTK2122; TK1170D-Tmut

Discussion

Several reports have demonstrated that deletion of genes encoding R5CMT leads to sensitivity to heat stress in rice and worms.[65,63] Molecular studies have shown that m^5C increases the stability of mRNAs through interactions with reader proteins that recognize and bind to m^5C.[10,11,12,13] m^5C also has inherent properties that result in increased hydrophobicity and melting temperature of CG hydrogen bonds[4], and therefore m^5C likely strengthens secondary structures and overall stability. High temperatures are a constant insult to *T. kodakarensis*, and it is likely that the dense m^5C epitranscriptome provides much needed stability to RNAs.

mRNAs have been increasingly identified as modification targets. Site-specific, and even sub-stoichiometric, modification of mRNA is linked to core biological functions, responses to environmental changes, and human disease.[1,9] How m^5C residues impact the functions of mRNA at the transcript level has been historically challenging to address due to low abundance RNA and substoichiometric modification frequencies in conventional model organisms. It is unclear how low frequency mRNAs modifications can be biologically relevant. Critical questions remain regarding how the epitranscriptome impacts cellular fitness and how modifications impact mRNA function.

I demonstrated that the m^5C epitranscriptome is extensive in *T. kodakarensis* and identified a collection of candidate m^5C sites ideal for detailed mechanistic studies. Efforts to introduce silent mutations at the m^5C site and produce RNAs that lack m^5C have largely been successful. Current efforts are focused on preparing RNA and protein from these strains for analysis of their abundance. If m^5C impacts RNA stability, we will likely see reduced levels of the

RNA that now lacks m^5C, and subsequently reduced translation of that mRNA. It is possible that m^5C directly impacts translatability, and we may see differential protein level but unchanged RNA abundance. These strains will facilitate investigations beyond what is proposed here, into the molecular consequences of individual m^5C residues on the fate of RNA once modified.

Materials and Methods

Strain construction

Procedures for generating strains are published in Gehring *et al.* (14) and Appendix 1 of this dissertation. For each gene targeted for modification or deletion, a plasmid containing the modification with an additional ~700 nt of up and downstream homology to the genome was purchased from Twist Bioscience. Each plasmid was transformed into *T. kodakarensis* strain TS559 and cells were plated on agmatine-free, rich medium. After 2-5 days of growth at 85°C, transformants were picked into rich liquid media lacking agmatine and grown overnight. Intermediate candidate strains are given a numerical identifier between 160-177, 260-277, etc. For each isolate, genomic DNA was extracted from 1 ml of fully grown culture by phenol/chloroform/isoamyl alcohol (25/24/1; v/v/v) followed by alcohol precipitation in an equal volume of isopropanol. Plasmid integration into the genome was confirmed by PCR using two primer pairs, with one primer from each pair having homology to the genome and the other primer to the plasmid. This ensures that the amplicon originates from a genomically integrated sequence.

Isolates where PCR indicated that the plasmid had integrated at the predicted loci were plated on minimal medium with the counter selectable marker, 6-methyl purine, and grown anaerobically for 2-5 days at 85°C. Colonies were then picked into rich liquid media and grown overnight. Candidate final strains were given the numerical identifier, 180-199, 280-299, etc. Genomic DNA was purified from each strain and screened for the deletion or C-tag at the target

loci. Candidate strains were identified *via* PCR. To confirm the presence of the C-tag, we used a set of primers that flank ~150 bp of the C-tag. Candidate strains where PCR indicated a ~40 nt increase in amplicon length were or will be confirmed *via* whole genome sequencing. For gene deletions, we used primer pairs that flanked the deletion loci (external primers) and resulted in a reduced amplicon size equivalent to the size of the deleted region. Primer sets both with homology to a region internal to the deletion loci (internal primers) resulted in PCR amplification in the parent strain but not the deletion strain. Two combinations of external and internal primers, which should not produce an amplicon in the deletion strain, were used to ensure the deletion loci was absent in the deleted strain.

To ensure the region targeted for modification or deletion was correctly modified and to ensure no off-target genome modifications are present, each candidate strain was/will be screened by whole genome sequencing on a Minion platform. For strains that have been sequenced, each strain was grown anaerobically overnight in 5 ml of rich medium with agmatine at 85°C. Cultures were pelleted and genomic DNA purified using the Monarch Genomic DNA Purification kit (NEB, cat# T3010S). Library preparation for each strain was completed using the Rapid Barcoding kit 96 (ONT, cat# SQK-RBK110.96). Sequencing was done on the MinIon Mk1C with the R9.4.1 flow cell. Fast5 files were converted to Fastq files using Guppy default settings. Fastq files were aligned to the TS559 reference genome using MiniMap2 and alignment files were compared to the reference genome using Medaka Variant Calling Pipeline via Neural Networks. Visual inspection of the modified loci on Integrated Genomics Viewer confirmed the deletion or correct insertion of the C-tag and/or C→T mutation.

Cell growth

All *T. kodakarensis* strains were grown anaerobically at 65, 75, 85, or 95°C in 10-15 mL artificial sea water with yeast, tryptone, pyruvate, sulfur, and agmatine.[15] All cultures were grown

to mid exponential growth phase (OD_{600} = ~0.4 - 0.5) in triplicate before harvest by centrifugation. Cell pellets were stored at -20°C. Total RNA or total protein were then collected from each cell pellet.

RNA collection

Cell pellets were resuspended in 1 mL TRIzol reagent and incubated at room temperature for 5 minutes before the addition of 200 µL of chloroform. Samples were mixed vigorously by scraping on a microcentrifuge block and incubated at room temperature for 5 minutes before centrifugation at 15K rpm for 15 minutes. The aqueous phase was taken into 2.7X the volume of Ethanol and 2 uL glycoblue. Samples were mixed vigorously by shaking and incubated at -20°C for at least 30 minutes before centrifugation at 15K rpm for 30 minutes. RNA pellets were aspirated, washed with 100% Ethanol, dried for no more than 5 minutes, and resuspended in DNase solution (1X DNaseI buffer and 3 µL DNaseI in a 20 µL volume), and incubated at 37°C for 1.5 hours. RNA was purified using the monarch RNA clean and concentrator kit (NEB, cat# T2030S or T2030L), quantified using the Qubit Broad Range RNA kit, and concentrated to 1 µg/µL.

Protein collection

Cell pellets were resuspended in 150 uL buffer A2 (1 M NaCl, 25 mM TrisHCl pH 7.5, 10% glycerol) and 2% SDS. Cell suspensions were incubated at room temperature for 5 minutes and sonicated ~10 seconds on/off for 5 minutes. Samples were quantified using the Qubit protein kit and concentrations normalized to 1 mg/mL. Samples were stored at -20°C.

Western blot

Protein from TK0183 C-strain and K-strain pairs were analyzed using a 4-20% criterion stain free SDS-PAGE (BioRad, cat# 5678094) in technical and biological triplicate. The amount of total protein needed to visualize the HA-tagged protein is listed in Table 4.3 (recommended protein load). The Kaleidoscope marker (BioRad, cat# 1610375EDU) and 0-1000 ng of the HA

standard (TK2304-6xHis-HA) were also loaded on the same SDS-PAGE and run at 170V for 42 minutes. Gels were imaged and transferred to a Trans-blot membrane (BioRad, cat# 1704157) using a Trans-blot Turbo Transfer system. All incubations were performed at room temperature with rocking. Membranes were soaked in 5% BSA for 1 hour and the BSA removed. The membrane was then incubated in the 10mL primary antibody solution (mouse anti-HA, diluted 1:10,000 in 1X TBST) then washed in 1X TBST for 10 minutes 3 times to remove the primary antibody. 10 mL of the secondary antibody solution (goat ant-imouse, 1:1000 dilution in 1X TBST) was applied to the membrane for 2 hours. The membrane was briefly washed with 1X TBST and then washed with Buffer W2 (1X Maleic Acid, 0.3% Tween) for 15 minutes. The membrane was then soaked in detection buffer (1 M TrisHCl pH 9.5, 1 M NaCl) for 5 minutes. The detection buffer was poured off and ~5 mL of NBT/BCIP (ThermoScientific, cat# 34042) was applied to the membrane. The membrane was incubated in the dark for ~15 minutes, then briefly washed with DI water, air dried overnight, and the blot quantified with ImageQuant software.

REFERENCES

1. Wilkinson, E., Cui, Y.-H. & He, Y.-Y. Roles of RNA Modifications in Diverse Cellular Functions. *Front Cell Dev Biol* **10,** 828683 (2022).
2. Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., Kurkowska, M., Shirvanizadeh, N., Destefanis, E., Groza, P., Avşar, G., Romitelli, A., Pir, P., Dassi, E., Conticello, S. G., Aguilo, F. & Bujnicki, J. M. MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Res.* **50,** D231–D235 (2021).
3. Jonkhout, N., Tran, J., Smith, M. A., Schonrock, N., Mattick, J. S. & Novoa, E. M. The RNA modification landscape in human disease. *RNA* **23,** 1754–1769 (2017).
4. Ontiveros, R. J., Stoute, J. & Liu, K. F. The chemical diversity of RNA modifications. *Biochem. J* **476,** 1227–1245 (2019).
5. Edelheit, S., Schwartz, S., Mumbach, M. R., Wurtzel, O. & Sorek, R. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. *PLoS Genet.* **9,** e1003602 (2013).
6. 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., Washburn, M. P., Isobe, T., Santangelo, T. J., Shalev-Benami, M., Meier, J. L. & Schwartz, S. Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. *Nature* **583,** 638–643 (2020).
7. Trixl, L. & Lusser, A. The dynamic RNA modification 5-methylcytosine and its emerging role as an epitranscriptomic mark. *Wiley Interdiscip. Rev. RNA* **10,** e1510 (2019).
8. Lui, L. M., Uzilov, A. V., Bernick, D. L., Corredor, A., Lowe, T. M. & Dennis, P. P. Methylation guide RNA evolution in archaea: structure, function and genomic organization of 110 C/D

box sRNA families across six Pyrobaculum species. *Nucleic Acids Res.* **46,** 5678–5691 (2018).

9. Höfer, K. & Jäschke, A. Epitranscriptomics: RNA Modifications in Bacteria and Archaea. *Microbiol Spectr* **6,** (2018).
10. Blanco, S., Bandiera, R., Popis, M., Hussain, S., Lombard, P., Aleksic, J., Sajini, A., Tanna, H., Cortés-Garrido, R., Gkatza, N., Dietmann, S. & Frye, M. Stem cell function and stress response are controlled by protein synthesis. *Nature* **534,** 335–340 (2016).
11. Yang, Y., Wang, L., Han, X., Yang, W.-L., Zhang, M., Ma, H.-L., Sun, B.-F., Li, A., Xia, J., Chen, J., Heng, J., Wu, B., Chen, Y.-S., Xu, J.-W., Yang, X., Yao, H., Sun, J., Lyu, C., Wang, H.-L., Huang, Y., Sun, Y.-P., Zhao, Y.-L., Meng, A., Ma, J., Liu, F. & Yang, Y.-G. RNA 5-Methylcytosine Facilitates the Maternal-to-Zygotic Transition by Preventing Maternal mRNA Decay. *Mol. Cell* **75,** 1188–1202.e11 (2019).
12. Chen, X., Li, A., Sun, B.-F., Yang, Y., Han, Y.-N., Yuan, X., Chen, R.-X., Wei, W.-S., Liu, Y., Gao, C.-C., Chen, Y.-S., Zhang, M., Ma, X.-D., Liu, Z.-W., Luo, J.-H., Lyu, C., Wang, H.-L., Ma, J., Zhao, Y.-L., Zhou, F.-J., Huang, Y., Xie, D. & Yang, Y.-G. 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. *Nat. Cell Biol.* **21,** 978–990 (2019).
13. Schumann, U., Zhang, H.-N., Sibbritt, T., Pan, A., Horvath, A., Gross, S., Clark, S. J., Yang, L. & Preiss, T. Multiple links between 5-methylcytosine content of mRNA and translation. *BMC Biol.* **18,** 40 (2020).
14. Gehring, A. M., Sanders, T. J. & Santangelo, T. J. Markerless Gene Editing in the Hyperthermophilic Archaeon Thermococcus kodakarensis. *Bio Protoc* **7,** (2017).
15. Scott, K. A., Williams, S. A. & Santangelo, T. J. Thermococcus kodakarensis provides a versatile hyperthermophilic archaeal platform for protein expression. *Methods Enzymol.* **659,** 243–273 (2021).

APPENDIX 1: THERMOCOCCUS KODAKARENSIS PROVIDES A VERSATILE HYPERTHERMOPHILIC ARCHAEAL PLATFORM FOR PROTEIN EXPRESSION[4]

Summary

Hyperthermophiles, typically defined as organisms with growth optima ≥ 80˚C, are dominated by the Archaea. Proteins that support life at the extremes of temperatures often retain substantial biotechnological and commercial value, but the recombinant expression of individual hyperthermophilic proteins is commonly complicated in non-native mesophilic hosts due to differences in codon bias, intracellular solutes and the requirement for accessory factors that aid in folding or deposition of metal centers within archaeal proteins. The development of versatile protein expression and facilitated protein purification systems in the model, genetically-tractable, hyperthermophilic marine archaeon *Thermococcus kodakarensis* provides an attractive platform for protein expression within the hyperthermophiles. The assortment of *T. kodakarensis* genetic backgrounds and compatible selection markers allow iterative genetic manipulations that facilitate protein overexpression and expedite protein purifications. Expression vectors that stably replicate both in *T. kodakarensis* and *Escherichia coli* have been validated and permit high-level ectopic gene expression from a variety of controlled and constitutive promoters. Biologically relevant protein associations can be maintained during protein purifications to identify native protein partnerships and define protein interaction networks. *T. kodakarensis* thus provides a versatile platform for the expression and purification of thermostable proteins.

[4] This chapter was published in July 2021 under the same title.

Scott KA, Williams SA, Santangelo TJ. *Thermococcus kodakarensis* provides a versatile hyperthermophilic archaeal platform for protein expression. Methods Enzymol. (2021).

Introduction

Well-regulated and robust bacterial- and fungal-based recombinant gene expression systems are commonly employed to overproduce and, ultimately, purify proteins in large quantities for biochemical and structural characterization.[1,2] In choosing a recombinant expression system, parameters such as cost, yield and ease of purification are considered. Additional parameters, including the assembly of metal-cluster binding sites, proper post-translational modifications, introduction of disulfide-linkages, protein secretion, post-translation N- and C-terminal processing, intein excision, the availability of specific chaperones, unique genetic codes, insertion into membranes of specific compositions and the availability of essential cofactors limit the choice of production host or demand overexpression in the native host.[3,4] Protein expression systems have been widely established and commercialized in several mesophilic species, including *Escherichia coli*, yeasts and insect cells (with baculoviruses). While these systems often permit production of cross-species, and occasionally cross-Domain recombinant protein expression, there remain instances wherein the canonical protein expression platforms fail to retain the necessary parameters for optimal protein expression. An increasing class of such proteins originate in the Archaea, wherein proteins often meant to survive the extremes of pressures, temperature, salinity and pH have unique expression and folding requirements that are not easily replicated outside of the archaeal Domain.

The dearth of genetic systems for most archaeal lineages limits controlled protein overexpression in most archaeal clades, but a few systems have been established that permit the controlled overexpression of proteins in archaeal cell lysates or whole cells.[5,6] Archaeal-derived *in vitro* translation systems, often-linked to *in vitro* transcription systems, provide a route to produce archaeal proteins in minute quantities but are impractical for moderate- or large-scale protein production. Many efforts have thus focused on the few

genetically-tractable clades of Archaea to develop techniques to aid in protein overexpression and to facilitate rapid and efficient purification of the desired protein products.[7] Controlled and inducible expression systems have long been established in halophilic (salt-loving) archaeal species,[8,9] but the near-saturating intracellular salinities often impair protein folding of non-halophilic proteins and impede downstream protein processing in almost all cases. Several methanogenic archaeal lineages have been manipulated to provide controlled and inducible protein expression platforms,[10] but none have been developed at scale, as the growth rates and complicated growth conditions of many methanogens limit the scale of culture growth. The remaining archaeal cell-based expression systems are found in species that grow at high (> 60˚C, thermophilic) or very high (> 80˚C, hyperthermophilic) temperatures. The thermophilic protein-expression systems established for the *Sulfolobales*[11] and the hyperthermophilic protein expression systems established for the *Thermococcales*[5,12–15] provide unique platforms for expression of thermostable proteins with a broad range of applications.

Hyperthermophilic archaea have significant potential as host systems to invigorate new biotechnological innovations and conversion chemistries.[16–20] Protein expression in hyperthermophilic archaeal systems motivates both basic and industrial efforts for greener chemistries, particularly towards the production of enzymes that survive the high temperatures common to many manufacturing processes. Microbial contamination is reduced or effectively eliminated by carrying out enzymatic conversions at high temperatures, and the increased half-life of thermotolerant or thermostable enzymes often reduces total enzyme requirements and costs.[21,22] Thermostable enzymes, particularly amylases and xylanases, are employed in bulk in biomass degradation and processing[22] as well as in pulp and fiber productions[23] and starch liquefaction processes[24]. Thermostable proteases are in high-demand for bulk addition to laundry detergents, for leather processing, as rennins for cheese-making, as meat-tenderizers, and for baking and brewing platforms. Thermostable lipases are employed in high-temperature, low-water-content solvent systems for hydrolysis activities, transesterification and ester

synthesis. Thermostable enzymes are more diversely used in smaller scale molecular studies, including PCR, next-generation sequencing techniques and manipulations of nucleic acids. Finally, many thermostable proteins are stably folded and display reduced long-range intramolecular dynamics at ambient and/or mesophilic conditions, properties that often aid or inform structural or biophysical studies.

Thermococcus kodakarensis has emerged as a common model species for the study of biological processes at high temperatures and as an expression platform for thermostable enzymes.[25] *T. kodakarensis* is an ecologically pervasive, fast-growing, anaerobic, heterotroph first isolated from a solfatara on Kodakara island, Japan[26,27] that can be grown to high cell densities with economical media requirements. The development of highly reproducible, accurate genomic and ectopic manipulation strategies for *T. kodakarensis* permits rational and iterative strain construction,[28–30] controlled protein overexpression,[5,6] and the introduction of protein tags or epitopes[31–33] that aid in rapid and cost-effective protein purifications. *T. kodakarensis* thus provides a readily available and versatile platform for the expression of thermostable proteins for functional studies[7] and a relatively untapped source of thermostable enzymes. Here, we outline current methodologies for expression and purification of native and recombinant proteins which have been developed for *T. kodakarensis*. These procedures take advantage of the natural competency of *T. kodakarensis* cells, reliable culturing techniques and commercialized approaches for affinity chromatography. We first introduce the general procedures that permit genetic manipulation of *T. kodakarensis*, then detail genomic-, then plasmid-based protein expression and purification systems.

2. Genetic systems for Thermococcus kodakarensis

The *Thermococcales* family is distributed globally and represents a well-studied branch within the *Euryarchaeota* that includes a growing list of species divided among three genera: the *Pyrococcus, Paleococcus* and *Thermococcus*.[34] *T. kodakarensis* was initially classified as

Pyrococcus kodakaraensis strain KOD1 but was reclassified as *T. kodakarensis* strain KOD1 after genome re-sequencing and phylogenetic analysis of the 16S ribosomal RNA (note that the original species designation was also changed from *kodakaraensis* to *kodakarensis*).[35] *T. kodakarensis* is naturally competent and, through homologous recombination, readily integrates exogenous DNA into the genome.[28,36] The natural competency of *T. kodakarensis* contributes to its utility as a platform for protein expression. Transformation procedures are not arduous and clonal populations can be selected and maintained on a variety of media using a range of selectable markers.[7] Specialized strains (Table A1.1) permit selections based on the restoration of prototrophy[28–30,37] and resistance to toxic nucleotide analogs[28,30] (Table A1.2). Complementary counter-selective pressures permit unlimited reuse of selective markers and iterative strain constructions, inclusive of both genomic modifications and plasmid retention.[28,38] While *T. kodakarensis* is not known to harbor any naturally-occuring plasmids, cryptic mini plasmids from related members of the *Thermococcales* have been modified to stably replicate in *T. kodakarensis*.[5,15] The development of genetic techniques has led to the emergence of *T. kodakarensis* as a versatile host for protein expression.

2.1 Genetic backgrounds and selectable markers

The choice of selection strategy and host genotype dictate initial plans for protein expression and purification within *T. kodakarensis*. When using *T. kodakarensis* strains with WT or near WT genotypes, only two validated selection markers exist that do not require specialized genotypic backgrounds. The first is based on hydroxymethylglutaryl coenzyme A (HMG-CoA) reductase activity that is essential for archaeal isoprenoid-based lipid biosynthesis.[39] HMG-CoA is inhibited with simvastatin or mevinolin (a simvastatin analog)[40] and overexpression of the *P. furiosus* HMG-CoA reductase *via* a constitutive promoter provides *T. kodakarensis* strains resistance to ~8 μM simvastatin[40] and >30 μM mevinolin[41]. The second universal selective pressure is based on amino acid auxotrophies. Unlike many other *Thermococcus* species, *T.*

kodakarensis lacks an arginine biosynthetic pathway and requires arginine supplementation for growth.[27,37] The *P. furiosus* genes PF0207 and PF0208 encode argininosuccinate synthase and argininosuccinate lyase, respectively, which together catalyze the synthesis of arginine from citrulline or aspartate.[7,42] *T. kodakarensis* strains that incorporate and express donor DNA carrying both PF0207 and PF0208 can grow with citrulline or aspartate in the absence of arginine in minimal medium.[42] These selection markers are compatible with all known strains; however, counter selection strategies independent of genotypic background have not yet been developed.

It is far more common to direct protein expression in one of several strains (Table A1.1) that have been constructed to allow auxotrophic–prototrophic selection of genomic integration of donor DNA by homologous recombination or retention of autonomously replicating plasmids.[5,15,43] Auxotrophic strains that incorporate donor DNA restoring prototrophy provide positive selection strategies. Counter selection strategies lead to the excision or loss of donor DNA, as the activity of counter selectable markers result in death of all cells still containing the counterselectable marker. KU216 lacks the orotidine-5′-phosphate decarboxylase (*pyrF)* gene and displays uracil auxotrophy in minimal medium.[29] KU216 also confers resistance to 5-fluoroorotic acid (5-FOA), allowing counter selection for cells that have excised *pyrF* encoded donor DNA. Building from KU216, stains KUW1 and KUWH1 additionally lack anthranilate synthase (TK0254), encoded by the *trpE* locus, conferring tryptophan auxotrophy.[29] KUWH1 further lacks *hisD,* providing histidine auxotrophy. Strain KW128 lacks *pyrF* and the *trpE* locus is replaced with *pyrF*, leading to tryptophan auxotrophy.[29] The absence of the amino acid corresponding to the induced auxotrophy provides a positive selection strategy based on the restoration of prototrophies provided in the donor DNA.

Multiple strains derived from KU216 were constructed to allow for sequential selection and counter selection of genomically integrated DNA. Such strains are useful for genetic manipulations *via* markerless allelic exchange (MAE) and are the backbone of genetic

techniques in *T. kodakarensis*. Strain TS517 is genotypically similar to KW128 but with an additional deletion of hypoxanthine/guanine phosphoribosyltransferase (TK0664), a purine scavenging protein.[30] Strain TS559 additionally lacks the pyruvoyl-dependent arginine decarboxylase (TK0149) which leads to agmatine auxotrophy. Integration of donor DNA encoding TK0149 results in agmatine prototrophy. Similar to TS517, ejection of donor DNA carrying TK0664 is selected for by isolating mutants resistant to 6MP.[43] Lastly, strain TS900 was derived from TS559 and is cleanly deleted for *pyrF* (unpublished). Selection and counter selection of transformants in TS900 are identical to that of TS559. The strains described here and their compatible selection strategies allow targeted mutagenesis of chromosomal DNA.

2.2 Targeted mutagenesis of chromosomal DNA

T. kodakaraensis is naturally competent for DNA uptake and incorporates donor DNA into its genome by homologous recombination,[28,29] allowing targeted chromosomal mutations for increased protein expression,[6] placement of extracellular secretion signals,[6] gene deletions,[44–46] epitope- or affinity-purification,[31,32] or allelic substitutions.[47] Genome manipulations are currently possible using MAE where transformation with plasmid DNAs that carry a positive selection marker (*i.e.*: TK0254 (*trpE*), TK2276 (*pyrF*) or TK0149 (*pdaD*)) are integrated into the genome resulting in tryptophan, uracil and agmatine prototrophic transformants, respectively (Figure A1.1).[29,30,43] After selection and confirmation of the intermediate genotype, growth of confirmed intermediate-genome containing cells in the presence of tryptophan, uracil, or agmatine permits a small percentage of cells to spontaneously remove the integrated plasmid sequences *via* homologous recombination.[28] These unique cells containing the desired final genotypes are now able to survive in the presence of 5-FOA or 6MP due to the loss of the counterselectable marker.[28,41] MAE, sometimes referred to as pop-in/pop-out, permits a tailored approach to expression and purification of proteins in a hyperthermophilic system.

Targeted chromosomal alterations are achieved by transforming a vector containing the desired modification flanked by several hundred bases (*i.e.*: > 500 bp) of up- and downstream

Table A1.1 *T. kodakarensis* strains and compatible selection strategies.

Strain	Genotype	Medium requirements	Available selection strategies	Reference(s)
KOD1	wild type	rich or minimal medium	HMG-CoA reductase and citrulline.	Morikawa et al. 1994, Fukui et al. 2005
KU216	Δ*pyrF*	rich or minimal medium with uracil	HMG-CoA reductase and citrulline. Transformation with DNA encoding *pyrF* confers uracil prototrophy.	Sato et al. 2005
KUW1	Δ*pyrF;* Δ*trpE*	minimal medium with uracil and tryptophan	HMG-CoA reductase and citrulline. Transformation with DNA encoding *pyrF* or *trpE* confers uracil or tryptophan prototrophy, respectively.	Sato et al. 2005
KUWH1	Δ*pyrF;* Δ*trpE;* Δ*hisD*	minimal medium with uracil, tryptophan and histidine	HMG-CoA reductase and citrulline. Transformation with DNA encoding *pyrF*, *trpE*, or *hisD* confers uracil, tryptophan, or histidine prototrophy, respectively.	Sato et al. 2005
KW128	Δ*pyrF;* Δ*trpE::pyrF*	rich medium or minimal medium with tryptophan	HMG-CoA reductase and citrulline. Compatible with MAE using tryptophan (selection) and 5-FOA (counter-selection).	Sato et al. 2005
TS517	Δ*pyrF;* Δ*trpE::pyrF;* Δ*HPRT*	rich medium or minimal medium with tryptophan	HMG-CoA reductase and citrulline. Compatible with MAE using tryptophan (selection) and 6MP (counter-selection).	Santangelo et al. 2010
TS559	Δ*pyrF;* Δ*trpE::pyrF;* Δ*pdaD;* Δ*HPRT*	rich medium with agmatine or minimal medium with tryptophan and agmatine	HMG-CoA reductase and citrulline. Compatible with MAE using agmatine (selection) and 6MP (counter-selection).	Santangelo et al. 2010
TS900	Δ*pyrF;* Δ*trpE;* Δ*pdaD;* Δ*HPRT*	rich medium with agmatine or minimal medium with tryptophan and agmatine	HMG-CoA reductase and citrulline. Compatible with MAE using (i) agmatine (selection) and 6MP (counter-selection); (ii) uracil (selection) and 5-FOA (counter-selection).	unpublished

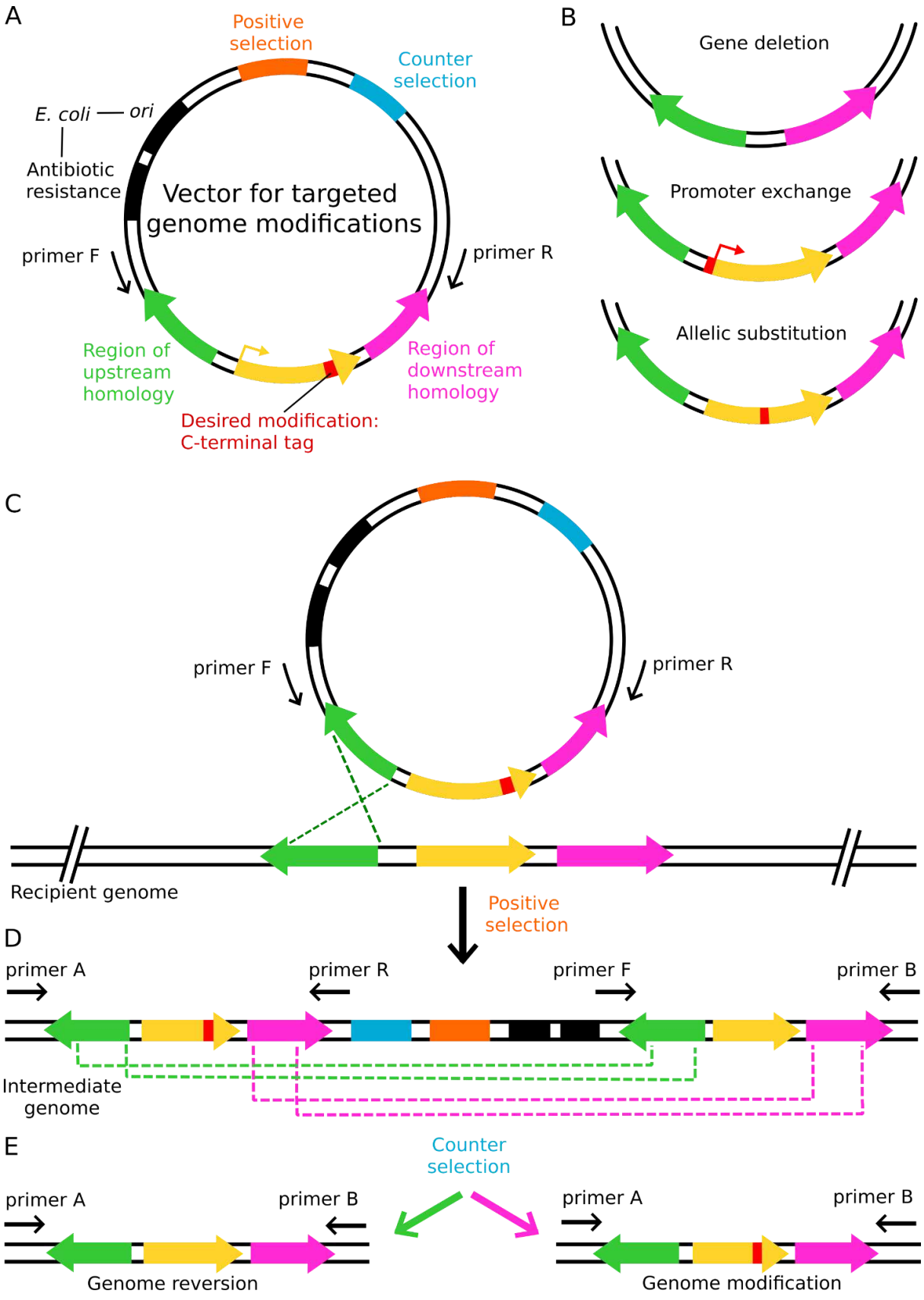

A
Positive selection
Counter selection
E. coli — ori
Antibiotic resistance
Vector for targeted genome modifications
primer F
primer R
Region of upstream homology
Region of downstream homology
Desired modification: C-terminal tag
B
Gene deletion
Promoter exchange
Allelic substitution
C
primer F
primer R
Recipient genome
Positive selection
D
primer A
primer R
primer F
primer B
Intermediate genome
E
Counter selection
primer A
primer B
Genome reversion
primer A
primer B
Genome modification

Figure A1.1 Sequential homologous recombination events permit modifications of chromosomal DNA. A vector designed for genomic integration **(A)** includes replication and selection elements for *E. coli* (black) but only selection (orange) and counter selection (blue) elements for *T. kodakarensis*. Regions of up (green) and downstream (pink) homology to the target locus sequences (yellow, with red modification) permit the vector to integrate adjacent to the target locus. Possible genomic modifications including **(B)** gene deletion(s), promoter exchange and allelic substitutions can be generated using identical pop-in/pop-out techniques. Genomic integration is mediated by either the region of upstream or downstream homology **(C)**, and the first recombination event (pop-in) is selected for using a positive selection marker **(D).** In this example, the region with upstream homology (green) is the recombination point. The intermediate genome is verified using primers pairs that have homology to the genome and the plasmid separately (primers A and F and primers R and B). The second recombination event, wherein the plasmid is ejected from the genome (pop-out), is selected for using a counter selectable marker **(E)**; there is an equal probability of recombination at the green or pink loci. In this example, recombination between the green loci yields the parent genome while recombination between the pink loci results in the desired modification. Depending on the modification, the final genome can be verified using PCR and amplicon or whole genome sequencing.of the final genotype can be achieved through PCR analysis and/or Sanger sequencing of the target loci.[38] Whole genome sequencing should be considered as off target recombination events are possible. While modification of promoter or expression sequences (*i.e.*: Shine-Delgarno sequences) may alter protein expression, many changes to chromosomal coding sequences are specifically designed to retain native protein expression. Ectopic shuttle vectors which stably replicate in *T. kodakarensis* and *E. coli* are often more practical for high level protein expression.

homology (Figure A1.1A).[43] Routine modifications including N- or C-terminal epitope- or affinity-tag additions, promoter exchange, gene deletions and allelic substitutions (Figure A1.1A, B) are encoded on a vector that has both a selectable and counter selectable marker compatible with MAE in *T. kodakarensis* (Figure A1.1A and Table A1.2). Components for autonomous replication and selection in *E. coli* are included; however, no *T. kodakarensis* origin of replication (*ori*) is necessary, as the plasmid is expected to integrate into the genome. After vector transformation, either the up- or downstream regions homologous to the target loci will integrate into the genome (Figure A1.1C). Prototrophic selection of plasmid integration (pop-in) *via* homologous recombination is mediated by the positive selectable marker (Table A1.2). The intermediate genome is confirmed *via* PCR using two primer sets (primers A/R and primers F/B, Figure A1.1D). Primers A and B are external to the site of plasmid integration and are not

homologous to any region on the integrative vector, while primers F and R have homology to the plasmid. Spontaneous recombination of the plasmid out of the genome (pop-out) is selected for using a counter selectable marker (Figure A1.1E and Table A1.2) and has a ~50% chance of reverting to the parent genome (Figure A1.1E, green) or including the modification (Figure A1.1E, pink). Confirmation of the final genotype can be achieved through PCR analysis and/or Sanger sequencing of the target loci.[38] Whole genome sequencing should be considered as off target recombination events are possible. While modification of promoter or expression sequences (*i.e.*: Shine-Delgarno sequences) may alter protein expression, many changes to chromosomal coding sequences are specifically designed to retain native protein expression. Ectopic shuttle vectors which stably replicate in *T. kodakarensis* and *E. coli* are often more practical for high level protein expression.

2.3 Autonomously replicating plasmids

Plasmids that autonomously replicate and express selectable markers in both *T. kodakaraensis* and *E. coli* can be easily manipulated and confirmed in *E. coli*, then transferred to *T. kodakaraensis* where manipulation of plasmids is more complicated. Shuttle vectors, when modified to encode expression cassettes under the control of a variety of constitutive promoters (Table A1.3 and Figure A1.2, purple), provide a mechanism for high levels of ectopic expression. The cell surface glycoprotein (*csg*, TK0895) is a highly expressed gene in *T. kodakarensis*, and therefore the P*csg* promoter, among others, can direct high level expression when inserted into the multiple cloning site (Figure A1.2, green).[48] Where an exogenous promoter is prefered, the P*hmtb* promoter from *Methanothermobacter thermautotrophicus* is compatible with *T. kodakarensis* expression machinery.[41] In the related species, *Pyrococcus furiosus,* some level of inducible expression can be achieved from the P*cipA* promoter.[49] With the exception of a riboswitch-mediated expression system, no single-gene controlled expression system currently exists in *T. kodakarensis*.[14] The endogenous P*fbpase* promoter can be activated by the addition of elemental sulfur (S) but like the P*cipA* promoter, such large scale

Table A1.2 Selectable markers for *T. kodakarensis.*

Selection	Genetic background	Selection requirements	Compatible strains	Reference(s)
tryptophan	*ΔtrpE* (ΔTK0254)	Minimal medium lacking tryptophan.	KUW1, KUWH1, KW128	Sato et al. 2003, Sato et al. 2005
simvastatin, mevinolin	None	None. Requires overexpression of HMG-CoA reductase (PF1848).	Any strain	Matsumi et al. 2007, Santangelo et al. 2008
agmatine	Δ*pdaD* (ΔTK0149)	Minimal or rich medium lacking agmatine supplementation.	TS559, TS900	Fukuda et al. 2008; Santangelo et al. 2008
histidine	Δ*hisD* (ΔTK0244)	Minimal medium lacking histidine.	KUWH1	Sato et al. 2005
6-methyl purine (6MP)	Δ*HPRT* (ΔTK0664)	Minimal medium containing 6MP.	TS517, TS559, TS900	Santangelo et al. 2010
uracil, 5-fluororatic acid (5-FOA)	*ΔpyrF* (ΔTK2276)	Minimal medium lacking uracil and supplemented with 5-FOA, respectively.	KU216, KW128, TS517, TS559, TS900	Sato et al. 2003
citrulline, aspartate	+*argG* (PF0207), +*argH* (PF0208)	Minimal medium lacking citrulline or aspartate.	Any strain	Atomi et al. 2004

changes to growth conditions alter the expression of many genes. Select changes to the ribosome binding site (RBS, 5′-AGGTGA) can act as another point to control protein expression levels.[5] No naturally occurring plasmids exist in *T. kodakarensis*.[50] All *T. kodakarensis–E. coli* shuttle vectors (Table A.4) rely on one of two plasmids isolated from close relatives in the *Thermococcales* for replication (Figure A1.2, red). The cryptic mini plasmid pTN1 isolated from *T. nautilus* encodes a rolling-circle replication initiator protein (Rep74) and a p24 protein, although p24 is not essential.[51] The fusion of pTN1 to *E. coli* vectors provide the minimal replicative sequences for both species and selectable markers for *E. coli*, but not *T. kodakarensis*.[5] Analogous to pTN1, pTP2 from *Thermococcus prieurii* contains five putative open reading frames, one of which encodes a rolling circle replication gene that also supports

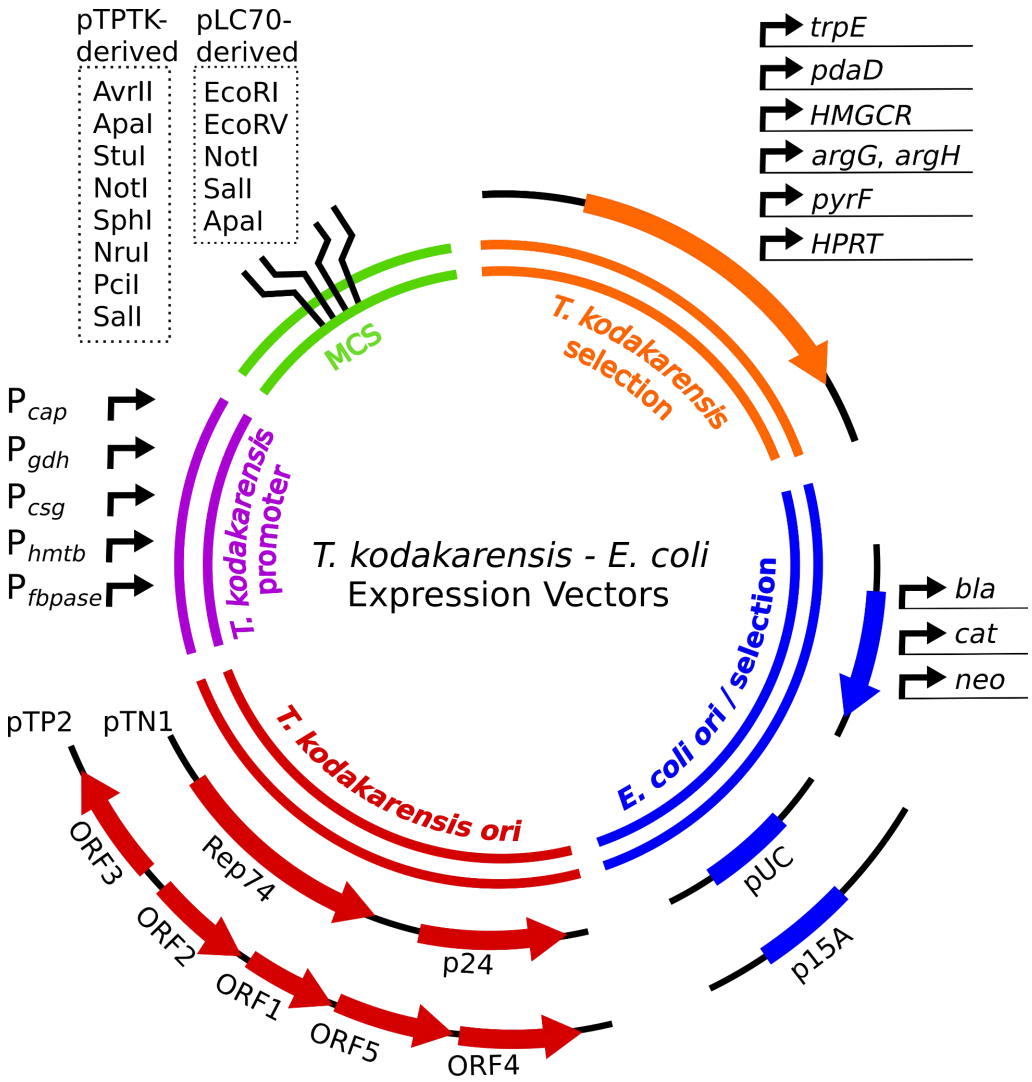


Figure A1.2 Autonomously replicating expression vectors permit stable replication and gene expression in *T. kodakarensis* and *E. coli*. Constitutive or inducible promoters (purple) drive high levels of ectopic gene expression of sequences cloned into the multiple cloning site (MCS, green). Selections in *T. kodakarensis* (orange) rely on restorations of prototrophies (agmatine, tryptophan, arginine, uracil), statin resistance (mevinolin, simvastatin) or resistance to toxic nucleoside analogues (6MP, 5-FOA). Expression vectors retain either a pUC or p15A origin (*ori*) and an antibiotic selection cassette(s) for replication in *E. coli* (blue): *bla* (ampicillin resistance); *cat* (chloramphenicol resistance); *neo* (kanamycin resistance). The entire pTN1 or pTP2 plasmid sequences (red) permit autonomous replication in *T. kodakarensis*.

Table A1.3 Promoters compatible with gene expression in *T. kodakarensis.*

Promoter	Gene ID	Origin	Description	Reference(s)
P*hmtb*	MT0254	*Methanothermobacter thermautotrophicus*	Constitutive; Histone B.	Santangelo et al. 2007
P*cap*	TK2279	*Thermococcus kodakarensis*	Constitutive; CDP-alcohol phosphatidyltransferase.	Rodrigues et al. 2007
P*gdh*	TK1431	*Thermococcus kodakarensis*	Constitutive; Glutamate dehydrogenase.	Santangelo et al. 2008
P*csg*	TK0895	*Thermococcus kodakarensis*	Constitutive; Cell surface glycoprotein.	Takemasa et al. 2011
P*fbpase*	TK2164	*Thermococcus kodakarensis*	Inducible; Fructose-1,6-bisphosphatase. Involvement in central metabolism hinders tight regulation.	Sato et al. 2004
P*cipA*	PF0190	*Pyrococcus furiosus*	Inducible; membrane-bound glycoprotein, cold-induced protein A. Induction *via* heat/cold shock or other stress has significant biological effects.	Basen et al. 2012

Table A1.4: *T. kodakarensis* - *E. coli* shuttle vectors.

Plasmid	Size (bp)	*T. kodakarensis* copy number	*E. coli* copy number	*T. kodakarensis* selection(s)	*E. coli* selection(s)	Description	Genetic background	Ref.
pLC64	9034	~3/chromosome	~500-700/cell	tryptophan	ampicillin, kanamycin	pTN1 + pCR2.1-TOPO::P*cap*-TrpE	Δ*trpE* (ΔTK0254)	Santangelo et al. 2008
pLC70	10495	~3/chromosome	~500-700/cell	statins, tryptophan	ampicillin, kanamycin	pTN1 + pCR2.1-TOPO::P*cap*-TrpE; P*gdh*-HMG-CoA	Δ*trpE* (ΔTK0254)	Santangelo et al. 2008
pLC71	9190	~3/chromosome	~500-700/cell	statins, tryptophan	ampicillin, kanamycin	pTN1::Δp24 + pCR2.1-TOPO::P*cap*-TrpE; P*gdh*-HMG-CoA	Δ*trpE* (ΔTK0254)	Santangelo et al. 2008
pTS543	8196	~3/chromosome	~500-700/cell	agmatine	ampicillin, kanamycin	pTN1 + pCR2.1-TOPO::P*hmtB*-PdaD	Δ*pdaD* (ΔTK0149)	Santangelo et al. 2010
pTPTK1	5489	~50/cell	~20-30/cell	statins	chloramphenicol	pTP2 + pBAD33::P*gdh*-HMG-CoA	Any strain	Catchpole et al. 2018
pTPTK2	5444	~50/cell	~20-30/cell	tryptophan	chloramphenicol	pTP2 + pBAD33::P*cap*-TrpE	Δ*trpE* (ΔTK0254)	Catchpole et al. 2018
pTPTK3	4710	~50/cell	~20-30/cell	agmatine	chloramphenicol	pTP2 + pBAD33::P*pdad*-PdaD	Δ*pdaD* (ΔTK0149)	Catchpole et al. 2018

plasmid replication in *T. kodakarensis.*[15] Both pTN1- and pTP2-based plasmids can be maintained within the same cell, providing two distinct and simultaneous means of selection. *T. kodakarensis* is polyploid, retaining ~7-19.[52] pTN1-based plasmids have been measured to exist at a copy number of ~3 copies per chromosome[5] while pTP2-based plasmids have been observed at ~50 copies per cell[53] (Table A.4). Stable replication of plasmids in *E. coli* is dependent on a pUC or p15A origin (Figure A.2, blue), with a copy number of ~500-700 and ~20-30 per cell, respectively (Table A.4).[5,15,54,55]

Naturally occurring replicative plasmids (*i.e.*: pTP2 and pTN1; Figure A.2, red) have been modified to contain cassettes for expression of selection markers in both *T. kodakaraensis* and *E. coli* (Table A1.4). Ampicillin, chloramphenicol and kanamycin resistance genes (*bla*, *cat* and *neo*, respectively; Figure A.2, blue) have previously been used for selection in *E. coli*, while a variety of prototrophic and statin-resistance strategies have been validated in *T. kodakarensis* (Figure A.2, orange).[5,15] The pLC plasmid series (*e.g.*: pLC64, pLC70, pLC71 and pTS543) can provide statin resistance or restore tryptophan or agmatine protoprophy to transformants.[5,30,56] The pLC64 plasmid was constructed by ligating together pTN1 and the commercially available pCR2.1-TOPO. pLC64 contains the trpE gene under control of the constitutive promoter, P*cap*.[57] Plasmids pLC70 and then pLC71 were derived by adding HMG-CoA from *P. furiosus* under the *T. kodakarensis* P*gdh* promoter, providing an additional mechanism for statin resistance. Several restriction enzyme sites are available for digestion at the MCS (Figure A.2, green).

The pTPTK plasmid series (*e.g.*: pTPTK1, pTPTK2, pTPTK3) are a result of the ligation of pTP2 and pBAD33 and have been modified to enable statin-resistance or restore tryptophan or agmatine prototrophy under a native or constitutive promoter (Figure A.2, purple).[15] Restriction enzyme recognition sites are present in the multiple cloning site; however, *Stu*I and *Sal*I digestion should be avoided in pTPTK2 as their recognition sequences are present in the *trpE* cassette. With the exception of agmatine prototrophy, all other prototrophic selection strategies require cells to be grown in minimal medium. Use of all prototrophic selective markers

requires a mutant genetic background that induces auxotrophy for the selection nutrient.[28–30,41]

3. Purifying proteins and identifying protein interactions

Recombinant protein expression in a genetically accessible, native hyperthermophile such as *T. kodakarensis* provides access to protein production and research applications in a novel setting. Hyperthermophilic protein expression can produce thermostable proteins in their native state, with biologically relevant associations and modifications (*e.g.*: in complex and/or containing post-translational modifications). Protein purifications *via* liquid chromatography (LC) can be applied (1) to purify proteins at industrial scale (typically from ectopic vectors) or (2) to describe native-level protein expression or identify protein–protein partnerships at research scale.

3.1 Protein production and purification

Ectopic expression vectors permit easy introduction and expression of heterologous genes and are commonly used for applications involving large-scale protein production. *T. kodakarensis* cultures grow rapidly at 85˚C, doubling every ~40 minutes and will achieve a final cell density of ~10^{8-9} cells/mL in rich media. Optical density (OD) at 600 nm provides a convenient method to monitor cell growth with cultures reaching ~1.0-1.3 absorbance units at maximum cell densities. *T. kodakarensis* strains harboring the expression vector(s) may be cultured at a range of temperatures (~65-95˚C) and in media containing the necessary nutritional and selectable components using batch fermentation to yield large quantities of biomass. Growth conditions (*e.g.*: temperature) and media additives should be selected for the specific strain (Table A1.1) and the compatible genetic markers (Table A1.2). Industrial-scale protein purifications typically require greater culture volumes, and pilot trials at small scale are recommended for process development. Larger culture volumes (>10 L) are ideally grown within a bioreactor. Protein production in minimal medium is possible, although culture yields are dramatically reduced resulting in minimal biomass. Cultures that have reached late exponential

or early stationary growth phases are rapidly cooled and harvested, and the biomass is processed for protein purification. Cell pellets can be stored at -80˚C indefinitely prior to processing.

Cell lysis to separate cytoplasmic from membrane-bound or extracellular products is achieved by sonication in physiologically relevant buffers followed by centrifugation to pellet membranes, while retaining cytoplasmic proteins in the supernatant. Purification of the desired protein can be obtained by liquid chromatography, and conditions specific to the protein of interest should be considered (*e.g.*: a protein bound strongly to DNA may require elution at higher salinity or unique pHs). Reverse phase, ion exchange, size exclusion or affinity chromatography can be employed for a variety of purification techniques. Purification efficacy can be assessed by SDS-PAGE and western blotting, if respective antibodies are available. The addition of epitope tags, such as the hemagglutinin epitope (HA) to the protein of interest can be utilized for immunodetection if protein-specific antibodies are not available. Tags which include a cleavage site can be used for structure specific applications (*e.g.*: crystallography).

3.2 Native protein-expression and retention of in vivo protein partnerships

Purification of tagged-protein complexes and retention of native protein interactions established *in vivo* permits identification of associated protein partnerships. Chromosomal MAE techniques can be used to introduce sequences to the genome that encode affinity- (His_6) and epitope- (HA) tags to protein targets and result in retention of native promoters and native levels of target proteins within *T. kodakarensis*. Steady-state protein levels retain protein partnerships that can be identified through the rapid, gentle, co-purification of the tagged-target protein in association with its natural protein partners.

To maintain the greatest number of protein interactions, cultures should be chilled rapidly in an ice-water bath to preserve as many interactions in their native conformational state as possible. Subjecting cultures to dry ice-mediated cooling is an alternative when exceptionally delicate interactions (*e.g.*: protein–RNA associations) are of interest. Pelleted and resuspended

cells can be sonicated briefly to disturb membranes while maintaining protein–protein interactions. The downstream chromatography must be completed as rapidly as possible to preserve the maximum number of protein partnerships. Purified fractions can be assessed for the presence of the tagged protein by SDS-PAGE and western blotting. Fractions containing the protein of interest are typically pooled, quantified for total protein and trypsin digested for peptide analysis on liquid chromatography tandem mass spectrometry (LC-MS/MS). The high-confidence peptide identification and low-input requirements offered by LC- MS/MS provides a simple and economical route to establish protein interactions in complex protein mixtures.[31,32]

In contrast to large-scale protein production strategies (section 3.1), a single liter or less of *T. kodakarensis* culture grown in rich medium to mid-exponential phase is typically sufficient to identify protein-protein interaction networks. Meaningful interpretation of proteomics data requires reproducibility and biochemical validation. Repeated identification of target and co-purifying proteins, with a minimum of two unique peptides per protein, in multiple biological replicates (three are recommended) provides a high level of confidence in *bona fide in vivo* partnerships.[31] Reciprocal pulldowns or known protein associations (*e.g.*: established protein complexes) can be used to confirm interactions *in vivo*. Given that a minor number of *T. kodakarensis* proteins have a natural affinity for some purification matrices, care should be taken to ensure that co-purifying proteins are not present in identical purifications from near isogenic control strains (*e.g.*: the parental strain grown and processed under the same conditions).

4. Reagents, recipes and equipment

A list of reagents, recipes and equipment are provided for use in the protocols outlined in section 5.

4.1 Reagents

1. S; elemental sulfur; flowers of sulfur
2. Agmatine
3. Gelzan (Sigma-Aldrich)
4. 6-methylpurine (Sigma-Aldrich)
5. Cysteine
6. Glutamic acid
7. Glycine
8. Arginine
9. Proline
10. Asparagine
11. Histidine
12. Isoleucine
13. Leucine
14. Lysine
15. Threonine
16. Tyrosine
17. Alanine
18. Methionine
19. Phenylalanine
20. Serine
21. Tryptophan
22. Aspartic acid
23. Glutamine
24. Valine
25. $NiSO_4$
26. EDTA

27. Guanidine HCl

4.2 Recipes

1. Trace minerals (1,000x; 1 L)
    a. 0.5 g $MnSO_4 \cdot H_2O$
    b. 0.1 g $CoCl_2 \cdot 6\ H_2O$
    c. 0.1 g $ZnSO_4 \cdot 7\ H_2O$
    d. 0.01 g $CuSO_4 \cdot 5\ H_2O$
    e. 0.01 g $AlK(SO_4)_2 \cdot 12\ H_2O$
    f. 0.01 g H_3BO_3
    g. 0.01 g $Na_2MoO_4 \cdot 2\ H_2O$
2. Artificial sea water containing yeast extract and tryptone (1.0x ASW-YT; 1 L)
    a. 5 g tryptone
    b. 5 g yeast extract
    c. 20 g NaCl
    d. 3 g $MgCl_2 \cdot 6\ H_2O$
    e. 6 g $MgSO_4 \cdot 7\ H_2O$
    f. 800 mg $(NH_4)_2SO_4$
    g. 160 mg $NaHCO_3$
    h. 240 mg $CaCl_2 \cdot 2\ H_2O$
    i. 400 mg KCl
    j. 336 mg KH_2PO_4
    k. 40 mg NaBr
    l. 16 mg $SrCl_2 \cdot 6\ H_2O$
    m. 8 mg $Fe(NH_4)_2(SO_4)_2 \cdot 6\ H_2O$
    n. 5 mL trace minerals

o. 5 g pyruvate (Optional)

3. Polysulfides (500x; 15 mL)
    a. 10 g $Na_2S•9\ H_2O$
    b. 3 g sulfur

        ** See Note 1.*

4. KOD1 Vitamins (200x; 1 L)
    a. 0.2 g niacin
    b. 0.08 g biotin
    c. 0.2 g pantothenate
    d. 0.2 g lipoic acid
    e. 0.08 g folic acid
    f. 0.2 g *p*-aminobenzoic acid
    g. 0.2 g thiamine
    h. 0.2 g riboflavin
    i. 0.2 g pyridoxine
    j. 0.2 g cobalamin
5. 20 amino acid mix (20x; 1L)
    a. 1 g cysteine
    b. 1 g glutamic acid
    c. 1 g glycine
    d. 500 mg arginine
    e. 500 mg proline
    f. 400 mg asparagine
    g. 400 mg histidine
    h. 400 mg isoleucine
    i. 400 mg leucine

j. 400 mg lysine

k. 400 mg threonine

l. 400 mg tyrosine

m. 300 mg alanine

n. 300 mg methionine

o. 300 mg phenylalanine

p. 300 mg serine

q. 300 mg tryptophan

r. 200 mg aspartic acid

s. 200 mg glutamine

t. 200 mg valine

** See Note 2.*

6. Solid-media plates (prepared for each transformation)

    a. 50 mL 2x media formulation (*e.g.*: ASW-YT), autoclave

    b. 50 mL water with 1 g Gelzan, autoclave

    c. Following separate autoclaving, immediately mix components *a* and *b*, 100 µL KOD1 vitamins, 200 µL polysulfides and any additional necessary components (*i.e.*: strain TS559 transformants rely on the absence of agmatine for positive selection)

    d. Mix anaerobically and pour ~25 mL into each of 4 glass petri dishes

7. Lysis buffer (Buffer A)

    10 mM Tris-HCl, pH 8

    0.5 M NaCl

    10% (v/v) glycerol

8. Elution buffer (Buffer B)

    10 mM Tris-HCl, pH 8

0.1 M NaCl

0.5 M imidazole

10% (v/v) glycerol

4.3 Equipment

1. Microcentrifuge and microcentrifuge tubes
2. Floor model centrifuge, compatible rotors and centrifuge tubes
3. Heating block
4. Anaerobic chambers (Coy Laboratory Products)
5. 85˚C incubator
6. Anaerobic microbiology culturing equipment
    a. Glass serum bottles
    b. Septa
    c. Aluminum crimp seals
    d. Glass petri plates

        **See Note 3.*
7. Sonicator
8. AKTA or similar liquid chromatography system
9. HiTrap chelating column (Cytiva)
10. SDS-PAGE equipment
    a. Gels
    b. Running buffer
    c. Apparatus
    d. Compatible protein stains
11. Western blotting equipment
    a. Membrane

b. Transfer system

c. Primary antibody (Invitrogen HA Tag Monoclonal Antibody, 26183)

d. Secondary antibody

e. Imaging system

5. Protocols

5.1 Targeted mutagenesis of chromosomal DNA

Markerless-modifications of *T. kodakarensis* chromosomal DNA typically rely on integration of a non-replicative plasmid harboring the desired modification into the genome, followed by excision of most plasmid sequences from the genome to yield a strain with a modified genomic sequence. Detailed instructions for the construction of integrative plasmids have been provided elsewhere[43]; herein we focus on the steps for transformation and confirmation of genomic modifications. All procedures are carried out anaerobically. For simplicity, we do not detail the specific selective and counter-selective procedures nor the exact media formulations of solid or liquid mediums, as the multitude of genetic techniques for *T. kodakarensis* provide many options depending on the host strain employed (Tables 1 and 2).

Transformation and selection for plasmid integration into the genome

1. Grow an MAE-compatible *T. kodakarensis* strain (Table A1.1) at 85˚C for 12 hours in ASW-YT medium supplemented with S, KOD1 vitamins and any necessary additives. Harvest *via* centrifugation and resuspend cells in 3 mL of 0.8X ASW, transfer to a microcentrifuge tube, and place on ice.
2. Add ~2 µg of plasmid DNA (in < 15 µL) to a 200 µL aliquot of concentrated cell suspension and incubate on ice for 30 minutes.
3. Heat shock the cell and plasmid mixture at 85˚C for 45 seconds using a heating block, then immediately incubate on ice for an additional 30 minutes.
4. Plate transformed cells on a selection medium and grow anaerobically at 85˚C for ~2-5

days.

Confirmation of appropriate plasmid integration and intermediate genome formation

5. Colonies resulting from transformations are anticipated to have integrated the plasmid sequences into the genome. Cells from distinct colonies should be picked and inoculated in 5 mL ASW-based medium containing KOD1 vitamins while ensuring maintenance of the positive selection(s) employed.
6. 5 mL cultures are grown overnight anaerobically at 85˚C to yield sufficient biomass for 1 mL genomic DNA preparations following established procedures.[43]
7. Perform diagnostic PCRs using genomic DNA.

    ** See Note 4.*

Counter selection, confirmation of plasmid excision and final genome confirmation

8. 5 mL ASW-based medium grown cultures of confirmed intermediate strains, supplemented to permit spontaneous excision of the selective and/or counterselective makers, should be grown overnight and the resulting cells spread on solid-medium containing the appropriate counter-selective pressures and allowed ~4 days of anaerobic growth at 85˚C.
9. Resulting colonies are anticipated to have spontaneously excised the plasmid from their genomes and either retained the desired modification or reverted back to the parental genome (Figure A.1E). Cells from distinct colonies should be picked, inoculated in 5 mL ASW-based medium and grown anaerobically overnight at 85˚C to yield sufficient biomass for 1 mL genomic DNA preparations following established procedures.[43]
10. Amplicons generated from primers pairs complementary to genomic sequences permit diagnostic confirmation of the desired final genotype. Depending on the desired modification, PCR followed by gel electrophoresis can be used for size discrimination of the target region. If the modification produces a size difference that is difficult to discern *via* PCR (*i.e.*: short sequences encoding affinity- or epitope-tags), differential restriction

digestion patterns of the PCR amplicon originating from the region of interest may assist in identifying desired genotypes. When diagnostic PCRs indicate the presence of the desired modification, genotypes should be confirmed *via* Sanger sequencing of PCR amplicon(s) that span the target loci or whole genome sequencing (*e.g.*: Oxford Nanopore Sequencing).

5.2 Transformation of T. kodakarensis–E. coli shuttle vectors

** See Note 5.*

1. Grow an MAE-compatible *T. kodakarensis* strain (Table A1.1) at 85˚C for 12 hours in ASW-YT medium supplemented with S, KOD1 vitamins and any necessary additives. Harvest *via* centrifugation and resuspend cells in 3 mL of 0.8X ASW, transfer to a microcentrifuge tube, and place on ice.
2. Add ~2 µg of plasmid DNA (in < 15 µL) to a 200 µL aliquot of concentrated cell suspension and incubate on ice for 30 minutes.
3. Heat shock the cell and plasmid mixture at 85˚C for 45 seconds using a heating block, then immediately incubate on ice for an additional 30 minutes.
4. Spread transformed cells on a solid-medium compatible with the selection marker(s) (Table A1.2). Incubate the plated cells anaerobically for ~2-6 days at 85˚C.
5. Pick cells from a discrete colony into 10 mL of the appropriate ASW-based medium required for the chosen selectable marker(s). Allow the culture to grow anaerobically at 85˚C overnight.

   **See Note 6.*

5.3 Ectopic protein expression

1. Inoculate 1 L of ASW-based medium supplemented with S, KOD1 vitamins and any necessary additives with a 10 mL overnight culture.

   **See Note 7.*

2. For expression reliant on constitutive promoters, cells should be harvested *via*

centrifugation immediately prior to entry into the stationary phase to maximize cell density. For expression reliant on inducible promoters, culture conditions should be altered to permit protein expression when the optical density has reached OD_{600} ≈ 0.2-0.3, and cells should be harvested *via* centrifugation prior to stationary phase.

**See Note 8.*

5.4 Large-scale single protein purification

** See Note 9.*

Cell lysis

1. Cell pellets are thawed (if frozen) and resuspended with ~3 mL per g of biomass in lysis buffer (Buffer A).
2. Sonicate the cell suspension on ice at half-maximal output for 10 seconds on/30 seconds off for a total of ~5-60 minutes. Suspension will be grey and homogenous when cells are sufficiently lysed. Avoid unnecessary sonication.
3. Generate a clarified cell lysate (CCL) by centrifugation (9,000 x *g* for 20 minutes) to pellet membranes while retaining cytoplasmic proteins in the supernatant.

**See Note 10.*

Purification via affinity chromatography

1. Prepare the AKTA chromatography unit (or similar) by ensuring all tubing is flushed and a fraction collector is prepared.
2. Charge HiTrap chelating column with 1 M $NiSO_4$.
3. Clean flow paths and HiTrap column with 5 column volumes of Buffer B followed by 5 column volumes of Buffer A.
4. Load CCL.

**See Note 11.*

5. Wash the column with 5 column volumes of Buffer A.

**See Note 12.*

6. Elute the target protein with an imidazole gradient and collect the eluant in small (~1 mL) fractions. Aliquots of each fraction can be resolved *via* SDS-PAGE and/or western blotting to identify fractions containing the target protein.

   **See Note 13.*

7. Flush the entire chromatography system with 5 column volumes Buffer B, followed by 10 column volumes of ultra-pure H_2O. To regenerate the chelating column, first remove any remaining proteins with 5 column volumes of 6 M guanidine HCl, then flush the column with H_2O, and finally strip the bound Ni^{2+} from the column with 2 column volumes 0.5 M EDTA. Store the chromatography apparatus and HiTrap column in 20% ethanol.

5.5 Rapid and gentle purification to identify protein partnerships

**See Note 14.*

Identification of in vivo protein partnerships

1. Chromatography fractions containing the tagged protein of interest are pooled, quantified, precipitated and trypsin-digested prior to the separation and identification of peptides *via* LC-MS/MS.
2. Peptides are identified using multidimensional protein identification technology (MuDPIT) and mapped to the *T. kodakarensis* proteome as described.[31,32]

6. Advantages and future perspectives

T. kodakarensis has emerged as a model species for the study of biological processes at high temperatures and as an expression platform for thermostable enzymes, owing mainly to the development of highly reproducible, accurate genomic and ectopic manipulation strategies.[25] While numerous advantages of a native hyperthermophilic expression system are evident, the *T. kodakarensis* platform could be improved.

Although genomic manipulation strategies in *T. kodakarensis* are simple and sufficiently straightforward to allow many researchers to generate large numbers of unique strains, genome

modification systems that use CRISPR/Cas are not yet available for *T. kodakarensis*. Most archaeal species (including the *Thermococcales*) encode CRISPR elements[35] and efforts to establish a thermostable CRISPR/Cas system for use in *T. kodakarensis* may accelerate strain construction. The development of additional cost-effective selection strategies, particularly non-prototrophic selections that could be universally deployed regardless of host genotype, would speed strain construction and add to the versatility of genetic techniques for *T. kodakarensis*. One of the most important tools for exploiting *T. kodakarensis* as an expression platform will be the development of a tightly-controlled, inducible promoter system. Available riboswitch-mediated regulatory expression systems are useful and are commonly employed within the *T. kodakarensis* research community[14,47] but lack the dynamic range desired for rapid and efficient control of protein expression.

Despite the challenges imposed by high temperature environments, the rapid development of tools for genetic manipulation for *T. kodakarensis* have provided the scientific community with a versatile protein expression platform conducive to the production and study of thermostable proteins. The combination of autonomously replicating ectopic expression systems, controlled protein overexpression and the ability to introduce genomic modifications makes *T. kodakarensis* a versatile platform for the expression of thermostable proteins and the evaluation of *in vivo* protein partnerships.

Notes

1. The powder is mixed with up to 15 mL H_2O and heated until the solution is homogeneous and free of particulates.
2. Individual amino acids can be excluded to allow prototrophic selections.
3. Many plasticware items melt upon extended incubation at hyperthermophilic temperatures.
4. Amplicons generated from primer pairs complementary to genomic and integrated plasmid sequences permit diagnostic confirmation of the desired intermediate

genotype(s). It is imperative that the primer sets used in the diagnostic PCRs include one primer that is complementary to genomic DNA sequences and a second primer complementary to sequences of the integrated plasmid (Figure A.1D). Such pairing ensures that resulting amplicons originate from a genomically integrated plasmid. Depending on how the plasmid integrated into the genome (either upstream or downstream to the modification site) two intermediate genomes are possible.[38] Whenever possible, identification of separate strains containing each potential intermediate genotype is desirable as each intermediate strain can yield the desired final genotype.

5. Methodologies for generating a *T. kodakarensis–E. coli* shuttle vector with selectable phenotypes and replicative origins are outlined elsewhere.[5,15]
6. Maintenance of a retained plasmid can be confirmed by growth on selective media, PCR, or by plasmid purification and transformation back into E. coli.
7. Cultures should be started with a 1:100 inoculum.
8. Secreted protein targets can be recovered from the culture media.
9. Both purification strategies (5.4 and 5.5) are reliant on genetically modified strains encoding affinity and epitope tags on the protein of interest. Six-histidines (His_6; affinity-tag) at the N- or C-terminus allow for protein purification using a chelating column charged with nickel. Bound proteins are eluted using an imidazole gradient. A hemagglutinin sequence (HA; epitope-tag) is encoded between the His_6 and protein coding sequence to allow for easy identification of the target protein via western blotting. The following procedures should be done at 4˚C.
10. The CCL can be stored at 4˚C but should not be frozen if protein structure or function is of interest.
11. The His_6 tagged protein will bind to the Ni^{2+}-charged column.
12. The tagged protein should not be in the wash.

13. Subsequent purifications to increase protein purity may be required. Size exclusion often complements affinity chromatography and is recommended.
14. Epitope and affinity (HA and His_6) tagged proteins can be rapidly purified through single-step chromatography to retain and identify in vivo protein partnerships. Cultures, clarified cell lysates and protein fractions are processed as described in section 5.4 with care given to reducing sonication time, using gentle centrifugation, and proceeding as quickly as possible through all steps.[31,32]

REFERENCES

1. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* **5**, 172 (2014).
2. Kim, H., Yoo, S. J. & Kang, H. A. Yeast synthetic biology for the production of recombinant therapeutic proteins. *FEMS Yeast Res.* **15**, 1–16 (2015).
3. Duong-Ly, K. C. & Gabelli, S. B. Explanatory chapter: troubleshooting recombinant protein expression: general. *Methods Enzymol.* **541**, 209–229 (2014).
4. Peleg, Y. & Unger, T. Resolving bottlenecks for recombinant protein expression in *E. coli*. *Methods Mol. Biol.* **800**, 173–186 (2012).
5. Santangelo, T. J., Cubonová, L. 'ubomíra & Reeve, J. N. Shuttle vector expression in *Thermococcus kodakaraensis*: contributions of cis elements to protein synthesis in a hyperthermophilic archaeon. *Appl. Environ. Microbiol.* **74**, 3099–3104 (2008).
6. Takemasa, R., Yokooji, Y., Yamatsu, A., Atomi, H. & Imanaka, T. *Thermococcus kodakarensis* as a host for gene expression and protein secretion. *Appl. Environ. Microbiol.* **77**, 2392–2398 (2011).
7. Farkas, J. A., Picking, J. W. & Santangelo, T. J. Genetic techniques for the archaea. *Annu. Rev. Genet.* **47**, 539–561 (2013).
8. Haque, R. U., Paradisi, F. & Allers, T. *Haloferax volcanii* for biotechnology applications: challenges, current state and perspectives. *Appl. Microbiol. Biotechnol.* **104**, 1371–1382 (2020).
9. Martínez-Espinosa, R. M. Heterologous and homologous expression of proteins from haloarchaea: denitrification as case of study. *Int. J. Mol. Sci.* **21**, (2019).
10. Guss, A. M., Rother, M., Zhang, J. K., Kulkarni, G. & Metcalf, W. W. New methods for tightly regulated gene expression and highly efficient chromosomal integration of cloned genes for

Methanosarcina species. *Archaea* **2**, 193–203 (2008).

11. Schocke, L., Bräsen, C. & Siebers, B. Thermoacidophilic *Sulfolobus* species as source for extremozymes and as novel archaeal platform organisms. *Curr. Opin. Biotechnol.* **59**, 71–77 (2019).
12. Farkas, J., Chung, D., DeBarry, M., Adams, M. W. W. & Westpheling, J. Defining components of the chromosomal origin of replication of the hyperthermophilic archaeon *Pyrococcus furiosus* needed for construction of a stable replicating shuttle vector. *Appl. Environ. Microbiol.* **77**, 6343–6349 (2011).
13. Walker, J. E. & Santangelo, T. J. Analyses of in vivo interactions between transcription factors and the archaeal RNA polymerase. *Methods* **86**, 73–79 (2015).
14. Speed, M. C., Burkhart, B. W., Picking, J. W. & Santangelo, T. J. An archaeal fluoride-responsive riboswitch provides an inducible expression system for hyperthermophiles. *Appl. Environ. Microbiol.* **84**, (2018).
15. 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).
16. Adams, M. W. & Kelly, R. M. Finding and using hyperthermophilic enzymes. *Trends Biotechnol.* **16**, 329–332 (1998).
17. Crosby, J. R. *et al.* Extreme thermophiles as emerging metabolic engineering platforms. *Curr. Opin. Biotechnol.* **59**, 55–64 (2019).
18. Straub, C. T. *et al.* Biotechnology of extremely thermophilic archaea. *FEMS Microbiol. Rev.* **42**, 543–578 (2018).
19. Counts, J. A. *et al.* Physiological, metabolic and biotechnological features of extremely thermophilic microorganisms. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **9**, (2017).
20. Zeldes, B. M. *et al.* Extremely thermophilic microorganisms as metabolic engineering platforms for production of fuels and industrial chemicals. *Front. Microbiol.* **6**, 1209 (2015).

21. Suresh, A. *et al.* Recent advancements in the synthesis of novel thermostable biocatalysts and their applications in commercially important chemoenzymatic conversion processes. *Bioresour. Technol.* **323**, 124558 (2021).
22. Kumar, S., Dangi, A. K., Shukla, P., Baishya, D. & Khare, S. K. Thermozymes: Adaptive strategies and tools for their biotechnological applications. *Bioresour. Technol.* **278**, 372–382 (2019).
23. Bhardwaj, N., Kumar, B. & Verma, P. A detailed overview of xylanases: an emerging biomolecule for current and future prospective. *Bioresources and Bioprocessing* **6**, 1–36 (2019).
24. de Souza, P. M. & de Oliveira Magalhães, P. Application of microbial α-amylase in industry - A review. *Braz. J. Microbiol.* **41**, 850–861 (2010).
25. Atomi, H. & Reeve, J. Microbe Profile: *Thermococcus kodakarensis*: the model hyperthermophilic archaeon. *Microbiology* **165**, 1166–1168 (2019).
26. Morikawa, M., Izawa, Y., Rashid, N., Hoaki, T. & Imanaka, T. Purification and characterization of a thermostable thiol protease from a newly isolated hyperthermophilic *Pyrococcus sp*. *Appl. Environ. Microbiol.* **60**, 4559–4566 (1994).
27. 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**, 263–267 (2004).
28. Sato, T., Fukui, T., Atomi, H. & Imanaka, T. Targeted gene disruption by homologous recombination in the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1. *J. Bacteriol.* **185**, 210–220 (2003).
29. Sato, T., Fukui, T., Atomi, H. & Imanaka, T. Improved and versatile transformation system allowing multiple genetic manipulations of the hyperthermophilic archaeon *Thermococcus kodakaraensis*. *Appl. Environ. Microbiol.* **71**, 3889–3899 (2005).
30. Santangelo, T. J., Cubonová, L. 'ubomíra & Reeve, J. N. *Thermococcus kodakarensis*

genetics: TK1827-encoded beta-glycosidase, new positive-selection protocol, and targeted and repetitive deletion technology. *Appl. Environ. Microbiol.* **76**, 1044–1052 (2010).

31. Li, Z., Santangelo, T. J., Cuboňová, L. 'ubomíra, Reeve, J. N. & Kelman, Z. Affinity purification of an archaeal DNA replication protein network. *MBio* **1**, (2010).
32. Burkhart, B. W., Febvre, H. P. & Santangelo, T. J. Distinct physiological roles of the three ferredoxins encoded in the hyperthermophilic archaeon *Thermococcus kodakarensis*. *MBio* **10**, (2019).
33. Li, Z. *et al.* A novel DNA nuclease is stimulated by association with the GINS complex. *Nucleic Acids Res.* **39**, 6114–6123 (2011).
34. Schut, G. J. *et al.* The order thermococcales and the family thermococcaceae. *The Prokaryotes: Other Major Lineages of Bacteria and the Archaea* 363–383 (2014).
35. Fukui, T. *et al.* Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. *Genome Res.* **15**, 352–363 (2005).
36. Čuboňováa, L. *et al.* An archaeal histone is required for transformation of *Thermococcus kodakarensis*. *J. Bacteriol.* **194**, 6864–6874 (2012).
37. Fukuda, W., Morimoto, N., Imanaka, T. & Fujiwara, S. Agmatine is essential for the cell growth of *Thermococcus kodakaraensis*. *FEMS Microbiol. Lett.* **287**, 113–120 (2008).
38. Hileman, T. H. & Santangelo, T. J. Genetics techniques for *Thermococcus kodakarensis*. *Front. Microbiol.* **3**, 195 (2012).
39. Cabrera, J. A., Bolds, J., Shields, P. E., Havel, C. M. & Watson, J. A. Isoprenoid synthesis in *Halobacterium halobium*. Modulation of 3-hydroxy-3-methylglutaryl coenzyme a concentration in response to mevalonate availability. *J. Biol. Chem.* **261**, 3578–3583 (1986).
40. 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. *J. Bacteriol.* **189**, 2683–2691 (2007).

41. Santangelo, T. J. *et al.* Polarity in archaeal operon transcription in *Thermococcus kodakaraensis*. *J. Bacteriol.* **190**, 2244–2248 (2008).
42. Santangelo, T. J. & Reeve, J. N. Genetic tools and manipulations of the hyperthermophilic heterotrophic archaeon *Thermococcus kodakarensis*. in *Extremophiles Handbook, ed. K. Horikoshi (Tokyo: Springer), 567–582* 567–582 (2010).
43. Gehring, A. M., Sanders, T. J. & Santangelo, T. J. Markerless gene editing in the hyperthermophilic archaeon *Thermococcus kodakarensis*. *Bio Protoc* **7**, (2017).
44. Gehring, A. M. *et al.* Genome replication in *Thermococcus kodakarensis* independent of Cdc6 and an origin of replication. *Front. Microbiol.* **8**, 2084 (2017).
45. Santangelo, T. J., Cuboňová, L. 'ubomíra & Reeve, J. N. Deletion of alternative pathways for reductant recycling in *Thermococcus kodakarensis* increases hydrogen production. *Mol. Microbiol.* **81**, 897–911 (2011).
46. Ishino, S. *et al.* Biochemical and genetical analyses of the three mcm genes from the hyperthermophilic archaeon, *Thermococcus kodakarensis*. *Genes Cells* **16**, 1176–1189 (2011).
47. Sanders, T. J. *et al.* FttA is a CPSF73 homologue that terminates transcription in Archaea. *Nat Microbiol* **5**, 545–553 (2020).
48. Kanai, T. *et al.* Overproduction of the membrane-bound [NiFe]-hydrogenase in *Thermococcus kodakarensis* and its effect on hydrogen production. *Front. Microbiol.* **6**, 847 (2015).
49. Basen, M., Sun, J. & Adams, M. W. W. Engineering a hyperthermophilic archaeon for temperature-dependent product formation. *MBio* **3**, e00053–12 (2012).
50. Krupovic, M., Gonnet, M., Hania, W. B., Forterre, P. & Erauso, G. Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new *Thermococcus* plasmids. *PLoS One* **8**, e49044 (2013).
51. Soler, N. *et al.* The rolling-circle plasmid pTN1 from the hyperthermophilic archaeon

Thermococcus nautilus. *Mol. Microbiol.* **66**, 357–370 (2007).

52. 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).
53. Gorlas, A., Krupovic, M., Forterre, P. & Geslin, C. Living side by side with a virus: characterization of two novel plasmids from *Thermococcus prieurii*, a host for the spindle-shaped virus TPV1. *Appl. Environ. Microbiol.* **79**, 3822–3828 (2013).
54. Lin-Chao, S., Chen, W. T. & Wong, T. T. High copy number of the pUC plasmid results from a Rom/Rop-suppressible point mutation in RNA II. *Mol. Microbiol.* **6**, 3385–3393 (1992).
55. Del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* **62**, 434–464 (1998).
56. Santangelo, T. J., Cubonová, L. 'ubomíra, James, C. L. & Reeve, J. N. TFB1 or TFB2 is sufficient for *Thermococcus kodakaraensis* viability and for basal transcription in vitro. *J. Mol. Biol.* **367**, 344–357 (2007).
57. Rodrigues, M. V. *et al.* Bifunctional CTP:inositol-1-phosphate cytidylyltransferase/CDP-inositol:inositol-1-phosphate transferase, the key enzyme for di-myo-inositol-phosphate synthesis in several (hyper)thermophiles. *J. Bacteriol.* **189**, 5405–5412 (2007).

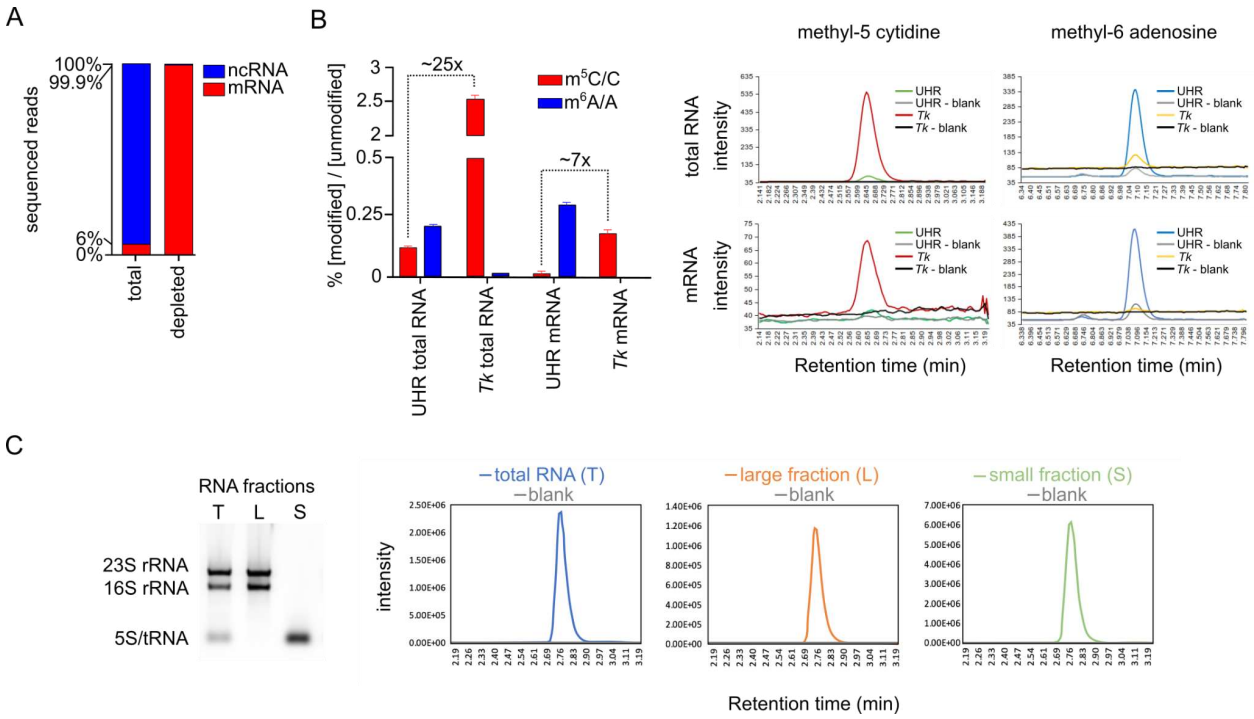


Figure A2.1 m^5C dominates the epitranscriptomic landscape in *T. kodakarensis.* (A) RNA was collected from cells and the rRNA and tRNAs depleted. Total RNA includes ~6% mRNAs whereas virtually all post-depletion RNA is mRNA. **(B)** Methyl-5 cytidine without phosphate groups has a retention time of ~2.65 min. Mass spectrometry analysis of total and depleted RNA from a Universal Human Reference (UHR) and *T. kodakarensis* indicate m^5C is 25 fold more abundant in total RNA and 13 fold more abundant in mRNAs in *T. kodakarensis*. In contrast, m^6A is much more abundant in human cells and excluded from *T. kodakarensis* mRNAs. **(C)** Total, large and small fractions of RNA from *T. kodakarensis* indicate that m^5C is highly abundant in rRNA and tRNAs.

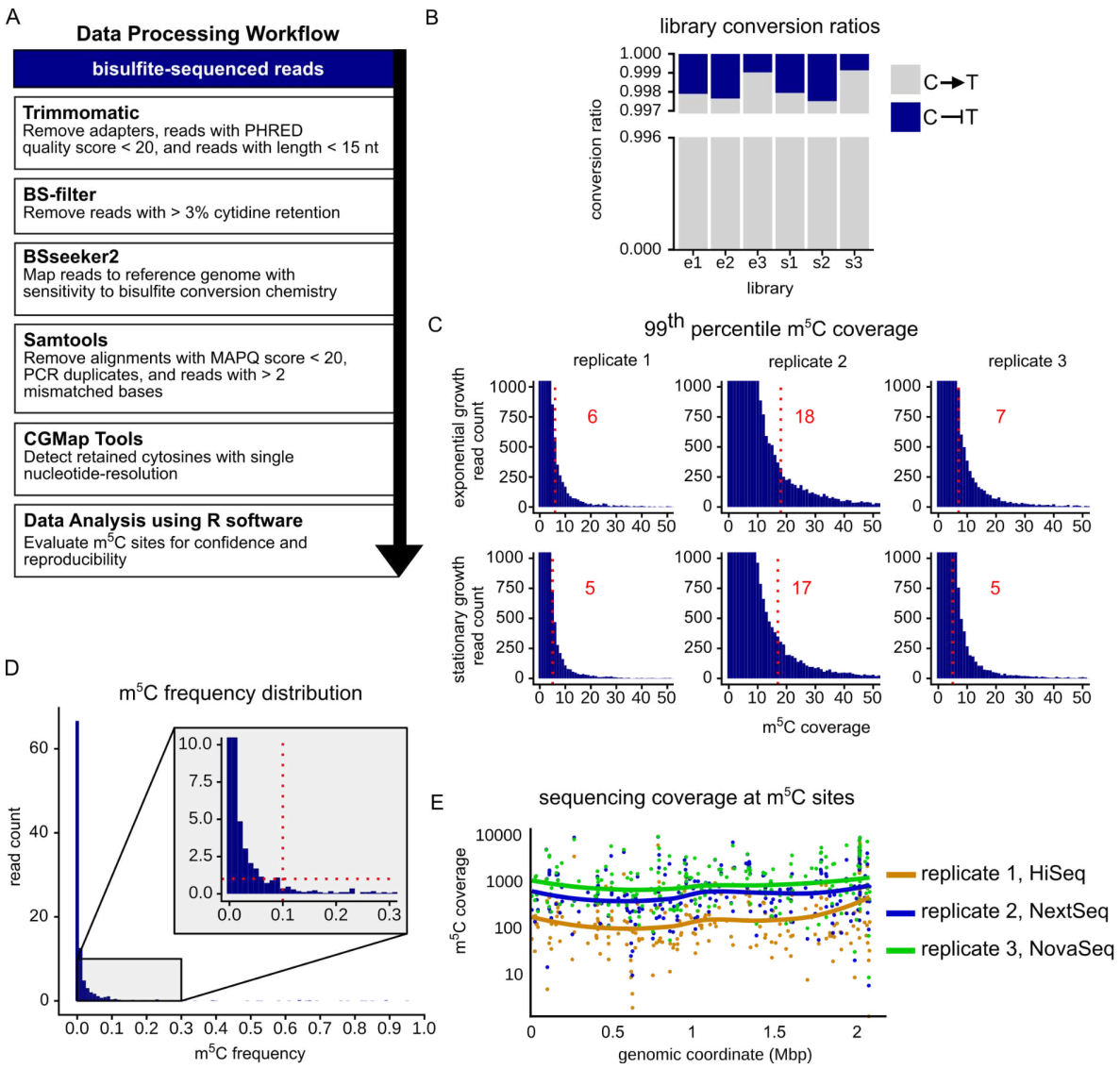


Figure A2.2 Bisulfite sequencing of RNA permits comprehensive detection of high confidence and reproducible m^{5}C sites. RNA was collected from cells grown to exponential or stationary growth phase and bisulfite-sequenced on an Illumina platform. **(A)** Reads sequenced with low quality were removed and adapters trimmed. A custom python script was used to remove reads where cytidine constituted >3% of the nucleotides. The reads were then mapped to the reference genome using BSseeker2 where C to T conversions are not counted as mis-matched bases. The aligned reads were further filtered where low MAPQ scores, PCR duplicates, and alignments with > 2 non-C-to-T mismatches were removed. Cytidine retention at single nucleotide resolution was quantified using CGmaptools. Finally, CGmaps were analyzed using R software. **(B**) The library-wide cytidine conversion ratio was quantified for three replicates from each biological condition. **(C)** For each sequenced library, we calculated the 99th percentile for m^{5}C coverage (listed in red text) for all genomically encoded cytidines. **(D)** A histogram illustrates the m^{5}C frequency of all candidate m^{5}C sites. A sharp decline followed by a steady pattern of m^{5}C frequencies are observed at ~10% modification frequency. **(E)** The m^{5}C coverage at all high-confidence and reproducible m^{5}C sites is extraordinarily high in all sequenced libraries, ranging from 100-1000X coverage on average.

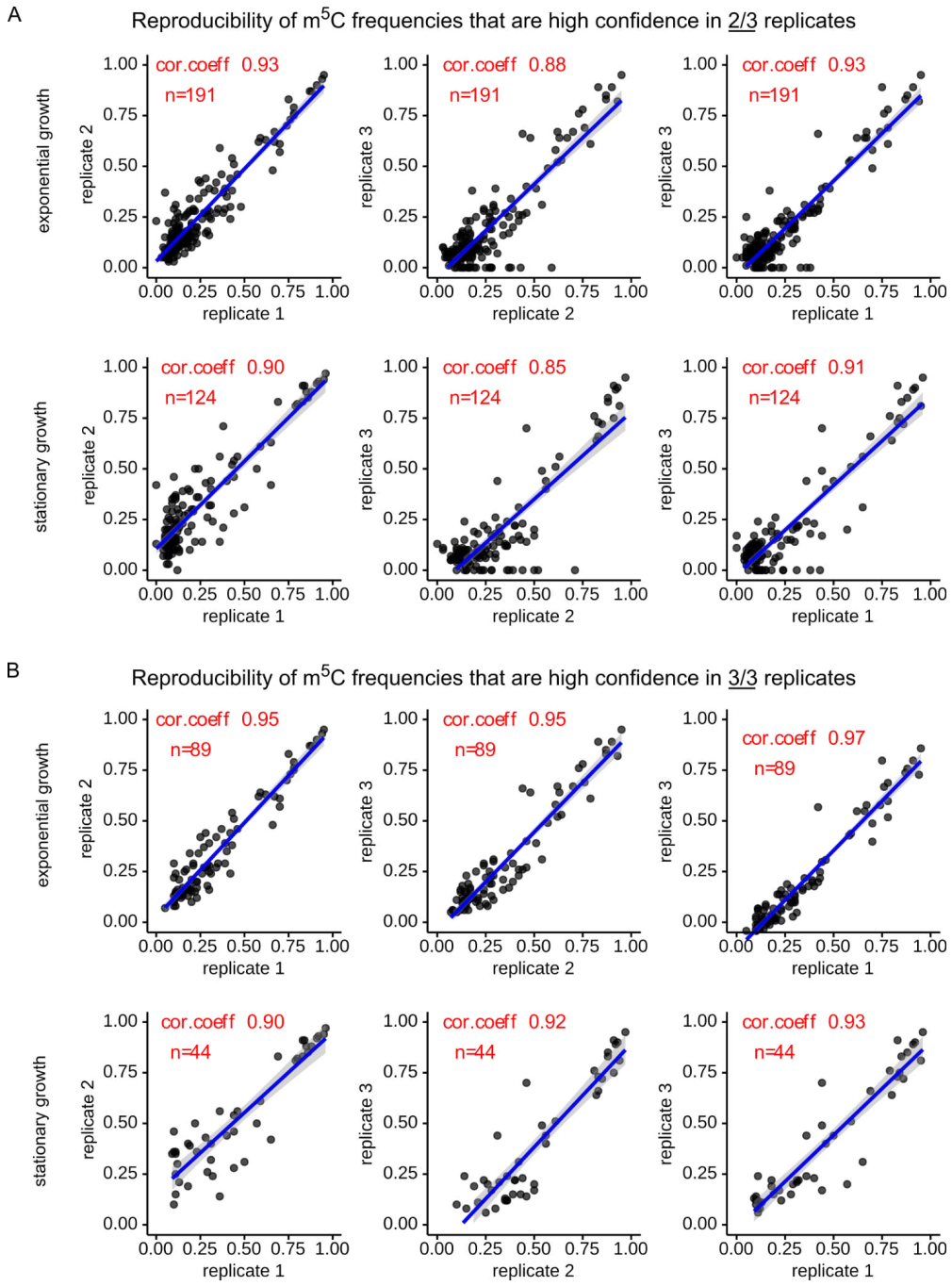


Figure A2.3 Linear regression of m^{5}C frequencies indicated highly reproducible modification frequencies. Linear regression and Pearson's correlation coefficients (cor. coeff) were used to compare the site-specific modification frequencies of high confidence and reproducible m^{5}C sites in each replicate. The analysis was performed for sites reproducible in **(A)** 2/3 and **(B**) 3/3 replicates. The analysis was performed using R software.

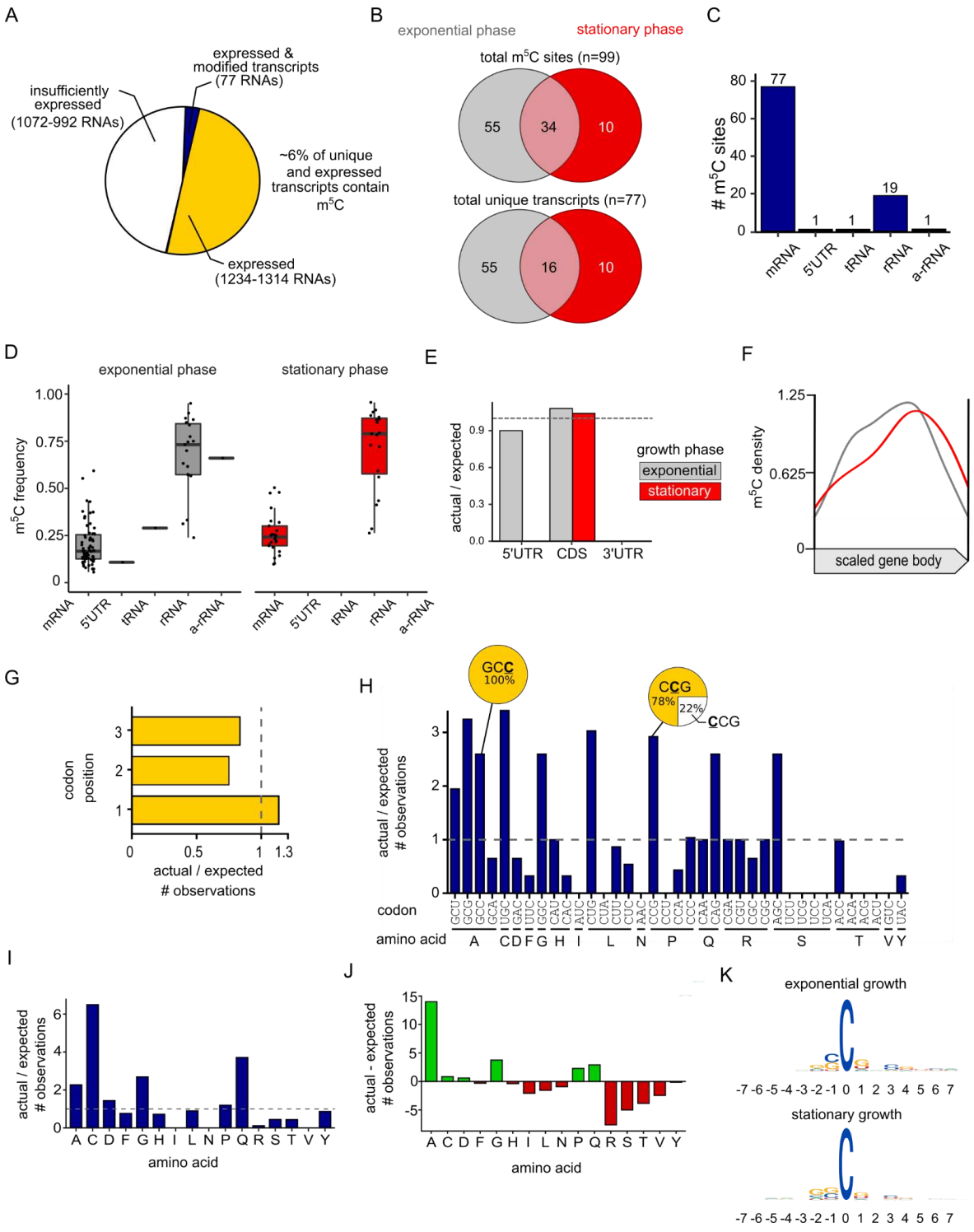

A
expressed &
modified transcripts
(77 RNAs)
insufficiently
expressed
(1072-992 RNAs)
~6% of unique
and expressed
transcripts contain
m5C
expressed
(1234-1314 RNAs)
B
exponential phase
stationary phase
total m5C sites (n=99)
55
34
10
total unique transcripts (n=77)
55
16
10
C
m5C sites
77
1
1
19
1
mRNA
5'UTR
tRNA
rRNA
a-rRNA
D
exponential phase
stationary phase
m5C frequency
E
actual / expected
growth phase
exponential
stationary
5'UTR
CDS
3'UTR
F
m5C density
scaled gene body
G
codon
position
actual / expected
observations
H
GCC
100%
CCG
78%
22%
CCG
actual / expected
observations
codon
amino acid
I
actual / expected
observations
amino acid
J
actual - expected
observations
amino acid
K
exponential growth
stationary growth

Figure A2.4 Analysis of m^5C sites present in 3/3 replicates reveal largely similar conclusions compared to m^5C sites present in at least 2/3 replicates. Parallel analysis of the context in which m^5C is found throughout the transcriptome was performed for m^5C sites found in all three replicates. **(A)** We detected 1234-1314 RNAs with expression at or above the thresholds set for the minimum coverage required to call m^5C sites in all three replicates. Between two biological conditions, 6% of expressed transcripts contained at least one m^5C site. **(B)** The number of m^5C sites and unique transcripts were calculated in two biological conditions, and distinct m^5C-epitranscriptomic profiles were observed between exponential and stationary growth phases. **(C)** m^5C are similarly distributed throughout a diverse set of RNAs and **(D**) average modification frequencies are nearly identical when compared to analysis of m^5C sites present in at least 2/3 replicates. **(E)** We compared the actual number observed and the expected number of m^5C sites in different regions of an mRNA. We expect the actual / expected number of m^5C sites to be equal to 1 if the m^5C site is distributed throughout the transcriptome randomly. Unlike m^5C sites present in at least 2/3 replicates, there were no sites mapped to 3'UTRs, so we did not observe an enrichment in this region when only analyzing sites present in all 3 replicates. **(F)** When m^5C mapped to coding sequences, there was no apparent positional bias, **(G)** nor did we observe a strong bias in codon position. **(H)** The number of m^5C sites mapped to each codon was compared to the number expected by random chance. Obvious deviations from expectation (actual / expected = 1) were similarly observed between sites present in 2/3 or 3/3 replicates. **(I)** The actual / expected and **(J)** difference between actual and expected (actual - expected = 0 where no bias is detected) number of m^5C sites mapping to particular amino acid codons was calculated and indicate clear bias in m^5C positioning in select amino acid contexts. **(K)** Logos sequence analysis of nucleotides adjacent to m^5C sites do not reveal a strong RNA sequence context.

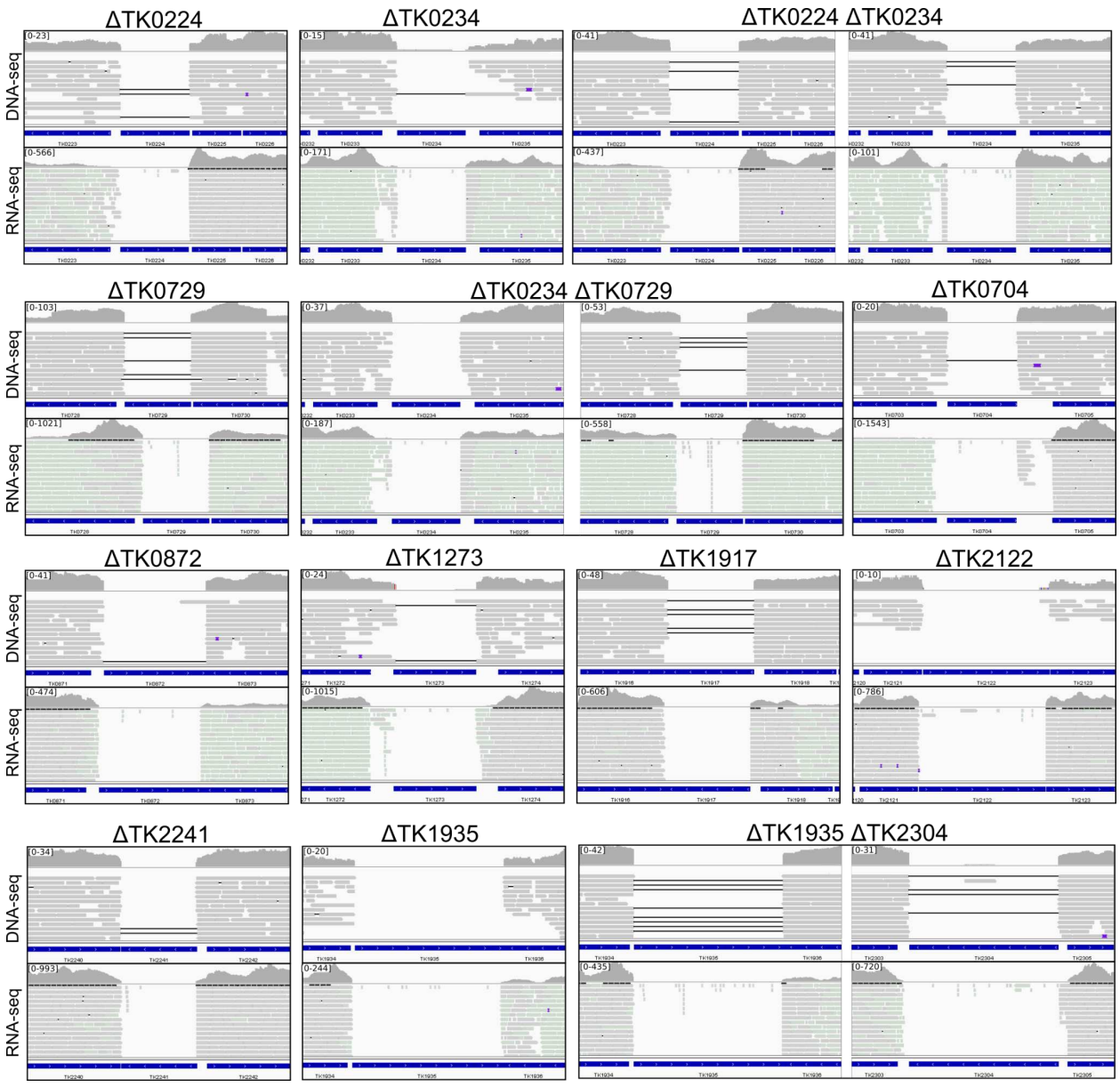


Figure A2.5 Genes encoding putative RMTases were genomically deleted. Genomic deletions were initially screened using PCR for preliminary confirmation (data not shown), and final confirmation was performed using Minion whole genome sequencing (DNA-seq) and Illumina RNA bisulfite sequencing (RNA-seq). Visual inspection of DNA and RNA sequences aligned to the *T. kodakarensis* reference genome using Integrative Genomics Viewer. In each window, the coverage and read tracks are stacked. The RNA-seq alignments are displayed using bisulfite-mode. Reads absent from the deleted loci indicate markerless removal of the loci encoding the target RMTase. In the DNA-seq windows, black lines connect contiguous reads and indicate a gapped alignment, further confirming the target DNA region is removed. The boundaries of each gene and their TK gene IDs are represented by blue bars, and white arrows illustrate the direction in which the gene is transcribed. The coverage range for each window is listed in brackets.

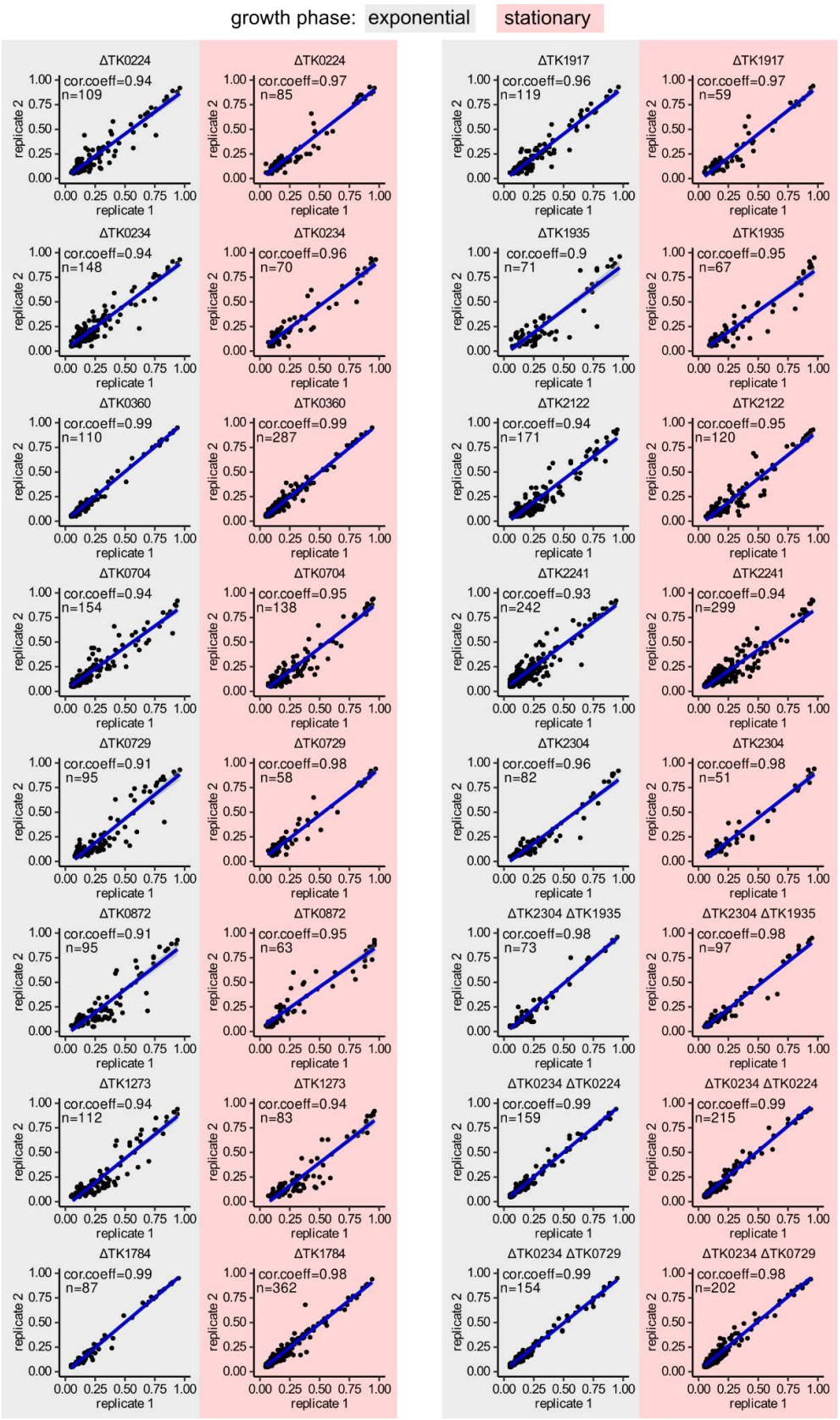

growth phase: exponential stationary
ΔTK0224
cor.coeff=0.94
n=109
ΔTK0224
cor.coeff=0.97
n=85
ΔTK1917
cor.coeff=0.96
n=119
ΔTK1917
cor.coeff=0.97
n=59
ΔTK0234
cor.coeff=0.94
n=148
ΔTK0234
cor.coeff=0.96
n=70
ΔTK1935
cor.coeff=0.9
n=71
ΔTK1935
cor.coeff=0.95
n=67
ΔTK0360
cor.coeff=0.99
n=110
ΔTK0360
cor.coeff=0.99
n=287
ΔTK2122
cor.coeff=0.94
n=171
ΔTK2122
cor.coeff=0.95
n=120
ΔTK0704
cor.coeff=0.94
n=154
ΔTK0704
cor.coeff=0.95
n=138
ΔTK2241
cor.coeff=0.93
n=242
ΔTK2241
cor.coeff=0.94
n=299
ΔTK0729
cor.coeff=0.91
n=95
ΔTK0729
cor.coeff=0.98
n=58
ΔTK2304
cor.coeff=0.96
n=82
ΔTK2304
cor.coeff=0.98
n=51
ΔTK0872
cor.coeff=0.91
n=95
ΔTK0872
cor.coeff=0.95
n=63
ΔTK2304 ΔTK1935
cor.coeff=0.98
n=73
ΔTK2304 ΔTK1935
cor.coeff=0.98
n=97
ΔTK1273
cor.coeff=0.94
n=112
ΔTK1273
cor.coeff=0.94
n=83
ΔTK0234 ΔTK0224
cor.coeff=0.99
n=159
ΔTK0234 ΔTK0224
cor.coeff=0.99
n=215
ΔTK1784
cor.coeff=0.99
n=87
ΔTK1784
cor.coeff=0.98
n=362
ΔTK0234 ΔTK0729
cor.coeff=0.99
n=154
ΔTK0234 ΔTK0729
cor.coeff=0.98
n=202
replicate 1
replicate 2
0.00 0.25 0.50 0.75 1.00

Figure A2.6 Linear regression of modification frequencies between replicates in strains deleted for putative RMTases. Linear regression and Pearson's correlation coefficients (cor. coeff) were used to compare the site-specific modification frequencies of high confidence and reproducible m^5C sites in each strain across replicates and growth phases. The analysis was performed for sites reproducible. The analysis was performed using R software.

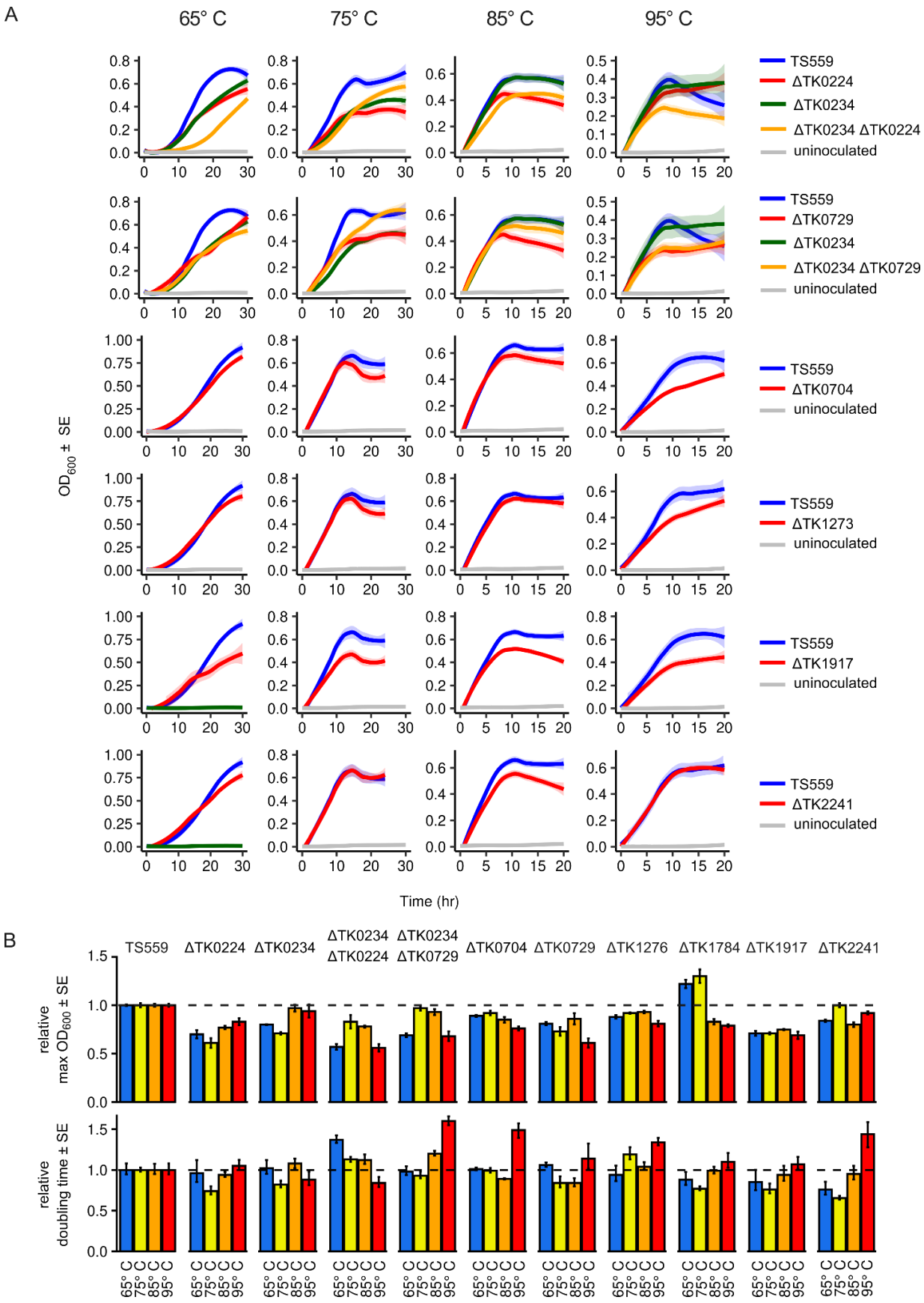

A
65° C
75° C
85° C
95° C
TS559
ΔTK0224
ΔTK0234
ΔTK0234 ΔTK0224
uninoculated
TS559
ΔTK0729
ΔTK0234
ΔTK0234 ΔTK0729
uninoculated
TS559
ΔTK0704
uninoculated
TS559
ΔTK1273
uninoculated
TS559
ΔTK1917
uninoculated
TS559
ΔTK2241
uninoculated
OD600 ± SE
Time (hr)
B
TS559
ΔTK0224
ΔTK0234
ΔTK0234 ΔTK0224
ΔTK0234 ΔTK0729
ΔTK0704
ΔTK0729
ΔTK1276
ΔTK1784
ΔTK1917
ΔTK2241
relative max OD600 ± SE
relative doubling time ± SE
65° C
75° C
85° C
95° C

Figure A2.7 Phenotypic analysis of 10 putative RMTases indicate a role for the epitranscriptome in thermophily. (A) Head-to-head growth competitions were performed at 65°, 75°, 85°, and 95° C for each strain deleted for a single of 2 putative RMTase and parent strain TS559. **(B)** The maximum optical density (Max OD_{600}) and the rate of growth (doubling time) for each culture is illustrated relative to parent strain TS559.

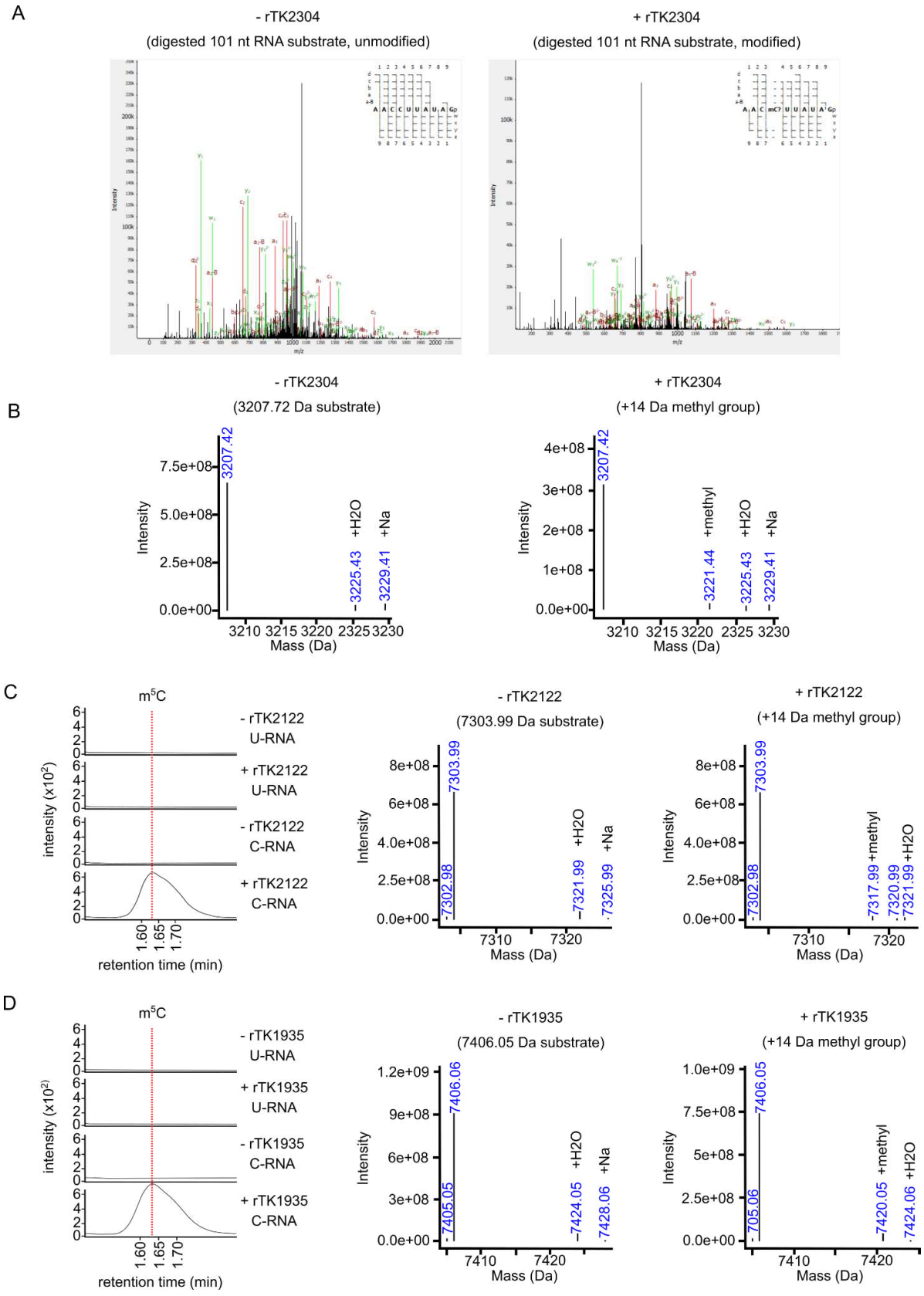

A
- rTK2304
(digested 101 nt RNA substrate, unmodified)
+ rTK2304
(digested 101 nt RNA substrate, modified)
B
- rTK2304
(3207.72 Da substrate)
+ rTK2304
(+14 Da methyl group)
Intensity
Mass (Da)
3207.42
+H2O
+Na
+methyl
C
m5C
- rTK2122
U-RNA
+ rTK2122
U-RNA
- rTK2122
C-RNA
+ rTK2122
C-RNA
intensity (x10²)
retention time (min)
(7303.99 Da substrate)
(+14 Da methyl group)
D
- rTK1935
U-RNA
+ rTK1935
U-RNA
- rTK1935
C-RNA
+ rTK1935
C-RNA
(7406.05 Da substrate)
(+14 Da methyl group)

Figure A2.8 Mass spectrometry analysis R5CMTs demonstrate *bonafide* MTase activity. **(A)** RNase T1 digestion results in the modified fragment AAC**C**UUAUAG being released from the 101 nt oligonucleotide. The underlined C is the anticipated target cytidine for modification by rTK2304. The mass spectra of the substrate fragments incubated with or without rTK2304 enzyme are shown. **(B)** Mass deconvolution of TK2304-treated 101nt substrate RNaseT1 Product (AACCUUAUAGp, 3207.4265 Da) indicates a mass shift consistent with a methyl group (~14 Da) at the expected cytidine. The relative methylated product abundance is 5.85%. Analysis of a 23 nt RNA substrate modified by **(C)** rTK2122 or **(D)** rTK1935 indicate a mass shift consistent with m^5C.

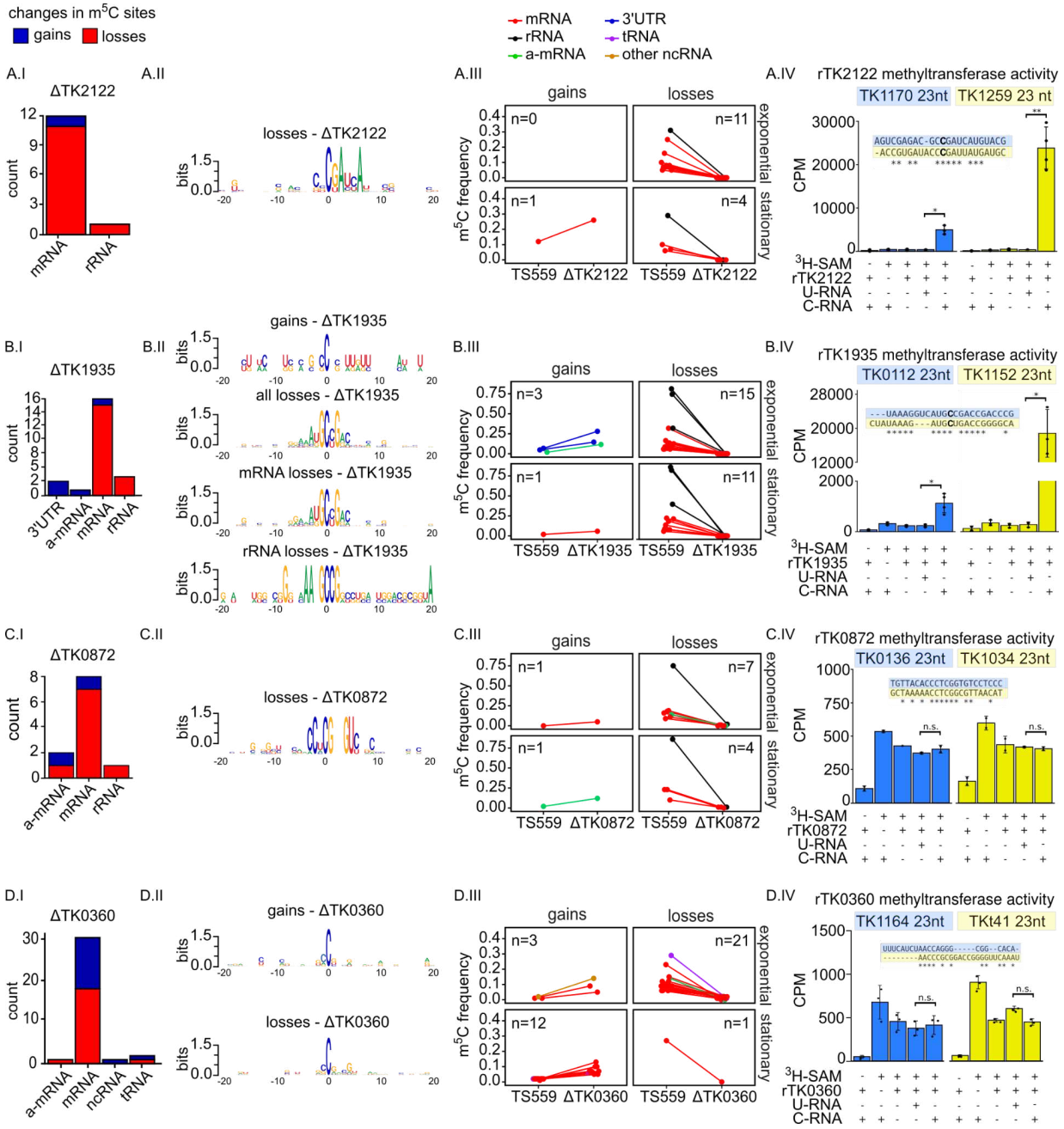


Figure A2.9 Four additional R5CMTs contribute to the maintenance of the m^{5}C-epitranscriptome in *T. kodakarensis*. Genes encoding **(A)** TK2122, **(B)** TK1935, and **(C)** TK0872 exhibit both gains and losses (subfigures I) in m^{5}C sites upon gene deletion. Sites where losses or gains were detected are defined within sequence contexts (subfigures II). The change in modification frequencies of individual sites between the parent and deletion strains are illustrated (subfigures III). Methyltransferase activity assays were completed for each enzyme using RNA substrates identified *in vivo* to be methylated by the select enzyme (subfigures IV). Methyltransferase activity is measured in counts per minute (CPM). The substrates used are 23 nt oligonucleotides with the target cytidine at the center in the C-RNA. Sequence alignments between the two 23 nt substrates are displayed with stars denoting a match. * $p < 0.05$, **$p < 0.01$, not significant (n.s.).

Table A2.2 Differential m^5C sites detected in strains deleted for putative RMTase, not thought to install m^5C.

		exponential phase[b]		stationary phase[c]		
Gene ID	Description[a]	rel. gains	rel. losses	rel. gains	rel. losses	# losses/ gains
TK0224	class I SAM-dependent MTase, UbiE/COQ5 family	3	0	0	0	0
TK0234	MTase domain-containing protein	3	0	0	0	5
TK0704	class I SAM-dependent MTase, UbiE/COQ5 family	0	0	0	0	10
TK0729	class I SAM-dependent MTase, UbiE/COQ5 family	1	0	0	0	20
TK1273	class I SAM-dependent MTase, UbiE/COQ5 family	0	0	0	0	30
TK1784	predicted SAM-dependent MTase, DUF890 family	1	10	34	0	
TK1917	class I SAM-dependent MTase	1	0	0	0	
TK2241	class I SAM-dependent MTase, UbiE/COQ5 family	2	0	20	0	
TK0234; TK0224	double deletion	10	5	25	1	
TK0234; TK0729	double deletion	7	3	12	0	

[a] The NCBI description and protein family (Uniprot and InterPro predictions) are provided.

[b,c] BS-seq of individual or double deletion of non-essential RMTases, we identified relative losses and gains (≥ 2X fold change) in m5C sites between exponential and stationary growth phase cells.

Table A2.2 RNA substrates analyzed by LC-MS/MS

Recombinant Enzyme	RNA substrate name	RNA substrate sequence[a]
rTK2304	TK1911_23nt_C	UAGCGAAGAAC**C**UUAUAGUUGCU
rTK2304	TK1911_23nt_U	UAGCGAAGAAC**U**UUAUAGUUGCU
rTK2304	TK1911_101nt_C	AGGGCAAGGAGCAGGGCCUUCGUGACAGGUCGGAGAUGAUAGCG AAGAAC**C**UUAUAGUUGCUGGAAAAAUUGAGCCUGAGGGGACAUC UAAAAAGGUUGAG
rTK2304	TK1911_101nt_U	AGGGCAAGGAGCAGGGCCUUCGUGACAGGUCGGAGAUGAUAGCG AAGAAC**U**UUAUAGUUGCUGGAAAAAUUGAGCCUGAGGGGACAUC UAAAAAGGUUGAG
rTK2122	TK1259_23nt _C	ACCGUGAUACC**C**GAUUAUGAUGC
rTK2122	TK1259_23nt _U	ACCGUGAUACC**U**GAUUAUGAUGC
rTK1935	TK1152_23nt_C	CUAUAAAGAUG**C**UGACCGGGGCA
rTK1935	TK1152_23nt_U	CUAUAAAGAUG**U**UGACCGGGGCA

[a] The cytidine targeted for methylation or the uridine substitution are underlined and bolded.

Table A3.1 RNA substrates for *in vitro* methyltransferase assays.

description	length	sequence
target cytidine at center	23 nt	UUAAUUGGAUU**C**AACGCCGGGAA
target cytidine substituted for uracil	23 nt	UUAAUUGGAUU**U**AACGCCGGGAA
target cytdine at center	46 nt	GGAGCG[U]GCGGUUUAAUU[G]GAU[U]**C**AACGCCGGGAACCUCACCGGGG
target cytidine substituted for uracil	46 nt	GGAGCG[U]GCGGUUUAAUU[G]GAU[U]**U**AACGCCGGGAACCUCACCGGGG
partially modified, target cytdine at center	46 nt	GGAGCG[Ψ]GCGGUUUAAUU[G]GAU[m1Ψ]**C**AACGCCGGGAACCUCACCGGGG
modified, target cytdine at center	46 nt	GGAGCG[Ψ]GCGGUUUAAUU[m1G]GAU[m1Ψ]**C**AACGCCGGGAACCUCACCGGGG
target cytdine at center	201 nt	CGUUAAGCCCGCCGCCUGGGGAGUACGGCCGCAAGGCUGAAACUUAAAGGAAUUGGCGGGGGGAGCACUACAAGGGGUGGAGCGUGCGGUUUAAUUGGAUU**C**AACGCCGGGAACCUCACCGGGGGCGACGGCAGGAUGAAGGCCAGGCUGAAGGUCUUGCCGGACACGCCGAGAGGAGGUGCAUGGCCGCCGUCAGCUCGUA
target cytidine substituted for uracil	201 nt	CGUUAAGCCCGCCGCCUGGGGAGUACGGCCGCAAGGCUGAAACUUAAAGGAAUUGGCGGGGGGAGCACUACAAGGGGUGGAGCGUGCGGUUUAAUUGGAUU**U**AACGCCGGGAACCUCACCGGGGGCGACGGCAGGAUGAAGGCCAGGCUGAAGGUCUUGCCGGACACGCCGAGAGGAGGUGCAUGGCCGCCGUCAGCUCGUA
target cytidine at center	301 nt	CGGGCUAGGUGUCGGGCGAGCUUCGAGCUCGCCCGGUGCCGGAGGGAAGCCGUUAAGCCCGCCGCCUGGGGAGUACGGCCGCAAGGCUGAAACUUAAAGGAAUUGGCGGGGGAGCACUACAAGGGGUGGAGCGUGCGGUUUAAUUGGAUU**C**AACGCCGGGAACCUCACCGGGGGCGACGGCAGGAUGAAGGCCAGGCUGAAGGUCUUGCCGGACACGCCGAGAGGAGGUGCAUGGCCGCCGUCAGCUCGUACCGUGAGGCGUCCACUUAAGUGUGGUAACGAGCGAGACCCGCGCCCCCAG
target cytidine substituted for uracil	301 nt	CGGGCUAGGUGUCGGGCGAGCUUCGAGCUCGCCCGGUGCCGGAGGGAAGCCGUUAAGCCCGCCGCCUGGGGAGUACGGCCGCAAGGCUGAAACUUAAAGGAAUUGGCGGGGGAGCACUACAAGGGGUGGAGCGUGCGGUUUAAUUGGAUU**U**AACGCCGGGAACCUCACCGGGGGCGACGGCAGGAUGAAGGCCAGGCUGAAGGUCUUGCCGGACACGCCGAGAGGAGGUGCAUGGCCGCCGUCAGCUCGUACCGUGAGGCGUCCACUUAAGUGUGGUAACGAGCGAGACCCGCGCCCCCAG
target cytidine at center	401 nt	AGGGGAGCGAACCGGAUUAGAUACCCGGGUAGUCCUGGCUGUAAAGGAUGCGGGCUAGGUGUCGGGCGAGCUUCGAGCUCGCCCGGUGCCGGAGGGAAGCCGUUAAGCCCGCCGCCUGGGGAGUACGGCCGCAAGGCUGAAACUUAAAGGAAUUGGCGGGGGAGCACUACAAGGGGUGGAGCGUGCGGUUUAAUUGGAUU**C**AACGCCGGGAACCUCACCGGGGGCGACGGCAGGAUGAAGGCCAGGCUGAAGGUCUUGCCGGACACGCCGAGAGGAGGUGCAUGGCCGCCGUCAGCUCGUACCGUGAGGCGUCCACUUAAGUGUGGUAACGAGCGAGACCCGCGCCCCCAGUUGCCAGUCCUCCCCGCUGGGGAGGAGGCACUCUGGGGGGACCGCCGGCG

target cytidine substituted for uracil	401 nt	AGGGGAGCGAACCGGAUUAGAUACCCGGGUAGUCCUGGCUGUAAAGGAUGCGGG CUAGGUGUCGGGCGAGCUUCGAGCUCGCCCGGUGCCGGAGGGAAGCCGUUAAGC CCGCCGCCUGGGGAGUACGGCCGCAAGGCUGAAACUUAAAGGAAUUGGCGGGGG AGCACUACAAGGGGUGGAGCGUGCGGUUUAAUUGGAUU**U**AACGCCGGGAACCUC ACCGGGGGCGACGGCAGGAUGAAGGCCAGGCUGAAGGUCUUGCCGGACACGCCG AGAGGAGGUGCAUGGCCGCCGUCAGCUCGUACCGUGAGGCGUCCACUUAAGUGU GGUAACGAGCGAGACCCGCGCCCCCAGUUGCCAGUCCUCCCCGCUGGGGAGGAG GCACUCUGGGGGGACCGCCGGCG
target cytidine at -46	966 nt	AUUCCGGUUGAUCCUGCCGGAGGCCACUGCUAUGGGGGUCCGACUAAGCCAUGC GAGUCAUGGGGCGCGCUCUGCGCGCACCGGCGGACGGCUCAGUAACACGUCGGU AACCUACCCUCGGGAGGGGGAUAACCCCGGGAAACUGGGGCUAAUCCCCCAUAG GCCUGAGGUACUGGAAGGUCCUCAGGCCGAAAGGGGCAUCUGCCCGCCCGAGGA UGGGCCGGCGGCCGAUUAGGUAGUUGGUGGGGUAACGGCCCACCAAGCCGAAGA UCGGUACGGGCCAUGAGAGUGGGAGCCCGGAGAUGGACACUGAGACACGGGUCC AGGCCCUACGGGGCGCAGCAGGCGCGAAACCUCCGCAAUGCGGGCAACCGCGAC GGGGGGACCCCCAGUGCCGUGGCAUAGCCACGGCUUUUCCGGAGUGUAAAAAGC UCCGGGAAUAAGGGCUGGGCAAGGCCGGUGGCAGCCGCCGCGGUAAUACCGGCG GCCCGAGUGGUGGCCGCUAUUAUUGGGCCUAAAGCGUCCGUAGCCGGGCCCGUA AGUCCCUGGCGAAAUCCCACGGCUCAACCGUGGGGCUUGCUGGGGAUACUGCGG GCCUUGGGACCGGGAGAGGCCGGGGGUACCCCUGGGGUAGGGGUGAAAUCCUAU AAUCCCAGGGGGACCGCCAGUGGCGAAGGCGCCCGGCUGGAACGGGUCCGACGG UGAGGGACGAAGGCCAGGGGAGCGAACCGGAUUAGAUACCCGGGUAGUCCUGGC UGUAAAGGAUGCGGGCUAGGUGUCGGGCGAGCUUCGAGCUCGCCCGGUGCCGGA GGGAAGCCGUUAAGCCCGCCGCCUGGGGAGUACGGCCGCAAGGCUGAAACUUAA AGGAAUUGGCGGGGGAGCACUACAAGGGGUGGAGCGUGCGGUUUAAUUGGAUU**C** AACGCCGGGAACCUCACCGGGGGCGACGGCAGGAUGAAGGCCAGGCUG
target cytidine substituted for uracil	966 nt	AUUCCGGUUGAUCCUGCCGGAGGCCACUGCUAUGGGGGUCCGACUAAGCCAUGC GAGUCAUGGGGCGCGCUCUGCGCGCACCGGCGGACGGCUCAGUAACACGUCGGU AACCUACCCUCGGGAGGGGGAUAACCCCGGGAAACUGGGGCUAAUCCCCCAUAG GCCUGAGGUACUGGAAGGUCCUCAGGCCGAAAGGGGCAUCUGCCCGCCCGAGGA UGGGCCGGCGGCCGAUUAGGUAGUUGGUGGGGUAACGGCCCACCAAGCCGAAGA UCGGUACGGGCCAUGAGAGUGGGAGCCCGGAGAUGGACACUGAGACACGGGUCC AGGCCCUACGGGGCGCAGCAGGCGCGAAACCUCCGCAAUGCGGGCAACCGCGAC GGGGGGACCCCCAGUGCCGUGGCAUAGCCACGGCUUUUCCGGAGUGUAAAAAGC UCCGGGAAUAAGGGCUGGGCAAGGCCGGUGGCAGCCGCCGCGGUAAUACCGGCG GCCCGAGUGGUGGCCGCUAUUAUUGGGCCUAAAGCGUCCGUAGCCGGGCCCGUA AGUCCCUGGCGAAAUCCCACGGCUCAACCGUGGGGCUUGCUGGGGAUACUGCGG GCCUUGGGACCGGGAGAGGCCGGGGGUACCCCUGGGGUAGGGGUGAAAUCCUAU AAUCCCAGGGGGACCGCCAGUGGCGAAGGCGCCCGGCUGGAACGGGUCCGACGG UGAGGGACGAAGGCCAGGGGAGCGAACCGGAUUAGAUACCCGGGUAGUCCUGGC UGUAAAGGAUGCGGGCUAGGUGUCGGGCGAGCUUCGAGCUCGCCCGGUGCCGGA GGGAAGCCGUUAAGCCCGCCGCCUGGGGAGUACGGCCGCAAGGCUGAAACUUAA AGGAAUUGGCGGGGGAGCACUACAAGGGGUGGAGCGUGCGGUUUAAUUGGAUU**U** AACGCCGGGAACCUCACCGGGGGCGACGGCAGGAUGAAGGCCAGGCUG

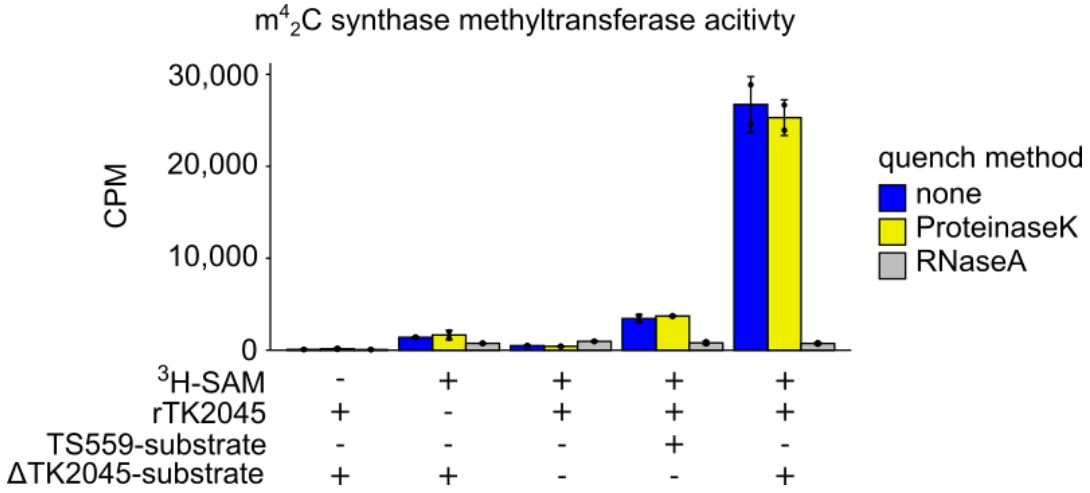


Figure A3.1 Quenched methyltransferase activity assay. Methyltransferase activity of $m^4{}_2C$ synthase retained on substrates after selective degradation of protein or RNA.

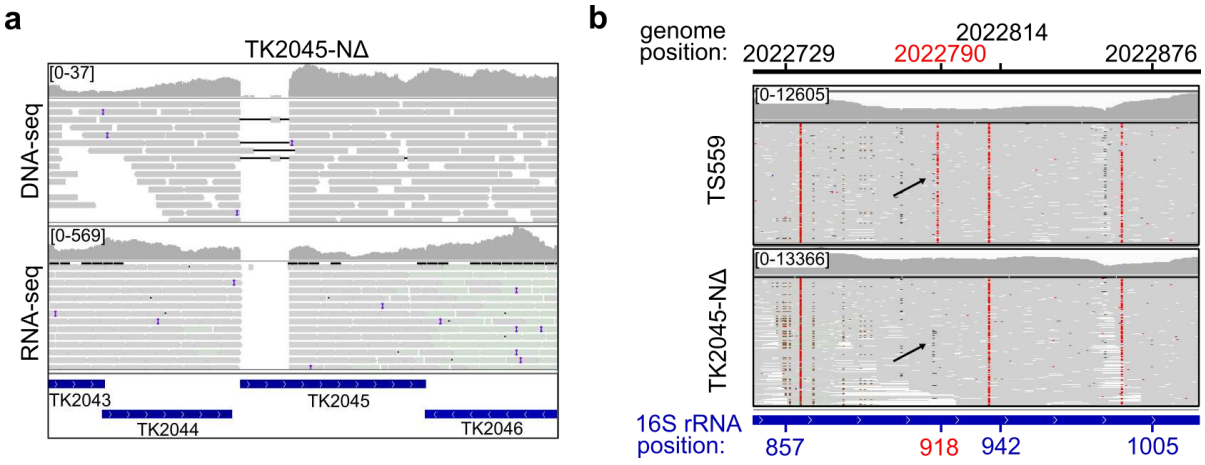


Figure A3.2 The N-terminal lobe is essential. a, Genomic deletion of the N-terminal region of TK2045 was confirmed using whole genome sequencing (DNA-seq) and RNA-bisulfite sequencing (RNA-seq). Black lines that connect gray bars represent single reads with a gapped alignment. **b,** Bisulfite sequenced RNA collected from TS559 or TK2045-NΔ was aligned to a reference genome where all cytidines were replaced with thymines, and visualized using Integrative Genomics Viewer. Sites of modification are indicated where a cytidine is retained (red) within a sequenced read.