METHODS OF PARTITIONING, BIOGENESIS, AND SELECTING FOR NATURAL AND ENGINEERED COA-RNA

A Dissertation

presented to

The Faculty of the Graduate School

At the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Jordyn Kaye Lucas

Dr. Donald H. Burke-Agüero, Dissertation Advisor

May 2023

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

METHODS OF PARTITIONING, BIOGENESIS, AND SELECTING FOR NATURAL AND ENGINEERED COA-RNA

presented by Jordyn K Lucas

a candidate for the degree of Doctor of Philosophy of Biochemistry,

and hereby certify that, in their opinion, it is worthy of acceptance.

_____

Dr. Donald Burke-Agüero

_____

Dr. James Amos-Landgraf

_____

Dr. Xiao Heng

_____

Dr. Lloyd Sumner

DEDICATION

I dedicate this work to my mother, Lezlee Kaye Raiford. Words hardly begin to describe the overwhelming appreciation and love I feel towards you. You have always been my number one supporter and source of guidance. You taught me to embrace challenges, never fear failure, and the value of hard work and perseverance. Above all else, you encouraged and believed in me when I was at my lowest, giving me the strength to push forward. Thank you for the many years of support, sacrifices, and unconditional love. I am thankful everyday to have you as my parent.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude for my advisor, Dr. Donald Burke-Agüero, for his feedback, guidance, and teachings that made me into the scientist I am today. This endeavor would not have been possible without my doctoral committee, who generously provided their knowledge and expertise through the years. I am also thankful to Dr. Margaret Lange for her advise and encouragement. Many thanks to Dr. Rick Ryan for serving as a professional mentor since my days as an undergraduate student.

I am also grateful to my peers in the Burke lab for many wonderful memories working in lab, their frequent help and feedback, and significant moral support. Dr. Kwaku Tawiah's infectious joy made even my most stressful days a little bit brighter and with her sharp humor and brilliant mind, Dr. Phuong Nguyen was an unforgettable friend and mentor. I must give special thanks to Dr. Paige Gruenke for being an incredible friend (and peer) for the duration of my graduate experience. Thank you for answering every late-night phone call, always double-checking my math, pushing me to try to new things (like running a half-marathon), and several years of joy and friendship. I'm also extremely grateful to my childhood friend, Courtney Mellanby, for always supporting and uplifting me during my toughest times.

Lastly, I would be remiss in not mentioning my family, especially my parents, Lezlee Raiford and John Lucas, and my sister, Dana Lucas. Their encouragement over the years has kept my spirits and motivation high, especially during stressful times. I could not have

undertaken this journey without their love and support. I'd also like to thank my Aunt Kathy for her uncanny timing of supportive messages and advice throughout the years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

More than a decade ago, RNAs with NAD+, CoA, and acylated CoA caps were identified. Since then, studies have described NAD+'s protective, cap-like function in bacteria and its role in promoting mRNA degradation in eukaryotic cells. However, the identities, functional roles, and mechanisms of biogenesis of CoA-RNA have not yet been explored. NAD-RNAs are generated primarily by co-transcriptional capping where NAD+ is inserted into the +1 position of transcripts in place of ATP. However, this co-transcriptional model is unlikely to generate CoA-RNAs in cells because the required non-canonical initiator nucleotide for co-transcription, 3' dephospho CoA (dpCoA), is estimated to be ~200 fold less concentrated in cells than NAD+ and is therefore unlikely to outcompete ATP for the +1 position of transcripts. Thus, this work demonstrates that post-transcriptional capping by enzyme phosphopantetheine adenylyltransferase (PPAT) is a possible mechanism to generate CoA-RNA. Additionally, because having a reliable method to partition CoA-RNAs from other total RNA is a crucial step for studying and making use of them, this work describes the development of a CoA Capture Seq method for separating CoA-RNAs from total RNA. Although the CoA Capture Seq method described in this work needs further optimization before it is suitable for identifying endogenous CoA-RNAs, it was adapted and used successfully to establish *in vivo* post-transcriptional capping of RNAs by PPAT to generate CoA-RNAs.

I further investigated methods of *in vivo* biogenesis of CoA-RNA by performing a selection under *in vivo* like conditions to select for RNAs capable of capping themselves with 4' phosphopantetheine (pPant) to become CoA-RNAs (CoAzymes) and RNAs which serve

as the best substrate to be capped enzymatically by PPAT. The selection conditions were designed to mimic intracellular conditions (neutral pH, fewer number of ions included, lower ion concentrations, etc) to increase the probability of selecting for RNAs which retained functionality in cells. Before starting the selection, five different reverse-transcriptases (RTs) were tested and optimized under various reaction conditions with RNA library templates of varying structure, to determine which RT introduced the least amount of inter-library bias. The RT analysis revealed that BST 3.0 DNA Polymerase was the best choice for the RT step of the selection due to its excellent processivity, significant yield, and low-inter library bias. After 12 rounds of selection, no significant increase in CoAzyme or PPAT capping activity was observed, thus selection rounds were prepared for HTS to evaluate the library pool's progression throughout the course of the selection. Unfortunately, the HTS analysis revealed no convergence or enrichment of specific sequences or clusters of sequences. Additionally the diversity of sequence reads in each round was also inconsistent and the enrichment analysis revealed the inconsistencies in population structure throughout the selection. These data suggest the selection failed, which is likely related to the overly stringent selection parameters, especially the buffer conditions. Overall, this work illustrates the importance of selection parameters, especially the selection buffer and using RTs that introduce minimal bias, for successful selection outcomes.

CHAPTER 1: Introduction and review of the literature

Nucleotide Cofactors & Metabolite-linked RNAs

RNA is an essential biological molecule which plays roles in protein translation, catalysis, regulation of gene expression, and more. The capping and chemical modification of the 5' end of RNA affect RNA translation efficiency, localization, stability, and processing (1, 2). In eukaryotes, the 5' 7-methylguanylate cap is responsible for mRNA stability, translation, and export from the nucleus, as well as processing events such as poly(A) tailing (3). RNA capping was previously thought to be unique to eukaryotes, until recent studies observed bacteria and archaea making use of metabolic cofactors NAD^+, CoA, and CoA-thioesters to generate metabolite capped-RNAs (4–6). Since then, several studies have investigated the identities, functions, and mechanisms of biogenesis of NAD-RNA. However, CoA-RNAs largely remain a mystery in the field.

Cells from all three Domains of life generate several nucleotide analogs such as cyclic adenosine monophosphate (cAMP), nicotinamide adenine dinucleotide (NAD+), and coenzyme A (CoA). These analogs play pivotal roles as signaling molecules, energy carriers, and enzyme cofactors. Nucleotide cofactors with free 3' hydroxyl groups (e.g., NAD+, FAD, and 3' dephospho CoA) (Figure 1.1) were also found suitable as non-canonical initiator nucleotides (NCINs) for *in vitro* RNA transcription (7, 8) to generate CoA-RNA, NAD-RNA, and FAD-RNA, representing a possible mechanism of cofactor-RNA biogenesis *in vivo*.

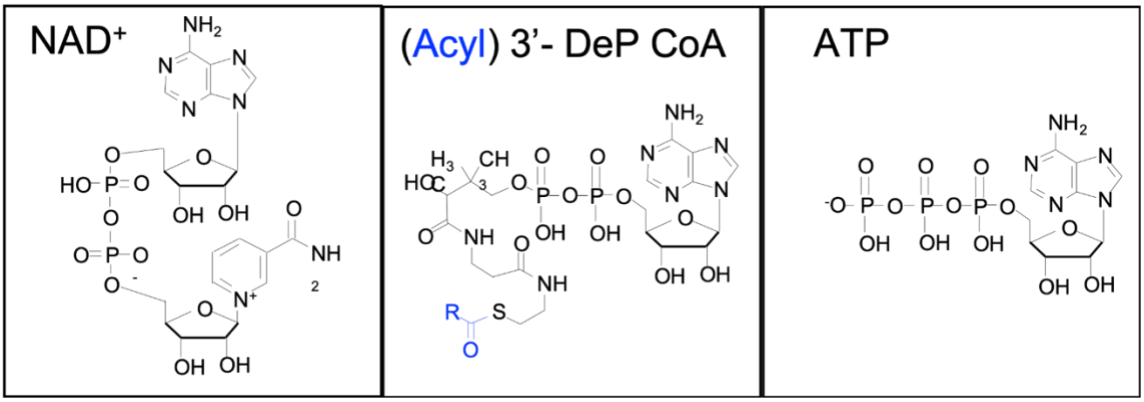


Figure 1.1 Initiator nucleotides. Metabolite cofactors and non-canonical initiator nucleotide NAD+ is shown on the left. Acylated 3'-dephospho CoA (dpCoA) (middle) is a non-canonical initiator nucleotide which can be acylated (blue). dpCoA differs only slightly from the metabolite cofactor Coenzyme A which carries a 3' phosphate group. ATP (right) is a canonical initiator transcript of RNA.

NAD-RNA

NAD-RNA was initially discovered by fully digesting isolated and size fractionated cellular RNAs with nuclease P1 into mononucleotides and using a mass spectrometry based technique to identify the NAD modification (4). However, as a result of the required hydrolysis step, Chen et al. could not identify the specific sequences of RNAs carrying the NAD modification. Thus, soon after their initial discovery, a 'NAD-Capture Seq' method (Figure 1.2) was developed to isolate NAD-RNAs from total cellular RNA and determine the sequence identities of NAD-RNAs, (9). This method utilized adenosine diphosphate-ribosylcyclase (ADPRC) to catalyze the selective transglycosylation reaction of NAD+ with "clickable" alkynyl alcohols. The transglycosylated RNA was then biotinylated, captured on streptavidin beads, and prepared for high-throughput sequencing by reverse transcription and PCR amplification (Figure 1.2). This NAD Capture Seq method was used by many groups to identify RNAs carrying NAD+ caps across all three domains of life (6, 10–13).

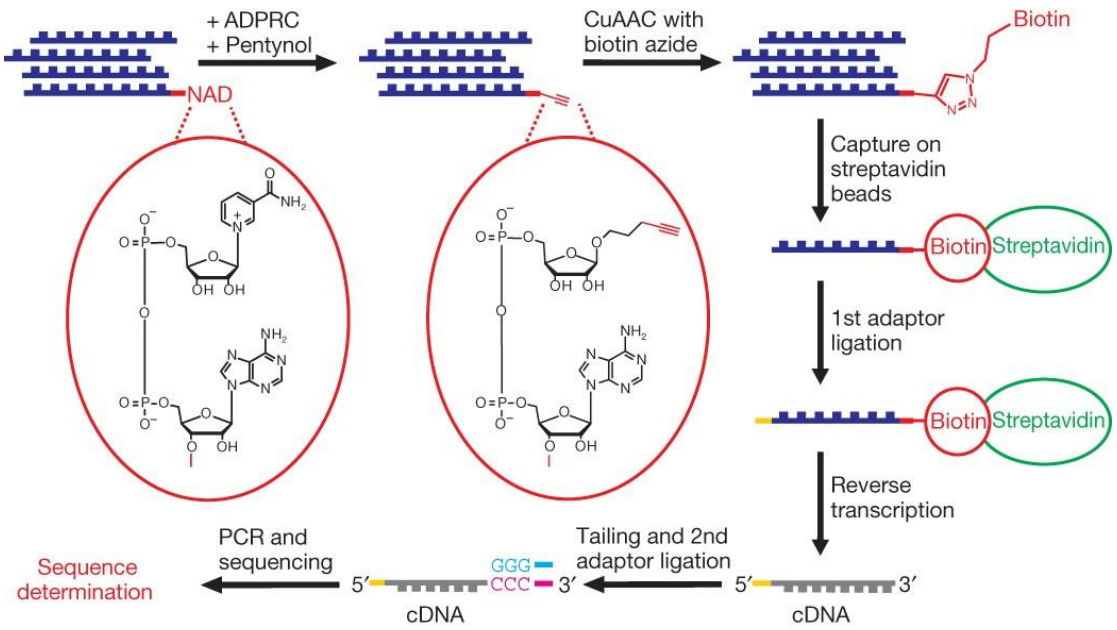


Figure 1.2. NAD+ Capture Seq. This figure was adapted from Cahová et al. NAD-RNAs have an NAD moiety that is transglycosylated by ADPRC. Click chemistry is used to biotinylate transglycosylated RNA products. Biotinylated products are captured on beads and prepped for HTS.

Biogenesis

Co-transcriptional (10, 11, 14, 15) and post-transcriptional (4, 5, 12) mechanisms have been proposed to explain *in vivo* RNA capping with NAD+ (Figure 1.3). Although there is strong precedent for the existence of both methods of NAD-RNA biogenesis, co-transcriptional capping mechanisms are well-established and have significant supporting data. Whereas post-transcriptional mechanisms of NAD+ capping remain largely speculative and unsupported by published data.

Co-transcriptional capping occurs during RNA transcription, when NAD+ (or another NCIN) is substituted for ATP, the canonical nucleotide, and is incorporated as the +1 position, generating NADylated RNA. Previous studies determined several RNA polymerases, including *E. coli* RNAP, T7 RNAP, and human mitochondrial RNAP (8, 14, 16–19) are capable of co-transcriptionally capping RNA transcripts with adenosine cofactors such as NAD+, FAD, and dpCoA. Intracellular NAD+ levels in *E. coli* typically fluctuate between 4-7 mM while intracellular ATP levels typically fluctuate between 1-5 mM. Furthermore, RNA polymerase in *E. coli* has a K_m of ~0.38 mM for NAD+, about 10 times lower than intracellular NAD+ levels (20). Although RNA polymerase in *E. coli* has a K_m of ~0.090 mM for ATP, its canonical substrate, it is still feasible for NAD+ to both be substituted for ATP and be incorporated by the RNA by the polymerase into the +1 position of RNA transcripts. The ~4x difference in RNAP affinities for NAD+ and ATP and their similar intracellular concentrations predict around a quarter of transcripts with +1A to have an NAD+ cap. Interestingly, the Jaschke group observed approximately 25% NAD-capping for RNAI transcripts in bacterial cells with NAD+ de-capping enzyme

NudX knocked out (9), corresponding closely with the predicted quantities of cellular NAD-RNA, further supporting co-transcriptional model as the primary mechanism for NAD-RNA biogenesis. Furthermore, the promoter sequences of RNA transcripts were found to significantly impact the efficiency of NAD+ capping, where NCIN initiation with adenosine analogs only took place at +1 A promoters (14). Bird et al. also determined that the – 1 position of promoters also impacts capping efficiency.

Identification & Functional roles

Cahová et al isolated RNA from bacterial cells to identify NAD-RNAs using their NAD+ Capture Seq method. The enriched NAD-RNA sequences consisted of mostly of sRNAs and 5' fragments of mRNAs encoding for proteins involved in cellular metabolism, stress response pathways, and lesser known proteins (9). The most abundantly NAD-capped RNA in *E. coli* was RNAI, an sRNA for replication control of ColE1 plasmid (21), where ~13% of RNAI transcripts carried an NAD+ cap. Interestingly, the bacterial NAD+ cap conferred resistance to RNA degradation by RNase E and phosphohydrolases, indicating that bacterial NAD+ caps serve a protective role. Additionally, it was reported that NudC, a nudix phosphohydrolase which hydrolyses NAD(H) into NMN(H) and AMP (22), serves as an NAD decapping enzyme (Figure 1.3). It's possible that the specific decapping of NAD-RNAs by NudC could be a mechanism for selective degradation of capped RNAs. Frindert et al. also elucidated a co-transcriptional mechanism for generating NAD-RNA *in vivo* whereby NAD+ serves as a non-canonical initiator nucleotide (NCIN) and is incorporated into the +1 position of transcripts. NAD-RNAs in gram positive bacteria *B. subtilis* was also isolated and identified, with the majority of NAD+ capped RNAs

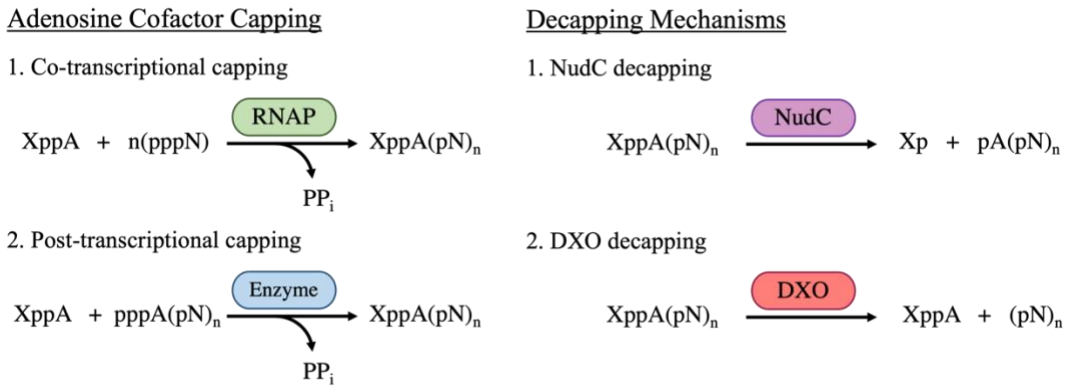


Figure 1.3. Capping and decapping mechanisms for cofactor-RNAs. Possible capping mechanisms (left) for NAD+, dpCoA, and other adenosine analog cofactors. Specific decapping mechanisms (right) reported for cellular NAD-RNAs. The 'X' of XppA in the decapping mechanisms (right) represent nicotinamide. This figure is adapted from Julius et al (19).

recognized as mRNAs (11). Interestingly, NAD-caps in *B. subtilis* conferred resistance against RNA degradation, similar to previous observations in *E. coli*, implying NAD+ caps serve protective roles in bacteria.

Several studies also investigated and established the presence of NAD-RNAs in eukaryotes. In *S. cerevisiae*, specific nuclear and mitochondrial mRNAs were enriched for carrying NAD+ caps (13). NAD+ caps in *S. cerevisiae* were also found on pre-mRNAs and mature mRNAs with introns, suggesting a co-transcriptional model for biogenesis. Additionally, Walters et al. observed NAD+ capping of mitochondrial transcripts known to not undergo any 5' end processing, further supporting a co-transcriptional mechanism for NAD-RNA biogenesis. A large number (up to 6,000) NAD+ capped RNAs were identified in *Arabidopsis* (10, 15). These NAD-RNAs were widespread across the *Arabidopsis* genome and their abundance was highly correlated to the levels of free cellular NAD+, consistent with a co-transcriptional mechanism of NAD capping. For example, little to no chloroplast NAD-RNA was identified, consistent with low levels of NAD+ in chloroplasts. On the other hand, mitochondrial NAD-RNA were abundant, consistent with the largest intracellular NAD+ pool (~2mM) being located in the mitochondria (23). Thus, these data support of co-transcriptional method of NAD incorporation during transcription, rather than a post-transcriptional processing mechanism.

The NAD-RNAs isolated from mammalian cells (HEK293T, human kidney tissue) revealed preferential NAD+ capping of small nucleolar (snoRNA) and small Cajal body RNAs (scaRNAs) (12). Jiao et al. also reported DXO decapping enzymes modulated

cellular levels of NAD-RNAs (Figure 1.3). Interestingly, NAD+ caps were observed on intronic sno/scaRNAs which normally contain a monophosphate on their 5' end, which some believed could be indicative of an alternate NAD-capping mechanism. Specifically, these data could indicate that mammalian cells may make use of an alternate, post-transcriptional capping mechanism, different from co-transcriptional methods, to generate NAD-RNAs. However, these RNAs are generated by eukaryotic RNA Polymerase II which has been shown to generate NAD-RNAs co-transcriptionally *in vitro* (24). Additionally, other groups demonstrated a direct correlation between cellular NAD+ levels in human cells and the levels of NAD-capped RNA (25), which is more consistent with a co-transcriptional method of NAD+ capping, and no further data exists to support a post-transcriptional mechanism at this time.

Finally, NAD-RNAs were also found in Archaea, suggesting the evolutionary conservation of NAD+ capping across all three domains of life (6). NAD-RNA was more abundant in *m. bakeri* cultures where RNA was isolated during stationary phase. The increased levels of NAD+ capping during stationary phase is hypothesized to be a result of higher cellular [NAD]:[ATP] ratios. During stationary phase, intracellular ATP concentrations decrease and intracellular NAD+ concentrations increases which leads to an increase in NAD+ capping by co-transcriptional mechanisms (14, 26). These data further support a co-transcriptional mechanism of NAD-RNA biogenesis.

CoA-RNA

In contrast to the expansive work done to elucidate the identities, roles, and biogenesis mechanisms of NAD-RNA, virtually no additional information concerning biological CoA-RNAs has been published following their initial discovery in 2009 (5). The identities, functional roles, and mechanisms to generate natural CoA-RNAs remain elusive more than a decade after their debut in the field. NAD-RNAs, on the other hand, were a simpler research target for two reasons: 1. NAD-RNAs are more abundant in cells than CoA-RNAs and 2. An NAD Capture Seq method for isolating NAD-RNAs from total RNA was established (9). Thus CoA-RNAs have not received the same level attention in part due to lack of a well-established methods for capturing and sequencing CoA-RNAs from total cellular RNA. Also, unlike NAD-RNA which is fairly abundant in cells, intracellular levels of CoA-RNA are hypothesized to be much lower. This creates an additional challenge: a successful CoA Capture Seq methods needs to be extremely effective at capturing and detecting very small quantities of CoA-RNA.

Capture methods

Other groups, including the Huang group, verbally concede that they have tried (and failed) to isolate and sequence cellular CoA-RNAs, but few groups are willing to report the details of such failures in their publications. However, one study which used a mass spectrometry based technique, dubbed 'CapQuant', provided some insights on cellular CoA-RNAs (27). While developing the CapQuant method, several controls were used to determine the limit of detection (LOD) for each capping molecule including 5 'metabolite' caps: NAD+, FAD, UDP-glucose (UDP-Glc), UDP-N-acetylglucosamine (UDP-GlcNAc), and 3' dephospho

CoA (dpCoA). NAD+ had a LOD of ~0.72 fmols whereas dpCoA had a LOD of ~2.3 fmols. For context, the estimated amount of NAD-capped Ms1 (non-coding small RNA) was ~155 fmol in myobacteria *M. smegmatis* (6), more than 200 fold higher than the limit of detection for NAD+ caps. Thus, any successful CoA Capture Seq method would need to be extremely effective at capturing and detecting very small quantities of CoA-RNA.

Biogenesis

The primary mechanism for cellular NAD-RNA formation is co-transcriptional non-canonical nucleotide initiation (14), a mechanism in which NAD+ is incorporated as the +1 nucleotide generating an RNA with an NAD+ 'cap'. Intracellular NAD+ levels in *E. coli* typically fluctuate between 4-7 mM while intracellular ATP levels typically fluctuate between 1-5 mM. Furthermore, RNA polymerase in *E. coli* has a K_m of ~0.38 mM for NAD+, about 10 times lower than intracellular NAD+ levels (20) hence it is feasible for NAD+ to both be substituted for ATP and be incorporated by the RNA by the polymerase into the +1 position of RNA transcripts. However, Coenzyme A cannot serve as an NCIN or be incorporated into the +1 position of transcripts due to the phosphate on the 3' hydroxyl. In fact, the NCIN for CoA-RNAs is dpCoA (Figure 1.1), a metabolic precursor in the CoA synthesis pathway that is present in small quantities in the cell. One study which looked at the anaerobic bacterium *C. kluyveri* reported dpCoA and NAD+ intracellular concentrations to be ~20μM and 12 mM respectively (28). Although the dpCoA concentration in *E. coli* is unknown, if it is comparable to quantities measured in *C. kluyveri*, low levels of dpCoA would make competing with ATP for the +1 spot of RNA transcripts difficult. Therefore, non-canonical nucleotide initiation by dpCoA is most likely

not a significant mechanism for generating cellular CoA-RNAs. Furthermore, even if dpCoA occasionally served as a NCIN for co-transcriptional CoA capping of transcripts, the generated CoA-RNAs would be well below the limit of detection for existing CoA-Capture Seq methods based on the hypothesized intracellular levels of dpCoA.

CoA-RNAs could also be generated through post-transcriptional capping by an enzyme called phosphopantetheine adenylyltransferase (PPAT) (Figure 1.3). PPAT is part of the CoA biosynthesis pathways and is responsible for turning substrates ATP and phosphopantetheine (pPant) into dpCoA. Several of the enzymes in the CoA biosynthesis pathways, including CoaA and PPAT, have demonstrated relaxed substrate specificities by allowing a range of modifications on pantetheine possible (29–31). Therefore, it is possible that PPAT could use a non-canonical substrate in place of ATP, for instance ATP-RNA. Thus, PPAT could be responsible for post-transcriptional capping mechanisms of CoA-RNAs, which is a possibility that is explored in Chapter 3.

Selection Conditions & Library Design

Systematic Evolution of Ligands through EXponential enrichment (SELEX) is an *in vitro* evolution technique used to identify new aptamers, ribozymes, and other functional nucleic acid modules (32, 33). Selections begin with the synthesis of a oligonucleotide starting library of fixed length that consists of 5’ and 3’ constant regions and a randomized region in between. If the starting library is RNA, the library will be transcribed, incubated with a specific ligand, and active/bound RNA will be partitioned from unactive sequences and

amplified (Figure 1.4). This process is repeated for several rounds before libraries from various rounds are sequenced to determine the identities of active RNAs.

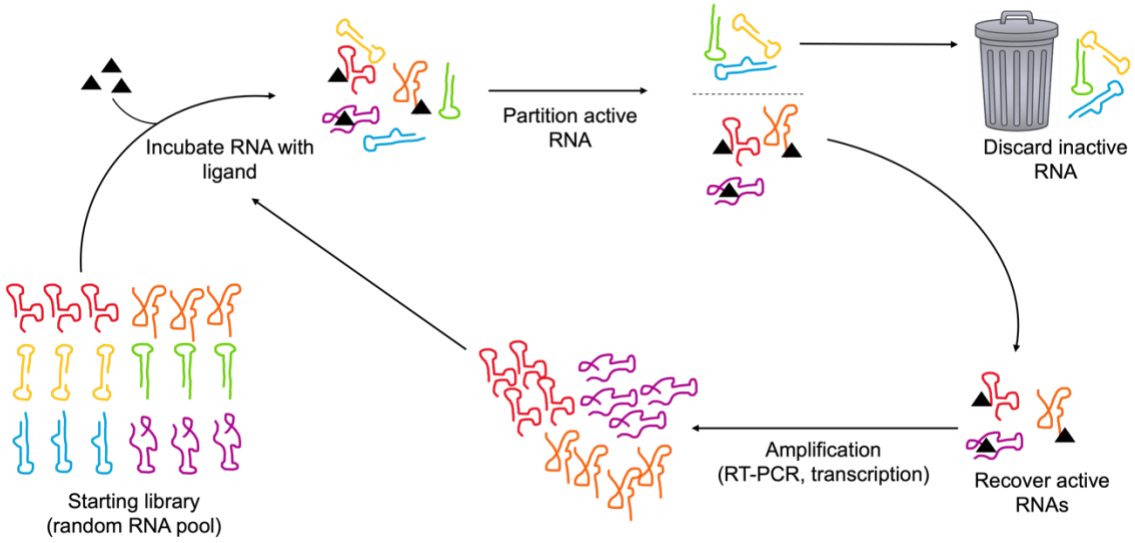


Figure 1.4. Schematic of SELEX. A randomized starting library is incubated with a ligand. Active sequences are partitioned and inactive sequences are discarded. After recovery of active sequences, the sequences are amplified (RT-PCR and transcription) before going into the next round of selection. This process is repeated for several rounds.

Although the underlying concepts behind SELEX are straightforward, preparation and successful completion of selections can prove challenging. The conditions in which a selection is performed generally establish limitations for a selected molecule's activity. Additionally, library designs at the onset of a selection can shape selection outcomes, and a successful selection is often directly related to the structural and sequence diversity of the starting library and robust amplification to uphold it (34, 35). Finally, amplification biases, especially reverse transcription, when left unchecked can sabotage a selection to the point of failure. Thus, there are many important factors to consider beyond the binding target or desired activity when preparing for a selection.

Selection conditions

The intended use of aptamers or ribozymes after they are selected, should be carefully considered when designing a selection. Specifically, it is important to consider the selection conditions because selecting in an environment that is vastly different from the intended reaction environment, may result in reduced or no activity (36). For example, selecting for ribozymes in buffers containing 20 mM $MgCl_2$ is a poor selection strategy if the selected ribozymes are intended for *in vivo* purposes where free magnesium concentrations are ~ 300 μM. Furthermore, the conditions used during a selection often determine the conditions for optimal activity of a selected RNA or DNA molecules. Thus, it is important to carefully consider the intended reaction conditions for selected molecules when choosing the selection conditions.

Choosing selection conditions becomes particularly tricky when the selected molecules are intended for *in vivo* use, as is increasingly the case. Some of the important selection conditions to consider are selection buffers and reaction temperature. For selection buffers, some of the key factors are pH of the buffer, the number of included ions, the concentrations of ions, and crowding reagents. pH plays a significant role for both aptamer and ribozyme selections as it influences the protonation of bases which may be involved in catalysis or structural formation (37). Ions play crucial roles in selections by stabilizing RNA structures and improving catalysis for ribozymes. Previous studies have established that different ions are beneficial for different types of RNA structures. For example, potassium ions are reported to be required for the stable formation of G-Quadruplexes, however potassium and sodium have also been reported to weaken electrostatic forces which can reduce non-specific binding when they are present low concentrations or disrupt important, specific electrostatic interactions at high concentrations (38–40). Magnesium can stabilize RNA-ligand interactions and hairpin and helical structures. Magnesium has also been demonstrated to play crucial roles in catalysis. For instance, a study which examined the mechanism of catalysis by a hammerhead ribozyme observed a linear relationship between activity and the magnesium concentration (41). However, not all ions are beneficial in a selection. One study reported that some metal ions, including Fe^{3+} and Al^{3+}, had a negative impact on selections and were observed to disrupt both RNA structure and function (42). The use of crowding reagents in a selection buffer can simulate cellular folding conditions for the RNA, allowing for secondary and tertiary structures to form which may not be achievable in buffer only. Previous studies demonstrated that larger PEGs stabilized compact RNA structures and strengthened ligand-binding as compared to

RNAs in buffer only (43). Considering the complexity and many variables for selection buffers, each element of the buffer should be thoughtfully considered within the context of the final purpose of the selected molecules.

The chosen reaction conditions are also very important during a selection. Temperature can also prove either beneficial or detrimental: higher temperature may promote improved ribozyme catalysis, however, it can also produce higher RNA degradation. Interestingly, it was reported that increasing the reaction temperature had a significant impact on the K_d of selected aptamers, with a reaction temperature of 37◦C being most favorable (39). Thus, it is especially important to consider biologically relevant temperatures for the selections of aptamers and ribozymes with intended *in vivo* use.

Structured libraries

Many library designs for various types of selections utilize a fully randomized region flanked by constant regions on the 5' and 3' ends. Functional RNAs with binding (aptamers) or catalytic (ribozymes) activity are typically more structured than random RNA sequences. Experimental and computational studies of nucleic acid libraries support the correlation between increased secondary structure and increased function (44–49). As a result, functional RNAs have high information content, deriving from information required to specify both generic structural elements and uniquely specified nucleotides that are often found in regions responsible for activity, such as binding sites or catalytic active sites. However, when starting a selection from a fully randomized library, the probability of key nucleotides being independently specified within an appropriate structural context

is low. Incorporating basic structural elements into a starting library randomized region allows exploration of sequence space within a context that already contains those structural elements which ultimately increase the fraction of the starting library that can't actually attain the desired function. Thus incorporating structural elements into a starting library can actually increase the probability of a successful selection.

Several groups have reported successful selection outcomes arising from starting libraries which contained predefined structural elements (44, 48–55). For example, a selection performed against neurotransmitter precursors using starting libraries that incorporated three-way junction (3WJ) scaffolds derived from natural and synthetic aptamers revealed that aptamers that preserved the 3WJ scaffold showed higher specificity and affinity for their ligands than those that lost this feature (54). Another study showed when a partially structured library containing a stable stem-loop was directly compared with a fully randomized RNA library during a selection for aptamers to GTP, more aptamers emerged from the partially structured library, and those that retained the incorporated stem-loop structure displayed some of the highest affinities (44). These and other studies demonstrate that incorporating structural elements into starting library pools provides and opportunity to enrich the selection pool for active RNAs. However, amplification biases against structured templates can negate the benefits from structured libraries, thus minimizing these biases is crucial for all selections and especially for those using structured library designs.

Amplification Bias

Although there is a clear correlation between RNA structure and function, highly structured RNA is susceptible to amplification biases, particularly during reverse transcription. During a selection, this bias can discriminate against structured RNAs even if they have superior performance in the functional step of a given selection. Thus, reverse transcription bias can sabotage selections to the point of failure by discriminating against the very sequences that are being selected for.

Polymerase processivity and fidelity are the primary characteristics responsible for reverse transcription bias. Low RT processivity is especially detrimental to highly structured templates and can lead to incomplete readthrough or low yield of full length product during cDNA synthesis that precludes subsequent PCR (56, 57). Although the RT bias can be reduced against highly structured templates by increasing reaction temperatures to partially denature structured regions, most RTs do not perform well at elevated temperatures which in turn impacts the cDNA yield (58). On the other hand, low polymerase fidelity causes nucleotide misincorporations leading to an accumulation of mutations that favor amplification rather than the intended function (59). Furthermore, mutations that disrupt the designed structural regions will result in more efficient cDNA synthesis by RTs that struggle with structured templates and allow these less structured mutants to dominate the selected pool. In one instance where this phenomenon was previously observed with SuperScript III (SSIII), the amplification bias was resolved by switching to an RT that favored readthrough of structured RNAs (54). Thus, even low levels of amplification bias

can cripple the best library designs and selection outcomes, thus the choice of RT for selections is critical.

Outline of dissertation

Chapter 2 focuses on minimizing amplification bias during reverse transcription (RT) for *in vitro* selections. I directly compared five different reverse-transcriptases on highly-structured RNA templates susceptible to RT bias to determine which enzymes introduced the least bias. This work demonstrates BST 3.0 DNA Polymerase introduced little bias among structured templates, exhibited excellent processivity, and generatec large quantities of full length cDNA product. I also performed an *in vitro* selection to compare how different RTs impacted selection outcomes and determined BST had low inter-library and mutational bias during the selection, making it an excellent choice for RT reactions and *in vitro* selections. This work was done with the assistance of Paige R. Gruenke and under the supervision of Dr. Donald Burke-Agüero. This work has been submitted and is under revision at *RNA.*

Chapter 3 reports a possible post-transcriptional capping mechanism to generate CoA-RNAs. This work established *in vitro* capping of CoA-RNA by the PPAT enzyme and the RNA substrate requirements for capping. I utilized the CoA Capture Seq method and observed preferential *in vivo* PPAT capping of RNA that met the established *in vitro* capping requirements. This work was performed with Dr. Krishna Sapkota and Matt F. Lichte and under the supervision of Dr. Faqing Huang and Dr. Donald Burke-Agüero. The work will be submitted as a manuscript for publication.

Chapter 4 describes a method development for the separation and identification of cellular CoA-RNAs. I detailed a nine-step process for purifying cellular RNA, partitioning CoA-RNAs from total RNA, and preparing RNA for high-throughput sequencing and discuss steps in need of further improvement. This method was also modified and used to isolate and partition CoA-RNAs in chapter 3. This work was performed with the assistance of Matt F. Lichte and under the supervision of Dr. Donald Burke-Agüero. This work will be adapted for submission as a method development paper.

Chapter 5 describes an *in vivo* like selection using structured libraries to identify RNAs which can self-cap with CoA and RNAs that serve as the best substrates to be capped by enzyme PPAT. After many rounds of selection, I prepared several selection from each trajectory for high-throughput sequencing (HTS). The HTS analysis revealed that no convergence or enrichment of specific sequences occurred even after 12 selection rounds, indicating selection conditions may have been too stringent. This work was performed under the supervision of Dr. Donald Burke-Agüero.

Chapter 6 is a perspective on the capture methods, identities, and mechanisms of biogenesis of CoA-RNAs and impact of selection conditions and library designs on selection outcomes.

REFERENCES

1. Topisirovic, I., Svitkin, Y. V., Sonenberg, N., and Shatkin, A. J. (2011) Cap and cap-binding proteins in the control of gene expression. *Wiley Interdiscip. Rev.*

RNA. **2**, 277–298

2. Hui, M. P., Foley, P. L., and Belasco, J. G. (2014) Messenger RNA degradation in bacterial cells. *Annu. Rev. Genet.* **48**, 537–559
3. Cowling, V. H. (2010) Regulation of mRNA cap methylation. *Biochem. J.* **425**, 295
4. Chen, Y. G., Kowtoniuk, W. E., Agarwal, I., Shen, Y., and Liu, D. R. (2009) LC/MS analysis of cellular RNA reveals NAD-linked RNA. *Nat. Chem. Biol.* **5**, 879–881
5. Kowtoniuk, W. E., Shen, Y., Heemstra, J. M., Agarwal, I., and Liu, D. R. (2009) A chemical screen for biological small molecule-RNA conjugates reveals CoA-linked RNA. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7768–7773
6. Ruiz-Larrabeiti, O., Benoni, R., Zemlianski, V., Hanišáková, N., Schwarz, M., Brezovská, B., Benoni, B., Hnilicová, J., Kaberdin, V. R., Cahová, H., Vítězová, M., Převorovský, M., and Krásný, L. (2021) NAD+ capping of RNA in Archaea and Mycobacteria. *bioRxiv*. 10.1101/2021.12.14.472595
7. Barvík, I., Rejman, D., Panova, N., Šanderová, H., and Krásný, L. (2017) Non-canonical transcription initiation: The expanding universe of transcription initiating substrates. *FEMS Microbiol. Rev.* **41**, 131–138
8. Huang, F. (2003) Efficient incorporation of CoA, NAD and FAD into RNA by in vitro transcription. *Nucleic Acids Res.* 10.1093/nar/gng008
9. Cahová, H., Winz, M. L., Höfer, K., Nübel, G., and Jäschke, A. (2015) NAD captureSeq indicates NAD as a bacterial cap for a subset of regulatory RNAs. *Nature*. **519**, 374–377

10. Wang, Y., Li, S., Zhao, Y., You, C., Le, B., Gong, Z., Mo, B., Xia, Y., and Chen, X. (2019) NAD+-capped RNAs are widespread in the Arabidopsis transcriptome and can probably be translated. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12094–12102

11. Frindert, J., Zhang, Y., Nübel, G., Kahloon, M., Kolmar, L., Hotz-Wagenblatt, A., Burhenne, J., Haefeli, W. E., and Jäschke, A. (2018) Identification, Biosynthesis, and Decapping of NAD-Capped RNAs in B. subtilis. *Cell Rep.* **24**, 1890-1901.e8

12. Jiao, X., Doamekpor, S. K., Bird, J. G., Nickels, B. E., Tong, L., Hart, R. P., and Kiledjian, M. (2017) 5′ End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding. *Cell.* **168**, 1015-1027.e10

13. Walters, R. W., Matheny, T., Mizoue, L. S., Rao, B. S., Muhlrad, D., and Parker, R. (2017) Identification of NAD+ capped mRNAs in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 480–485

14. Bird, J. G., Zhang, Y., Tian, Y., Panova, N., Barvík, I., Greene, L., Liu, M., Buckley, B., Krásný, L., Lee, J. K., Kaplan, C. D., Ebright, R. H., and Nickels, B. E. (2016) The mechanism of RNA 5′ capping with NAD+, NADH and desphospho-CoA. *Nature*. **535**, 444–447

15. Zhang, H., Zhong, H., Zhang, S., Shao, X., Ni, M., Cai, Z., Chen, X., and Xia, Y. (2019) NAD tagSeq reveals that NAD+-capped RNAs are mostly produced from a large number of protein-coding genes in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12072–12077

16. Li, N., Yu, C., and Huang, F. (2005) Novel cyanine-AMP conjugates for efficient 5′ RNA fluorescent labeling by one-step transcription and replacement of [γ-

32P]ATP in RNA structural investigation. *Nucleic Acids Res.* **33**, 1–8

17. Huang, F., Wang, G., Coleman, T., and Li, N. (2003) Synthesis of adenosine derivatives as transcription initiators and preparation of 5′ fluorescein- and biotin-labeled RNA through one-step in vitro transcription. *RNA*. **9**, 1562–1570
18. Julius, C., and Yuzenkova, Y. (2019) Noncanonical RNA-capping: Discovery, mechanism, and physiological role debate. *Wiley Interdiscip. Rev. RNA*. **10**, e1512
19. Julius, C., Riaz-Bradley, A., and Yuzenkova, Y. (2018) RNA capping by mitochondrial and multi-subunit RNA polymerases. *Transcription*. **9**, 292–297
20. Julius, C., and Yuzenkova, Y. (2017) Bacterial RNA polymerase caps RNA with various cofactors and cell wall precursors. *Nucleic Acids Res.* **45**, 8282–8290
21. Lacatena, R. M., and Cesareni, G. (1981) Base pairing of RNA I with its complementary sequence in the primer precursor inhibits ColE1 replication. *Nature*. **294**, 623–626
22. Frick, D. N., and Bessman, M. J. (1995) Cloning, purification, and properties of a novel NADH pyrophosphatase. Evidence for a nucleotide pyrophosphatase catalytic domain in MutT-like enzymes. *J. Biol. Chem.* **270**, 1529–1534
23. Igamberdiev, A. U., and Gardeström, P. (2003) Regulation of NAD- and NADP-dependent isocitrate dehydrogenases by reduction levels of pyridine nucleotides in mitochondria and cytosol of pea leaves. *Biochim. Biophys. Acta - Bioenerg.* **1606**, 117–125
24. Bird, J. G., Zhang, Y., Tian, Y., Panova, N., Barvík, I., Greene, L., Liu, M., Buckley, B., Krásný, L., Lee, J. K., Kaplan, C. D., Ebright, R. H., and Nickels, B. E. (2016) The mechanism of RNA 5′ capping with NAD+, NADH and

desphospho-CoA. *Nature*. **535**, 444–447

25. Grudzien-Nogalska, E., Bird, J. G., Nickels, B. E., and Kiledjian, M. (2018) "NAD-capQ" detection and quantitation of NAD caps. *RNA*. **24**, 1418–1425

26. Zhang, H., Zhong, H., Wang, X., Zhang, S., Shao, X., Hu, H., Yu, Z., Cai, Z., Chen, X., and Xia, Y. (2021) Use of NAD tagSeq II to identify growth phase-dependent alterations in E. coli RNA NAD+ capping. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 2026183118

27. Wang, J., Alvin Chew, B. L., Lai, Y., Dong, H., Xu, L., Balamkundu, S., Cai, W. M., Cui, L., Liu, C. F., Fu, X. Y., Lin, Z., Shi, P. Y., Lu, T. K., Luo, D., Jaffrey, S. R., and Dedon, P. C. (2019) Quantifying the RNA cap epitranscriptome reveals novel caps in cellular and viral RNA. *Nucleic Acids Res.* 10.1093/NAR/GKZ751

28. Thauer, R. K., Jungermann, K., and Decker, K. (1977) Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol. Rev.* **41**, 100

29. Nazi, I., Koteva, K. P., and Wright, G. D. (2004) One-pot chemoenzymatic preparation of coenzyme A analogues. *Anal. Biochem.* **324**, 100–105

30. Sapkota, K., and Huang, F. (2018) Efficient one-pot enzymatic synthesis of dephospho coenzyme A. *Bioorg. Chem.* **76**, 23–27

31. Stewart, C. J., Thomas, J. O., Ball, W. J., and Aguirre, A. R. (1968) Coenzyme A Analogs. III. The Chemical Synthesis of Desulfopantetheine 4′-Phosphate and Its Enzymatic Conversion to Desulfo-coenzyme A. *J. Am. Chem. Soc.* **90**, 5000–5004

32. Ellington, A. D., and Szostak, J. W. (1990) *In vitro selection of RNA molecules that bind specific ligands*

33. Tuerk, C., and Gold, L. (1990) Systematic Evolution of Ligands by Exponential

Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science (80-.).* **249**, 505–510

34. Coleman, T. M., and Huang, F. (2002) RNA-catalyzed thioester synthesis. *Chem. Biol.* **9**, 1227–1236

35. Sabeti, P. C., Unrau, P. J., and Bartel, D. P. (1997) Accessing rare activities from random RNA sequences: the importance of the length of molecules in the starting pool. *Chem. Biol.* **4**, 767–774

36. Komarova, N., and Kuznetsov, A. (2019) Inside the black box: What makes Selex better? *Molecules*. 10.3390/molecules24193598

37. Thompson, I. A. P., Zheng, L., Eisenstein, M., and Soh, H. T. (2020) Rational design of aptamer switches with programmable pH response. *Nat. Commun. 2020 111*. **11**, 1–7

38. Lin, P. H., Chen, R. H., Lee, C. H., Chang, Y., Chen, C. S., and Chen, W. Y. (2011) Studies of the binding mechanism between aptamers and thrombin by circular dichroism, surface plasmon resonance and isothermal titration calorimetry. *Colloids Surf. B. Biointerfaces*. **88**, 552–558

39. McKeague, M., McConnell, E. M., Cruz-Toledo, J., Bernard, E. D., Pach, A., Mastronardi, E., Zhang, X., Beking, M., Francis, T., Giamberardino, A., Cabecinha, A., Ruscito, A., Aranda-Rodriguez, R., Dumontier, M., and DeRosa, M. C. (2015) Analysis of In Vitro Aptamer Selection Parameters. *J. Mol. Evol.* **81**, 150–161

40. Amano, R., Takada, K., Tanaka, Y., Nakamura, Y., Kawai, G., Kozu, T., and Sakamoto, T. (2016) Kinetic and Thermodynamic Analyses of Interaction between

a High-Affinity RNA Aptamer and Its Target Protein. *Biochemistry*. **55**, 6221–6229

41. Inoue, A., Takagi, Y., and Taira, K. (2004) Importance in catalysis of a magnesium ion with very low affinity for a hammerhead ribozyme. *Nucleic Acids Res.* **32**, 4217–4223

42. Zivarts, M., Liu, Y., and Breaker, R. R. (2005) Engineered allosteric ribozymes that respond to specific divalent metal ions. *Nucleic Acids Res.* **33**, 622

43. Tyrrell, J., Weeks, K. M., and Pielak, G. J. (2015) Challenge of Mimicking the Influences of the Cellular Environment on RNA Structure by PEG-Induced Macromolecular Crowding. *Biochemistry*. **54**, 6447–6453

44. Davis, J. H., and Szostak, J. W. (2002) Isolation of high-affinity GTP aptamers from partially structured RNA libraries. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11616–11621

45. Chushak, Y., and Stone, M. O. (2009) In silico selection of RNA aptamers. *Nucleic Acids Res.* **37**, e87

46. Kim, N., Hin, H. G., and Schlick, T. (2007) A computational proposal for designing structured RNA pools for in vitro selection of RNAs. *RNA.* **13**, 478–492

47. Carothers, J. M., Oestreich, S. C., and Szostak, J. W. (2006) Aptamers selected for higher-affinity binding are not more specific for the target ligand. *J. Am. Chem. Soc.* **128**, 7929–7937

48. Chizzolini, F., Luiz, #, Passalacqua, F. M., Oumais, M., Dingilian, A. I., Szostak, J. W., and Lupták, A. L. (2020) Large Phenotypic Enhancement of Structured Random RNA Pools. *Cite This J. Am. Chem. Soc*. 10.1021/jacs.9b11396

49. Carothers, J. M., Oestreich, S. C., Davis, J. H., and Szostak, J. W. (2004) Informational Complexity and Functional Activity of RNA Structures. *J. Am. Chem. Soc.* **126**, 5130–5137

50. Ishikawa, J., Matsumura, S., Jaeger, L., Inoue, T., Furuta, H., and Ikawa, Y. (2009) Rational optimization of the DSL ligase ribozyme with GNRA/receptor interacting modules. *Arch. Biochem. Biophys.* **490**, 163

51. Seelig, B., and Szostak, J. W. (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature*. **448**, 828–831

52. Ekland, E. H., Szostak, J. W., and Bartel, D. P. (1995) Structurally Complex and Highly Active RNA Ligases Derived from Random RNA Sequences. *Science (80-.).* **269**, 364–370

53. Pobanz, K., and Lupták, A. (2016) Improving the odds: Influence of starting pools on in vitro selection outcomes. *Methods*. **106**, 14–20

54. Porter, E. B., Polaski, J. T., Morck, M. M., and Batey, R. T. (2017) Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors. *Nat. Chem. Biol.* **13**, 295–301

55. Luo, X., McKeague, M., Pitre, S., Dumontier, M., Green, J., Golshani, A., Derosa, M. C., and Dehne, F. (2010) Computational approaches toward the design of pools for the in vitro selection of complex aptamers. *RNA.* **16**, 2252–2262

56. Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A., and Weeks, K. M. (2015) Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc. 2015 1011*. **10**, 1643–1669

57. Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J. W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T., and Chang, H. Y. (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*. **519**, 486–490

58. Mayer, G., Müller, J., and Lünse, C. E. (2011) RNA diagnostics: real-time RT-PCR strategies and promising novel target RNAs. *Wiley Interdiscip. Rev. RNA*. **2**, 32–41

59. Thiel, W. H., Bair, T., Wyatt Thiel, K., Dassie, J. P., Rockey, W. M., Howell, C. A., Liu, X. Y., Dupuy, A. J., Huang, L., Owczarzy, R., Behlke, M. A., McNamara, J. O., and Giangrande, P. H. (2011) Nucleotide bias observed with a short SELEX RNA aptamer library. *Nucleic Acid Ther.* **21**, 253–263

CHAPTER 2: Minimizing amplification bias during reverse transcription for *in vitro* selections

This work has been submitted and is under review for publication at RNA. *Authors include Jordyn K. Lucas, Paige R. Gruenke, and Donald H. Burke.*

ABSTRACT

Systematic Evolution of Ligands through EXponential enrichment (SELEX) is widely used to identify functional nucleic acids, such as aptamers and ribozymes. Ideally, selective pressure drives enrichment of sequences that display the function of interest (binding, catalysis, etc). However, amplification biases from reverse transcription can overwhelm this enrichment and leave some functional sequences at a disadvantage, with cumulative effects across multiple rounds of selection. Libraries that are designed to include structural scaffolds can improve selection outcomes by sampling sequence space more strategically, but they are also susceptible to such amplification biases, particularly during reverse transcription. Therefore, we tested five reverse transcriptases (RTs) – ImProm-II, Marathon RT (MaRT), TGIRT-III, SuperScript IV (SSIV), and BST 3.0 DNA polymerase (BST) – to determine which enzymes introduced the least bias. We directly compared cDNA yield and processivity for these enzymes on RNA templates with varying degrees of structure under various reaction conditions. In these analyses, BST exhibited excellent processivity, generated large quantities of full length cDNA product, displayed little bias among templates with varying structure and sequence, and performed well on long, highly structured viral RNAs. Additionally, six RNA libraries containing either strong, moderate,

or no incorporated structural elements were pooled and competed head-to-head in six rounds of an amplification-only selection without external selective pressure using either SSIV, ImProm-II, or BST during reverse transcription. High-throughput sequencing established that BST maintained the most neutral enrichment values, indicating low inter-library bias over the course of six rounds, relative to SSIV and ImProm-II, and it introduced minimal mutational bias.

INTRODUCTION

RNA is a highly versatile molecule that can be developed into tools for gene regulation, biosensors, targeted drug delivery, high-specificity binding, catalysis, and many other purposes for molecular medicine, synthetic biology, and other biotechnologies. As the need for new functional RNAs has expanded, so too have the methods and strategies for discovering these molecules. Systematic Evolution of Ligands through EXponential enrichment (SELEX) is an *in vitro* evolution technique used to identify new aptamers, ribozymes, and other functional nucleic acid modules (1, 2). Although the underlying concepts behind SELEX are straightforward, preparation and successful completion of selections can prove challenging. Library designs at the onset of a selection can shape selection outcomes, and a successful selection is often directly related to the structural and sequence diversity of the starting library and robust amplification to uphold it (3, 4).

Functional RNAs such as aptamers, riboswitches, and ribozymes are typically more structured than random RNA sequences. Experimental and computational studies of nucleic acid libraries support the correlation between increased secondary structure and

increased function (5–10). As a result, functional RNAs have high information content, deriving from information required to specify both generic structural elements and uniquely specified nucleotides that are often found in regions responsible for activity, such as binding sites or catalytic active sites. When starting from a fully randomized library, the probability of these key nucleotides being independently specified within an appropriate structural context is low. In contrast, when a starting library includes basic structural elements from the beginning, sequence space exploration takes place within a context that already contains those structural elements, potentially increasing the fraction of the library that can attain the desired function and increasing the probability of a successful selection.

Engineering starting pools to contain structural elements is not a new concept and the success of these engineered structured starting pools has already been seen (5, 9–16). For example, when a partially structured library containing a stable stem-loop was directly compared with a fully randomized RNA library during a selection for aptamers to GTP, more aptamers emerged from the partially structured library, and those that retained the incorporated stem-loop structure displayed some of the highest affinities (5). Similarly, a selection performed against neurotransmitter precursors using starting libraries that incorporated three-way junction (3WJ) scaffolds derived from natural and synthetic aptamers revealed that aptamers that preserved the 3WJ scaffold showed higher specificity and affinity for their ligands than those that lost this feature (15). In another study, structures derived from Group I intron P4-P6 were incorporated into a starting library to select for a *trans*-acting RNA ligase (17). The Szostak and Lupták groups used native PAGE fractionation to enrich for structured, functional RNAs, and this pre-enrichment

increased the phenotypic potential and allowed the randomized regions of their libraries to explore more complex structures, which can potentially enable higher level activity (9). Their fractionation method for purifying structured RNAs increased the phenotypic potential and allowed the randomized regions of their libraries to explore more complex structures which can potentially enable higher level activity. These and other examples demonstrate that engineering starting pools with structural elements provides an opportunity to enrich the selection pool for active RNAs.

However, highly structured RNA is susceptible to amplification biases, particularly during reverse transcription, which can sabotage a selection to the point of failure by discriminating against those RNAs – even those with superior performance in the functional step of a given selection. Reverse transcription bias is tied closely to two polymerase performance characteristics of the enzyme: its processivity and fidelity. Low RT processivity is especially detrimental to highly structured templates and can lead to incomplete readthrough or low yield of full length product during cDNA synthesis that precludes subsequent PCR (18)(19). Increasing the temperature of the reverse transcription reaction can reduce bias against highly structured templates by partially denaturing structured regions; however, many of the commonly used RTs do not function well at elevated temperatures, which in turn can negatively impact the cDNA product yield (20). Low fidelity by some RTs leads to nucleotide misincorporations, which can lead to an accumulation of mutations that favor amplification rather than the intended function (21). Mutations that disrupt the designed structural regions will result in more efficient cDNA synthesis by RTs that struggle with structured templates and allow these less structured

mutants to dominate the selected pool (Figure 2.1A). This phenomenon was observed during initial selections using the 3WJ scaffolded libraries above when cDNA synthesis was carried out using SuperScript III (SSIII). Amplification bias was resolved by switching to an RT that favored readthrough of structured RNA, which greatly improved retention of structural design features and improved selection outcomes(15). However, even low levels of amplification bias can cripple the best library designs, hence for pre-determined structure to play a significant favorable role in selection outcomes, it is critical that amplification bias be very low. Although RT processivity and yield on structured RNA templates are critical determinants of selection outcomes, a systematic comparison of RT suitability has not been performed, to our knowledge.

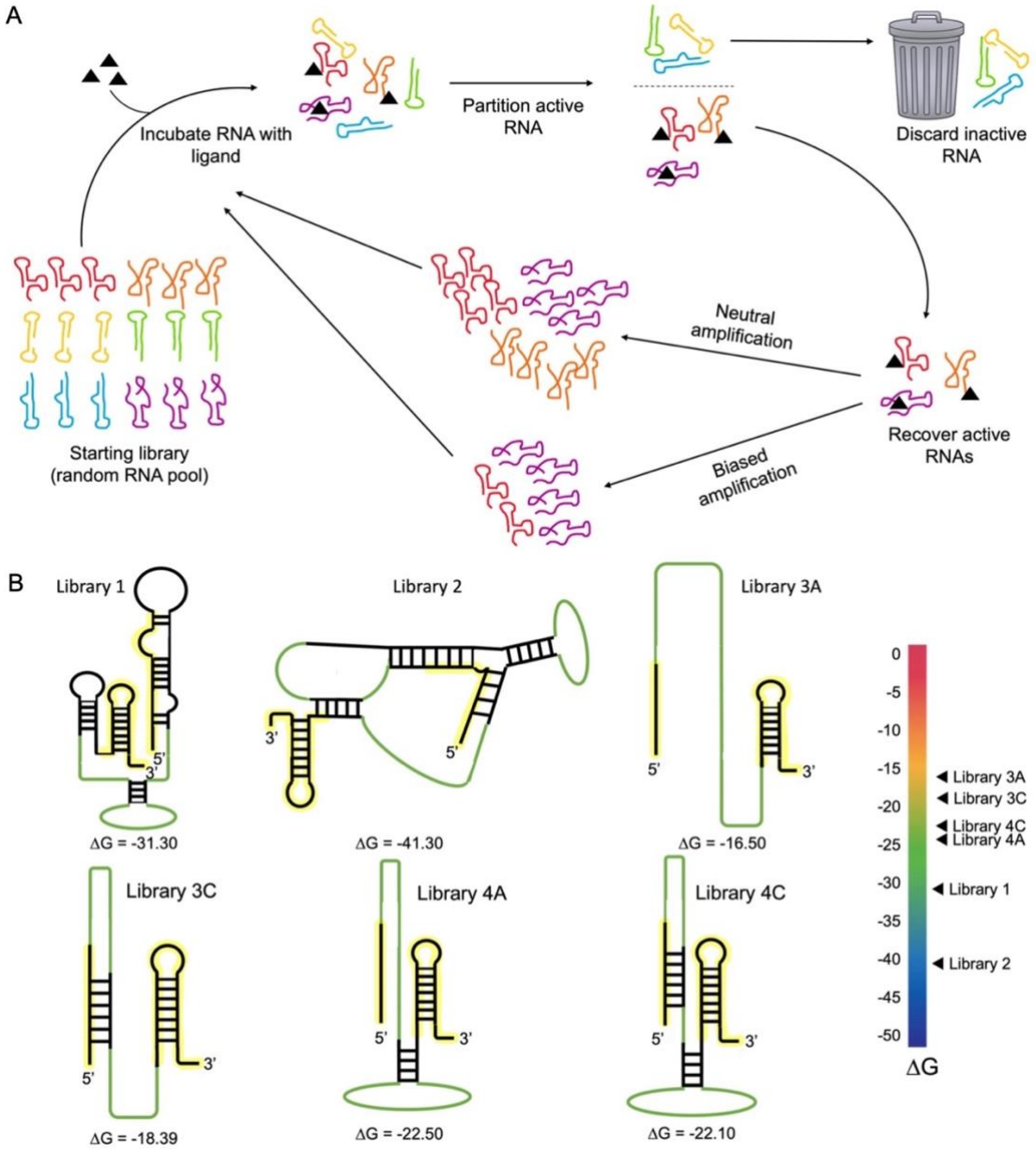


Figure 2.1. RNA sequences with different structures can experience varying amounts of amplification bias during reverse transcription and PCR. (A) SELEX schematic illustrating the impact of neutral and biased amplification on library pools during selection rounds. (B) Designed secondary structural architectures of the six RNA libraries studied here, where black represents defined sequence and green represents random regions.

Primer binding sites are highlighted in yellow. Calculated ΔG values for designed structural elements and overall secondary structures were predicted using Mfold (22). On the right, libraries are arranged according to their ΔG values from most (bottom) to least (top) stable.

Here we directly compared the activities of five RTs under multiple reaction conditions with various RNA templates to identify the best RT and its optimized conditions for future selections. We chose ImProm-II, SuperScript IV (SSIV), TGIRT-III, Marathon RT (MaRT), and BST 3.0 DNA Polymerase (BST) for these comparisons. ImProm-II originates from Avian Myeloblastosis Virus (AMV) RT and has been used frequently in successful aptamer and ribozyme selections by our group (23–25) and others (26–28). SSIV (and its predecessor SSIII) originated from Moloney murine leukemia virus (M-MLV) and is also commonly used in aptamer and ribozyme selections (29–34). TGIRT-III, originating from a mobile type II intron, is reported to have higher processivity, fidelity, and thermostability than some retroviral RTs (35). MaRT, which also originated from a mobile type II intron, has displayed excellent processivity on highly structured templates, some of which are multiple kilobases in length (36). BST 3.0 DNA Polymerase is derived from DNA Polymerase I from the bacterium *Geobacillus stearothermophilus* (37). BST 3.0 contains 5' to 3' DNA or RNA dependent DNA polymerase activity and has strong strand displacement activity (38, 39). Relative to its wildtype predecessor, BST 3.0 boasts improvements in RT activity, amplification speed, inhibitor tolerance, and thermostability. It is also notable for its helicase-like activity and its use in loop-mediated isothermal amplification (LAMP)(40). Also, BST 3.0 has demonstrated activity at higher temperatures in addition to excellent processivity (38). Here we show that of the tested enzymes, BST 3.0 has the highest processivity, lowest bias between structured and non-structured RNA templates, and highest yield for full length product. High-throughput sequencing data demonstrate that BST 3.0 has equivalent fidelity to other enzymes and also introduced less bias than ImProm-II RT or SSIV RT after multiple rounds of an 'amplification-only'

selection, making this enzyme especially attractive for exploration as a standard RT for *in vitro* selections.

RESULTS

Assessing reverse transcriptase bias, processivity, and yield for library templates

To probe the potential impact of RT bias on library evolution during selections, we utilized RNA libraries that all contained the same 5' and 3' constant regions and overall length but with varying degrees of built-in structure (Figure 2.1B). The 3' end primer binding site for the libraries incorporated a stem-loop derived from a SHAPE cassette to minimize differences in annealing of 3' primers and RT initiation (41). Initial analysis focused on libraries 1 and 3A which represent the most and least structured RNAs, respectively. As such, we expected any bias associated with readthrough of structured RNA should be especially pronounced in comparing results for these two libraries. Library 1 is a highly structured RNA derived from an aptamer domain of the *B. subtilis xpt-pbuX* guanine riboswitch. Our library design is similar to a library used previously to select aptamers with high affinity for neurotransmitters (15), based on the hypothesis that incorporating secondary and tertiary structural scaffolds provides a more advantageous starting point for selections and results in more highly functional species. In contrast, library 3A is a highly randomized RNA with little to no pre-existing structural elements and would represent a 'Null Hypothesis' for selections that utilized Library 1.

A panel of RTs were evaluated under various conditions to identify enzymes and optimized conditions that generate the greatest amount of full length cDNA product. To this end,

cDNA products for each RNA template were visualized and quantified with respect to net yield and processivity. Reactions performed with ImProm-II (Figure 2.2) revealed very low yield of full length product and low average processivity even under a variety of reaction conditions (Figure 2.3A). SSIV RT reactions produced similar levels of full length product with library 1 and library 3A (Figure 2.3B) resulting in similar yield and processivity values between the two templates indicating less template related bias as compared to ImProm-II reactions. Although increasing the SSIV reaction temperature from 55°C to 65°C reduced template bias, the higher temperatures also negatively impacted the yield and processivity, consistent with previous reports of activity loss at higher temperatures (34). Even under optimized conditions, SSIV reactions produced no significant improvement in yield or processivity compared to reactions using ImProm-II. In our hands, both TGIRT-III (Figure 2.3C) and MaRT (Figure 2.3D) exhibited lower processivity and yield than anticipated based on previous reports, in which both of these group II intron derived RTs outperform many retroviral RTs including SSIV (35, 36, 42, 43). However, those experiments primarily used long RNA templates ranging upwards of 6 kb, and the longer templates present significantly different performance demands than the 136 nt templates used here, potentially accounting for some of the disparity between relative yields and processivities observed here and those reported previously. Amplification of long RNAs is crucial for many types of research including transcription of mRNAs for gene therapy and vaccines, RNA structure mapping, and gaining further understanding of long RNAs. Thus while MaRT, TGIRT-III, and/or other intron II derived RTs may be good choices for experiments in which extremely long templates must be

reverse transcribed, they do not appear to be the best choice for aptamer or ribozyme selections, wherein templates are typically 70 to 200 nucleotides.

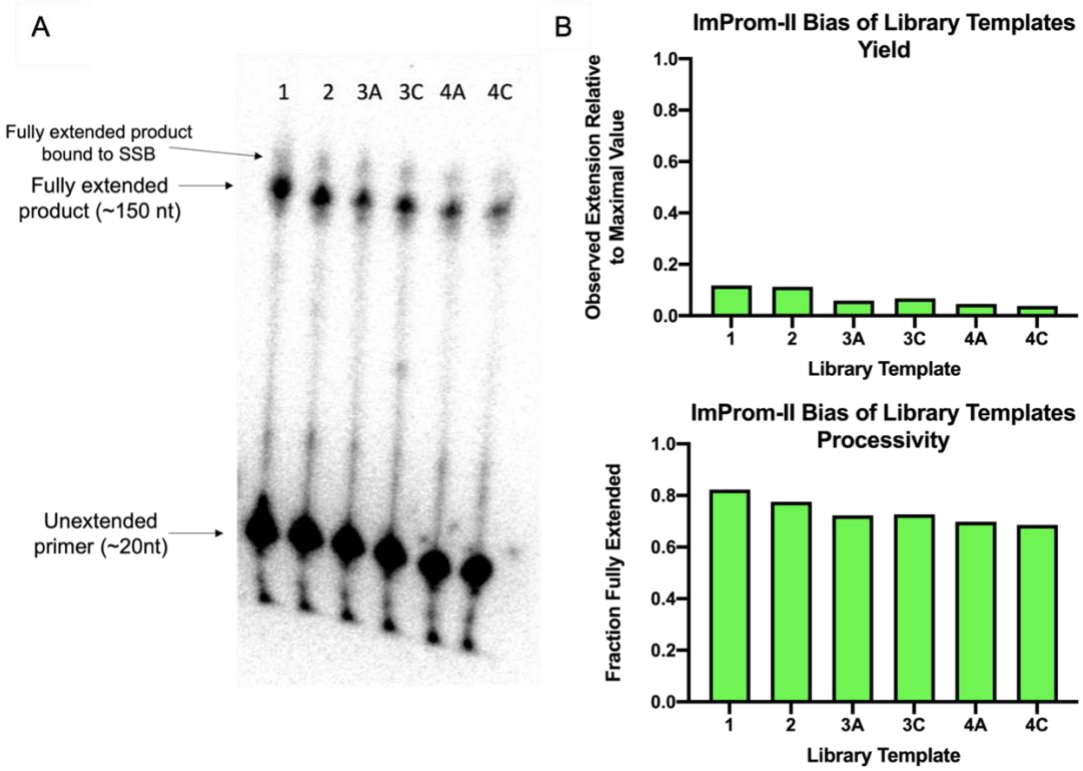


Figure 2.2. Comparison of reverse transcriptase bias between six RNA library templates with various amounts of structure. The primer extension assay was performed using ImProm-II RT. After a 5 min incubation at 25°C for primer-template annealing, the reaction was performed for 1 hr at 42°C using 15 pmol of RNA library template and 30 pmol of γ^{32}P labeled reverse primer. Libraries 1, 2, 3A, 3C, 4A, 4C (left to right) were used for these reactions. N=1 (A) To directly compare the reverse transcriptase bias between the six RNA library templates, the reactions were run on the same PAGE gel and visualized using a phosphorimager. (B) The yield and processivity for each reaction were quantified and plotted. Yield is determined by the ratio of fully extended primer to the unextended primer. Processivity is the ratio of full extended product to all (partially and fully) extended primer.

In contrast to the ‘professional’ RTs, BST dramatically outperformed the other four enzymes in full length product formation and had equivalent or better processivity (Figure 2.3E). To illustrate this, the data from Figures 2.3A-E were plotted in a multi-variable graph where each point represents the quantified yield and processivity for a given reaction condition, template, and enzyme. The orange points that represent BST data are in the top right corner of the graph, further illustrating that the yield (*y*-axis) and processivity (*x*-axis) were both significantly higher for BST in comparison to the other four enzymes. From these data, reaction conditions (reaction time and temperature) were identified for each enzyme that minimized bias between the two templates (Libraries 1 and 3a) and resulted in the highest yield and processivity. These optimized reaction conditions were then used for each enzyme in subsequent experiments.

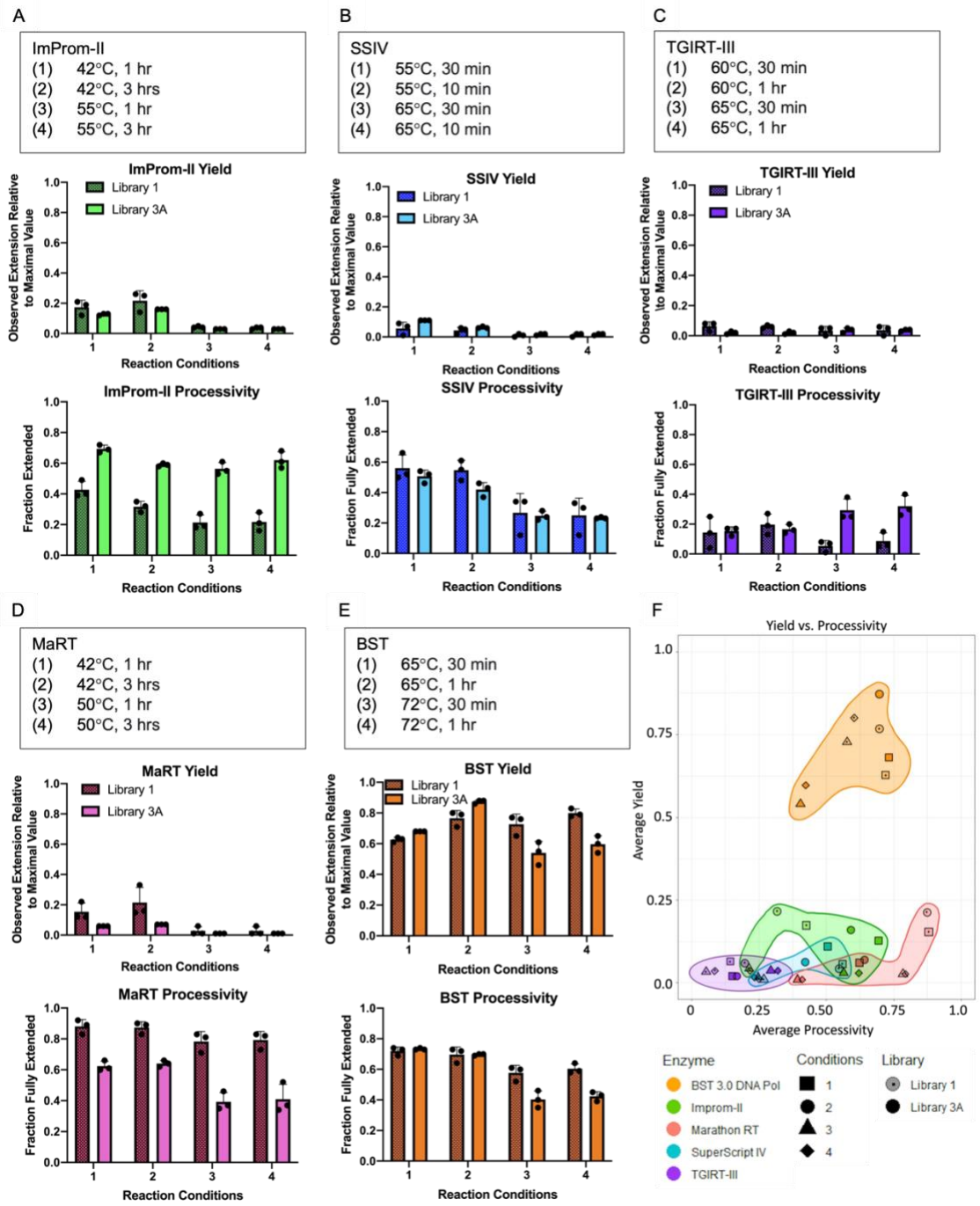


Figure 2.3. Comparison of RT products (yields and processivities) from ImProm-II RT, SSIV RT, TGIRT-III, MaRT, and BST 3.0 DNA Polymerase by primer extension assays with γ32P labeled primer. Yield is determined by the ratio of fully extended primer to the unextended primer. Shading of the bars indicates RNA template: library 1 (darker,

left) or library 3A (lighter, right). Net processivity is calculated as the ratio of fully extended product to the sum of all partially or fully extended primer. Data are shown for N=3 independent experiments for panels A-E. Yields and net processivities under various reaction temperatures and times for (A) ImProm-II RT (B) SuperScript IV (C) TGIRT-III (D) MaRT and (E) BST 3.0 DNA Polymerase. (F) Multi-variable plot comparing the RTs using average net processivity (x-axis) and yield (y-axis) values from panels A-E. Colors indicate which RT enzyme was used, shapes indicate reaction conditions from 2A-E, and shadings indicate which library templates were used. The colored clouds around each of the five RT data sets are artistic renderings to aid visualization of groupings for each enzyme.

RT additives can reduce template-based bias and slightly enhance activities for some RTs

Various additives have been suggested to improve reverse transcription yields. Betaine (trimethylglycine) decreases the melting temperatures of DNA and RNA duplexes while simultaneously stabilizing proteins to prevent their thermal denaturation, making it an excellent additive for PCR, reverse transcription, and sequencing reactions at elevated temperature (44). Like betaine, trehalose also serves a protective function towards the enzymes by reducing movement within the protein backbone to reduce thermal unfolding (50). Single stranded binding protein (SSB) has been previously used as a reverse transcription additive to increase the sizes of the cDNA products generated, thereby favoring completion of more full length product (46).

Primer extension assays for ImProm-II, SSIV, and BST were performed in the presence and absence of betaine, trehalose, and/or SSB. ImProm-II showed slight improvement in processivity but there was no notable improvement for full length product formation (Figure 2.4A). SSIV appeared to have reduced bias between structured (library 1) and unstructured (library 3A) templates in the presence of betaine, trehalose, and SSB as compared to reactions under the same conditions without any additives present (Figure 2.4B), but again with no notable improvement in yield in the presence of additives. BST continued to outperform the other enzymes in yield and processivity and also demonstrated a slight increase in yield and processivity in the presence of betaine, trehalose, and SSB. Furthermore, BST also showed minimal bias between the structured and unstructured RNA templates (Figure 2.4C).

Additionally, we questioned whether enzyme concentration was responsible for reduced cDNA yield especially for ImProm-II and SSIV. Therefore, we performed primer extension assays using the optimal conditions identified in Figure 2.4A-C with either normal (1X) amounts of enzyme or 3X enzyme and directly compared yield and processivity (Figure 2.4D). Interestingly, addition of three times the suggested enzyme concentration did not improve cDNA yield or processivity for any of the three enzymes. These data suggests that the differences observed in yield and processivity are not a result of some RT reactions having less enzyme than others, but rather a true reflection of each enzyme's performance.

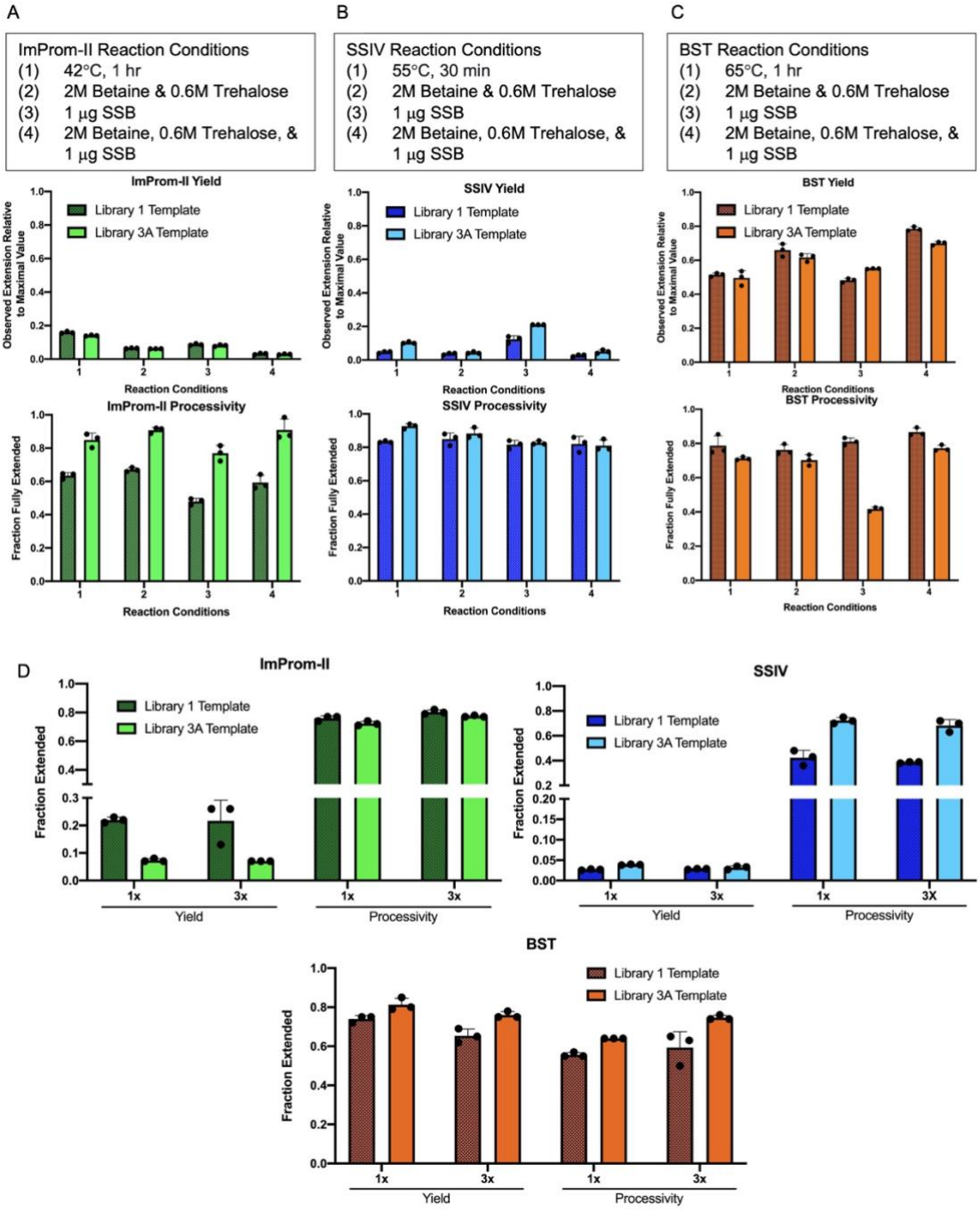


Figure 2.4. Comparison of RT products (yield and processivity) from ImProm-II RT, SuperScript IV RT, and BST 3.0 DNA Polymerase by primer extension assay with γ32P labeled primer using the best reaction conditions (temperature and time) from

Figure 2.3. Shading of the bars indicates which template, library 1 (darker, left) or library 3A (lighter, right), was used. Yields and net processivities were calculated as in Fig 2 and were measured in the presence of 2M betaine, 0.6M trehalose, and/or 1 μg of Single Stranded Binding Protein (SSB) for (A) ImProm-II RT (B) SuperScript IV and (C) BST 3.0 DNA Polymerase. (D) Impact of enzyme quantity on RT products (yield and processivity). Primer extension assays were performed as in A-C using either 1x enzyme (as used in A-C) or 3x enzyme amounts. The optimal conditions identified in A-C were used for these primer extension assays: ImProm-II reactions were performed under condition 1, SSIV reactions were performed under condition 3, and BST reactions were performed under condition 4. N=3 for A-D.

Comparison of RT bias using the six library designs as templates

The results above establish that BST efficiently reverse transcribes both highly structured (Library 1) and relatively unstructured (Library 3A) RNA templates, and that it does so more effectively than the other RTs. To evaluate whether BST could be similarly applied to other RNA templates, four additional RNA libraries were tested, each with unique and variable structural elements (Figure 2.1B). Reverse transcription reactions with all six templates were compared using BST under the optimized conditions and in the presence of additives. Notably, the observed yield and processivity for the six RNA library templates was remarkably similar (Figure 2.5A and 2.5B), with the greatest differences arising between library 3A (62% yield, 66% processivity) and library 4C (78% yield, 82% processivity). Even with these differences, the ratio of highest to lowest remains less than 1.26 for both yield and processivity, which translates to a very modest cumulative effect of <4-fold even after six selection cycles. This was a stark contrast to results with ImProm-II (Figure 2.2), where all six library templates had observed yields less than 20%. Figure 2.5C further illustrates this point by plotting the average processivity and yield values and observing the spread of data points (or lack thereof in this case). The points cluster together as a result of their similar values. This clustering indicates that BST is expected to introduce relatively little amplification bias during a selection as compared to the other RTs.

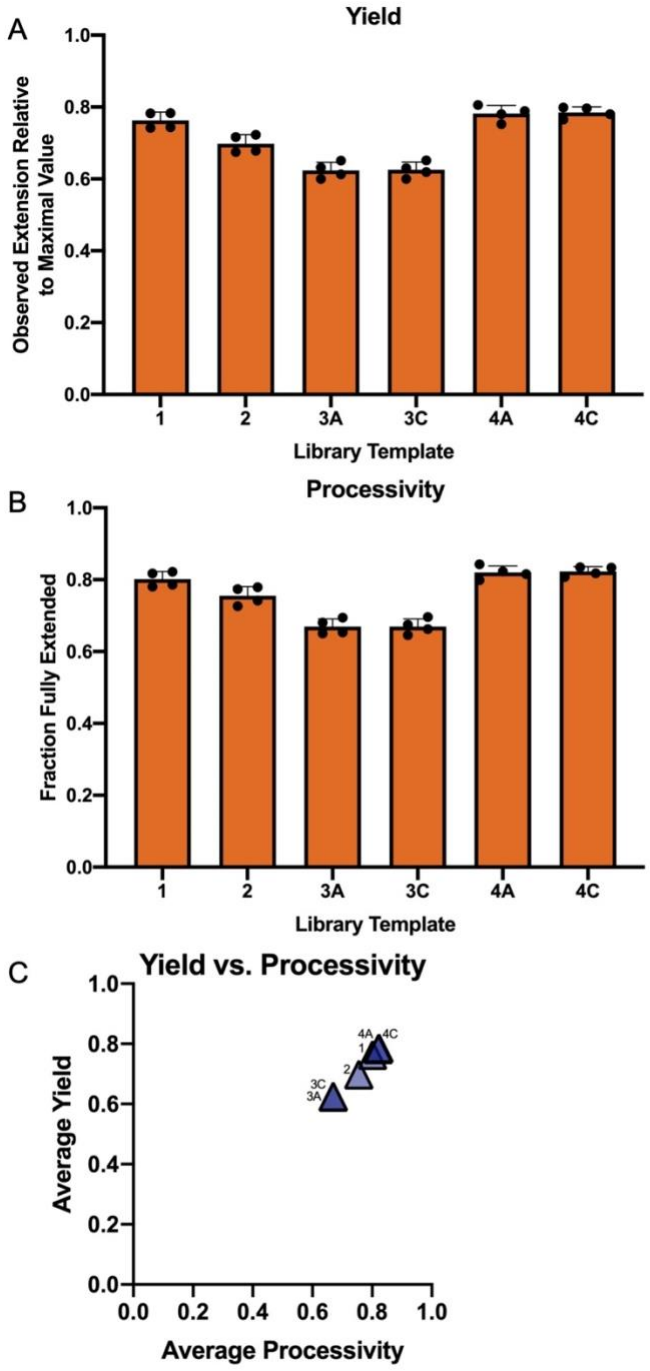


Figure 2.5. Comparison of reverse transcriptase bias for BST 3.0 DNA polymerase across six RNA library templates with various types and amounts of structure, as detailed in Figure 2.1B. Primer extension assays were performed using optimal conditions identified in figure 2.4C (65°C for 1 hour in the presence of 2M Betaine & 0.6M Trehalose,

and 1 μg SSB). The same batch of γ^{32}P labeled primer was used for all six templates. N=3 for A-C (A) Comparison of the yield of RT products from the six different templates. (B) Comparing processivity of BST 3.0 for the six different library templates. (C) Mean yield from panel A is plotted against mean processivity from panel B. Tight clustering of all pairs of values shows overall performance and low inter-library bias under optimized conditions.

BST 3.0 DNA polymerase can be used to reverse transcribe long, structured viral RNAs

BST's activity at higher temperatures, ability to generate large quantities of full length product, low bias among RNA templates with varying structural elements, and high processivity make it an excellent candidate to be used in the reverse transcription step of selections for functional RNA. However, it was unclear based on these data whether BST would perform similarly in reactions containing long, structured RNAs. Therefore, a primer extension assay was done to probe the enzyme's activity (under optimized conditions) with the ~350 nucleotide RNA from the HIV-1 5' long terminal repeat (LTR) as the template. This is a well-studied and highly structured RNA. As seen in Figure 2.6A, hardly any partial-length products can be observed between the top band of fully extended product and bottom band of remaining, unextended primer. BST demonstrated an ability to maintain high yield and processivity (Figure 2.6B) despite the template size doubling and containing many complex structured regions.

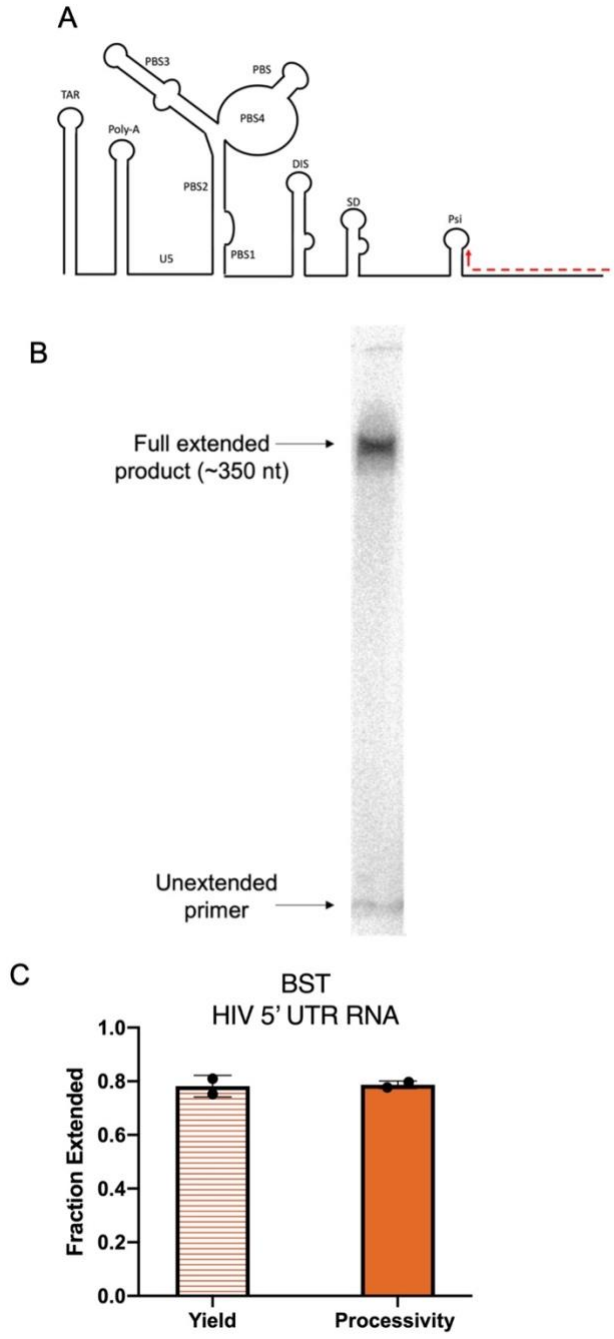


Figure 2.6. Testing BST 3.0 DNA Polymerase on a structured viral RNA. Yield and net processivity were calculated as in Fig 2. (A) Structure of the RNA template (HIV 5' UTR). Reverse primer binding site indicated in red. The primer extension assay was performed using BST 3.0 DNA polymerase at 65°C for 1 hour in the presence of 2M

Betaine & 0.6M Trehalose, and 1 μg SSB (optimal conditions identified in figure 2.4C). The γ^{32}P labeled primer were used for HIV 5' LTR RNA template. (B) Phosphorimage of primer extension assay run on a 5% denaturing PAGE. (C) Quantification of phosphorimage to calculate the yield and processivity of BST 3.0 DNA polymerase for a viral template. n=2

BST 3.0 DNA polymerase outperforms SSIV and ImProm-II in six rounds of an 'amplification-only' selection

To evaluate enzyme performance within the context of a selection, we utilized six libraries (Figure 2.1B) with the same 5' and 3' constant regions and overall length but with varying degrees of structure. Each of the six libraries amplified similarly during PCR (Figure 2.7) with minimal differences observed in band intensity after 30 rounds of PCR. Because the PCR and transcription steps were performed under identical conditions for the three trajectories, we anticipate that any library-specific biases observed from one trajectory to another can be attributed to the RT step. Libraries 1 and 2 are highly structured RNAs, analogous to those used in selections in which starting library pools are engineered to have structured regions. We expected these libraries to become progressively depleted over the course of the selection for the SSIV and ImProm-II trajectories. Libraries 3A and 3C (Figure 2.1B) are highly randomized RNAs with little to no pre-existing structural elements and represent a 'Null Hypothesis' with regards to structure. We expected to see enrichment of these libraries if templates with less structure were favored by RTs. Libraries 4A and 4C (Figure 2.1B) are intermediate in terms of pre-determined structural elements, carrying hairpin loop structures near the 3' ends and are therefore expected to fall between the more structured and fully randomized templates.

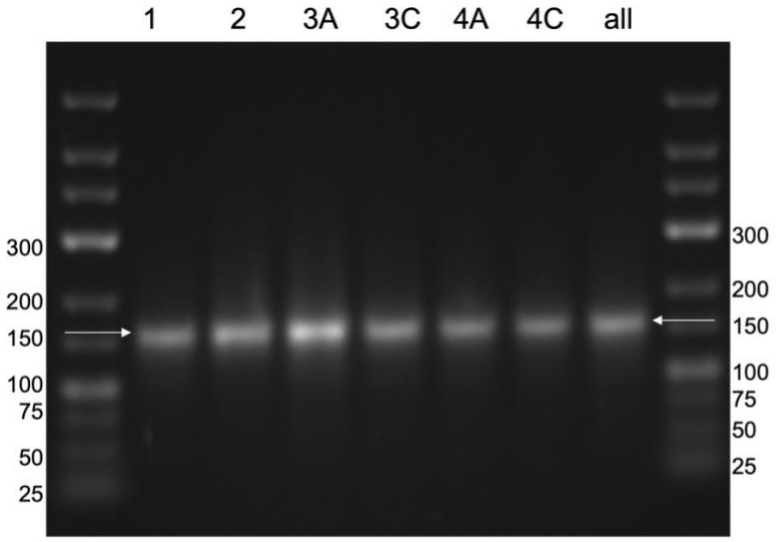


Figure 2.7. Minimizing PCR bias between 6 different libraries. PCR for the 6 different selection libraries was optimized to minimize undesired products, primer dimers, and discrepancy in PCR amplification. 5 cycles of PCR were performed, and products were run on an ethidium bromide stained 2% agarose gel and visualized with UV-Vis. The observed bands match the predicted size of 156 base pairs and no undesired products were observed. Although there were some small differences in band intensity (eg. library 3A), no large discrepancy in PCR product or amplification was observed between the six libraries.

These six libraries were mixed to form a single, pooled starting library and subjected to six rounds of an 'amplification-only' selection that excluded a partition step for biochemical function to evaluate the cumulative impact of amplification biases from reverse transcription (Figure 2.8A). During the selection, the starting pool was reverse transcribed by three different RTs – ImProm-II, SSIV, and BST – then amplified by PCR and transcribed back into RNA. This process was repeated for six rounds before sending each round from the ImProm-II, SSIV, and BST trajectories for high-throughput sequencing (HTS). The HTS data was used to monitor any systematic drift away from the original distribution of fractional representation for each library. Only the RT step differed among the three trajectories, so we anticipate that any library-specific biases observed from one trajectory to another can be attributed to the RT step.

Although there were some differences in read quantity between selection rounds (Tables 2.1, 2.3, and 2.5), clear trends emerged from this dataset. ImProm-II experienced the most drastic round-to-round variations when looking at fractional representations of each library (Figure 2.8B and 2.8C). For instance, library 1 went from being the most represented library in rounds 1-4, to one of the most depleted by round 6. Additionally, library 4C became increasingly favored during the ImProm-II selection and ultimately made up the largest fraction of total sequence by round 6. These ImProm-II HTS data were consistent with observations above in which ImProm-II introduced large inter-library bias among the six library templates, favoring unstructured libraries (Figure 2.2). SSIV performed well with minimally structured templates such as library 4C and completely unstructured templates such as library 3A. These less structured templates consistently made up the

largest fraction of total reads throughout the SSIV selection (Figure 2.8B), while the most structured templates (libraries 1 and 2) made up lower and lower fractions of the total reads throughout the selection (Figure 2.8C). This outcome agrees with the observations above indicating SSIV's preference for less structured templates. In contrast to the other two trajectories, the BST selection showed less drift in the fractional representation of each of the six libraries during the selection, particularly when comparing enrichment values for each input library in round 6 relative to round 1 (Figure 2.8B, 2.8C, and Tables 2.2, 2.4, and 2.6). Notably for the BST selection, the highly structured library 1 made up 25-40% of the fraction of total sequences in any given round without any systematic depletion across the six rounds, and the minimally structured library 4C made up between 26-38% of the fraction of total sequences without any systematic enrichment. Although library 2 is also a structured library, it was depleted throughout the BST selection. However, library 2's depletion was consistent across all three selections and was less dramatic in BST trajectories (Figure 2.8D) (see discussion).

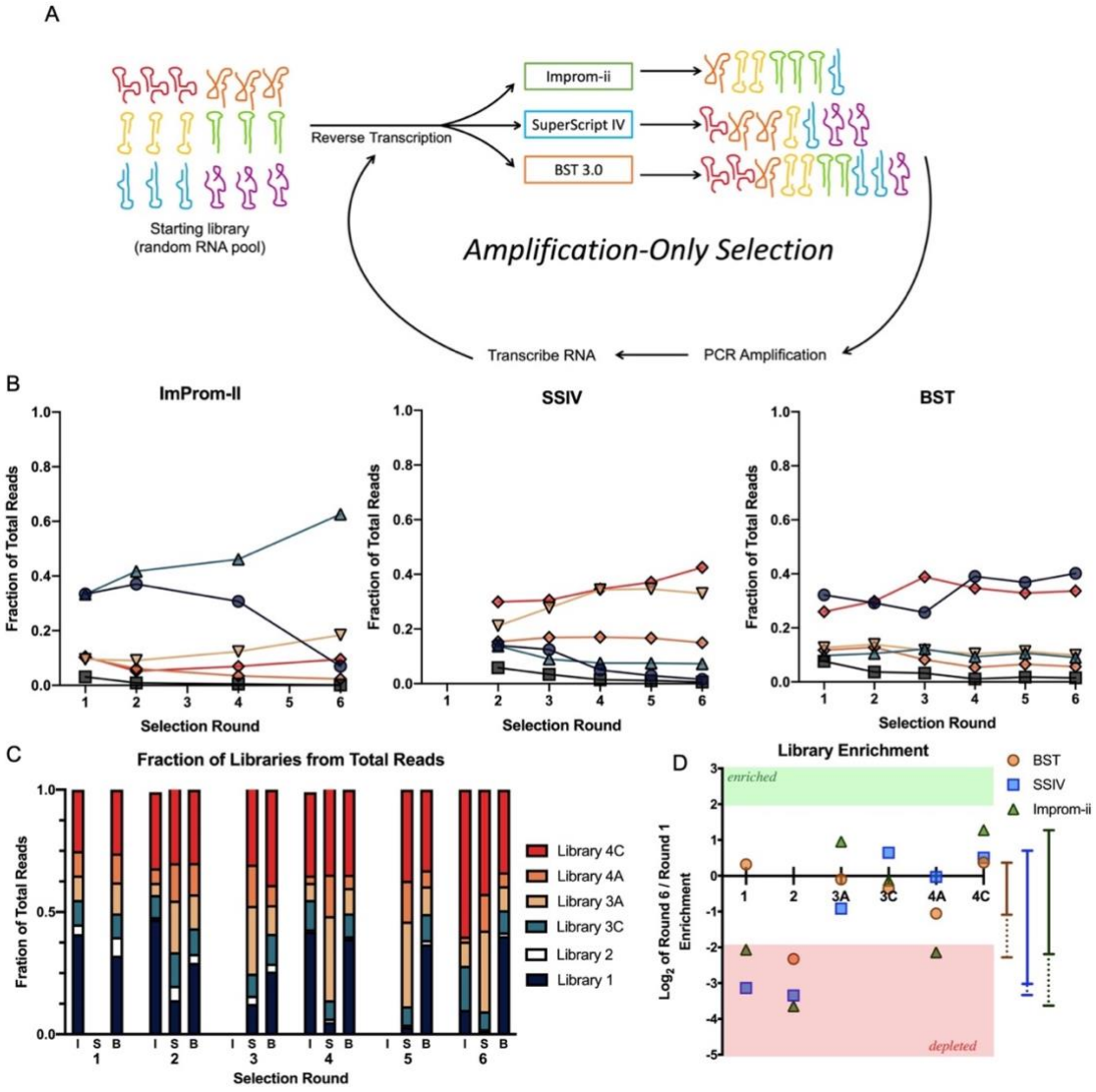


Figure 2.8. Impact of different reverse transcriptase on selection outcomes. (A) General schematic of the amplification-only selection. Three selection trajectories were performed either using ImProm-II, SSIV, or BST. All other conditions and handling steps were nearly identical between each selection trajectory. (B) Each of the six libraries was identified by sequence markers and then quantified to determine the fraction of total reads for each library. The changes in library fractions are plotted over the course of 6 rounds of selection. Any rounds with fewer than 1,000 total unique processed reads were excluded

from these data sets. (C) Bar graph indicating what fraction of total reads each library consisted of in each selection round. Each group of three shows data for ImProm-II (I), SSIV (S), and BST (B). (D) Log_2 enrichment values comparing Round 6 to Round 1 were determined for the ImProm-II and BST trajectories. SSIV enrichment values compared Round 6 to Round 2 due to low number of reads in round 1. Points falling within the red region are ≥4-fold depleted. None of the points fell within the green region (≥4-fold enriched). Range of values is shown on the right. Solid line encompasses range of values excluding library 2. Dotted lines include enrichment values for library 2.

ImProm-II	**Round 1**	**Round 2**	**Round 3**	**Round 4**	**Round 5**	**Round 6**
Raw total reads	133,124	4,613	7,970	416,663	15,825	2,228
Long sequences	17,443	2,417	7,224	25,422	15,448	922
Unique long sequences	11,572	2,211	5,747	17,561	9,973	892
Short sequences	27,880	334	58	40,974	54	89
Unique short sequences	21,626	224	39	29,871	53	67
Total processed sequence reads	87,801	1,862	688	350,267	323	1,217
Unique processed sequences	67,029	1,472	613	267,184	323	935

Table 2.1. ImProm-II high-throughput sequencing raw data and processing. Data processing was performed using cutadapt to trim the 5' and 3' constant regions from sequences and to discard any uncut sequences or sequences with lengths not within ± 9 nt of the expected size (90 nt) after trimming. Raw total reads is the number of sequences prior to any processing. Long and short sequences did not fit within the ± 9 nt parameter. Total processed sequence reads were analyzed using FASTAptameR 2.0.

ImProm-II	**Round 1 RPM**	**Round 2 RPM**	**Round 3 RPM**	**Round 4 RPM**	**Round 5 RPM**	**Round 6 RPM**	**Enrichment (R6/R1)**
Library 1	412,080	469,925	88,663	424,288	9,288	98,603	0.24
Library 2	41,093	13,426	5,814	8,291	-	3,287	0.08
Library 3A	95,101	90,763	59,593	123,246	18,576	184,059	1.94
Library 3C	104,862	52,095	15,988	69,584	37,152	95,316	0.91
Library 4A	101,388	60,687	27,616	34,340	9,288	23,007	0.23
Library 4C	245,476	313,104	802,326	340,252	925,697	595,727	2.43

Table 2.2. ImProm-II pre-processed and analyzed HTS data. Library reads shown in reads per million (RPM). The processed reads were analyzed in FASTAptameR2.0 and each read was identified as belonging to one of the six libraries using sequence markers. The enrichment values of each of the six libraries calculated by comparing round 6 to round 1.

SSIV	**Round 1**	**Round 2**	**Round 3**	**Round 4**	**Round 5**	**Round 6**
Raw total reads	1,363	151,377	85,360	79,385	82,293	309,736
Long sequences	1,124	12,912	8,606	7,833	7,810	148,716
Unique long sequences	1,165	9,660	6,870	6,405	6,374	113,528
Short sequences	12	4,313	5,982	3,429	3,130	12,304
Unique short sequences	9	3,226	4,653	2,599	2,323	9,430
Total processed sequence reads	227	134,152	70,772	68,123	71,353	148,716
Unique processed sequences	189	102,421	56,624	51,989	54,201	113,528

Table 2.3. SSIV high-throughput sequencing raw data and processing. Data processing was performed using cutadapt to trim the 5' and 3' constant regions from sequences and to discard any uncut sequences or sequences with lengths no within ± 9 nt of the expected size (90 nt) after trimming. Raw total reads is the number of sequences prior to any processing, long and short sequences did not fit within the ± 9 nt parameter, and the total processed sequence reads were analyzed using FASTAptameR2.0.

SSIV	**Round 1 RPM**	**Round 2 RPM**	**Round 3 RPM**	**Round 4 RPM**	**Round 5 RPM**	**Round 6 RPM**	**Enrichment (R6/R2)**
Library 1	180,617	140,140	124,908	50,020	29,165	15,923	0.11
Library 2	61,674	58,642	33,686	14,698	11,408	5,769	0.1
Library 3A	114,537	137,732	89,612	75,470	74,755	72,938	0.53
Library 3C	132,159	211,074	277,384	343,626	346,292	330,287	1.56
Library 4A	136,564	153,006	169,036	170,273	167,029	149,856	0.98
Library 4C	374,449	299,407	305,375	345,914	371,351	425,227	1.42

Table 2.4. SSIV pre-processed and analyzed HTS data. Library reads shown in reads per million. The processed reads were analyzed in FASTAptameR2.0 and each read was identified as belonging to one of the six libraries using sequence markers. The enrichment values of each of the six libraries calculated by comparing round 6 to round 2 instead of round 1 due to low number of reads in round 1.

BST	**Round 1**	**Round 2**	**Round 3**	**Round 4**	**Round 5**	**Round 6**
Raw total reads	92,494	11,212	7,383	10,134	161,986	78,395
Long sequences	8,071	5,813	5,854	6,356	11,252	14,897
Unique long sequences	6,652	4,808	4,792	5,061	8,907	9,912
Short sequences	731	84	36	112	5,911	658
Unique short sequences	591	67	32	85	4,100	500
Total processed sequence reads	83,692	5,315	1,493	3,666	144,823	62,840
Unique processed sequences	64,627	4,020	1,182	2,774	113,385	47,653

Table 2.5. BST high-throughput sequencing raw data and processing. Data processing was performed using cutadapt to trim the 5' and 3' constant regions from sequences and to discard any uncut sequences or sequences with lengths no within ± 9 nt of the expected size (90 nt) after trimming. Raw total reads is the number of sequences prior to any processing, long and short sequences did not fit within the ± 9 nt parameter, and the total processed sequence reads were analyzed using FASTAptameR2.0.

BST	**Round 1 RPM**	**Round 2 RPM**	**Round 3 RPM**	**Round 4 RPM**	**Round 5 RPM**	**Round 6 RPM**	**Enrichment (R6/R1)**
Library 1	322,181	292,380	257,200	390,616	368,243	402,132	1.25
Library 2	76,196	36,877	32,150	11,184	17,925	15,213	0.2
Library 3A	96,987	105,174	121,902	93,290	106,468	90,484	0.93
Library 3C	126,762	139,040	117,883	103,928	113,228	98,727	0.78
Library 4A	117,824	127,940	82,384	54,010	65,556	56,779	0.48
Library 4C	260,050	298,589	388,480	346,972	328,580	336,665	1.29

Table 2.6. BST pre-processed and analyzed HTS data. Library reads shown in reads per million. The processed reads were analyzed in FASTAptameR2.0 and each read was identified as belonging to one of the six libraries using sequence markers. The enrichment values of each of the six libraries calculated by comparing round 6 to round 1.

DISCUSSION

This work provides an initial roadmap for evaluating RTs for amplification biases, focusing on simple measures of yield and processivity, selection outcomes, and fidelity. Of the five RTs we tested, BST exhibited the lowest bias between structured and non-structured templates and retained high processivity and activity to generate large quantities of full length cDNA product for templates of varying sizes and structures, including good yields for large, structured RNAs ranging from 100-350 nt. The 'amplification-only' selection results indicate that BST is a good candidate for cDNA synthesis during *in vitro* selections, as it introduces the least amount of inter-library bias between differently structured templates. Enrichment values for the six libraries generally showed the least amount of library-specific enrichment/depletion as compared to selections using SSIV and ImProm-II. BST also had comparable fidelity to SSIV and ImProm-II, suggesting that fidelity did not greatly impact the outcome of the amplification only selection (Figures 2.9B and 2.9C).

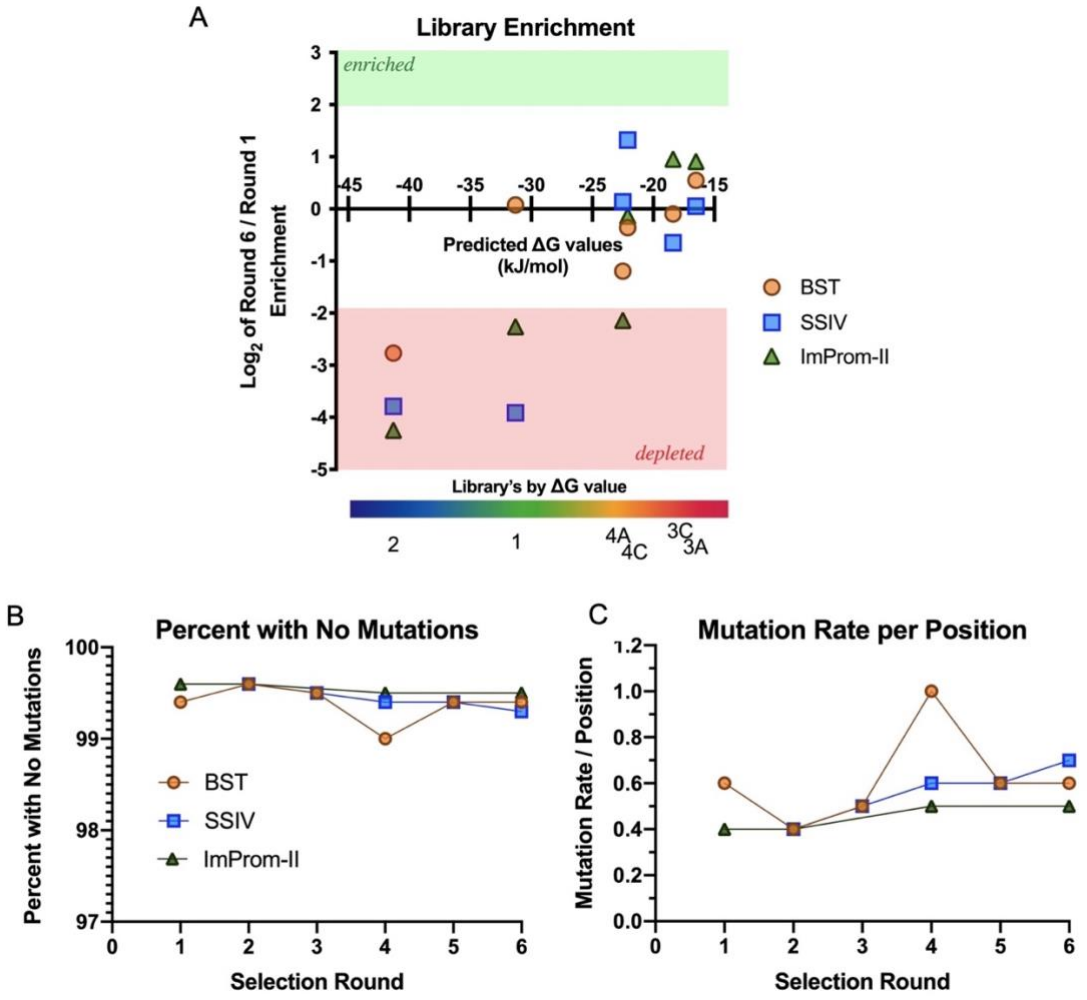


Figure 2.9. Alternative perspectives on amplification-only selection. (A) Log_2 enrichment values comparing Round 6 to Round 1 were determined for the ImProm-II and BST trajectories. SSIV enrichment values compared Round 6 to Round 2 due to low number of reads in round 1. Points falling within the red region are ≥4-fold depleted. None of the points fell within the green region (≥4-fold enriched). Libraries are arranged according to their predicted ΔG values based on their secondary structure (most negative ΔG values to most positive from left to right). As the predicted ΔG values get increasingly stable, the depletion increases. (B) Any rounds with fewer than 1,000 total unique processed reads were excluded from these data sets. Percent of nucleotides with no

mutations was calculated in each round for each trajectory and plotted. (C) Approximate mutation rate per position was determined for each selection round and plotted.

Despite using a single starting RNA pool for all three selections, reverse transcribing the pool back into cDNA during round 1 gave markedly different fractions of total reads among the six input libraries for the three RTs. For example, library 1 was poorly represented in the SSIV trajectory from the start as indicated by total fractions in round 2 (Figure 2.8C). Whereas Improm-ii had relatively high library representation in round 1, it slowly favored less structured sequences over the course of the selection, ultimately leading to structured libraries 1 and 2 being strongly underrepresented by the end (Figure 2.8C). For each of the three enzymes, the magnitude of the depletion roughly correlated with the calculated ΔG values for the designed elements of each library (Figure 2.9A), again indicating that low predicted ΔG values (stable structures) were responsible for the highest depletion values for each enzyme. Additional insights emerge by comparing the spread of relative enrichment across the different input libraries (maximum log enrichment - minimum log enrichment values, vertical lines on the right side of Figure 2.8D). The BST trajectory has a far smaller spread than the other two trajectories, particularly when library 2 (dashed vertical lines) is omitted. Cumulative error rates for library 1 across the six rounds were indistinguishable (0.3 to 0.7%) for the three trajectories (Figure 2.9B-C and Tables 2.7 – 2.9).

ImProm-II	**Round 1 RPM**	**Round 2 RPM**	**Round 3 RPM**	**Round 4 RPM**	**Round 5 RPM**	**Round 6 RPM**
Library 1 Total	412,080	469,925	N.D.	424,288	N.D.	98,603
Library 1 No Mutations	351,454	396,885	N.D.	333,691	N.D.	76,417
Percent of Sequences with ≥ 1 Mismatches	14.7	15.5	N.D.	21.4	N.D.	22.5
Adjusted Fidelity	99.6	99.6	N.D.	99.5	N.D.	99.5
Cumulative Mutation Rate/Position	0.4	0.4	N.D.	0.5	N.D.	0.5

Table 2.7. ImProm-II fidelity for library 1. Fraction of un-mutated and mutated library 1 sequences were determined for each round. Fraction of sequences with more than one mismatch was calculated by dividing library 1 with ≥ 1 mutations (library 1- no mutations subtracted from library 1) by total library 1 sequences. The mutation rate per position was calculated by diving the fraction of sequences with ≥ 1 mutations by the total number of nucleotides (42). Adjusted fidelity is (1 - mutation rate/position). In cases where the read count was too low, values were not determined (N.D.).

SSIV	**Round 1 RPM**	**Round 2 RPM**	**Round 3 RPM**	**Round 4 RPM**	**Round 5 RPM**	**Round 6 RPM**
Library 1 Total	N.D.	140,140	124,908	50,020	29,165	15,923
Library 1 No Mutations	N.D.	115,488	99,149	38,314	21,905	11,317
Percent of Sequences with ≥ 1 Mismatches	N.D.	17.6	20.6	23.4	24.9	28.9
Adjusted Fidelity	N.D.	99.6	99.5	99.4	99.4	99.3
Cumulative Mutation Rate/Position	N.D.	0.4	0.5	0.6	0.6	0.7

Table 2.8. SSIV fidelity for library 1. Fraction of un-mutated and mutated library 1 sequences were determined for each round. Fraction of sequences with more than one mismatch was calculated by dividing library 1 with ≥ 1 mutations (library 1- no mutations subtracted from library 1) by total library 1 sequences. The mutation rate per position was calculated by diving the fraction of sequences with ≥ 1 mutations by the total number of nucleotides (42). Adjusted fidelity is (1 - mutation rate/position). In cases where the read count was too low, values were not determined (N.D.).

BST	**Round 1 RPM**	**Round 2 RPM**	**Round 3 RPM**	**Round 4 RPM**	**Round 5 RPM**	**Round 6 RPM**
Library 1 Total	322,181	292,380	257,200	390,616	368,243	402,132
Library 1 No Mutations	247,421	240,640	208,305	230,769	282,614	305,490
Percent of Sequences with ≥ 1 Mutations	23.2	17.7	19.0	40.9	23.3	24.0
Adjusted Fidelity	99.4	99.6	99.5	99.0	99.4	99.4
Cumulative mutation Rate/Position	0.6	0.4	0.5	1.0	0.6	0.6

Table 2.9. BST fidelity for library 1. Fraction of un-mutated and mutated library 1 sequences were determined for each round. Fraction of sequences with more than one mismatch was calculated by dividing library 1 with ≥ 1 mutations (library 1- no mutations subtracted from library 1) by total library 1 sequences. The mutation rate per position was calculated by diving the fraction of sequences with ≥ 1 mutations by the total number of nucleotides (42). Adjusted fidelity is (1 - mutation rate/position).

Library 2's depletion from the BST trajectory was surprising within the context of the observed low inter-library bias and successful primer extension assays using library 2 as template (Figure 2.5). At least three models can be proposed to explain this observation. First, the RNA:primer ratios were different in the selection (20 pmol RNA: 30 pmol primer) and *in vitro* primer extension assays (25 pmol RNA:50 pmol primer). The selection used lower concentration of reverse primer to reduce the amount of non-specific products and primer-dimers formed during the subsequent PCR step. Perhaps higher concentrations of primer are required for sufficient primer binding and subsequent reverse transcription of library 2. Second, library 2 is predicted to have the most stable structure (most negative ΔG value) and is probably therefore the most challenging template for all three enzymes. Third, the primer extension assays in Figure 2.5 used an individual library as the sole template in each reaction, whereas the RT reaction during the selection used a mixed pool of all six libraries for templates. It may be that less structured templates (more positive predicted ΔG values) can be amplified more quickly whereas very stable highly structured templates require more time for proper primer annealing and complete amplification especially when they compete with preferred less structured templates. Although BST performed well on 5 out of 6 libraries, perhaps further optimization for selection conditions and pooled templates could further reduce inter-library bias for selections.

BST is inexpensive and easy to handle with simple reaction conditions, standardized pre-made buffers and solutions, and straight-forward quenching procedures (as outlined in the materials and methods), making it attractive for adoption for *in vitro* selections in addition to its well-established use for LAMP and potential extrapolation to broader applications,

such as RNA structural probing and cellular RNA library preparation for transcriptomics analysis. Several new RTs have recently been reported in the literature that have potential applications in specialized selections, such as RT-C8 evolved from a variant of the DNA polymerase from *Thermococcus gorgonarius* from the Holliger and Taylor groups for use with XNAs (47, 48) and RTX (an RT evolved *in vitro* from the B family DNA polymerase KOD) from the Ellington group (49), among others. Early reports of these RTs note robust activity on difficult RNA templates and in some cases their use in library amplification for *in vitro* selections. Although RT-C8 and RTX can be purified in house from bacteria carrying the appropriate plasmids, these enzymes are not yet commercially available and their potential impact on amplification bias with respect to structured templates is unknown. However, BST's strong performance in reverse transcribing RNA templates with variable degrees of structure makes it an attractive go-to RT enzyme for selections.

MATERIAL AND METHODS

RNA Transcripts

DNA templates (Table 2.10) were ordered from Integrated DNA Technologies and amplified by PCR using *Pfu* DNA polymerase. Sizes of double-stranded DNA (dsDNA) templates were confirmed by agarose gel electrophoresis (Figure 2.7). Each RNA was transcribed *in vitro* from the amplified PCR products using the Y639F T7 RNA polymerase (50), *in vitro* transcription buffer (1x = 50 mM Tris-HCl pH 7.5, 15 mM $MgCl_2$, 5 mM DTT, and 2 mM spermidine), and 2 mM each of ATP, UTP, GTP, CTP. Transcription reactions were incubated at 37°C overnight (approximately 16 hrs) and terminated by the addition of denaturing gel loading dye (90% formamide, 50 mM EDTA and 0.01% of

xylene cyanol and bromophenol blue). Transcripts were subsequently purified by denaturing polyacrylamide gel electrophoresis (5-8% TBE-PAGE, 8 M urea). Bands corresponding to the expected product sizes were visualized by UV shadow, excised from the gel, and eluted by tumbling overnight at 4°C in 300 mM sodium acetate pH 5.4. Eluates were ethanol precipitated, resuspended in nuclease-free water, and stored at -20°C until further use. A NanoDropOne spectrophotometer (Thermo Fisher Scientific) was used to determine specific RNA concentrations for all assays. The ΔG values from Figure 1B were estimated using Mfold (22) by forcing the depicted structured regions to pair and leaving randomized regions unfolded.

For testing BST 3.0 on structured viral genomic RNA, we used the previously stated sequences of 5' HIV UTR (51). The reverse primer aligns to positions 333-352 the Human Immunodeficiency Virus-1 (HIV-1) genome (NCBI Reference Sequence: NC_001802.1) and produces a cDNA product which maps to 1-352 of the HIV genome.

Name	Sequence (5' → 3')
Library 1	**AGGACCGGCCUAAACGGCAUUGC**ACUCCGCCGUAGGUAG CGNNNNNNNNNCGUGNNNNNNNNNNNNNNNNNNCAC GNNNNNNNNNNACCAUUCGAAAGAGUGGGACGCAA**ACCA AUCCGCUUCGGCGGAUACA**
Library 2	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNGAUGGN NNNNNNNNNNGCAAUGCCGUCAUGGCAANNNNNNNNNNN NNNNUUGCCAUGUGGCCGNNNNNNNNNNNNNNNUC**ACC AAUCCGCUUCGGCGGAUACA**
Library 3A	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN**ACC AAUCCGCUUCGGCGGAUACA**
Library 3C	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNGCCGGUNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN**ACC AAUCCGCUUCGGCGGAUACA**
Library 4A	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNGGUCCNNNNNNNNNNNNNNGG**ACC AAUCCGCUUCGGCGGAUACA**
Library 4C	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNGGUCCNNNNNGCCGGUNNNNNNNNNNGG**ACC AAUCCGCUUCGGCGGAUACA**
Library Reverse Primer	TGTATCCGCCGAAGCGGATTGGT
Library Forward Primer	GCGTAATACGACTCACTATTAGGACCGGC
5' HIV UTR Reverse Primer	CCGACGCTCTCGCACCCATCTC

Table 2.10. Library and primer sequences. Primer binding sites of the library sequences are bolded and underlined. Regions with red letters represent incorporated structural regions. Segments highlighted in yellow were used to identify library of origin during high-throughput sequencing analysis. Note, library 3A's yellow region was fully randomized and therefore library 3A sequences were identified through process of elimination.

Radiolabeling primers

500 pmol of reverse primer (Integrated DNA Technologies) was radiolabeled using 1 μL of 250 μCi of γ^{32}P ATP (Perkin Elmer) and 50 units of T4 polynucleotide kinase (PNK, New England Biolabs) at 37°C for 90 min in 1X T4 PNK Buffer. Following incubation, 5' end-labeled primers were PAGE purified. A NanoDropOne spectrophotometer (Thermo Fisher Scientific) was used to determine concentration of primers prior to use in primer extension assays.

Primer extension assays

Primer extension assays were performed using γ^{32}P 5' end-labeled reverse primer and the corresponding RNA templates (see above). Figures 2.3-2.5 used a primer:RNA ratio of 2:1 (50 pmol:25 pmol) and Figure 2.8 used a ratio of 1:2 (25 pmol:50 pmol). All primer extension reactions were quenched with 2 volumes (60 μL) of denaturing gel loading buffer (95% formamide with 50 mM EDTA and 0.01% of bromophenol blue and xylene cyanol), heated to 95°C for 90 sec and analyzed by denaturing PAGE. Reactions that included additives used final concentrations of 2 M betaine (Sigma), 0.6 M trehalose (Sigma), and/or 1μg of Single Stranded Binding Protein ('SSB,' Thermo Fisher).

Reaction conditions were chosen based on manufacturer's recommended protocols for optimal cDNA synthesis. For ImProm-II RT (Promega) reactions, primer-template complexes were pre-assembled and incubated at 70°C for 5 min and cooled on ice for an additional 5 min. Reverse transcription was then initiated in 1x ImProm-II RT buffer (Promega), 10.6 mM $MgCl_2$, 0.6 mM each dNTP, and 240 U ImProm-II. SSIV RT

(Thermo Fisher) reactions were initiated as previously described (29). Briefly, 0.5 mM of dNTP mix, 1x SSIV buffer (Thermo Fisher), 7.5 mM DTT, template-primer, and 300 U SSIV were incubated at a final volume of 30 μL. TGIRT-III (InGex) reactions were performed as previously described (35) with minor changes. Briefly, primer-template complexes were pre-assembled in 1x annealing reaction buffer (1 mM Tris-HCl, pH 7.5, 1 mM EDTA) at 82°C for 2 min and cooled to 25°C with a 10% ramp (0.1°C /sec). Reverse transcription was initiated in 1x RT reaction buffer (450 mM NaCl, 5 mM $MgCl_2$, 20 mM Tris-HCl, pH 7.5), 5 mM DTT, and 10 pmol TGIRT-III enzyme in a volume of 30 μL. After a pre-incubation at 30 min at 25°C, 1 μL of 25 mM dNTPs was added. Prior to quenching, 1 μL of 5M NaOH was added to the reaction mixture and incubated at 95°C for 3 min to interrupt very strong binding of the TGIRT-III enzyme to the RNA. The reaction was then cooled to room temperature and neutralized with 1 μL of 5M HCl and quenched as described above (52, 53). MarRT (Kerafast) reactions were performed as previously described with minor changes (54). Primer-template complexes were pre-assembled and incubated at 95°C for 30 sec and snap cooled on ice. Reverse transcription was initiated in 1x reaction buffer (50 mM Tris-HCl, pH 8.3, 200 mM KCl, 2 mM $MgCl_2$, 5 mM DTT, 20% glycerol), 0.33 mM of a dNTP mix, and 20 U MaRT enzyme in a reaction volume of 30 μL. BST 3.0 DNA Polymerase (New England Biolabs) reactions were initiated without a pre-assembly step unless otherwise noted. Reactions included labeled primer, template, 1X isothermal buffer (New England Biolabs), 6 mM $MgSO_4$, 1.6 mM dNTP mix, and 16 U BST enzyme in a 25 μL reaction volume.

Analysis of primer extension assays

Primer extension reactions were analyzed by denaturing PAGE. Gels were exposed and scanned for ^{32}P using a Typhoon FLA 9000 phosphorimager (GE Healthcare Life Sciences). Full length product, partially extended product, and unextended primer were quantified by measuring band intensities using Multigauge software (Fujifilm) and plotted in Prism. Yield was calculated by dividing full length product by the sum of unextended primer and partial/fully extended product. In assays where primer:RNA ratios were 2:1, yield was multiplied by 2 to account for excess unincorporated primer. Processivity was calculated by dividing full length product by the sum of all partial or fully extended product, excluding unextended primer.

Amplification only selection

Approximately 100 pmol each of the six libraries were combined to make a single, pooled RNA starting library. After round 1, 20 pmol of RNA (consisting of the mixed libraries) and 30 pmol of reverse primer were used for the reverse transcription steps. ImProm-II, SSIV, and BST reverse transcription reactions were set up and quenched as described above, using optimized conditions identified in data from Figures 2.3 and 2.4. Reactions with ImProm-II were incubated at 42°C for 1 hr without any additives. SSIV reactions were incubated in the presence of 1μg single stranded binding protein at 55°C for 30 min. BST reactions were incubated at 65°C for 1 hr with 2 M betaine, 0.6 M trehalose, 1 μg single stranded binding protein. Following reverse transcription, 75% of the cDNA (15 μL of the 20 μL SSIV and ImProm-II RT reactions and 22.5 μL of the 30 μL BST reactions) was used as template for PCR (15 cycles) by adding 30 pmol forward primer (no additional

reverse primer beyond the carryover excess reverse primer from the RT reaction) and Pfu DNA polymerase as described above. Amplified product was run on a 2% agarose gel to confirm product size before being transcribed for 4 hrs at 37°C for the next cycle as described above. 30 μL of the 100 μL PCR reaction volume served as template in a 100 μL transcription. Transcriptions were set up as described above. Transcribed RNA was run on a PAGE gel and ethanol precipitated as described above before being used in reverse transcription reactions.

HTS sequencing and data analysis

Libraries were prepared for sequencing using a series of PCR steps to append Illumina adapters and sequencing indices for multiplexing of the libraries as previously described (55). Primers used to append the Illumina adapters and sequencing indices can be found in Table 2.11. Sequencing was performed on an Illumina NextSeq 500 (University of Missouri Genomics Technology Core). Although paired-end reads generated Read 1 and Read 2 for each selection round, a single 150 nt read provided enough coverage such that no additional information was gained through Read pairing. Populations were demultiplexed, and the relevant sequence information was found and used from Read 1 (5′ constant region, 90 nt library sequence, 3′ constant region) and all data shown represents reads from Read 1 only. Data preprocessing was performed using cutadapt (56) to trim 5′ and 3′ constant regions and to discard any uncut sequences or sequences with lengths not within ± 9 nt of the expected size (90 nt) after trimming (Tables 2.1, 2.3, and 2.5). These populations were then analyzed using FASTAptameR 2.0 (57, 58) to count and normalize reads (FASTAptameR-Count) and find library motifs (FASTAptameR-Motif Search)

(found in Tables 2.2, 2.4, and 2.6) across multiple rounds to calculate library RPM (total reads that contained library motif/total reads in round * 1,000,000) and enrichment values (ratio of RPM in round y to RPM in round x). These populations were then analyzed using FASTAptameR 2.0 (57, 58) to count and normalize reads (FASTAptameR-Count) and find library motifs (FASTAptameR-Motif Search) (found in Tables 2.2, 2.4, and 2.6) across multiple rounds to calculate library RPM (total reads that contained library motif/total reads in round * 1,000,000) and enrichment values (ratio of RPM in round y to RPM in round x). For libraries 1 and 2, the two longest stretches of designed sequence were used as independent search terms (highlighted yellow regions, Table 2.10). Searches were performed as OR functions to allow detection of sequences that had accumulated one or more mutations in one (but not both) segment. Libraries 3C, 4A, and 4C had short defined sequences so the random region was used in the search terms to ensure the defined sequences were located at the indicated positions. Library 3A was fully randomized and therefore did not have sequences that could be used for detection; hence, it was identified through process of elimination after identifying sequences belonging to the other five libraries.

For the mutational analysis, FASTAptameR-Motif Search was used to search for library 1 using the same terms as described above using AND functions (both segments had to be present without mismatches) to calculate the library 1 RPM containing no mutations (total reads containing no mutations/ total reads per round *1,000,000). The difference in the numbers of hits found by the two methods (OR – AND) provides the approximate number of total reads with one mismatch in one or the other segment (Tables 2.7-2.9). Dividing

by the total nucleotides in the two segments (42) gives the estimated cumulative mutation rate per position $\mu = (OR - AND)/42$.

Name	Sequence (5’ → 3’)
Forward Primer for HTS	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGACCGGCCTAAACGGCATT
Reverse Primer 1 for HTS	CAGACGTGTGCTCTTCCGATCTGTATCCGCCGAAGCGGATTGG
Reverse Primer 2 for HTS	CAAGCAGAAGACGGCATACGAGAT**NNNNNN**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTATCCGCCGAAGCGGATTGG
5’ Universal HTS Adapter	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
3’ Indexed HTS Adapter	CAAGCAGAAGACGGCATACGAGAT**NNNNNN**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Table 2.11. Sequences for the high-throughput sequencing primers used to append the Illumina adapters and their respective sequencing indices. We used the NEBNext Index (1-38) Primers for Illumina. 38 reverse primers (corresponding to the 38 indices) were used in the second PCR for high-throughput sequencing preparation. The index region is indicated by the six red N region in the reverse primer 2 for HTS sequence. Index sequences are from the instruction manual for the NEBNext Multiplex Small RNA Library Prep Set 1, Set 2, Index Primers 1-48 and Multiplex Compatible (https://rb.gy/0dbqe9).

ACKNOWLEDGEMENTS

We would like to thank Dr. Margaret Lange for her critical read of the manuscript, Skyler Kramer and Kevin Muñoz-Forti for their insights, contributions, and help with generating Figure 2.3F, members of the Burke and Lange laboratories for suggestions and constructive feedback throughout the project, and Eric Strobel (@NADFDGD) and others in the Twitterverse for suggesting that we investigate BST. This research was supported by NASA Exobiology grant NNX17AE88G and NASA Interdisciplinary Consortium for Astrobiology Research "Bringing RNA to Life" grant 80NSSC21K0596 to D.H.B. and by University of Missouri Life Sciences Fellowship and Wayne L. Ryan Graduate Fellowship from the Ryan Foundation to P.R.G.

REFERENCES

1. Ellington, A. D., and Szostak, J. W. (1990) *In vitro selection of RNA molecules that bind specific ligands*
2. Tuerk, C., and Gold, L. (1990) Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science (80-.).* **249**, 505–510
3. Coleman, T. M., and Huang, F. (2002) RNA-catalyzed thioester synthesis. *Chem. Biol.* **9**, 1227–1236
4. Sabeti, P. C., Unrau, P. J., and Bartel, D. P. (1997) Accessing rare activities from random RNA sequences: the importance of the length of molecules in the starting pool. *Chem. Biol.* **4**, 767–774
5. Davis, J. H., and Szostak, J. W. (2002) Isolation of high-affinity GTP aptamers

from partially structured RNA libraries. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11616–11621

6. Chushak, Y., and Stone, M. O. (2009) In silico selection of RNA aptamers. *Nucleic Acids Res.* **37**, e87
7. Kim, N., Hin, H. G., and Schlick, T. (2007) A computational proposal for designing structured RNA pools for in vitro selection of RNAs. *RNA.* **13**, 478–492
8. Carothers, J. M., Oestreich, S. C., and Szostak, J. W. (2006) Aptamers selected for higher-affinity binding are not more specific for the target ligand. *J. Am. Chem. Soc.* **128**, 7929–7937
9. Chizzolini, F., Luiz, #, Passalacqua, F. M., Oumais, M., Dingilian, A. I., Szostak, J. W., and Lupták, A. L. (2020) Large Phenotypic Enhancement of Structured Random RNA Pools. *Cite This J. Am. Chem. Soc*. 10.1021/jacs.9b11396
10. Carothers, J. M., Oestreich, S. C., Davis, J. H., and Szostak, J. W. (2004) Informational Complexity and Functional Activity of RNA Structures. *J. Am. Chem. Soc.* **126**, 5130–5137
11. Ishikawa, J., Matsumura, S., Jaeger, L., Inoue, T., Furuta, H., and Ikawa, Y. (2009) Rational optimization of the DSL ligase ribozyme with GNRA/receptor interacting modules. *Arch. Biochem. Biophys.* **490**, 163
12. Seelig, B., and Szostak, J. W. (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature*. **448**, 828–831
13. Ekland, E. H., Szostak, J. W., and Bartel, D. P. (1995) Structurally Complex and Highly Active RNA Ligases Derived from Random RNA Sequences. *Science (80-.).* **269**, 364–370

14. Pobanz, K., and Lupták, A. (2016) Improving the odds: Influence of starting pools on in vitro selection outcomes. *Methods*. **106**, 14–20

15. Porter, E. B., Polaski, J. T., Morck, M. M., and Batey, R. T. (2017) Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors. *Nat. Chem. Biol.* **13**, 295–301

16. Luo, X., McKeague, M., Pitre, S., Dumontier, M., Green, J., Golshani, A., Derosa, M. C., and Dehne, F. (2010) Computational approaches toward the design of pools for the in vitro selection of complex aptamers. *RNA.* **16**, 2252–2262

17. Ikawa, Y., Tsuda, K., Matsumura, S., and Inoue, T. (2004) De novo synthesis and development of an RNA enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13750

18. Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A., and Weeks, K. M. (2015) Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc. 2015 1011*. **10**, 1643–1669

19. Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J. W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T., and Chang, H. Y. (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*. **519**, 486–490

20. Mayer, G., Müller, J., and Lünse, C. E. (2011) RNA diagnostics: real-time RT-PCR strategies and promising novel target RNAs. *Wiley Interdiscip. Rev. RNA.* **2**, 32–41

21. Thiel, W. H., Bair, T., Wyatt Thiel, K., Dassie, J. P., Rockey, W. M., Howell, C. A., Liu, X. Y., Dupuy, A. J., Huang, L., Owczarzy, R., Behlke, M. A., McNamara,

J. O., and Giangrande, P. H. (2011) Nucleotide bias observed with a short SELEX RNA aptamer library. *Nucleic Acid Ther.* **21**, 253–263

22. Markham, N. R., and Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.* 10.1093/NAR/GKI591

23. Gruenke, P. R., Aneja, R., Welbourn, S., Ukah, O. B., Sarafianos, S. G., Burke, D. H., and Lange, M. J. (2022) Selection and identification of an RNA aptamer that specifically binds the HIV-1 capsid lattice and inhibits viral replication. *Nucleic Acids Res.* **50**, 1701–1717

24. Alam, K. K., Chang, J. L., Lange, M. J., Nguyen, P. D. M., Sawyer, A. W., and Burke, D. H. (2018) Poly-Target Selection Identifies Broad-Spectrum RNA Aptamers. *Mol. Ther. - Nucleic Acids*. **13**, 605–619

25. Gruenke, P. R., Alam, K. K., Singh, K., and Burke, D. H. (2020) 2'-fluoro-modified pyrimidines enhance affinity of RNA oligonucleotides to HIV-1 reverse transcriptase. *RNA.* **26**, 1667–1679

26. Kang, H. S., Huh, Y. M., Kim, S., and Lee, D.-K. (2009) Isolation of RNA Aptamers Targeting HER-2-overexpressing Breast Cancer Cells Using Cell-SELEX. *RNA Aptamers Bull. Korean Chem. Soc*. **30**, 1827

27. Dua, P., Kang, S., Shin, H. S., Kim, S., and Lee, D. K. (2018) Cell-SELEX-Based Identification of a Human and Mouse Cross-Reactive Endothelial Cell-Internalizing Aptamer. *https://home.liebertpub.com/nat*. **28**, 262–271

28. Rahimi, F., and Bitan, G. (2010) Selection of Aptamers for Amyloid β-Protein, the Causative Agent of Alzheimer's Disease. *JoVE (Journal Vis. Exp.* 10.3791/1955

29. Saha, R., Verbanic, S., and Chen, I. A. (2018) Lipid vesicles chaperone an

encapsulated RNA aptamer. *Nat. Commun. 2018 91*. **9**, 1–11

30. Ning, L., Wang, X., Xu, K., Song, S., Li, Q., and Yang, X. (2019) A novel isothermal method using rolling circle reverse transcription for accurate amplification of small RNA sequences. *Biochimie*. **163**, 137–141

31. Akoopie, A., Arriola, J. T., Magde, D., and Müller, U. F. (2021) A GTP-synthesizing ribozyme selected by metabolic coupling to an RNA polymerase ribozyme. *Sci. Adv.* 10.1126/SCIADV.ABJ7487/SUPPL_FILE/SCIADV.ABJ7487_SM.PDF

32. Moretti, J. E., and Müller, U. F. (2014) A ribozyme that triphosphorylates RNA 5′-hydroxyl groups. *Nucleic Acids Res.* **42**, 4767–4778

33. Pressman, A. D., Liu, Z., Janzen, E., Blanco, C., Müller, U. F., Joyce, G. F., Pascal, R., and Chen, I. A. (2019) Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA. *J. Am. Chem. Soc.* **141**, 6213–6223

34. Strobel, E. J., Cheng, L., Berman, K. E., Carlson, P. D., and Lucks, J. B. (2019) A mechanism for ligand gated strand displacement in ZTP riboswitch transcription regulation. *bioRxiv*. 10.1101/521930

35. Mohr, S., Ghanem, E., Smith, W., Sheeter, D., Qin, Y., King, O., Polioudakis, D., Iyer, V. R., Hunicke-Smith, S., Swamy, S., Kuersten, S., and Lambowitz, A. M. (2013) Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA*. **19**, 958–970

36. Zhao, C., Liu, F., and Pyle, A. M. (2018) An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA*. **24**, 183–185

37. Agustriana, E., Nuryana, I., Laksmi, F. A., Dewi, K. S., Wijaya, H., Rahmani, N., Yudiargo, D. R., Ismadara, A., Helbert, Hadi, M. I., Purnawan, A., and Cameliawati Djohan, A. (2022) Optimized expression of large fragment DNA polymerase I from Geobacillus stearothermophilus in Escherichia coli expression system. *Prep. Biochem. Biotechnol.* 10.1080/10826068.2022.2095573

38. Wang, G., Ding, X., Hu, J., Wu, W., Sun, J., and Mu, Y. (2017) Unusual isothermal multimerization and amplification by the strand-displacing DNA polymerases with reverse transcription activities. *Sci. Reports 2017 71*. **7**, 1–10

39. Kabir, M. S., Clements, M. O., and Kimmitt, P. T. (2015) RT-Bst: an integrated approach for reverse transcription and enrichment of cDNA from viral RNA. *Br. J. Biomed. Sci.* **72**, 1–6

40. Padzil, F., Mariatulqabtiah, A. R., Tan, W. S., Ho, K. L., Isa, N. M., Lau, H. Y., Abu, J., and Chuang, K. P. (2022) Loop-mediated isothermal amplification (Lamp) as a promising point-of-care diagnostic strategy in avian virus research. *Animals*. 10.3390/ani12010076

41. Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006) Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc. 2006 13*. **1**, 1610–1616

42. Lentzsch, A. M., Stamos, J. L., Yao, J., Russell, R., and Lambowitz, A. M. (2021) Structural basis for template switching by a group II intron–encoded non-LTR-retroelement reverse transcriptase. *J. Biol. Chem.* 10.1016/J.JBC.2021.100971

43. Xu, H., Yao, J., Wu, D. C., and Lambowitz, A. M. (2019) Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer

formation and bias correction. *Sci. Rep.* 10.1038/s41598-019-44457-z

44. Santoro, M. M., Liu, Y., Khan, S. M. A., Hou, L. X., and Bolen, D. W. (1992) Increased Thermal Stability of Proteins in the Presence of Naturally Occurring Osmolytes. *Biochemistry*. **31**, 5278–5283

45. Spiess, A. N., and Ivell, R. (2002) A highly efficient method for long-chain cDNA synthesis using trehalose and betaine. *Anal. Biochem*. **301**, 168–174

46. Perales, C., Cava, F., Meijer, W. J. J., and Berenguer, J. (2003) Enhancement of DNA, cDNA synthesis and fidelity at high temperatures by a dimeric single-stranded DNA-binding protein. *Nucleic Acids Res.* **31**, 6473–6480

47. Houlihan, G., Arangundy-Franklin, S., Porebski, B. T., Subramanian, N., Taylor, A. I., and Holliger, P. (2020) Discovery and evolution of RNA and XNA reverse transcriptase function and fidelity. *Nat. Chem. 2020 128*. **12**, 683–690

48. Hervey, J. R. D., Freund, N., Houlihan, G., Dhaliwal, G., Holliger, P., and Taylor, A. I. (2022) Efficient synthesis and replication of diverse sequence libraries composed of biostable nucleic acid analogues. *RSC Chem. Biol.* **3**, 1209–1215

49. Choi, W. S., He, P., Pothukuchy, A., Gollihar, J., Ellington, A. D., and Yang, W. (2020) How a B family DNA polymerase has been evolved to copy RNA. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 21274–21280

50. Sousa, R., and Padilla, R. (1995) A mutant T7 RNA polymerase as a DNA polymerase. *EMBO J.* **14**, 4609–4621

51. Sampathkumar, R., Scott-Herridge, J., Liang, B., Kimani, J., Plummer, F. A., and Luo, M. (2017) HIV-1 Subtypes and 5 LTR-Leader Sequence Variants Correlate with Seroconversion Status in Pumwani Sex Worker Cohort. *Viruses*. **10**, 4

52. Nottingham, R. M., Wu, D. C., Qin, Y., Yao, J., Hunicke-Smith, S., and Lambowitz, A. M. (2016) RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *RNA.* **22**, 597–613

53. Qin, Y., Yao, J., Wu, D. C., Nottingham, R. M., Mohr, S., Hunicke-Smith, S., and Lambowitz, A. M. (2016) High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *RNA.* **22**, 111–128

54. Guo, L. T., Adams, R. L., Wan, H., Huston, N. C., Potapova, O., Olson, S., Gallardo, C. M., Graveley, B. R., Torbett, B. E., and Pyle, A. M. (2020) Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-Generation Reverse Transcriptase. *J. Mol. Biol.* **432**, 3338–3352

55. Ditzler, M. A., Lange, M. J., Bose, D., Bottoms, C. A., Virkler, K. F., Sawyer, A. W., Whatley, A. S., Spollen, W., Givan, S. A., and Burke, D. H. (2013) High-throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase. *Nucleic Acids Res.* **41**, 1873–1884

56. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10–12

57. Alam, K. K., Chang, J. L., and Burke, D. H. (2015) FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Mol. Ther. Nucleic Acids*. **4**, e230

58. Kramer, S. T., Gruenke, P. R., Alam, K. K., Xu, D., and Burke, D. H. (2022) FASTAptameR 2.0: A web tool for combinatorial sequence selections. *Mol. Ther. Nucleic Acids*. **29**, 862–870

Chapter 3: Post-transcriptional capping generates Coenzyme A-linked RNA

This chapter will be modified for submission. Anticipated authors include Krishna Sapkota, Jordyn K. Lucas*, Matthew F. Lichte, Yan-Lin Guo, Donald H. Burke and Faqing Huang. * KS and JKL contributed equally.*

ABSTRACT

More than a decade after initial reports of CoA-linked RNA in bacteria, their biogenesis, metabolism, functional roles, and sequence identities remain unknown. While co-transcriptional insertion via non-canonical initiation has been demonstrated for NAD+, that mechanism is unlikely for CoA-linked RNAs due to low intracellular concentration of the required initiator nucleotide, 3' dephospho CoA (dpCoA). Instead, we found that phosphopantetheine adenylyltransferase (PPAT), an enzyme of CoA biosynthetic pathway, accepts RNA transcripts as its acceptor substrate and transfers 4'-phosphopantetheine to yield CoA-RNA post-transcriptionally. Synthetic natural (RNAI) and artificial (22nt stem-loop) RNAs were used to characterize the essential features of RNA that are needed for it to serve as PPAT substrate. RNAs with 4-10 unpaired nucleotides at the 5' terminus served as PPAT substrates, but RNAs having <4 unpaired nucleotides did not undergo capping. No capping was observed when the +1A was changed to G or when 5' triphosphate was removed by RNA pyrophosphohydrolase (RppH), suggesting that the enzyme recognizes pppA-RNA as an ATP analog. However, no significant differences in binding affinities

were observed between PPAT and +1A, +1G, or 5'OH (+1A) RNA, indicating that productive enzymatic recognition is likely driven by local positioning effects and not by overall binding affinity. The rate of capping was independent of the number of unpaired nucleotides in the range of 4-10 nucleotides. The capping reaction was strongly inhibited by ATP as the value of k_{obs} was reduced by ~90% when equimolar amounts ATP and substrate RNA were present. Dual bacterial expression of candidate RNAs with different 5' structural features, followed by CoA-RNA CaptureSeq, revealed >10-fold enrichment of the better PPAT substrate, consistent with *in vivo* CoA-capping of RNA transcripts by PPAT. Overall, this study suggests post-transcriptional RNA capping as a possible mechanism for the biogenesis of CoA-RNAs in bacteria.

INTRODUCTION

Unlike eukaryotic mRNA, for which the canonical 7-methylguanosine (m^7G) cap is a ubiquitous feature, bacterial transcripts are not typically associated with 5' caps. This view changed as Gram negative *Escherichia coli* and Gram positive *Streptomyces venezulae* were both found to cap some of their RNAs with metabolic cofactors NAD^+, CoA, and CoA-thioesters at the 5' end (1, 2). Studies on the functions, diversity, and the mechanism of capping are illuminating the biological significance of these non-canonical caps (3–9). For example, NAD^+ capping in *E. coli* was found to protect RNA from RNase E mediated degradation (10), while in eukaryotes NAD-ylation promoted RNA decay by DXO-mediated deNAD-ylation (6). Both co-transcriptional and post-transcriptional mechanisms

have been proposed to explain RNA capping with NAD^+ and 3'-dephospho-CoA (dpCoA) (1, 11, 12). Given suitable promoters with A in the +1 position, several RNA polymerases can initiate transcription *in vitro* with adenosine cofactors such as NAD^+, FAD, and dpCoA, including T7 RNAP, *E. coli* RNAP and human mitochondrial RNAP (12–15). Moreover, T7 RNAP was also shown to initiate transcription efficiently with several non-biological adenosine containing molecules *in vitro* (14, 16, 17). These findings substantiate the co-transcriptional mechanism of capping, essentially via competition with the canonical NTP initiator nucleotide.

Mechanistic investigation into transcription initiation by non-canonical initiator nucleotides (NCINs) has focused on NAD^+-linked RNAs, fueled in part by the availability of capture methods for these RNA species (10, 18). *E. coli* RNAP was shown to initiate transcription with NAD^+ under a consensus promoter HRR$\mathbf{A}_{+1}$SWW (19). This mode of NAD-ylation is highly efficient and yields up to ≈15% of NAD-linked transcripts *in vivo* at high [NAD+]/[ATP] ratios (20). Intracellular ATP concentrations decrease during stationary phase, when the [NAD+]/[ATP] ratio can be higher than 2:1, allowing for more efficient NADylation of RNAs. The Km value of NAD^+ utilization by *E. coli* RNAP (0.38 mM) is an order of magnitude lower than the intracellular NAD^+ concentration, which typically fluctuates between 4-7 mM in *E. coli* (21). Although ATP is the preferred substrate (Km ≈ 0.09 mM for *E. coli* RNAP), the comparable concentrations and Km values of these substrates make it feasible for *E. coli* RNAP to incorporate NAD+ into the +1

position of RNA transcripts. Thus, a co-transcriptional mechanism could account for much of the NAD-RNA found in cells. Further supporting this notion, Cahová et al. reported around 26% of sRNA RNAI to be capped with NAD in the absence of de-capping enzyme nudix phosphohydrolase NudC(10). The situation for CoA-RNA is dramatically different from NAD-RNA, as the intracellular concentration of dpCoA is approximately 100-200 times lower than that of NAD^+ (22, 23), making transcription initiation with dpCoA very unlikely. Thus, to the extent that CoA-RNA transcripts exist naturally or can be engineered to form within bacterial transcriptomes, they are more likely to arise from post-transcriptional mechanisms.

CoA is an indispensable metabolic cofactor participating in diverse acyl transfer reactions, from the TCA cycle and fatty acid metabolism to metabolite biosynthesis and gene regulation. In *E. coli*, the *de novo* synthesis of CoA involves five enzymatic reactions starting with the phosphorylation of pantothenic acid (vitamin B5) by pantothenate kinase (PanK/CoaA). The CoaBC complex converts the resulting phosphopantothenate (Pant) to phosphopantetheine (pPant) by sequential cysteinylation and decarboxylation. Adenylation of pPant by phosphopantetheine adenylyltransferase (PPAT) yields dpCoA, which is then phosphorylated at 3'-OH by CoaE, completing the pathway.

Several of the CoA biosynthetic enzymes – including CoaA and PPAT – exhibit relaxed substrate specificities that allow a broad range of modifications on pantetheine, making the

enzymatic synthesis of diverse CoA analogs possible (24, 25), and desulfopantetheine has been shown to be incorporated into CoA by beef liver enzymes (26). Due to the critical requirement of CoA in metabolism, these analogs have proven useful for the development of novel therapeutics, including antibiotics (27, 28). We asked whether PPAT's ability to accept non-canonical substrates extends beyond pantetheine to include ATP-RNAs, thus providing a post-transcriptional capping mechanism for the biosynthesis of CoA-RNAs.

We report here that PPAT accepts ATP-RNA in place of ATP, yielding CoA-RNA *in vitro*. We have characterized the essential structural features of the substrate RNA at the 5' terminus and determined that a 5' triphosphate, a 5'-terminal adenosine, and a minimum of four unpaired nucleotides must be present at the RNA 5' terminus for the PPAT mediated pPant capping. Furthermore, using the *in vitro* substrate requirements as a guide, we expressed, partitioned, and sequenced D2*, a poor PPAT RNA substrate, and D7*, a good PPAT RNA substrate from *E. coli* cells. Interestingly, RNA samples that underwent a sulfur-partition method to isolate CoA-RNAs, showed more than a 10-fold increase in D7*:D2* RNA ratios as compared to the total RNA samples, suggesting D7* templates may have been preferentially capped by PPAT *in vivo* to generate CoA-RNAs. Overall, both the *in vitro* and *in vivo* experiments suggest that bacterial CoA-RNAs may be generated as a product of post-transcriptional capping by PPAT.

RESULTS

E. coli PPAT enzyme can cap native RNA *in vitro* to form CoA-RNA *(data generated by K. Sapkota)*

RNAI is an antisense RNA that regulates the replication and copy number of some plasmids in *E. coli*, such as ColE1 (29), and it is the most abundant NAD-ylated RNA in *E. coli* (10). We tested whether RNAI can serve as a PPAT substrate and be capped with pPant to yield CoA-RNA *in vitro*. A DNA template for RNAI transcription *in vitro* was prepared by PCR. We fused the φ2.5 promoter and a dinucleotide AG upstream of the 109 nt RNAI sequence to meet the requirements of transcription by T7 RNAP with adenosine in the +1 position.

Based on our prior experience with CoA-RNA transcripts formed either co-transcriptionally or through the action of self-capping ribozymes, the transfer of pPant to RNAI is expected to slow its migration on PAGE so that it moves as N+1 (110 nt) RNA. However, the large size of RNAI makes it challenging to resolve 109 nt RNA (AG-RNAI) from 110 nt (CoA-G-RNAI) unambiguously by PAGE. The internal radiolabelling of RNA by using ATP [α-^{32}P] is also not ideal as the gel mobility of CoA RNA-product is equivalent to 3'-extended N+1 RNA, which may alternatively be generated by the transcription itself (30). To make the visualization of the CoA-RNA product simple and straightforward, we synthesized a pPant analog BKPP tagged with ^{14}C and biotin (Figure

3.1) so that only the RNA that has BKPP cap will be visible upon exposure to a phosphor screen.

When BKPP and synthetic RNAI with +1A were used as PPAT substrates, we observed a transfer of the pPant analog to RNA, yielding CoA-RNAI analog *in vitro* (lane 2, Figure 3.2a). Since BKPP contains a biotin, addition of streptavidin retarded the electrophoretic mobility of the product on the gel (lane 3 Figure 3.2a & lane 4 Figure 3.2b), which further confirmed that PPAT successfully capped RNAI with pPant analog BKPP. We did not observe the BKPP capping of RppH-treated RNAI (lanes 4 & 5, Figure 3.2). RppH treatment removes 5' pyrophosphate from pppRNAI yielding pRNAI, which did not meet the requirements to be an enzyme substrate. To further investigate whether an ATP at +1 position is a strict requirement, we prepared RNAI with GTP at +1 position by *in vitro* transcription under T7 φ6.5 promoter. When pppG-RNAI was used as a PPAT substrate, the enzyme did not catalyze the transfer of BKPP to RNA. These experiments established that RNA requires a 5' triphosphate and an adenosine at +1 position to serve as a PPAT substrate and get capped with CoA.

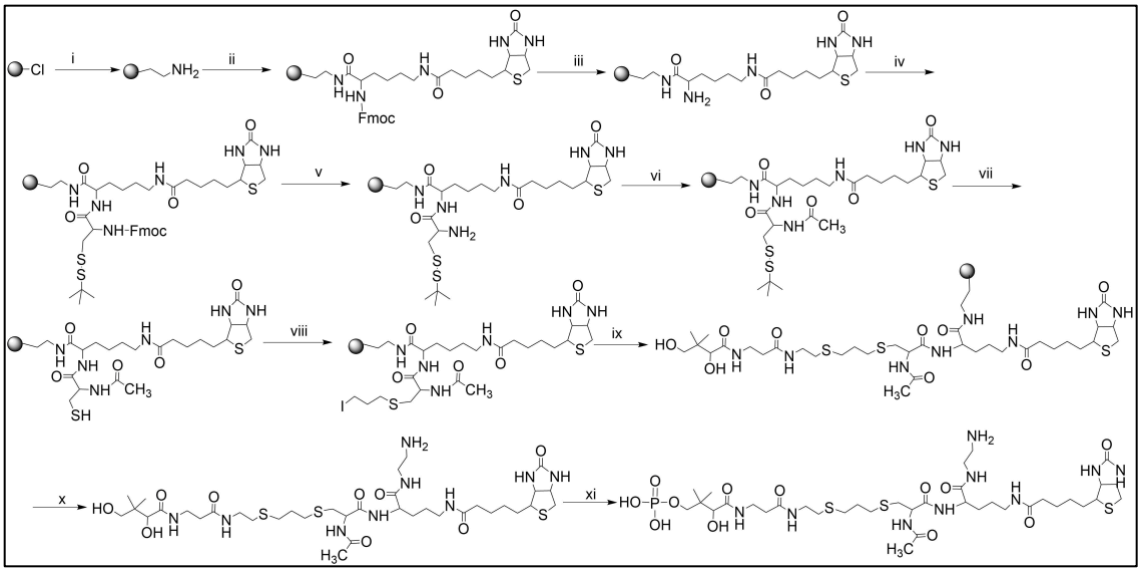


Figure 3.1. Synthesis of Biotin-Lys-14C-phosphopantetheine (BKPP). Phosphopantetheine analog BKPP was synthesized by multistep solid phase synthesis. Reagents and conditions: (i) ethylene diamine, DMF, 10 min, rt. (ii) N^{α}-Fmoc-N^{ε}-biotinyl-L-lysine, HCTU, NMM/DMF, 10-30 min, rt. (iii) and (v) 20% piperidine in DMF, rt, 5 min. (iv) fmoc-cys(stbu)-OH, HCTU NMM/DMF, 10-30 min, rt. (vi) [1-^{14}C] NaOAc, HCTU, DMF, 1 h, rt. (vii) 1 M DTT, 60 °C, 3 h. (viii) 1,3-diiodopropane, DMF, RT, 30 min, rt (ix) pantetheine, DMF, rt, 30 min. (x) 5% TFA/DCM, rt, 10 min. (xi) PanK, ATP, 37 °C. See Supplementary Methods for details of reaction conditions. *Figure and data generated by K. Sapkota*

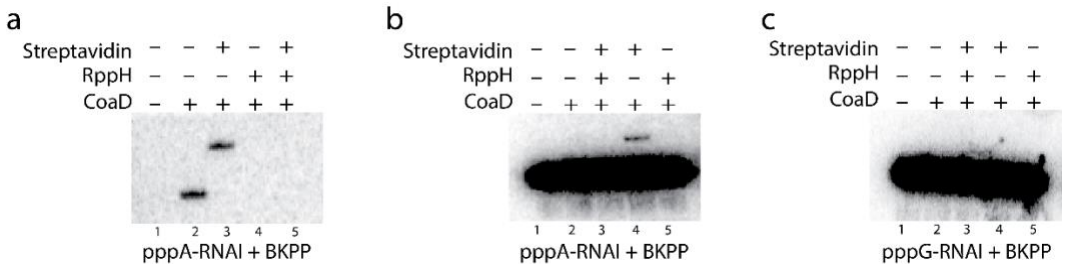


Figure 3.2. PPAT accepts E. coli native RNA as a substrate to yield CoA-RNA post-transcriptionally. RNAI was prepared by *in vitro* transcription under T7 φ2.5 (ATP initiated) and φ6.5 (GTP initiated) promoters. All capping reactions were carried out at 37 °C for 4 h in a buffer containing 20 mM tris pH 7.5, 100 mM NaCl, 5 mM $MgCl_2$, 200 μM BKPP and 500 nM PPAT. Reaction products were resolved by 8% PAGE in denaturing conditions and visualized by phosphorimaging. (a) *In vitro* capping assays using 5 μM RNAI (ATP initiated) as a PPAT substrate. Only the capped-RNAs are visible as phosphopantetheine analog BKPP is labeled with ^{14}C. The mobility of the product RNA was retarded after the addition of streptavidin and the signal disappeared when 5'(p)RNA, generated by RppH hydrolysis of 5' pyrophosphate, was used as a PPAT substrate. (b) *in vitro* capping assays using ^{32}P labeled RNAI (ATP initiated). ATP [α-^{32}P] was used to internally label RNA during *in vitro* transcription. The product is visible only after the addition of streptavidin as the gel could not resolve (ppp)RNAI and BKPP-RNAI. (c) Internally ^{32}P labeled pppG-RNAI was prepared by transcription under T7 φ6.5 promoter.

PPAT mediated BKPP capping was not observed when ATP at +1 position is replaced with GTP. *Figure and data generated by K. Sapkota.*

PPAT accepts 5mer RNA as its substrate and caps with pPant to form genuine CoA-RNA *(data generated by K. Sapkota)*

To investigate post-transcriptional RNA capping by PPAT to yield genuine CoA-RNA, a 5mer RNA of sequence pppAGGAA was prepared by abortive transcription *in vitro* as shown in Figure 3.3a-c, and genuine pPant was synthesized using recombinant PanK/CoaA (Supplemental Figure 3.1). After incubating both substrates with PPAT, both the control and reaction were treated with nuclease P1 to yield the corresponding nucleotide 5' monophosphates and separated by ion pairing reverse phase HPLC. The P1-digested PPAT reaction products produced an extra peak on HPLC having the same retention time as that of authentic dpCoA and an adenosine-like UV signature (Figure 3.3d). When the nuclease P1 treated samples (both the control and the reaction) were analyzed by MALDI-ToF, we observed a peak in the PPAT reaction having m/z of 686.47, corresponding to $[dpCoA-H]^-$ (expected m/z=686.14) (Figure 3.3e).

PPAT transfers both BKPP and pPant to 5' terminus of RNA with similar rates *(data generated by K. Sapkota)*

Having established that PPAT can accept pppA-RNA as a substrate in place of ATP and transfer both BKPP and pPant to RNA 5' terminus, we next compared the kinetics of capping with the synthetic BKPP relative to those of the natural substrate pPant. To determine the apparent turnover number (k_{obs}) of the purified recombinant enzyme,

formation of genuine dpCoA from pPant and ATP was monitored by HPLC and plotted against time. By this approach, k_{obs} was calculated to be ~13 min^{-1}.

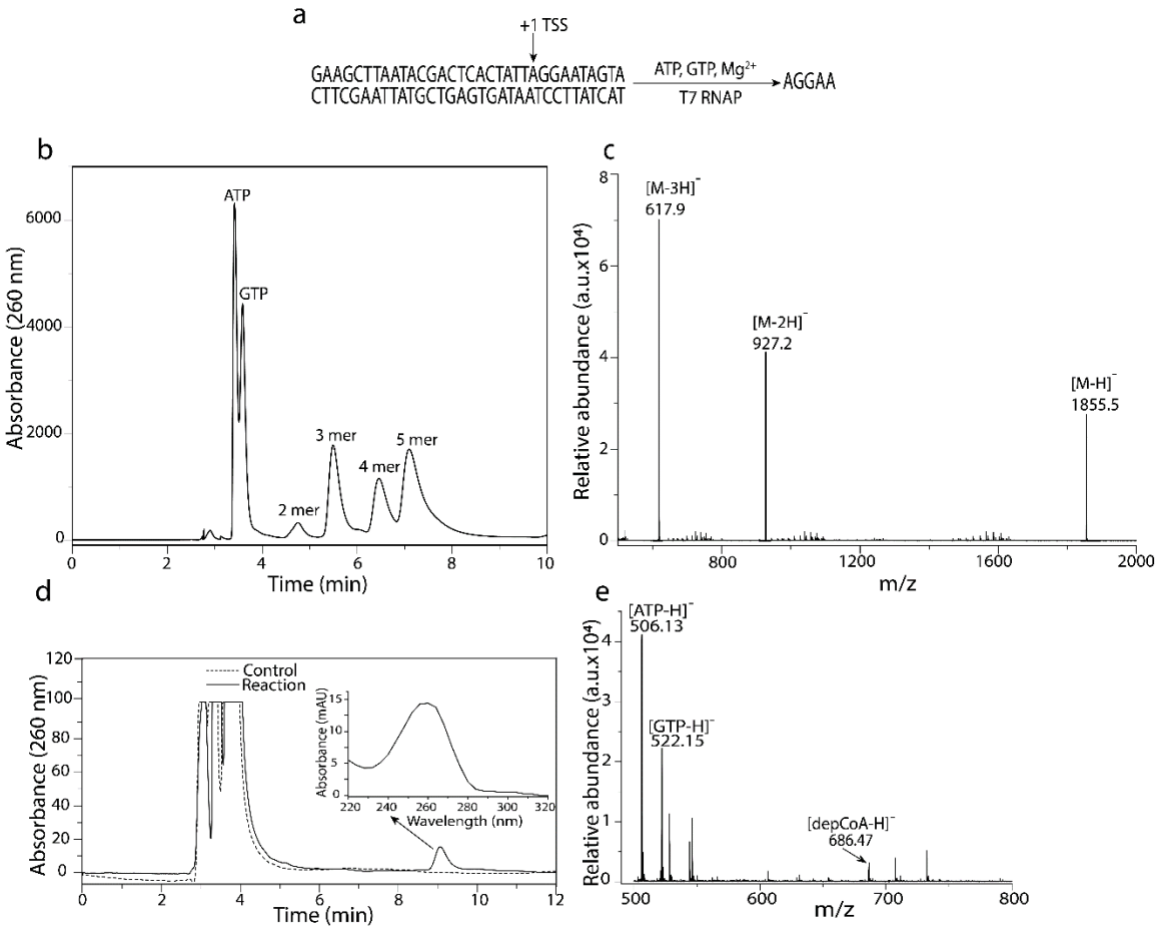


Figure 3.3. In vitro CoA capping of RNA catalyzed by PPAT. (a) Abortive in vitro transcription was used to prepare 5mer RNA. (b) Ion pairing HPLC was used to obtain pure 5mer RNA. HPLC conditions were: 4.6 x 250 mm C18 column (Econosphere), flowrate:1 mL/min, solvents: 90% 0.1 M triethylamine-acetate buffer pH 7.0 and 8% acetonitrile in isocratic conditions. The peaks were lyophilized and (c) characterized by mass spectrometry. (d) HPLC was used to analyze the PPAT-catalyzed capping of 5mer RNA by phosphopantetheine. *In vitro* capping assays were performed under the same conditions as in figure 1. For the control, reaction was quenched immediately (~ 0 min) by

heating at 80 °C and stored at -20 °C until further processing. After 4 h reaction at 37 °C, RNAs were digested with nuclease P1 and separated by ion pairing HPLC using 90% TEA-acetate buffer pH 7.0 and 5% acetonitrile as a mobile phase. HPLC analysis of the reaction showed a peak having the same retention time as dephospho CoA and a characteristic 'adenosine' UV signature. (e) The product was further characterized by mass spectrometry, which showed a peak corresponding to dpCoA. *Figure and data generated by K. Sapkota*.

We next prepared the same small RNA as above with internal radiolabel by abortive *in vitro* transcription in the presence of ATP [α-^{32}P] (Supplemental Figure 3.2) and gel purification. For small substrate RNAs, the substrate (pppRNA) and the product (CoA-RNA) can easily be resolved by denaturing PAGE. Therefore, after incubating substrates with PPAT for the specified times at 37°C, reactions were quenched by freezing in -20°C and separated by 20% denaturing PAGE (Figure 3.4a and 3.4b). Reaction yield was calculated from band intensities, and reaction progress was assessed by plotting the reaction yield versus time (Figure 3.4c and 3.4d). The CoA-RNA product increased approximately linearly for 2 hr and then the rate of product formation decreased. The k_{obs} of the reaction with natural substrate pPant ($2.02x10^{-3}$ min^{-1}) was calculated to be ~4 times faster than with the synthetic pPant analog BKPP ($5.01x10^{-4}$ min^{-1}). Nevertheless, the pPant transfer to pppRNA is ~6000 times slower when compared to the value above for ATP.

The ability of RNA to serve as PPAT substrate is determined by the structure at its 5' terminus *(data generated by K. Sapkota)*

Even though the PPAT-mediated chemistry is identical with pppA (ATP) and pppA-RNA (ATP-RNA, e.g. RNAI and 5mer RNA), we reasoned that access the enzyme's active site may be blocked if a large stem-loop structure is present very close to the 5' terminus. The 5mer RNA used above to study the kinetics of pPant and BKPP capping does not fold into any significant secondary structure. In contrast, 109 mer *E. coli* RNAI has two structures

as predicted by mfold (Supplemental Figure 3.3). Both structures have 3 stem-loops and a stretch of either 4 or 10 unpaired nucleotides at the 5’ terminus.

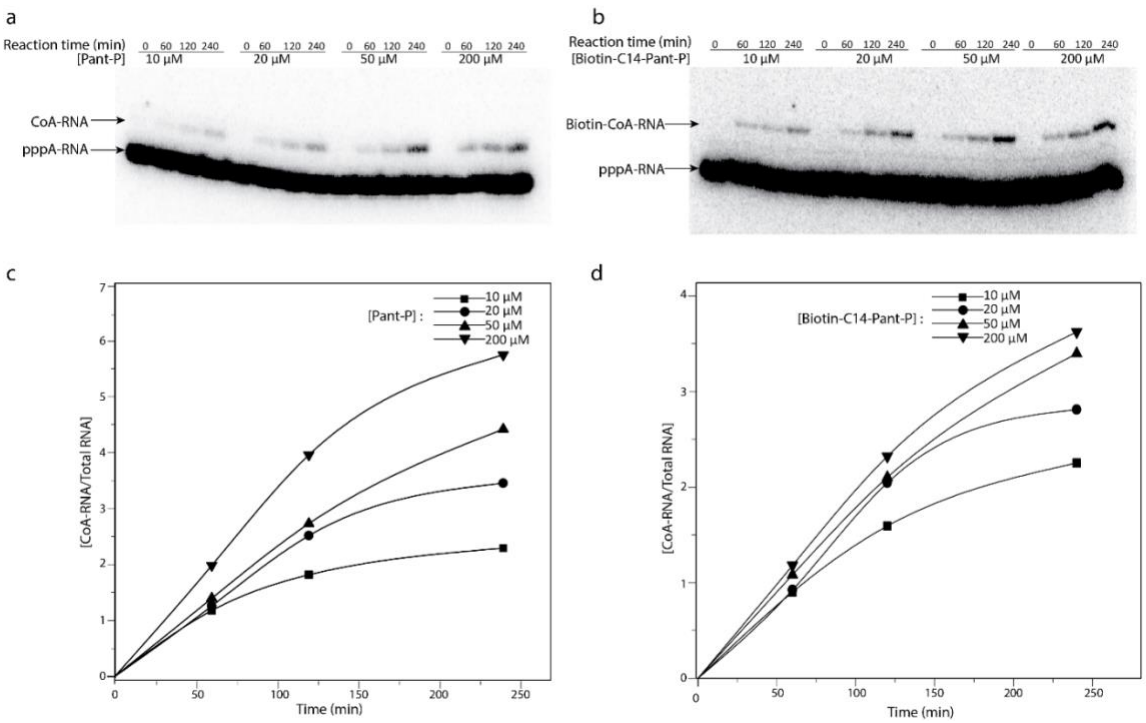


Figure 3.4. Comparison of natural pPant substrate and synthetic BKPP analog for their ability to serve as PPAT substrates. Internally ^{32}P labeled 5mer RNA was prepared by *in vitro* transcription and used to study PPAT kinetics. *In vitro* capping assays were carried out in a PPAT buffer containing 10 µM 5mer RNA, 500 nM PPAT, and 10-200 µM of either (a) pPant or (b) BKPP. Reactions were carried out at 37 °C for the specified time using specified concentrations of pant-p. The reactions were quenched by freezing at -20 °C and separated by 20% denaturing PAGE. Gels were dried, exposed to phosphor screen overnight and visualized after exposing the gel to phosphor screen overnight. Bands were quantified by volume analysis feature of Quantity One software. The ratios of (c) CoA-

RNA to total RNA and (d) biotin-CoA-RNA to total RNA were plotted against reaction time (min). *Figure and data generated by K. Sapkota*.

To study the effect of RNA structure near the 5' terminus on CoA capping, we designed six different 22 nt RNAs that are predicted to have only one thermodynamically favorable secondary structure, with a stable stem-loop at systematically varied distances from the 5' terminus (Figure 3.5a). For example, D2 RNA has a stretch of two unpaired nucleotides before the stem-loop, while D3, D4, D5, D7 and D10 RNAs have stretches of 3, 4, 5, 7, and 10 unpaired nucleotides at the 5' end (Figure 3.5a). To simplify product visualization and analysis, BKPP was used as pPant analog so that only the RNAs that serve as PPAT acceptor substrates are labeled with [^{14}C]. Gel images of reaction products showed that D4, D5, D7 and D10 RNAs, which have ≥ 4 unpaired nucleotides at the 5' terminus, underwent PPAT catalyzed BKPP capping (Figure 3.5b). However, the D2 and D3 RNAs, which have shorter unpaired stretches of 'AG' (2 nt) or 'AGG' (3 nt) at 5' terminus, did not undergo capping (Figure 3.5b and 3.5c). These results established that an RNA needs a stretch of at least 4 unpaired nucleotides at the 5' terminus to undergo PPAT mediated capping to form CoA-RNA.

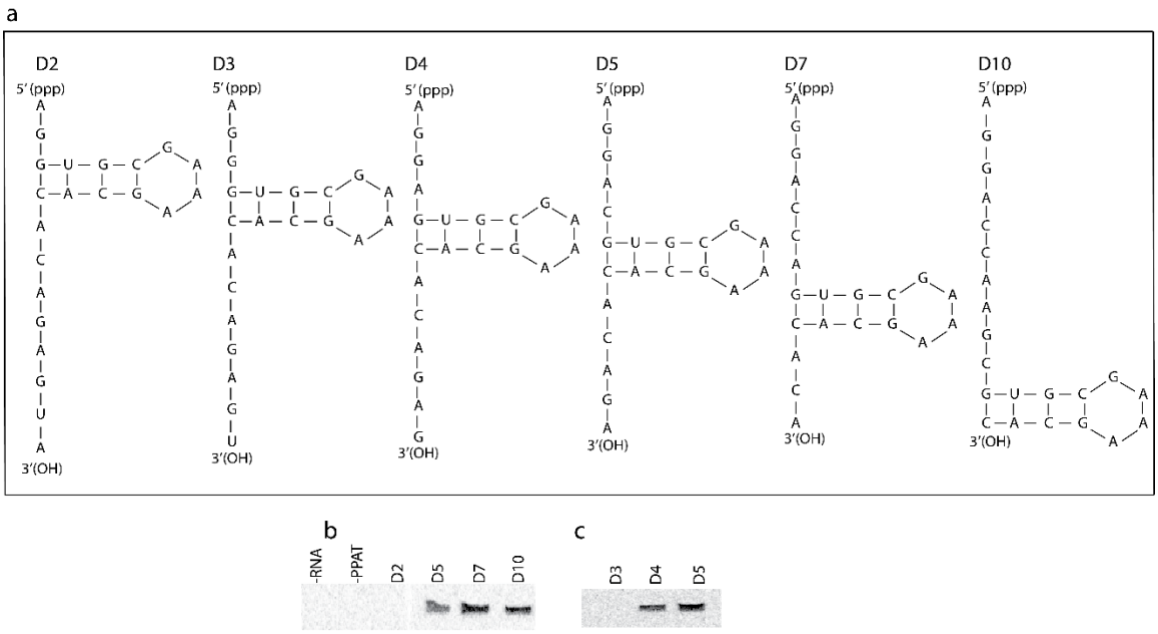


Figure 3.5. Structural requirement at the 5' terminus of substrate RNA. (a) Six RNA sequences (22mer each) were designed and their structures were predicted by mfold (31). All RNAs share a stable stem-loop structure, but the number of non-base paired nucleotides at the 5' terminus varies. All six RNAs were prepared by *in vitro* transcription under T7 φ2.5 promoter. (b) Four RNAs— D2, D5, D7, and D10 — that have 2, 5, 7 and 10 unpaired nucleotides, respectively, at the 5' terminus were tested as PPAT substrates. The reactions were carried out at 37 °C for 4 h in a buffer containing 20 mM tris pH 7.5, 100 mM NaCl, 5 mM $MgCl_2$, 200 μM BKPP, and 500 nM PPAT. The products were visualized by phosphorimaging after separating them in 12% denaturing PAGE. As BKPP is labeled with ^{14}C, only RNAs that served as a PPAT substrate can be visualized. The enzyme did not accept D2 RNA as its substrate while D5, D7, and D10 RNA were accepted. (c) The exact

number of non-structured nucleotides at the 5’ termini for an RNA to undergo PPAT catalyzed CoA capping were determined by using D3, D4 and D5 RNAs. The enzyme recognized D4 and D5 RNAs as substrates but not D3 RNA. Capping reactions were set in a reaction buffer containing 10 μM RNA, 200 μM BKPP and 0.5 μM PPAT, and incubated at 37 °C for 2 h. After removing buffer and salts, the RNAs were resolved by 12% PAGE and exposed to a phosphor screen for a week. *Figure and data generated by K. Sapkota.*

RNA binding affinity to PPAT is not determined by the 5' terminus *(data generated by J. Lucas)*

The BKPP incorporation results above establish that CoA capping by PPAT requires RNAs to have an unstructured 5' terminus and a 5' triphosphate. We wondered whether RNA binding to PPAT occurred primarily in the active site and whether RNAs lacking the proper chemical and structural features were incapable of binding to PPAT. Therefore, we radiolabeled RNA substrates and tested their binding to PPAT using a nitrocellulose filter binding assay. Interestingly, the D7 RNA bound to PPAT only slightly better than D2, especially at the μM PPAT concentration used in the capping reactions, even though D2 does not meet the required 5' terminus structural requirements for the capping reaction (Figure 3.6a). Longer RNA substrates D2* and D7* (102-107 nt) were designed to mimic the key features of D2 and D7 RNAs: +1A, same single-stranded 5' termini with 2 or 7 unpaired 5' nucleotides, followed by stable stem. No significant difference in binding to PPAT was observed between these two RNAs (Figure 3.6b). Finally, the potential importance of the chemical composition of the 5' terminus was evaluated by measuring binding at 1.0 and 3.0 μM PPAT for RNAs that carried either pppG or HO-A in the +1 position. Neither of these modifications reduced RNA binding to PPAT (Figure 3.6c), even though both are incompatible with capping by PPAT. We conclude that overall RNA binding to PPAT does not occur solely in the active site, and that productive interactions that lead to capping are controlled by local positioning effects and not by overall affinity.

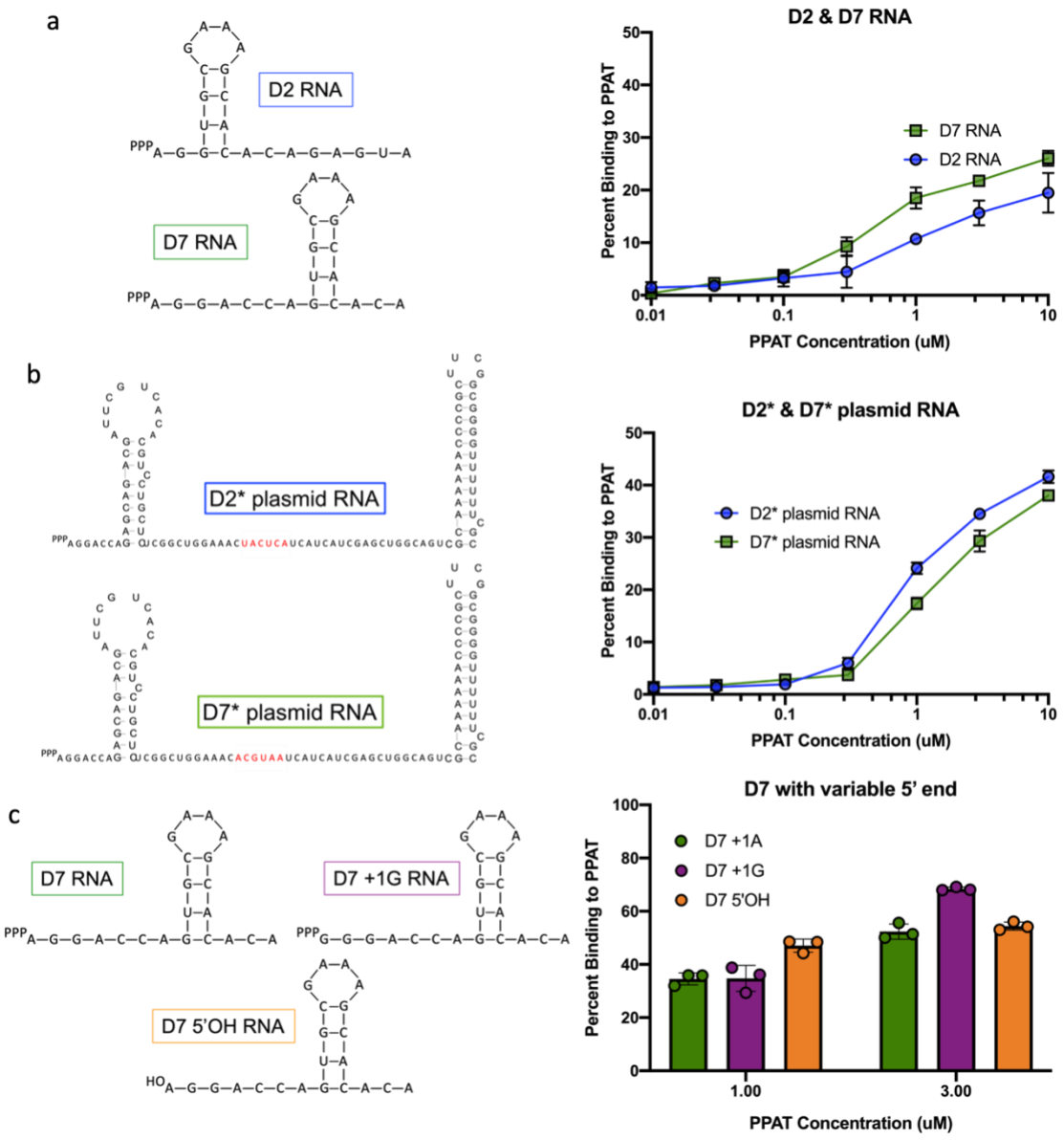


Figure 3.6. RNA binding to PPAT. Comparing binding of different RNAs to purified CoaD/PPAT using nitrocellulose binding assays. Trace amounts of 3' end α^{32}P dCTP radiolabeled RNA were incubated with varying concentrations of CoaD/PPAT at 37°C for 15 min. Predicted secondary structures (mfold) are shown to the left of its corresponding

graphs. N=3 for all binding assays. (a) D2 and D7 RNA binding to PPAT, (b) D2* and D7* RNA binding to PPAT, and (c) D7, D7 +1G, and D7 5’ OH binding to PPAT. *Figure and data generated by J. Lucas.*

The number of unpaired nucleotides at the 5' terminus does not affect capping kinetics *(data generated by K. Sapkota)*

We next examined how capping kinetics are affected by the number of unpaired nucleotides at the 5' terminus of substrate RNA. D4 RNA, which meets the minimum requirements to undergo capping, and the D10 RNA, which has the maximum number of unpaired nucleotides at the 5' terminus among our designed RNAs, were used as PPAT acceptor substrates. Product yields increased with time for both D4 and D10 RNAs, as expected (Figure 3.7a). Band intensities were quantified by comparing with a standard that was spotted onto the gel before drying and exposed for the same period as the sample. Reaction yields for both RNAs were ~1% for 1 h reaction and increased to ~4.5% when incubated for 4 h. When reaction yield (CoA-RNA/total RNA) was plotted against time, both D4 and D10 RNAs were found to have similar reactivities with PPAT (Figure 3.7b), with k_{obs} values of 2.13×10^{-3} min^{-1} and 2.06×10^{-3} min^{-1}, respectively. These data clearly indicate that the rate of pPant transfer is independent of the number of free nucleotides at 5' unpaired region for the range of 4-10 nt. If the minimum requirement of ≥ 4 nucleotides at 5' unpaired region is met, the reaction proceeds with the same speed regardless of the number of nucleotides present.

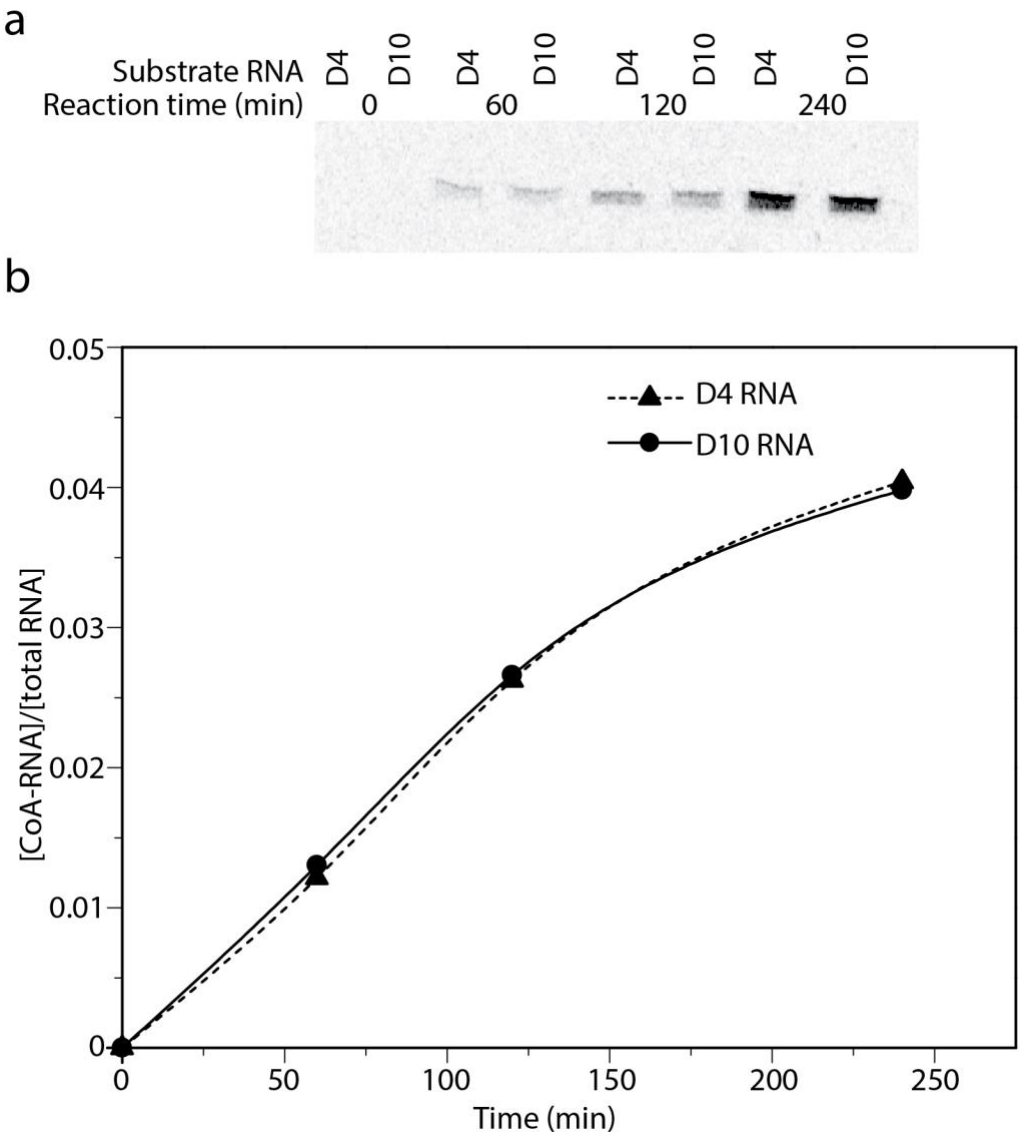


Figure 3.7. Effect of the number of unpaired nucleotides at the 5' terminus of substrate RNA on reaction kinetics. Two representative RNAs – D4 and D10 – having a stretch of 4 and 10 unpaired nucleotides, respectively, at their 5' termini were used to study PPAT kinetics. (a) D4 and D10 RNA at 10 μM concentration were incubated at 37°C with 200 μM BKPP and 500 nM PPAT for the specified times (1-4 hr). After ethanol

precipitation, products were separated by 12% denaturing PAGE and visualized by phosphorimaging. (b) The ratio of CoA-RNA to total RNA (reaction yield) was plotted against incubation time (min) to visualize the reaction progress over time. The product bands were quantified by comparing with a series of standards having known concentration and radioactivity. A linear increase in yield with time was observed for both D4 and D10 RNA with similar slope. *Figure and data generated by K. Sapkota.*

ATP inhibits the transfer of pPant to pppA-RNA by PPAT *(data generated by K. Sapkota)*

Since both ATP and RNA are PPAT substrates with different efficiencies of reactivity (6000-fold k_{obs} difference, from above), we investigated how the presence of ATP affect pPant transfer to pppA-RNA. When the same 5mer pppA-RNA as above was incubated with pPant and PPAT in the presence of varying concentrations (0-20 μM) of ATP, RNA capping with CoA was inhibited in a concentration dependent manner (Figure 3.8a). Reaction yields were plotted against ATP concentration (Figure 3.8b). In the absence of ATP, the reaction proceeded with k_{obs} of 1.9×10^{-3} min^{-1} and 4.5% of substrate RNA was converted to CoA-RNA in 4 hr. When 2 μM ATP was included in the reaction, both k_{obs} and reaction yield dropped by ~two-thirds to 0.64×10^{-3} min^{-1} and 1.5%, respectively. The value of k_{obs} further decreased by ~85, ~90, and ~95% when [ATP] was increased to 5, 10, and 20 μM, respectively. When [ATP] was increased to 1 mM, CoA-RNA product was not detectable. These results imply that post-transcriptionally generation of CoA-RNA would experience significant competition from intracellular ATP, given that the intracellular concentration of ATP in *E. coli* is ≥1 mM.

***In vivo* capping of D2 and D7 RNA substrates by PPAT** *(data generated by J. Lucas)*

To determine whether PPAT was capable of *in vivo* capping, RNA substrates D2* and D7* were cloned into a plasmid for constitutive dual expression under separate promoters along with the *coaD* gene for inducible expression of PPAT. The two RNAs are identical except

for their 5'termini and internal 6 nt index in the middle of the transcript to enable differentiation of the two transcripts (Supplemental Table 3.3). We reasoned that if PPAT acts upon these RNAs within bacterial cells, then it will cap D7* transcripts more than D2*

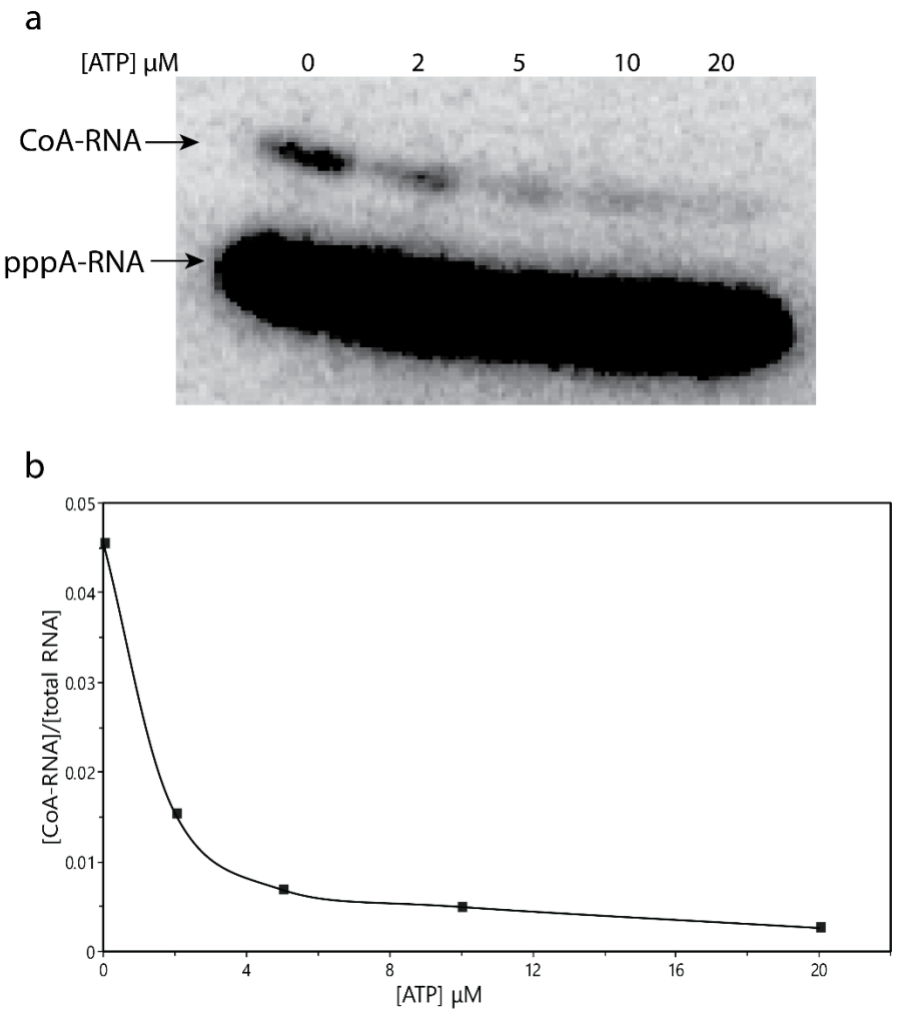


Figure 3.8. Effect of ATP on PPAT catalyzed phosphopantetheine transfer to RNA. ^{32}P labeled 5mer RNA (5 μM) was incubated with 200 μM pPant, 500 nM PPAT and the specified concentration of ATP at 37 °C for 4 h in the standard reaction buffer. The 5mer substrate RNA and the product (CoA-RNA) were separated by 20% denaturing PAGE. The gel was dried, exposed to phosphor screen overnight and visualized by phosphorimaging. (b) Volume analysis tool of Quantity One software was used to quantify the bands

corresponding to CoA-RNA product and the product-to-substrate ratio was plotted against ATP concentration. The PPAT catalyzed phosphopantetheine transfer on RNA was inhibited by ATP and no CoA-RNA product was observed when the reaction contains 10-fold excess of ATP relative to RNA. *Figure and data generated by K. Sapkota.*

transcripts, with the consequence that the ratio of CoA-D7*:CoA-D2* will be enriched for D7* and depleted for D2* relative to the ratio for total RNA.

A 'CoA Capture Seq' method was developed to determine these ratios (Figure 3.9a). Briefly, total RNA was isolated from four cultures of transformed *E. coli* carrying the dual expression plasmid (two biological replicates of induced cultures and two of uninduced cultures). Each of the four samples was split for processing as technical replicates. Contaminating DNA was removed with DNase and each samples was split again for processing two different ways to recover either total RNA (no partition step) or sulfur-containing RNA (partition step on a tri-layer mercury gel, as previously described) (32–35). Following elution from the mercury layer of the gel and ethanol precipitation, both total and sulfur RNAs were reverse transcribed, PCR amplified, and the 16 samples were prepared for high throughput sequencing. Read counts for D7* and D2* RNAs were determined from their respective indices, and these values were converted to ratios. Of the 16 sampled populations, six had insufficient reads and were discarded. Ten populations yielded >2000 processed reads identified as D7* or as D2* (five for total RNA and five for sulfur-partitioned RNA) and were used in the analysis (Supplemental Tables 3.1-3.2).
Clear trends emerged from this dataset. In the total RNA samples, D2* consistently made up around 80% of the read counts whereas D7* comprised around 20% (Figure 3.9b), corresponding to a 1:5 ratio of D7*:D2*. In contrast, for the sulfur-containing RNA samples recovered from APM gels, which should contain the CoA-capped RNAs, D7*

made up ~71% of the read counts (Figure 3.9c), corresponding to a 2.45:1 ratio of CoA-D7*:CoA-D2*. Induction of plasmid PPAT expression did not noticeably impact the amount of D7 plasmid RNA that was capped. These data show more than a 10-fold increase in D7*:D2* RNAs between the sulfur RNA samples and total RNA samples, exactly in the direction predicted based on the suitability of these two RNAs for capping by intracellular PPAT.

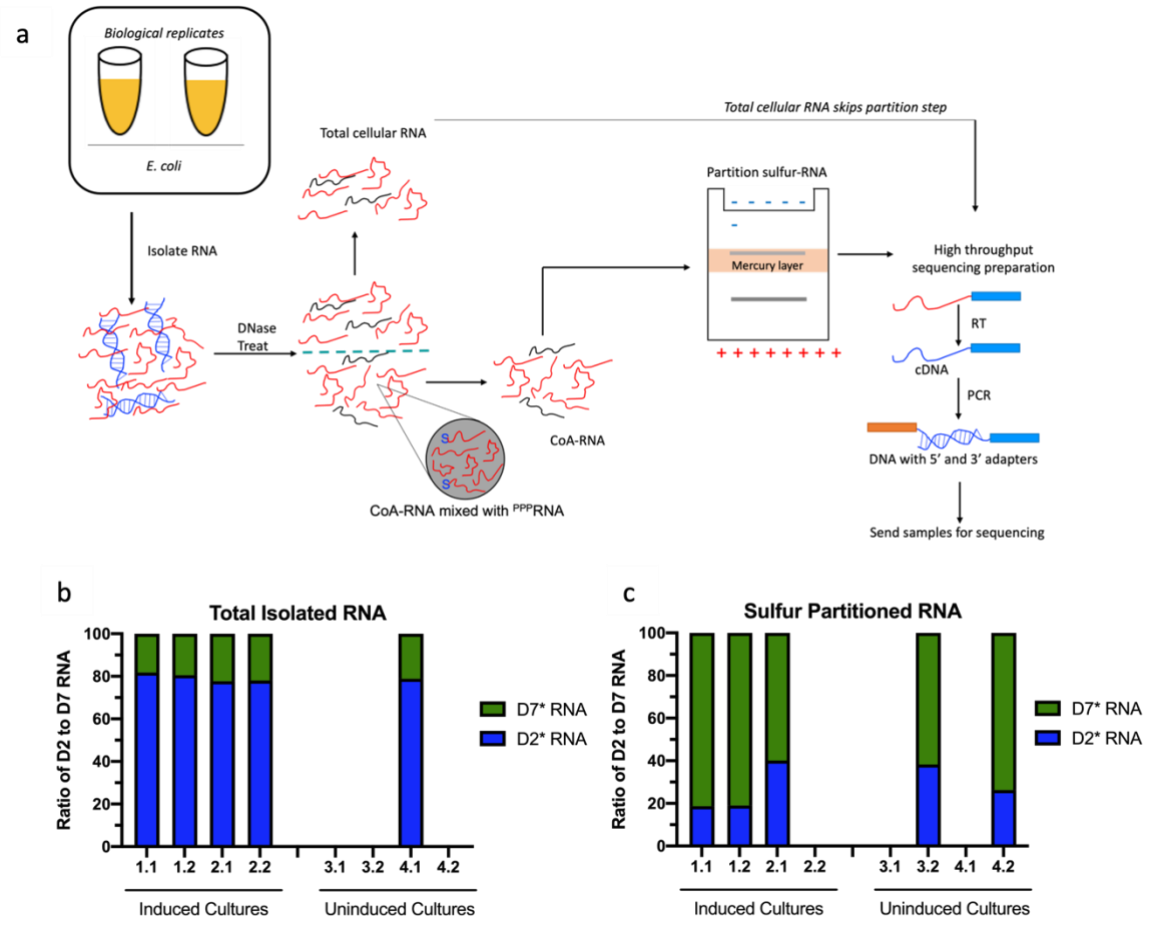


Figure 3.9. Capturing cellular RNAs capped in vivo by PPAT. (A) CoA Capture Seq Method schematic summarizing the various steps to isolate, prepare, partition, and sequence CoA-RNAs from total RNA. RNA was isolated from bacterial cultures (two cultures grown separately as biological replicates) and then separated into two technical replicates prior to DNase treatment. Sulfur samples were then partitioned on an APM gel. Following purification from the APM gel, samples undergo RT-PCR in preparation for high-throughput sequencing. D2* (blue) and D7* (green) were identified by their

respective index sequences and counted for (B) total RNA and (C) CoA-RNA partitioned samples. The ratio of D2 to D7 was then calculated for each sample. Any samples with fewer than 2,000 total unique processed reads were excluded from these data sets. *Figure and data generated by J. Lucas.*

DISCUSSION

By using *in vitro* transcribed RNAs of diverse sizes (5mer to 109mer), we have demonstrated that RNA can be capped post-transcriptionally as CoA-RNA by the activity of PPAT. The substrate RNA requires three distinct features at the 5' terminus: a triphosphate group on the terminal nucleotide, adenosine in the +1 position, and a stretch of four or more unpaired nucleotides. These findings represent a possible novel mode of post-transcriptional RNA capping as demonstrated by our CoA-CaptureSeq data (Figure 3.9). Although PPAT was known to accept a broad range of modifications to the pantetheine/ pantothenate substrate (24, 25, 36), our results demonstrated its ability to take modifications on the ATP substrate – albeit at substantially reduced efficiency – since pppA-RNA can be considered as an ATP analog that carries a bulky 3' modification in the form of a long stretch of nucleotides. This atypical activity of PPAT generated CoA-RNA *in vitro*, and possibly contributes to CoA-RNA biogenesis *in vivo*. In principle, similar PPAT-catalyzed reactions could occur within bacteria under certain circumstances for RNAs that meet the three requirements mentioned above, such as *E. coli* RNAI.

The activity of PPAT is regulated by free CoA though feedback inhibition. To achieve this regulation, the enzyme binds CoA and dpCoA using distinct binding modes but at overlapping sites. The pantetheine arm and the adenosine group of CoA bound to the enzyme are oriented differently than those of bound dpCoA (Supplemental Figure 3.4) (37–39). This difference probably prevents the enzyme's activity on CoA, as PPAT was previously reported to be unable to catalyze pyrophosphorolysis of CoA (CoA + ppi →

CoA + ppi pant-p + pppAp) (40). However, CoA was reported to bind differently to two different trimers of PPAT hexamer (39), suggesting that it may utilize at least two distinct binding modes to achieve feedback inhibition. In the case of a bound pppA-RNA chain, the additional nucleotides on the 3' phosphate may force the terminal pppA into a different orientation to produce the weaker activity on RNA relative to ATP. It is tempting to speculate that a stretch of four or more unpaired nucleotides may provide the necessary flexibility to correctly position an RNA and that once the +1A of RNA gets into the enzyme's active site, the chemistry is the same as that of ATP. Interestingly, the binding data indicated no obvious difference in RNA binding related to the structure or chemical composition of the 5' terminus (Figure 3.6). Therefore, PPAT capping requirements may be a reflection of substrate orientation and chemical reactivity in the active site, but not necessarily due to overall RNA substrate binding affinity. Ultimately, detailed structural and mechanistic studies are needed to fully uncover the molecular features of enzyme-RNA interaction.

PPAT-mediated formation of CoA-RNA in cells could be limited by intracellular ATP, based on three observations. First, ATP strongly inhibited the activity of PPAT on RNA to form CoA-RNA *in vitro*. For example, capping activity was reduced by ~95% when ATP was present in only two-fold higher concentration than the RNA. Second, PPAT acts upon ATP as the acceptor substrate ~6000 fold faster than it does with pppA-RNA as the acceptor substrate. Third, the ATP reaction quickly forms dpCoA, which not only acts as

a feedback inhibitor but also reduces the concentration of available pPant. However, our CoA-RNA CaptureSeq data from *in vivo* experiments revealed a >10-fold change in D7*/D2* ratios in the sulfur-containing RNA samples relative to the ratio in total RNA samples, in line with the expectation that D7* RNA would be preferentially capped with CoA by PPAT in cells. Because D2* and D7* RNAs were expressed from strong constitutive promoters, they may accumulate at high levels that aided competition with intracellular ATP and allowed for detectable capping. While some of the mechanistic details remain to be resolved, our *in vitro* and *in vivo* data indicate a possible role for PPAT in post-transcriptional CoA capping in bacterial cells.

Although the biological significance of CoA-RNA is not yet clear, at least three speculative models suggest themselves. One possibility is that the CoA-RNA might function as a sensor for a cell's energy state. Specifically, growth conditions that support abundant ATP are expected to inhibit PPAT activity on RNA, preventing CoA-RNA formation, while low-ATP conditions could favor CoA-RNA. The second possibility arises from reports of thioester forms such as acetyl CoA-RNA, succinyl CoA-RNA, and malonyl CoA-RNA (1). These CoA-thioesters are high-energy intermediates in acyl transfer reactions and their presence at the RNA 5' end could direct the RNA's reactivity. Unlike CoA-RNA which can be generated either co-transcriptionally by RNAP or post-transcriptionally by PPAT, thioester-CoA-RNA could only be synthesized post-transcriptionally by thioesterification of CoA-RNA since thioester-dpCoA that can possibly act as transcription initiators are not

known to present in *E. coli*. It is intriguing to speculate that CoA-RNA might be a substrate for one or more acyl CoA synthetases to drive post-transcriptional acylation of CoA-RNA into thioester-CoA-RNA. Finally, CoA-RNA might represent a molecular fossil from an RNA world in which RNA served in both genetic and catalytic capacities before the emergence of the contemporary ribonucleoprotein (RNP) world (41). Chemically reactive moieties such as the thiol in CoA, nicotinamide in NAD^+, and isoallaxazine in FAD – can expand the catalytic repertoire of RNA. For example, we recently described a flavin-binding RNA aptamer that shifts the reduction potential of the bound cofactor, dramatically enhancing its intrinsic reactivity relative to that of free flavin (42). Bound organic cofactors may have similarly aided ribozyme catalytic diversity during an RNA world, especially when attached covalently. Indeed, CoA-RNA transcripts have been used during *in vitro* selections to isolate ribozymes that promote thioester formation and aminoacylation (43, 44). Other ribozymes have been isolated that promote self-capping with pPant, FMN, and NMN to form CoA-RNA, FAD-RNA, and NAD-RNA conjugates (45). Similar coupling reactions might have occurred during an RNA world to provide RNA an extra layer of reactivity. With respect to the biology of CoA-RNA conjugates, it remains to be seen what mechanisms drive their production and what roles they may play in extant, engineered, or emergent biological forms.

MATERIALS AND METHODS

The detailed experimental procedure for BKPP synthesis is described in supplementary methods.

Expression and purification of enzymes. Recombinant version of two CoA biosynthetic enzymes, CoaA and PPAT, were expressed and purified according to our published protocol (25). Briefly, the plasmids encoding *coaA* and *PPAT* genes (addgene # 50386 & 50388) (36) were transformed separately into Bl21(DE3) strain of *E. coli*. For each protein, a single colony was isolated and grown in LB/Kan media until OD600 reached 0.6. Protein expression was induced by adding 500 μM IPTG at 37°C for 4 hr. Cells were harvested, lysed by sonication (20s on, 40s rest on ice, 5 cycles), and centrifuged at 40,000xg to clear the lysate. The supernatant was loaded into a Ni-NTA resin preequilibrated with 50 mM tris, 300 mM NaCl and 10 mM Imidazole pH 8.0 and washed extensively, and the bound protein was eluted by the same buffer containing 200 mM imidazole. Membrane filters of 10,000 Da cut-off were used to remove imidazole, concentrations were estimated by UV spectroscopy, and proteins stored in -20°C until further use.

RNA transcripts. DNA templates (Supplemental Table 3.3) were ordered from Integrated DNA Technologies (IDT). Each RNA was transcribed *in vitro* by mixing 100 pmol of top strand oligo and 100 pmol of bottom strand oligo, using the Y639F T7 RNA polymerase (46), *in vitro* transcription buffer (1x = 50 mM Tris-HCl pH 7.5, 15 mM $MgCl_2$, 5 mM

DTT, and 2 mM spermidine), and 2 mM each of ATP, UTP, GTP, CTP. Transcription reactions were incubated at 37°C overnight (approximately 16 hrs) and terminated by the addition of denaturing gel loading dye (90% formamide, 50 mM EDTA and 0.01% of xylene cyanol and bromophenol blue). Transcripts were subsequently purified by denaturing polyacrylamide gel electrophoresis (5-8% TBE-PAGE, 8 M urea). Transcriptions were also carried out by using high yield transcription kit (Epicentre) following manufacturer's protocol. Bands corresponding to the expected product sizes were visualized by UV shadow, excised from the gel, and eluted by tumbling overnight at 4°C in 300 mM sodium acetate pH 5.4. Eluates were ethanol precipitated, resuspended in nuclease-free water, and stored at -20°C until further use. A NanoDropOne spectrophotometer (Thermo Fisher Scientific) was used to determine specific RNA concentrations for all assays.

Preparation of RNAI. A template DNA for RNAI transcription was prepared by PCR. Gene specific primers were used to amplify the 109 bp region of RNAI from *EcPPAT* plasmid. A dinucleotide 'AG' (for φ2.5 promoter) or 'GG' (for φ6.5 promoter) were added during PCR to meet the requirements of T7 transcription. The PCR reaction was concentrated to 10x by Zymo DNA clean and concentrator kit following manufacturer's protocol. Transcription was carried out at 37°C for 3 hr. A representative 20 µL transcription contained 2 µL 10x buffer, 2 µL 100 mM DTT, 6 µL NTP mix (25 mM each), 0.5 µL RNase inhibitor, 2 µL 10x template DNA and 2 µL T7 RNA polymerase. RNA was

purified by Zymo RNA clean and concentrator kit following manufacturer's protocol, quantified by UV and stored in -20°C until further use.

Preparation of 5mer RNA. An abortive *in vitro* transcription with modifications was used to prepare 5mer RNA. A dsDNA template was prepared by annealing a DNA oligo containing T7 φ2.5 promoter and an appropriate sequence to encode 10mer RNA with its complementary oligo (Figure 3.3a). *In vitro* transcription reaction was set to contain 10 μM template DNA and 25 mM ATP and GTP each (UTP and CTP were omitted to prevent run-off synthesis), in addition to the common components mentioned above. 1 μL [α-^{32}P]ATP (Perkin Elmer) was included to radiolabel RNA internally. Transcriptions produced a mixture of 2mer, 3mer, 4mer and 5mer RNAs, from which the 5mer was either gel purified from 20% denaturing PAGE (for radiolabeled 5mer) or purified by ion-pairing reverse phase HPLC (for non-radiolabeled 5mer) (Figure 3.3b), lyophilized, and stored at -20°C until further use. MALDI-ToF in negative ion mode confirmed the identity of the purified 5mer RNA, which showed a peak having m/z of 1855.5 (expected 1855), along with two other peaks having m/z of 927.2 and 617.9, corresponding to the 5mer RNA with -2 and -3 charges, respectively (Figure 3.3c). Additional details in supplementary methods.

Preparation of 22mer RNA. The dsDNA templates were prepared by annealing corresponding DNA oligos of designed sequences (D2, D3, D4, D5, D7 and D10) with their complementary oligos. Transcription was carried out at 37°C for 3 hr by as above.

RNA precipitation was carried out at -20°C for 1 hr by adding NaOAc (0.3 M final) and 3 vol. of EtOH. The pellet was resuspended in nuclease free water, quantified by UV and stored in -20°C until further use.

Synthesis of pPant. Recombinant PanK/CoaA was used to synthesize pPant. We first synthesized ox-pPant by phosphorylating pantethine in a PanK/CoaA catalyzed reaction and purifying the disulfide-linked product by reverse phase HPLC. Reduction of ox-pPant by TCEP yielded pPant in high purity.

***In vitro* PPAT assay.** All PPAT reactions were performed at 37°C in a reaction buffer (20 mM Tris, pH 7.5, 100 mM NaCl, 10 mM KCl, 1 mM $MgCl_2$). Substrate concentrations and reaction times were varied: typically, 5-10 μM RNA, 0-200 μM pPant/BKPP, 0-1 μM PPAT. Reactions were loaded directly onto the urea-PAGE (12%-20%), run for 1hr-4hr at 15 W, dried, and exposed to phosphor screen for 1-7 days. Bands were visualized by phosphorimager and quantified by using volume analysis tool of Quantity One software (Bio-Rad).

3' end radiolabeling RNA transcripts. To radiolabel the 3' end of RNA transcripts a 100 pmol of RNA was mixed with 100 pmol of reverse primer (Supplemental Table 3.3), 1 μL of 250 μCi of α^{32}P dCTP (Perkin Elmer), 1X isothermal buffer (New England Biolabs), 6 mM $MgSO_4$, 1.6 mM dNTP mix, and 16 U BST enzyme in a 25 μL reaction volume. BST

reactions incubate at 65°C for 1 hr and are heat inactivated at 80°C for 20 min. Radiolabeled transcripts were subsequently purified by ethanol precipitation and stored at -20°C until further use.

Nitrocellulose filter binding assays. Radiolabeled RNA was counted using a liquid scintillation counter to measure DPM. The RNA used in each binding assay sample was 30,000 DPM or higher. RNA was incubated with no PPAT to determine background binding of RNA to the nitrocellulose filter and an unfiltered 'no wash' sample was measured to determine the total amount of radioactivity present in each binding assay reaction. To decrease non-specific nucleic acid binding to the nitrocellulose filters, prior to use filters are incubated in 0.5 M KOH for 20 min, washed with MilliQ water, and incubated in 1X binding buffer for 45 min (47). Trace amounts of radiolabeled and refolded RNA was incubated with varying concentrations of PPAT in 1X binding buffer (20 mM Tris pH 7.5, 100 mM NaCl, 10 mM KCl, 1 mM $MgCl_2$) at 37°C for 15 minutes. RNA:PPAT complexes were partitioned from unbound RNA by filtering the samples through a pre-wet, KOH-treated nitrocellulose filter under vacuum and immediately washing with 1 mL of 1X binding buffer. Three replicates were performed for each binding assay. Radioactive RNA bound to PPAT on the filters was counted by placing the nitrocellulose filters into scintillation vials, adding 3 mL Emulsifier-safe liquid scintillation fluid (Perkin Elmer, Waltham, MA), and measuring DPM using a liquid scintillation counter.

RNA expression plasmids. Plasmid pJKL1 (VectorBuilder, Chicago, Illinois) was designed for inducible expression of CoaD/PPAT and constitutive expression of D2* and D7* RNA transcripts (Figure 3.6b and Supplemental Figure 3.5). Plasmid sequence was confirmed by Sanger Sequencing (University of Missouri DNA Core Facility). Purified plasmid was transformed into BL21(DE3)pLysS chemically competent cells and colonies were grown on Ampicillin (50 μg/mL) agar plates for 16 hr in a 37°C incubator shaking at 250 rpm. Single colonies were inoculated into 5 mL of 2XYT Ampicillin media and grown at 37°C for 16 hr. 1 mL overnight culture was added to 4 different 50 mL Falcon Tubes containing 25 mL of 2XYT Ampicillin media (two biological replicate each for 'induced' and 'uninduced' cultures). These were incubated in a 37°C incubator with shaking at 250 rpm. After the cultures reached an OD_{600} of 0.6, IPTG was added to a final concentration of 10 mM to two of the four cultures to induce PPAT expression and incubated at 37°C for an additional hour. The other two cultures remained uninduced but were incubated at 37°C for the same amount of time. OD_{600} was measured hourly and all cultures were put on ice when OD_{600} reached 1.45-1.6.

RNA isolation from bacteria. Using OD_{600} values, approximately equivalent numbers of bacterial cells were harvested by centrifugation at 4,000 xg at 4°C for 15 minutes. Pellets was resuspended in 5mL lysozyme buffer (50 mM Tris HCl pH 7.6, 250 mM NaCl, 0.1 mM EDT). Lysozyme (ThermoFisher Scientific) was added to a final concentration of 0.2

mg/mL and allowed to incubate for 10 min at room temperature. 3 volumes of TRIzol (ThermoFisher Scientific) was added to the total cell lysate and vortexed before incubating on ice for 5 min. Cold (4°C) chloroform (ThermoFisher Scientific, 1/5 of the total TRIzol volume) was added and the sample was briefly vortexed to mix. Samples were centrifuged at 12,000 xg at 4°C for 15 min. The top aqueous layer (~10 mL) was carefully removed to a fresh 50 mL falcon tube and 2 volumes of cold chloroform were added to remove any leftover phenol. The samples were vortexed then centrifuged at 12,000 xg at 4°C for 15 min. The top aqueous layer was removed and an equal volume of cold isopropanol (ThermoFisher Scientific) was added and the sample was vortexed. Samples were incubated on ice for 5 min before being centrifuged at 16,000 xg at 4°C for 30 min. The supernatant was discarded and the pellet was washed with cold 70% ethanol and centrifuged at 16,000 xg at 4°C for 30 min. Pellets containing the RNA were dried and resuspended in 1 mL of MilliQ water and stored at -20°C until further use.

TURBO DNase treatment. Samples of isolated RNA with nucleic acid concentrations greater than 20 μg/mL were diluted to 10 μg/50 μL. 5 μL of sample were set aside after 0, 1, and 2 DNase treatments to be used for PCR as template (without reverse transcription) to evaluate carryover DNA contamination. TURBO DNase (ThermoFisher Scientific) reactions with up to 10 μg of nucleic acid were assembled in 1X TURBO DNase buffer with 2 U of TURBO DNase and incubated at 37°C for 30 minutes. To stop the reaction, 5 μL of TURBO DNase Inactivation Reagent (ThermoFisher Scientific) was added and

incubated at room temperature for 5 min with intermittent mixing to keep the reagent in solution. Samples were spun on a tabletop centrifuge for ~90 sec to separate the DNase treated samples (supernatant) and inactivation reagent. The supernatant was collected and concentrated by ethanol precipitation. PCR to test the effectiveness of TURBO DNase treatments used 16S RNA primers.

APM gels & elution. [(N-Acryloylamino) Phenyl] Mercuric Chloride (APM) stock solution was made as previously described (33). To pour a tri-layer APM gel, the first layer of polyacrylamide (about 20 mL) was poured in a gel casing standing upright. 1 mL of MilliQ water was added directly after pouring the bottom layer to create a smooth interface between layers. After allowing the first layer to polymerize for approximately 30 min, water was removed and the second, APM-containing layer was added, consisting of 1 mL of polyacrylamide, 1 μL TEMED, 10 μL 10% ammonium persulfate (APS), and 200 μL of APM stock solution. The APM layer was covered by a fresh 1 mL layer of water and allowed to polymerize for approximately 30 min. Excess water was then removed and the final 10 mL layer of polyacrylamide was added along with the comb for generating wells, and allowed to polymerize for 30 min.

RNA samples were loaded onto the APM gel and run in 1X TBE at 30 watts. CoA-RNAs were visualized by UV-shadow at the top APM interface and excised from the gel. Gel pieces were 'minced' into very small pieces and added to 1X APM elution buffer (0.5 M

ammonium acetate, 0.5 M DTT, 10 mM EDTA) and tumbled overnight at 4°C. The gel slurry was then loaded into a pre-wetted 100,000 Da molecular weight cutoff filter (ThermoFisher Scientific) and spun at 14,000 xg for 15 min to separate eluted RNA (eluate) from the gel pieces (retentate). The columns were washed with 200 μL of 1X APM elution buffer and spun at 14,000 xg for an additional 15 min. Flow-through was collected into a fresh tube. 1 μL of 10 μg/μL glycogen and 1 mL of cold ethanol were added precipitate the CoA-RNA. RNA was resuspended in nuclease-free water and stored at -20°C until further use.

Illumina sequencing and data analysis.

The purified RNA was reverse transcribed using BST 3.0 DNA Polymerase (New England Biolabs). Reactions included a specific reverse primer for both D2* and D7* RNA templates, 1X isothermal buffer (New England Biolabs), 6 mM $MgSO_4$, 1.6 mM dNTP mix, and 16 U BST enzyme in a 25 μL reaction volume and incubated at 72°C for 1 hr. The primers for reverse transcription and PCR anneal to the same sequences on D2* and D7* (Supplemental Table 3.3). The RT and PCR steps were used to append Illumina adapters and sequencing indices for multiplexing of the libraries as previously described (48). Primers used to append the Illumina adapters and sequencing indices can be found in Supplemental Table 3.3. Sequencing was performed on an Illumina NextSeq 500 (University of Missouri Genomics Technology Core). Although paired-end reads generated Read 1 and Read 2 for each selection round, a single 300 nt read provided enough

coverage such that no additional information was gained through read pairing. Populations were demultiplexed, and the relevant sequence information was found and used from Read 1 (5′ HTS primer binding sequence, 6 nt index, 3′ HTS primer binding sequence), and all data shown represents reads from Read 1 only. Data preprocessing was performed using cutadapt (49) to trim 5′ and 3′ PBS and to discard any uncut sequences or sequences with lengths not within ± 3 nt of the expected size (26 nt) after trimming (Supplemental Tables 3.1-3.2). These populations were then analyzed using FASTAptameR 2.0 (50, 51) to count and normalize reads (FASTAptameR-Count) and to find the 6 nt index motifs (FASTAptameR-Motif Search) (found in Supplemental Tables 3.1-3.2) for all samples to determine counts for D2 and D7.

ACKNOWLEDGEMENTS

We thank Mr. Jarrett Faulkner for the insightful discussion. This work is supported by development fund from the University of Southern Mississippi (FH) and by NASA Interdisciplinary Consortium for Astrobiology Research (ICAR) grant 80NSSC21K0596 (DHB).

REFERENCES

1. Kowtoniuk, W. E., Shen, Y., Heemstra, J. M., Agarwal, I., and Liu, D. R. (2009) A chemical screen for biological small molecule-RNA conjugates reveals CoA-linked RNA. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7768–7773

2. Chen, Y. G., Kowtoniuk, W. E., Agarwal, I., Shen, Y., and Liu, D. R. (2009) LC/MS analysis of cellular RNA reveals NAD-linked RNA. *Nat. Chem. Biol.* **5**, 879–881

3. Zhang, H., Zhong, H., Zhang, S., Shao, X., Ni, M., Cai, Z., Chen, X., and Xia, Y. (2019) NAD tagSeq reveals that NAD+-capped RNAs are mostly produced from a large number of protein-coding genes in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12072–12077

4. Wang, Y., Li, S., Zhao, Y., You, C., Le, B., Gong, Z., Mo, B., Xia, Y., and Chen, X. (2019) NAD+-capped RNAs are widespread in the Arabidopsis transcriptome and can probably be translated. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12094–12102

5. Grudzien-Nogalska, E., Bird, J. G., Nickels, B. E., and Kiledjian, M. (2018) "NAD-capQ" detection and quantitation of NAD caps. *RNA*. **24**, 1418–1425

6. Jiao, X., Doamekpor, S. K., Bird, J. G., Nickels, B. E., Tong, L., Hart, R. P., and Kiledjian, M. (2017) 5′ End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding. *Cell*. **168**, 1015-1027.e10

7. Walters, R. W., Matheny, T., Mizoue, L. S., Rao, B. S., Muhlrad, D., and Parker, R. (2017) Identification of NAD+ capped mRNAs in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 480–485

8. Barvík, I., Rejman, D., Panova, N., Šanderová, H., and Krásný, L. (2017) Non-canonical transcription initiation: The expanding universe of transcription

initiating substrates. *FEMS Microbiol. Rev.* **41**, 131–138

9. Frindert, J., Zhang, Y., Nübel, G., Kahloon, M., Kolmar, L., Hotz-Wagenblatt, A., Burhenne, J., Haefeli, W. E., and Jäschke, A. (2018) Identification, Biosynthesis, and Decapping of NAD-Capped RNAs in B. subtilis. *Cell Rep.* **24**, 1890-1901.e8
10. Cahová, H., Winz, M. L., Höfer, K., Nübel, G., and Jäschke, A. (2015) NAD captureSeq indicates NAD as a bacterial cap for a subset of regulatory RNAs. *Nature*. **519**, 374–377
11. Bird, J. G., Basu, U., Kuster, D., Ramachandran, A., Grudzien-Nogalska, E., Towheed, A., Wallace, D. C., Kiledjian, M., Temiakov, D., Patel, S. S., Ebright, R. H., and Nickels, B. E. (2018) Highly efficient 5' capping of mitochondrial RNA with nad + and NADH by yeast and human mitochondrial RNA polymerase. *Elife*. 10.7554/eLife.42179
12. Bird, J. G., Zhang, Y., Tian, Y., Panova, N., Barvík, I., Greene, L., Liu, M., Buckley, B., Krásný, L., Lee, J. K., Kaplan, C. D., Ebright, R. H., and Nickels, B. E. (2016) The mechanism of RNA 5′ capping with NAD+, NADH and desphospho-CoA. *Nature*. **535**, 444–447
13. Julius, C., and Yuzenkova, Y. (2019) Noncanonical RNA-capping: Discovery, mechanism, and physiological role debate. *Wiley Interdiscip. Rev. RNA*. **10**, e1512
14. Huang, F. (2003) Efficient incorporation of CoA, NAD and FAD into RNA by in vitro transcription. *Nucleic Acids Res.* 10.1093/nar/gng008
15. Julius, C., Riaz-Bradley, A., and Yuzenkova, Y. (2018) RNA capping by

mitochondrial and multi-subunit RNA polymerases. *Transcription*. **9**, 292–297

16. Li, N., Yu, C., and Huang, F. (2005) Novel cyanine-AMP conjugates for efficient 5′ RNA fluorescent labeling by one-step transcription and replacement of [γ-32P]ATP in RNA structural investigation. *Nucleic Acids Res.* **33**, 1–8

17. Huang, F., Wang, G., Coleman, T., and Li, N. (2003) Synthesis of adenosine derivatives as transcription initiators and preparation of 5′ fluorescein- and biotin-labeled RNA through one-step in vitro transcription. *RNA*. **9**, 1562–1570

18. Winz, M. L., Cahová, H., Nübel, G., Frindert, J., Höfer, K., and Jäschke, A. (2017) Capture and sequencing of NAD-capped RNA sequences with NAD captureSeq. *Nat. Protoc.* **12**, 122–149

19. Vvedenskaya, I. O., Bird, J. G., Zhang, Y., Zhang, Y., Jiao, X., Barvík, I., Krásný, L., Kiledjian, M., Taylor, D. M., Ebright, R. H., and Nickels, B. E. (2018) CapZyme-Seq Comprehensively Defines Promoter-Sequence Determinants for RNA 5′ Capping with NAD+. *Mol. Cell*. **70**, 553-564.e9

20. Bird, J. G., Zhang, Y., Tian, Y., Panova, N., Barvík, I., Greene, L., Liu, M., Buckley, B., Krásný, L., Lee, J. K., Kaplan, C. D., Ebright, R. H., and Nickels, B. E. (2016) The mechanism of RNA 5′ capping with NAD+, NADH and desphospho-CoA. *Nature*. **535**, 444–447

21. Julius, C., and Yuzenkova, Y. (2017) Bacterial RNA polymerase caps RNA with various cofactors and cell wall precursors. *Nucleic Acids Res.* **45**, 8282–8290

22. Jackowski, S., and Rock, C. O. (1984) Metabolism of 4’-phosphopantetheine in

Escherichia coli. *J. Bacteriol.* **158**, 115

23. Bennett, B. D., Kimball, E. H., Gao, M., Osterhout, R., Van Dien, S. J., and Rabinowitz, J. D. (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nat. Chem. Biol.* **5**, 593–599
24. Nazi, I., Koteva, K. P., and Wright, G. D. (2004) One-pot chemoenzymatic preparation of coenzyme A analogues. *Anal. Biochem.* **324**, 100–105
25. Sapkota, K., and Huang, F. (2018) Efficient one-pot enzymatic synthesis of dephospho coenzyme A. *Bioorg. Chem.* **76**, 23–27
26. Stewart, C. J., Thomas, J. O., Ball, W. J., and Aguirre, A. R. (1968) Coenzyme A Analogs. III. The Chemical Synthesis of Desulfopantetheine 4′-Phosphate and Its Enzymatic Conversion to Desulfo-coenzyme A. *J. Am. Chem. Soc.* **90**, 5000–5004
27. Van Der Westhuyzen, R., Hammons, J. C., Meier, J. L., Dahesh, S., Moolman, W. J. A., Pelly, S. C., Nizet, V., Burkart, M. D., and Strauss, E. (2012) The antibiotic CJ-15,801 is an antimetabolite that hijacks and then inhibits CoA biosynthesis. *Chem. Biol.* **19**, 559–571
28. Mercer, A. C., Meier, J. L., Hur, G. H., Smith, A. R., and Burkart, M. D. (2008) Antibiotic evaluation and in vivo analysis of alkynyl Coenzyme A antimetabolites in Escherichia coli. *Bioorganic Med. Chem. Lett.* **18**, 5991–5994
29. Xu, F., Sue, L. C., and Cohen, S. N. (1993) The Escherichia coli pcnB gene promotes adenylylation of antisense RNAI of ColE1-type plasmids in vivo and degradation of RNAI decay intermediates. *Proc. Natl. Acad. Sci. U. S. A.* **90**,

6756–6760

30. Milligan, J. F., Groebe, D. R., Witherell, G. W., and Uhlenbeck, O. C. (1987) Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic Acids Res.* **15**, 8783

31. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415

32. Biondi, E., and Benner, S. A. (2018) Artificially Expanded Genetic Information Systems for New Aptamer Technologies. *Biomedicines*. 10.3390/BIOMEDICINES6020053

33. Biondi, E., and Burke, D. H. (2012) Separating and analyzing sulfur-containing RNAs with organomercury gels. *Methods Mol. Biol.* **883**, 111–120

34. Igloi, G. L. (1988) Interaction of tRNAs and of Phosphorothioate-Substituted Nucleic Acids with an Organomercurial. Probing the Chemical Environment of Thiolated Residues by Affinity Electrophoresis. *Biochemistry*. **27**, 3842–3849

35. Rhee, S. S., and Burke, D. H. (2004) Tris(2-carboxyethyl)phosphine stabilization of RNA: Comparison with dithiothreitol for use with nucleic acid and thiophosphoryl chemistry. *Anal. Biochem.* **325**, 137–143

36. Strauss, E., and Begley, T. P. (2002) The antibiotic activity of N-pentylpantothenamide results from its conversion to ethyldethia-coenzyme A, a coenzyme A antimetabolite. *J. Biol. Chem.* **277**, 48205–48209

37. Izard, T., and Geerlof, A. (1999) The crystal structure of a novel bacterial

adenylyltransferase reveals half of sites reactivity. *EMBO J.* **18**, 2021–2030

38. Izard, T. (2002) The crystal structures of phosphopantetheine adenylyltransferase with bound substrates reveal the enzyme's catalytic mechanism. *J. Mol. Biol.* **315**, 487–495

39. Izard, T. (2003) A novel adenylate binding site confers phosphopantetheine adenylyltransferase interactions with coenzyme A. *J. Bacteriol.* **185**, 4074–4080

40. Geerlof, A., Lewendon, A., and Shaw, W. V. (1999) Purification and characterization of phosphopantetheine adenylyltransferase from Escherichia coli. *J. Biol. Chem.* **274**, 27105–27111

41. Gilbert, W. (1986) Origin of life: The RNA world. *Nature*. **319**, 618

42. Samuelian, J. S., Gremminger, T. J., Song, Z., Poudyal, R. R., Li, J., Zhou, Y., Staller, S. A., Carballo, J. A., Roychowdhury-Saha, M., Chen, S. J., Burke, D. H., Heng, X., and Baum, D. A. (2022) An RNA aptamer that shifts the reduction potential of metabolic cofactors. *Nat. Chem. Biol.* **18**, 1263–1269

43. Li, N., and Huang, F. (2005) Ribozyme-catalyzed aminoacylation from CoA thioesters. *Biochemistry*. **44**, 4582–4590

44. Coleman, T. M., and Huang, F. (2002) RNA-catalyzed thioester synthesis. *Chem. Biol.* **9**, 1227–1236

45. Huang, F., Bugg, C. W., and Yarus, M. (2000) RNA-catalyzed CoA, NAD, and FAD synthesis from phosphopantetheine, NMN, and FMN. *Biochemistry*. **39**, 15548–15555

46. Sousa, R., and Padilla, R. (1995) A mutant T7 RNA polymerase as a DNA polymerase. *EMBO J.* **14**, 4609–4621

47. McEntee, K., Weinstock, G. M., and Lehman, I. R. (1980) recA protein-catalyzed strand assimilation: Stimulation by Escherichia coli single-stranded DNA-binding protein. *Proc. Natl. Acad. Sci. U. S. A.* **77**, 857–861

48. Ditzler, M. A., Lange, M. J., Bose, D., Bottoms, C. A., Virkler, K. F., Sawyer, A. W., Whatley, A. S., Spollen, W., Givan, S. A., and Burke, D. H. (2013) High-throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase. *Nucleic Acids Res.* **41**, 1873–1884

49. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10–12

50. Alam, K. K., Chang, J. L., and Burke, D. H. (2015) FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Mol. Ther. Nucleic Acids*. **4**, e230

51. Kramer, S. T., Gruenke, P. R., Alam, K. K., Xu, D., and Burke, D. H. (2022) FASTAptameR 2.0: A web tool for combinatorial sequence selections. *Mol. Ther. Nucleic Acids*. **29**, 862–870

SUPPLEMENTAL INFORMATION

Supplementary Methods

Solid phase synthesis of BKPP *(K. Sapkota)*

The synthesis of BKPP was performed at 0.1 mmol scale. Numbers below refer to steps in overview given in Scheme 1 of the main text.

i. – iii. Bead-immobilized biocytin. Amino functionalized resin (100 mg, 120 micomoles capacity) was swelled in dry DMF for 1 h. 190 mg (325 μmol) of N^{α}-Fmoc-N^{ε}-biotinyl-L-lysine (Chem-Impex International, catalog #04988) was activated by adding 130 mg (315 μmol, 0.96 eq.) of HCTU in 1 mL 20% NMM/DMF. The reaction produced yellow colored solution upon incubation at room temperature (rt) for 5 min. It was added to the beads and agitated for 30 min. The solvent was drained and the resin was washed 3x with 5 mL DMF. Unreacted amino groups on the beads were capped by adding 100 μl/ 1 mmol acetic anhydride in 1 mL DMF to the beads followed by agitation at rt for 30 min. Deprotection of Fmoc was carried out by adding 1 mL 20% piperidine in DMF to the beads. The reaction was incubated for 5 min at rt. The deprotection product was collected by draining the solvent and quantified by UV spectrophotometry, which showed nearly quantitative loading (~98%) of Fmoc-Lys(biotin)-OH on the beads.

iv. – v. Coupling of fmoc-cys(stbu)-OH. The deprotected amino group was used to react with HCTU activated carboxyl group of Fmoc-L-Cys(stbu)-OH, forming an amide bond. Commercially available Fmoc-Fys(Stbu)-OH (Chem Impex International, catalog number 02403) 172 mg (400 μmol) and 157 mg (380 μmol) HCTU was dissolved in 1 mL 20% NMM/DMF and allowed to react at rt for 5 min. It was then added to the beads and agitated

for 1 hr at rt. The solvent was drained and the resin was washed 3x with 2 mL DMF. Fmoc deprotection was carried out as above. Quantification by UV spectrophotometry showed quantitative yield of the coupling reaction.

vi. Loading of ^{14}C acetate. The reactive amino group generated through fmoc deprotection was used to install a radiolabeled acetate as a reporter tag. Acetic Acid, Sodium Salt, [1-^{14}C] was obtained as a generous gift from Dr. Philip Bates as 3 mCi/mL aqueous solution. Similarly, non-radiolabeled sodium acetate, 5 mg (60 μmol) was dissolved in 0.6 mL water as a 0.1 M solution, to which was added 30 μl Acetic Acid, Sodium Salt, [1-^{14}C]. The solution was frozen and lyophilized to obtain 62 μmol of sodium acetate with 100 μCi radioactivity. 25 mg (60 μmol) HCTU and 0.5 mL 20% NMM/DMF were added to NaOAc to activate the carboxyl group. The reaction was vortexed at rt for 10 min until completely dissolved. The activated ^{14}C-acetate was added to the beads and allowed to react for 1 hr with gentle agitation. Non-radiolabeled sodium acetate, 41 mg, was activated in the same way and added to the beads to cap the unreacted amino group. The solvent was drained and the radiolabeling yield was found to be ~70% as determined by liquid scintillation counting.

vii. Stbu deprotection. The deprotection of S-t-butyl (stub) protection group was carried out by DTT-mediated disulfide reduction. DTT (77 mg) was dissolved in 0.5 mL DMF and added to the beads. 30 μL DIPEA was also added to create a basic environment for the

reduction. The reaction was carried out at 60 °C. A small portion of the beads were taken, deprotected and analyzed by HPLC to monitor the reaction progress. The reaction was completed after 2 h of incubation at 60 °C as shown by HPLC analysis. Solvent was drained and the resin was washed 3x with 4 mL DMF.

viii. Coupling of 1,3-diiodopropane. The reactive thiol group on the beads was reacted with 1,3 diiodopropane to yield an iodo-functional group on the beads. 191 mg (650 µmol) 1,3 diiodopropane (Oakwood chemical, catalog # 003098) in 500 µL DMF was added to the beads and reacted at rt for 30 min with gentle agitation. Excess solvent and unreacted diiodopropane were drained and the resin was washed 3x with 4 mL DMF. The iodo functional group generated this way was used to react with the pantetheine thiol in the next step of synthesis.

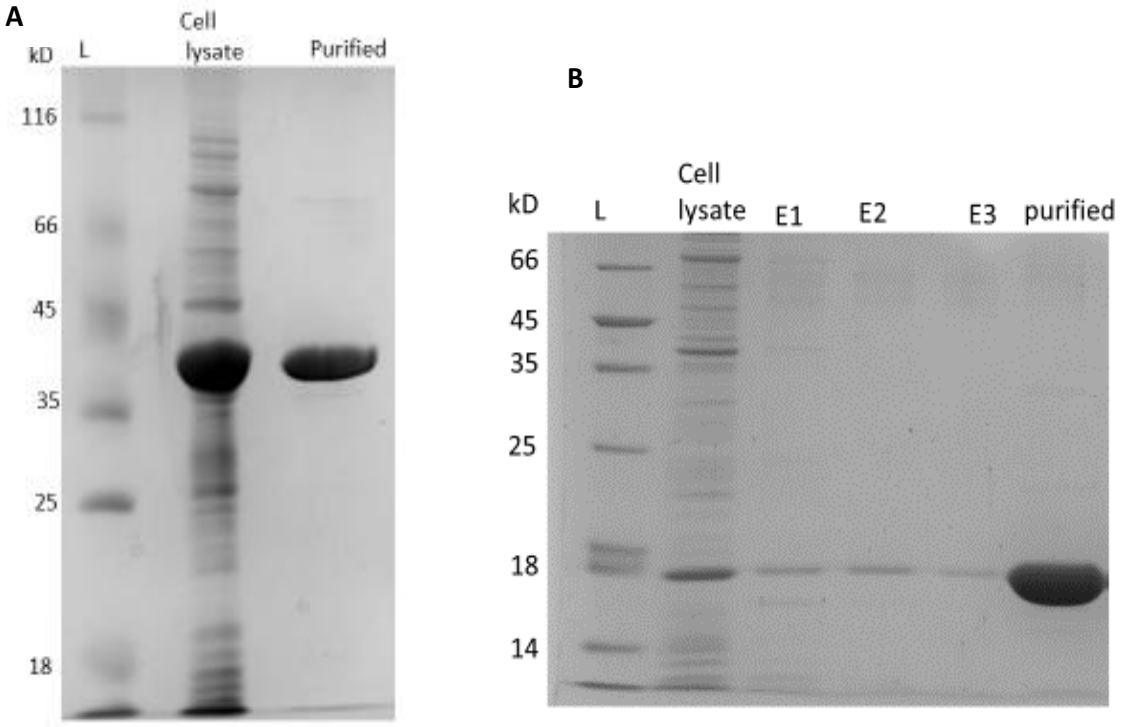
ix. Loading of Pantetheine. Pantethine (Ox-Pant, Chem-Impex International, catalog number 00240) was reduced with DTT to yield pantetheine. 150 mg (270 µmole) Ox-pant was dissolved in 2 mL DMF, 63 mg (405 µmoles) of DTT was added (Sigma Aldrich), and the reduction was carried out at rt for 1 hr with gentle stirring. The pantetheine product was precipitated by adding 5 mL diethyl ether, washed, and dried in a desiccator. It was then dissolved in 1 ml DMF and added to the resin. The reaction was carried out at rt for 30 min with gentle agitation. The solvent was then drained and the resin was washed 3x with 2 mL DMF, completing the solid phase synthesis of BKPP.

x. Deprotection from the beads. The compound was deprotected form the trityl resin by adding 1 mL deprotection cocktail composed of TFA/DCM/TIPS in 1:18:1 ratio at rt for 10 min. The solution was drained directly in 10 mL ether to precipitate the product. Beads were washed 3x with 0.5 mL DCM and drained into diethyl ether. The precipitate was collected by centrifugation and dried at reduced pressure in a desiccator to yield 78 mg (76%) of BKP as a white gummy solid.

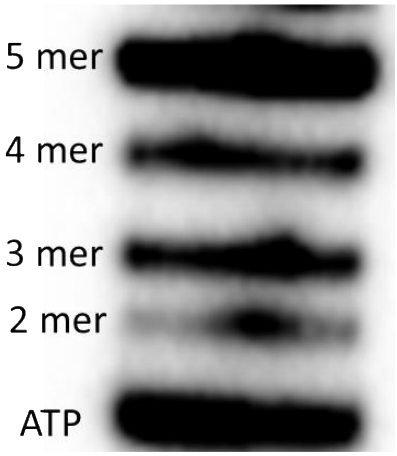
xi. Phosphorylation of BKP to form BKPP.

The phosphorylation of BKP was carried out by recombinant pantothenate kinase (PanK) purified from *E. coli* (Fig S1a). BKP (7.8 mg, 10 µmoles) was dissolved in 1 mL reaction buffer (50 mM Tris pH 7.0, 10 mM KCl, 1 mM $MgCl_2$) and 11 mg (20 µmole) ATP was added. The reaction was started by adding 0.2 mg PanK and incubated at 37 °C. Reaction progress was analyzed by HPLC and found to be completed in 2 hr. Activity of the enzyme in phosphorylating BKP was found to be comparable to its natural substrate pantetheine. The BKPP product was purified to its highest purity by reverse phase HPLC. The reaction was loaded into a 4.6 x 250 mm C18 column equilibrated with 10% 40 mM KH_2PO_4 at a flow rate of 1 mL/min. The column was washed with 5% acetonitrile for 5 min to remove ATP and ADP. The product was eluted with 50:50 MeCN: Water with no buffer. It was then dried under reduced pressure at 60 °C to yield 7.7 mg (90%) of pure BKPP product.

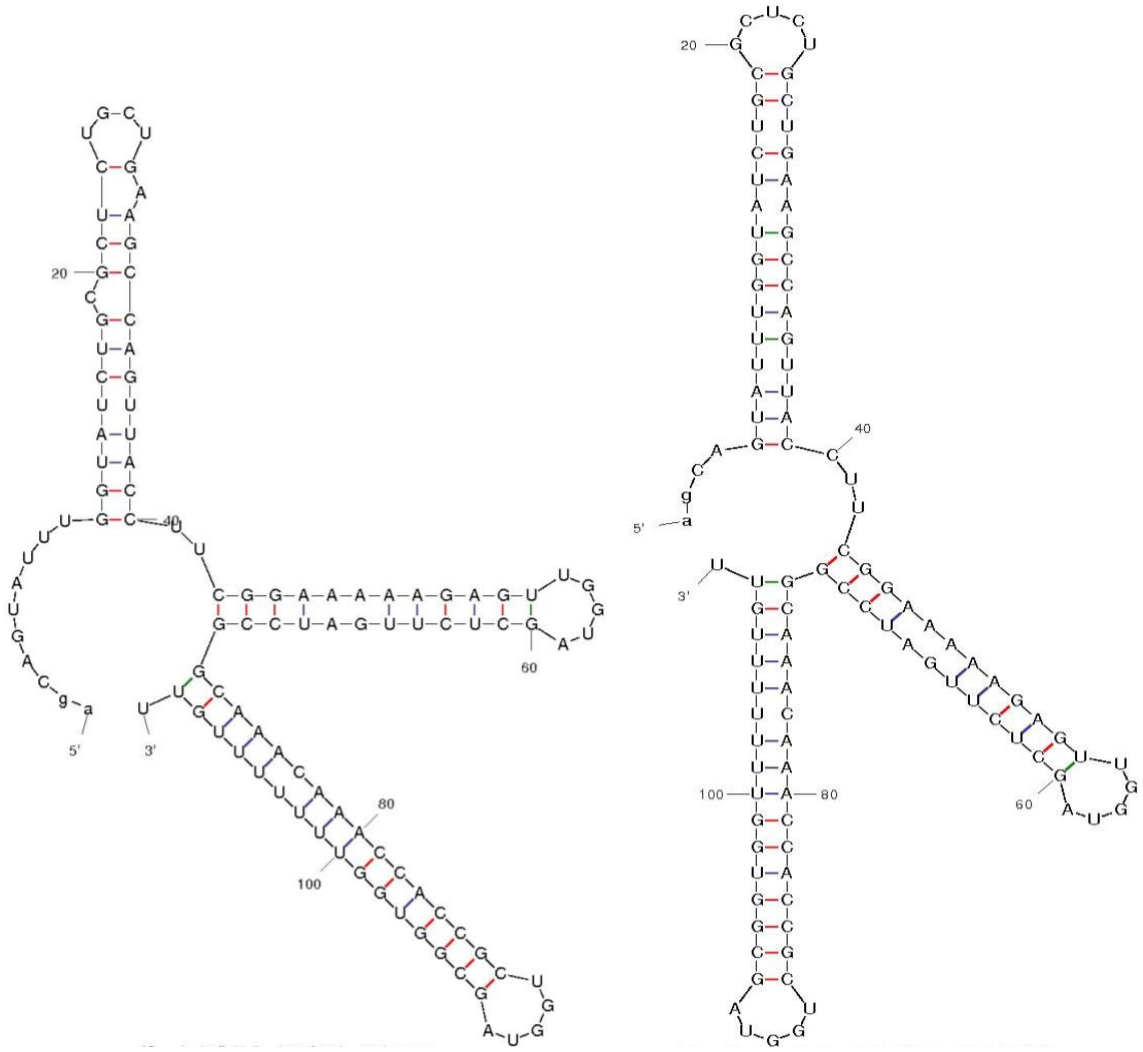
The concentration was further analyzed by liquid scintillation counting and was stored as 5 mM aqueous solution having radioactivity of ~50,000 CPM/µL.



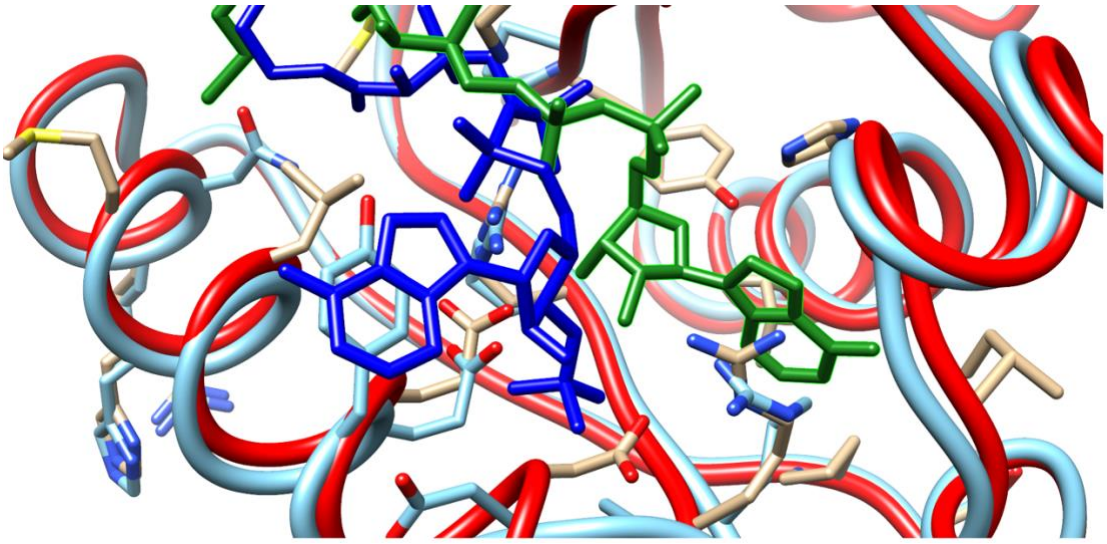
Supplemental Figure 3.1. Expression and purification of recombinant PanK (CoaA) and PPAT (CoaD). A) PanK was expressed in BL21(DE3) cells as His-tagged recombinant protein and purified by Ni-NTA chromatography. A dark single band in lane 3 shows the good purity of the PanK after purification. B) The PPAT enzyme was purified same way as PanK. A dark band below 18,000 Da marker corresponds to PPAT. After eluting with 200 mM imidazole, both enzymes were concentrated by a membrane filter of 10,000 Da cutoff and stored at -20 °C in 1x storage buffer containing 50% glycerol. *Figure and data generated by K. Sapkota.*



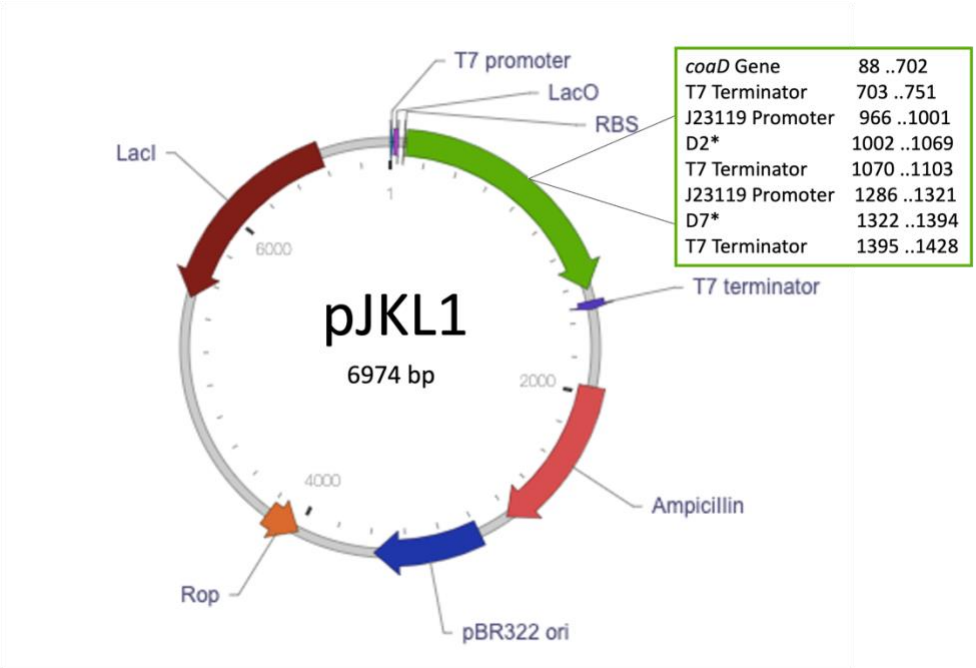
Supplemental Figure 3.2. Purification of 5mer RNA by denaturing PAGE. RNA was prepared by standard *in vitro* transcription using a synthetic DNA template (Figure 3.3a, main text). Only two of the four nucleotides (ATP and GTP) were used in the reaction to abort it after 5 nucleotides. Transcription produced a heterogenous mixture of 2 mer, 3 mer, 4 mer and 5 mer which were separated on 20% denaturing PAGE of 0.4 mm thickness. The band corresponding to 5 mer was excised and purified by crush and soak method. *Figure and data generated by K. Sapkota.*



Supplemental Figure 3.3. Predicted secondary structures of E. coli RNAI. Mfold server was used to predict the secondary structures that showed two possible structures of comparable free energy having a stretch of either 10 or 4 unpaired nucleotides at the 5' terminus. Not shown are the AG or GG dinucleotides appended to the 5' ends to aid *in vitro* transcription with T7 RNA Polymerase. *Figure and data generated by K. Sapkota.*



Supplemental Figure 3.4. Overlay of binding of the CoaD product dpCoA (green) and feedback competitive inhibitor CoA (blue) in one protomer of the PPAT hexamer. CoA and dephospho-CoA bind differently, as the cysteamine part of pantetheine arm orients in opposite directions and CoA adenosine does not bind in the adenylate binding site of PPAT. *Figure and data generated by K. Sapkota.*



Supplemental Figure 3.5. Plasmid map and sequence of pJKL1. pJKL1 pET-{PPAT, term, D2+D7}, 6974 bp. Vector Builder ID: VB221111-1136vyp. Plasmid map (above) and sequence (below) 5' → 3'. *Figure and design generated by J. Lucas.*

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGA
AATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGCAAAAACGGGCGAT
TTATCCGGGTACTTTCGATCCCATTACCAATGGTCATATCGATATCGTGACGC
GCGCCACGCAGATGTTCGATCACGTTATTCTGGCGATTGCCGCCAGCCCCAGT
AAAAAACCGATGTTTACCCTGGAAGAGCGTGTGGCACTGGCACAGCAGGCAA
CCGCGCATCTGGGGAACGTGGAAGTGGTCGGGTTTAGTGATTTAATGGCGAA
CTTCGCCCGTAATCAACACGCTACGGTGCTGATTCGTGGCCTGCGTGCGGTGG
CAGATTTTGAATATGAAATGCAGCTGGCGCACATGAATCGCCACTTAATGCC
GGAACTGGAAAGTGTGTTTCTGATGCCGTCGAAAGAGTGGTCGTTTATCTCTT
CATCGTTGGTGAAAGAGGTGGCGCGCCATCAGGGCGATGTCACCCATTTCCT
GCCGGAGAATGTCCATCAGGCGCTGATGGCGAAGTTAGCGTAGCGTTGGATC
CGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCA
CCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCT
GCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGG
TCTTGAGGGGTTTTTTGCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCC
CTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTAAT
TCGTAACTGCTGCTGGGATTACACATGGCATGGATGAGCTCTACAAATAATG
AAACGAATTCAAGCTTGATATCATTCAGGACGAGCCTCAGACTCCAGCGTAA

CTGGACTGCAATCAACTCACTTTGACAGCTAGCTCAGTCCTAGGTATAATGCT
AGCAGGTGCAGACGATTCGTCACACGTCCTGCACACGGCTGGAAACTACTCA
TCATCATCGAGCTGGCAGTCGCAAAAAACCCCGCTTCGGCGGGGTTTTTTCGC
GTAACTGCTGCTGGGATTACACATGGCATGGATGAGCTCTACAAAGGTCAAT
ACACTACATGGCGTGATTTCATATGCGCGATTGCTGATCCCCATGTGTATCAC
TGGCAAACTGTGATGGACGACACCGTCAGTGCGTCCGTCGCGCAGGCTCTCG
ATGAGCTGATGCTTTGGGCCGAGGATTGACAGCTAGCTCAGTCCTAGGTATA
ATGCTAGCAGGACCAGTGCAGACGATTCGTCACACGTCCTGCACACGGCTGG
AAACACGTAATCATCATCGAGCTGGCAGTCGCAAAAAACCCCGCTTCGGCGG
GGTTTTTTCGCGGATCCGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGC
TGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGG
GTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATATCCCGCAA
GAGGCCCGGCAGTACCGGCATAACCAAGCCTATGCCTACAGCATCCAGGGTG
ACGGTGCCGAGGATGACGATGAGCGCATTGTTAGATTTCATACACGGTGCCT
GACTGCGTTAGCAATTTAACTGTGATAAACTACCGCATTAAAGCTTATCGATG
ATAAGCTGTCAAACATGAGAATTCTTGAAGACGAAAGGGCCTCGTGATACGC
CTATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCAGGTGG
CACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATAC
ATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAA
TATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCC
CTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAA
AGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTG
GATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCC
AATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTG
ACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTA
AGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACT
TACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAA
CATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAA
GCCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACA
ACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACA
ATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCG
GCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGG
GTCACGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATC
GTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGAC
AGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCA
AGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAG
GATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTG
AGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCT
TGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACC
GCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGA
AGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTA

GCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTC
GCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCT
TACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGC
TGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCG
AACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGG
GAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCG
CACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGG
TTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCG
GAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTT
GCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATA
ACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGAC
CGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGGGCGCCTGATGCGGTA
TTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCAATGGTGCACTCTCAGT
ACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTA
CGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGC
CCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTC
TCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGA
GGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTC
TGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCT
GGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACT
GATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGA
AACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTT
ACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCA
GAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGT
GTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGG
TGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAA
GACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTC
ACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGC
CAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGCC
AGGACCCAACGCTGCCCGAGATGCGCCGCGTGCGGCTGCTGGAGATGGCGG
ACGCGATGGATATGTTCTGCCAAGGGTTGGTTTGCGCATTCACAGTTCTCCGC
AAGAATTGATTGGCTCCAATTCTTGGAGTGGTGAATCCGTTAGCGAGGTGCC
GCCGGCTTCCATTCAGGTCGAGGTGGCCCGGCTCCATGCACCGCGACGCAAC
GCGGGGAGGCAGACAAGGTATAGGGCGGCGCCTACAATCCATGCCAACCCG
TTCCATGTGCTCGCCGAGGCGGCATAAATCGCCGTGACGATCAGCGGTCCAA
TGATCGAAGTTAGGCTGGTAAGAGCCGCGAGCGATCCTTGAAGCTGTCCCTG
ATGGTCGTCATCTACCTGCCTGGACAGCATGGCCTGCAACGCGGGCATCCCG
ATGCCGCCGGAAGCGAGAAGAATCATAATGGGGAAGGCCATCCAGCCTCGC
GTCGCGAACGCCAGCAAGACGTAGCCCAGCGCGTCGGCCGCCATGCCGGCG
ATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGG
CTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCAT
CGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCC

GGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGA
CGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCT
CAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTA
ATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCT
GCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGC
CAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCA
CCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAG
CAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTG
TCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCC
GGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACC
AGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAA
ACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGAT
TGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAG
AACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAG
ATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTG
ATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGG
CAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAG
CCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCG
ACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGC
GCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTG
GAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCA
CGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGC
GTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGAT
AAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTC
ACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGT
TTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCC
TGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCG
CAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGG
GGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCG
AGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCA
CCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGA
TCTCGATCCCGCGAAAT

Supplemental Tables

Total Isolated RNA Samples	**+I 1.1**	**+I 1.2**	**+I 2.1**	**+I 2.2**	**-I 3.1**	**-1 3.2**	**-I 4.1**	**-I 4.2**
Raw total reads	142	27,205	19,775	30,591	265	1,281	11,952	926
Discarded Long reads (>29nt)	130	3,075	1,600	3,893	259	1,218	1,879	601
Discarded Short reads (<23nt)	-	14	6	14	-	2	2	3
Total processed sequence reads	12	24,116	18,169	26,684	6	61	10,071	322
D2* + D7* RNA Total Counts	11	23,210	17,513	25,738	6	58	9,668	307
D2* RNA	9	18,674	13,610	20,088	4	42	7,626	246
D7* RNA	2	4,536	3,903	5,650	2	16	2,042	61
Percent of D2* RNA	81.8	80.5	77.7	78.1	66.7	72.4	78.9	80.1
Percent of D7* RNA	18.2	19.5	22.3	22.0	33.3	27.6	21.1	19.9

Supplemental Table 3.1. High-throughput sequencing raw data and processing for total isolated RNA samples (without sulfur fractionation). Data processing was performed using cutadapt to trim the 5' and 3' constant regions from sequences and to discard any uncut sequences or sequences with lengths not within ± 3 nt of the expected size (26 nt) after trimming. Raw total reads is the number of sequences prior to any processing, long and short sequences did not fit within the ± 3 nt parameter, and the total processed sequence reads were analyzed using FASTAptameR2.0. Of the total processed

reads, those that contained the index sequence for D2* or D7* RNA were identified and counted. The percent of D2* and D7* was determined by dividing the individual D2* or D7* counts by the total number of (D2* + D7*) counts. Column headers indicate sources of each RNA sample. +I indicates plus induction and -I indicates no induction. The first number indicates which culture (1-4) the sample was isolated from and the second number indicates which technical replicate the sample came from. For instance, sample "+I 1.2" was from the second technical replicate of the first culture that was induced to express PPAT. Columns shown in light gray indicate samples with fewer than 2,000 total unique processed reads and are not plotted in figure 8. *Table and analyses generated by J. Lucas*

Sulfur Partitioned Samples	+I 1.1	+1 1.2	+I 2.1	+1 2.2	-I 3.1	-1 3.2	-I 4.1	-I 4.2
Raw total reads	19,444	25,809	27,779	945	8,658	17,253	806	18,960
Discarded Long reads (>29nt)	3,050	3,163	3,556	933	1,343	6,550	804	3,058
Discarded Short reads (<23nt)	8,618	8,288	3,927	0	6,177	8,262	0	4,243
Total processed sequence reads	7,776	14,358	20,296	12	1,138	2,441	2	11,659
D2* + D7* RNA Total Counts	6,081	12,112	18,502	11	322	1,564	2	10,873
D2* RNA	1,136	2,303	7,415	11	183	599	1	2,856
D7* RNA	4,945	9,809	11,087	0	139	965	1	8,017
Percent of D2* RNA	18.7	19.0	40.1	100.0	56.8	38.3	50.0	26.3
Percent of D7* RNA	81.3	81.9	60.0	0.0	43.2	61.7	50.0	73.7

Supplemental Table 3.2. High-throughput sequencing raw data and processing for sulfur partitioned RNA samples. The same data processing was performed as described in Supplemental Table 3.1. *Table and analyses generated by J. Lucas.*

Name	**Sequence (5'→3')**
D2 RNA	AGGUGCGAAAGCACACAGAGUA
D3 RNA	AGGGUGCGAAAGCACACAGAGU
D4 RNA	AGGAGUGCGAAAGCACACAGAG
D5 RNA	AGGACGUGCGAAAGCACACAGA
D7 RNA	AGGACCAGUGCGAAAGCACACA
D10 RNA	AGGACCAAGCGUGCGAAAGCAC
Type II P +1A top strand	GCGTAATACGACTCACTATTAGG
Type III P +1G top strand	GCGTAATACGACTCACTATAGAG
D2 3' end labeling oligo	GGTACTCTGTTATGGG
D7 3' end labeling oligo	GGTGTGTGCTCGATAT
Plasmid RNA 3' end labeling oligo	GGGACTGCCAGCTCGATGATGA
Plasmid RNA 3' end labeling oligo	GGGACTGCCAGCTCGATGATGA
D2* RNA	**AG**GAGCAGACGAUUCGUCACACGUCCUGCUCUCGGCUG GAAACU***ACUCA***UCAUCAUCGAGCUGGCAGUCGCAAAAA ACCCCGCUUCGGCGGGGUUUUUUCGC
D7* RNA	**AGGACCA**GAGCAGACGAUUCGUCACACGUCCUGCUCUC GGCUGGAAACU***ACGUAA***UCAUCAUCGAGCUGGCAGUCGC AAAAAACCCCGCUUCGGCGGGGUUUUUUCGC
Forward Primer for HTS	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTA CACGACGCTCTTCCGATCTGACGATTCGTCACACGTCCTG CTCTCGG
Reverse Primer 1 for HTS	CAGACGTGTGCTCTTCCGATCCCGAAGCGGGGTTTTTTGC GACTGCCA
Reverse Primer 2 for HTS	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGG AGTTCAGACGTGTGCTCTTCCGATCTGTATCCGCCGAAGC GGATTGG
5' Universal HTS Adapter	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTA CACGACGCTCTTCCGATCT
3' Indexed HTS Adapter	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGG AGTTCAGACGTGTGCTCTTCCGATC

Supplemental Table 3.3. RNA and oligo sequences. ”Top strand” oligos were mixed with bottom strand oligos (reverse complement of RNA sequences shown, plus antisense to “top strand” oligo) for D2-10 as template for transcriptions. Note, the +1 nucleotide for D2-D7 was determined during transcription by the choice of which “top strand” oligo was used during template assembly for the transcription reaction (underlined). The default depiction in this table is +1A for D2-D10. In the promoter sequences, red regions are not part of the promoter and underlined regions indicate the start of RNA sequence (+1→+3). The bolded and italicized blue regions of the plasmid D2* and D7* RNA indicate the 6 nucleotide indexes that differentiate plasmid D2* and D7* RNA apart from their 5’ end sequences. Segments highlighted in yellow were used during high-throughput sequencing analysis to identify whether a given amplicon originated from D2* or D7* RNA. Underlined regions of plasmid D2* and D7* RNA indicate where the sequencing primers bind. Sequences for the high-throughput sequencing primers used to append the Illumina adapters and their respective sequencing indices. We used the NEBNext Index (1-16) Primers for Illumina. 16 reverse primers (corresponding to the 16 indices) were used in the second PCR for high-throughput sequencing preparation. The index region is indicated by the six red N region in the reverse primer 2 for HTS sequence. Index sequences are from the instruction manual for the NEBNext Multiplex Small RNA Library Prep Set 1, Set 2, Index Primers 1-48 and Multiplex Compatible (https://rb.gy/0dbqe9). *Table generated by J. Lucas.*

Chapter 4: A method for identifying and partitioning Coenzyme A linked RNAs

This work will constitute a majority of the material that will be submitted for publication after further developments outlined at the end of the chapter. Anticipated co-authors include Jordyn K. Lucas, Matthew F. Lichte, Donald H. Burke-Agüero and potentially others.

ABSTRACT

Natural and synthetic coenzyme A-linked RNAs (CoA-RNAs) are of interest both for their undetermined biological roles and their potential uses in synthetic biology. Therefore, having a reliable method to partition CoA-RNAs from other total RNA is a crucial step for studying and making use of them. Here, we developed a method for separating CoA-RNAs from total RNA by making use of the free sulfur on CoA and mercury-containing polyacrylamide gels. Taking advantage of this CoA-RNA partition method, we developed key portions of a system to separate potential CoA-RNAs from total cellular RNAs and prepared them for high-throughput sequencing in an effort to identify endogenous CoA-RNAs. Unfortunately, existing methods for ligating pre-adenylated sequencing adapters to RNA were extremely inefficient and highly biased against RNAs with 3' end structures, negatively impacting the results of this experiment. This, in combination with what we hypothesize to be very low existing quantities of natural CoA-RNAs, prevented clear

detection and identification of biological CoA-RNAs. Further optimization of the ligation of pre-adenylated adapters to RNAs may improve outcomes enough to allow for possible identification of endogenous CoA-RNAs. Furthermore, we used a simplified version of this method to isolate known RNA sequences from cells by omitting the adapter ligation step to partition CoA-RNAs from cells (Chapter 3), demonstrating its usefulness and viability.

INTRODUCTION

Cells from all three Domains of life generate several nucleotide analogs such as cyclic adenosine monophosphate (cAMP), nicotinamide adenine dinucleotide (NAD+), and coenzyme A (CoA). These analogs play pivotal roles as signaling molecules, energy carriers, and enzyme cofactors. Nucleotide analogs with free 3' hydroxyl groups (e.g., NAD+, FAD, and 3' dephospho CoA) were also found suitable as non-canonical initiator nucleotides (NCINs) for *in vitro* RNA transcription (1) to generate CoA-RNA, NAD-RNA, and FAD-RNA. More interestingly, previous studies reported the presence of naturally occurring cofactor-linked RNAs in bacterial cells, in which NAD+, CoASH and acyl-CoAs were reported to be present in the most 5' position of transcripts (2, 3). Since then, studies have described NAD+'s protective, cap-like function in bacteria and its role in promoting mRNA degradation in eukaryotic cells (3–7). In contrast, the identities and functional roles, if any, of natural CoA-linked RNA have not yet been explored. CoA-RNAs have not received the same level attention in part due to lack of a well-established methods for

capturing and sequencing CoA-RNAs from total cellular RNA. Also, unlike NAD-RNA which is fairly abundant in cells, intracellular levels of CoA-RNA are hypothesized to be much lower, such that any successful CoA Capture Seq method would need to be extremely effective at capturing and detecting very small quantities of CoA-RNA.

The primary reason for the hypothesized difference in abundance between NAD-RNAs and CoA-RNAs is related to the mechanism by which they can be generated. The primary mechanism for cellular NAD-RNA formation is non-canonical nucleotide initiation (8). Simply, when RNA is being transcribed, NAD+ is substituted for ATP, the canonical nucleotide, and is incorporated as the +1 nucleotide generating an RNA with an NAD+ 'cap'. Intracellular NAD+ levels in *E. coli* typically fluctuate between 4-7 mM while intracellular ATP levels typically fluctuate between 1-5 mM. Furthermore, RNA polymerase in *E. coli* has a K_m of ~0.38 mM for NAD+, about 10 times lower than intracellular NAD+ levels (9). Although RNA polymerase in *E. coli* has a K_m of ~0.090 mM for ATP, its canonical substrate, it is still feasible for NAD+ to both be substituted for ATP and be incorporated by the RNA by the polymerase into the +1 position of RNA transcripts. The ~4x difference in RNAP affinities for NAD+ and ATP and their similar intracellular concentrations predict around a quarter of transcripts with +1A to have an NAD+ cap. Interestingly, the Jaschke group observed approximately 25% NAD-capping for RNAI transcripts in bacterial cells with NAD+ de-capping enzyme NudX knocked out (10), corresponding closely with the predicted quantities of cellular NAD-RNA.

Meanwhile, CoA-RNAs are at a great disadvantage. Coenzyme A cannot serve as an NCIN or be incorporated into the +1 position of transcripts due to the phosphate on the 3' hydroxyl. In fact, the NCIN for CoA-RNAs is actually 3' dephospho CoA (dpCoA), a metabolic precursor in the CoA synthesis pathway that is present in vanishingly small quantities in the cell. For context, acetyl-CoA and malonyl-CoA, the primary forms of Coenzyme-A in cells, have reported intracellular concentrations between 20-600 μM and 4-90 μM respectively in *E. coli* (11). Even with an unrealistic 1:1 ratio of dpCoA to acetyl-CoA at its highest concentration, dpCoA would be 10 times less concentrated than NAD+. Furthermore, it's unlikely that dpCoA would be available at concentrations anywhere near 600 μM. In fact, one study reported dpCoA intracellular concentrations to be 20 μM and NAD+ concentrations were 12 mM in the anaerobic bacterium *C. kluyveri* (12). The low levels of dpCoA would make competing with ATP for the +1 spot of RNA transcripts difficult. Therefore, non-canonical nucleotide initiation by dpCoA is most likely not a significant mechanism for generating cellular CoA-RNAs.

An alternative CoA capping mechanism is post-transcriptional capping by an enzyme called phosphopantetheine adenylyltransferase (PPAT). PPAT is part of the CoA biosynthesis pathways and is responsible for turning substrates ATP and phosphopantetheine into 3' dephospho CoA. It's possible that cellular ATP-RNA may serve as an analog for ATP in select contexts, that it can be used as a substrate for PPAT,

and that it is capped with phosphopantetheine to become CoA-RNA. The Huang group performed several *in vitro* experiments that demonstrate that PPAT is capable of capping ATP-RNAs *in vitro* to generate CoA-RNAs, and we observed RNAs which meet the *in vitro* RNA substrate requirements of PPAT being preferentially capped to form CoA-RNAs (unpublished work, reference chapter 3 Figure 3.5). Thus, it is possible that PPAT is responsible for CoA capping of RNAs by a post transcriptional capping mechanism in cells.

Although there is precedent for the existence of cellular CoA-RNAs and post-transcriptional mechanisms for their capping seems plausible, their identities have remained elusive. The absence of a reliable CoA Capture Seq method has been a roadblock to exploring many longstanding hypotheses about cellular CoA-RNAs. It is possible that CoA-RNAs use their prosthetic groups to enhance RNA's chemical diversity for a variety of reactions (such as metabolic chemistry) or that they evolved to differentiate RNAs for actions such as signaling, trafficking, or regulation. Perhaps the identities and quantities of CoA-capped transcripts are be affected by various growth conditions and stressors as was seen with NAD-RNAs. However, the functions, roles, and identifies of cellular CoA-RNAs will remain speculative without a method for isolating and identifying them.

Here we describe the status of a method in development to capture and sequence *known* CoA-RNAs from cells. We developed and optimized RNA isolation and DNase treatments

to achieve the greatest RNA yield from bacterial cells. We also established a reliable method to de-acylate acyl-CoA-RNAs with cysteamine. Additionally, using previously described methods to serve as a guide (13), we used APM gels to separate CoA-RNAs from total RNAs, a key component of the CoA Capture Seq method. Also, we describe some of the key remaining challenges with this method, including low efficiency of ligation of pre-adenylated adapters to RNA templates and non-specific PCR products, both of which contribute to low quality high-throughput sequencing data.

RESULTS

CoA Capture Seq Overview

To capture and identify cellular CoA-RNAs we developed our CoA Capture Seq method (Figure 4.1). The first step is to grow the bacterial cultures and isolate the RNA from cells. Next the RNA is treated with TURBO DNase to remove any contaminating genomic DNA. Following DNase treatment, six reference RNAs are added to the isolated RNA. These reference RNAs will serve as internal controls for the final high-throughput sequencing (HTS) data. After separating the samples into total RNA and CoA-RNA, the CoA-RNA undergoes a de-acylation step to remove any acetyl groups on the sulfur of the CoA that prevent the CoA's sulfur from interacting with the mercury layer of the APM gel. Both total and CoA-RNAs are ligated to pre-adenylated adapters required for sequencing before partitioning CoA-RNAs from total RNA with an APM gel. After eluting CoA-RNAs from

the APM layer, the RNA samples are reverse transcribed and amplified before being sent to the University of Missouri, Columbia DNA Core for HTS.

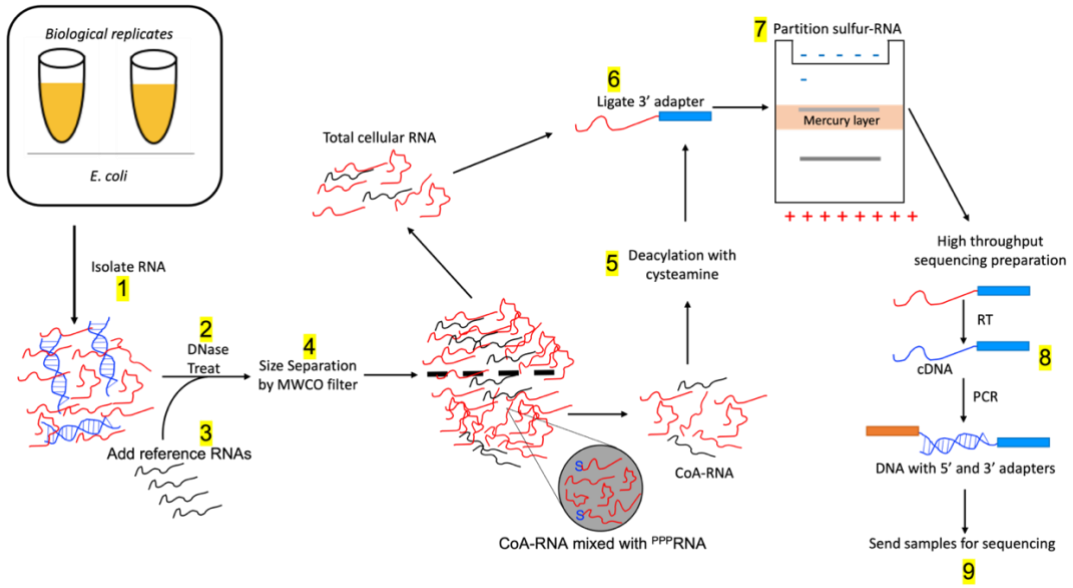


Figure 4.1. CoA-Capture Seq Schematic. Schematic summarizing the various steps to isolate, prepare, partition, and sequence cellular CoA-RNAs from total cellular RNA. **(1)** RNA is isolated from bacterial cultures (two cultures are grown separately as biological replicates) and later separated into two technical replicates. **(2)** RNA is DNase treated, **(3)** reference RNAs are added, **(4)** size fractionated with a 100 kDa molecular weight cutoff filter, and each replicate sample is split into two separate tubes, which then become "total cellular RNA" or "sulfur-RNAs". Sulfur- RNAs are then **(5)** de-acylated. Then both total and sulfur RNAs samples **(6)** ligated to adapters and **(7)** are then partitioned on an APM gel following. Following **(8)** purification from the APM gel, **(9)** samples undergo RT-PCR in preparation for high-throughput sequencing.

Step 1: RNA isolation

As a first step towards trying to determine the functions provided by a CoA modification, we aimed to identify which bacterial transcripts carry a CoA cap under various growth conditions. We isolated RNA from bacterial cultures grown to exponential and stationary phase as well as cultures grown up in stressful, acidic environments (materials and methods). We hypothesized that identifying which transcripts have a CoA modification under differing cellular conditions could elucidate whether CoA capping occurs randomly or is the result of a concerted cellular effort. Each bacterial growth condition (exponential, stationary, acid stress, etc) had two cultures grown to serve as biological replicates. RNA was isolated from the bacterial cultures as described in the materials and methods. Isolated RNA samples were nanodropped to confirm successful RNA isolation as well as to determine the relative nucleic acid concentration before moving on to next steps.

Step 2: TURBO DNase treatments of isolated cellular RNA removes contaminating genomic DNA.

Following isolation of RNA, it is imperative to remove any contaminating genomic or plasmid DNA. If left unchecked, the leftover DNA can persist in the samples until they are submitted for HTS, which can generate false positive signals in the HTS data. Thus, following the RNA isolations step, all samples were treated with TURBO DNase to remove any contaminating DNA. Aliquots of each sample were set aside before DNase treatments and after the first and second treatments and used as template for PCR to assess the

effectiveness of the TURBO DNase treatment. Because we expected there to be some contaminating DNA, we anticipated there would PCR product from samples that had not been DNase treated. As shown in Figure 4.2, before treatment (denoted as “0”), PCR bands can be observed in all samples indicating the presence of contaminating genomic DNA; however, following two DNase treatments, PCR products were no longer observable, suggesting that all genomic DNA had been eradicated by the DNase treatments. These data indicate that our isolated RNA no longer contained any DNA and was ready to continue on to the next steps in preparation for HTS.

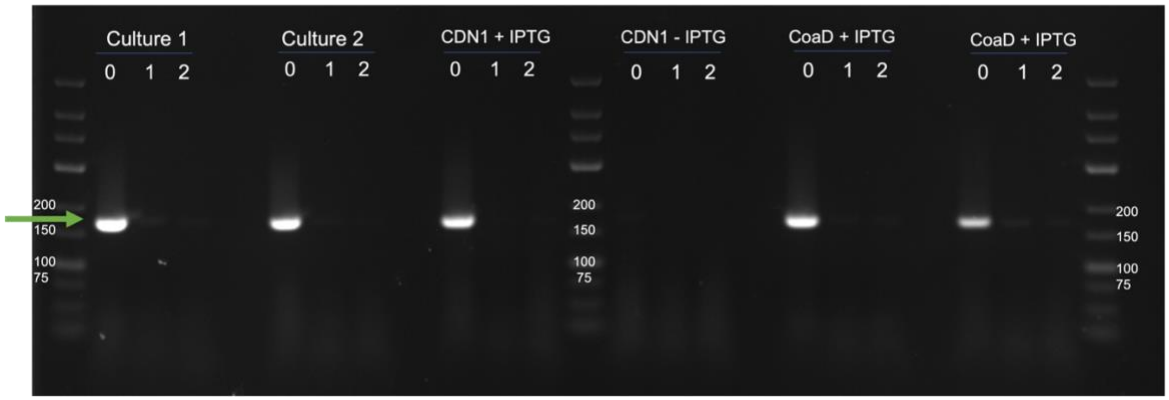


Figure 4.2. TURBO DNase treatments remove contaminating genomic DNA from RNA isolation preps. 30 cycles of PCR were performed with 16S RNA primers after 0, 1, and 2 TURBO DNase treatments and run on a 1% agarose stained with ethidium bromide and visualized with UV-Vis. The predicted band size for PCR products is ~167 bp, as indicated by the green arrow. CDN1 and CoaD are plasmids with inducible expression of PPAT, an enzyme that may be responsible for post-transcriptional capping of CoA-RNAs. CDN1 plasmid also expresses two known RNAs at high concentrations. The cultures were grown from cells containing plasmids with inducible expression of PPAT. +/- IPTG indicates presence or absence of induction step. Absence of visible bands in lanes with 1 and 2 TURBO DNase treatments indicated that no contaminating DNA remained to serve as a template.

Step 3: Generating and adding reference RNAs as controls for HTS.

Following DNase treatment, six reference RNAs (Figure 4.3A) were added into each sample. Half of the reference RNAs were "positive" controls containing a sulfur from a CoA cap generated by priming in vitro transcription reactions with dpCoA and purified on an APM gel to separate CoA-RNAs from ATP-RNAs (Figure 4.3B). The control RNAs were added in quantities of 1, 10, 100 fmol to RNA isolated from 20 x 10^8 cells to serve as references for RNA quantities during the HTS data analysis. We used very small quantities of control RNAs deliberately with the intention of establishing a limit of detection for this CoA Capture Seq method. For instance, if we were able to detect reference RNAs spiked in at 1 fmol and higher, this would provide an approximate limit of detection for this method. Additionally, if no cellular CoA-RNAs can be detected using this method (but our reference RNAs can be) it may indicate that either CoA-RNAs do not exist *or* they are present in quantities smaller than 1 fmol and are therefore beyond our limit of detection.

The reference RNAs have identical sequences and lengths (Table 4.1) with the exception of a 6 nt index. By minimizing the differences between each reference RNA, the biases experienced during the CoA Capture Seq method are expected to be similar for each reference RNA. For example, if the reverse transcription and PCR are biased against some sequences or structures, the differential bias between each reference RNA should be negligible due to their similarity in sequence, size, and structure. Therefore any differences observed in the reference RNAs during HTS analysis will be a result of handling (e.g. APM

partitioning). Also, the positive (CCS1-3, CoA-containing) and negative (CCS4-6, 5’ ATP) reference RNAs are intended to serve as internal controls within the HTS data set. For example, within the APM-partitioned samples, we expect the positive RNA controls to be enriched and the negative RNA controls to be deplete relative to the un-partitioned, total RNA samples. Overall, we were able to generate the positive and negative RNA controls (transcription gel not shown) and were able to add them into our isolated RNA samples to serve as internal controls for HTS.

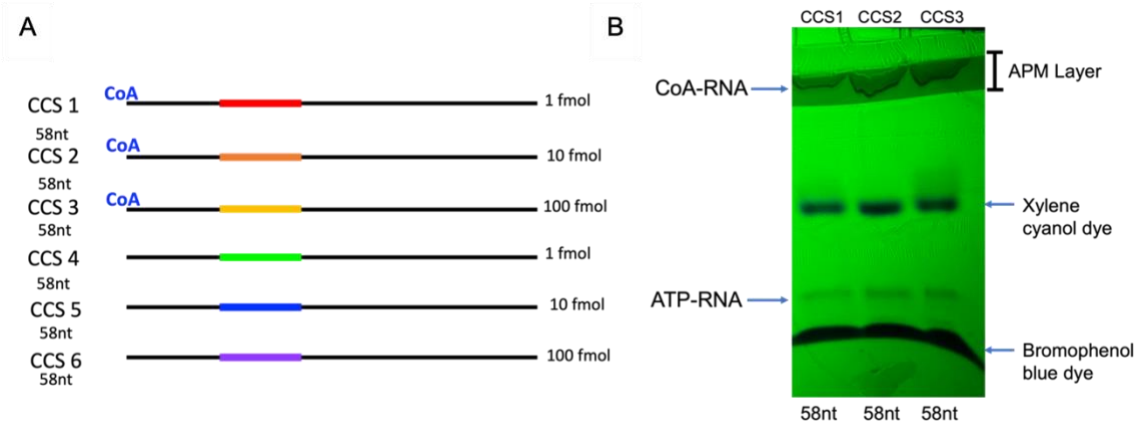


Figure 4.3. Reference RNAs for CoA Capture Seq. A. Six reference RNAs for CoA Capture Seq Method. CCS 1-6 are all identical in sequence and length with the exception of a 6 nt index (indicated by varying colors). Reference RNAs were added in at staggered concentrations indicated on the right. **B.** APM gel for positive CoA reference RNAs. Imaged by UV-Vis. CoA-RNAs (visible in the APM layer) were excised and precipitated.

Name:	Sequence 5' → 3':
CCS-1	AGGAGUCGUGCUGCCGCGAACUCCUCUACCAACAAGCGCG CGAUCCUCGCCAAGUAG
CCS-2	AGGAGUCGUGCUGCCGCGUGACAGUCUACCAACAAGCGCG CGAUCCUCGCCAAGUAG
CCS-3	AGGAGUCGUGCUGCCGCGGAAGUCUCUACCAACAAGCGCG CGAUCCUCGCCAAGUAG
CCS-4	AGGAGUCGUGCUGCCGCGACUCCAUCUACCAACAAGCGCG CGAUCCUCGCCAAGUAG
CCS-5	AGGAGUCGUGCUGCCGCGCGUAGGUCUACCAACAAGCGCG CGAUCCUCGCCAAGUAG
CCS-6	AGGAGUCGUGCUGCCGCGUACUUUUCUACCAACAAGCGCG CGAUCCUCGCCAAGUAG
CCS Forward Primer	GCG**TAATACGACTCACTATT**AGGAGTCGTGCTGC
CCS Reverse Primer	CTACTTGGCGAGGATCGCGCG

Table 4.1. Reference RNA sequences. Sequences and primers for the reference RNAs. Underlined and red portions are the 6 nt indeces used to differentiate the RNAs during HTS analysis. Bolded and underlined portion of the forward primer is the T7 Type II promoter that encodes for a +1 A.

Step 4: Size separation of isolated RNA by a molecular weight cutoff filter.

When the Liu group first identified metabolite-linked RNAs, they included a size exclusion step excluding RNAs larger than 300 nt (2). Thus, although RNAs larger than 300 nt may carry a CoA cap, it has been previously observed that RNAs 300 nt and smaller already carry CoA caps. Therefore, when designing our CoA Capture Seq, we incorporated a size exclusion step to focus on 'smaller' RNAs. A 300 nt-long RNA is approximately 97,000 kDa, therefore we used a 100 kDa molecular weight cutoff filter (as described in the materials and methods) as our method of size exclusion. 'Large' and 'small' size fractions were nanodropped following size fractionation to confirm presence of RNAs in both fractions. Additionally, RT-PCR was performed using primers for RNAs that were expected to fractionate into either the large (recA, ~1000 nt) and small fractions (gadY, ~100 nt) (data not shown). Generally, we observed successful size fractionation with very minimal large RNA contamination into small RNA fractions as indicated by faint PCR product bands from RT-PCR (data not shown).

Step 5: Cysteamine de-acylates acyl-CoA-RNAs and allows them to be partitioned by APM gel.

In addition to being a 5' cap of RNAs, CoA is more famously known as a protein coenzyme, especially within the context of fatty acid synthesis and oxidation and the citric acid cycle. Within a metabolic context, acetyl CoA is the true star, delivering acetyl groups to the citric acid cycle for energy production in cells. Interestingly, when the Liu group

first identified CoA-RNAs in cells, they also identified succinyl-CoA-RNAs and acetyl-CoA-RNAs (2). While identifying these acyl-CoA-RNA derivatives is of great interest, the acyl-group attachment to the sulfur of CoA poses a problem: acylated-CoA-RNAs do not have a free sulfur to interact with the mercury layer of an APM gel and therefore cannot be partitioned. However, we developed a method for deacylation using cysteamine. Cysteamine can be easily generated in large quantities by reducing cystamine with tris(2-carboxyethyl)phosphine (TCEP) (Figure 4.4A). Once generated, cysteamine can be easily reacted with acetyl-CoA-RNAs to transfer the acyl group to cysteamine, leaving the sulfur of the CoA free to interact with the APM layer (Figure 4.4B). This deacylation method was tested by former undergraduate student Matt Lichte. RNAs were generated by *in vitro* transcription with dephosphorylated succinyl-CoA as the NCIN to generate succinyl-CoA-RNAs, and samples were run on an APM gel with and without prior deacylation treatment (data not shown). We observed more RNA in the APM layer from deacylated samples as compared to untreated succinyl-CoA, indicating that treatments with cysteamine were successful in removing acyl groups to free CoA's sulfur to interact with mercury in the APM layer. We expect that it will be possible to elucidate the identities of acyl-CoA-RNAs from CoA-RNAs by looking at the relative enrichment and depletion of individual sequences in HTS data between treatment samples for de-acylated and acylated (no treatment) APM-partitioned samples.

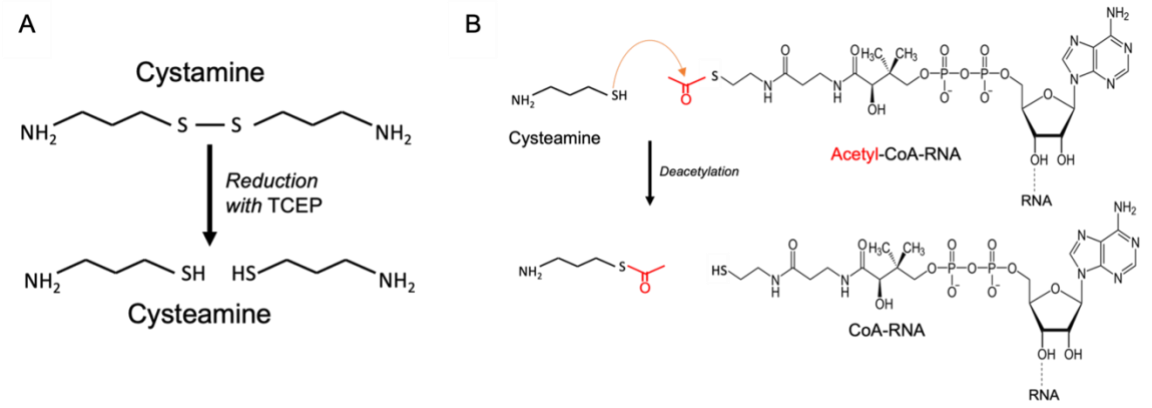


Figure 4.4. Deacylation of acyl-CoA-RNAs by cysteamine. A. Cystamine is reduced with TCEP to generate cysteamine. **B.** Cysteamine is used to deacylate acetyl-CoA-RNAs. Acetyl group (red) is transferred to cysteamine leaving the CoA sulfur free to interact with mercury in the APM layer during the subsequent partitioning step.

Step 6: Ligation of pre-adenylated sequencing adapters to RNA is inefficient and biased against RNA structure.

Cys-tRNA poses a challenge for the CoA Capture Seq method. It is relatively abundant in cells and carries a sulfur from the Cysteine amino acid. As a result, Cys-tRNAs can interact with mercury in the APM layer and be partitioned alongside CoA-RNAs. Cys-tRNAs are so abundant relative to CoA-RNAs that when partitioned and sequenced alongside other CoA-RNAs, the Cys-tRNAs will likely make up most if not all of the HTS output, effectively washing out any CoA-RNA signal. Thus, it is critical that during our CoA Capture Seq method the tRNAs be excluded from HTS.

Ligating an adenylated sequencing adapter to the 3' end of our RNA samples using T4 RNA ligase 2 (truncated, K227Q) (14, 15) solves the Cys-tRNA problem because ligation to the 3' end of RNAs is only successful for RNAs without blocked ends. Therefore, due to Cys-tRNA's 3' blocked end, no adapters are expected to be ligated precluding Cys-tRNA from being sequenced. Also, we make use of the conserved regions of the sequencing adapter to anneal the same reverse primer to the different cellular RNAs during reverse transcription (Figure 4.5A and Table 4.2).

However, before the adapters can be ligated to the RNA using T4 RNA Ligase 2, they must first be pre-adenylated (Figure 4.5A). Using pre-adenylated adapters makes the ligase reaction ATP-independent, which in turn prevents non-specific ligation. Buying pre-

adenylated adapters is somewhat costly; however, there are established methods for adenylating DNA (16, 17), which we adopted for our CoA Capture Seq method. Adapters can be pre-adenylated through a polynucleotide kinase treatment and incubation with T4 RNA ligase I. It's important to note that the 3' end of the adapters have a C3 spacer to prevent self-ligation or adapter-adapter ligation. Once pre-adenylated, the adapters can be ligated to DNA or RNA when incubated overnight at 16°C with T4 RNA ligase 2 (truncated, K227Q). We use a total of 8 adapters which are identical except for a six nt index region which serves as a barcode, allowing us to distinguish between different samples or partitions.

Although seemingly straightforward, the strategy of ligating pre-adenylated adapters to RNA proved to be quite challenging and produced the greatest number of roadblocks for the CoA Capture Seq method. Despite extraordinary troubleshooting, we frequently observed low ligation efficiency of our adapters to RNA templates. The pre-adenylation reaction to generate adenylated adapters was reported to have nearly 100% adenylation efficiency (18, 19). Although we used the same materials and methods outlined, it is possible that the low ligation efficiency between adenylated adapter and RNA template is not a complete reflection of poor ligation but also a result of poor adenylation. Incomplete pre-adenylation may occur during the reaction resulting in a mixed population of both adenylated and non-adenylated adapters. Obviously, adapters that are not adenylated are incapable of ligation, which in turn affects RNA-ligation efficiency.

In Figure 4.5B, despite having a 1:1 ratio of adapter to RNA, the RNA is never fully ligated. Initially, it was unclear whether the low efficiency of ligation is a result of poor adenylation or simply low ligation reactivity. Setting up ligation reactions with adenylated adapters in great excess (10:1 vs 1:1) did slightly improve ligation efficiency (Figure 4.5C). If we assume only 25% successful adenylation -- a drastically lower efficiency than what is reported – a 4:1 (adapter : RNA) ratio for ligation should be sufficient to generate ligated product if robust ligation is occurring. However, as shown in Figure 4.5C, even with 10:1 ratio large quantities of unligated RNA are observed. Therefore, although incomplete adenylation may be contributing to the overall problem, it is likely that the majority of the issue arises during the ligation step. Furthermore, other groups have reported many problems associated with this particular ligation step, affirming that ligation is the probably culprit (20–23).

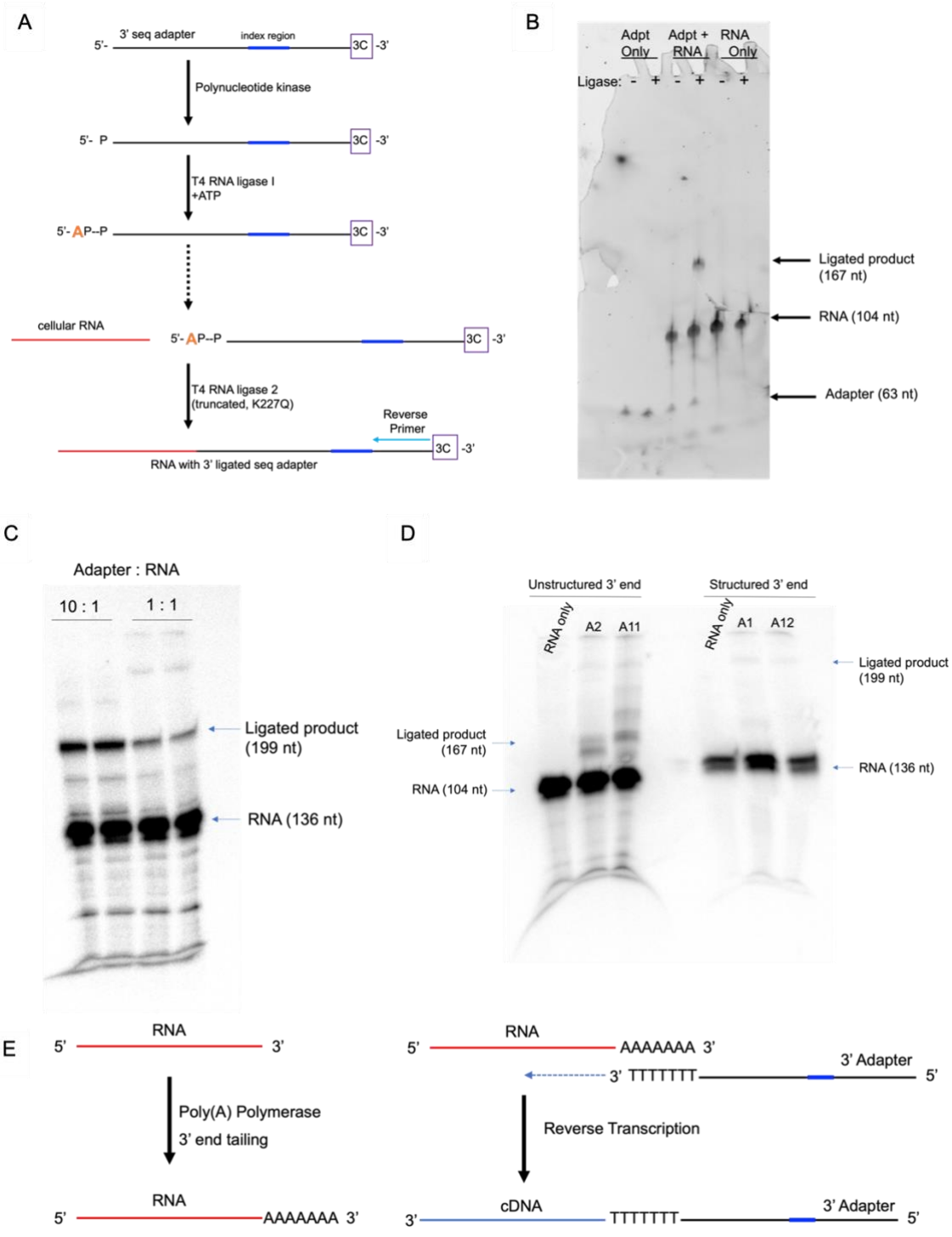

A
3' seq adapter
index region
5'-
3C
-3'
Polynucleotide kinase
5'- P
T4 RNA ligase I
+ATP
5'- AP--P
cellular RNA
T4 RNA ligase 2
(truncated, K227Q)
Reverse
Primer
RNA with 3' ligated seq adapter
B
Adpt
Only
Adpt +
RNA
RNA
Only
Ligase: - + - + - +
Ligated product
(167 nt)
RNA (104 nt)
Adapter (63 nt)
C
Adapter : RNA
10 : 1
1 : 1
Ligated product
(199 nt)
RNA (136 nt)
D
Unstructured 3' end
Structured 3' end
RNA only
A2
A11
RNA only
A1
A12
Ligated product
(199 nt)
Ligated product
(167 nt)
RNA (136 nt)
RNA (104 nt)
E
RNA
5'
3'
Poly(A) Polymerase
3' end tailing
RNA
5'
AAAAAAA 3'
RNA
5'
AAAAAAA 3'
3' Adapter
3' TTTTTTT
5'
Reverse Transcription
cDNA
3'
TTTTTTT
3' Adapter
5'

Figure 4.5. Pre-adenylating adapters and ligating them to RNA. A. Schematic overview for pre-adenylating 63 nt 3' end sequencing adapters. The 3' end of the adapters have a C3 spacer (indicated by 3PC in purple) to prevent self-ligation or adapter-adapter ligation. Adapters are differentiated by an internal 6 nt index indicated in blue. Adenylated adapters are ligated to cellular RNA (red). Reverse primer (light blue) for reverse transcription binds to the adapter. **B.** Adenylated adapter ligation to RNA. RNA and adenylated adapter do not self-ligate as indicated by the pairs on lanes on the far left and far right, respectively. 50 pmol of RNA and adenylated adapter (1:1) were used. RNA used was from Paige Gruenke and is predicted to have an unstructured 3' end. **C.** RNA was radiolabeled with γ^{32}P ATP Ligation reactions were tested in duplicate with the indicated ratios of adenylated adapter : RNA. **D.** Ligations were set up with (1:1) ratios of adenylated adapter : RNA. RNA was radiolabeled with γ^{32}P ATP. The same RNA from B was used as the unstructured 3' end RNA. The structured 3' end RNA has a hairpin stem loop from SHAPE cassette at its most 3' end (stable structure). Adapter used was indicated by "A" followed by a number to indicate the index sequence. Adapters are identical except for a 6 nt index. Products from (B-D) were run on a 5% PAGE gel, stained with ethidium bromide, and visualized on a Typhoon FLA 9000. **E.** Schematic overview for an alternative method of adding the 3' sequencing adapter without ligation. This method first polyadenylates the 3' end of RNA (right). Next the polyadenylated RNA is reverse transcribed with a 3' adapter primer including a 3'-oligo-dT region.

It has been reported that adenylated-adapter ligation is highly biased against RNAs with structure on their 3' ends (24, 25). We also observed this bias by directly comparing ligation of an RNA with no predicted 3' structure and an RNA with a stable hairpin stemloop structure derived from SHAPE cassette on its 3' end. As shown in Figure 4.5D, the unstructured 3' end RNA has observably better ligation than its structured counterpart. Although ligation was improved slightly by increasing the amount of adenylated adapter used during ligations, the ligation of the 3' end sequencing adapter remains our greatest challenge with few existing solutions.

In an attempt to bypass this bottleneck, we also tried an alternative method that omitted a ligation step (26). In this method, following partitioning by APM gel, RNAs were polyadenylated using *E. coli* poly(A) polymerase (New England Biolabs) and 3' sequencing adapters were annealed through the use of an oligo-dT region (Figure 4.5E). RNAs were subsequently reverse transcribed and amplified by PCR. Although this method seemed promising, there were some key problems. First, by omitting the 3' adapter ligation step, Cys-tRNAs is expected to be partitioned alongside other sulfur-RNAs and likely make up the majority of the HTS signal. Second, this particular method required more primers and led to a much higher background signal in the HTS data from primer-dimers and non-specific PCR product. Overall, methods to avoid the ligation step generated more problems than solutions and resulted in low quality HTS data which could not even be analyzed. Unfortunately, despite our troubleshooting, the ligation step remains a bottleneck in the

CoA Capture Seq method because it limits the number of RNAs that can be observed from the sequencing data and introduces bias against RNAs whose 3’ end have structure, leading some sequences in the HTS populations to be severely underrepresented.

Name:	Sequence 5' → 3':
16s forward primer	GCCTTCGGGTTGTAAAGTACTTTCAGC
16s reverse primer	TGCGTGCGCTTTACGCC
HTS FWD Primer + dG Region	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCT**GGGCGGG**
HTS Reverse Primer	CAAGCAGAAGACGGCATACGAGAT
5' Universal HTS Adapter	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCT
3' Indexed HTS Adapter	CAAGCAGAAGACGGCATACGAGAT**NNNNNN**GTGAC TGGAGTTCAGACGTGTGCTCTTCCGATC

Table 4.2. Sequences for PCR primers and HTS adapters. 16s primers were used in PCRs before and after TURBO DNase treatments. Sequences for the high-throughput sequencing primers used to append the Illumina adapters and their respective sequencing indices. We used the NEBNext Index (1-27) Primers for Illumina. 27 adapters were pre-adenylated (corresponding to the 27 indices) and ligated to the 3' end of RNA. During the RT step, the HTS reverse primer bound to the ligated adapter. The index region of the adapters is indicated by the six red N region in the reverse primer 2 for HTS sequence. Index sequences are from the instruction manual for the NEBNext Multiplex Small RNA Library Prep Set 1, Set 2, Index Primers 1-48 and Multiplex Compatible (https://rb.gy/0dbqe9). The HTS forward primer, also used during the RT step, has a poly G region (in blue) which is used to bind to the poly C region the RT adds to the 3' end of the first strand.

Step 7: APM gels can be used to separate sulfur-containing RNAs from non-sulfur RNAs.

A key component of this project is partitioning of CoA-RNAs from other RNAs. Our method to partition CoA-RNAs takes advantage of the free sulfur on Coenzyme A. Mercury is a metal ion which forms a coordinate covalent bond with sulfur. We generated an [(N-Acryloylamino) Phenyl] Mercuric Chloride (APM) stock solution as previously described (27–29) and used it to prepare three-layered PAGE gels whereby the middle layer contained APM. APM gels are a standard method for partitioning sulfur-containing RNAs and have been used successfully by the Burke group (30–34). Incorporating APM into the middle layer of PAGE gels allows sulfur-containing RNAs (e.g. CoA-RNAs) to bond to the mercury in turn reducing their migration through the gel (Figure 4.6A). Meanwhile, the migration of non-sulfur RNAs is unaffected and they migrate through the mercury-containing layer, which allows for effective separation of sulfur and non-sulfur RNAs.

To directly test whether APM gels were an effective method to partition CoA-RNAs from other RNAs, we transcribed RNA from oligos containing a T7 type II promoter as previously described (1) to ensure into the +1 position would incorporate ATP or ATP nucleotide analogs such as dpCoA (Figure 4.6B). Transcriptions included dpCoA as the non-canonical initiator nucleotide and lowered ATP concentrations. This allowed dpCoA to compete with ATP for the +1 site to yield a mix of CoA-RNAs and ATP-RNAs. This

transcription mix was then run on a 3 layered APM PAGE gel to confirm APM gels were a reliable partition method for CoA-RNAs. As shown in Figure 4.6C, RNA transcribed in the presence of dpCoA produced CoA-RNAs that were captured in the APM layer, as indicated by a faint band at the top of the APM layer. In Figure 4.6D, radiolabeled RNAs that were being selected to cap themselves with phosphopantetheine to form CoA-RNAs (see Chapter 5) were reacted over a time course and run on an APM gel. The phosphorimage of the APM gel shows that the reacted CoA-RNAs remain trapped within the APM layer while non-sulfur RNAs (ATP-RNAs) continue their migration into the third (non-Hg) layer of PAGE. Overall, these data confirmed that APM gels were a reliable method for partitioning sulfur-containing RNAs from total RNAs.

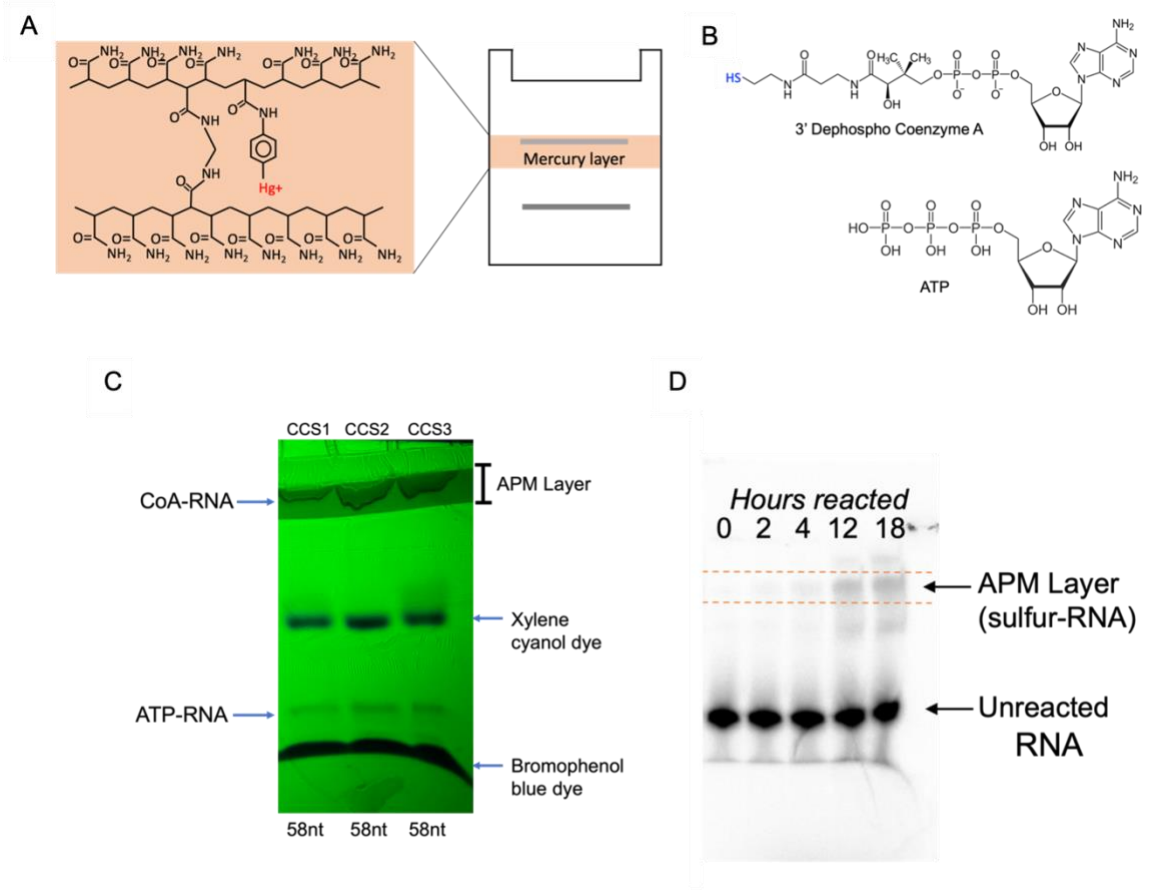


Figure 4.6. APM Gels partition CoA-RNAs from total RNA. A. Schematic of a three-layered APM gel (right) and a depiction of a one APM unit (left) with mercury shown in red. Sulfur-containing RNAs interact with mercury and remain in the APM layer and non-sulfur RNAs migrate through the APM layer. **B.** Chemical structures of 3' dpCoA and ATP. The free sulfur of dpCoA is shown in blue. DpCoA is an ATP analog and is used as a non-canonical initiator nucleotide during transcriptions to generate CoA-RNAs. **C.** UV-Shadow of an APM gel. Three dpCoA transcriptions were set up using the reference controls. CoA-RNA can be observed at the top of the APM layer and ATP-RNA product

can be observed between the dye fronts. **D.** Phosphorimage of an APM gel. Gel samples were from round 7B of a selection to find RNAs that could self-cap with phosphopantetheine to become CoA-RNAs (see Chapter 5). RNA was radiolabeled with α^{32}P dCTP and reacted with phosphopantetheine over a time course and increasing amounts of CoA-RNA was formed and captured in the APM layer.

Step 8: RT-PCR prepared samples for HTS.

Following the sulfur-partitioning step, RNAs were reverse transcribed and amplified by PCR and submitted for high-throughput sequencing. However, because the identities of the CoA-RNA sequences (and total RNA) were variable and unknown, a standard reverse transcription reaction would not be sufficient for our purposes because unlike the adapter ligated 3' ends, the 5' ends of the RNA templates were individual and unknown preventing amplification by PCR. In recent years, several groups have made use of polymerases with terminal transferase activity to append 5'-adapters to their RNA templates for HTS (26, 35). Therefore, we used SMARTScribe (Takara Bio), a Moloney Murine Leukemia Virus (MMLV) reverse transcriptase derivative with terminal nucleotidyltransferase activity that adds additional nucleotides (primarily dC) that are not encoded by the template to the 3' end of the first strand (cDNA) (Figure 4.7A). We took advantage of this unique feature by incorporating a dG region (GGGCGGG) on the 3' end of the 5' sequencing primer (Table 4.2), which allowed binding to the 3' dC region of the cDNA templates. A single C was added before and after 3 consecutive G's in the dG region of our primer to facilitate chemical synthesis. Following the RT step, the cDNA product was amplified by PCR using primers that annealed to the 5' and 3' sequencing adapters.

Because the isolated RNAs are different sizes, we did not expect to see a single band after PCR. However, due to the size exclusion step whereby we removed RNAs of ~300 nt or greater, there should not be any PCR products larger than ~500 base pairs (larger size due

to the appended sequencing adapters). As shown in Figure 4.7B, there are no discrete bands after PCR, as expected. Surprisingly, there was no obvious difference between the total cellular RNA and sulfur-RNA samples. Considering that the total RNA samples have a much higher concentration of RNA compared to partitioned sulfur-RNA (where all RNAs without a sulfur are removed), we expected to see more visible PCR product in the total RNA samples. However, the absence of obvious PCR bands or smears may be a reflection of a poor ligation efficiency. Recall that the reverse primer binds the 3' adapter (which is ligated to RNAs) during first strand synthesis in the first part of the RT step. However, if the ligation reaction failed for the majority of RNAs, then very little product would be generated and amplified through RT-PCR. This would result in very little PCR product being formed, as is observed in Figure 4.7B.

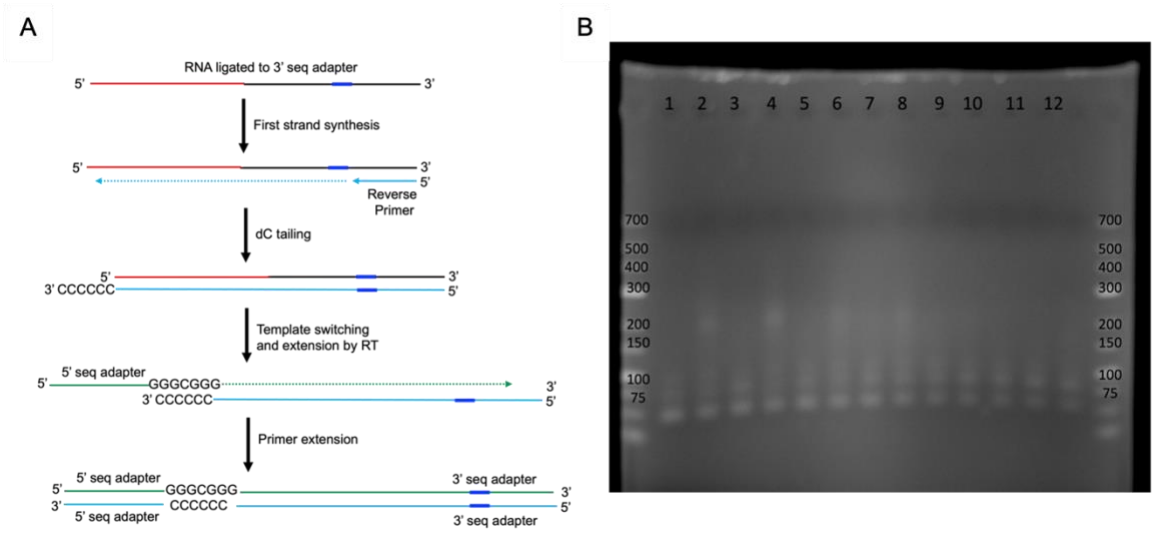


Figure 4.7. Reverse transcription and PCR to prepare cellular RNAs for high-throughput sequencing. A. Schematic overview of reverse transcription step using SMARTScribe RT. Cellular RNA (red) ligated to the 3' sequencing adapter from **Step 6**. is used as template for the RT reaction. A reverse primer binds to the 3' conserved region of the 3' seq adapter to synthesize the first strand. SMARTScribe adds an untemplated polyC region to the 3' end of the first strand. Making use of the dC region on the first strand, a forward primer with the 5' seq adapter and a dG region (Table 4.2) anneals to the first strand. Following template switching and extension, cDNA (bottom) is used as template for PCR. **B.** PCR products were run on a 1% agarose gel stained with ethidium bromide and visualized by UV-Vis. Lanes 1, 3, 5, 7, 9, 11 are total RNA and lanes 2, 4, 6, 8, 10, 12 are sulfur-RNA partitioned by APM gel. Samples were derived from 6 different bacterial cultures grown under varying conditions, split into total or sulfur samples (eg. lane 1-2 originated from the same culture).

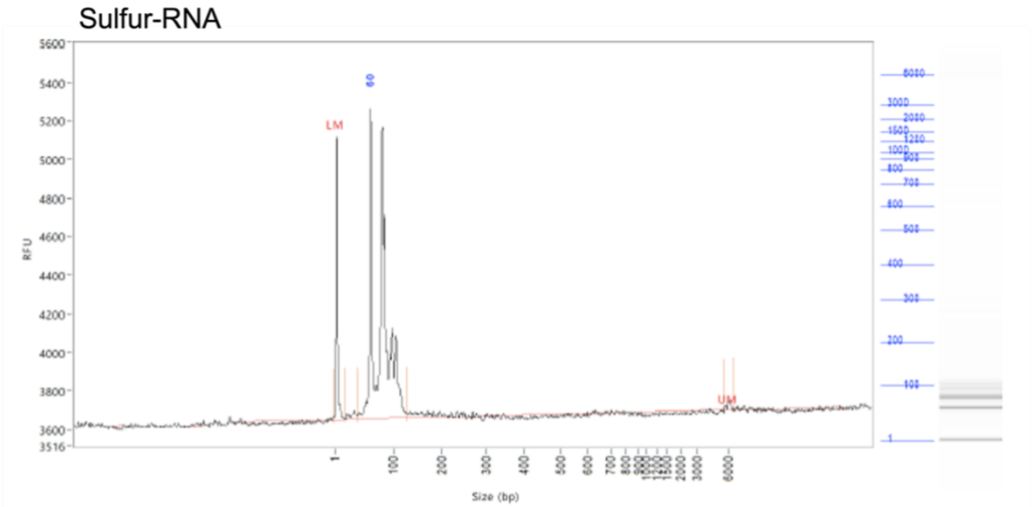
Step 9: Fragment analysis and high-throughput sequencing data.

In preparation for sequencing our samples, the University of Missouri DNA Core used an ABI 3730x1 DNA Analyzer to size the DNA fragments in each sample. As shown in Figure 4.8A-B, the fragment analysis indicated that samples fell within the expected size range (100-500 bp). Despite no clear differences being detectable from the PCR product run on the agarose gel, the fragment analysis did reveal greater diversity of sizes and slightly higher concentrations of nucleic acid in the total RNA sample as compared to the sulfur-RNA sample. As these were anticipated results, samples were then sequenced on an Illumina NextSeq 500 to generate 300 bp paired end reads.

In total, three different iterations of the CoA Capture Seq were performed and the respective samples were sent for HTS. In the first iteration the adapter ligation step had undergone virtually no optimization which ultimately impacted the amount of RNA that was reverse transcribed and amplified by PCR. As a result, the HTS data that came back from the first iteration was almost exclusively made up of primer-dimer and non-sensical sequences. We were unable to identify any sequences matching our reference RNAs either. Thus, we deduced that the first iteration of the CoA Capture Seq had failed and needed further optimization. For the second iteration of the CoA Capture Seq, we believed the adapter ligation has been the primary culprit of our problems in the first iteration, and so we used an alternative method that polyadenylated the RNA sequences and used a reverse primer with an oligo dT region to bind during the RT step. Unfortunately, the HTS data

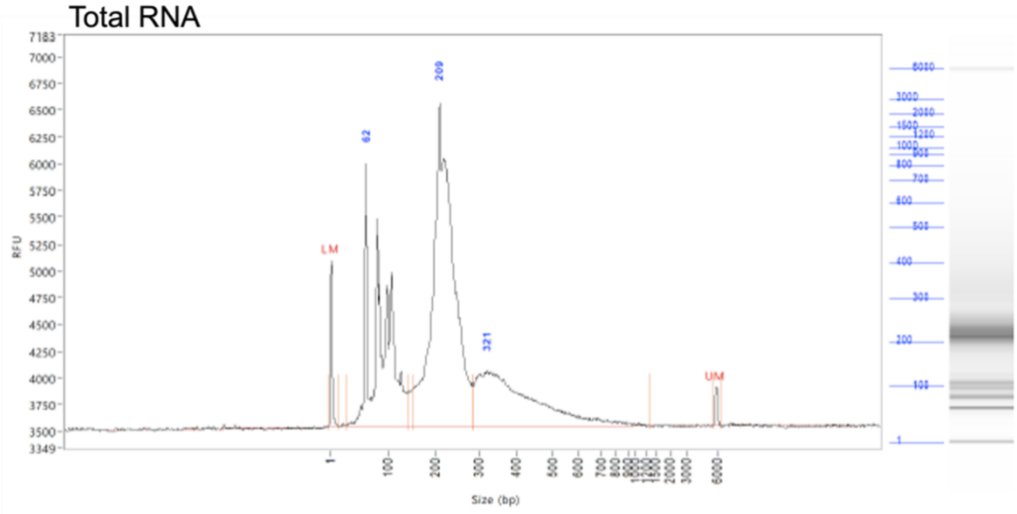
that came back from this iteration did not contain any useful data related to cellular RNAs. Once again, we saw a large number of primer-dimers and could not identify any of our reference RNAs, despite increasing the quantities in which they had been added by 10 fold for this iteration. In our third iteration of the CoA Capture Seq, we focused more on trying to optimize the adapter ligation step, but to little avail. Unfortunately, the data from this high-throughput sequencing run was not analyzable. In short, the quality of data was so low that it could not be interpreted in any way. We believe the poor HTS data is the result of two issues; first, low ligation efficiency of the 3' adapter to RNAs impacts the amount of RNA that can proceed through the CoA Capture Seq pipeline. Second, during the PCR amplification, with no real template available as a result of the ligation issues, non-specific PCR products are likely formed over the course of 30 cycles. To further complicate matters, because our PCR primers contain the sequencing adapters, the non-specific PCR products contain these sequences and thus can be sequenced as well. Thus, until further optimization of the adapter ligation and RT-PCR steps can be completed, this method will remain unreliable for isolating unknown sequences of CoA-RNAs from cells.

A



Peak	Size (bp)	Conc. (ng/uL)	From (bp)	To (bp)	Avg. Size (bp)	CV%	RFU	Corr. Peak Area
1	1 (LM)	0.0116	0	17	1	155.53	1466	9.651
2	60	0.6862	39	127	78	20.64	1610	47.649
3	6000 (UM)	0.0004	5517	6530	6002	3.34	49	0.308

B



Peak	Size (bp)	Conc. (ng/uL)	From (bp)	To (bp)	Avg. Size (bp)	CV%	RFU	Corr. Peak Area
1	1 (LM)	0.0116	0	14	1	172.00	1552	10.125
2	62	1.4560	30	143	91	25.67	2457	106.060
3	209	2.5816	153	287	218	12.66	3026	188.054
4	321	1.0224	287	1315	411	29.04	522	74.478
5	6000 (UM)	0.0020	5619	6455	5970	1.95	364	1.784

Figure 4.8. Fragment analysis of sulfur-RNAs and total-RNA. Representative fragment analysis of **A.** sulfur RNAs and **B.** total RNA. Each analysis was run with an upper (6000 bp) and lower (1 bp) size markers (indicated in red on the graphs) as controls for quantification of fragment sizes. Various peaks indicate various DNA fragment sizes (listed below the graph). Note that fragment analysis was performed for all samples.

DISCUSSION

Overall, we developed several key components of a method to capture and sequence known CoA-RNAs from cells. We also established a reliable method to de-acylate acyl-CoA-RNAs with cysteamine which is both affordable and fast. Notably, we also have a reliable method for the separation of sulfur-RNAs from total RNAs, a key cornerstone of this method. However, there still remains a few challenges in this method which impact the overall success of this method. Specifically, the low efficiency of ligation of pre-adenylated adapters to RNA templates in combination with already low CoA-RNA quantities drastically impacts the outcomes of high-throughput sequencing (HTS) results. The issues with ligation efficiency are exasperated by the RT-PCR steps, which rely on the presence of the 3' sequencing adapter for amplification. Without it, non-specific PCR primer-dimer type complexes form and make up the majority of the sequencing data.

Future work

Optimizing the ligation step may address these issues and allow for successful identification of cellular CoA-RNAs. Specifically, some studies have explored the impact of adapter sequences on ligation efficiency. Also, it has been reported that the two terminal bases on the 3' adapter dramatically impact ligation efficiency (23), instigating the use of randomized adapter pools as a means to reduce ligation bias from truncated T4 RNA ligase 2 (20–22). One possible future direction to optimize ligation would be to incorporate randomized adapter pools as a method to combat ligation bias. Although this would address

ligation bias, it does not change ligation efficiency which remains a fundamental problem. Perhaps studying ligation reaction parameters such as PEG%, adapter saturation, temperature, and reaction time could be useful to better optimize the ligation step; however, other groups have performed similar studies (36) from which we established our initial ligation protocol. Therefore, it seems unlikely that further optimization of these parameters would make a large impact as we already performed our reactions under the most optimal conditions.

Although the CoA Capture Seq method was originally designed to isolate naturally occurring cellular RNAs, we have since used a modified version to isolate and sequence specific CoA-RNAs expressed in cells (chapter 3, Figure 3.9). In this case, because we are interested in specific, known sequences of CoA-RNAs, the 3' adapter ligation step is unnecessary. Instead specific forward and reverse primers can be annealed directly during reverse transcription and PCR. In the future, this method can continue to be used for alternative purposes as described, or perhaps with thorough troubleshooting, the ligation efficiency can be resolved and the CoA Capture Seq can be used to identify naturally occurring CoA-RNAs.

MATERIALS AND METHODS

RNA transcripts

DNA templates were ordered from Integrated DNA Technologies and amplified by PCR using *Pfu* DNA polymerase. Sizes of double-stranded DNA (dsDNA) templates were confirmed by agarose gel electrophoresis. Each RNA was transcribed *in vitro* from the amplified PCR products using the Y639F T7 RNA polymerase (37), *in vitro* transcription buffer (1x = 50 mM Tris-HCl pH 7.5, 15 mM $MgCl_2$, 5 mM DTT, and 2 mM spermidine), and 2 mM each of ATP, UTP, GTP, CTP. Transcription reactions were incubated at 37°C for approximately 16 hrs. Transcriptions to generate CoA-RNAs used the same conditions with few changes: ATP concentration was reduced to 0.5 mM, reactions include 2 mM of 3' dephospho coenzyme A, and 5 mM tris(2-carboxyethyl)phosphine (TCEP) was used in place of DTT to prevent DTT from binding the APM layer of the PAGE gels which are run in subsequent steps. For Succinyl-CoA RNA transcriptions, 3'-dephospho Succinyl-CoA was generated by phosphatase treatment with FastAP (ThermoFisher Scientific) used according to manufacturer guidelines. Dephosphorylated Succinyl-CoA was separated from Succinyl-CoA by HPLC. Transcription reactions were incubated at 37°C overnight (approximately 16 hrs) and terminated by the addition of denaturing gel loading dye (90% formamide, 50 mM EDTA and 0.01% of xylene cyanol and bromophenol blue). Transcripts were subsequently purified by denaturing PAGE (5-8% TBE-PAGE, 8 M urea) or partitioned by APM gel when relevant. Bands corresponding to the expected product sizes were visualized by UV shadow, excised from the gel, and eluted by tumbling overnight at 4°C in 300 mM sodium acetate pH 5.4. Eluates were ethanol precipitated, resuspended in nuclease-free water, and stored at -20°C until further use. A NanoDropOne

spectrophotometer (Thermo Fisher Scientific) was used to determine specific RNA concentrations for all assays.

APM gels & elution

The [(N-Acryloylamino) Phenyl] Mercuric Chloride (APM) stock solution was made as previously described (13). To pour an APM gel, the first layer of polyacrylamide (about 20 mL) was poured in a gel casing standing upright. 1 mL of MilliQ water is added directly after pouring the bottom layer to create a smooth interface between layers. The first layer polymerizes for approximately 30 min. The second, APM containing layer consists of 1 mL of polyacrylamide, 1 μL TEMED, 10 μL 10% ammonium persulfate (APS), and 200 μL of APM stock solution. After removing the 1 mL of water from the gel casing, the APM layer is added following by a fresh 1 mL of water. After the gel to polymerizes for approximately 30 min, the excess water is removed and the final 10 mL layer of polyacrylamide is added and the wells are inserted. After polymerizing for 30 min the APM gel is ready to be run in 1X TBE at 30 watts.

We purified CoA-RNAs from an APM gels by first visualizing the bands by UV-shadow then excising them from the gel. Gel pieces were 'minced' into very small pieces and added to 1X APM elution buffer (0.5 M ammonium acetate, 0.5 M DTT, 10 mM EDTA) and tumbled overnight at 4°C. The gel slurry was then loaded into a pre-wetted 100kDa molecular weight cutoff filter (ThermoFisher Scientific) and spun at 14,000 xg for 15 min.

The columns were washed with 200 μL of 1X APM elution buffer and spun at 14,000 xg for an additional 15 min. The flow-through was collected and placed into a fresh tube. 1 μL of glycogen and 1 mL of cold ethanol to ethanol precipitate the CoA-RNA. RNA was resuspended in nuclease-free water and stored at -20°C until further use.

RNA isolation

HB101 K-12 strain *E. coli* cultures were grown in duplicate to serve as biological replicates. Starter 10 mL cultures were grown in 2xYT media in a 37°C incubator shaking at 250 rpm for ~16 hr. 1 mL of starter culture was used to start 25 mL cultures for RNA isolation. 25 mL cultures were grown in 2xYT media or an acid stress media (2xYT with lactic acid, pH 5.2) in a 37°C incubator shaking at 250 rpm and OD_{600} was measured frequently by a NanoDropOne spectrophotometer (Thermo Fisher Scientific). Cultures were grown to mid-exponential phase (OD_{600}: ~0.5) or stationary phase (OD_{600}: ~2.0). Using OD_{600} values, approximately equivalent numbers of bacterial cells were harvested by centrifugation at 4,000 xg at 4°C for 15 minutes. Pellets was resuspended in 5mL lysozyme buffer (50 mM Tris HCl pH 7.6, 250 mM NaCl, 0.1 mM EDT) and lysozyme (ThermoFisher Scientific) was added to a final concentration of 0.2 mg/mL and allowed to incubate for 10 min at room temperature. 3 volumes of TRIzol (ThermoFisher Scientific) was added to the total cell lysate and vortexed before incubating on ice for 5 min. 1/5 of the total TRIzol volume of cold (4°C) chloroform (ThermoFisher Scientific) was added and the sample was briefly vortexed to mix. Samples were centrifuged at 12,000 xg at 4°C

for 15 minutes. The top aqueous layer (~10 mL) was carefully removed to a fresh 50 mL falcon tube and 2 volumes of cold chloroform were added to remove any leftover phenol. The samples were vortexed then centrifuged at 12,000 xg at 4°C for 15 minutes. The top aqueous layer was removed and an equal volume of cold isopropanol (ThermoFisher Scientific) was added and the sample was vortexed. Samples were incubated on ice for 5 min before being centrifuged at 16,000 xg at 4°C for 30 minutes. The supernatant was discarded and the pellet was washed with cold 70% ethanol and centrifuged at 16,000 xg at 4°C for 30 minutes. Pellets containing the RNA were dried and resuspended in 1 mL of MilliQ water and stored at -20°C until further use.

TURBO DNase treatment

Samples of isolated RNA with nucleic acid concentrations greater than 20 μg/mL were diluted to 10 μg/50 μL. 5 μL of sample were set aside after 0, 1, and 2 DNase treatments to be used for PCR as template. TURBO DNase (ThermoFisher Scientific) reactions with up to 10 μg of nucleic acid were assembled in 1X TURBO DNase buffer with 2 U of TURBO DNase and incubated at 37°C for 30 minutes. To stop the reaction, 5 μL of TURBO DNase "Inactivation Reagent" (ThermoFisher Scientific) was added and incubated at room temperature for 5 min with intermittent mixing to keep the reagent in solution. Samples were spun on a tabletop centrifuge for ~90 sec to separate the DNase treated samples (supernatant) and inactivation reagent. The supernatant was collected and

concentrated by ethanol precipitation. PCR to test the effectiveness of TURBO DNase treatments used 16S RNA primers and was setup as described above.

100kDa molecular weight cutoff filtration

To recover 'small' RNAs (300 nt and smaller), we used Amicon Ultra-0.5 Centrifugal 100KDa Molecular Weight Cutoff Filters (Millipore Sigma). Filters were pre-wet with 500 μL of water and spun at 14,000 xg for 10 min at room temperature. Samples were loaded into pre-wet filters and spun at 14,000 xg for 30 min at room temperature. After collecting the flow-through, the columns were washed with 400 μL of water and spun at 14,000 xg for 30 min at room temperature. Flow-through was pooled and concentrated into a smaller volume by ethanol precipitation.

Deacylation reactions by cysteamine

Cysteamed was prepared by incubating cysteamine (Millipore Sigma) with 1 M tris(2-carboxyethyl)phosphine (TCEP) in a 1:4 molar ratio at room temperature for 15 min. Isolated RNA was incubated with 10 mM cysteamine at room temperature for 30 minutes.

Pre-adenylation and ligation of HTS adapters

Adapter DNA oligos were ordered with a 3' C3 spacer blocking group and a 5' phosphate group. For adapter DNA oligos ordered without a 5' phosphate group, 50 pmol of DNA was added 5 μL of 10X T4 PNK reaction buffer (New England Biolabs), 0.5 μL 10 mM

ATP, and 2 μL T4 Polynucleotide Kinase (PNK) (New England Biolabs). PNK reactions were incubated at 37°C for 30 min and heat inactivated at 65°C for 20 min. The adenylation reaction was carried out in a 50 μL reaction volume containing 100 pmol of adapter DNA, 1X T4 RNA ligase buffer (New England Biolabs), 35% PEG8000, 1 mM ATP, and 30 units of T4 RNA ligase 1 (New England Biolabs). Adenylation was performed at 37°C for 1 hour and heat inactivated at 65°C for 20 min. Note: Adenylation reactions were also tested under alternate conditions where they were incubated at room temperature for ~16 hrs (less optimal conditions). Additionally, adapter adenylation was also performed using the 5' DNA adenylation kit (New England Biolabs) according to the manufacturer protocol, however we found the adenylation with the kit to be less affective.

Ligations were set up by mixing pre-adenylated adapter and template RNA with (concentration varied depending on ratio to RNA) 2 μL truncated T4 RNA Ligase 2 K227Q (New England Biolabs, Ipswich, MA), 10 μL of 50% PEG8000, and water into a 40 μL reaction. Ligation reactions were incubated at 25°C for 4 hours. Note: Ligation reactions were also tested under alternate conditions where they were incubated at 16°C for ~16 hrs (less optimal conditions). The enzyme was heat inactivated at 65°C for 20 min. However, these conditions were found to be less affective and were ultimately not used in the majority of experiments.

RT and PCR for HTS preparation

For first strand synthesis, up to 5 μg of RNA was mixed with 20 pmol of reverse primer to a final volume of 4.25 μL, heated to 70°C for 3 min and cooled to 42°C for 2 min. Then we added 5.75 μL of a master mix consisting of 1X first-strand buffer (Takara Bio), 0.5 μL of 100 mM DTT, 0.25 μL RNase inhibitor, 20 pmol dG forward primer, 1 μL 50x dNTP mix (10 mM each dATP, dGTP, dCTP, dTTP), and 1 μL SMARTScribe (Takara Bio). Reactions were incubated at 42°C for 90 min and terminated by heat inactivation at 68°C for 10 min. After inactivation, RT reactions were used as template for PCR. PCR was performed as described above for 30 cycles. Sequencing was performed on an Illumina NextSeq 500 (University of Missouri Genomics Technology Core) to generate 300 bp paired-end reads.

ACKNOWLEDGEMENTS

We would like to acknowledge Matt F. Lichte for his contributions to this work. We also thank Dr. Phuong Nguyen for her assistance in generating APM stocks.

REFERENCES

1. Huang, F. (2003) Efficient incorporation of CoA, NAD and FAD into RNA by in vitro transcription. *Nucleic Acids Res.* 10.1093/NAR/GNG008
2. Kowtoniuk, W. E., Shen, Y., Heemstra, J. M., Agarwal, I., and Liu, D. R. (2009) A chemical screen for biological small molecule-RNA conjugates reveals CoA-

linked RNA. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7768–7773

3. Chen, Y. G., Kowtoniuk, W. E., Agarwal, I., Shen, Y., and Liu, D. R. (2009) LC/MS analysis of cellular RNA reveals NAD-linked RNA. *Nat. Chem. Biol.* **5**, 879–881

4. Jiao, X., Doamekpor, S. K., Bird, J. G., Nickels, B. E., Tong, L., Hart, R. P., and Kiledjian, M. (2017) 5′ End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding. *Cell*. **168**, 1015-1027.e10

5. Kiledjian, M. (2018) Eukaryotic RNA 5′-End NAD+ Capping and DeNADding. *Trends Cell Biol.* **28**, 454–464

6. Winz, M. L., Cahová, H., Nübel, G., Frindert, J., Höfer, K., and Jäschke, A. (2017) Capture and sequencing of NAD-capped RNA sequences with NAD captureSeq. *Nat. Protoc.* **12**, 122–149

7. Walters, R. W., Matheny, T., Mizoue, L. S., Rao, B. S., Muhlrad, D., and Parker, R. (2017) Identification of NAD+ capped mRNAs in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 480–485

8. Bird, J. G., Zhang, Y., Tian, Y., Panova, N., Barvík, I., Greene, L., Liu, M., Buckley, B., Krásný, L., Lee, J. K., Kaplan, C. D., Ebright, R. H., and Nickels, B. E. (2016) The mechanism of RNA 5′ capping with NAD+, NADH and desphospho-CoA. *Nature*. **535**, 444–447

9. Julius, C., and Yuzenkova, Y. (2017) Bacterial RNA polymerase caps RNA with

various cofactors and cell wall precursors. *Nucleic Acids Res.* **45**, 8282–8290

10. Cahová, H., Winz, M. L., Höfer, K., Nübel, G., and Jäschke, A. (2015) NAD captureSeq indicates NAD as a bacterial cap for a subset of regulatory RNAs. *Nature*. **519**, 374–377
11. Takamura, Y., and Nomura, G. (1988) Changes in the intracellular concentration of acetyl-CoA and malonyl-CoA in relation to the carbon and energy metabolism of Escherichia coli K12. *J. Gen. Microbiol.* **134**, 2249–2253
12. Thauer, R. K., Jungermann, K., and Decker, K. (1977) Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol. Rev.* **41**, 100
13. Biondi, E., and Burke, D. H. (2012) Separating and analyzing sulfur-containing RNAs with organomercury gels. *Methods Mol. Biol.* **883**, 111–120
14. Persson, H., Søkilde, R., Pirona, A. C., and Rovira, C. (2017) Preparation of highly multiplexed small RNA sequencing libraries. *Biotechniques*. **63**, 57–64
15. Mann, D. G. J., King, Z. R., Liu, W., Joyce, B. L., Percifield, R. J., Hawkins, J. S., LaFayette, P. R., Artelt, B. J., Burris, J. N., Mazarei, M., Bennetzen, J. L., Parrott, W. A., and Stewart, C. N. (2011) Switchgrass (Panicum virgatum L.) polyubiquitin gene (PvUbi1 and PvUbi2) promoters for use in plant transformation. *BMC Biotechnol.* **11**, 72
16. Chen, Y. R., Zheng, Y., Liu, B., Zhong, S., Giovannoni, J., and Fei, Z. (2012) A cost-effective method for Illumina small RNA-Seq library preparation using T4 RNA ligase 1 adenylated adapters. *Plant Methods*. **8**, 41

17. Lama, L., and Ryan, K. (2016) Adenylylation of small RNA sequencing adapters using the TS2126 RNA ligase I. *RNA*. **22**, 155–161

18. Song, Y., Liu, K. J., and Wang, T. H. (2015) Efficient synthesis of stably adenylated DNA and RNA adapters for microRNA capture using T4 RNA ligase 1. *Sci. Rep.* **5**, 1–8

19. Zhelkovsky, A. M., and McReynolds, L. A. (2011) Simple and efficient synthesis of 5′ pre-adenylated DNA using thermostable RNA ligase. *Nucleic Acids Res.* **39**, e117–e117

20. Sorefan, K., Pais, H., Hall, A. E., Kozomara, A., Griffiths-Jones, S., Moulton, V., and Dalmay, T. (2012) Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence*. **3**, 4

21. Sun, G., Wu, X., Wang, J., Li, H., Li, X., Gao, H., Rossi, J., and Yen, Y. (2011) A bias-reducing strategy in profiling small RNAs using Solexa. *RNA*. **17**, 2256–2262

22. Zhang, Z., Lee, J. E., Riemondy, K., Anderson, E. M., and Yi, R. (2013) High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biol.* **14**, R109

23. Jayaprakash, A. D., Jabado, O., Brown, B. D., and Sachidanandam, R. (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res.* **39**, e141

24. Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y., and Robb, G. B. (2012) Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. *Nucleic Acids Res.* **40**, e54

25. Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J. T. G., Nusbaum, J. D., Morozov, P., Ludwig, J., Ojo, T., Luo, S., Schroth, G., and Tuschl, T. (2011) RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA.* **17**, 1697–1712

26. Turchinovich, A., Surowy, H., Serva, A., Zapatka, M., Lichter, P., and Burwinkel, B. (2014) Capture and Amplification by Tailing and Switching (CATS), Anultrasensitive ligation-independent method for generation of DNA libraries for deep sequencing from picogram amounts of DNA and RNA. *RNA Biol.* **11**, 817–828

27. Biondi, E., and Benner, S. A. (2018) Artificially Expanded Genetic Information Systems for New Aptamer Technologies. *Biomedicines*. 10.3390/BIOMEDICINES6020053

28. Igloi, G. L. (1988) Interaction of tRNAs and of Phosphorothioate-Substituted Nucleic Acids with an Organomercurial. Probing the Chemical Environment of Thiolated Residues by Affinity Electrophoresis. *Biochemistry*. **27**, 3842–3849

29. Rhee, S. S., and Burke, D. H. (2004) Tris(2-carboxyethyl)phosphine stabilization of RNA: Comparison with dithiothreitol for use with nucleic acid and thiophosphoryl chemistry. *Anal. Biochem.* **325**, 137–143

30. Saran, D., Held, D. M., and Burke, D. H. (2006) Multiple-turnover thio-ATP hydrolase and phospho-enzyme intermediate formation activities catalyzed by an RNA enzyme. *Nucleic Acids Res.* **34**, 3201–3208

31. Biondi, E., Nickens, D. G., Warren, S., Saran, D., and Burke, D. H. (2010) Convergent donor and acceptor substrate utilization among kinase ribozymes. *Nucleic Acids Res.* **38**, 6785–6795

32. Biondi, E., Poudyal, R. R., Forgy, J. C., Sawyer, A. W., Maxwell, A. W. R., and Burke, D. H. (2013) Lewis acid catalysis of phosphoryl transfer from a copper(II)-NTP complex in a kinase ribozyme. *Nucleic Acids Res.* **41**, 3327–3338

33. Burke, D. H., and Rhee, S. S. (2010) Assembly and activation of a kinase ribozyme. *RNA.* **16**, 2349–2359

34. Biondi, E., Maxwell, A. W. R., and Burke, D. H. (2012) A small ribozyme with dual-site kinase activity. *Nucleic Acids Res.* **40**, 7528–7540

35. Adamopoulos, P. G., Tsiakanikas, P., Stolidi, I., and Scorilas, A. (2022) A versatile 5′ RACE-Seq methodology for the accurate identification of the 5′ termini of mRNAs. *BMC Genomics*. 10.1186/S12864-022-08386-Y

36. Song, Y., Liu, K. J., and Wang, T. H. (2014) Elimination of ligation dependent artifacts in T4 RNA ligase to achieve high efficiency and low bias microRNA capture. *PLoS One*. 10.1371/journal.pone.0094619

37. Sousa, R., and Padilla, R. (1995) A mutant T7 RNA polymerase as a DNA polymerase. *EMBO J.* **14**, 4609–4621

Chapter 5: Probing the role of library design in SELEX for ribozymes

There are currently no plans to submit this work for publication, unless significant additional data is generated in the future.

ABSTRACT

Although CoA-RNAs were reported to exist in cells more than a decade ago, their identities and mechanisms of biogenesis remain unknown. Possible mechanisms to generate CoA-RNA have been tested and verified *in vitro* including co-transcriptional capping (1, 2), post-transcriptional capping by PPAT (Chapter 3), and self-capping by ribozyme catalysis (3, 4). However, the previous studies which selected and identified self-capping ribozymes were performed under non-biological conditions, and the studies which demonstrated *in vitro* co-transcriptional CoA capping utilized 3' dephospho CoA (dpCoA) concentrations more than ~50 fold greater than the predicted intracellular concentration of dpCoA. Thus it remained unclear if these mechanisms were feasible for *in vivo* generation of CoA-RNA. Here, we performed selections under challenging *in vivo*-like conditions to identify self-capping CoA ribozymes (CoAzymes) and RNAs which can serve as a substrate to be post-transcriptionally capped by PPAT under cellular conditions. Six libraries with varying degrees of structure were pooled to form a single starting library for the selection. Each library was designed with/without specific structural features which we hypothesized would provide an advantage in either selection trajectory. After 12 rounds of selection

with no significant increase in CoAzyme or PPAT capping activity, several rounds were sent for high-throughput sequencing (HTS). HTS data analysis revealed no convergence or enrichment of specific sequences or clusters of sequences, indicating the selection had failed. We speculated that the overly stringent *in vivo* like conditions were too challenging for the majority of active sequences especially at the start of the selection, which ultimately sabotaged any successful selection outcomes. Additionally, HTS data demonstrated highly inconsistent populations between selection rounds, which could be indicative of non-specific capture of random sequence or capture of non-specifically capped RNA sequences.

INTRODUCTION

Nucleotide cofactors such as cyclic adenosine monophosphate (cAMP), nicotinamide adenine dinucleotide (NAD+), and coenzyme A (CoA) play pivotal roles as signaling molecules, energy carriers, and enzyme cofactors. In addition to these functions, nucleotide analogs with free 3' hydroxyl groups (e.g., NAD+, FAD, and 3' dephospho CoA) were also found suitable as non-canonical initiator nucleotides (NCINs) for *in vitro* RNA transcription (5) to generate CoA-RNA, NAD-RNA, and FAD-RNA. Previous studies reported the presence of naturally occurring cofactor-linked RNAs in bacterial cells, in which NAD+, CoASH and acyl-CoAs were reported to be present in the most 5' position of transcripts (6, 7). Although the function and biogenesis of NAD-RNAs has been increasingly explored and understood (7–11), the identities, functional roles, and mechanism of biogenesis for CoA-linked RNA have not yet been explored.

CoA-RNA can be generated by several possible mechanisms. First, CoA-RNA can be made co-transcriptionally whereby 3' dephospho CoA (dpCoA) is incorporated into the +1 position of transcripts resulting in CoA-RNA. Although co-transcriptional generation of CoA-RNA has already been established *in vitro* (5), it is an unlikely mechanism within the context of *in vivo* biogenesis of CoA-RNA. Although dpCoA values in *E. coli* remain unknown, one study reported dpCoA intracellular concentrations to be 20 μM and NAD+ concentrations were 12 mM in the anaerobic bacterium *C. kluyveri* (12). Intracellular NAD+ levels in *E. coli* typically fluctuate between 4-7 mM, comparable to intracellular ATP levels (1-5 mM). However, assuming similar levels of dpCoA in *E. coli* as was observed in *C. kluyveri*, dpCoA would struggle to compete with ATP for the +1 spot of RNA transcripts difficult. Therefore, co-transcriptional capping is likely not a significant mechanism for generating cellular CoA-RNAs and for this project is of lesser interest. A second mechanism to generate CoA-RNAs is through post-transcriptional modification (Figure 5.1A). ATP-RNA transcripts can be post-transcriptionally capped with phosphopantetheine by phosphopantetheine adenylyltransferase (PPAT) to become CoA-RNAs. PPAT is an enzyme in the CoA biosynthesis pathways responsible for generating dpCoA from ATP and phosphopantetheine. Interestingly, pre-liminary data demonstrate PPAT's *in vitro* capping capabilities using ATP-RNA as a non-canonical substrate and possible *in vivo* capping of specific RNA substrates (unpublished data, Chapter 3 Figures 3.2 & 3.4). Thus, further understanding the properties (sequential, structural, chemical, etc)

of RNAs which can serve as capping substrates for PPAT is of great interest. A third mechanism for CoA-capping is self-capping ribozymes (Figure 5.1B). Catalytically active RNAs with a +1A can self-phosphopantetheinylate themselves as self-capping ribozymes, referred to as "CoAzymes", to become CoA-RNAs. Previous selections performed by the Huang group were able to identify CoAzymes capable of self-capping with phosphopantetheine to form CoA-RNAs (13). However, the CoAzymes identified by the Huang group require up to 5 mM Mn^{2+} and/or 5 mM Ca^{2+} to be active. *E. coli* cells have a free calcium concentration around 100 nM and Manganese is considered toxic. Therefore, these previously identified CoAzymes are unlikely to retain activity within *E. coli*'s cells and do not represent a suitable *in vivo* method of CoA-RNA biogenesis.

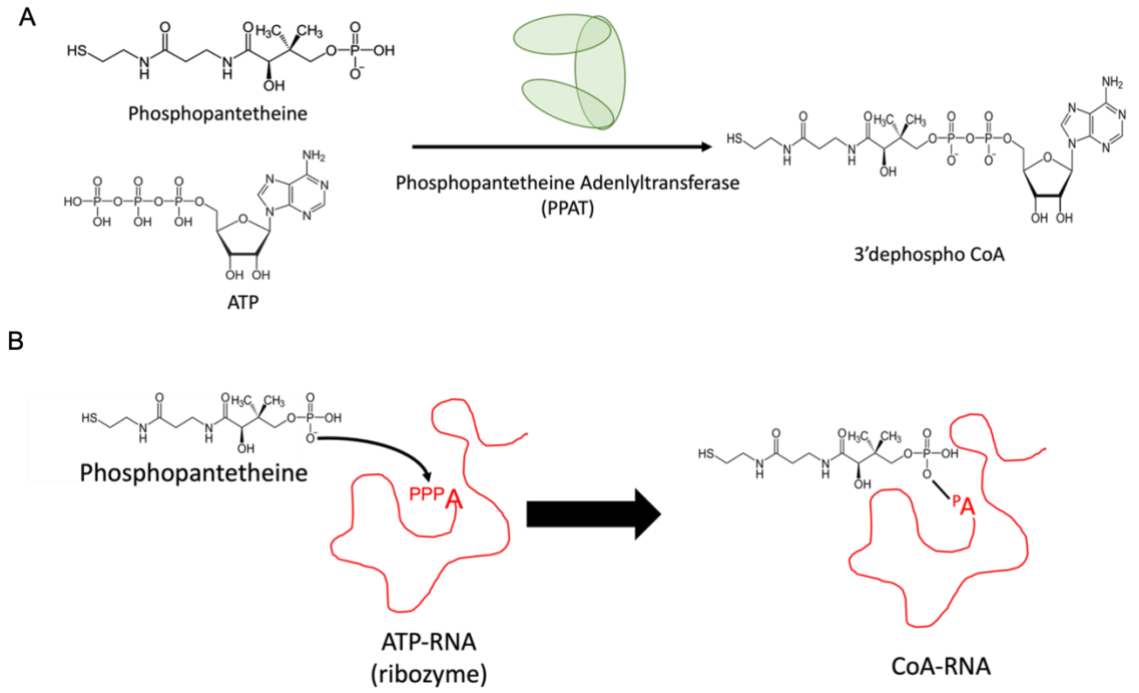


Figure 5.1. CoA capping by ribozymes and PPAT. (A) Schematic of the canonical reaction of PPAT. PPAT uses substrates ATP and phosphopantetheine to generate product 3' dephospho CoA. In place of its canonical substrate ATP, PPAT can also use ATP-RNAs to generate CoA capped RNAs. (B) Schematic of a CoAzyme (ribozyme) self-capping RNA. Phosphopantetheine is the substrate for CoAzymes to self-cap and become CoA-RNAs.

To gain further insight into self-capping ribozymes (CoAzymes) and which RNAs serve as the best substrates to be post-transcriptionally capped by PPAT, we performed a selection under *in vivo* like conditions to identify CoAzymes and RNAs capable of serving as PPAT capping substrates which will retain their functionality in cells. The selection conditions were carefully considered because selecting for RNAs in an environment that is vastly different from the intended reaction environment (e.g. bacterial cells), may result in reduced or no activity later on (14). Therefore, we used selection buffers and reactions times that were consistent with cellular conditions to enrich for sequences that would retain activity *in vivo.* Our selection buffers were had a physiological pH of ~7.5, with only biologically relevant ions present at their intracellular concentrations. Crowding reagents were also included in the selection buffer to mimic cellular crowding. Additionally, reaction conditions were performed at 37◦C to keep conditions as *in vivo* like as possible.

Although *in vivo* like conditions can be challenging for selections, we hypothesized that resulting RNAs would retain activity in cells compared to RNAs selected under 'standard' *in vitro* conditions. Here we laid the groundwork for a platform in which RNAs can be selected *in vitro* using *in vivo*-like conditions to generate functional RNA molecules which are likely to retain activity (binding, catalysis, etc) in cells. Six different libraries were intentionally designed with significant, moderate, or few structural features and represented various hypotheses for selection outcomes. Thus, we also discussed the use of intentional library designs for selections, possible mechanisms by which CoA-RNAs can

be generated (i.e. self-made or post transcriptional modification) intracellularly and *in vitro*, and how we gained insights from HTS data for improving future selection outcomes.

RESULTS

Library design

We aimed to further understand CoA-RNA biogenesis mechanisms by selecting for RNAs capable of self-capping (CoAzymes) and serving as substrates to be post-transcriptionally capped by PPAT. Careful consideration was taken to generate six different libraries with varying structural elements which were eventually pooled to form a single starting library (Figure 5.2A). Libraries used the same primer binding sites and they were the same length (Table 5.1). Also, to prevent the 3' end from interfering with predetermined structural scaffold motifs, a portion of SHAPE cassette sequence was used for the 3' end sequence which formed a semi-stable hairpin structure (15). Each library was designed with a trajectory and hypothesis in mind, though all six libraries were pooled to form a single starting library used in both trajectories.

Libraries 1 and 2 were designed within the context of the CoAzyme selection trajectory. Because CoAzymes use their 5' end (the +1 A) as a substrate to which phosphopantetheine is added, we assumed that the positioning of the 5' end would be critical for activity. Specifically, we believed the 5' end must be near the active site (likely a structured region) and not pointed to the exterior of all structure. Thus, we hypothesized that active RNAs

from the CoAzyme trajectory would have structure stabilizing the 5’ end (acceptor), a donor binding site, and an active site in close proximity to the 5’ end. Keeping this hypothesis in mind, we intentionally incorporated structural elements into libraries 1 and 2. Using the Batey group’s scaffold as inspiration (16), library 1 was designed using an existing crystal structure (PDB ID: 3IWN) (17) to specifically relocate the 5’ end into the interior of the structural scaffold near the small molecule binding site region (Figure 5.2B). Library 2 was designed to contain a three-way-junction (3WJ) motif with the 5’ end buried between two structured regions. Because the designs incorporated into libraries 1 and 2 stabilized the 5’ end, including a possible donor binding site, and buried the 5’ end in structure, we expected these libraries to be over-abundant in final populations for the CoAzyme trajectory and depleted for PPAT substrate trajectories.

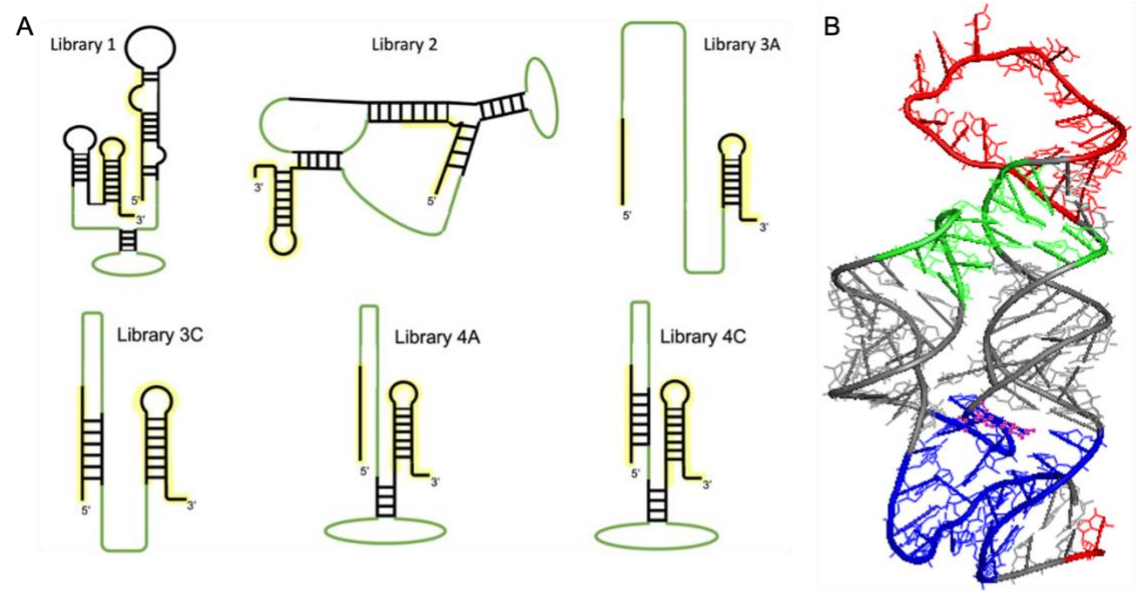

A
Library 1
Library 2
Library 3A
Library 3C
Library 4A
Library 4C
B
C
D
CoaD
G
5'
3'
E
CoaD
3'
G
5'

Figure 5.2. Six library designs with incorporated structure. (A) Simplified schematic of designed secondary structural architectures of the six RNA libraries studied here, where black represents defined sequence and green represents random regions. Primer binding sites are highlighted in yellow. RNAs are all the same length and use the same PBS. (B) Crystal structure of GMP riboswitch derived scaffold (PDB ID: 31WN) used for library 1. Library 1 is a circular permutation of the GMP riboswitch Batey scaffold (16). Blue indicates three-way junction where small molecules tend to bind. Red indicates differences in the crystal structure sequence form the sequence displayed in the Batey paper. Green residues form tertiary structural elements. Gray represent the remaining structural elements. Pink dots surround the residue that was chosen to be the 5' end in library design 1. (C) Crystal structure of CoaD bound to 3' dephospho CoA (PDB ID: IB6T) (18). Circled in yellow is the 3' hydroxyl group where an RNA chain would connect. CoaD is depicted as gray orbs. (D) General schematic for libraries 3A-3C design and rationale. (E) General schematic for libraries 4A-4C design and rationale.

Libraries 3A, 3C, 4A, and 4C were designed with the PPAT RNA substrate trajectory in mind, paying special mind to the active site of PPAT by examining its crystal structure (Figure 5.2C). As indicated by the crystal structure (PDB ID: IB6T), an RNA has little room for maneuvering as indicated by the position of the 3' hydroxyl of the ribose (circled in yellow). Therefore, we hypothesized that 5' RNA structure may govern active site accessibility and/or orientation. Furthermore, we predicted any RNA substrate of PPAT will likely require single-stranded structure on the 5' end, possibly with some stabilizing structures just downstream of the 5' end (Figure 5.2D). We expected the remainder of the RNA chain will serve to form interactions with the enzyme to stabilize the required, precarious positioning of the 5'end into the active site. These principles were used to incorporate intentional (or lack of) structure for libraries 3A-4C. The key differences between libraries 3A, 3C, 4A, 4C was the number of base pairs used to form a stem near the 5' end for the purpose of stabilizing it. Libraries 3A and 4A had no stem on the 5' end to account for the possibility that such a structure may actually be inhibitory for the ideal substrate (Figure 5.2A). Libraries 3C and 4C, however, had 6 base pairs forming a stem 3 nucleotides shy of the 5' end (Figure 5.2A). Interestingly, our general hypothesis that PPAT's RNA substrates required single stranded 5' ends was confirmed several months after the selection was started (unpublished data, Chapter 3, Figure 3.4). However, it was determined that PPAT preferentially capped RNA substrates with 4 or more unpaired nucleotides on the 5' terminus, predicting that libraries 3C and 4C would not serve as capping substrates of PPAT. Libraries 4A and 4C contain an additional structural motif: a

second base pair region which brings the 3’ end of the RNA back away from the enzyme (Figure 5.2E). Because of the base paired region in the 5’ end, the RNA chain will angle back towards the active site, possibly crowding the small area further. This design was expected to be favored in proportion to the extent that a second steam motif near the 3’ end helped the RNA chain wind back away from the active site and interact with other parts of the enzyme. Because the designs incorporated into libraries 3A-4C have single-stranded 5’ ends and downstream 5’ end stabilizing structures we expected these libraries to be highly abundant in the final populations for the PPAT substrate trajectory and depleted for the CoAzyme trajectory.

Name	Sequence (5' → 3')
Library 1	**AGGACCGGCCUAAACGGCAUUGC**ACUCCGCCGUAGGUAG CGNNNNNNNNNCGUGNNNNNNNNNNNNNNNNNNCAC GNNNNNNNNNNACCAUUCGAAAGAGUGGGACGCAA**ACCA AUCCGCUUCGGCGGAUACA**
Library 2	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNGAUGGN NNNNNNNNNNGCAAUGCCGUCAUGGCAANNNNNNNNNNN NNNNUUGCCAUGUGGCCGNNNNNNNNNNNNNNNUC**ACC AAUCCGCUUCGGCGGAUACA**
Library 3A	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN**ACC AAUCCGCUUCGGCGGAUACA**
Library 3C	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNGCCGGUNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN**ACC AAUCCGCUUCGGCGGAUACA**
Library 4A	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNGGUCCNNNNNNNNNNNNNNGG**ACC AAUCCGCUUCGGCGGAUACA**
Library 4C	**AGGACCGGCCUAAACGGCAUUGC**NNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNGGUCCNNNNNGCCGGUNNNNNNNNNNGG**ACC AAUCCGCUUCGGCGGAUACA**
Library Reverse Primer	TGTATCCGCCGAAGCGGATTGGT
Library Forward Primer	GCGTAATACGACTCACTATTAGGACCGGC
5' HIV UTR Reverse Primer	CCGACGCTCTCGCACCCATCTC

Table 5.1. Library and primer sequences. Primer binding sites of the library sequences are bolded and underlined. Regions with red letters represent incorporated structural regions. Segments highlighted in yellow were used to identify library of origin during high-

throughput sequencing analysis. Note, library 3A's yellow region was fully randomized and therefore library 3A sequences were identified through process of elimination.

Selection Strategy

To separate CoA-capped RNAs from unreactive RNAs we developed a selection strategy that use [(N-Acryloylamino) Phenyl] Mercuric Chloride (APM) layer PAGE gels as a partition method (Figure 5.3). The six libraries were pooled to form a single starting library and the two selection trajectories (CoAzyme and PPAT substrate) were performed in parallel. In the CoAzyme trajectory RNA was incubated with phosphopantetheine and in the PPAT substrate trajectory, RNA was incubated with both phosphopantetheine and active PPAT enzyme. CoA-capped RNAs were partitioned from uncapped RNAs by APM gel. The sulfur of the phosphopantetheine formed a coordinate covalent bond with mercury in the middle layer of the APM gel, which slowed the migration of CoA-RNAs allowing their separation from unreacted RNAs. Following their purification, CoA-RNAs were reverse transcribed and amplified. This cycle was repeated for several rounds for each trajectory.

Optimizing PCR for the selection

Before beginning the selection, it was necessary to establish some basic protocols such as number of cycles to be used during PCR and which reverse transcriptase to use. A series of pilot PCRs were run to determine the optimal number of PCR cycles to be use during the selection. Ideally, we wanted to use the smallest number of PCR cycles that would still provide robust amplification without producing significant non-specific product ("PCR monsters"). Therefore, we performed pilot PCRs to confirm the best number of cycles for

the libraries. As shown in Figure 5.4A, 5 cycles of PCR generated PCR product of the correct size with no notable non-specific products. Although 7 cycles of PCR (Figure 5.4B) produced similar results to 5 cycles, there was some minor smearing in the lanes indicative of non-specific products. Furthermore, 7 cycles was not required to successfully amplify and transcribe the libraries, thus 5 cycles of PCR was used throughout the selection. 9, 12, and 16 cycles clearly produced large quantities of incorrect product as indicated by bands between 180-250 bp (Figure 5.4B-C).

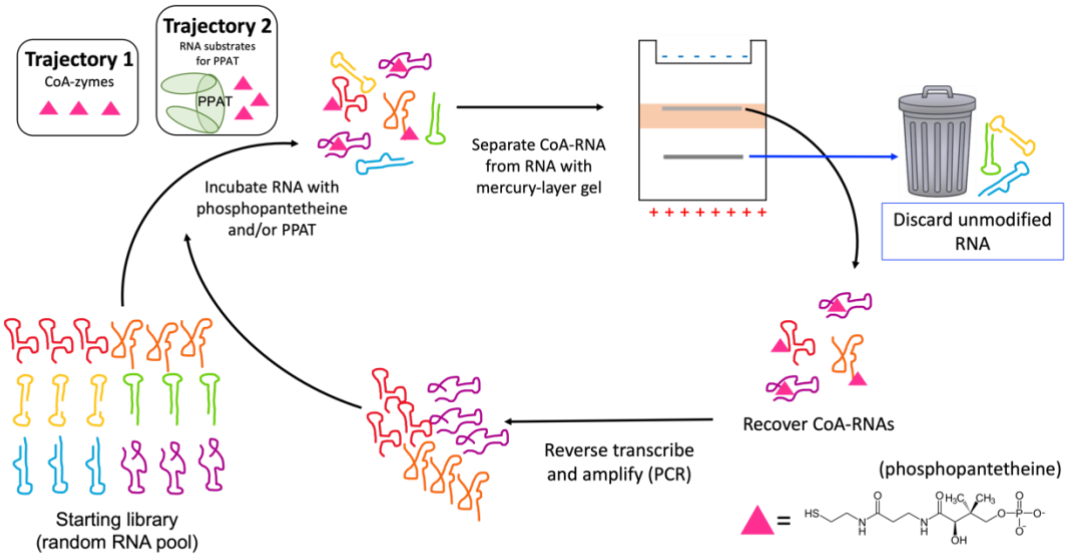


Figure 5.3. CoAzyme and PPAT RNA Substrate SELEX Schematic. Schematic for SELEX. A random RNA pool of six different libraries was mixed and used as the starting library. Two trajectories of the selection (CoAzyme and PPAT Substrates) were performed in parallel. RNAs in the CoAzyme selections were incubated with substrate phosphopantetheine. RNAs in the PPAT substrate trajectory were incubated with phosphopantetheine and enzyme PPAT. Reacted CoA-RNAs were separated from unreactive ^{PPP}A-RNAs by APM gel. CoA-RNAs were excised and purified from the APM layer, reverse transcribed, amplified, and transcribed back into RNA. Several rounds of selection were performed for each trajectory in this manner.

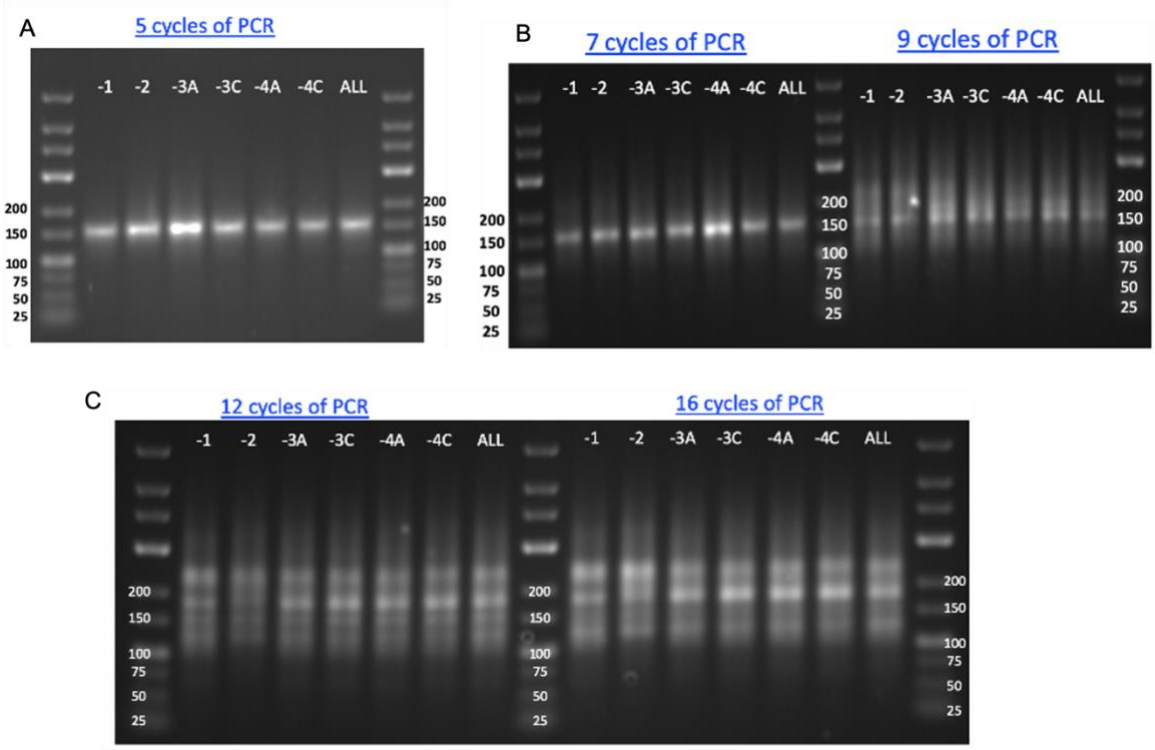


Figure 5.4 Pilot PCR for six libraries. Each library was used as an individual template (lanes 1-6) for PCR except lane 7, where are six libraries templates were mixed ('all') prior to PCR. PCR product was run on a 2% agarose gel stained with ethidium bromide and visualized by UV-Vis. Predicted band size is 153 bp. The difference between gels is the number of PCR cycles performed: (A) 5 cycles, (B) 7 and 9 cycles, and (C) 12 and 16 cycles.

Reducing reverse transcription bias between the six libraries

When designing the libraries for the CoAzyme selection, several designs and structural elements were incorporated into the libraries to test specific hypotheses. For example, for the self-capping CoAzyme trajectory we hypothesized that a library design which stabilizes the 5' end, includes a donor binding site, and buries the 5' end within structure would dominate the CoAzyme trajectory. Therefore, elements of this hypothesis were incorporated into our library 1 design (reference figure 7). A key feature of this selection is testing structural hypotheses and observing how they perform, are enriched/depleted, or mutate throughout the course of the selection. However, if certain steps such as reverse transcription are highly biased against the very structural elements that were intentionally incorporated into the libraries, then the results of the selection will be skewed inaccurately (not as a result of fitness, but as a result of bias), as we observed with ImProm-II (Figure 5.5A). This lead to an extensive comparison of various RTs and various reaction conditions (see Chapter 2). Ultimately, BST 3.0 DNA polymerase was found to have the highest yield during reverse transcriptase and the least amount of inter-library bias to the six libraries (Figure 5.5B). Further optimization of reaction conditions gave rise to minimal inter-library bias and high cDNA yield and these optimized conditions were used for the selection.

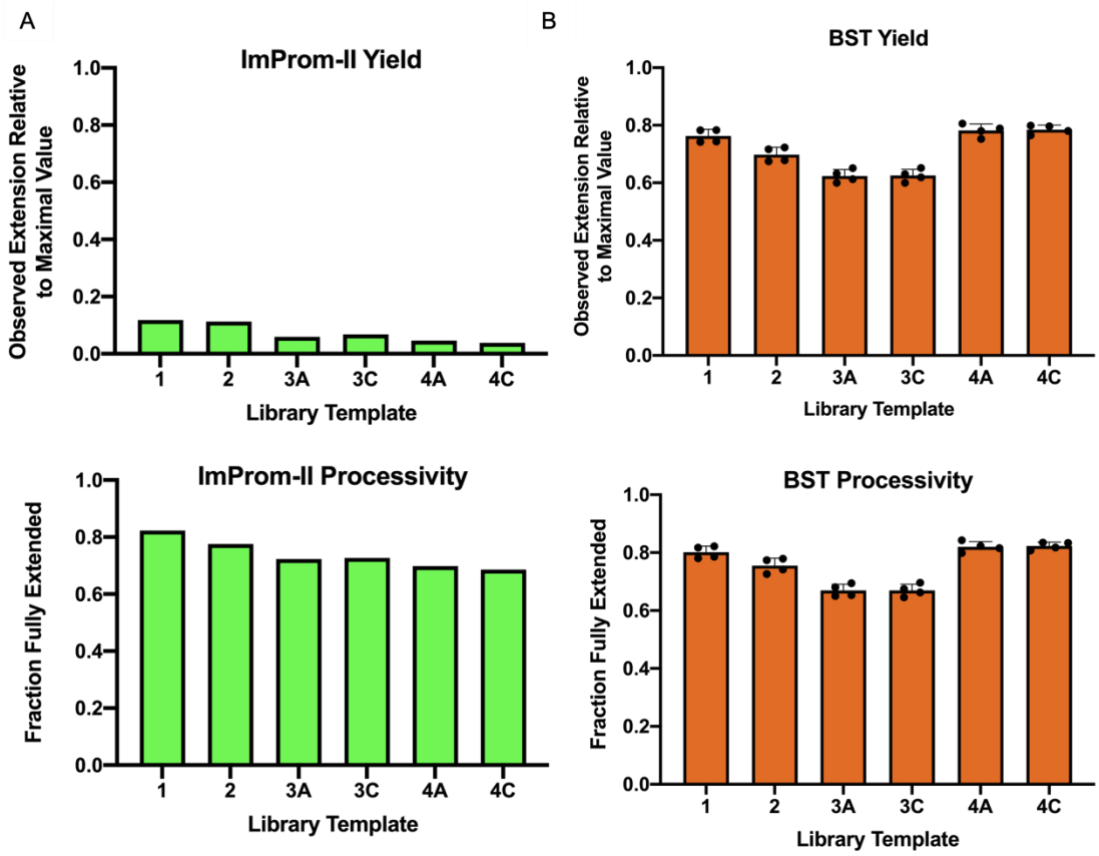


Figure 5.5. Reverse transcriptases have different inter-library bias. Comparison of reverse transcriptase bias between six RNA library templates with various amounts of structure. The primer extension assays were performed using ImProm-II reverse transcriptase or BST 3.0 DNA Polymerase as described in the materials and methods. Library 1, 2, 3A, 3C, 4A, 4C (left to right) were used for these reactions. (A) ImProm-II (green), N=1 (*adapted from Chapter 2, Figure 2.2B)* and (B) BST 3.0 DNA polymerase (orange), N=3 yield and processivity for six library templates (*adapted from Chapter 2, Figure 2.5 A-B).*

CoAzyme self-capping is time-dependent.

After several rounds of selection, we observed an accumulation of CoA-RNA in the APM layer of PAGE gels during the partitioning step for both trajectories. One possibility is that this data indicated RNA capping and of RNA serving as a substrate for PPAT. However, it is also possible that these data were reflective of a non-specific chemical reaction or the work of a *trans* acting ribozyme. We wondered how much of the CoA-RNA being observed was background signal and whether the CoAzyme reaction for self-capping was time-dependent. Therefore, during rounds 7A and 11A we removed aliquots of reacting RNA at various time points, including a no incubation (0 hr) sample to observe background signal, ran them on an APM gel (Figure 5.6A) and quantified the amount of reacted RNA in the APM layer (Figure 5.6B). Over the course of 7 hours (round 7A), larger quantities of CoA-RNA were observed in the APM layer, indicative of a time-dependent reaction. The background signal observed at 0 hours (no incubation reaction) made up less than 0.2% of the total RNA control sample from round 7A. Interestingly, little more than 2% of the RNA from round 7A was reactive after a 7 hr incubation period and ~1% of the RNA form round 11A was active after 4 hrs (Figure 5.6B). Looking at the activity across several rounds, it was clear that there was no significant increase in CoA capping activity in either trajectory by round 12A. Thus, we decided to restart the selection from round 5 and increase the stringency more slowly for subsequent rounds to try and retain active sequences.

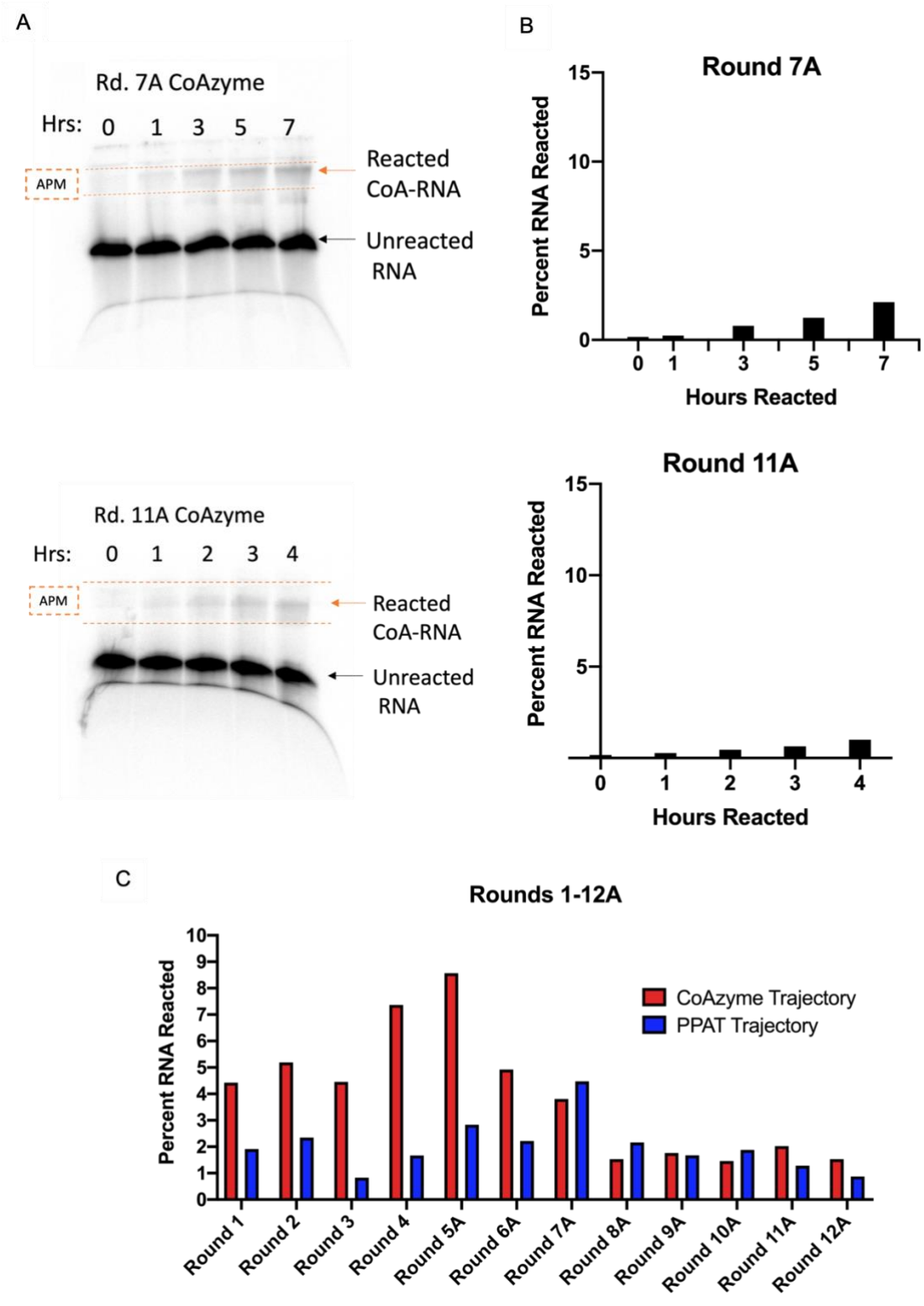

A
Rd. 7A CoAzyme
Hrs: 0 1 3 5 7
APM
Reacted
CoA-RNA
Unreacted
RNA
B
Round 7A
Percent RNA Reacted
15
10
5
0
0 1 3 5 7
Hours Reacted
Rd. 11A CoAzyme
Hrs: 0 1 2 3 4
APM
Reacted
CoA-RNA
Unreacted
RNA
Round 11A
Percent RNA Reacted
15
10
5
0
0 1 2 3 4
Hours Reacted
C
Rounds 1-12A
Percent RNA Reacted
10
9
8
7
6
5
4
3
2
1
0
CoAzyme Trajectory
PPAT Trajectory
Round 1
Round 2
Round 3
Round 4
Round 5A
Round 6A
Round 7A
Round 8A
Round 9A
Round 10A
Round 11A
Round 12A

Figure 5.6. Round 7A & 11A CoAzyme trajectory time-course. (A) Time-course of CoAzyme RNA during reaction step of round 7A (upper) and round 11A (lower). Aliquots of internally radiolabeled α^{32}P RNA were removed and terminated at various time points and run on an APM gel. Gel was visualized by phosphorimaging on a Typhoon FLA 9000. (B) Quantification of round 7A (upper) and round 11A (lower) gel image. The band intensities of reacted and unreacted RNA and percent RNA reacted were determined by dividing reacted RNA by (reacted + unreacted). (C) Quantified percent reacted RNA from CoAzyme and PPAT trajectories for rounds 1 through 12B.

Selection stringency may have negatively impacted selection outcomes

Low observed activity in round 7A and 11A may have been related to selection stringency. If increased too quickly, high stringency may have prevented the few active sequences in the population from evolving and/or enriching (Table 5.2). Alternatively, the low observed activity could be a result of the selection populations requiring additional rounds to provide active sequences time to enrich within the population and produce a greater signal (percent reacted RNA). Therefore, we performed four additional rounds of selection after round 7A (Figure 5.6A) with decreasing pPant concentrations and reaction times. However, we observed even less activity in Rnd 11A than in Rnd 7A. Specifically, we observed ~1% reacted RNA (Figure 5.6B) in round 11A. Furthermore, there was a drop off in percent reacted RNA that persisted over the course of several rounds that may have corresponded to rapid increase in stringency (Figure 5.6C). Relative to round 1, the reaction time in round 7A was ~3 fold shorter and the substrate (phosphopantetheine) concentration an order of magnitude smaller (Table 5.2). Whereas round 11A had a reaction time that was ~5 fold less relative to round 1 and the pPant concentration was more than 20 fold less than round 1. Perhaps in earlier rounds, active sequences were present but high stringency drove them out. Therefore, we decided repeat the selection starting from round 5 and increase the stringency at a much slower rate in hopes of retaining and evolving active species within the populations.

Selection Round	Reaction Time (hrs)	[pPant] (μM)	Mutagenic PCR
1	24	3330	
2	24	3330	
3	22	3330	
4	22	3330	x
5A	22	666	x
6A	7	666	x
7A	7	333	x
8A	4	333	x
9A	4	150	x
10A	4	150	x
11A	4	150	x
5B	24	666	
6B	24	666	x
7B	18	666	
8B	18	666	
9B	18	666	
10B	12	333	x
11B	12	333	x
12B	12	333	

Table 5.2 Selection reaction conditions per round. Reaction times and substrate (phosphopantetheine) concentrations for each selection round. Rounds where mutagenic PCR was performed are indicated by an "x", otherwise standard PCR was performed as outlined in the materials and methods. Rounds that were sent for high-throughput sequencing are highlighted in yellow. Note, selection was re-started after round 4 and the stringency was increased more slowly for rounds 5B-12B.

Selections with lowered stringency had similar levels of activity as starting library

Using DNA from round 4, fresh RNA was transcribed for round 5B and less stringent conditions were used during initial rounds and stringency was increased at a slower rate (Table 5.2). Round 5B's CoAzyme activity was much higher (~11%) than activity observed in rounds 7A and 12A (Figure 5.7A-B). This is likely due to the 24 hr incubation period, higher concentrations of pPant substrate (666 μM in 5B vs 150 μM in 12A), and larger presence of RNAs with lower-level activity (which would have been lost during high-stringency rounds like 7A-12A.

To better estimate the progress of the selection, we directly compared CoAzyme activity from round 12B to starting library CoAzyme activity. We observed ~9.5% of reacted RNA from round 12B and ~7% reactive RNA from the starting library after 11 hours (Figure 5.7B-C). Furthermore, we observed low activity (percent reacted RNA) over the course of the several repeated selection rounds (5B-12B) without notable increases (Figure 5.7E). Given the minimal difference between round 12B's activity and the starting library activity, no additional rounds of selection were performed as there was no strong indication of significant selection progress.

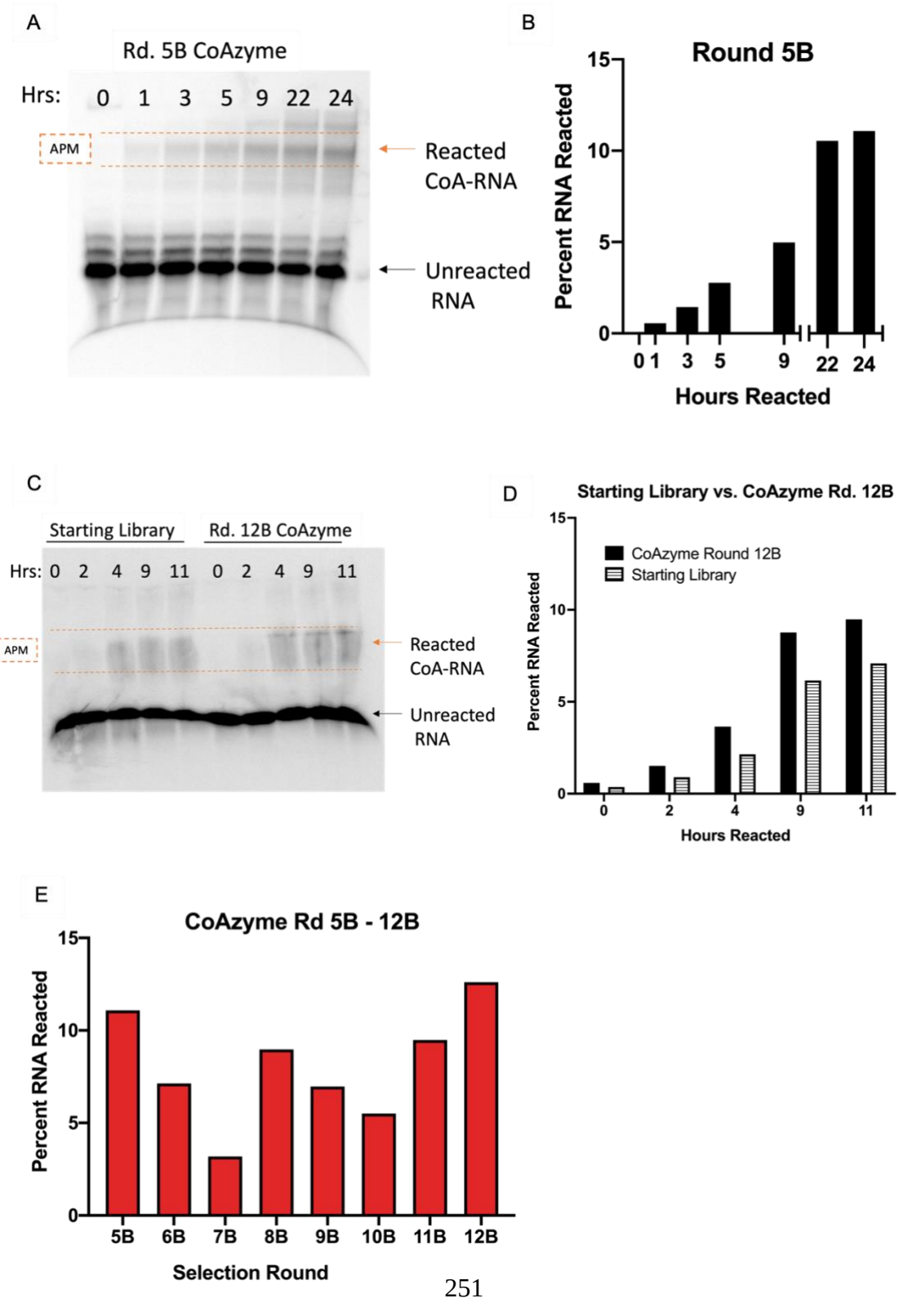

A
Rd. 5B CoAzyme
Hrs: 0 1 3 5 9 22 24
APM
Reacted CoA-RNA
Unreacted RNA
B
Round 5B
Percent RNA Reacted
15
10
5
0
0 1 3 5 9 22 24
Hours Reacted
C
Starting Library
Rd. 12B CoAzyme
Hrs: 0 2 4 9 11 0 2 4 9 11
APM
Reacted CoA-RNA
Unreacted RNA
D
Starting Library vs. CoAzyme Rd. 12B
CoAzyme Round 12B
Starting Library
Percent RNA Reacted
15
10
5
0
0 2 4 9 11
Hours Reacted
E
CoAzyme Rd 5B - 12B
Percent RNA Reacted
15
10
5
0
5B 6B 7B 8B 9B 10B 11B 12B
Selection Round

Figure 5.7. Round 5B and 12B CoAzyme trajectory time-course. Time-course of CoAzyme RNA during reaction step of round 5B (A) and round 12B and starting library (C). Aliquots of internally radiolabeled α^{32}P RNA were removed and terminated at various time points and run on an APM gel. Gel was visualized by phosphorimaging on a Typhoon FLA 9000. Quantification of round 5B (B), round 12B and starting library (D) gel image. The band intensity of reacted and unreacted RNA and percent RNA reacted was determined by dividing reacted RNA by (reacted + unreacted). (E) Quantified percent reacted RNA from CoAzyme trajectory for rounds 5B through 12B.

Counted reads for ranked unique sequences reveals little convergence of populations

Although the data we observed showed no indication of significant selection progress, we wondered if enrichment occurred at low levels, not detectable through analysis of APM gels. Therefore, to gain further insights about the selection progress and how stringency may have impacted populations over the course of the selection, we sent several rounds of selection from each trajectory (Table 5.2) for high throughput sequencing (HTS). The HTS data was preprocessed to remove constant regions and to discard any sequences not within ± 9 nt of the expected size (90 nt) (Tables 5.3-4.3). The data was further analyzed using an open-source bioinformatic toolkit called FASTAptameR 2.0 which is specifically designed for HTS analysis of combinatorial selections (19, 20).

To determine whether the selection populations had converged on specific sequences or if the populations remained largely diverse after several rounds of selection we used FASTAptameR-Count to count, normalize, and rank unique sequences in each selection round. Line plots of reads for each unique sequence, sorted by rank, (Figure 5.8) were generated for critical selection rounds: 2A (early in the selection), 5A/5B (divergent points), and 11A/12B (selection end). In cases where selection rounds had less than 1,000 total processed reads, the closest following selection round was used instead (eg. 7A instead of 5A). In the earliest stage of a selection, little to no convergence of a population is expected. However, convergence is likely to be observed over the course of a successful selection, especially in the final rounds of that selection. For instance, the selection might

have converged on a large number of enriched molecules or on a smaller number of extremely robust molecules. Plots of the ranked unique sequences in order of their total number of reads makes either scenario for convergence very clear.

CoAzyme	**Round 2A**	**Round 5A**	**Round 7A**	**Round 9A**	**Round 11A**	**Round 5B**	**Round 7B**	**Round 9B**	**Round 11B**	**Round 12B**
Raw total reads	37,951	396,889	5,564	568,227	486,729	8,687	351,884	449,901	148,933	39,408
Long sequences	28,374	33,690	2,221	68,704	53,798	7,879	32,639	40,363	15,986	18,886
Short sequences	295	10,210	159	24,999	26,246	43	20,738	59,772	126,099	7,515
Total processed sequence reads	9,282	352,989	3,184	474,524	406,685	765	298,507	349,766	6,848	13,007
Unique processed sequences	7,380	269,892	2,464	361,798	307,023	685	224,893	254,550	6,229	10,093

Tables 5.3. High-throughput sequencing raw data and processing for CoAzyme trajectory. Data processing was performed using cutadapt to trim the 5' and 3' constant regions from sequences and to discard any uncut sequences or sequences with lengths no within ± 9 nt of the expected size (90 nt) after trimming. Raw total reads is the number of sequences prior to any processing; long and short sequences did not fit within the ± 9 nt parameter; and the total processed sequence reads were analyzed using FASTAptameR2.0. Columns in light gray indicate rounds with less than 1,000 total processed reads which were omitted during data analysis.

PPAT	**Round 2A**	**Round 5A**	**Round 7A**	**Round 9A**	**Round 11A**	**Round 5B**	**Round 7B**	**Round 9B**	**Round 11B**	**Round 12B**
Raw total reads	11,737	665	76,020	9,523	509,088	587,330	271,211	608,106	9,140	420,839
Long sequences	7,938	438	17,569	2,362	69,513	49,212	16,478	76,252	5,469	32,256
Short sequences	250	10	9,441	537	35,766	24,386	14,213	49,003	980	122,873
Total processed sequence reads	3,549	217	49,010	6,624	403,809	513,732	240,520	482,851	2,691	265,710
Unique processed sequences	2,800	189	37,449	5,335	307,285	387,602	183,026	362,978	2,066	198,140

Tables 5.4. High-throughput sequencing raw data and processing for PPAT Capping trajectory. Data processing was performed as described in Table 5.3

Throughout the CoAzyme trajectory (Figure 5.8A) the total number of reads for the most highly ranked unique sequences never exceeds 40 reads, even in the final rounds of selection (11A or 12B). In a successful selection, the highest ranked sequences can have more than 50,000 reads. Furthermore, of the 100 top ranked unique sequences, the lower 75 ranked sequences have less than 10 reads each. Thus, the CoAzyme selection was enormously diverse, and no convergence was observed. Similarly, the PPAT trajectory (Figure 5.8B) showed little convergence across several rounds of selection. Even though round 12B had the highest counts for a unique sequence (~130 reads), its plot was indicative of enormous diversity and little to no convergence. PPAT round 12B had ~198,000 total processed reads (Table 5.4), but the highest ranked unique sequence from that round had only ~130 reads which is less than 0.1% of the total reads in the round 12B population. Overall, there was no meaningful convergence in any of the selection rounds for either trajectory, thus successfully completing this selection will require many additional rounds of selection and/or serious changes to the selection parameters.

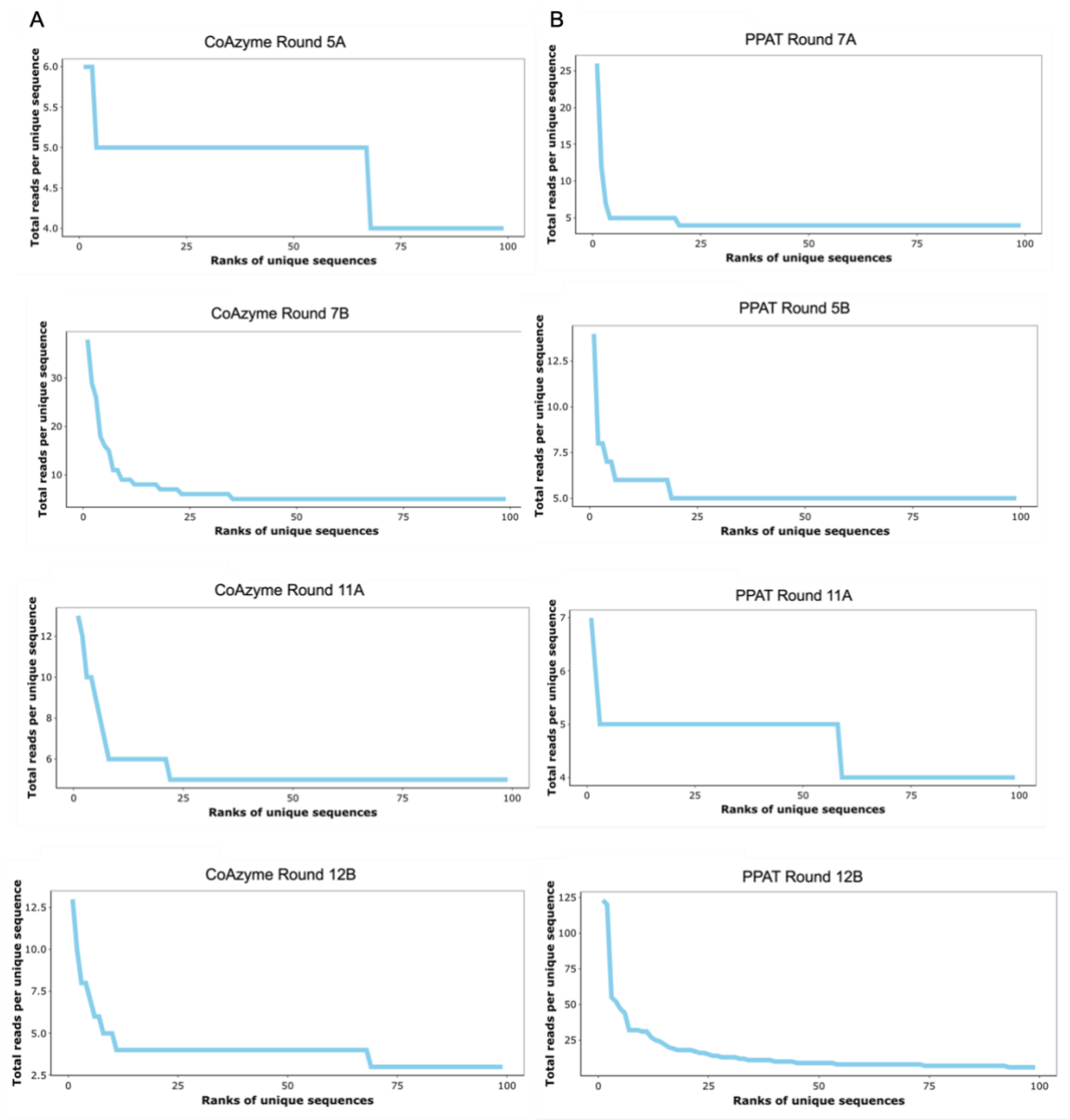


Figure 5.8. Total reads of ranked unique sequences. After counting the occurrence of each unique sequence (FASTAptameR-Count) from various selection rounds, the 100 highest ranking unique sequences were identified. Line plots of reads for each unique sequence (sorted by rank) were then plotted using interactive plotting tools

(FASTAptameR-Count, reads per rank) for (A) CoAzyme selection rounds and (B) PPAT substrate selection rounds.

No significant enrichment of individual sequences was observed

To measure the consistency of the population structure between two given selection rounds, selection rounds were compared using the FASTAptamer-Enrich function and plotted against each other using RPM values of specific sequences to generate a scatter plot. This analysis provided insight about whether specific sequences were enriched or depleted throughout the course of the selections. Specifically, the location and spread of the scatter on the plot relative to a diagonal line at y = x is indicative the magnitude of enrichment or depletion of specific sequences between rounds. Ordinarily, these plots have a huge number of data points on and around the x = y diagonal line, indicating some level of consistency in the population structure across rounds of selection, and some data points from enriched or depleted sequences falling above and below the line. However, the most notable property of the RPM scatter plots (Figure 5.9) is the lack of scatter falling near the x = y line.

The CoAzyme plots (Figure 5.9A) comparing rounds 5A to 11A and PPAT plots comparing rounds 7A to 11A (Figure 5.9B) have virtually no overlap of sequences between rounds, with little scatter falling on or near the x = y line, and the majority of individual sequences being present in either (but not both) selection rounds. In this case, it was not that these sequences were enriched or depleted, but rather that they were not even present in the other selection round. For CoAzyme plots comparing rounds 7B to 12B (Figure 5.9A) and PPAT plots comparing rounds 5B to 12B (Figure 5.9B) there was some minor

overlap of sequences between sequence rounds (around the x = y line), however there was no significant enrichment of few, robust sequences or even mild enrichment numerous, lower-performing sequences. These data taken alongside the ranked reads data (Figure 5.8) are strongly indicative of this selections failure to converge on/enrich for sequences capable of self-capping or serving as a capping substrate.

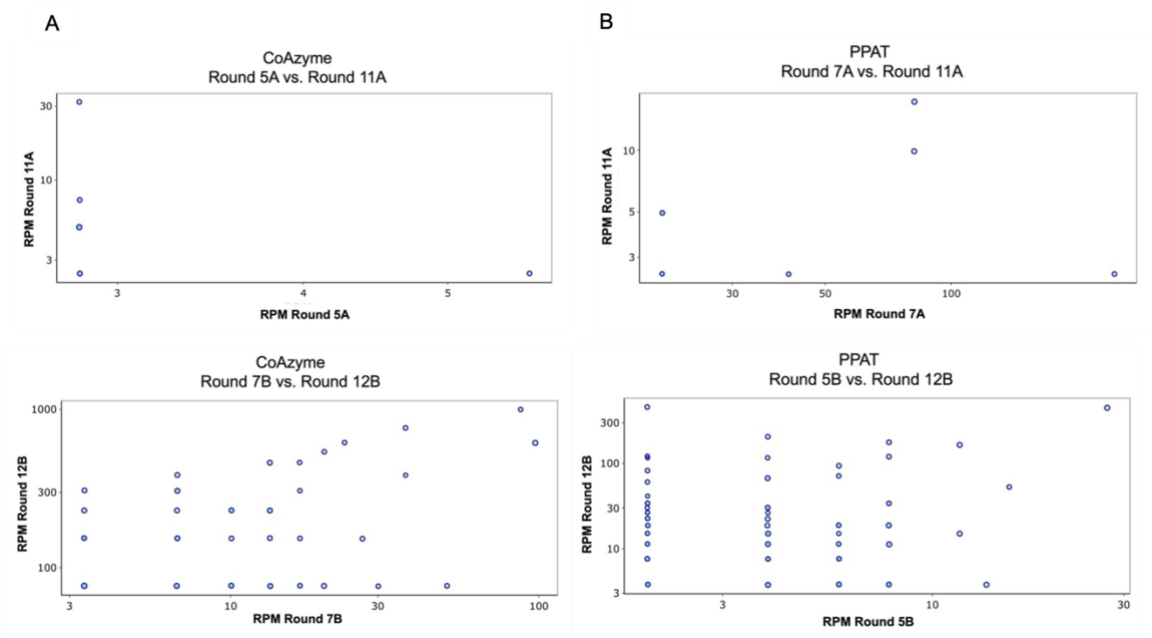


Figure 5.9. Enrichment of sequences between various selection rounds. Two selection rounds were directly compared using the FASTAptameR-Enrich function to determine the enrichment and depletion of specific sequences during the selection. Plots with very few data points indicate few conserved sequences between the compared rounds. Early rounds (2A), divergent rounds (5A/5B), and final rounds (11A/12B) of the (A) CoAzyme and (B) PPAT selection trajectories were directly compared. Ordinarily in these plots, values falling along the x = y diagonal line indicate no enrichment or depletion, indicative of consistent population structures between rounds. However, …

Impact of library design on selection outcomes

Before beginning this selection, six libraries were designed with various structural elements, each of which represented a hypotheses related to selection outcomes. For the CoAzyme selection, libraries 1 and 2 were designed with highly structured elements with the 5' end to be buried in structure. These structural elements were chosen based on our hypothesis that the 5' end would need to be in close proximity to an active site (a highly structured region). For the PPAT, libraries 3A, 3C, 4A, and 4C were designed with little to no structure directly on the 5' end, but with downstream stabilizing structures. For the PPAT substrate selection, we hypothesized that PPAT would require RNA substrates to have a single-stranded 5' end to fit into the active site as a result of steric constraints and that structural elements downstream of the 5' end could stabilize the RNAs positioning into the active site.

To ascertain whether either selection trajectory was dominated by specific library designs, each of the six libraries was identified using sequence markers (Table 5.1) and the FASTAptamer-Motif search function. The fraction of total reads (Tables 5.5 – 5.6) was determined for the six libraries in each round of the CoAzyme and PPAT trajectories (Figure 5.10A-C). Interestingly, library 3A makes up the largest fraction of total reads for both the CoAzyme and PPAT trajectory. However, due to library 3A's fully randomized nature, it has no sequence markers by which it can be unambiguously identified. Thus, the other five libraries are identified, and whatever sequences remain are assumed to be library

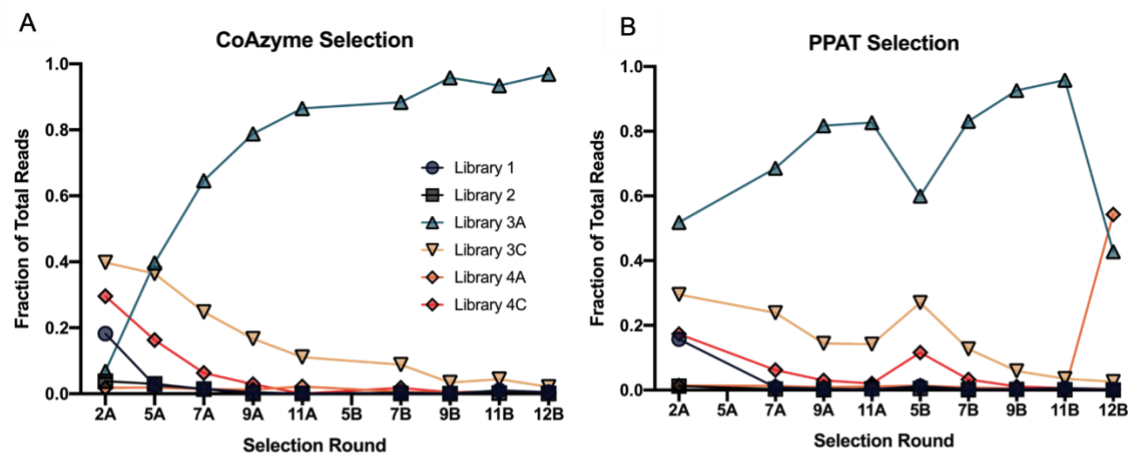
3A. However, because mutagenic PCR was used during the selection, it is likely that many of the sequence regions which are used to identify the other five libraries accumulated mutations, thereby omitting them from being counted. This may have falsely skewed the fraction of total reads count to favor library 3A slightly.

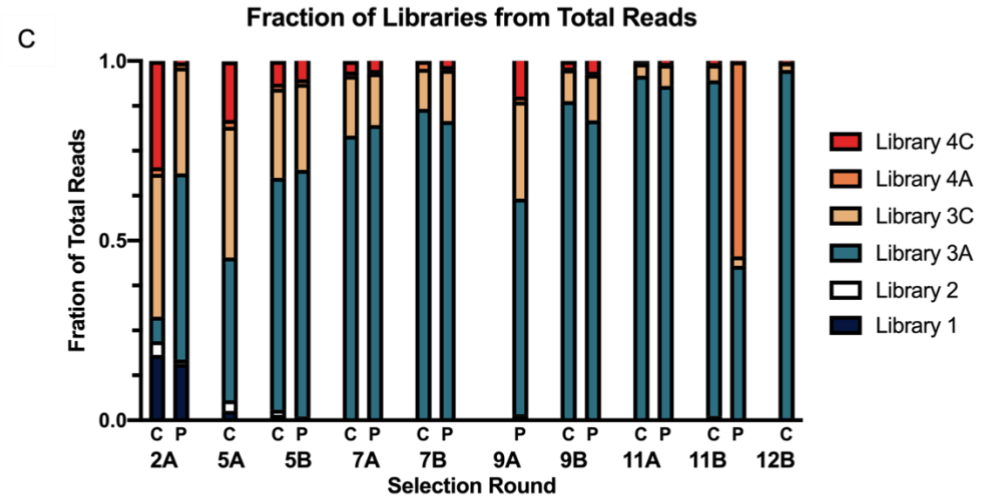
CoAzyme	**Round 2A**	**Round 5A**	**Round 7A**	**Round 9A**	**Round 11A**	**Round 5B**	**Round 7B**	**Round 9B**	**Round 11B**	**Round 12B**
Library 1	182,396	26,491	14,761	4,179	1,436	10,458	2,750	1,224	9,638	4,844
Library 2	37,600	29,658	13,505	181	929	13,072	1,109	246	1,898	692
Library 3A	68,089	396,817	646,357	787,886	864,753	698,039	883,768	958,132	933,703	969,017
Library 3C	398,082	364,377	248,430	167,245	110,680	197,386	87,686	34,037	43,516	21,296
Library 4A	17,776	19,284	14,133	11,569	22,081	19,608	6,295	2,696	4,089	1,615
Library 4C	296,057	163,373	62,814	28,938	120	61,438	18,392	3,665	7,155	2,537

Tables 5.5. Processed and analyzed CoAzyme HTS data. CoAzyme trajectory processed data. Library reads shown in reads per million. The processed reads (Tables 5.3-5.4) were analyzed in FASTAptameR2.0 and each read was identified as belonging to one of the six libraries using sequence markers.

PPAT	**Round 2A**	**Round 5A**	**Round 7A**	**Round 9A**	**Round 11A**	**Round 5B**	**Round 7B**	**Round 9B**	**Round 11B**	**Round 12B**
Library 1	156,664	18,433	7,876	4,378	2,310	12,090	3,018	3,223	5,946	2,476
Library 2	12,398	-	3,183	755	4,336	4,802	1,472	1,555	743	226
Library 3A	517,892	622,120	685,738	817,331	827,044	600,190	831,241	925,782	958,380	427,978
Library 3C	295,294	244,240	239,237	143,569	141,668	269,940	127,382	59,107	34,931	26,243
Library 4A	14,088	23,041	12,610	8,605	11,087	13,746	7,974	4,517	1,115	542,565
Library 4C	172,725	110,599	62,416	30,495	20,200	116,123	33,403	10,593	5,574	3,214

Tables 5.6. Processed and analyzed PPAT HTS data. Data processing done as described above in Table 5.5.


A
CoAzyme Selection
Fraction of Total Reads
1.0
0.8
0.6
0.4
0.2
0.0
2A 5A 7A 9A 11A 5B 7B 9B 11B 12B
Selection Round
Library 1
Library 2
Library 3A
Library 3C
Library 4A
Library 4C
B
PPAT Selection
Fraction of Total Reads
Selection Round


C
Fraction of Libraries from Total Reads
Fration of Total Reads
1.0
0.5
0.0
C P C C P C P C P P C P C P C P C
2A 5A 5B 7A 7B 9A 9B 11A 11B 12B
Selection Round
Library 4C
Library 4A
Library 3C
Library 3A
Library 2
Library 1

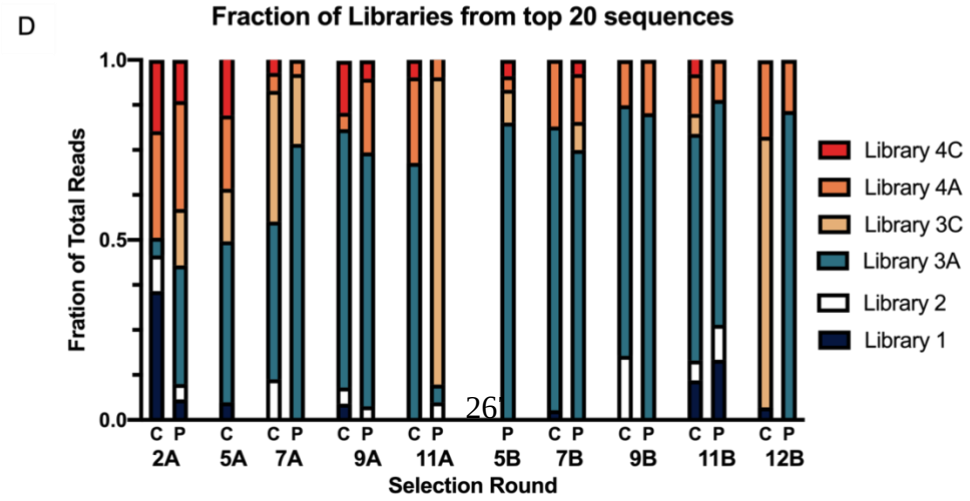

D
Fraction of Libraries from top 20 sequences
Fration of Total Reads
1.0
0.5
0.0
C P C C P C P C P P C P C P C P
2A 5A 7A 9A 11A 5B 7B 9B 11B 12B
Selection Round
Library 4C
Library 4A
Library 3C
Library 3A
Library 2
Library 1

Figure 5.10. Analysis of HTS to observe selection progress. Each of the six libraries was identified by sequence markers and then quantified to determine the fraction of total reads for each library. The changes in library fractions are plotted for each round of selection for (A) CoAzyme and (B) PPAT trajectories. Any rounds with fewer than 1,000 total unique processed reads were excluded from these data sets. (C) Bar graph indicating what fraction of total reads each library consisted of in each selection round. Each group of three shows data for CoAzyme (C) and PPAT (P) trajectory. (D) The top 20 most abundant sequences from selection rounds were identified as one of the six libraries. The bar graph indicates what fraction of the top 20 most abundant sequences each library consisted of in various selection rounds.

To determine if the fraction totals for the six libraries for the total read count was similar to the fraction totals for the top 20 most abundant sequences in each selection round for both trajectories, the top sequences were analyzed separately (Tables 5.7 – 5.8). Because much fewer sequences were being analyzed, a more detailed analysis of each individual sequence ensued which allowed the presence of mutations amongst the identifying sequences for each library. These data analysis of the the top 20 sequences were plotted in a bar graph to indicate the fraction of total reads each library made up in various selection rounds (Figure 5.10D). The data were relatively consistent with the total read count analysis (Figure 5.10C), with the exceptions of PPAT round 11A and CoAzyme round 12B where library 3C made up the largest fraction of total reads. Although it is interesting to speculate about the impact of library designs on selection outcomes, these data do not allow for any concrete conclusions. For the CoAzyme selection we hypothesized that libraries containing structured elements and 5' ends buried in structure (libraries 1 & 2) would be enriched. On the other hand, we hypothesized that libraries with single-stranded 5' ends and possible downstream stabilizing structures (libraries 3A, 3C, 4A, & 4C) would be enriched for the PPAT trajectory. Therefore, any distinct enrichment or depletion of specific library designs over the course of the selection would be informative about the possible structural requirements for a given selection. However, within the top 20 most abundant sequences, the library design diversity is relatively high. For example, round 11B for the CoAzyme and PPAT trajectories has 4-5 of the library designs present, further supporting previous data which indicated a very high sequence and population diversity

throughout the selection. There were very few of the same sequence in one round to the next, resulting in virtually no enrichment of depletion of specific sequences. This may be indicative of a random, rather than intentional capture, of RNAs during the partition step or it could indicate that there is specific capture of sequences that are being non-specifically capped. Overall, the lack of enrichment and convergence accompanied by high levels of sequence diversity and population inconsistencies between rounds indicate the selection failed.

CoAzyme Top 20 Seq	**Round 2A**	**Round 5A**	**Round 7A**	**Round 9A**	**Round 11A**	**Round 7B**	**Round 9B**	**Round 11B**	**Round 12B**
Library 1	0.358	0.049	0.000	0.045	0.000	0.027	0.000	0.110	0.035
Library 2	0.099	0.000	0.113	0.045	0.000	0.000	0.178	0.055	0.000
Library 3A	0.049	0.447	0.438	0.718	0.714	0.788	0.696	0.630	0.000
Library 3C	0.000	0.146	0.363	0.000	0.000	0.000	0.000	0.055	0.752
Library 4A	0.296	0.204	0.050	0.045	0.238	0.185	0.126	0.110	0.212
Library 4C	0.198	0.155	0.038	0.145	0.048	0.000	0.000	0.041	0.000

Tables 5.7 CoAzyme trajectory library fractions from top 20 sequence reads. The top 20 most abundant sequences for selection rounds were determined using FASTAptameR-Count**.** The top 20 sequences from each round were identified as one of the six libraries using sequence markers. The fraction of the reads from the top 20 sequences was determined for each library. Rounds with less than 1,000 total processed reads were omitted.

PPAT Top 20 Seq	**Round 2A**	**Round 7A**	**Round 9A**	**Round 11A**	**Round 5B**	**Round 7B**	**Round 9B**	**Round 11B**	**Round 12B**
Library 1	0.057	0.000	0.000	0.000	0.000	0.000	0.000	0.167	0.000
Library 2	0.043	0.000	0.038	0.049	0.000	0.000	0.000	0.097	0.000
Library 3A	0.329	0.767	0.705	0.049	0.826	0.750	0.852	0.625	0.858
Library 3C	0.157	0.194	0.000	0.854	0.091	0.078	0.000	0.000	0.000
Library 4A	0.300	0.039	0.205	0.049	0.038	0.133	0.148	0.111	0.142
Library 4C	0.114	0.000	0.051	0.000	0.045	0.039	0.000	0.000	0.000

Tables 5.8. PPAT trajectory library fractions from top 20 sequence reads. This analysis was performed as described in Table 5.7.

DISCUSSION

Two goals of this selection were to select for RNAs under *in vivo*-like conditions to generate RNAs which retained functionality in cells and to better understand possible mechanisms of CoA-biogenesis by identifying RNAs capable of self-capping or serving as a substrate for post-transcriptional capping by PPAT. However, as the HTS data indicated, after 12 rounds of selection, there was no enrichment or convergence on active sequences in the populations. One possibility is that the selection parameters were too stringent. However, stringency is a broad term which encompasses many parameters including reaction times, substrate concentration, reaction volume, buffer composition, etc. After we observed decreasing CoA-capping activity over the course of the initial 12 rounds of selection (Figure 5.6C), we repeated the selection starting from round 4 and reduced the stringency (reaction time and substrate concentration). However, even with less stringent conditions the CoA-RNA capping activity remained low through round 12B of both trajectories (Figure 5.7E). These data were a strong indication of no significant selection progress which was further confirmed by the HTS data analysis.

The HTS ranked reads data (Figure 5.8) revealed no significant enrichment for few, robust sequences or even mild enrichment numerous, lower-performing sequences, ultimately illuminating the selections failure to converge on/enrich for sequences capable of self-capping or serving as a capping substrate. The diversity of sequence reads in each round was also inconsistent; the enrichment analysis revealed the inconsistencies in population

structure throughout the selection (Figure 5.9). There were very few of the same sequence in one round to the next, resulting in virtually no enrichment of depletion of specific sequences. This may be indicative of a random, rather than intentional capture of RNAs during the partition step. Alternatively, specific capture of non-specific capped of RNAs could also be responsible for these types of data. Though CoA-RNAs appeared clearly within the APM layer (Figure 5.6A), other RNAs can be trapped at the interfaces between the APM and PAGE layers of the gel and are often excised along with the APM layer. Perhaps some the inconsistencies in sequence identity between rounds is attributed to the fully random capture of RNAs at the gel interfaces. Although the APM gels have been reliable for other experiments involving CoA-RNAs (Chapter 4, Figure 4.3B), alternative CoA-RNA capture methods like thiopropryl sepharose columns could be employed in future selections rounds. Thiopropyl sepharose columns have been used for the reliable separation of CoA-RNA from ATP-RNA previously (5) and could be used a feasible alternative to APM gels. However, it is possible that instead of random capture, specific capture of non-specifically capped RNAs took place. In this scenario, low-level, background chemical reactions which result in CoA-capped RNA could be responsible for non-specific capping (and therefore capture) of RNAs. Therefore, this model suggests that the time-dependent CoA-RNA formation that was observed (Figures 5.6 & 5.7) could be a result of background activity, and not specific CoAzyme or PPAT capping activity.

The poor selection outcome is most likely related to the *in vivo*-like conditions of the selection were too stringent. The *in vivo* like selection conditions included ion concentrations similar to cellular conditions, crowding reagents, reaction incubations at 37°C, and buffers with a pH of 7.5. Perhaps, these selection conditions were too restrictive and ultimately stood in way of successful selection outcomes. Although reduced ion concentrations can negatively affect an RNAs ability to catalyze reactions, a recent study showed that addition of biologically relevant concentrations of glutamate-magnesium complexes to RNA (aptamers and ribozymes) stabilized the RNA, reduced RNA degradation, and promoted higher catalysis (21). We took advantage of chelated magnesium from glutamate-magnesium complexes, to try to provide optimal conditions for RNA activity, stability, and folding whilst still maintaining *in vivo* like conditions. However, even with slightly higher $MgCl_2$ concentrations, other ions like calcium and manganese remain low or omitted in the selection buffer. As was previously reported, *in vitro* self-capping CoAzymes required high (~5 mM) calcium and manganese for activity(13). Perhaps, future selections should begin using buffers with higher, non-biologically relevant ion concentrations and slowly titrate in the *in vivo*-like buffer, providing sequences time to evolve. Temperature can also prove either beneficial or detrimental: higher temperature may promote improved ribozyme catalysis, however, it can also produce higher RNA degradation. It is possible that increased reaction temperatures in coordination with long reaction times lead to increased degradation of many RNAs, negatively impacting the selection outcomes. However, selecting for RNAs

under biologically relevant temperatures is a critical component of generating RNAs which retain their functionality in cells. Furthermore, there was little observed degradation of the RNA on the APM gels, suggesting temperature was not a significant issue. The use of crowding reagents can simulate cellular folding conditions for the RNA, allowing for secondary and tertiary structures to form which may not be achievable in buffer only. Previous studies demonstrated that larger PEGs stabilized compact RNA structures and strengthened ligand-binding as compared to RNAs in buffer only (22). Therefore, it is unlikely that crowding reagents drastically impacted the selection outcomes. To summarize, any future selection attempts should be initiated under much less stringent selection conditions which should only be increased to become more *in vivo*-like after observing significant CoAzyme and PPAT capping activity.

MATERIALS & METHODS

RNA Transcripts

DNA templates (Table 5.1) were ordered from Integrated DNA Technologies and amplified by PCR using *Pfu* DNA polymerase. Sizes of double-stranded DNA (dsDNA) templates were confirmed by agarose gel electrophoresis. Each RNA was transcribed *in vitro* from the amplified PCR products using the Y639F T7 RNA polymerase (23), *in vitro* transcription buffer (1x = 50 mM Tris-HCl pH 7.5, 15 mM $MgCl_2$, 5 mM DTT, and 2 mM spermidine), and 2 mM each of ATP, UTP, GTP, CTP. During the selection, RNAs were internally radiolabeled during transcriptions and reactions included 0.25 mM CTP and 1

μL of 250 μCi of α^{32}P CTP (Perkin Elmer). Transcription reactions were incubated at 37°C overnight (approximately 16 hrs) and terminated by the addition of denaturing gel loading dye (90% formamide, 50 mM EDTA and 0.01% of xylene cyanol and bromophenol blue). Transcripts were subsequently purified by denaturing polyacrylamide gel electrophoresis (5-8% TBE-PAGE, 8 M urea). Bands corresponding to the expected product sizes were visualized by UV shadow, excised from the gel, and eluted by tumbling overnight at 4°C in 300 mM sodium acetate pH 5.4. Eluates were ethanol precipitated, resuspended in nuclease-free water, and stored at -20°C until further use. A NanoDropOne spectrophotometer (Thermo Fisher Scientific) was used to determine specific RNA concentrations for all assays.

Mutagenic PCR

Rounds amplified by mutagenic PCR are indicated in Table 5.2 and mutagenic PCR protocol was based loosely on (24, 25). Mutagenic PCR reactions include 1X mutagenic PCR mix (7 mM $MgCl_2$, 50 mM KCL, 10 mM Tris pH 8.3, 0.01% glycerol), 0.5 mM $MnCl_2$, 1x dNTP mix (0.2 mM dGTP, 0.2 mM dATP, 1 mM dCTP, 1 mM dTTP), 250 pmol of forward and reverse primer, 20 U Taq DNA polymerase, and water to a final volume of 100 μL. A regular, non-mutagenic PCR is set up as described above and 12.5 μL of the PCR product is used at template for the first mutagenic PCR, amplified for 5 cycles. Then 12.5 μL of the first mutagenic PCR is used as template for a second mutagenic

PCR for 5 cycles. This process is repeated until four mutagenic PCRs have been completed. 100 μL of the fourth mutagenic PCR are used as template for the transcription reaction.

PPAT protein purification

The plasmid encoding for *coaD* gene (addgene # 50388)(26) was transformed into Bl21(DE3) strain of *E. coli*. Protein purification was performed as previously described (27). A single colony was isolated and grown in 2xYT/Kan media until OD600 reached 0.6. The protein expression was induced by adding 500 μM IPTG at 37 °C for 4 hrs. Cells were harvested, lysed by sonication (20 sec on, 40 sec rest on ice, 5 cycles), and centrifuged at 40,000 xg to clear the lysate. The supernatant was loaded into a Ni-NTA resin preequilibrated with 50 mM Tris-HCl, 300 mM NaCl and 10 mM Imidazole pH 8.0, washed extensively and the bound protein was eluted by same buffer containing 200 mM imidazole. Membrane filters of 10,000 Da cut-off were used to remove imidazole, protein concentration was quantitated by UV spectroscopy and purified PPAT was stored in 50% glycerol at -80 °C until further use.

APM gels and elution

The [(N-Acryloylamino) Phenyl] Mercuric Chloride (APM) stock solution was made as previously described (28). To pour an APM gel, the first layer of polyacrylamide (about 20 mL) was poured in a gel casing standing upright. 1 mL of MilliQ water is added directly after pouring the bottom layer to create a smooth interface between layers. The first layer

polymerizes for approximately 30 min. The second, APM containing layer consists of 1 mL of polyacrylamide, 1 μL TEMED, 10 μL 10% ammonium persulfate (APS), and 200 μL of APM stock solution. After removing the 1 mL of water from the gel casing, the APM layer is added following by a fresh 1 mL of water. After the gel to polymerizes for approximately 30 min, the excess water is removed and the final 10 mL layer of polyacrylamide is added and the wells are inserted. After polymerizing for 30 min the APM gel is ready to be run in 1X TBE at 30 watts. Gel pieces were 'minced' into very small pieces and added to 1X APM elution buffer (0.5 M ammonium acetate, 0.5 M DTT, 10 mM EDTA) and tumbled overnight at 4°C. The gel slurry was then loaded into a pre-wetted 100kDa molecular weight cutoff filter (ThermoFisher Scientific) and spun at 14,000 xg for 15 min. The columns were washed with 200 μL of 1X APM elution buffer and spun at 14,000 xg for an additional 15 min. The flow-through was collected and placed into a fresh tube. 1 μL of glycogen and 1 mL of cold ethanol to ethanol precipitate the CoA-RNA. RNA was resuspended in nuclease-free water and stored at -20°C until further use.

CoAzyme and PPAT substrate selection

To generate the starting library, six different library designs were amplified in a 2.5 mL PCR for 5 rounds and transcribed in 2.5 mL reactions independently. For the first round of selection, 1 nmol of each library design was pooled to form a single starting library of 6 nmol (~4.8 x 10^{15} molecules) and subsequent rounds used 1 nmol of pooled RNA (~8 x 10^{14} molecules). Round 1 RNA was incubated at 37°C for 24 hrs in 1X Selection Buffer

(21, 29) (0.1 mM EDTA, 1 mM TCEP, 96 mM glutamic acid, 13.3 mM Mg^{2+}, 150 mM K^{+}, 10 mM Na^{+}, 67.6 mM Cl^{-}, 20% PEG8000, pH 7.5), 3.5 mM phosphopantetheine, and 50 μM PPAT (for PPAT substrate trajectory only) in a volume of 2.5 mL. Subsequent rounds had reaction volumes of 1 mL. Phosphopantetheine (pPant) was synthesized and provided by the Krishnamurthy group. The pPant concentration and reaction time were reduced over the course of the selection to increase stringency (Table 5.2). We purified CoA-RNAs from unreacted RNAs by APM gel partitioning. The gels were phosphorimaged using a Typhoon FLA 9000 and the CoA-RNAs were excised and purified from the APM layer as described above. For various timepoints, 25 μL aliquots of the reaction were removed and immediately terminated with 1X denaturing gel loading dye and stored at -20°C until they were loaded and run on an APM gel. Unreacted RNA and CoA-RNA in the APM layer (reacted RNA) were quantified by measuring band intensities using Multigauge software (Fujifilm) and plotted in Prism. Recovered RNA was reverse transcribed using BST 3.0 DNA Polymerase (New England Biolabs). RT reactions included recovered RNA, 50 pmol of reverse primer, 1X isothermal buffer (New England Biolabs), 6 mM $MgSO_4$, 1.6 mM dNTP mix, and 40 U BST enzyme in a 100 μL reaction volume, incubated at 72°C for 1 hr, and were heat inactivated at 80°C for 20 min. 100 μL of the cDNA was amplified by PCR using Pfu DNA polymerase to generate transcription template for the next round of selection. Some rounds were subjected to mutagenic PCR (described above) as indicated by Table 5.2.

HTS sequencing and data analysis

Libraries were prepared for sequencing using a series of PCR steps to append Illumina adapters and sequencing indices for multiplexing of the libraries as previously described (30). Primers used to append the Illumina adapters and sequencing indices can be found in Table 5.9. Sequencing was performed on an Illumina NextSeq 500 (University of Missouri Genomics Technology Core). Although paired-end reads generated Read 1 and Read 2 for each selection round, a single 150 nt read provided enough coverage such that no additional information was gained through Read pairing. Populations were demultiplexed with respect to Rnd and selection trajectory, and the relevant sequence information was found and used from Read 1 (5′ constant region, 90 nt library sequence, 3′ constant region) and all data shown represent reads from Read 1 only. Data preprocessing was performed using cutadapt (31) to trim 5′ and 3′ constant regions and to discard any uncut sequences or sequences with lengths not within ± 9 nt of the expected size (90 nt) after trimming (Supplemental Tables 1, 3, and 5). These populations were then analyzed using FASTAptameR 2.0 (19, 20) to count and normalize reads (FASTAptameR-Count) and find library motifs (FASTAptameR-Motif Search) (Tables 5.5 - 5.8) across multiple rounds to calculate library RPM [(total reads that contained library motif)/(total reads in round) * 1,000,000] and enrichment values (ratio of RPM in round y relative to RPM in round x). Using interactive 'reads per rank' plotting (FASTAptameR-Count), line plots of reads for each unique sequence (sorted by rank) were generated for various selection rounds. To observe enrichement of individual sequences across populations, two FASTA files

generated from FASTAptameR-Count were compared to generate scatter plots of RPM (FASTAptameR-Enrich).

Name	Sequence (5’ → 3’)
Forward Primer for HTS	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGACCGGCCTAAACGGCATT
Reverse Primer 1 for HTS	CAGACGTGTGCTCTTCCGATCTGTATCCGCCGAAGCGGATTGG
Reverse Primer 2 for HTS	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTATCCGCCGAAGCGGATTGG
5’ Universal HTS Adapter	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
3’ Indexed HTS Adapter	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Table 5.9. Sequences for the high-throughput sequencing primers used to append the Illumina adapters and their respective sequencing indices. We used the NEBNext Index (1-38) Primers for Illumina. 38 reverse primers (corresponding to the 38 indices) were used in the second PCR for high-throughput sequencing preparation. The index region is indicated by the six red N region in the reverse primer 2 for HTS sequence. Index sequences are from the instruction manual for the NEBNext Multiplex Small RNA Library Prep Set 1, Set 2, Index Primers 1-48 and Multiplex Compatible (https://rb.gy/0dbqe9).

ACKNOWLEDGEMENTS

We would like to thank members of the Burke and Lange laboratories for suggestions and constructive feedback throughout the project.

REFERENCES

1. Huang, F. (2003) Efficient incorporation of CoA, NAD and FAD into RNA by in vitro transcription. *Nucleic Acids Res.* 10.1093/nar/gng008
2. Bird, J. G., Zhang, Y., Tian, Y., Panova, N., Barvík, I., Greene, L., Liu, M., Buckley, B., Krásný, L., Lee, J. K., Kaplan, C. D., Ebright, R. H., and Nickels, B. E. (2016) The mechanism of RNA 5′ capping with NAD+, NADH and desphospho-CoA. *Nature*. **535**, 444–447
3. Coleman, T. M., and Huang, F. (2002) RNA-catalyzed thioester synthesis. *Chem. Biol.* **9**, 1227–1236
4. Huang, F., Bugg, C. W., and Yarus, M. (2000) RNA-catalyzed CoA, NAD, and FAD synthesis from phosphopantetheine, NMN, and FMN. *Biochemistry*. **39**, 15548–15555
5. Huang, F. (2003) Efficient incorporation of CoA, NAD and FAD into RNA by in vitro transcription. *Nucleic Acids Res.* 10.1093/NAR/GNG008
6. Kowtoniuk, W. E., Shen, Y., Heemstra, J. M., Agarwal, I., and Liu, D. R. (2009) A chemical screen for biological small molecule-RNA conjugates reveals CoA-linked RNA. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7768–7773

7. Chen, Y. G., Kowtoniuk, W. E., Agarwal, I., Shen, Y., and Liu, D. R. (2009) LC/MS analysis of cellular RNA reveals NAD-linked RNA. *Nat. Chem. Biol.* **5**, 879–881

8. Jiao, X., Doamekpor, S. K., Bird, J. G., Nickels, B. E., Tong, L., Hart, R. P., and Kiledjian, M. (2017) 5′ End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding. *Cell*. **168**, 1015-1027.e10

9. Kiledjian, M. (2018) Eukaryotic RNA 5′-End NAD+ Capping and DeNADding. *Trends Cell Biol.* **28**, 454–464

10. Winz, M. L., Cahová, H., Nübel, G., Frindert, J., Höfer, K., and Jäschke, A. (2017) Capture and sequencing of NAD-capped RNA sequences with NAD captureSeq. *Nat. Protoc.* **12**, 122–149

11. Walters, R. W., Matheny, T., Mizoue, L. S., Rao, B. S., Muhlrad, D., and Parker, R. (2017) Identification of NAD+ capped mRNAs in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 480–485

12. Thauer, R. K., Jungermann, K., and Decker, K. (1977) Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol. Rev.* **41**, 100

13. Coleman, T. M., and Huang, F. (2002) RNA-catalyzed thioester synthesis. *Chem. Biol.* **9**, 1227–1236

14. Komarova, N., and Kuznetsov, A. (2019) Inside the black box: What makes Selex better? *Molecules*. 10.3390/molecules24193598

15. Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006) Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc. 2006 13*. **1**, 1610–1616

16. Porter, E. B., Polaski, J. T., Morck, M. M., and Batey, R. T. (2017) Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors. *Nat. Chem. Biol.* **13**, 295–301

17. Kulshina, N., Baird, N. J., and Ferré-D'Amaré, A. R. (2009) Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nat. Struct. Mol. Biol. 2009 1612*. **16**, 1212–1217

18. Izard, T., and Geerlof, A. (1999) The crystal structure of a novel bacterial adenylyltransferase reveals half of sites reactivity. *EMBO J.* **18**, 2021–2030

19. Alam, K. K., Chang, J. L., and Burke, D. H. (2015) FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Mol. Ther. Nucleic Acids*. **4**, e230

20. Kramer, S. T., Gruenke, P. R., Alam, K. K., Xu, D., and Burke, D. H. (2022) FASTAptameR 2.0: A web tool for combinatorial sequence selections. *Mol. Ther. Nucleic Acids*. **29**, 862–870

21. Yamagami, R., Sieg, J. P., and Bevilacqua, P. C. (2021) Functional Roles of Chelated Magnesium Ions in RNA Folding and Function. *Biochemistry*. **60**, 2374

22. Tyrrell, J., Weeks, K. M., and Pielak, G. J. (2015) Challenge of Mimicking the Influences of the Cellular Environment on RNA Structure by PEG-Induced

Macromolecular Crowding. *Biochemistry*. **54**, 6447–6453

23. Sousa, R., and Padilla, R. (1995) A mutant T7 RNA polymerase as a DNA polymerase. *EMBO J.* **14**, 4609–4621

24. Cadwell, R. C., and Joyce, G. F. (1994) Mutagenic PCR. *Genome Res.* 10.1101/gr.3.6.S136

25. Bartel, D. P., and Szostak, J. W. (1993) Isolation of new ribozymes from a large pool of random sequences. *Science (80-.).* **261**, 1411–1418

26. Strauss, E., and Begley, T. P. (2002) The antibiotic activity of N-pentylpantothenamide results from its conversion to ethyldethia-coenzyme A, a coenzyme A antimetabolite. *J. Biol. Chem.* **277**, 48205–48209

27. Miller, J. R., Ohren, J., Sarver, R. W., Mueller, W. T., De Dreu, P., Case, H., and Thanabal, V. (2007) Phosphopantetheine adenylyltransferase from Escherichia coli: Investigation of the kinetic mechanism and role in regulation of coenzyme A biosynthesis. *J. Bacteriol.* **189**, 8196–8205

28. Biondi, E., and Burke, D. H. (2012) Separating and analyzing sulfur-containing RNAs with organomercury gels. *Methods Mol. Biol.* **883**, 111–120

29. Yamagami, R., Bingaman, J. L., Frankel, E. A., and Bevilacqua, P. C. Cellular conditions of weakly chelated magnesium ions strongly promote RNA stability and catalysis. 10.1038/s41467-018-04415-1

30. Ditzler, M. A., Lange, M. J., Bose, D., Bottoms, C. A., Virkler, K. F., Sawyer, A. W., Whatley, A. S., Spollen, W., Givan, S. A., and Burke, D. H. (2013) High-

throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase. *Nucleic Acids Res.* **41**, 1873–1884

31. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10–12

Chapter 6: Frontiers & Perspectives

The work from this thesis provides new insights into methods for isolating and identifying cellular CoA-RNAs as well as mechanisms by which CoA-RNA can be generated both *in vitro* and *in vivo*. Additionally, this work investigated the impact of library design, selection conditions, and reveres-transcription bias on selection outcomes.

Main Conclusions, Insights, and Future Directions

Chapter 2

Previous studies have established the significant impact of reverse transcriptase (RT) bias on selection outcomes, especially when structural elements are incorporated into a starting library design (1). Specifically, amplification biases from reverse transcription can overwhelm enrichment of sequences displaying the function of interest and leave some functional sequences at a disadvantage, with cumulative effects across multiple rounds of selection. This work provides an initial roadmap for evaluating RTs for amplification biases, focusing on simple measures of yield and processivity, selection outcomes, and fidelity. In this thesis, I directly compared five RTs –ImProm-II, Marathon RT (MaRT), TGIRT-III, SuperScript IV (SSIV), and BST 3.0 DNA polymerase (BST) – to determine which enzymes introduced the least bias. Of the five RTs tested, BST was the most notable with excellent processivity and yield and little bias among templates of varying structure. Furthermore, BST also performed well on long, highly structured viral RNAs. This study also directly compared selection outcomes from an amplification-only selection using

either BST, ImProm-II, or SSIV for reverse transcription. Six RNA libraries containing either strong, moderate, or no incorporated structural elements were pooled and competed head-to-head in six rounds of selection. High-throughput sequencing (HTS) data indicated that BST introduced low inter-library and mutational bias over the course of six rounds of selection. The HTS analysis also demonstrated that BST maintained the mot neutral enrichment values.

Overall, of the five RTs we tested, BST exhibited the lowest bias between structured and non-structured templates and retained high processivity and activity to generate large quantities of full length cDNA product for templates of varying sizes and structures, including good yields for large, structured RNAs ranging from 100-350 nt. The amplification-only selection results indicate that BST is a good candidate for cDNA synthesis during *in vitro* selections, as it introduced the least amount of inter-library bias between differently structured templates. Furthermore, BST is inexpensive and easy to handle with simple reaction conditions, standardized pre-made buffers and solutions, and straight-forward quenching procedures, making it attractive choice for *in vitro* selections in addition to its well-established use for loop-mediated isothermal amplification (LAMP) (2) and potential extrapolation to broader applications, such as RNA structural probing and cellular RNA library preparation for transcriptomics analysis.

Since I began this study, several new RTs have were reported in the literature that have potential applications in specialized selections, including RT-C8 evolved from a variant of the DNA polymerase from *Thermococcus gorgonarius* from the Holliger and Taylor groups for use with XNAs (3, 4) and RTX (an RT evolved *in vitro* from the B family DNA polymerase KOD) from the Ellington group (5), among others. Early reports of these RTs note robust activity on difficult RNA templates and in some cases their use in library amplification for *in vitro* selections. Although RT-C8 and RTX can be purified in house from bacteria carrying the appropriate plasmids, these enzymes are not yet commercially available and their potential impact on amplification bias with respect to structured templates is unknown. However these new RTs, and the directed evolution techniques to generate them, represent a new and interesting direction for the field. It is possible that new and improved RTs that reverse transcribe even more difficult (long, structured) template RNAs will be reported in the coming years. In the future, it may be useful to directly compare these new RTs to BST.

Chapter 3

NAD+ and CoA capped RNAs were identified in cells more than a decade ago (6, 7). Since their initial discovery, various studies have elucidated the identities and mechanisms of biogenesis of NAD-RNAs (8–13). Previous studies have described a co-transcriptional method to generate NAD-RNA, where NAD+ serves as a non-canonical initiator nucleotide (NCIN) during transcription and is incorporated into the +1 position of

transcripts (14). However, co-transcriptional capping is an unlikely method to generate cellular CoA-RNA due to low intracellular concentration 3' dephospho CoA (dpCoA), the required NCIN for co-transcription. Therefore, an alternative method of capping, like post-transcriptional capping, may be more feasible for synthesis of endogenous CoA-RNAs. Thus, this study established *in vitro* and possible *in vivo* post-transcriptional CoA capping of RNA by enzyme phosphopantetheine adenylyltransferase (PPAT).

PPAT is an enzyme of CoA biosynthetic pathway that catalyzes dpCoA from substrates 4' phosphopantetheine (pPant) and ATP. This work established that ATP-RNA can serve as a non-canonical substrate to PPAT to generate CoA-RNA *in vitro*. Testing various RNA substrates, we determined the essential features of RNA for it to serve as a PPAT substrate: 4-10 unpaired nucleotides at the 5' terminus, an A at the +1 position, and a 5' triphosphate. From these data, we speculate that PPAT likely recognized pppA-RNA as an ATP analog. Interestingly, when testing binding of various RNA substrates, no significant differences in binding affinities were observed between PPAT and +1A, +1G, or 5'OH (+1A) RNA, indicating that productive enzymatic recognition is likely driven by local positioning effects and not by overall binding affinity. Furthermore, I observed possible CoA-capping by PPAT *in vivo*. Dual bacterial expression of candidate RNAs with different 5' structural features, followed by CoA-RNA CaptureSeq, revealed >10-fold enrichment of the better PPAT substrate, consistent with *in vivo* CoA-capping of RNA transcripts by PPAT.

Overall, this study established post-transcriptional RNA capping as a possible mechanism for the biogenesis of CoA-RNAs in bacteria.

To further strengthen these observations, future work could determine if RNAI, a natural RNA known to be capped by NAD+ *in vivo* and observed to serve as a PPAT RNA substrate *in vitro*, would preferentially accumulate in the sulfur-contain fraction relative to other RNAs. Additionally, it would be interesting to observe the impact of *in vivo* CoA capping under stationary phase conditions with reduced ATP concentrations, as ATP was observed to inhibit *in vitro* PPAT capping of RNA substrates. Overall, these data represent a big step forward to establishing mechanisms of biogenesis of cellular CoA-RNAs and a small step towards identifying CoA-RNAs. This work may also help shift the field's focus away from co-transcriptional biogenesis models for generating CoA-RNA. It is interesting to speculate that future work from other groups could manipulate PPAT expression *in vivo* to increase quantities of CoA-capped RNAs in cells, allowing for effective capture and identification of those species.

Chapter 4

In addition to exploring the mechanisms by which CoA-RNA could be generated, this thesis also includes the development of a method to elucidate the identities of cellular CoA-RNAs.

The work outlined in this chapter describes a method to separate CoA-RNAs from total cellular RNA, prepare them for high-throughput sequencing (HTS), and subsequently use the HTS data to determine the identities of CoA capped RNAs. Unfortunately, I observed that current methods for ligating pre-adenylated sequencing adapters to RNA were extremely inefficient and highly biased against RNAs with 3' end structures, negatively impacting the outcome of this method. Other groups have observed similar ligation results and reported that the two terminal bases on the 3' adapter and structural elements on the 3' of the RNA significantly impact ligation efficiency (15–18). Ultimately, despite significant troubleshooting and several attempts to identify CoA-RNAs, the CoA Capture Seq method was unsuccessful in determining the identities of endogenous CoA-RNAs.

However, we developed several key components of the CoA Capture Seq method which was sufficient for capturing and identifying specific, known CoA-RNAs from cells. Specifically, the CoA Capture Seq method was adapted and used in chapter 3 to successfully isolate known CoA-RNA sequences from cells by omitting the adapter ligation step, demonstrating its usefulness and viability. However, the key remaining challenge for this method's development is the low efficiency of ligation of pre-adenylated adapters to RNA templates in combination with already low CoA-RNA quantities. Additionally, the issues with ligation efficiency are further exasperated by the RT-PCR steps, which rely on the presence of the 3' sequencing adapter for amplification. Without it, non-specific PCR primer-dimer type complexes form and make up the majority of the

sequencing data. Therefore, additional work to improve the ligation efficiency will be crucial for this method's success for the purpose of identifying unknown CoA-RNA sequences. Future work could optimize the ligation step by incorporating randomized adapter pools to combat ligation bias as previously reported (16–18). However, this method would only address ligation bias, not ligation efficiency which remains a significant issue. Further optimization of the ligation reaction parameters including PEG%, adapter saturation, temperature, and reaction time could be useful for better optimize the ligation step; however, other groups have performed similar studies (19) from which we established our initial ligation protocol.

Although not the original intended purpose, this method has been successfully used for isolating known sequences of CoA-RNA from cells which provided novel and interesting insights about *in vivo* CoA capping mechanisms (chapter 3). This method can continue to be used for alternative purposes as described, or perhaps with thorough troubleshooting, the ligation efficiency can be resolved and the CoA Capture Seq can be used to identify naturally occurring CoA-RNAs. Overall, this method has already successfully provided insights about capping mechanisms of cellular CoA-RNA and with future optimization, this method may be capable of elucidating the identities cellular CoA-RNAs as well.

It is possible that future work to identify CoA-RNAs may focus on specific cellular RNAs, such as RNAI, instead of trying to pull out fewer, unknown CoA-RNAs from total cellular

RNA. Specifically, it may be useful to focus on RNAs which are known to be capped by NAD+ and RNAs which have the ability to be capped with CoA (i.e. they contain a ATP in the +1 position of the transcript). Focusing on specific, known RNAs provides an advantage as it effectively eliminates the adapter ligation step, which I (and others) demonstrated to be a major bottleneck in the capture method. Alternatively, the field could continue to focus on capturing unknown CoA-RNA transcripts from total cellular RNA, in which case I would expect the development of more sensitive capture methods with lower limit of detections for CoA-RNAs.

Chapter 5

Some mechanisms to generate CoA-RNA have been tested and verified *in vitro* including co-transcriptional capping (14, 20), post-transcriptional capping by PPAT (Chapter 3), and self-capping by ribozyme catalysis (21, 22). However, the previous studies which selected and identified self-capping ribozymes were performed under non-biological conditions, and the studies which demonstrated *in vitro* co-transcriptional CoA capping utilized dpCoA concentrations more than ~50 fold greater than the predicted intracellular concentration of dpCoA. Thus it remained unclear if these mechanisms were feasible for *in vivo* generation of CoA-RNA. Therefore, in this study I performed selections under *in vivo*-like conditions to identify self-capping CoA ribozymes (CoAzymes) and RNAs that can serve as a substrate to be post-transcriptionally capped by PPAT under cellular conditions. Two goals of this selection were to generate RNAs which retained functionality in cells and to better

understand possible mechanisms of CoA-biogenesis by identifying RNAs capable of self-capping or serving as a substrate for post-transcriptional capping by PPAT. Once no significant increase in CoAzyme or PPAT capping activity was observed after 12 rounds of selection, I prepared several selection rounds for HTS to analyze the data and evaluate the library pool's progression throughout the course of the selection. HTS analysis revealed no enrichment or convergence and high sequence diversity in all rounds across the selection, ultimately indicative of a failed selection.

The HTS analysis revealed no convergence or enrichment of specific sequences or clusters of sequences. Some enrichment of specific sequences is expected if active sequences are slowly enriching during the duration of a selection, thus, the obvious lack of enrichment points an unsuccessful selection. Additionally the diversity of sequence reads in each round was also inconsistent and the enrichment analysis revealed the inconsistencies in population structure throughout the selection. There were very few of the same sequence in one round to the next, resulting in virtually no enrichment of depletion of specific sequences. This may be indicative of a random, rather than intentional capture, of RNAs during the partition step or it could indicate that there is specific capture of sequences that are being non-specifically capped. Overall, the lack of enrichment and convergence accompanied by high levels of sequence diversity and population inconsistencies between rounds lead me to conclude the selection failed.

I speculate that selection's failure is a result of overly stringent selection parameters. The types and concentrations of ions included in a selection buffer are known to have significant impacts on RNA structure, RNA-ligand interactions, and RNA (23–26). I used an *in vivo* like buffer where the types and concentrations of ions included were drastically reduced, which may be responsible for the absence of CoA capping activity. Future selection attempts may benefit from making use of alternate partition methods, such as thiopropyl sepharose columns (20), to decrease the amount of non-specific capture that occurs from RNAs collecting at the mercury-layer interface of mercury-layered PAGE gels. Additionally, I recommend initiating future selections using buffers with higher, non-biologically relevant ion concentrations and slowly titrating in the *in vivo*-like buffer, providing sequences time to evolve. Overall, despite the selection failing, this study provided insights about the impact of selection conditions on selection outcomes which can be used to generate fresh strategies for selecting for RNAs which retain activity *in vivo*.

REFERENCES

1. Porter, E. B., Polaski, J. T., Morck, M. M., and Batey, R. T. (2017) Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors. *Nat. Chem. Biol.* **13**, 295–301
2. Padzil, F., Mariatulqabtiah, A. R., Tan, W. S., Ho, K. L., Isa, N. M., Lau, H. Y., Abu, J., and Chuang, K. P. (2022) Loop-mediated isothermal amplification (Lamp) as a promising point-of-care diagnostic strategy in avian virus research. *Animals*.

10.3390/ani12010076

3. Houlihan, G., Arangundy-Franklin, S., Porebski, B. T., Subramanian, N., Taylor, A. I., and Holliger, P. (2020) Discovery and evolution of RNA and XNA reverse transcriptase function and fidelity. *Nat. Chem. 2020 128*. **12**, 683–690
4. Hervey, J. R. D., Freund, N., Houlihan, G., Dhaliwal, G., Holliger, P., and Taylor, A. I. (2022) Efficient synthesis and replication of diverse sequence libraries composed of biostable nucleic acid analogues. *RSC Chem. Biol.* **3**, 1209–1215
5. Choi, W. S., He, P., Pothukuchy, A., Gollihar, J., Ellington, A. D., and Yang, W. (2020) How a B family DNA polymerase has been evolved to copy RNA. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 21274–21280
6. Kowtoniuk, W. E., Shen, Y., Heemstra, J. M., Agarwal, I., and Liu, D. R. (2009) A chemical screen for biological small molecule-RNA conjugates reveals CoA-linked RNA. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7768–7773
7. Chen, Y. G., Kowtoniuk, W. E., Agarwal, I., Shen, Y., and Liu, D. R. (2009) LC/MS analysis of cellular RNA reveals NAD-linked RNA. *Nat. Chem. Biol.* **5**, 879–881
8. Julius, C., Riaz-Bradley, A., and Yuzenkova, Y. (2018) RNA capping by mitochondrial and multi-subunit RNA polymerases. *Transcription*. **9**, 292–297
9. Jiao, X., Doamekpor, S. K., Bird, J. G., Nickels, B. E., Tong, L., Hart, R. P., and Kiledjian, M. (2017) 5′ End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding. *Cell*. **168**, 1015-

1027.e10

10. Julius, C., and Yuzenkova, Y. (2019) Noncanonical RNA-capping: Discovery, mechanism, and physiological role debate. *Wiley Interdiscip. Rev. RNA.* **10**, e1512

11. Julius, C., and Yuzenkova, Y. (2017) Bacterial RNA polymerase caps RNA with various cofactors and cell wall precursors. *Nucleic Acids Res.* **45**, 8282–8290

12. Jäschke, A., Höfer, K., Nübel, G., and Frindert, J. (2016) Cap-like structures in bacterial RNA and epitranscriptomic modification. *Curr. Opin. Microbiol.* **30**, 44–49

13. Wang, J., Alvin Chew, B. L., Lai, Y., Dong, H., Xu, L., Balamkundu, S., Cai, W. M., Cui, L., Liu, C. F., Fu, X. Y., Lin, Z., Shi, P. Y., Lu, T. K., Luo, D., Jaffrey, S. R., and Dedon, P. C. (2019) Quantifying the RNA cap epitranscriptome reveals novel caps in cellular and viral RNA. *Nucleic Acids Res.* 10.1093/NAR/GKZ751

14. Bird, J. G., Zhang, Y., Tian, Y., Panova, N., Barvík, I., Greene, L., Liu, M., Buckley, B., Krásný, L., Lee, J. K., Kaplan, C. D., Ebright, R. H., and Nickels, B. E. (2016) The mechanism of RNA 5′ capping with NAD+, NADH and desphospho-CoA. *Nature.* **535**, 444–447

15. Jayaprakash, A. D., Jabado, O., Brown, B. D., and Sachidanandam, R. (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res.* **39**, e141

16. Sorefan, K., Pais, H., Hall, A. E., Kozomara, A., Griffiths-Jones, S., Moulton, V., and Dalmay, T. (2012) Reducing ligation bias of small RNAs in libraries for next

generation sequencing. *Silence*. **3**, 4

17. Sun, G., Wu, X., Wang, J., Li, H., Li, X., Gao, H., Rossi, J., and Yen, Y. (2011) A bias-reducing strategy in profiling small RNAs using Solexa. *RNA.* **17**, 2256–2262

18. Zhang, Z., Lee, J. E., Riemondy, K., Anderson, E. M., and Yi, R. (2013) High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biol.* **14**, R109

19. Song, Y., Liu, K. J., and Wang, T. H. (2014) Elimination of ligation dependent artifacts in T4 RNA ligase to achieve high efficiency and low bias microRNA capture. *PLoS One*. 10.1371/journal.pone.0094619

20. Huang, F. (2003) Efficient incorporation of CoA, NAD and FAD into RNA by in vitro transcription. *Nucleic Acids Res.* 10.1093/nar/gng008

21. Coleman, T. M., and Huang, F. (2002) RNA-catalyzed thioester synthesis. *Chem. Biol.* **9**, 1227–1236

22. Huang, F., Bugg, C. W., and Yarus, M. (2000) RNA-catalyzed CoA, NAD, and FAD synthesis from phosphopantetheine, NMN, and FMN. *Biochemistry*. **39**, 15548–15555

23. Lin, P. H., Chen, R. H., Lee, C. H., Chang, Y., Chen, C. S., and Chen, W. Y. (2011) Studies of the binding mechanism between aptamers and thrombin by circular dichroism, surface plasmon resonance and isothermal titration calorimetry. *Colloids Surf. B. Biointerfaces*. **88**, 552–558

24. McKeague, M., McConnell, E. M., Cruz-Toledo, J., Bernard, E. D., Pach, A.,

Mastronardi, E., Zhang, X., Beking, M., Francis, T., Giamberardino, A., Cabecinha, A., Ruscito, A., Aranda-Rodriguez, R., Dumontier, M., and DeRosa, M. C. (2015) Analysis of In Vitro Aptamer Selection Parameters. *J. Mol. Evol.* **81**, 150–161

25. Amano, R., Takada, K., Tanaka, Y., Nakamura, Y., Kawai, G., Kozu, T., and Sakamoto, T. (2016) Kinetic and Thermodynamic Analyses of Interaction between a High-Affinity RNA Aptamer and Its Target Protein. *Biochemistry*. **55**, 6221–6229

26. Inoue, A., Takagi, Y., and Taira, K. (2004) Importance in catalysis of a magnesium ion with very low affinity for a hammerhead ribozyme. *Nucleic Acids Res.* **32**, 4217–4223

Vita

Jordyn Kaye Lucas, daughter of Lezlee Raiford and John Lucas, was born on June 25, 1994 in Saint Louis, Missouri. She received her high school diploma from Villa Duchesne in 2012. She received her Bachelors of Science in Biochemistry at Purdue University in West Lafayette, Indiana in 2016. Jordyn completed her Biochemistry PhD in April 2023 under the mentorship of Dr. Donald Burke-Agüero.