Abstract

Reengineering the Uridine-54 tRNA Methyltransferase, TrmA, to Covalently Capture New RNA Substrates

Tyler Smith

2019

Cellular RNAs can be regulated by enzymes that covalently modify specific RNA substrates. Altering the RNA specificities of these enzymes can provide powerful tools to study and manipulate cellular RNAs. I studied the RNA binding properties of the uridine-54 tRNA methyltransferase, TrmA, using a mutant of the enzyme (TrmA*) that covalently binds substrate RNA. This covalent binding protein allowed me to use high-throughput sequencing methods to examine the RNA substrates of TrmA in a way that would be impossible to do using traditional methyltransferase assays. In addition to substrate screening, I employed rational design based on the crystal structure of TrmA and its substrate to reengineer the enzyme and substrate RNA specificity. This engineered interaction resulted in a substantial change in TrmA binding specificity, but not a complete change. Combining rational design of TrmA and high-throughput substrate screening, I discovered a triple mutant of the substrate RNA (C56A A58G C60U) that was bound by a TrmA* double mutant (E49R R51E) but not by the wild type enzyme.

In parallel to my work exploring the specificity of TrmA, I have worked to use TrmA* to form covalent RNA-protein adducts. One possible application of this would be to tag an RNA of interest and use TrmA* to pull it down. In order to study RNA-protein interactions, researchers employ a variety of affinity tags to capture an RNA of interest.

The ability to purify RNA-protein complexes are limited by low affinity binding, which limits the wash conditions that are available to remove background contamination. Because TrmA* covalently binds RNA, I am less limited in the wash conditions and have successfully pulled down tagged RNA *in vivo* and *in vitro* using TrmA* coupled with fully denaturing washes. While I have not solved the problem of capturing RNA-protein complexes from crosslinked extracts, I have laid the foundations for future researchers to use this covalent affinity tagging system. Combining my discoveries of altered TrmA specificity with covalent capture will allow researchers to biochemically purify an RNA of interest with fully denaturing washes in order to identify bound proteins with high confidence.

Another application I have explored with TrmA* is to tag an RNA transcript both 5' and 3' with TrmA substrate RNA and coexpress TrmA*. This forms protein crosslinks *in vivo*, which could function in a manner similar to sno-lncRNAs to block exonuclease degradation and stabilize transcripts. My preliminary results expressing TrmA with a tagged luciferase transcript show that this may be possible and could be a valuable tool for stabilizing RNAs of interest. My doctoral research on reengineering TrmA provides the groundwork for future researchers to covalently bind RNA with TrmA in a variety of ways similar, but not restricted to, those described here.

Reengineering the Uridine-54 tRNA Methyltransferase, TrmA, to Covalently Capture New RNA Substrates

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctorate of Philosophy

by
Tyler Samuel Smith

Dissertation Director: Matthew D. Simon

December 2019

Contents

List of Figures

List of Tables

Acknowledgements

My research mentors: To my advisor, Dr. Matt Simon, thank you for being a constant source of encouragement throughout my Ph.D. and supporting all of my exploration with TrmA, teaching, and consulting. I cannot count the number of times I went into our meetings feeling frustrated with the roadblocks I was encountering, but left with a renewed drive to tackle the next set of problems. I would also like to thank my research committee, Dr. Joan Steitz and Dr. Antonio Giraldez, for your frequent feedback and advice when I reached each major shift in my research. Thank you for always providing fresh perspectives and actionable next steps on the problems I had spent each term struggling with. Many thanks to Dr. Wendy Gilbert and her lab members Cassandra Schaening and Dr. Thomas Carlile for their help and feedback while designing the TrmA substrate library.

My labmates: Science can be a frustrating journey. Thank you all for being such an encouraging and humorous bunch to cushion the blows. I will be forever grateful to have been part of the Simon Lab. Alec, thank you for always having my back when it comes to bad jokes or generally harassing whoever happened to be unfortunate enough to sit in our office. Meaghan and Michelle, I'm sorry. Jeremy, thanks for always support whatever shenanigans were happening in the lab. To Will, thank you for backing Fun Shirt Fridays. To Martin, thank you for your frequent advice after mini-meetings. Those short conversations in the office doorway always helped to carve out the problem I was working on. Maddie, thank you for all of your hard work working out the nitty gritty details of

TrmA pull downs with me that first summer. To Lea and Josh, my rotation students, thank you for your help exploring new avenues and applications for TrmA and your continued feedback throughout the research. To Erin, thank you for paving the way as the "first pancake" from the lab and your guidance throughout graduate school.

My community: To CPC, the community and love I have been a part of here has been a glimpse of heaven on earth. Thank you for your encouragement, teaching, and ceaseless prayers. To Cary and Evan, thanks for always being great friends to drink beers and play Magic with. To the GPSCY staff, thanks for being a second family to Laura and me and being a great place to take a break.

My family: To my dad, Dr. Clay Smith, thank you for being such a wonderful example for me throughout my scientific career and for your training, advice, and feedback through this entire journey. Getting to present and publish our work on BBS5 together will forever be a highlight of my scientific career. To my mom, Jenny, thank you for your constant encouragement and unconditional love, both to me and my growing family. To my siblings, Blake and Madison, thank you for being some of my best friends and such wonderful aunts and uncles. To my son, Jack, thank you for always being excited to see me walk in the door after a long day. Your smile has made this last year of graduate school a true joy and you will always be my favorite alarm clock. Finally, I thank my wife, Laura. Your support throughout the ups and downs of graduate school has been an invaluable constant in an otherwise very uncertain time. Thank you for being such a loving mom to Jack, for your wonderful cooking, for pushing me to never be complacent in any area of my life, and for making me laugh more times than I can count.

Chapter 1

Introduction

Throughout my research, I have been interested in improving techniques to biochemically purify and identify the functions of long noncoding RNA (lncRNA) using a covalent RNA-trap. In order to understand this process, one needs to understand the context of lncRNA biology, current efforts to purify RNA, and existing RNA modifying proteins.

1.1 Noncoding RNA

1.1.1 The categorization of noncoding RNA

The functions of nucleic acids and RNA in particular have always been less obvious than the more chemically diverse group of proteins and amino acids. The apparent simplicity of RNA and DNA, consisting of primarily four monomeric building blocks each, compared to the complexity of proteins, consisting of twenty canonical amino acid monomers, directed many biochemists to believe that nucleic acids were too simple to perform many biological functions. Evidence suggesting that DNA (and not proteins) carries genetic information emerged in 1944, hinting at more important roles for nucleic acids[1]. While DNA was determined to be the critical archive of genetic information, roles for RNA remained mysterious until the discoveries of rRNA[2], tRNA[3], and mRNA[4] in the following decades. After these discoveries, RNAs were categorized into two major categories – protein coding mRNA and RNA that was transcribed but did not contribute to protein coding[5], noncoding RNA (ncRNA). This broad category of ncRNA contains RNA that

1

functions in many critical biological roles such as protein translation (e.g. rRNA and tRNA) and regulation of gene expression (e.g. miRNA, piRNA, siRNA)[6].

1.1.2 Long noncoding RNA

Long noncoding RNAs (lncRNAs) are defined as ncRNA that exceed 200 nucleotides in length[7,8]. Several now-categorized lncRNAs had been studied previously, but it was the beginning of high throughput sequencing and cDNA analysis methods such as FANTOM[9,10] and RIKEN[11] that illuminated the surprising number of lncRNA transcripts in higher-order eukaryotes[12]. While a few lncRNA species are well studied, the category as a whole, partly due to the large number of lncRNAs and their arbitrary definition, remains largely unstudied. Two of the lncRNAs that are best characterized regulate dosage compensation (e.g. *roX2* in *Drosophila melanogaster*[13] and *Xist* in mammals[14,15]) . Another example is 7SK ncRNA, which modifies the activity of transcription factors[16,17]. While the functions of most lncRNA remain poorly understood, most models of lncRNA function assumes their function require interactions with proteins, for example through sequestering protein factors as a decoy, scaffolding large protein complexes, or guiding proteins to specific sites in the cell[6,18].

1.1.3 RNA conformation and structure

RNA is different from DNA because it is mostly produced in a single stranded form and has an additional 2'-hydroxyl that can engage in hydrogen bonding and metal chelation. These differences underlie RNAs ability to adopt complicated secondary and higher order

structures that are generally driven by base pairing[19–21]. Highly structured RNA motifs play important roles in RNA function through generating environments conducive to catalysis (such as in the ribosome) and structures that recruit various protein cofactors to form functional complexes (such as in the spliceosome)[22]. Many techniques have been developed to determine RNA conformation and structure (SHAPE[23], DMS Mapping[24,25], Targeted Structure-Seq[26]). These techniques leverage properties such as solvent accessibility and 2'-OH flexibility to infer properties of RNA on a transcriptome scale, which can be collected using sequencing information.

1.1.4 RNA-protein interactions

RNA is ubiquitously bound by proteins. Even cotranscriptionally, RNA is bound by proteins responsible for RNA processing, localization, export, or other biological functions[27]. The basis of these interactions revolves around the use of modular RNA binding motifs[28]. The most common of these is the RNA recognition motif (RRM), which is found in all kingdoms of life[29,30]. The RRM binds single stranded RNA sequence specifically, such as those occurring in the U1A spliceosomal proteins, which recognizes a AUUGCAC sequence[31]. While many RNA binding proteins act by sequence recognition, many also bind structural elements. For example, the phage coat PP7 protein binds an operator hairpin containing a specific sequence in the loop region[32]. Another structure-based interaction occurs in the widely used Cas9. The endonuclease has a conserved arginine cluster, which binds the variable guide region of guide RNA and drives the structural changes necessary for cleavage[33]. RNA-protein interactions play a critical biological role and are incredibly

3

abundant. Just proteins containing RRMs represent 0.5-1% of human genes[28,34] and understanding the role of various RNA-protein interactions is critical to fully understanding the biological role of many RNAs.

1.2 RNA-protein copurification

In order to probe the biological functions of a lncRNA, one approach is to biochemically purify the RNA of interest with any protein cofactors. As proteins are generally better studied and understood than lncRNAs, the identity of co-purifying proteins is likely to provide insight into lncRNA function and mechanism. As is frequently done to identify factors that copurify with proteins, one approach is to tag the RNA of interest, and use this tag to selectively enrich the RNA with associated factors by capturing the tagged molecule on a solid support such as a resin. Unfortunately, when compared to the strength and variety of affinity tags available for protein study (e.g. K_D of 60-200 pM for antibodies[35], 10 fM for streptavidin-biotin[36], and 50 nM for 6-His tags[37]), the affinities of available RNA tags are considerably weaker (e.g. K_D of 70 nM for the streptavidin binding motif[38], 1 nM for MS2 coat protein[39], and 1 nM for phage coat PP7[40]). In addition to the difficulties associated with weaker affinity tags, biochemical purification of lncRNA is complicated by the low abundance of many lncRNAs (some as low as 2-4 copies per cell[41]), which necessitates very high levels of enrichment to be able to detect the co-purifying molecules over the background contamination in the experiment. Another challenge is that some biologically relevant RNA-protein complexes dissociate rapidly and do not

remain bound to each other during handling, further reducing signal from these purifications. This dissociation can be overcome by crosslinking samples with UV or formaldehyde before lysis, but these treatments complicate sample handling and can create biases in sample capture[42,43]. Lysis of cells during crosslinked or non-crosslinked sample preparation also allows for nonbiological interactions to occur in the sample, increasing the noise in any purifications of RNA[20]. While these nonspecific interactors are not crosslinked to tagged RNA in such an experiment, the ability of researchers to remove these contaminants is limited by the wash conditions that are tolerated by the affinity tags. Despite the current difficulties, capturing RNA remains a significant interest in the field with many advances to be made. Technologies are continuing to be developed to bring the toolkit available to RNA capture in line with the well-developed protein capture resources.

1.3 RNA modifying proteins

1.3.1 The biological role of RNA modifications

RNA modification occurs extensively throughout all forms of RNA. At least 110 types of RNA modifications have been reported[44], 93 of which have been found in tRNA[45]. The high abundance and conservation of the location of these modification in structurally functional RNA such as tRNA and rRNA indicates that modified ribonucleotides present an advantage for highly specialized structures and chemical properties that these molecules require[46]. Even within mRNA, a class of molecules where their roles are thought to rely more on primary sequence rather than structure and chemical reactivity, at least

13 types of modifications have been reported[47]. Some of these modifications, such as the ubiquitous 7-methylguanasine in mRNA 5' caps, stabilize transcripts and protect against degradation[48]. Other mRNA modifications such as 6-methyladenosine are thought to destabilize transcripts[49] and alter protein translation of modified transcripts[50].

RNA modifications on rRNA and tRNA are found on structurally and functionally important regions of the molecules. Two of the most common modifications of these RNA are the modification of uridine to either 5-methyluridine (m^{5}U) or pseudouridine (Ψ). These modifications have been shown to significantly affect mRNA translation and protein synthesis[19,51,52]. In *E. coli* m^{5}U modification occurs in one position in tRNA (U54, located in the loop of the T-arm) and two positions in the 23S ribosomal RNA (U747 and U1939). While these three modifications are installed by different modifying enzymes (TrmA, RumB, and RumA respectively)[53,54], the mechanisms of m^{5}U modification is highly conserved, with all three enzymes using a structurally conserved active site to perform the same chemical steps. While crystal structures are only available for TrmA and RumA, all three enzymes utilize S-adenosylmethionine (SAM) as a methyl donor for chemical reactions. Sequence alignments show that all of these reactions occur within a SAM methyltransferases protein fold[55].

TrmA shares significant mechanistic similarity with the yeast cytosine methyltransferase Trm4. The chemical mechanisms of both of these proteins proceed through a covalent intermediate with the substrate nucleobase[56,57]. Each methyltransferase first activates its

substrate uridine through a Michael addition of cysteine to carbon 6. A methyl group is then transferred to the substrate from a SAM cofactor to form a covalently bound RNA-protein intermediate. This covalent intermediate is released by β-elimination by Glu-358 in TrmA and Cys-360 in Trm4.

1.3.2 Covalent RNA-protein traps using RNA modifying proteins

In both TrmA and Trm4, researchers have removed the basic residue responsible for β-elimination. These mutations trap the enzyme-RNA complex in the bound form, creating a denaturation resistant RNA-protein complex[58,59]. Trapping the proteins in a covalent intermediate has been used to study the interactions of TrmA with its substrate RNA in the intermediate form and to better understand the mechanism of methylation[58]. While this is a very interesting use of this mutant, creating a covalent RNA-protein trap could be applied in many fields outside of crystallography to generate useful tools to study RNA-protein interactions.

1.3.3 Engineered RNA-protein interactions

Because of the biological importance of RNA-protein interactions, the ability to engineer and manipulate these interactions has been the topic of much research interest. In order to be able to manipulate RNA-protein interactions in a predictable way, it is often helpful to understand the processes by which specificity is created in the RNA binding protein[60]. One simple approach to study RNA-protein interactions has been to focus on the RNA side of interactions. Because RNA contains easily readable information content, techniques such as systematic evolution of ligands by exponential enrichment (SELEX) can be used to

find RNA aptamers that have a desired property, such as protein binding[61]. On the protein side of this engineering, work has focused around using structural information to generate point mutants of proteins in residues interacting with RNA. By screening a wide range of point mutations, researchers are able to create a variety of different binding strengths and specificities, which can be used to further alter RNA-protein interactions *in vivo*[62–64].

1.4 Overview

My doctoral research focused on two separate protein engineering studies of TrmA: altering the substrate specificity of TrmA and using TrmA to covalently capture RNA for the purpose of tagging and capturing RNA. My first aim was to establish TrmA as a tool to covalently bind a specific RNA sequence based on the T-loop of tRNA, hereafter referred to as the TrmA-substrate RNA tag (TSR-tag). Due to the highly conserved nature of tRNA and modifying enzymes, the TSR-tag is modified by endogenous methyltransferases, and therefore a TSR-tag corresponding to the wild-type tRNA T-loop sequence is unlikely to be ideal. Furthermore, capturing the TSR-tag with a version of TrmA that retains its native substrate preference is likely to introduce complications as TrmA can bind cellular tRNA which is highly abundant. Combined, these two factors have the potential to substantially reduce signal and increase noise in enrichment experiments, particularly if the target RNA has few copies per cell. With the assistance of Madeline Zoltek, I sought to create a TrmA-RNA substrate pair that would be orthogonal to the natural TrmA-tRNA specificity. The reengineering of TrmA to recognize new RNA substrates described in Chapter 2. In parallel

to these efforts, I sought to establish TrmA as a capture reagent to capture TSR-tagged RNAs in two different formats. First, in Chapter 3 I describe my efforts toward capturing TSR-tagged RNAs using immobilized TrmA on solid support. To accomplish this, I expressed TrmA variants that include a peptide sequence that is biotinylated and applied these biotinylated proteins to a streptavidin containing resin, thereby creating a resin for covalent capture of RNAs that have TSR-tags. Second, I describe the capture of the abundant lncRNA, 7SK, using this RNA-binding resin, as well as capturing 7SK that is bound by TrmA in cells. While I was unable to capture RNA-protein complexes as I had hoped, in Chapter 4 I describe preliminary results for applications of TrmA in RNA capture, transcript stabilization, and RNA subspecies enrichment as well as the limitations and future considerations for applying TrmA to such experiments. The combination of a reengineered TrmA and a method for applying TrmA to covalently bind RNA transcripts of interest lay a foundation for a generally applicable, highly specific, and high stringency capture of RNA, which can be used to study many lncRNA and associated proteins by future researchers.

Chapter 2

Reengineering a tRNA methyltransferase to covalently capture new RNA substrates

This chapter is an excerpt from:

2.1 Author contributions

I performed all experiments, with assistance from Madeline Zoltek in designing the library of substrates for high throughput study of TrmA* activity.

2.2 Summary

I describe the irreversible reaction of TrmA E358Q (TrmA*) with specific RNA sequences leading to a covalent TrmA*-RNA adduct. To make this reaction useful as a tool, I sought to engineer the substrate specificity of TrmA* to react with RNA sequences other than the T-arm of tRNA. By using an alternative RNA sequence, these RNAs can avoid being modified by endogenous methyltransferases when expressed in cells. I describe approaching this problem using rational design to generate TrmA mutants with altered

substrate specificity. While the RNA-substrate reactivities were not easily predicable, I was able to use a library of 2000 RNA substrates to discover a TrmA* mutant (E49R R51E) and RNA substrate (C56A A58G C60U) pair where the RNA substrate is specific for the mutant but not wild-type TrmA*.

2.3 Introduction

Cellular RNAs can be regulated by enzymes that covalently modify specific RNA substrates. Altering the RNA specificities of these enzymes can provide powerful tools to study and manipulate cellular RNAs. The specificity of enzymes can be controlled through programing with guide RNAs, including the C/D box small nucleolar RNAs that direct O2'-methylation[65,66], miRNAs that can direct Argonaut proteins to cleave mRNAs[67–71] and CRISPR-Cas9, which can be directed to cleave RNA[72] as well as DNA[73]. Most RNA-modifying enzymes, however, do not use RNA-guided strategies and instead recognize their substrates through contacts at their protein-RNA interface[74–78]. Compared with protein-DNA and protein-protein interfaces, the engineerability of these protein-RNA interfaces is less well explored, and therefore it is unclear how easily these RNA modifying enzymes can be re-purposed. While there are abundant encouraging examples where RNA-binding proteins have been re-engineered to bind specific RNAs (e.g., PUFs[79–81], zinc fingers[82]), the properties of enzyme-RNA interfaces can be more highly constrained[83] raising additional challenges. To explore whether I could re-engineer the specificity of an RNA-modifying enzyme, I first examined TrmA, a tRNA methyltransferase that methylates uridine-54 of the T-arm of tRNA – a 17-nucleotide stem loop that is highly structurally

conserved, but has a variety of base identities within the stem-loop even within the tRNA population of a single species[84]. I performed both rational design and high-throughput screening to identify a mutant version of TrmA that can covalently capture a non-tRNA 17-mer RNA substrate that is not recognized by the wild-type enzyme.

I focused my study on TrmA because its structure, activity, and mechanism are well studied[56,58,85,86]. While TrmA recognizes its substrates through a combination of RNA-conformation and base-specific contacts, it has been demonstrated to have high tolerance for substrate mutations that do not significantly alter substrate structural conformations[58,85,87]. TrmA forms a covalent intermediate with its substrate RNA (**Scheme 2.1**)[56]. By using mutagenesis to remove the basic residue (glutamate-358) responsible for resolving this intermediate, the resulting TrmA mutant (hereafter TrmA*) can be trapped in a covalent complex with its substrate. This covalent complex has been used to trap the protein-RNA complex for structural studies[58] and similar reactivity of TrmA homologues (NSUN) has been used to capture and identify RNA substrates[78]. I reasoned that trapping the TrmA* covalent intermediate could be useful to more generally covalently tag RNAs, and this utility would be enhanced if the RNA substrate specificity TrmA* could be re-engineered to recognize an alternative sequence other than the highly conserved tRNA stem loop.
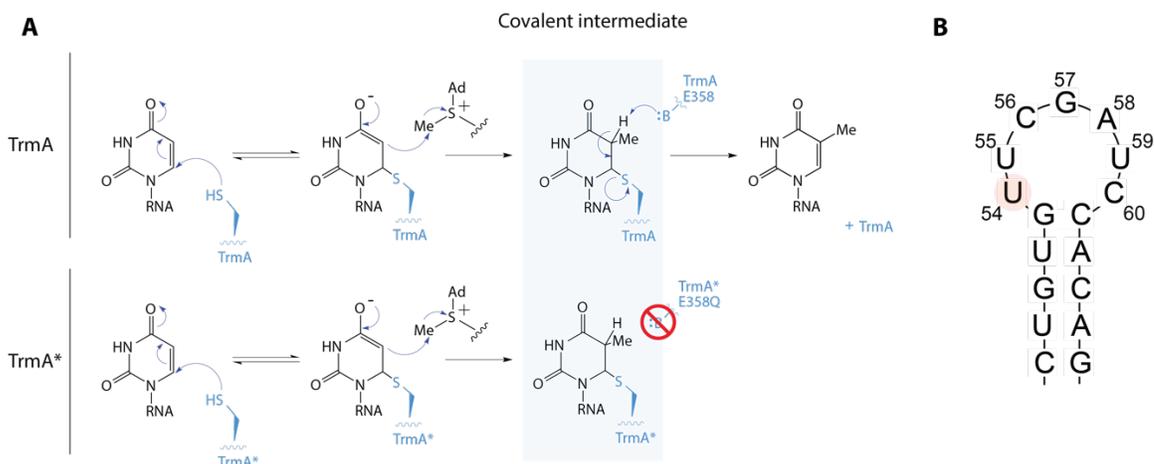
Scheme 2.1 (B) TrmA first activates its substrate uridine-54 through a Michael addition of TrmA cysteine-324 to carbon 6 of U54 to form a covalently bound RNA-protein intermediate. A methyl group is then transferred to carbon 5 of U54 from an S-adenosyl methionine cofactor. The covalent intermediate is released by β elimination by TrmA glutamate-358 (E358)[56]. By creating TrmA mutant E358Q (TrmA*), the basic residue responsible for β elimination is removed, trapping the covalently bound RNA-protein intermediate. (B) TrmA minimal substrate RNA (TSR) sequence (T-arm of tRNA^phe)[53].

2.4 Results

2.4.1 TrmA* covalent capture and rate measurements

In order to explore the utility of TrmA* as a covalent RNA binding protein, I first sought to characterize the ability TrmA* to form covalent complexes with a minimal substrate loop, the T-arm (nt 48-66) of bacterial tRNA^Phe, hereafter referred to as the TrmA substrate RNA (TSR). N-terminus-hexahistidine-tagged TrmA* was expressed from *E. coli* and affinity purified using nickel-NTA resin. Using a synthetic 5' Cy3-labeled TSR, TrmA* reactivity was measured by observing the retarded mobility of fluorescent TSR by urea-PAGE. To control for nonspecific or noncovalent binding, I added a 5' Cy5-labeled TSR with

the mutation U54C (TSR-U54C) to the same reaction mixture as an internal control. By reacting TrmA and TrmA* with equimolar TSR and TSR-U54C I confirmed that TrmA* binds TSR covalently and specifically (**Figure 2.1A**). In order to characterize the activity of TrmA*, the protein was incubated with an excess of fluorescent TSR over time and separated by urea-PAGE (**Figure 2.1B**). Measuring the fluorescence intensity of the shifted RNA and total sample allowed me to determine the fraction of RNA bound for each timepoint and to calculate the apparent second order rate constant (k_{obs}) of the TrmA* reaction (**Figure 2.1C**) (**Table 2.1**). These values were fit from the data using the fluorescent intensity of the shifted vs unshifted bands as described in **Appendix A**.

Protein	TSR	G57C	C60G	C56A A58G C60U
TrmA*	7.69	0.32	1.39	$< 5 \times 10^{-5}$
	(6.52, 9.05)	(0.30, 0.35)	(1.06, 1.77)	
E49R	0.56	0.39	0.89	$< 5 \times 10^{-5}$
	(0.54, 0.65)	(0.35, 0.43)	(0.64, 1.17)	
R51E	7.10	0.16	3.27	0.10
	(5.74, 8.83)	(0.15, 0.18)	(2.58, 4.18)	(0.09, 0.11)
E49R R51E	3.18	0.25	1.73	0.84
	(2.51, 4.00)	(0.23, 0.26)	(1.19, 2.45)	(0.78, 0.90)

Table 2.1: Measured values of k_{obs} ($nM^{-1}s^{-1}$) for TrmA mutants and TSR substrates (95% CI)
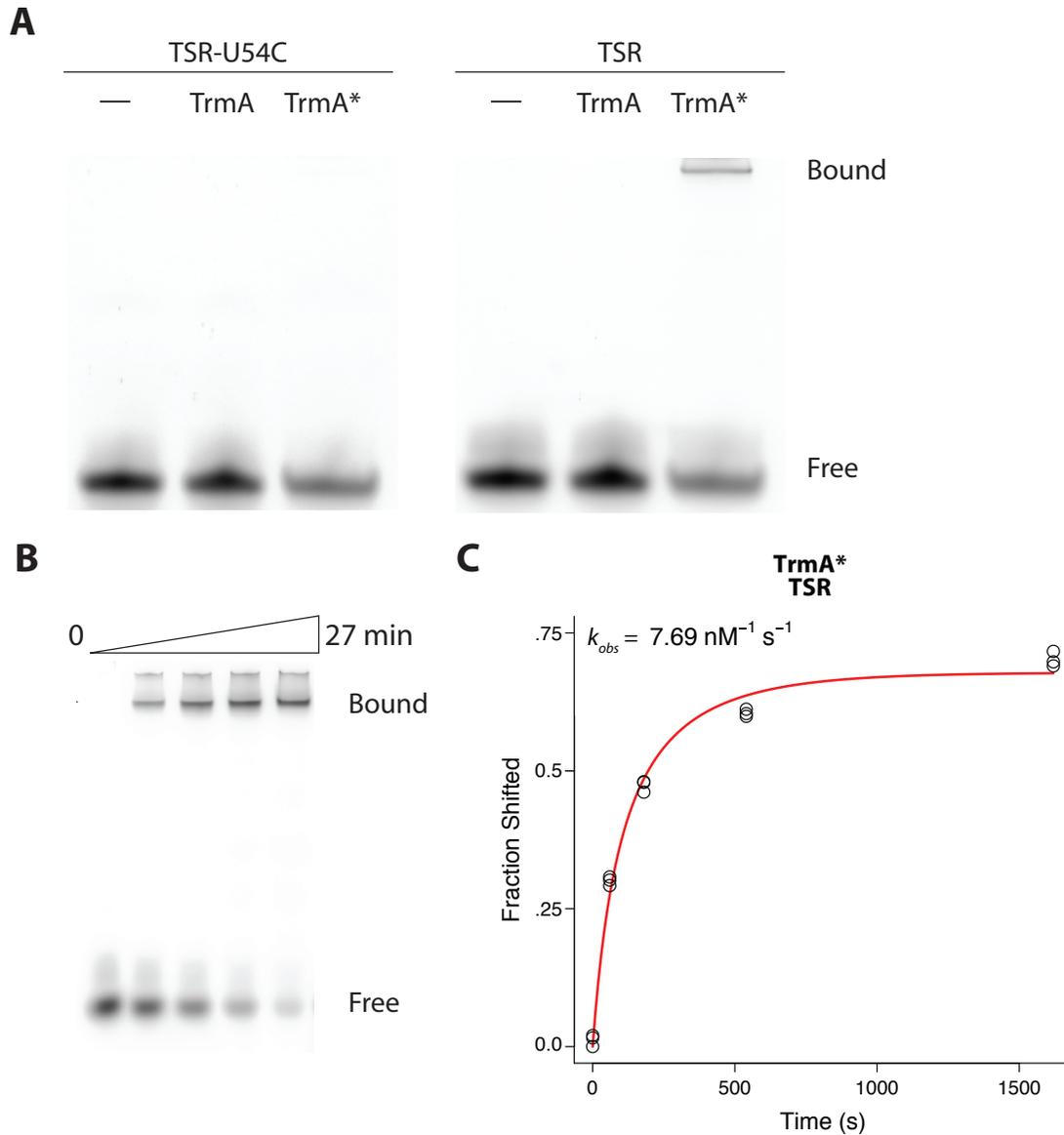
Figure 2.1: **TrmA* binds substrate RNA (TSR) covalently**. (A) Urea-PAGE separated reactions of no protein, TrmA, and TrmA* reacted with equimolar fluorescently labeled TSR and TSR-U54C RNA. (B) Urea-PAGE separated reactions of TrmA* with 20 pmol TSR RNA for 0, 1, 3, 9, and 27 minutes. (C) Triplicate data from (B) fit by non-linear regression to determine the rate of TrmA* covalent binding of TSR RNA, 7.69 $nM^{-1}s^{-1}$.

2.4.2 TrmA* specificity can be altered by rational design

In order to develop a TrmA*-TSR pair that is orthogonal to the TSR sequence of the T-arm of tRNA, I first searched for related methyltransferases that already possessed alternate RNA substrate specificity. The binding pocket of TrmA has significant structural homology to the rRNA uridine methyltransferase RumA[54,88,89]. Furthermore, each of these enzymes modify RNA stem-loops with a seven-membered loop. Examining a structural alignment of the enzyme-substrate pairs[55,58,89] reveals important differences. Hydrogen bonds between Glu-49 of TrmA and G57 of tRNA[Phe] stabilize the flipped conformation of U54 into the active site of TrmA. In RumA, these interactions are replaced with hydrogen bonds between Arg-132 of RumA and C1942 of its rRNA substrate (**Figure 2.2A**). I hypothesized that mutating both the RNA substrate and the TrmA* enzyme to match the interactions found in RumA (TrmA* E49R and TSR-G57C) would generate a protein-substrate pair with orthogonal reactivity. TrmA* E49R demonstrated reduced activity, but consistent with my hypothesis, TrmA* E49R shows increased reactivity with TSR-G57C versus TSR, resulting in a nearly 10-fold change in specificity from TrmA*. While TrmA* is still reactive with TSR-G57C, its reactivity is significantly reduced, particularly in the presence of competing TSR RNA (**Figure 2.2B**). While these results are encouraging, TSR-G57C is still a reasonable substrate for TrmA*, demonstrating incomplete reengineering of the protein-RNA binding interaction to create an enzyme that reacts with an RNA substrate that will not be modified by endogenous TrmA-family methyltransferases.
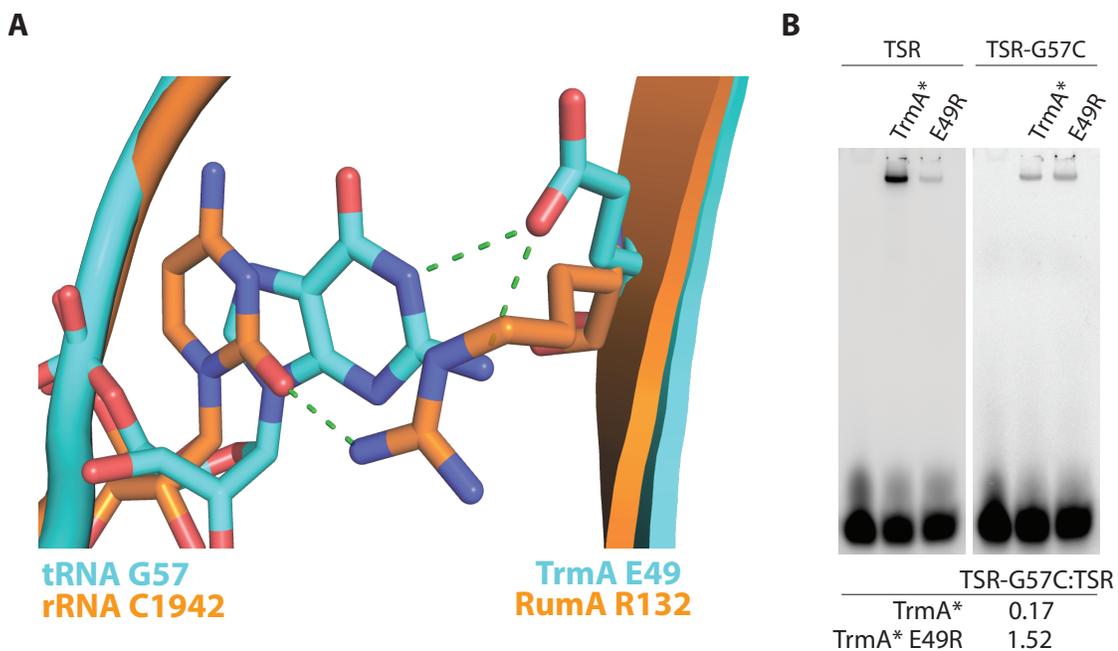
Figure 2.2: **TrmA* binding specificity alteration by rational design.** (A) TrmA Glu-49 interactions with tRNA G57 (cyan) superimposed over rRNA methyltransferase, RumA, Arg-132 interactions with 23S rRNA C1942 (orange). (B) Top: Urea-PAGE separated reactions of no protein, TrmA*, and designed TrmA* E49R reacted with excess equimolar RNA mixtures of TSR and designed TSR-G57C. Bottom: ratio of the fraction bound of TSR-G57C and TSR for TrmA* and TrmA* E49R

Next, I examined the TrmA-RNA structure and identified other base-specific interactions that might provide a handle for changing TrmA substrate specificity. One site of interest was the hydrogen bonding interaction between Arg-51 of TrmA and C60 of tRNA[Phe]. I pursued an analogous approach as the design of TrmA* E49R and TSR-G57C described above; I hypothesized that TrmA* R51E would remove a critical hydrogen bond that could be rescued with C60G mutation in the TSR RNA substrate. Consistent with my hypothesis, TrmA* R51E was more reactive than TrmA* with TSR-C60G in isolation. However, TrmA* R51E did not exhibit preferential binding to TSR-C60G over TSR and TrmA* still

demonstrated significant reactivity with C60G (**Table 2.1**). I conclude that I was able

engineer TrmA* mutants with altered TSR specificities, but that the TSR sequences from

rational design retained significant activity with wild-type TrmA* suggesting they would

be substrates for endogenous methyltransferases. Next, I sought RNA substrates that are

reactive with TrmA* mutants but not wild-type TrmA*.

2.4.3 TrmA specificity can be determined using high-throughput screening*

In order to investigate a wide spectrum of RNA substrates for TrmA* and mutants, I

sought to create a higher-throughput screen for probing TrmA* binding specificity. While

likely substrates would retain some features of the TrmA* TSR, it is possible that multiple

mutations could together shift specificity toward a TrmA* mutant and away from the

wild-type enzyme. As there are 540 different possible triple mutants possible starting

from a 6-mer RNA loop, I estimated that identifying such RNAs would require screening

hundreds-to-thousands of well-defined substrates. Such screens have been previously

developed to probe the substrate specificity of pseudouridine synthase enzymes[90] using

an array of synthetic DNA substrates that can be converted to RNA using in vitro

transcription. Based on this precedent, I developed a synthetic DNA library encoding 1955

unique TSR mutants including all single, double, and triple mutants of positions 55-60 as

well as all quadruple and quintuple mutants of positions 55-59 (**Appendix B**).

Furthermore, I included 27 positive control and 20 negative control sequences each with

unique barcodes. This duplication was designed to prevent biases created by barcodes or

stochastic effects in library preparation.

To simultaneously screen these libraries, I developed a sequencing-based assay (**Figure 2.3**). Covalent binding of TrmA* creates a peptide adduct on the modified uridine. Even after proteinase K degradation of TrmA*, this adduct blocks reverse transcriptase (discussed further in **Chapter 4**). Reverse transcription and sequencing of TrmA substrates would be complicated by our inability to reverse transcribe past this adduct. To circumvent this challenge, Madeline Zoltek and I designed the substrate library and included unique 12-nucleotide barcodes to identify each variant. This barcode is located on the 3' side of each substrate loop and at a location remote from where TrmA* reacts. These barcode regions can be reverse transcribed, amplified and sequenced in order to determine the identity of bound RNA species without having to reverse transcribe past the substrate loop. The synthetic substrate library was *in vitro* transcribed, purified, and reacted with TrmA*, TrmA* E49R, TrmA* R51E, double mutant TrmA* E49R R51E, and an unreactive mutant of TrmA* where the reactive cystine is mutated to alanine (TrmA* C324A). After the reaction with TrmA* mutants, for either 30 seconds or 30 minutes, RNA samples were purified and reverse transcribed.
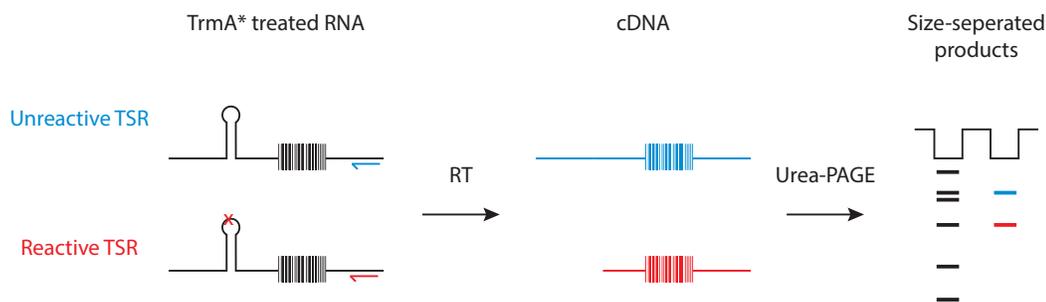
Figure 2.3: **TSR library screen workflow.** *In vitro* transcribed TSR libraries (schematized as a stem loop with cartoon barcode region) are reacted with TrmA* mutants and then reverse transcribed. TrmA* covalent attachment blocks reverse transcription, creating two different sizes of cDNA corresponding to unreactive and reactive TSR substrates, which can be separated by urea-PAGE. Separated populations can then be cut from gels and prepared for sequencing.

To identify those sequences that acted as substrates versus those that did not react, I used the property that the TrmA* reaction leads to a well-defined, truncated reverse transcription product. The cDNA product of the reactive sequences is shorter and can be separated from unreactive species by urea-PAGE. Full-length and truncated cDNAs were sequenced as were the input RNA libraries.

Sample reads were demultiplexed and perfect matches of barcode reads were counted and assigned to corresponding TSR loops. All of the active variants of TrmA* with the exception of TrmA* E49R showed enrichment of all 26 positive controls and did not enrich any of the 20 negative control samples (**Figure 2.4A**). Using DESeq2 to compare reacted and unreacted TSR libraries, substrates more highly represented in bound samples with p < .05 were labeled as enriched. From these enriched samples, I determined the sequence motifs enriched and depleted from each binding reaction. The results from TrmA* are very similar to the naturally occurring T-arm substrates of TrmA* (**Figure 2.4B**) (TrmA substrates determined by aligning all known T-arm tRNA sequences[84]). Furthermore, while TrmA* E49R reactive RNA demonstrates similar trends, consistent with my earlier experiments I found that TSR C57 is enriched with E49R relative to wild-type TrmA*.
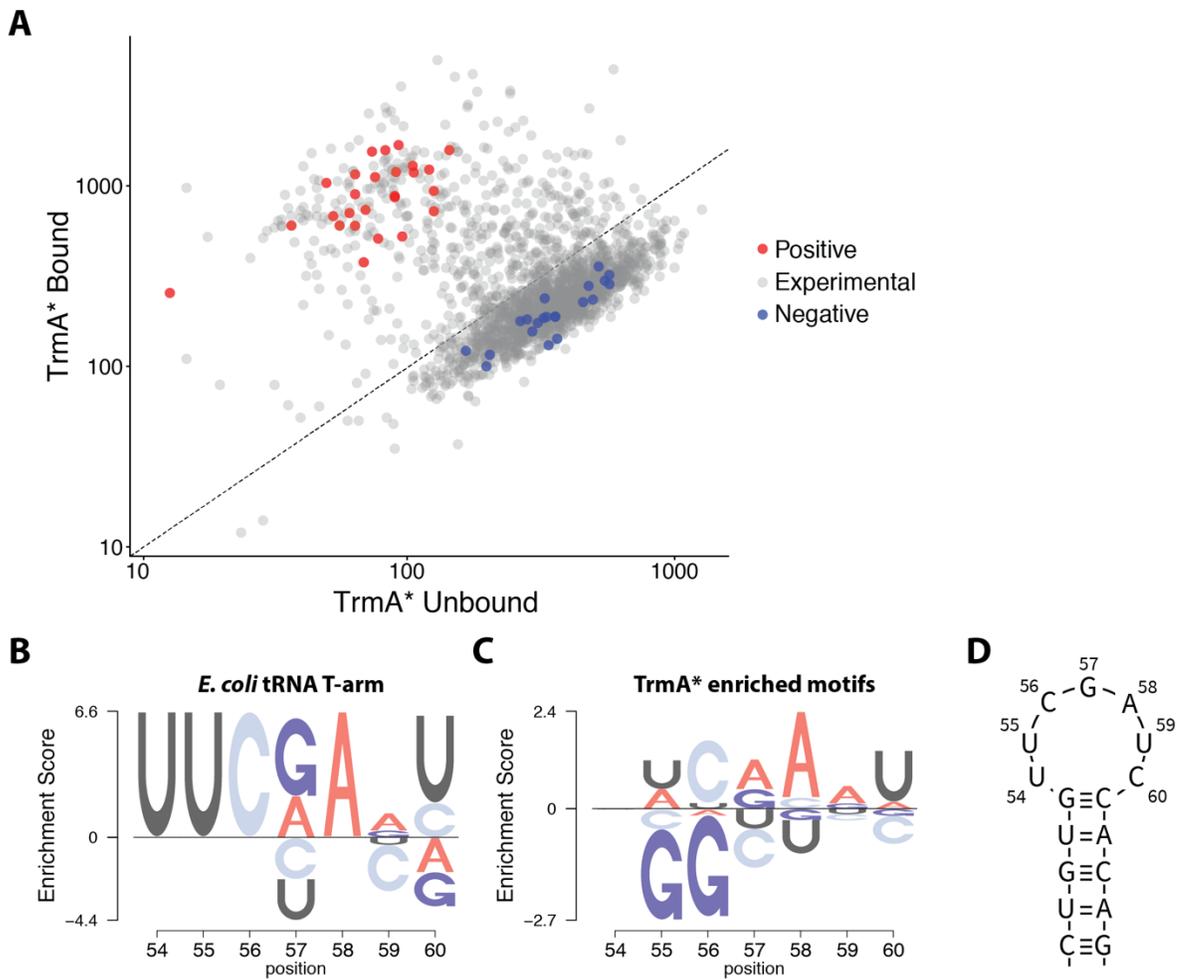
Figure 2.4: **TrmA* binding of 1955 TSR mutants in vitro shows enrichment of tRNA-like substrates**. (A) Scatterplot of TrmA* bound versus unbound TSR sequencing read counts. Red: Positive control tRNA[phe]-like substrates. Blue: tRNA[phe]-U54C negative control RNA substrates. (B) Consensus motif of *E. coli* tRNA T-arm substrates of TrmA[84]. (C) Consensus motif of 165 TrmA* enriched TSR mutants. (D) tRNA[Phe] T-arm stem loop sequence.

2.4.4 TrmA* E49R R51E has novel specificity for TSR C56A A58G C60U

As part of the analysis I searched for RNA sequences that were uniquely enriched by a TrmA* variant but not reactive with wild-type TrmA*. The triple mutant C56A A58G C60U was enriched by TrmA* E49R R51E in both 30-second and 30-minute long reactions, but was not significantly enriched in any of the other TrmA* variants. Previous work has

shown that each of the single mutations present in this substrate are reactive with TrmA, but have significantly reduced reactivity[87]. If these mutants have combinatorial effects, reactivity with TrmA* would be dramatically reduced. In order to confirm that this finding was not an artifact of sequencing or sample handling, I used a synthetic fluorescent substrate loop containing these mutations and a 5'-Cy5 label, TSR-TM. In my analysis of TrmA* mutants, TrmA* and TrmA* E49R showed no observable reactivity with TSR-TM. TrmA* R51E displayed low activity and TrmA* E49R R51E showed moderate activity as predicted from high-throughput screen (**Figure 2.5A**). TrmA* E49R R51E exhibits a greater than 10,000-fold higher rate of binding TSR-TM than TrmA* (**Figure 2.5B**) (**Table 2.1**).
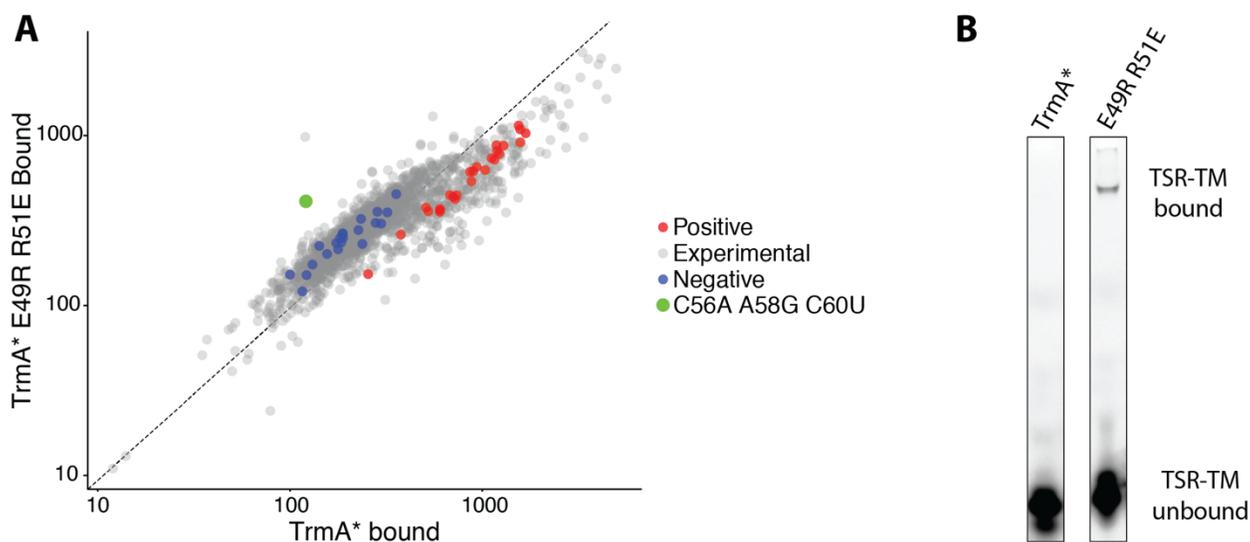
Figure 2.5: **TrmA* E49R R51E has specificity for RNA substrate C56A A58G C60U (TSR-TM) that is significantly different from TrmA*.** (A) Scatterplot of TrmA* E49R R51E bound versus TrmA* bound TSR RNA populations. (B) Urea-PAGE separated reactions of TrmA* and TrmA* E49R R51E reacted with excess TSR-TM RNA.

2.5 Discussion

The TrmA*-RNA substrate interactions are engineerable, with both predictable and surprising changes to substrate specificity. Using rational design and structural homology to related enzymes, I created a TrmA* mutant with altered substrate specificity (TrmA* E49R). Another structure-based approach to redesign TrmA* yielded a mutant that maintained the same specificity as wild type. Combining the rationally-designed mutants and high-throughput screening of RNA substrates, I identified an RNA substrate that reacts exclusively with non-wildtype variants of TrmA*. TSR-TM binding by TrmA* is dramatically reduced compared to TSR, which makes it a very good candidate to explore for use as a covalent capture reagent in tandem with TrmA* E49R R51E. It is possible that this interaction can be further developed to generate a completely orthogonal system in which the TrmA* mutant does not bind TSR or endogenous tRNA substrates, but will likely require high-throughput screening of TrmA* mutants.

Chapter 3

Covalent capture of RNA using an engineered RNA methyltransferase.

3.1 Author contributions

I performed all experiments in this chapter.

3.2 Summary

As described in Chapter 2, using a single point mutation, I am able to trap a covalent intermediate of the uridine methyltransferase and its 17-nucleotide RNA substrate. Using this covalent adduct I am able to conjugate TrmA-fusion protein to RNA both in cells and *in vitro.* By adding affinity tags including biotinylation to TrmA*, this approach has allowed me to efficiently capture tagged RNA and remove off-target RNA background with high stringency, denaturing washes. This method could be generally applied to target and capture any RNA of interest by addition of the 17-nucleotide substrate loop.

3.3 Introduction

Biochemical isolation of specific RNA sequences is valuable for several reasons including the determination of RNA structure and intermolecular interactions. In order to isolate RNA, many techniques have been developed using RNA binding proteins, chemically modified nucleotides, complementary oligos, and small-molecule binding aptamers

engineered into RNA[91–94]. The use of tagging approaches based on naturally occurring interactors, such as RNA-binding proteins, benefits from the binding specificity developed by evolution. However, previously developed approaches are based on relatively week interactions, in some cases, likely due to the requirements for efficient turnover for biological function. In contrast, chemical modifications can provide high strength interactions, but are limited by low specificity or an inability to be used *in vivo* due to low bioavailability. An ideal system would leverage the strength of covalent modification with the specificity of a selected biological system.

tRNA is heavily modified in order to correctly fold, to participate efficiently in translation and to be recognized by aminoacyl-tRNA synthetases[95]. The enzymes that install these modifications recognize specific tRNA folds and sequences to specifically modify tRNA. The T-arm of tRNA is the clover "leaf" 3' of the anticodon loop and contains two conserved types of modification, pseudouridine and methyuridine[47]. This methyluridine is generated by the methylation of uridine-54 by a uridine methyltransferase. In *E. coli* this is reaction is performed by TrmA[53]. This enzyme proceeds through a covalent intermediate with the RNA that can be trapped by a single amino acid mutation[57]. In the wild type protein, the catalytic cysteine attacks the carbon in position six of uridine 54, which subsequently is methylated by S-adenosyl methionine (SAM) on carbon five[56]. After this methylation is completed, the basic glutamate in position 358 abstracts a proton from carbon five and releases TrmA to continue further rounds of catalysis (**Figure 3.1A**). As discussed in Chapter 2 and recreated in **Figure 3.1B** for ease of reading, by mutating the glutamate to a glutamine, the covalent protein-RNA adduct is stable to denaturing gel separation[96].
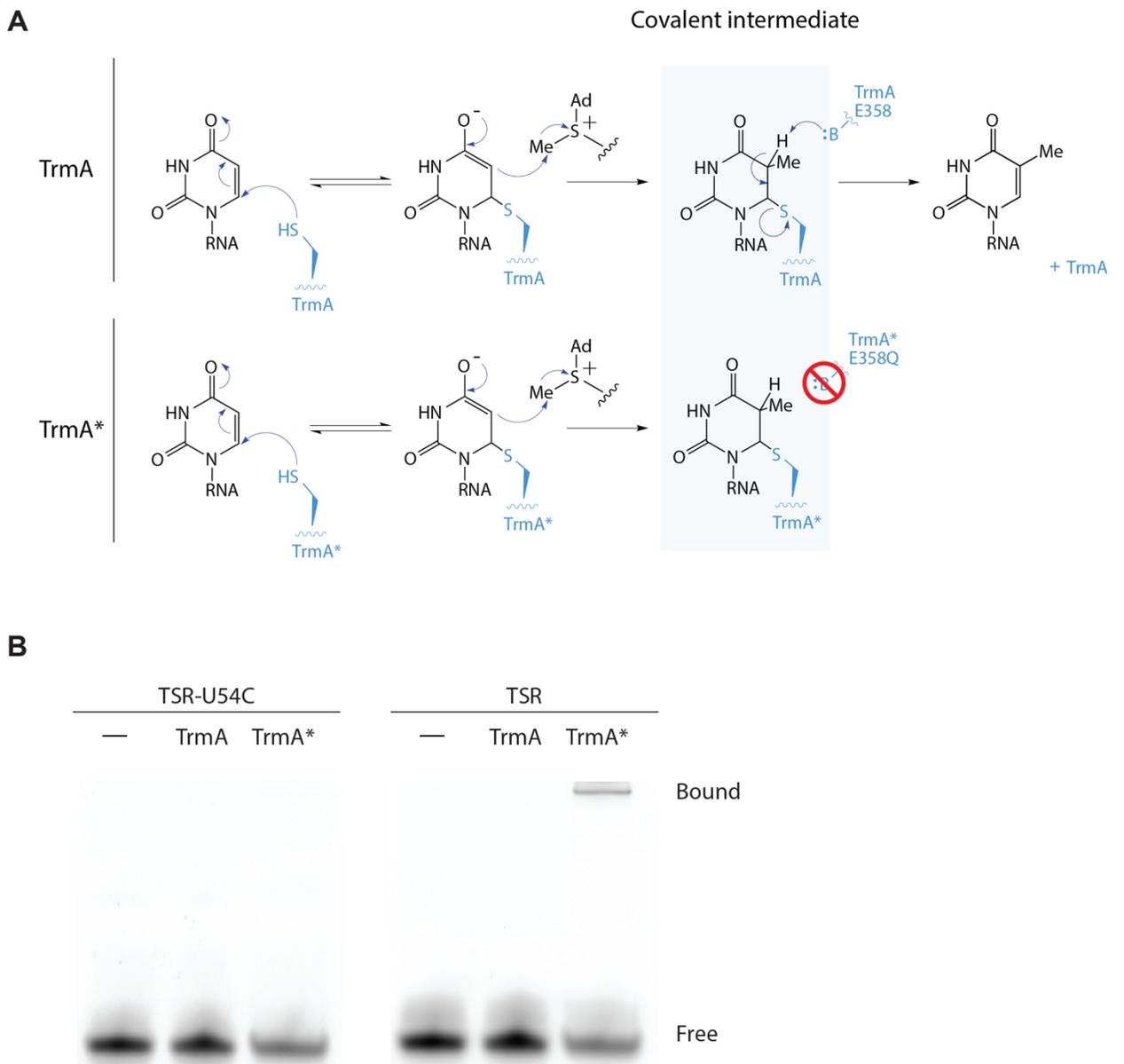
Figure 3.1: **TrmA mutant E358Q (TrmA*) binds RNA covalently.** (A) The mechanism of TrmA reaction proceeds through a covalent intermediate with its substrate. Removal of the basic glutamate 358 by mutation to glutamine traps this covalent intermediate. (B) TrmA simultaneously reacted with fluorescently-labeled TSR and TSR-U54C separated by 12% urea-PAGE. TrmA* captured RNA becomes significantly more retarded than unbound RNA.

One of the RNAs that I studied using TrmA* is the 7SK lncRNA. 7SK forms the core scaffold of the 7SK small nuclear ribonucleoprotein[97]. This RNP regulates transcription by binding and inhibiting CDK9[98,99]. The 7SK RNP is a good complex to study as it has previously been isolated through a variety of techniques both using sucrose gradient separation and affinity purification to study its structure and associated proteins[100]. Like many other ncRNAs, 7SK exists in multiple conformations that can be identified by mapping techniques[101]. In these proof-of-principle experiments, 7SK conformations were separated using antibodies against known protein cofactors, an experiment which would be impossible to perform on an RNA with unknown interaction partners. Other methods such as sucrose gradient purification[102], or even RIP[103] and CLIP[104] are effective for purifying some RNAs, but not generally applicable to less abundant RNA or those without known protein interactors. In order to analyze these types of unstudied RNA, a more efficient, stringent, and specific technique is necessary.

As discussed in Chapter 1, RNA-protein complexes are common and have been extensively studied. Through the modification of nucleoside bases, such as in uridine methyltransferases, these proteins and RNA form covalent intermediates that have been identified and even used as crystallographic tools to study the structural interactions of these types of reactions as is the case with the *E. coli* U54 tRNA methyltransferase, TrmA[58]. Using the covalent trap mutant of TrmA, TrmA*, I hoped to use create a covalent pull-down reagent to capture RNA-protein interactions.

3.4 Results

3.4.1 TrmA* can be used to create a covalent affinity capture resin

Using the covalent capture system for wild-type TrmA* and wild-type TSR characterized in Chapter 2, I sought to develop conditions to enrich tagged RNAs from cell lysate (**Figure 3.2A**). I created a vector (pLPCX) containing 7SK 5'-tagged with the TSR or TSR-U54C sequences used in *in vitro* fluorescent binding assays (**Chapter 2**). Placing affinity tags on the 5' end of 7SK has been shown to significantly increase RNA abundance relative 3' tagging[100]. To further control for potential nonspecific binding of TrmA* to RNA, I used site-directed mutagenesis to create an inactive version of TrmA* by adding a point mutation changing the reactive cysteine to an alanine (C324A). To generate recombinant protein, I coexpressed TrmA* or inactive TrmA* C324A with the biotin ligase BirA in *E. coli*. After purifying each biotinylated protein, I bound 20 µg of each purified protein to 100 µL streptavidin-agarose beads in PBS for 30 minutes at room temperature to create a TrmA*-affinity resin. Following binding, I washed these beads twice with 5 bead volumes PBS + 10 µM biotin to block any unoccupied streptavidin sites and minimize the binding of cellular biotinylated proteins. Finally, I equilibrated the beads with TrmA reaction buffer.

To test if the resin could enrich TSR-tagged 7SK, I incubated each affinity resin with cell lysate from hypotonic-lysed 293T extract (5×10^5 cells) transiently transfected with TSR-7SK or TSR-U54C-7SK. I carried out each reaction for 30 minutes at room temperature with constant agitation. After the reaction, I washed the resin with denaturing wash

buffers (1% SDS and 4M urea) and RNA was eluted by the addition of 40 µg proteinase K

and incubated at 40˚C for 30 minutes with agitation. The eluted RNA was isolated, reverse

transcribed, and analyzed by qPCR (see methods in **Chapter 5** for more details on wash

conditions and cDNA preparation). Primers unique to the transfected, tagged 7SK were

used to determine the extent of capture of different RNA species including 7SK, MALAT1,

and NEAT1. While NEAT1 and MALAT1 are significantly larger that 7SK, both are similarly

abundant to 7SK and serve as controls for off-target binding to the resin (**FIGURE 3.2B**).
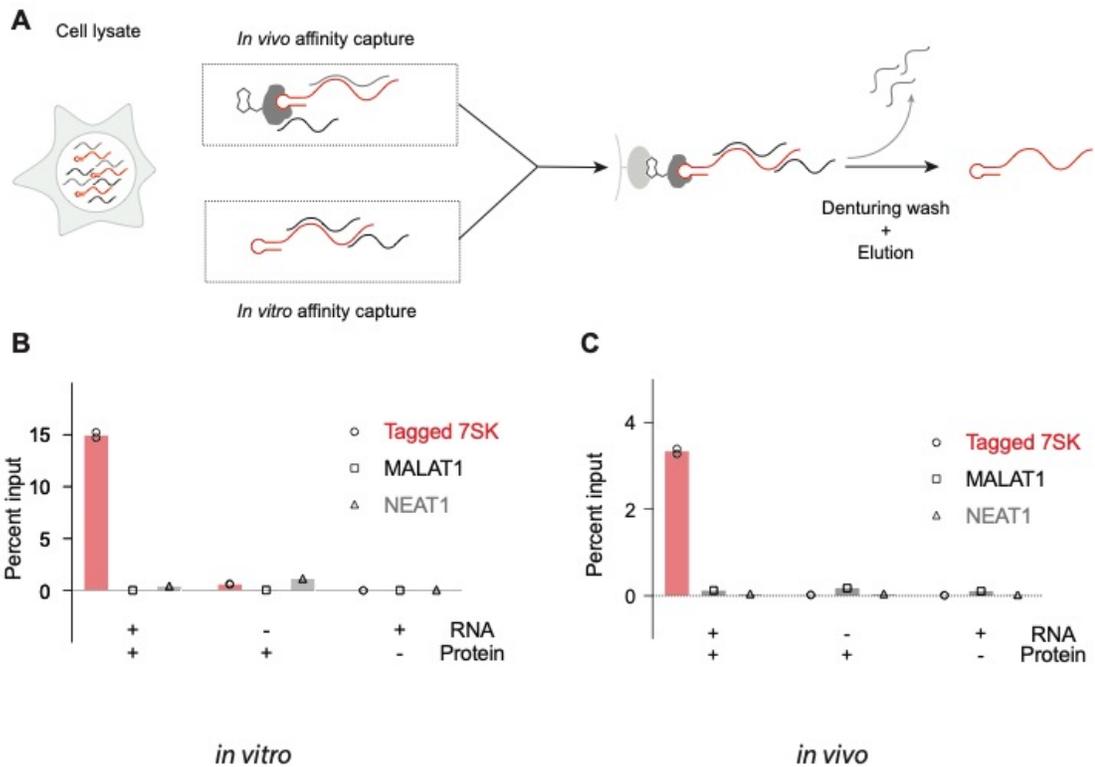
Figure 3.2: **TrmA* can be used to purify RNA from complex mixtures.** (A) Purification scheme for RNA bound by TrmA* *in vitro* or *in vivo*. Because both methods utilize the strength of TrmA* covalent binding and streptavidin-biotin interactions, washing under denaturing conditions (4M urea and 1% SDS) is possible. qPCR quantification of RNA pulled down by TrmA* capture resin (B) or bound by TrmA* *in vivo* and captured on streptavidin agarose (C). Primers specific for tagged 7SK (TSR-7SK or TSR-U54C-7SK) are used to quantify capture efficiency. MALAT1 and NEAT1 are used as background controls as both are abundant lncRNA species.

Through this capture system, I was able to enrich TSR-7SK ~30-fold over MALAT1 and NEAT1. Using this method of capture and stringent washes it is possible to capture RNA efficiently and in a manner that is capable of withstanding very highly denaturing wash conditions. Because TrmA* is covalently bound to TSR-7SK, the wash conditions are limited only by the strength of streptavidin-biotin interaction.

3.4.2 TrmA* can react with TSR RNA in living cells

In complementary work, I explored an alternative strategy where TrmA* was co-expressed with TSR-7SK to allow the reaction to proceed in living cells. By cloning TrmA* into a mammalian expression vector (pCDNA3) that results in TrmA* being expressed fused to a peptide motif that is biotinylated in cells. Successful conjugation leads to RNA biotinylation. 293T cells were transfected using calcium phosphate with TrmA* or TrmA* C324A and TSR-7SK or TSR-U54C-7SK as described previously. Twenty-four hours post transfection, cells were harvested by scraping, washing with PBS, and pelleting by centrifugation. Whole cell extracts were prepared by sonic disruption. This extract was applied to streptavidin-agarose beads and allowed to bind for 30 minutes at room temperature while rotating. After binding, samples were washed with the same series of washes, eluted, extracted, precipitated, and analyzed by qPCR as described previously for *in vitro* capture (**Figure 3.2C**). Similar to the results from *in vitro* capture, this method showed high levels of enrichment of 7SK over MALAT1 and NEAT1.

3.4.3 TrmA* enriches tagged RNA from the entire transcriptome

These strategies show that it is possible to efficiently capture RNA using TrmA* affinity capture. Due to the targeted nature of qPCR, these assays do not comprehensively address the numerous background RNAs that could conceivably be enriched from whole cell lysate. In order to examine the enriched RNAs globally, I prepared TrmA*/TSR-7SK samples as described for in-cell labeling above, but analyzed the resulting RNAs using NEBNext for Illumina library preparation prior to single end sequencing on an Illumina HiSeq 2500. Sequencing reads were demultiplexed, and processed using Cutadapt to remove adapter sequences. Processed reads were aligned to the human genome (hg19) using BowTie2 and sorted with bedtools. Read density was visualized in IGV with bedGraphs generated from bedtools. Examining aligned tracks at the genomic positions for 7SK, MALAT1, and NEAT1 confirms the same enrichment seen by qPCR (**Figure 3.3A**). In order to quantify different RNA species, I calculate the fragments per kilobase of transcript per million mapped reads (FPKM) for each transcript (hg19) using HISAT2. By graphing the FPKM for each transcript comparing the eluted samples to input cellular RNA, it is possible to visualize the enrichment of each RNA transcript (**Figure 3.3B**). In this preliminary analysis, cells transfected with TrmA* and TSR-7SK enriched 7SK more than 100-fold over background transcripts in the eluent relative to input. This suggests that TrmA* can be used to enrich a tagged RNA of interest from the complicated mixture of the cellular transcriptome.
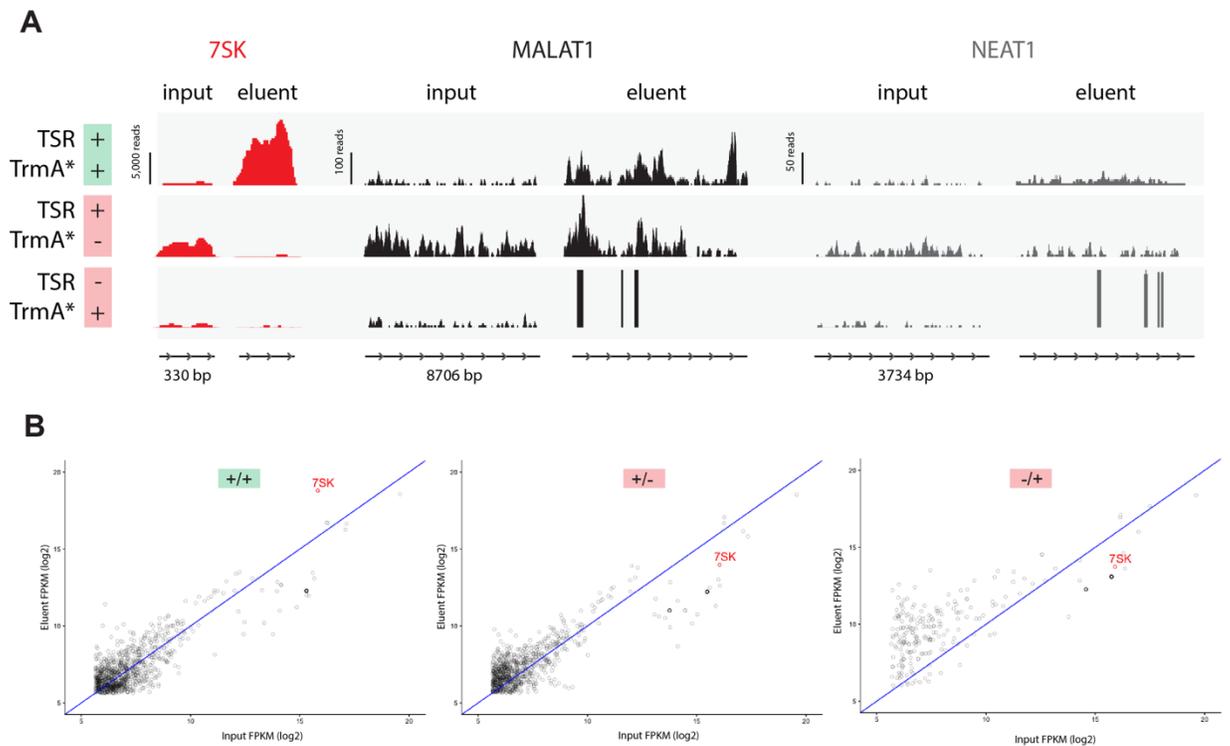
Figure 3.3: **TrmA* specifically enriches tagged RNA from the transcriptome.** TSR-7SK bound by TrmA* *in vivo* was sequenced to determine enrichment and to examine nonspecific background capture. (A) RNA seq tracks of RNA species quantified in Figure 3.2. (B) $\log_2$ FPKM values for each captured sample eluent and input are plotted. 7SK is labeled in red for each eluent/input sample. Eluent = Input is plotted in blue to easily visualize enrichment/depletion.

3.5 Discussion

In summary, I have established a technique for covalently capturing a targeted RNA on beads using a single mutated protein and a minimal 17-nucleotide substrate, which can be performed *in vitro* and *in vivo*. This capture of RNA on beads, which can be washed with high stringency, in principle allows isolation of a single RNA species with high levels of enrichment. The ability to biotinylate RNA using biotinylated TrmA* in cells allows specific enrichment of one species of RNA from a complex mixture. Both of these methods form a baseline for further researchers to develop a covalent capture of RNA-protein

complexes. Future experiments might seek to use RNA spike-ins in order to quantify the total amount of target RNA versus background as opposed to simply looking for relative levels of enrichment, which are less informative of overall background levels. Using crosslinkers such as formaldehyde or UV, it is likely possible to purify proteins bound to a TrmA*-tagged RNA in methods similar to those I have described here. This technique will allow high precision capture and identification of RNA associated proteins by introducing the 17-nucleotide substrate into an RNA of interest.

Chapter 4

TrmA* covalent capture applications and other considerations

4.1 Author contributions

I performed all experiments in this chapter.

4.2 Summary

In the process of conducting my research on TrmA and its ability to covalently capture specific RNA from complex mixtures, I sought to develop several applications for the technology. This chapter describes the preliminary results for several applications of TrmA* and my thoughts on possible roadblocks and considerations for future applications.

4.3 Introduction

TrmA* shows promise for covalently capturing RNA and purifying it from a complex mixture, as demonstrated in Chapter 3. Despite the high level of specificity and stringency of wash conditions available, I encountered challenges in my attempts to use TrmA* to enriched proteins crosslinked to an RNA of interest. I believe that this has been due to several limitations and complexities associated with protein pull down. These limitations include variability in the activity of TrmA* preparations, challenges in crosslink extract

preparation for enrichment studies, challenges in developing assays sensitive enough to observe enrichment and low yields of enrichment from crosslinked extracts. The following discussion summarizes my insights into these considerations as I developed applications for TrmA* in the course of my Ph.D.

4.4 Challenges enriching proteins from crosslinked extracts

Throughout my attempts to capture tagged RNAs (mainly 7SK), I have consistently observed the highest yields of tagged material from pull downs where TrmA* is co-expressed with TSR-RNA. While *in vitro* pull downs did produce enriched samples, the results were variable. The overall yield from in vitro capture experiments was an order of magnitude lower that I have seen with capture from co-expressed TrmA* in cells. Low yields are less problematic for analysis such as qPCR, which benefits from high sensitivity due to amplified signals, but most protein-based assays are significantly less sensitive and require more material or higher yields. One of the reasons I believe the *in vitro* pull downs are less effective is due to issues discussed in Chapter 2, namely that the TrmA substrate loops are methylated by endogenous methyltransferases, which prevent their being bound by TrmA* on beads. In the pull downs from co-expressed TrmA* in cells that I have performed on 7SK, I am overexpressing TrmA* simultaneously with tagged 7SK. This allows TrmA* to compete with endogenous methyltransferases and bind 7SK. With the new TrmA* E49R R51E and corresponding substrate, I believe that TrmA* capture of tagged RNA *in vitro* will be more effective by having a much higher fraction of RNA substrate loops unmethylated and therefore available for binding by TrmA*.

While TrmA* capture *in vivo* worked well for RNA in non-crosslinked extracts, I believe there are fundamental problems with *in vivo* capture of RNA from crosslinked extracts to correct, which are essential to fully realize the potential for a covalent tagging strategy. Specifically, the denaturing wash conditions that set the TrmA* approach apart from those associated with more traditional RNA affinity tags and crosslinking might interfere with each other. While the inactive TrmA* (C324A) that I have used in previous experiments does not react with RNA to form a covalent adduct, I predict that the protein would still fold normally and therefore bind substrate loops noncovalently. This is not normally a problem for my pull downs because I use denaturing washes to remove this non-covalent background binding. However, if I crosslink cell extracts that are expressing TrmA*, whether it be by UV or formaldehyde, any TrmA* that would be used as a negative control, such as C324A, that is associated with the substrate loop will now be covalently attached. This would make the background binding of TrmA* C324A indistinguishable in strength from the protein-based crosslinking of TrmA*. Similarly, I expect the same nonspecific binding to occur and result in crosslinking in the interactions between TrmA* and the TSR-U54C RNA loop.

While the use of an orthogonal TrmA* system will improve the ability of future studies to capture RNA *in vitro*, I believe that there are still significant difficulties to overcome with the crosslinked protein extracts. Crosslinked cell extracts can be more difficult to work with due to the variability of crosslinking, local concentration effects of chemical crosslinkers, and differences in solvent accessibility for proteins of interest. This is particularly true for formaldehyde crosslinked extracts, which would likely be used for

TrmA* RNA-protein capture. Because TrmA* relies on an accessible nucleotide in order to function, many aspects of crosslinking could prevent efficient capture of RNA *in vitro*. If the substrate loop or nearby regions is bound by any other proteins, the accessibility of the substrate RNA could be dramatically reduced. While crosslinkers such as formaldehyde are thought to react with a maximum of ~1% due to accessibility of reactive amines[105], the process of crosslinking is likely to modify the nucleotides of the substrate loop in a manner that will perturb TrmA* binding.

In order for TrmA* to bind RNA, the loop must be sufficiently flexible to adopt the flipped-out conformation of U54. If crosslinking alters the flexibility of this loop, it will reduce binding efficiency. For UV crosslinking in particular, there is reason to believe the process of crosslinking might chemically affect the substrate loop. Similar to the canonical DNA damage, a T-T dimer, uridine dimers are a frequent result of UV irradiation of RNA[106]. The TrmA substrate RNA contains two consecutive uridines, U54 and U55, which allow the possibility of such dimerization. Furthermore, because one of these uridines is the bound substrate, any dimerization would prevent covalent binding of TrmA*. In the capture of TrmA* substrates (**Chapter 2 Figure 4**), I showed that TrmA* readily binds RNA with adenosine in position 55. If UV crosslinking-induced U-U dimers are a concern, it may be worthwhile to include the U55A mutation in any future substrates (e.g. test the TSR-TM mutant with this addition: U55A C56A A58G C60U).

4.5 Substrate RNA pull down concerns

Another concern in application of TrmA* capture is the location of possible TSR insertions. As described in Chapter 1, many lncRNAs are thought to function through structural interactions[6,18]. Indeed, studying the structure of lncRNA is a focus of many labs[23,26]. Insertion of the TrmA* substrate into a lncRNA of unknown function or structure might perturb the function or interactions of that RNA. One of the appealing aspects of TrmA* is that the substrate is much smaller than other RNA affinity tags. This makes it less likely than other methods to significantly impact the structure of the tagged RNA and will also allow for easier insertion by CRISPR/Cas9 if used in such a manner. While these considerations illustrate several significant advantages over existing technologies, this system is not without disadvantages. For example, in capture of 7SK, the choice of 5' or 3' tagging has dramatic effects on the RNA transcription and biological interactions. Instability of 3'-tagged 7SK was observed (but not fully commented on) in previous work purifying 7SK and associated proteins[100]. In my own experience overexpressing 7SK 5'- or 3'-tagged, I observed dramatically lower expression of 3'-tagged 7SK than 5'-tagged 7SK. By RT-qPCR comparison normalized to total 7SK levels, 3'-TSR-tagged 7SK expression was approximately 600-fold lower than 5'-TSR-tagged 7SK. If an RNA of interest is not well-characterized to know what the best locations for substrate loop placement are, it is possible that a researcher applying TrmA* capture might unknowingly choose locations that disrupt RNA structure, protein binding, or transcriptional stability. For future studies, it may be best for researchers to insert the TSR into several different locations separately to determine the effects of TSR insertion empirically. Ideally, the TSR-RNA would be

tested to ensure it can rescue biological functions in a cell where the RNA is genetically ablated.

4.6 Crosslinked extract preparation

In the process of attempting to capture crosslinked RNA interactions with TrmA*, I sought to apply TrmA* in several different biological systems. Because I had already worked out conditions for efficient capture of 7SK from 293 cell extract, I hoped to first apply TrmA* capture with this lncRNA. After trying several different crosslinking conditions (.5%, 1%, 3% formaldehyde; .4, .8 J/cm^2 UV light) and capture of RNA both *in vitro* and *in vivo* (despite concerns about background binding as described previously), I was unsuccessful in capturing 7SK-associated proteins Hexim1 or CDK9 by western blot or in observing differences in protein samples by silver stain. Despite numerous different extract preparations ranging from hypotonic lysis, ChIP style extracts, CHART extract, and NEXSON (Nuclei EXtraction by SONication)[107] and many different attempts to vary washing conditions including my own fully denaturing washes, ChIP washes, PiCH[108], and ePiCH[109] washes, I was unsuccessful in even enriching 7SK RNA as I had been with non-crosslinked extracts. Without this enrichment, I could not be confident that any proteomics signal I would have seen did not result from non-specific binding or stochastic differences in the extracts.

4.7 Capture of RNA-protein complexes in other model systems

Because of the challenges I encountered when enriching TSR-7SK from crosslinked 293T extracts, I attempted TrmA* capture in other biological systems. I focused on the lncRNA roX2 in *D. melanogaster* embryonic cell culture (S2) because of its well-characterized protein binding partners and well-described genomic binding sites. This approach was potentially appealing because of the variety of different assays I could use as a proof of principle for my ability to apply TrmA* to capture RNA interactors. The most straightforward of these would be to capture any of the male sex lethal (MSL) complex proteins, MOF, MLE, and MSL1-3 known to form a complex with roX2[110]. Despite attempts to capture these complexes in crosslinked or native conditions, I failed to detect any of the proteins by western blot. Part of this difficult was due to low quality commercial antibodies for these targets, but another potential flaw in the assay revolved around the expression of roX2. roX2 structure has been studied and critical regions determined by protein interaction mapping and DMS probing[111], so I was confident in my ability to place the TSR RNA in a region of roX2 that was unlikely to affect structure or function. Furthermore, roX2 has previously been tagged with RNA affinity tags for purposes such as localizing DNA methyltransferases in DamID[112]. However, the only instances I have found for successful tagging of roX2 were with tags integrated genomically into whole organisms, not expressed from plasmids in *Drosophila* cell culture as I was performing. Perhaps tagging roX2 genomically would have resolved my issues with capture of roX2-associated proteins, but would add other complexities and heterogeneities to the process of roX2 capture and was outside my areas of expertise for the purposes of these studies.

4.8 Use of TrmA* to isolate RNA for structural studies

Another possible application of high efficiency purification of an RNA from complex mixture using TrmA* is studying the structures of specific subpopulations of RNA. 7SK has been shown to exist in two structural subpopulations depending on whether it is bound to BAF or Hexim1[113]. Furthermore, it is possible to alter Hexim binding in 7SK by altering nucleotides 43, 44, 65, and 66 of 7SK[114]. I hoped to use TrmA* to isolate a mutant form of 7SK and demonstrate that I was able to isolate this subpopulation and show that its structure by DMS probing was significantly different than total 7SK.

To start, I created a tagged 7SK mutant predicted to abolish HEXIM binding (A43G/U44C/A65G/U66C). 293T cells were then cotransfected by calcium phosphate with mutant TSR- or TSR-U54C 7SK and TrmA* or TrmA* C324A. 24 hours post transfection, cells were treated with 5 mL PBS +/- 0.5% dimethyl sulfide (DMS) for five minutes. After removing DMS, samples were washed with 5 mL quench buffer (50 mM Tris, pH 7.5, 100 mM NaCl, 3 mM MgCl2, and 40 mM β-mercaptoethanol) three times and finally washed in PBS where cells were harvested by scraping. After lysis by sonic disruption, cell extract was exposed to streptavidin beads and RNA extracted as described previously. The samples were then processed for DMS-seq and compared to inputs. Unfortunately, despite several attempts at performing DMS probing of 7SK and mutant 7SK I was not able to observe reproducible differences in the regions of 7SK known to have altered flexibility with either BAF (232 – 245) or Hexim1 (290 – 299) binding. Despite these negative results, I observed that in TrmA* bound RNA populations there was a very sharp

drop off of reverse transcription read-through at the site of TrmA* binding, despite having

treated the samples with proteinase K. This finding that TrmA* caused RT stops more than

90% of the time (**Figure 4.1)** could prove very useful in determining a way to separate

RNA that had be bound by TrmA* without the need to physically separate the RNA. This

drop off formed the foundation for the RT stop assay I used to generate the bound and

unbound populations of the high throughput RNA library screen in **Chapter 2**.
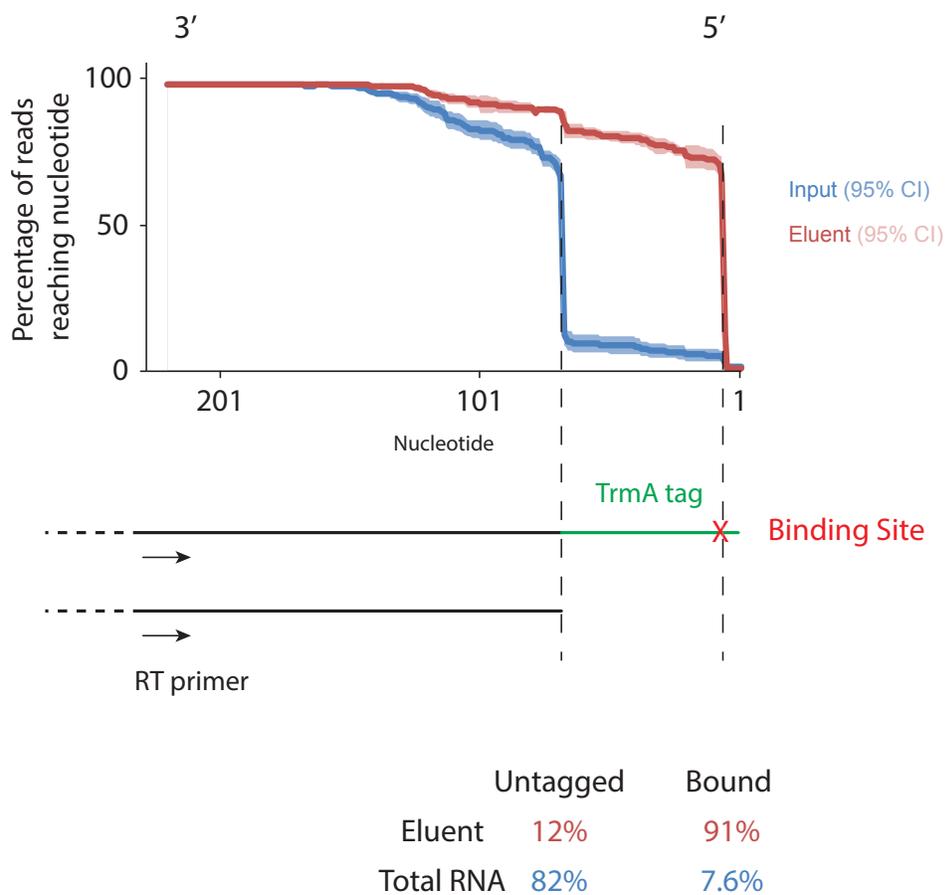
Figure 4.1: **TrmA* tag and RT-stops estimations.** 7SK and tagged 7SK reads are plotted by the percentage of reads that contain each nucleotide position. Drop off of RT reads is used to calculate the percentage of RNA that were either untagged (at nucleotide 70) or bound by TrmA* (at nucleotide 8)

4.9 TrmA* can create useful RNA-protein adducts *in vivo*

In addition to these purification methods, it is possible to append other protein moieties to RNA through fusion proteins with TrmA*. For example, RNA could be fluorescently labeled *in vivo* by the addition of fluorescent proteins such as GFP. By creating TrmA* fused to other affinity proteins it might be possible to generate novel RNA-protein complexes. I also hypothesize that TrmA* could also be used to stabilize RNA transcripts. The covalent adduct that TrmA* forms with its RNA substrate could possibly block 5'->3' by Xrn1 and/or 3'->5' decay by the exosome[115] in the same manner that it blocks reverse transcriptase as demonstrated in **Figure 4.1**. This strategy of RNA stabilization has already been observed in nature with particles such as sno-lncRNAs. These RNAs are thought to utilize the formation of snoRNPs on either end of the lncRNA transcript to block exonuclease degradation[116]. By placing TSR sequences on either side of a transcript of interest and coexpressing tagged RNA with TrmA* *in vivo*, RNA degradation might be reduced. Preliminary data with TSR-tagged luciferase genes suggest that this may be possible, but still requires substantial validation.

I designed an RNA transcript containing both renilla and firefly luciferase in series based on the commonly used internal ribosome entry site assay plasmid. I added TSR hairpins 5' of the IRES before the firefly luciferase and 3' of the end of the gene as illustrated in (**Figure 4.2A**). If TrmA* blocks exonucleases, coexpressing TrmA* with this transcript should stabilize the transcript, causing turnover to be lower and leading to higher levels of the RNA in these cells. Since the firefly luciferase is protected by TrmA sites on both

sides of the transcript and the renilla luciferase is protected only on the 3' end, I expected that the firefly luciferase RNA would be more highly stabilized by TrmA* overexpression. In order to test this hypothesis, I transfected 293T cells with the luciferase construct and either an active or inactive TrmA* expression vector. Because I was interested in observing degradation, I allowed the transfected cells to reach confluence and harvested 3 days after confluence was reached. I then isolated the total RNA with Trizol (Thermo) and ethanol precipitation and removed contaminated DNA with Turbo DNase (Thermo) treatment. I reverse transcribed the total RNA with SuperScript VILO Master Mix with random hexamer primers (Thermo) and quantified the abundance of renilla and firefly luciferase transcripts by qPCR primers for two different sites in each transcript as illustrated in (**Figure 4.2A,** rluc 1, rluc 2, fluc 1, fluc 2). I normalized each experiment to the total amount of RNA using MALAT1 as an indicator of total RNA abundance for each sample. I then compared the ratio of each luciferase to total RNA with and without TrmA* expression (**Figure 4.2B**). I saw that while the firefly luciferase was stabilized as expected, the renilla luciferase was more highly stabilized. There are several reasons that this could be the case. If a majority of the RNA degradation comes from 3' -> 5' exonuclease activity, any protected firefly luciferase transcripts will also protect the upstream renilla luciferase transcripts. Either way, this complexity left these results as ambiguous and further exploration of this hypothesis will require an alternative system.
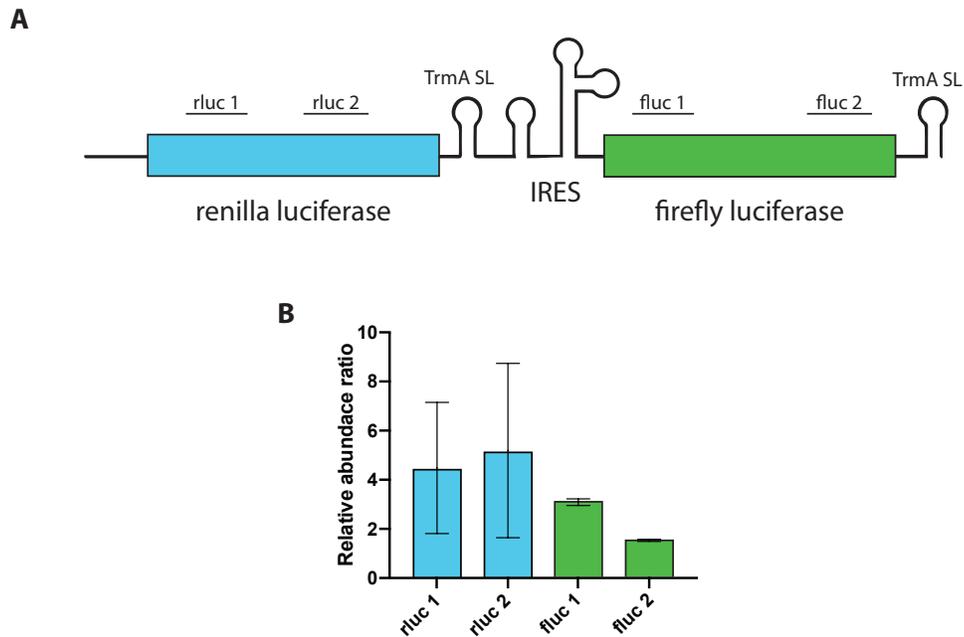
Figure 4.2: **TrmA* stabilizes tagged mRNA transcripts**. (A) Construct design schematic for expression in mammalian cells to test the stabilization of RNA by TrmA* binding. (B) Enrichment for each qPCR primer set in cells coexpressing TrmA* and luciferase genes normalized to cells expressing luciferase only.

4.10 Conclusions

While I have explored different ways to use TrmA* as a pull-down reagent to capture RNA-protein interactions over the course of my dissertation research, I have certainly not exhausted the parameter search space. Hopefully, this discussion will prove useful to future researchers that may hope to apply this technology, which I believe to be a promising tool. Perhaps the most straightforward application will be to improve the *in vitro* capture of RNA using TrmA* bound to solid phase. For this application in particular, I believe that the reengineered TrmA* E49R R51E and TSR-TM interaction pair described in **Chapter 2** will exhibit marked improvements. Furthermore, while the *in vitro* capture

of RNA may still suffer from difficulties associated with structural constraints and possible chemical modification of the substrate uridine, *in vitro* capture will prevent any possibility of crosslinker induced covalent attachment between inactive TrmA* or unreactive RNA substrates. Lastly, TrmA* can be applied to create RNA-protein adducts *in vivo* which can possibly applied to problems such as RNA transcript stabilization as preliminarily demonstrated with my experiments with tagged luciferase. With the ability to create covalent RNA-protein adducts, there many new routes to explore ranging from designed RNPs, fluorescent RNA transcript tagging, and engineered RNA localization.

Chapter 5

Methods and Data Analysis

5.1 Methods

5.1.1 TrmA cloning and mutagenesis

E. coli genomic DNA was extracted from DH5α cells with phenol-chloroform, precipitated, and resuspended in tris-buffered EDTA. TrmA was amplified by PCR, gel purified, restriction digested, and cloned into the hexahistidine-tagged bacterial expression vector pET28a. For *in vivo* biotinylated TrmA, the minimal BirA substrate was added by ligation of annealed and phosphorylated synthetic oligonucleotides. TrmA mutagenesis was performed by QuickChange II Site-Directed Mutagenesis (Agilent) using primers with 25-nucleotide homology on either side of the nucleotides to be mutated. TrmA* was first generated and subsequent mutations were generated from the same TrmA* background.

5.1.2 TrmA expression and purification

1 L of LB was inoculated with E. coli BL21(DE3) transformed with TrmA containing plasmids. Once the culture reached 0.6 OD_{600}, cells were induced with the addition of 1 mM IPTG for two hours and pelleted by centrifugation at 4000xg for 10 minutes. In TrmA mutant samples to be biotinylated, 20μM biotin was added with IPTG inductions.

Cells were solubilized by incubating for 15 minutes in B-PER, Bacterial Protein Extraction Reagent (Thermo) (1mL/g cell pellet) and cOmplete protease inhibitor cocktail (Roche) and lysate was cleared by centrifugation for 15 minutes at 15,000xg. Cleared lysate was applied to 5 mL Ni-NTA resin and incubated 1 hr at 4˚C. Samples were then washed 2x 5 minutes with 1 bed volume of B-PER + protease inhibitor cocktail and 5x 5 minutes with 1 bed volume of wash buffer. Following washes, samples were eluted for 30 minutes each in elution buffer containing 100, 150, and 250 mM imidazole. Fractions were analyzed by SDS-PAGE to determine purity of TrmA mutants and eluents were pooled and dialyzed into storage buffer and protein concentrations measured by Bradford assay (BioRad).

Protein purification buffers:

Solubilization

B-PER

Roche cOmplete protease inhibitor cocktail (manufacture recommended concentration)

Wash

50 mM NaCl

2 mM DTT

20 mM Tris-HCl, pH 7.0

Roche cOmplete protease inhibitor cocktail

50 mM imidazole

<u>Elution</u>

50 mM NaCl

2 mM DTT

20 mM Tris-HCl, pH 7.0

Roche cOmplete protease inhibitor cocktail

+ 100, 150, 250 mM imidazole

<u>TrmA storage buffer</u>

50 mM NaCl

2 mM DTT

20 mM Tris-HCl, pH 7.0

5.1.3 TrmA reactions

TrmA reactions were carried out in 1x TrmA reaction buffer with the addition of 1.5 mM

SAM and .16 U/µL SUPERase-IN (Invitrogen). Once reactions were complete, one

equivalent of Quench Solution was added and mixed thoroughly. Quenched samples were

then loaded onto gels for urea-PAGE analysis and visualized by fluorescent imaging.

<u>TrmA reaction buffer (5x)</u>

500 mM MOPS, pH 7.8

500 mM ammonium acetate

5 mM MgCl$_2$

25 mM DTT

Quench solution

8 M urea

10 mM EDTA

5.1.4 TrmA kinetics rate determination

TrmA reactions of 0, 1, 3, 9, and 27 minutes were performed on each RNA substrate and TrmA mutant pair. After quenching, samples were separated by urea-PAGE and fluorescently imaged. Signal intensities of bands in each lane were determined using the ImageJ. Fraction bound was calculated by dividing the intensity of the upper, bound, band by the total fluorescence of the lane.

Triplicate data of each reaction was plotted and fit by non-linear least squares regression fit to the equation:

$$[P] = \frac{[R]_0[T]_0 e^{k_{obs}t([T]_0 - [R]_0)} - [R]_0[T]_0}{[T]_0 e^{k_{obs}t([T]_0 - [R]_0)} - [R]_0}$$

$$Fraction\ shifted = [P]/([P] + [R])$$

Where, R is RNA, T is TrmA mutant, P is shifted product, and t is time calculated by dividing the intensity of the upper, bound, band by the total fluorescence of the lane.

Concentrations of RNA ($[R]_0$) added were specifically added. Concentrations of active TrmA added ($[T]_0$) and the rate of covalent binding k_{obs} were calculated by NLS and can be seen in **Table 5.2.1** and **Table 2.2** respectively.

5.1.5 TrmA competitive binding assay

TrmA reactions were carried out as described in the TrmA binding assay with equal concentrations of TSR (Cy3 labeled) and mutant TSR (Cy5 labeled) in excess of TrmA. Gels were fluorescently imaged and intensities measured using ImageJ as in TrmA kinetics binding. Fraction bound for each substrate were calculated as previously described and compared to the wild type substrate loop.

5.1.6 TrmA substrate library

The TrmA substrate library was designed with unique 12-nucleotide barcodes of Hamming distance 5 for each of 2000 substrate mutants. For a detailed description of all substrate variants see **Appendix B.** DNA template was synthesized by Twist Biosciences, amplified by PCR for 10 cycles in order to generate enough template for *in vitro* transcription, but not amplified so much as to create significant PCR biases of the library population. Lastly, the substrate DNA was agarose gel purified and eluted into Tris-buffered EDTA.

5.1.7 In vitro transcription of TrmA substrate library RNA

15 ng of gel purified template DNA were added to in vitro transcription buffer with 10 μg T7 RNA polymerase and incubated at 16˚C over-night. Template DNA was removed by Turbo DNase (Thermo Fischer Scientific) and reaction enzymes were subsequently removed by proteinase K digestion. RNA was isolated by phenol-chloroform extraction and ethanol precipitation and RNA pellet was resuspended in Tris-buffered EDTA and stored at -80˚C.

In vitro transcription buffer

75 mM Tris-HCl, pH 7.5

40 mM $MgCl_2$

2 mM Spermidine

5 mM DTT

5 mM NTPs

5.1.8 TrmA library screen

Equal units of each TrmA mutant were reacted with 5 μg *in vitro* transcribed TrmA

substrate RNA as described in the fluorescent substrate binding assays for either 30

seconds or 30 minutes. Following reaction, samples were treated with proteinase K and

RNA was isolated by ethanol precipitation. Purified RNA was reverse transcribed with

SuperScript III (Thermo Fischer). RNA was then degraded by addition of 3% volume of

5M NaOH and heating to 95˚C for 5 minutes. RNA-free cDNA was separated by urea-

PAGE. Bands corresponding to full length and truncated cDNA were excised and

extracted from the gel by excision, snap freezing in gel extraction buffer, and letting

samples diffuse overnight. DNA was precipitated by ethanol precipitation with sodium

acetate and glycogen carrier and resuspended in Tris-buffered EDTA. These samples

were then amplified by PCR with Phusion (NEB) for 14 cycles. Samples were agarose gel

purified and relative concentration measured by qPCR. Samples were diluted to

normalize DNA and then indexed for Illumina sequencing by PCR with Phusion (NEB)

amplifying for 4 cycles. Finally, relative concentrations of DNA were measured by qPCR

and samples were mixed evenly for sequencing.

<u>Gel extraction buffer</u>

20 mM Tris, pH 7.5

250 mM sodium acetate

1 mM EDTA

0.25% SDS

5.1.9 Mammalian cell culture and transfection

HEK 293T cells were grown in DMEM + 10% FBS in 10cm plates. Cells were passaged to 30-40% confluency before transfecting with calcium phosphate.

For calcium phosphate transfection, 10 µg of DNA were diluted into 437.5 µL final volume of ddH$_2$O. 62.5 µL of 2M CaCl$_2$ were then added to a final volume of 500 µL. DNA was then added dropwise to 500 µL of 2x HBS in a 50 mL conical tube with constant aeration and incubated 5 minutes at room temperature. 9 mL of growth media was then added to the precipitated DNA. 293T growth media was then removed and replaced with DNA containing growth media. Twenty-four hours post transfection, media was exchanged for fresh media. Cells were harvested forty-eight hours post transfection except in the case of the luciferase turnover assay in **Chapter 4**.

<u>2x HBS</u>

50 mM HEPES, pH 7.05

10 mM KCl

12 mM glucose

280 mM NaCl

1.5 mM Na_2HPO_4

5.1.10 Cell extract preparations

Hypotonic lysis

Confluent 10 cm plates of 293T cells were harvested by scraping into PBS and pelleted

by centrifugation at 1200xg at 4˚C. Supernatant was removed and cells resuspended in 1

mL ice-cold PBS. Cells were again pelleted and supernatant removed. In order to remove

all residual PBS, pelleted cells were centrifuged again and supernatant removed again.

Cells were then resuspended in 500 μL ice-cold hypotonic lysis buffer with 1 mM DTT

and 0.1 mM PMSF and incubated on ice for 5 minutes. Cells were then snap frozen in a

dry ice/ethanol bath and thawed at 37˚C. Freeze-thawing was repeated two more times

and lysed cells placed on ice and NaCl was gradually added to 400 mM final

concentration. Lysate was incubated on ice for 5 minutes before centrifugation at

20,000xg for 10 minutes at 4˚C to clear lysate.

Hypotonic lysis buffer

20 mM HEPES

2 mM MgCl$_2$

0.2 mM EGTA

10% glycerol

Sonic disruption

Pelleted cells were resuspended in 1 mL of RIPA lysis buffer and incubated on ice for 10 minutes. Resuspended cells were then lysed by sonic disruption at 50W (10% output power) in cycles of 1 second on, 2 seconds off for 30 seconds (10 cycles). This was repeated a total of three times to fully lyse cells. Cell lysates were then cleared by centrifugation at 20,000xg for 10 minutes at 4˚C.

RIPA lysis buffer

50 mM Tris, pH 8

150 mM KCl

0.1% SDS

1% Triton-X

5 mM EDTA

0.5% sodium deoxycholate

0.5 mM DTT

0.1 mM PMSF

20 U/mL SUPERase IN, RNase inhibitor (Thermo)

5.1.11 TrmA* in vitro capture

TrmA* bead preparation

25 µL of Pierce high capacity streptavidin agarose resin (Thermo) was washed 3x 500 µL with PBS. 20 µg of biotinylated and purified TrmA* (or desired mutant) was then applied to washed resin in 500 µL PBS and incubated, rotating, at room temperature for 30 minutes to bind. Resin was then washed once with 5 bead volumes of PBS + 1 mM biotin for 5 minutes in order to wash away nonspecific binding and to block unoccupied streptavidin sites. Beads were then washed 2x with 500 µL bead wash buffer and once with TrmA reaction buffer.

<u>Bead wash buffer</u>

2 M urea

20 mM HEPES, pH 7.4

RNA capture

33 µL of cleared lysate from tagged-7SK expressing cells was then mixed with TrmA reaction buffer into a total volume of 100 µL with 1.5 mM SAM and 20 U/mL SUPERase IN (Thermo). This mixture was then applied to TrmA* resin and allowed to react for 45 minutes at room temperature, rotating. Samples were then washed 2x with 500 µL PBS, 2x 500 µL low salt wash buffer, 2x 500 µL high salt wash buffer, and 100 µL elution buffer. RNA was eluted from beads by treating with 20 µg of proteinase K in 100 µL of elution buffer for 30 minutes at 40˚C with constant shaking.

Low salt wash buffer

4 M urea

1% SDS

20mM HEPES, pH 7.0

High salt wash buffer

4 M urea

1% SDS

250 mM NaCl

20mM HEPES, pH 7.0

Elution buffer

4 M urea

20mM HEPES, pH 7.0

5.1.12 TrmA* in vivo capture

100 µL Pierce high capacity streptavidin agarose resin was washed with PBS 2x 500µL.

400 µL cleared lysate expressing both tagged 7SK and biotinylated TrmA* was applied to

beads and incubated for 30 minutes at room temperature with constant rotation to

bind. As with *in vitro* binding, samples were then washed 2x with 500 µL PBS, 2x 500 µL

low salt wash buffer, 2x 500 µL high salt wash buffer, and 100 µL elution buffer. RNA

was eluted from beads by treating with 20 µg of proteinase K in 100 µL of elution buffer

for 30 minutes at 40˚C with constant shaking.

5.1.13 TrmA* capture quantification

RNA eluted from either *in vitro* or *in vivo* TrmA* capture was extracted with phenol-chloroform and pelleted by ethanol precipitation and centrifugation with a glycogen carrier and resuspended into 20 μL Tris-buffered EDTA. 5 μL of resuspended RNA was DNase treated in 10 μL using RQ1 DNase (Promega). DNase was heat inactivated and 5μL of the DNase reactions were reverse transcribed in 10 μL using SuperScript VILO (Thermo) to generate cDNA. cDNA was then analyzed by qPCR in 15 μL reactions containing 1 μL cDNA, 7.5 μL iTaq Universal SYBR Green Supermix (Biorad), and 1 μM target primers.

5.1.14 TrmA* capture sequencing

cDNA from TrmA* capture was prepared for sequencing on Illumina 2500 sequencer with the NEBNext DNA Library Prep Master Mix Set for Illumina (NEB) using standard protocols.

5.1.15 7SK DMS probing

90% confluent 10cm plates of 293T cells coexpressing TrmA* and tagged 7SK (or inactive variants of either transcript) were rinsed with PBS and then treated with 15 mL of 0.5% DMS in PBS for 5 minutes at room temperature. DMS was removed and then quenched by washing the cells 3x with DMS quench buffer. After rinses, cells were harvested by scrapping into 5 mL PBS. RNA was then extracted from cells as in other protocols using Trizol and ethanol precipitation.

DMS quench buffer

50 mM Tris, pH 7.5

100 mM NaCl

3 mM $MgCl_2$

40 mM β-mercaptoethanol

RNA was then reverse transcribed using primers targeted to 7SK. 1 µg of RNA was added
to 12 µL total volume with 10 pmol primer. This mixture was incubated at 70˚C for 5
minutes. To this, 3 µL of 5x FS buffer (Thermo) was added and incubated for 10 minutes
at 55˚C. 5 µL of RT master mix was then added to each reaction. Reaction mixtures were
then incubated for 45 minutes at 55˚C to reverse transcribe and then heat inactivated
for 5 minutes at 85˚C.

RT master mix (per reaction)

1 µL 5x FS buffer (Thermo)

1 µL 10 mM dNTPs

1 µL 100 mM DTT

0.5 µL SuperScript III (Thermo)

1.3 µL ddH_2O

0.2 µL RNaseOUT (Thermo)

cDNA was then prepared for sequencing. First, RNA was purified by SPRI (2.5x volumes), washed with ethanol, and resuspended in 16 µL water. Next, 3' adaptors were ligated onto purified cDNA. 6.8 µL DNA were added to 13.2 µL 3' adaptor ligation mix and incubated at 65˚C for 2 hours followed by 85˚C for 15 minutes.

3' adaptor ligation mix (per reaction)

1 µL 1 mM ATP

1 µL 50 mM MnCl$_2$

0.2 µL 100 µM adaptor (CAGACGTGTGCTCTTCCGATC)

8 µL 50% PEG 8000

 2 µL CircLigase buffer (Epicenter)

1 µL CircLigase (Epicenter)

Adaptor ligation products were then SPRI purified with 1x bead volume and again resuspended into 16 µL water. Products were then amplified 6 cycles (20 seconds 98˚C, 20 seconds 64˚C, 90 seconds 72˚C each) by PCR with Phusion master mix (NEB) in 20 µL reactions. Amplifications were purified by SPRI with 1.8x bead volumes and level measured by qPCR. Samples were then amplified with Phusion master mix again to add Illumina indexes to each sample. Each sample was amplified a variable number of cycles (4-8) in order to approximately normalized DNA concentrations for high-throughput sequencing. Amplified products were again purified by SPRI with 1.8x bead volumes and eluted in 16 µL of water. Finally, equal DNA concentrations (as measured by a final qPCR) were mixed for sequencing.

5.1.16 Luciferase stabilization and qPCR

TrmA substrate loops were added into a dual luciferase IRES reporter plasmid though restriction digestion and ligation with phosphorylated oligos. Tagged luciferase and TrmA*/TrmA* C324A containing constructs were transfected into 293T cells with calcium phosphate. Once confluent, cells were grown for an additional 3 days with daily media changes. Cells were harvested by scraping and RNA isolated by Trizol (Thermo). Levels of luciferase gene expression were determined by RT-qPCR. Total RNA concentration was normalized to abundance of MALAT1 transcripts measured by qPCR.

5.2 Data analysis

5.2.1 TrmA* binding quantification

TrmA* reactions with fluorescent oligos were separated by urea-PAGE and laser scanned on a GE Typhoon 9500. Images were then carried into ImageJ[117] and fluorescence intensity profile of each lane was measured using the 'analyze gel' tool. For each lane, I divided the image into sections corresponding to bound and unbound RNA products and integrated the intensity profile to calculate the total amount of fluorescence in each region. In order to determine the fraction of RNA bound by TrmA* from these values, I divided the total intensity of bound RNA by the total fluorescence signal for the lane. In instances in which I wanted to determine the molar quantity of RNA bound, such as for kinetics analysis, I multiplied the fraction of RNA bound by the known quantity (determined by synthetic yield post HPLC purification by IDT) of RNA added.

5.2.2 TrmA* binding kinetic analysis

After measuring the fraction of RNA bound for each triplicate TrmA* reaction, values were plotted using ggplot2[118]. Using the formula for product formation described in Appendix A, I used the nonlinear fit (NLS) function from the R stats package[119] to determine the quantity of active TrmA* initially present, $[T]_0$, in each reaction and most importantly, the k_{obs} for each enzyme. Because I was able to measure the amount of product formed and added a known quantity of TSR RNA, only $[T]_0$ and k_{obs} were fit to describe the binding reaction. In reactions with TSR and TSR-G57C, the same exact TrmA* proteins preparations and quantities were used. This allowed me to fit curves to both sets of reactions, but to maintain the same measured $[T]_0$ value for both reactions. The results of k_{obs} determination are described in **Chapter 2**. Calculated $[T]_0$ values for each reaction are shown in **Table 5.2.1**.

Protein	TSR	G57C	C60G	C56A A58G C60U
TrmA*	0.905	0.905	0.127	0.089
	(0.866, 0.946)	(0.866, 0.946)	(0.116, 0.143)	(0.080, 0.100)
E49R	0.326	0.326	0.067	0.117
	(0.219, 0.385)	(0.219, 0.385)	(0.058, 0.079)	(0.080, 0.100)
R51E	0.715	0.715	0.109	0.135
	(0.676, 0.756)	(0.676, 0.756)	(0.101, 0.117)	(0.115, 0.163)
E49R R51E	0.756	0.756	0.236	0.191
	(0.700, 0.822)	(0.700, 0.822)	(0.208, 0.274)	(0.177, 0.209)

Table 5.2.1: **Measured values of $[T]_0$ (μM) for TrmA mutants and RNA substrates (95% CI)**

5.2.3 TrmA* substrate sequencing data handling:

Paired end sequencing reads were demultiplexed using exact matches for each Illumina i5 and i7 index using the 'grep' command and divided into .fastq files for each sample, lane, and read direction. Each of these sample .fastq files was then searched for barcode sequences corresponding to each substrate loop. These matches were done using 'grep -c' for each barcode sequence (or the reverse compliment for reverse reads). Once all 2000 barcodes were searched for in each sample, the total number of matched reads was calculated by adding the total number of reads for each barcode. Because each sample contained different numbers of reads by sequencing, I normalized the counts for each barcode by dividing the barcode counts by the total number of reads that matched barcodes.

5.2.4 Correlation plot generation:

Pierson's R coefficients of correlation of all samples was determined by comparing the fractions of reads attributed to each substrate for each sample using the R package corrplot[120] and ordered using angular order of eigenvectors to determine which samples best correlated with each other.

5.2.5 Motif determination:

Enriched substrates for each TrmA mutant were determined using DESeq2[121] comparing bound samples to unbound samples (30 second and 30 minute incubations were used together for this analysis) with a significance cutoff of $p < 0.05$. The list of resulting

substrate loops was recorded and sequences of only the loop mutants were used for motif determination. For each position 55-60 of the loop, I calculated the relative fraction of each base at that location to create a matrix of probabilities for each base and position. Similarly, I compiled all of the sequences present in the library to create a matrix to correct for input biases. These values were then used to generate enriched and depleted sequence motifs using Logolas[122].

5.2.6 TrmA* capture quantification

TrmA* capture was quantified by measuring the transcript abundance by qPCR for both 10% input samples and bead eluents. For each TrmA* mutant and 7SK tag variant, eluent ct values were divided by input to determine a fraction of input yield for each sample.

5.2.7 TrmA* capture sequencing data analysis

Samples were demultiplexed and processed using Cutadapt to remove adapter sequences. Processed reads were aligned to the human genome (hg19) using BowTie2 and sorted with bedtools. After sorting, tracks were visualized in IGV with bedGraphs generated from bedtools. Using HISAT2, I generated FPKM values for each transcript and created a rank ordered list by sorting each gene from highest to lowest FPKM values. 7SK appeared as the highest FPKM value gene in the active TrmA* and reactive 7SK sample. Each list of FPKM values was used to compare eluent to input for each experimental condition and were used to generate scatter plots of the values with eluent on the y-axis and input on the x-axis. To better visualize each sample, I graphed the $\log_2$ FPKM values to help spread out the data more evenly.

5.2.8 TrmA* bound RT stops measurement

TrmA* bound 7SK RNA was prepared for targeted structure seq as described previously. Using published pipelines for this technique I was able to determine the fraction of reads that made it to each position of reverse transcription[123]. Using these values, I can plot the percentage of reads at each point and also calculate the change from one nucleotide to the next. By subtracting the fractional read value of a later nucleotide from the previous nucleotide, I can determine the stop rate for a given nucleotide.

Appendix A

Supplemental equations

TrmA* binding rate calculations and equations

TrmA* binding of RNA can be represented as a two-step reaction where RNA (R) first binds to TrmA* (T) noncovalently. TrmA* then reacts with the RNA to form the RNA-TrmA* complex (P).

$$R + T \rightleftharpoons RT \rightleftharpoons P$$
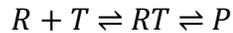
If I assume that the rate limiting step in this reaction is the chemical conversion of RT into P, the reaction can be simplified into:

$$R + T \xrightarrow{k_{obs}} P$$
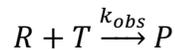
The rate of product formation can be defined from this reaction as:

$$\frac{d[P]}{dt} = -k_{obs}[R][T]$$

If the [P] is the only product formed, then [R] = [R]$_0$ – [P] and [T] = [T]$_0$ – [P], and the equation can be defined as:

$$\frac{d[P]}{dt} = -k_{obs}([R]_0 - [P])([T]_0 - [P])$$

Which can be rearranged to:

$$\frac{d[P]}{([R]_0 - [P])([T]_0 - [P])} = -k_{obs}dt$$

If I integrate this equation from t = 0 to time t, I get:

$$\int_0^{[P]} \frac{d[P]}{([R]_0 - [P])([T]_0 - [P])} = -k_{obs}\int_0^t dt$$

66

Using partial fractions, I can solve integrate left side of the reaction:

$$\int_0^{[P]} \frac{d[P]}{([R]_0 - [P])([T]_0 - [P])} = \frac{1}{[R]_0 - [T]_0}\left(\ln\frac{[R]_0}{[R]_0 - [P]} - \ln\frac{[T]_0}{[T]_0 - [P]}\right)$$

This then simplifies to:

$$\frac{1}{[R]_0 - [T]_0}\ln\frac{[R]_0[T]}{[T]_0[R]}$$

Therefore:

$$-k_{obs}t = \frac{1}{[R]_0 - [T]_0}\ln\frac{[R]_0[T]}{[T]_0[R]}$$

$$k_{obs}t([T]_0 - [R]_0) = \ln\frac{[R]_0[T]}{[T]_0[R]}$$

Substituting [R] = [R]$_0$ – [P] and [T] = [T]$_0$ – [P]:

$$e^{k_{obs}t([T_o]-[R]_0)} = \frac{[R]_0[T]}{[T]_0[R]} = \frac{[R]_0([T]_0 - [P])}{[T]_0([R]_0 - [P])}$$

$$([T]_0([R]_0 - [P])e^{k_{obs}t([T]_0-[R]_0)} = [R]_0([T]_0 - [P])$$

$$[T]_0[R]_0 e^{k_{obs}t([T]_0-[R]_0)} - [T]_0[P]e^{k_{obs}t([T]_0-[R]_0)} = [R]_0[T]_0 - [R]_0[P]$$

$$[T]_0[[R]_0]e^{k_{obs}t([T]_0-[R]_0)} - [R]_0[T]_0 = [T]_0[P]e^{k_{obs}t([T]_0-[R]_0)} - [R]_0[P]$$

Which finally simplifies to:

$$[P] = \frac{[R]_0[T]_0e^{k_{obs}t([T]_0-[R]_0)} - [R]_0[T]_0}{[T]_0e^{k_{obs}t([T]_0-[R]_0)} - [R]_0}$$

My observations of fraction shifted can be fit by substituting this value of [P] into:

$$Fraction\ shifted = [P]/([P] + [R])$$

Appendix B

TrmA substrate library design

Due to a variety of restrictions, it was not practical to synthesize a randomized library of all possible TrmA substrates, particularly for a synthetic DNA library in which each member is explicitly defined and synthesized. A convenient number for this process turned out to be 2,000 members. With this constraint it was not possible to sample all possible mutations of the TrmA substrate loop. I instead focused on designing substrates that were more similar to the normal TrmA substrate such as all of the single, double, and triple mutants of the loop region. In addition to these mutations, I created all possible mutations of nucleotides 55-59 by adding the quadruple and quintuple mutants of these positions. This left me with several other routes to continue to explore. Some parameters I was interested in investigating were if there were any loop structures besides the 5 membered step and 7 membered loop structure. To perturb these, I first chose to create all of the single and double insertions and deletions of the TrmA substrate. While these were unlikely to react with TrmA, they would be very surprising results and were a relatively small portion of overall sequence space. Finally, with the mutations I wanted to see if the stem structure and composition had any effects on TrmA binding. To do this, I varied the GC content and altered the length of the stem. Finally, I wanted to create an abundance of positive and negative controls to be sure that the barcodes and my own handling wasn't affecting the library. These would allow me to calibrate my results based on how well the positive and negative controls were separated from each other and

ensure that the TrmA* mutant added to each reaction was active. The breakdown of each

mutant category is show in **Table B.1.**

Type	Number of sequences
Single deletion	17
Double deletion (within loop)	15
Single insertion	72
Double insertion (within loop)	352
Single mutations of nt 55-60	18
Double mutations of nt 55-60	135
Triple mutations of nt 55-60	540
Quadruple mutants of nt 55-59	405
Quintuple mutants of nt 55-59	243
Altered stem composition	156
Positive controls	27
Negative controls	20
Total	2000

Table B.1: **TrmA substrate library composition by type of mutation.**

Appendix C

Primers and oligos

TrmA cloning primers:

AAAGGATCCATGACCCCCGAACACCT	TrmA with BamHI cut site 5'
GGGGAATTCTTACTTCGCGGTCAGTAATACG	TrmA with EcoRI cut site 3'
/5PHOS/CTAGCATGTCCGGCCTGAACGACATCTTCG	To add BriA biotinylation
AGGCTCAGAAAATCGAATGGCACGAAG	motif to TrmA by ligation
/5PHOS/GATCCTTCGTGCCATTCGATTTTCTGAGCC	To add BriA biotinylation
TCGAAGATGTCGTTCAGGCCGGACATG	motif to TrmA by ligation

TrmA mutagenesis primers:

CCCTACACGCACCATATGCAGTGCGGCGTAT	E358
ATACGCCGCACTGCATATGGTGCGTGTAGGG	E358Q
GGTCAGTCATTACCGGATGCGCGCGCGGTTCCGCATCTGGCACGATGGCGATG	E49R
CATCGCCATCGTGCCAGATGCGGAACCGCGCGCGCATCCGGTAATGACTGACC	E49R
GGTCAGTCATTACCGGATGCGCGCGCGGAGTTCGAGATCTGGCACGATGGCGATGACCTGT	R51E
ACAGGTCATCGCCATCGTGCCAGATCTCGAACTCCGCGCGCATCCGGTAATGACTGACC	R51E
GGTCAGTCATTACCGGATGCGCGCGCGGGTTCGAGATCTGGCACGATGGCGATGACCTGT	E49R R51E
ACAGGTCATCGCCATCGTGCCAGATCTCGAACCGCGCGCGCATCCGGTAATGACTGACC	E49R R51E
GCGTATTTTGTACATCTCCGCTAACCCGGAAACGTTATGC	C324A
GCATAACGTTTCCGGGTTAGCGGAGATGTACAAAATACGC	C324A

TrmA RNA substrate loops:

/5Cy5/rUrGrCrUrGrUrGrUrUrCrGrArUrCrCrArCrArGrC	TSR
/5Cy3/rUrGrCrUrGrUrGrGrCrUrCrGrArUrCrCrArCrArGrC	TSR-U54C
/5Cy3/rUrGrCrUrGrUrGrGrUrUrCrCrArUrCrCrArCrArGrC	TSR-G57C
/5Cy3/rUrGrCrUrGrUrGrGrUrUrCrGrArUrGrCrArCrArGrC	TSR-C60G
/5Cy3/rUrGrCrUrGrUrGrGrUrUrArGrGrUrUrCrArCrArGrC	TSR-TM (C56A A58G C60U)

TrmA substrate library primers:

GAAATTAATACGACTCACTATAGGGGC	Forward amplification primer
TGTCCTTGGTGCCCGAG	Reverse amplification primer
TGTCCTTGGTGCCCGAGTG	RT primer
CTACACGACGCTCTTCCGATCT	
CGTACTGGACCTGGCTTACG	PCR addition of sequencing primer
CTACACGACGCTCTTCCGATCT	Pre-indexing forward primer
CAGACGTGTGCTCTTCCGATC	Pre-indexing reverse primer

qPCR primers

Sequence	Name
CACAGCGGATGTGAGGC	TSR-7SK qPCR primer forward
CTAGCCAGCCAGATCAGC	TSR-7SK qPCR primer reverse
GAATTGCGTCATTTAAAGCCTAGTT	MALAT1 qPCR forward
GTTTCATCCTACCACTCCCAATTAAT	MALAT1 qPCR reverse
GGATCCTAGACCAGCATGCC	MALAT1 qPCR forward
AAAGGTTACCATAAGTAAGTTCCAGAAAA	MALAT1 qPCR reverse
CCTCCCTTTAACTTATCCATTCAC	NEAT1 qPCR forward
TCTCTTCCTCCACCATTACCA	NEAT1 qPCR reverse
GCCTGTTTCAGAGGGTTGTT	NEAT1 qPCR forward
ATGCTGATCTGCTGCGTATG	NEAT1 qPCR reverse
GTGTTGGGCGCGTTATTTATC	fLuc qPCR 1
TAGGCTGCGAAATGTTCATACT	fLuc qPCR 1
CGGAAAGACGATGACGGAAA	fLuc qPCR 2
CGGTACTTCGTCCACAAACA	fLuc qPCR 2
TAACGCGGCCTCTTCTTATTT	rLuc qPCR 1
GATTTGCCTGATTTGCCCATAC	rLuc qPCR 1
CATGGGATGAATGGCCTGATA	rLuc qPCR 2
CAACATGGTTTCCACGAAGAAG	rLuc qPCR 2

References

1. Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* **79**, 137–58 (1944).

2. Colter, J. S. & Brown, R. A. Preparation of Nucleic Acids from Ehrlich Ascites Tumor Cells. *Science (80-.).* **124**, 1077–1078 (1956).

3. Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I. & Zamecnik, P. C. A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.* **231**, 241–57 (1958).

4. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).

5. Salditt-Georgieff, M. & Darnell, J. E. Further evidence that the majority of primary nuclear RNA transcripts in mammalian cells do not contribute to mRNA. *Mol. Cell. Biol.* **2**, 701–7 (1982).

6. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77–94 (2014).

7. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**, 155–159 (2009).

8. Kapranov, P. *et al.* RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science (80-.).* **316**, 1484–1488 (2007).

9. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–63 (2005).

10. de Hoon, M., Shin, J. W. & Carninci, P. Paradigm shifts in genomics through the FANTOM projects. *Mamm. Genome* **26**, 391–402 (2015).

11. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).

12. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).

13. Franke, A. & Baker, B. S. The rox1 and rox2 RNAs Are Essential Components of the Compensasome, which Mediates Dosage Compensation in Drosophila. *Mol. Cell* **4**, 117–122 (1999).

14. Brockdorff, N. *et al.* The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515–26 (1992).

15. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379**, 131–137 (1996).

16. Yang, Z., Zhu, Q., Luo, K. & Zhou, Q. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature* **414**, 317–22 (2001).

17. Nguyen, V. T., Kiss, T., Michels, A. A. & Bensaude, O. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature* **414**, 322–325 (2001).

18. Rinn, J. L. & Chang, H. Y. Genome Regulation by Long Noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).

19. Moore, P. B. & Steitz, T. A. The involvement of RNA in ribosome function. *Nature* **418**, 229–235 (2002).

20. Mili, S. & Steitz, J. a. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA* **10**, 1692–4 (2004).

21. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **289**, 905–20 (2000).

22. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3**, (2011).

23. Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.* **1**, 1610–1616 (2006).

24. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).

25. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).

26. Fang, R., Moss, W. N., Rutenberg-Schoenberg, M. & Simon, M. D. Probing Xist RNA Structure in Cells Using Targeted Structure-Seq. *PLOS Genet.* **11**, e1005668 (2015).

27. Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nat. Rev. Mol. Cell Biol.* **3**, 195–205 (2002).

28. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479–490 (2007).

29. Maris, C., Dominguez, C. & Allain, F. H.-T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* **272**, 2118–2131 (2005).

30. Dreyfuss, G., Swanson, M. S. & Piñol-Roma, S. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem. Sci.* **13**, 86–91 (1988).

31. Oubridge, C., Ito, N., Evans, P. R., Teo, C.-H. & Nagai, K. Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **372**, 432–438 (1994).

32. Lim, F. & Peabody, D. S. RNA recognition site of PP7 coat protein. *Nucleic Acids Res.* **30**, 4138–44 (2002).

33. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–49 (2014).

34. Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251 (2006).

35. Landry, J. P., Ke, Y., Yu, G.-L. & Zhu, X. D. Measuring affinity constants of 1450 monoclonal antibodies to peptide targets with a microarray-based label-free

assay platform. *J. Immunol. Methods* **417**, 86–96 (2015).

36.   Green, M. Avidin and streptavidin. in *Methods Enzymol* (eds. Meir, W. & Edward, A. B.) **Volume 184**, 51–67 (Academic Press, 1990).

37.   Suman Lata, †, Annett Reichel, †, Roland Brock, ‡, Robert Tampé, † and & Jacob Piehler*, †. High-Affinity Adaptors for Switchable Recognition of Histidine-Tagged Proteins. (2005). doi:10.1021/JA050690C

38.   Walker, S. C., Scott, F. H., Srisawat, C. & Engelke, D. R. RNA Affinity Tags for the Rapid Purification and Investigation of RNAs and RNA–Protein Complexes. *Methods Mol. Biol.* **488**, 23 (2008).

39.   Johansson, H. E. *et al.* A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 9244–9 (1998).

40.   Carey, J., Cameron, V., De Haseth, P. L. & Uhlenbeck, O. C. Sequence-specific interaction of R17 coat protein with its ribonucleic acid binding site. *Biochemistry* **22**, 2601–2610 (1983).

41.   Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* **454**, 126–130 (2008).

42.   Riley, K. J. & Steitz, J. A. The 'Observer Effect' in genome-wide surveys of protein-RNA interactions. *Mol. Cell* **49**, 601–4 (2013).

43.   Schimmel, P. R. & Budzik, G. P. *Photo-cross-linking of protein-nucleic acid complexes*. *Methods in Enzymology* **46**, (Elsevier, 1977).

44.   Li, S. & Mason, C. E. The Pivotal Regulatory Landscape of RNA Modifications. *Annu. Rev. Genomics Hum. Genet.* **15**, 127–150 (2014).

45.   RajBhandary, U. L. & Söll, D. *tRNA : structure, biosynthesis, and function*. (Washington (D.C.) : ASM press, 1995).

46.   *Modification and Editing of RNA*. (American Society of Microbiology, 1998). doi:10.1128/9781555818296

47.   Cantara, W. A. *et al.* The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Res.* **39**, D195–D201 (2011).

48.   Shatkin, A. J. Capping of eucaryotic mRNAs. *Cell* **9**, 645–653 (1976).

49.   Wang, Y. *et al.* N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell Biol.* **16**, 191–198 (2014).

50.   Saletore, Y., Chen-Kiang, S. & Mason, C. E. Novel RNA regulatory mechanisms revealed in the epitranscriptome. *RNA Biol.* **10**, 342–346 (2013).

51.   Ofengand, J. Ribosomal RNA pseudouridines and pseudouridine synthases. *FEBS Lett.* **514**, 17–25 (2002).

52.   Agris, P. F. Decoding the genome: a modified view. *Nucleic Acids Res.* **32**, 223–238 (2004).

53.   Gu, X. & Santi, D. V. The T-arm of tRNA is a substrate for tRNA (m5U54)-methyltransferase. *Biochemistry* **30**, 2999–3002 (1991).

54.   Madsen, C. T., Mengel-Jørgensen, J., Kirpekar, F. & Douthwaite, S. Identifying the methyltransferases for m(5)U747 and m(5)U1939 in 23S rRNA using MALDI mass spectrometry. *Nucleic Acids Res.* **31**, 4738–46 (2003).

55.   Lee, T. T., Agarwalla, S. & Stroud, R. M. Crystal Structure of RumA, an Iron-Sulfur

Cluster Containing E. coli Ribosomal RNA 5-Methyluridine Methyltransferase. *Structure* **12**, 397–407 (2004).

56. Kealey, J. T., Gu, X. & Santi, D. V. Enzymatic mechanism of tRNA (m5U54)methyltransferase. *Biochimie* **76**, 1133–42 (1994).

57. and, M. Y. K. & Redman*, K. L. RNA Methyltransferases Utilize Two Cysteine Residues in the Formation of 5-Methylcytosine†. (2002). doi:10.1021/BI026055Q

58. Alian, A., Lee, T. T., Griner, S. L., Stroud, R. M. & Finer-Moore, J. Structure of a TrmA–RNA complex: A consensus RNA fold contributes to substrate selectivity and catalysis in m5U methyltransferases. *Proc Natl Acad Sci U S A* **105**, 6876–6881 (2008).

59. Redman, K. L. Assembly of Protein–RNA Complexes Using Natural RNA and Mutant Forms of an RNA Cytosine Methyltransferase. *Biomacromolecules* **7**, 3321–3326 (2006).

60. Chen, Y. & Varani, G. Engineering RNA-binding proteins for biology. *FEBS J.* **280**, 3734–54 (2013).

61. Goers, E. S., Voelker, R. B., Gates, D. P. & Berglund, J. A. RNA Binding Specificity of *Drosophila* Muscleblind [†]. *Biochemistry* **47**, 7284–7294 (2008).

62. Chmiel, N. H., Rio, D. C. & Doudna, J. A. Distinct contributions of KH domains to substrate binding affinity of Drosophila P-element somatic inhibitor protein. *RNA* **12**, 283–91 (2006).

63. Maris, C., Dominguez, C. & Allain, F. H.-T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* **272**, 2118–2131 (2005).

64. De Franco, S., O'Connell, M. R. & Vandevenne, M. Engineering RNA-Binding Proteins by Modular Assembly of RanBP2-Type Zinc Fingers. in 57–74 (Humana Press, New York, NY, 2018). doi:10.1007/978-1-4939-8799-3_5

65. Cavaillé, J., Nicoloso, M. & Bachellerie, J.-P. Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides. *Nature* **383**, 732–735 (1996).

66. Zhao, X. & Yu, Y.-T. Targeted pre-mRNA modification for gene silencing and regulation. *Nat. Methods* **5**, 95–100 (2008).

67. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855–862 (1993).

68. Stark, A., Brennecke, J., Russell, R. B. & Cohen, S. M. Identification of Drosophila MicroRNA Targets. *PLoS Biol.* **1**, e60 (2003).

69. Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001).

70. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–97 (2004).

71. Hammond, S. M., Boettcher, S., Caudy, A. A., Kobayashi, R. & Hannon, G. J. Argonaute2, a Link Between Genetic and Biochemical Analyses of RNAi. *Science (80-.).* **293**, 1146–1150 (2001).

72. Strutt, S. C., Torrez, R. M., Kaya, E., Negrete, O. A. & Doudna, J. A. RNA-dependent RNA targeting by CRISPR-Cas9. *Elife* **7**, (2018).

73. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331–338 (2012).

74. Cléry, A., Blatter, M. & Allain, F. H.-T. RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.* **18**, 290–298 (2008).

75. Zamore, P. D., Williamson, J. R. & Lehmann, R. The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. *RNA* **3**, 1421–33 (1997).

76. Manna, S. An overview of pentatricopeptide repeat proteins and their applications. *Biochimie* **113**, 93–99 (2015).

77. Manival, X. RNA-binding strategies common to cold-shock domain- and RNA recognition motif-containing proteins. *Nucleic Acids Res.* **29**, 2223–2233 (2001).

78. 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).

79. Zhang, C. & Muench, D. G. A Nucleolar PUF RNA-binding Protein with Specificity for a Unique RNA Sequence. *J. Biol. Chem.* **290**, 30108–18 (2015).

80. Zhao, Y.-Y. *et al.* Expanding RNA binding specificity and affinity of engineered PUF domains. *Nucleic Acids Res.* **46**, 4771–4782 (2018).

81. Cheong, C.-G. & Hall, T. M. T. Engineering RNA sequence specificity of Pumilio repeats. *Proc. Natl. Acad. Sci.* **103**, 13635 LP – 13639 (2006).

82. Font, J. & Mackay, J. P. Beyond DNA: Zinc Finger Domains as RNA-Binding Modules. in *Methods in molecular biology (Clifton, N.J.)* **649**, 479–491 (2010).

83. Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M. & Thornton, J. M. Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.* **29**, 943–54 (2001).

84. Sprinzl, M., Hartmann, T., Weber, J., Blank, J. & Zeidler, R. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **17**, r1–r172 (1989).

85. Gu, X., Ivanetich, K. M. & Santi, D. V. Recognition of the T-Arm of tRNA by tRNA (m 5 U54) -Methyltransferase Is Not Sequence Specific †. **2960**, 11652–11659 (1996).

86. Urbonavicius, J., Jager, G. & Bjork, G. R. Amino acid residues of the Escherichia coli tRNA(m5U54)methyltransferase (TrmA) critical for stability, covalent binding of tRNA and enzymatic activity. *Nucleic Acids Res.* **35**, 3297–3305 (2007).

87. Gu, X., Ivanetich, K. M. & Santi, D. V. Recognition of the T-Arm of tRNA by tRNA (m 5 U54)-Methyltransferase Is Not Sequence Specific †. *Biochemistry* **35**, 11652–11659 (1996).

88. Agarwalla, S., Kealey, J. T., Santi, D. V. & Stroud, R. M. Characterization of the 23 S Ribosomal RNA m 5 U1939 Methyltransferase from *Escherichia coli*. *J. Biol. Chem.* **277**, 8835–8840 (2002).

89. Lee, T. T., Agarwalla, S. & Stroud, R. M. A Unique RNA Fold in the RumA-RNA-Cofactor Ternary Complex Contributes to Substrate Selectivity and Enzymatic Function. *Cell* **120**, 599–611 (2005).

90. Carlile, T. M., Martinez, N.M. Schaening, C., Su, A., Bell, T. A., Zinshteyn, B., and Gilbert, W. V. mRNA structure determines modification by pseudouridine synthase 1. *Nat. Chem. Biol.* **(in press)**, (2019).

91. Alexander, S. C., Busby, K. N., Cole, C. M., Zhou, C. Y. & Devaraj, N. K. Site-Specific Covalent Labeling of RNA by Enzymatic Transglycosylation. *J. Am. Chem. Soc.* **137**, 12756–12759 (2015).

92. McDonald, R. I. *et al.* Electrophilic activity-based RNA probes reveal a self-alkylating RNA for RNA labeling. *Nat. Chem. Biol.* **10**, 1049–54 (2014).

93. Yoon, J.-H., Srikantan, S. & Gorospe, M. MS2-TRAP (MS2-tagged RNA affinity purification): tagging RNA to identify associated miRNAs. *Methods* **58**, 81–7 (2012).

94. Lim, F. & Peabody, D. S. RNA recognition site of PP7 coat protein. *Nucleic Acids Res.* **30**, 4138–4144 (2002).

95. Lorenz, C., Lünse, C. E. & Mörl, M. tRNA Modifications: Impact on Structure and Thermal Adaptation. *Biomolecules* **7**, (2017).

96. Redman, K. L. Assembly of protein-RNA complexes using natural RNA and mutant forms of an RNA cytosine methyltransferase. *Biomacromolecules* **7**, 3321–3326 (2006).

97. Perumal, K., Sinha, K., Henning, D. & Reddy, R. Purification, Characterization, and Cloning of the cDNA of Human Signal Recognition Particle RNA 3′-Adenylating Enzyme. *J. Biol. Chem.* **276**, 21791–21796 (2001).

98. Yang, Z., Zhu, Q., Luo, K. & Zhou, Q. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature* **414**, 317–322 (2001).

99. Nguyen, V. T., Kiss, T., Michels, A. A. & Bensaude, O. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature* **414**, 322–325 (2001).

100. Hogg, J. R. & Collins, K. RNA-based affinity purification reveals 7SK RNPs with distinct composition and regulation. *RNA* **13**, 868–880 (2007).

101. Flynn, R. A. *et al.* 7SK-BAF axis controls pervasive transcription at enhancers. *Nat. Struct. Mol. Biol.* **23**, 231–8 (2016).

102. Aviv, H. & Leder, P. Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. *Proc. Natl. Acad. Sci. U. S. A.* **69**, 1408–12 (1972).

103. Niranjanakumari, S., Lasda, E., Brazas, R. & Garcia-Blanco, M. A. Reversible cross-linking combined with immunoprecipitation to study RNA–protein interactions in vivo. *Methods* **26**, 182–190 (2002).

104. Jensen, K. B. & Darnell, R. B. CLIP: Crosslinking and ImmunoPrecipitation of In Vivo RNA Targets of RNA-Binding Proteins. in 85–98 (Humana Press, 2008). doi:10.1007/978-1-60327-475-3_6

105. Kim, T. H. & Dekker, J. Formaldehyde Cross-Linking. *Cold Spring Harb. Protoc.* **2018**, pdb.prot082594 (2018).

106. Fahr, E., Furst, G., Dorhofer, G. & Popp, H. The structure of the uridine dimer produced by UV irradiation of uridine. *Angew. Chem. Int. Ed. Engl.* **6**, 250–251 (1967).

107. Arrigoni, L. *et al.* Standardizing chromatin research: a simple and universal method for ChIP-seq. *Nucleic Acids Res.* **44**, e67–e67 (2016).

108. Déjardin, J. & Kingston, R. E. Purification of Proteins Associated with Specific

Genomic Loci. *Cell* **136**, 175–186 (2009).

109. Ide, S. & Dejardin, J. End-targeting proteomics of isolated chromatin segments of a mammalian ribosomal RNA gene promoter. *Nat. Commun.* **6**, 6674 (2015).

110. Gelbart, M. E. & Kuroda, M. I. Drosophila dosage compensation: a complex voyage to the X chromosome. *Development* **136**, 1399–1410 (2009).

111. Ilik, I. A. *et al.* Tandem Stem-Loops in roX RNAs Act Together to Mediate X Chromosome Dosage Compensation in Drosophila. *Mol. Cell* **51**, 156–173 (2013).

112. Cheetham, S. W. & Brand, A. H. RNA-DamID reveals cell-type-specific binding of roX RNAs at chromatin-entry sites. *Nat. Struct. Mol. Biol.* **25**, 109–114 (2018).

113. Flynn, R. A. *et al.* 7SK-BAF axis controls pervasive transcription at enhancers. *Nat. Struct. Mol. Biol.* **23**, 231–8 (2016).

114. Lebars, I. *et al.* HEXIM1 targets a repeated GAUC motif in the riboregulator of transcription 7SK and promotes base pair rearrangements. *Nucleic Acids Res.* **38**, 7749–7763 (2010).

115. Parker, R. & Song, H. The enzymes and control of eukaryotic mRNA turnover. *Nat. Struct. Mol. Biol.* **11**, 121–127 (2004).

116. Yin, Q.-F. *et al.* Long Noncoding RNAs with snoRNA Ends. *Mol. Cell* **48**, 219–230 (2012).

117. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

118. Wickham, H. *Ggplot2 : elegant graphics for data analysis*.

119. R Core Team. R: A Language and Environment for Statistical Computing. (2019).

120. Wei, T. & Simko, V. R package 'corrplot': Visualization of a Correlation Matrix. (2017).

121. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

122. Dey, K. K., Xie, D. & Stephens, M. Dey, K.K., Xie, D. and Stephens, M., 2017. A new sequence logo plot to highlight enrichment and depletion. *bioRxiv, p.226597* (2017).

123. Sexton, A. N., Wang, P. Y., Rutenberg-Schoenberg, M. & Simon, M. D. Interpreting Reverse Transcriptase Termination and Mutation Events for Greater Insight into the Chemical Probing of RNA. *Biochemistry* **56**, 4713–4721 (2017).