UNIVERSITY OF CALIFORNIA,
IRVINE

Development of methods for functional RNA characterization and discovery

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Kyle H. Cole

Dissertation Committee:
Professor Andrej Lupták, Chair
Professor Jennifer Prescher
Professor Chang Liu

2024

DEDICATION

To

my friends and my family

all that I am is because of you

“Maybe that’s enlightenment enough: to know that there is no final resting place of the mind; no moment of smug clarity. Perhaps wisdom is realizing how small I am, and unwise, and how far I have yet to go.” — Anthony Bourdain

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My journey into graduate school would not have been possible without the support and guidance of the leadership of the Master's in Biotechnology program in the Department of Molecular Biology and Biochemistry at UC Irvine. After finishing my undergraduate studies, I became unsure of my career trajectory and my passions. On a whim I applied to the Master's in Biotechnology program and was swiftly accepted. The support I had been given throughout that program has really created the scientific foundation that has led me to drafting this very thesis. Dr. Michael Cumsky has had unwavering support for me throughout my time at UC Irvine. Dr. Dana Aswad's passion for research has also been instrumental in my pursuit of the unknown. I am also eternally grateful to my wonderful TAs: Katherine Williams, Kara Paolilli, and Marie True. Thank you for sharing your firsthand experiences in science with me and the rest of my cohort of misfits. And to my cohort: I do not believe I would have succeeded without the friendships we have forged and the difficulties we surmounted together. A special thank you to Sumer Abdul-Hafiz and Amos Fung for being wonderful partners in crime as we mentored the underclassmen. We fought hard and we were able to inspire together. I have been fortunate enough to witness some of our past students succeed and become incredible scientists.

To Dr. Andrej Lupták, thank you for your unwavering support. When we first met, you welcomed me into your office as if I were an old friend of yours. It is an intimidating experience meeting someone as successful and intelligent as you, but you were so down-to-earth that I immediately felt at ease. That interaction led me to want to join your lab. I was already overly excited about the research but knowing that my advisor would be someone I could consider a friend was not expected. When I approached you about possibly pursuing a PhD following the master's program, you showed no apprehension to the idea. It gave me a lot of confidence about my decision. As you know, research is particularly challenging and can take a toll on a graduate student, any encouragement means the world. You never hesitated to be a supporting figure throughout this journey. When the Cellular and Molecular Bioscience program rejected my PhD application, you fought tooth-and-nail to ensure that I was admitted into MBB. I do not know where I would have ended up without that level of support. I will miss our semi-regular discussions and brainstorming sessions. Thank you for cultivating that intellectual creativity.

I appreciate my committee members, Dr. Chang Liu and Dr. Jennifer Prescher for their support. A special thank you to Jen as she has been a great mentor to me over the last few years. I look up to her as someone I aspire to be, and her love of science is unlike any other. I have always appreciated your advice and look forward to maintaining our relationship.

There is an incredible sense of comradery shared amongst people in an academic lab. It is a foreign concept to those that have not been in the trenches. It is almost cliché to discuss but some clichés exist for a reason. I have had the fortune of being amongst some of the absolute best people during my time in the Luptak lab. When I joined, I was apprehensive but had incredible mentorship that undoubtedly led to long-lasting friendships. Bao Ho, Luiz Passalacqua, Michael Vu, and Michael Wu were instrumental in molding me into a half-decent researcher and for pushing me to consider every variable. To this day, I know I can reach out and ask each of you for advice or support. Thank you to Alexandra Kent for being a valuable friend and one I could rely on for encouragement and advice, in addition to fun discussions about the universe. Brandon Latifi, it has been an absolute pleasure watching you evolve into an incredible biochemist, and

you are one of the hardest working people I know. Thank you to Claire Chen for being an absolute inspiration. She is without a doubt the hardest working individual I have ever met in my entire life. I am eternally motivated by her pursuit of science and whenever I am feeling a bit less motivated, I remember the effort she put into every experiment and day. I am also happy to have had the opportunity to work with Jay Gupta, Abigail Vandewalle, and Chely Garfio. While we only worked together for a brief time, I consider each of you a friend and really appreciated your helpful discussions. To Katrina Mogannam, Christelle Chan, and Mia Cao: it has been a pleasure watching each of you grow into fantastic researchers and I was so happy when each of you joined the lab. Thank you for being such great people and making every day a bit more fun.

The bioluminescent RNA tracking group has been the best collaboration I have ever been a part of. Thank you to the Oswald Steward and Jennifer Prescher labs for making every Thursday morning a wonderful time. Some genuinely insightful discussions with plenty of opportunities to produce cool and unique approaches to tracking and understanding RNA biology. Kevin Ng and Lila Halbers were pivotal in my growth as a scientist. Our teamwork was unmatched, and we accomplished so much in such little time. I honestly believe our work will be the foundational technology for live RNA tracking. Lila, thank you for being such a motivated and diligent partner in crime. We overcame some obstacles together and I will never forget what we achieved.

To my parents, Patty and Ron Cole, thank you for always loving me through it all. Your endless emotional, physical, and monetary support is not something I can put into words. You motivate me to be the best person I can possibly be and to push myself to achieve success. A zip code or a state away, my love for you is never-ending. To my sister Kelsey, I will always be your older brother and I will always see you as my baby sister. We have not always seen eye-to-eye, but I know I can always count on you. Thank you for bringing Rex and Reid into my life.

And to my absolute rock, my partner, Allyson Kay. You and Ted are stability when I need it most. You have seen the best and worst of me. You have witnessed me at my wits-end on many occasions, but your unwavering love has kept me grounded. From sleepless nights writing manuscripts or falling asleep on the couch after a long day, you have always been there to pick me back up. Ted, I know you are just a cat, but you have always brightened up my bad days when you come running to the door when I get home. I forgive you for regularly waking me up at 3 AM. No matter what our future holds together, I know the three of us will have the absolute best of times together. Thank you for being such an important part of my life and I am forever grateful for the love and support you have given me over the last four years. Love you both always.

I would like to thank Elsevier, Inc. for allowing me to use Chapter 1 of my thesis which was originally published in *Methods in Enzymology* and titled “High-throughput methods in aptamer discovery and analysis”. Chapter 4 of this thesis was originally published on *bioRxiv* and titled “A modular platform for bioluminescent RNA tracking” and reprinted in this thesis with permission from my co-authors. Chapter 6 of this thesis was originally published in *Scientific Reports* and titled “Single-tube collection and nucleic acid analysis of clinical samples for SARS-CoV-2 saliva testing” and reprinted in this thesis with permission from my co-authors. Full information of co-authors and contribution statements are listed preceding each chapter using previously published or forthcoming material. Funding for this research was provided by NASA ICAR (80NSSC21K0596), NASA Exobiology (80NSSC21K0488), NSF (1804220), and the W.M. Keck Foundation.

VITA

Kyle H. Cole

Education:

2018 – 2024	Ph.D. Biological Sciences, University of California, Irvine
2016 – 2018	M.S. Biotechnology, University of California, Irvine
2010 – 2015	B.S. Biotechnology, Cal Poly Pomona
2006 – 2010	A.A./A.S. General education, Moreno Valley College

Publications:

***Denotes equal contribution

†Denotes review/perspective

6. **Cole, K.H.***, Ho. B.T.*, Passalacqua, L.F.M., Nakata, M., Pham, J., Hua, C., Nguyen, P. (2024). Identification of mammalian cGMP aptamers through human genomic SELEX. *Manuscript in preparation.*

5. **Cole, K.H.***, Latifi, B.*, Vu, M.M.K.*, Lupták, A. (2024). A high-throughput workflow for experimental demarcation of functional RNA domains. *Manuscript in preparation.*

4. Ng, K.K.*, **Cole, K.H.***, Halbers, L.P.*, Chen, C.E.T., Fuller, E.B., Callicoatte, C., Metcalfe, M., Chen, C.C., Barhoosh, A.A., Reid-McLaughlin, E., Metcalfe, M., Kent, A.D., Steward, O., Lupták, A., Prescher, J. (2023). A modular platform for bioluminescent RNA tracking. *bioRxiv*. https://doi.org/10.1101/2022.07.02.498144

3. †**Cole, K.H.**, Mogannam, K., Lupták, A. (2023). Choreography of a self-splicing ribozyme. *Nature Catalysis New & Views.* 6, 291–293. https://doi.org/10.1038/s41929-023-00948-x

2. **Cole, K.H.***, Bouin, A., Ruiz, C. Semler, B.L., Inlay, M.A., Lupták, A. (2022). Single-tube collection and nucleic acid analysis of clinical samples for SARS-CoV-2 saliva testing. *Scientific Reports*, 12, 3951. https://doi.org/10.1038/s41598-022-07871-4

1. †**Cole, K. H.**, & Lupták, A. (2019). High-throughput methods in aptamer discovery and analysis. M*ethods in Enzymology*, Volume 621. https://doi.org/10.1016/bs.mie.2019.02.009

Presentations:

7. **International Conference on Learning and Memory**, Huntington Beach, CA, 2023
   Oral presentation – "*Investigating RNA dynamics with modular bioluminescent tags"*

6. **UCI Synthetic Biology/Chemical Biology Club**, 2022
   Oral presentation – "*A modular platform for bioluminescent RNA imaging"*

5. **UCI School of Pharmacy and Pharmaceutical Sciences Vertex Day**, 2022
   Oral presentation – "*A modular platform for bioluminescent RNA imaging"*

4. **UCI Synthetic Biology/Chemical Biology Club Symposium**, 2022
   Poster presentation – "*Bioluminescent method for tracking RNA dynamics in living systems"*

3. **RNA Society Annual Conference**, Boulder, Colorado, 2022
   Poster presentation – *"Bioluminescent method for tracking RNA dynamics in living systems"*

2. **UCI Pharmaceutical Sciences Retreat**, Lake Arrowhead, CA, 2022
   Poster presentation – *"Development and characterization of an aminoacylating ribozyme with codon specificity"*

1. **UCI Synthetic Biology/Chemical Biology Club Symposium**, 2020, 2021

   Poster presentation – "*Genomic SELEX for human cGMP RNA aptamers"*

Awards and honors:

2021	Robert Warner Award in Nucleic Acid Biochemistry University of California, Irvine, Dept. of Mol. Bio. and Biochem.
2017	Graduate Student Fellowship Award University of California, Irvine, Dept. of Mol. Bio. and Biochem.

Teaching experience:

Winter 2021	Undergraduate microbiology, teaching assistant
Fall 2020	Graduate protein structure and function, teaching assistant
Summer 2020	Undergraduate biochemistry, teaching assistant
Summer 2020	Undergraduate molecular biology, teaching assistant/guest lecturer
Winter 2020	Undergraduate biochemistry, teaching assistant
Winter 2020	Undergraduate microbiology, teaching assistant
Fall 2018	Undergraduate physical biochemistry, teaching assistant
Summer 2018	Undergraduate molecular biology, teaching assistant
Spring 2018	M.S. Biotechnology Management laboratory, head teaching assistant
Winter 2018	M.S. Biotechnology protein isolation and purification, head teaching assistant

Fall 2017	M.S. Biotechnology recombinant DNA technology, head teaching assistant

Skills:

Laboratory:	EMSA, icSHAPE/SHAPE-seq, T1/S1-nuclease footprinting, in-line probing, cleavage-interference mapping, fluorescence anisotropy, isothermal titration calorimetry (ITC), *in vitro* translation assays, RNA-seq/DNA-seq, DNA/RNA library preparation (Illumina), fundamental bioinformatic processing/analysis, PCR/RT-PCR, qPCR/RT-qPCR, LAMP/RT-LAMP, gel electrophoresis (agarose/PAGE), Western/Northern/Southern/dot blot, Gibson assembly/cloning, and DNA/RNA/protein chromatography (IMAC/IE/SEC), fluorescence and bioluminescence microscopy, instrumentation maintenance/setup, mammalian-cell tissue culture.
Software:	Fiji/ImageJ, Cytiva ImageQuant, Integrated Genome Browser (IGB), Integrated Genome Viewer (IGV), UCSF Chimera and ChimeraX, PyMOL, SnapGene, Jalview, Benchling, GraphPad Prism, OriginLab Pro, and Adobe Illustrator

ABSTRACT OF THE DISSERTATION

Development of methods for functional RNA characterization and discovery

by

Kyle H. Cole

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2024

Professor Andrej Lupták, Chair

Our understanding of the roles that non-coding, functional RNAs play in biological processes further develops each year. However, the inherent ability of RNAs to form complex three-dimensional structures that enable them to perform a variety of regulatory and catalytic processes, is what also makes them incredibly difficult to discover and characterize. Even with advancements in technology, new methodologies and techniques are required to better understand and discover their unique functionalities. The work presented within this thesis focuses on the characterization of functional RNAs in addition to the development of approaches to uncover new RNAs. First, I present a study in which a human-genomic selection is performed against the secondary-messenger cGMP and is further characterized. Then, a high-throughput selection-based approach is developed to elucidate the minimum domains necessary for RNA aptamer binding. My third focus was on developing a split-luciferase imaging platform to capture RNA dynamics in live cells and animals via bioluminescence. I also present the development of a synthetic ribozyme with the ability to aminoacylate tRNAs with non-canonical amino acids, and approaches to select for ribozymes with high turnover. Lastly, I present work on the development of an expedient single-tube collection approach for the analysis of viruses in saliva samples.

Chapter 1: Introduction

Publication note:

Part of this chapter was originally published in the book series *Methods in Enzymology.*

Cole, K.H., Lupták, A. High-throughput methods in aptamer discovery and analysis. *Methods in Enzymology* **621,** 329-346 (2019). https://doi.org/10.1016/bs.mie.2019.02.009.

1.1 Functional RNAs

Over the last seventy years, the role RNAs play in biological processes has become increasingly more complex and less defined. The notion that RNA was an intermediary between DNA and protein was consensus across molecular biology in the 1950s. The prevailing hypothesis proposed by Francis Crick suggested that unique RNA complexes, ribosomes[1], were involved in the synthesis of each protein product, and that cells contained thousands of these unique ribonucleic acid structures[2, 3]. In the early 1960s, work by several groups later supported the concept that an intermediary, unstable RNA element was the carrier of genetic information instead[4-6]. The discovery of this unstable RNA element, or messenger RNA (mRNA), convoluted the idea of the singular role RNAs possessed in biological processes as ribosomal RNA (rRNA) did not seem to be involved in the conveyance of the protein-code itself. Recognition of the existence of non-coding RNAs (ncRNAs) was further supported by the identification of the RNA that made up the "soluble" RNA fraction of cell lysates, known as transfer RNA (tRNA)[7].

Today, the number of ncRNAs identified between human and mouse models is estimated to be nearly 20,000[8-10]. The versatility of RNA, and the many structures it can take on, enable it to act as not only a genetic information carrier, but also a *cis*-and-*trans* activator of gene expression, binder of metabolites and/or proteins, and catalyst. This innate ability has given rise to several classes of ncRNAs[11]. In eukaryotes, the majority of ncRNAs can be classified into two major categories: housekeeping RNAs [rRNA, tRNA, splicesomal RNA, and small-nucleolar RNA (snoRNA)], and regulatory RNAs [micro RNA (miRNA), PIWI-interacting RNA (piRNA), small-interfering RNA (siRNA), and long non-coding RNA (lncRNA)] [12]. However, these classifications do not encompass all known functional RNAs in eukaryotes and do not include the diversity of those identified in bacteria and archaea.

In bacteria, regulatory RNA elements commonly found in the 5′-untranslated region (UTR) of genes control expression upon sensing of a metabolite or gene product. These regulators,

known as riboswitch, contain two domains – the aptamer domain which selectively binds a target ligand, and the expression platform[13, 14]. Upon binding or release of the target ligand, the aptamer domain "releases" the expression platform, allowing for changes in expression of the downstream gene at either a transcriptional or translational level (**Figure 1.1A**)[15, 16]. To date, 56 classes of riboswitches have been identified, many of which recognize enzyme cofactors such as *S*-adenosylmethionine[17], flavin mononucleotide[18], and coenzyme B_{12}[19]. While many riboswitches exist for enzyme cofactors, a generous portion of riboswitches recognize precursor molecules of RNAs[20]. This observation adds additional support towards the RNA World hypothesis[21-23]. It is postulated that riboswitches evolved during early life, further supported by their widespread abundance across bacterial and archaea[24].

However, to date, only one example of a riboswitch in eukaryotes (plants and fungi) has been identified, the thiamine pyrophosphate (TPP) riboswitch[25]. The TPP riboswitch in fungi exists within the 3′-UTR of its cognate mRNA and regulates alternative splicing of its gene product[26]. While the TPP riboswitch is the only known example of eukaryotic riboswitch, several groups have identified conserved aptamers within the human genome[27-30]. These human aptamers have been identified in both intronic and intergenic regions, but no apparent function has been identified. The adenosine-binding aptamers identified within intronic regions of the *RAB3C* and *FGD3* genes have been hypothesized to be associated with cotranscriptionally regulated processes, where occupancy of the aptamer by ATP determines the fate of the transcript[27]. This hypothesized mechanism would differ from the *cis* regulated gene expression inherent to traditional riboswitches. It is proposed that eukaryotic aptamers may be remnants of evolution, but that does not discount the possibility of allosteric regulation in eukaryotes by small-molecule metabolites or secondary messengers[31], a mechanism leveraged by many lncRNAs with protein, DNA, and RNA partners[12].

RNA's ability to form complex tertiary structures enables it to have the capacity to catalyze chemical reactions in the form of RNA enzymes known as ribozymes. The discovery of the first ribozymes, *Tetrahymena* group-I intron and RNase P, led to the 1989 Nobel Prize in Chemistry for the determination of the catalytic potential of RNA[32, 33]. Since, several classes of ribozymes have been identified to exist across all kingdoms of life and within viral genomes, including the discovery of that RNA catalyzed peptidyl transfer in the ribosome[34]. Most of these ribozymes catalyze self-cleaving, self-splicing, or self-ligation reactions, with the most studied being the self-cleaving[35]. Self-cleaving ribozymes like the HDV and HDV-like ribozymes, act in *cis,* causing self-scission via a transesterification reaction through general-acid base catalysis[36]. Several self-cleaving ribozymes have been identified within the human genome such as the HDV-like self-cleaving ribozyme found in the intron of the *CPEB3* gene[37] (**Figure 1.1B**) and the self-cleaving *Hovlinc* ribozyme identified in a lncRNA[38]. Recent studies have shown that some ribozymes play critical roles in mammalian biological processes[39]. The widespread occurrence of ribozymes has been identified through bioinformatics-based approaches, with few naturally occurring ribozymes being identified through experimental methods[40-42].

Many of the limitations in experimental discovery of ribozymes, in addition to riboswitches and their aptamers domains, is due to the unique chemistries and activities these RNAs possess. For example, the *Hovlinc* ribozyme was discovered by transcribing a pool of human genomic DNA and enriching for RNA that contained 5′-OH, a common product of self-scission and an uncommon characteristic of most RNA[35, 38]. Additionally, genomic riboswitches and their aptamers can be difficult to identify experimentally because their cognate ligands bind cotranscriptionally[43]. A commonly used approach to assist in the identification of genomic or synthetic ribozymes, aptamers, and riboswitches, is an iterative directed evolution method known as *in vitro* selection[44, 45]. This approach takes advantage of a pool of DNA with theoretical diversity of ~10^{16} unique sequences. This DNA pool is transcribed, and the product RNA pool is allowed then subjected to

conditions which select functional RNAs with the capacity to perform a given function. For aptamers and riboswitches, the pool may be incubated with a ligand covalently attached to a physical substrate, with the hope of capturing active RNAs on the medium. In the case of ribozymes, the desired catalytic phenotype may be selected for by gel electrophoresis or through covalent capture of the product of the reaction. The following section discusses some of these high-throughput selection approaches and the downstream analysis associated with *in vitro* selections.

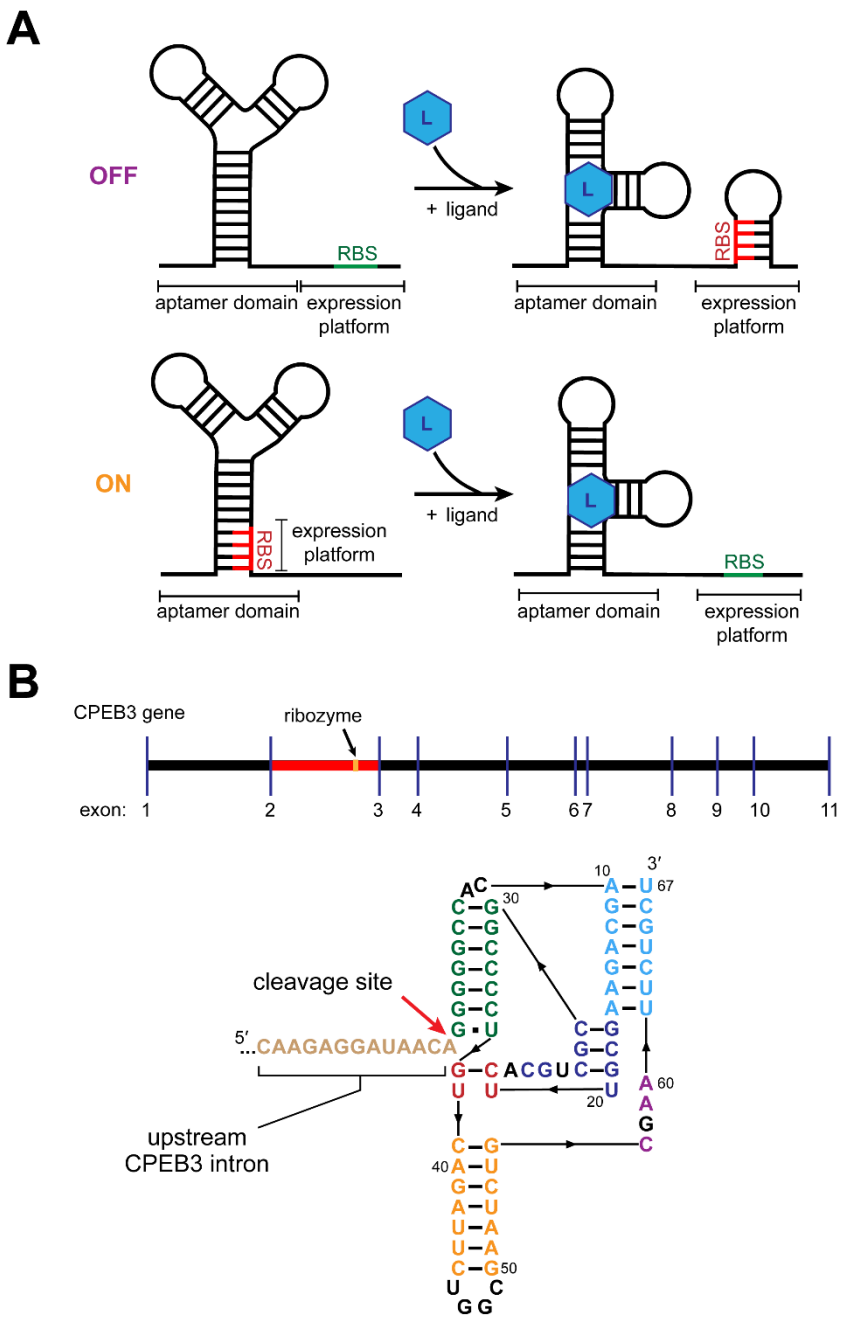


Figure. 1.1: Examples of functional, ncRNAs. (**A**) Examples of two translational riboswitches: Top, an "off" regulator where upon binding of the ligand (blue) to the aptamer domain during transcription, the ribosome binding site (RBS) is sequestered in a hairpin complex making it unavailable to the ribosome. A translational "on" riboswitch, which upregulates expression in the presence of the ligand. (**B**) *CPEB3* ribozyme in the second intron of the *CPEB3* gene. Self-cleavage (red arrow) occurs leading to alternative protein isoforms due to pre-mRNA degradation.

1.2 High-throughput methods in functional RNA discovery and analysis

In vitro selection of functional RNAs has gained popularity as a method for discovery, however, functional RNA discovery is hindered by the necessity of post-selection biochemistry[46]. High-throughput discovery methods aim to solve these issues through combinatorial approaches to identify ligand-targeting or catalytically active sequences while simultaneously determining biochemical and structurally relevant information.

Traditional methods of *in vitro* selection analysis reveal only a fraction of the potential number of RNAs for the target activity due to limitations in sequencing technologies. Target molecules are typically conjugated to agarose beads or oligonucleotide pools were 5′-biotin conjugated and separated with avidin-bound beads. After successive rounds of selection, enriched pools were collected, and restriction enzyme cut sites were introduced by traditional PCR or ligation. Pools were restriction-endonuclease digested, subsequently ligated into cloning vectors or TA-cloned and transformed into *E. coli*. To identify unique sequences, individual bacterial clones were sequenced via traditional dideoxynucleotide methods, such as Sanger sequencing[47, 48]. This strategy proved to be laborious, because individual colonies had to be sequenced to determine which clone was successfully transformed in the bacteria. If a specific sequence or structural motif dominated the enriched pool, the likelihood of identifying less abundant motifs was difficult due to the low throughput of this sequencing approach.

As the cost of high-throughput sequencing (HTS) decreased, deep sequencing was introduced to *in vitro* selection protocols, with the intent of improving functional RNA discovery and increasing data validation through improved sequencing depth[49]. The first applications of HTS to *in vitro* selections were methods for the identification of transcription factors[50-53]. Comparably, Zimmerman et al. developed a method of identifying naturally occurring RNA aptamers from *E. coli* genomic DNA and a similar approach was used to identify naturally occurring ATP-binding RNA aptamers[27, 54]. Cho et al. built on this methodology to select for growth factor specific DNA

aptamers and Ditzler et al. selected for RNA inhibitors of HIV reverse transcriptase[55, 56]. With increased sequencing depth and computational methods, the fitness landscape of an *in vitro* selection could be visualized, and evolutionary steps tracked between rounds of selection adding support for the RNA world hypothesis[57-59]. While these methods incorporate HTS for more expedient functional discovery, recent high-throughput discovery methods incorporate the determination of kinetic rate constants to ease the necessity for biochemical analysis (**Figure 1.2**). The objective of the subsequent sections is to provide an overview of post-selection sequence analysis methods and to provide insight into high-throughput discovery methods.

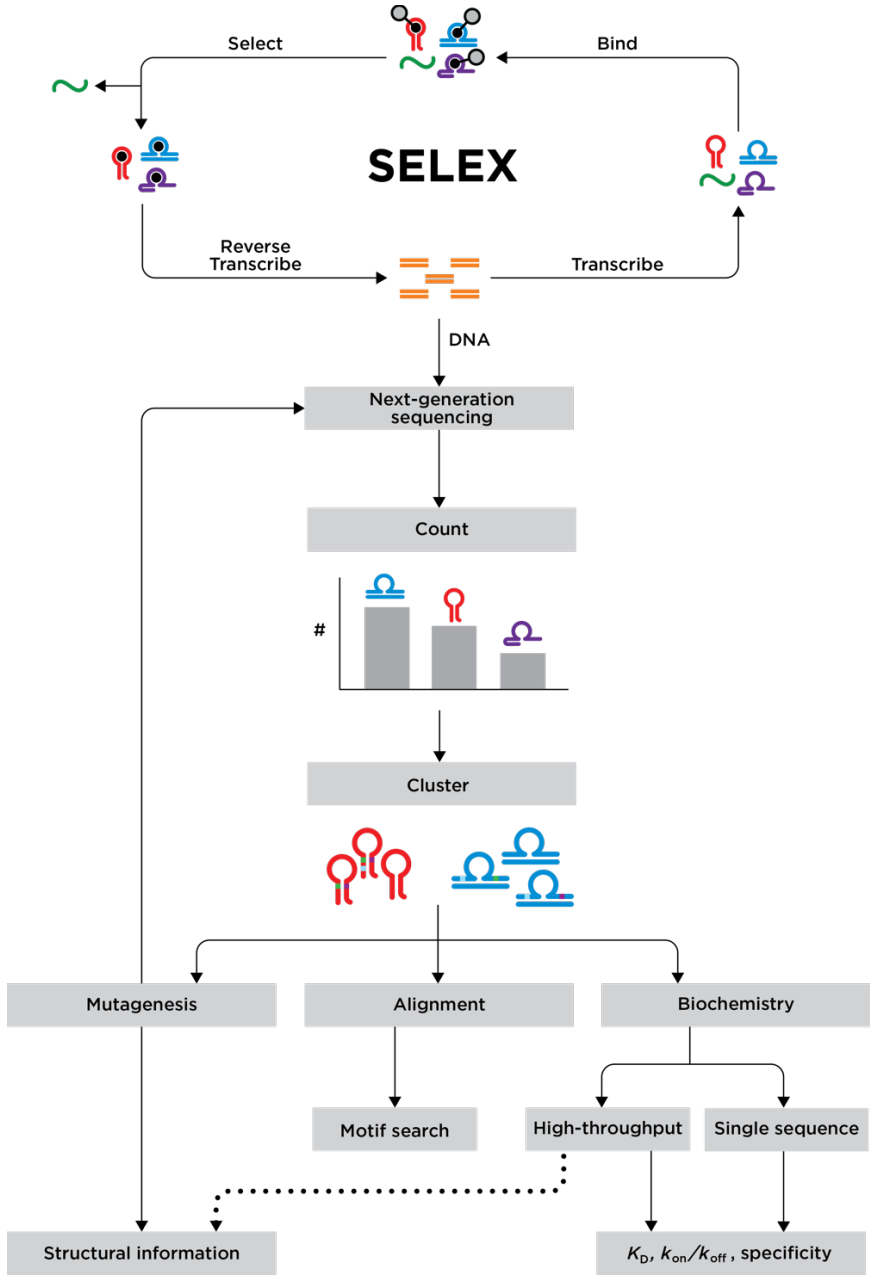


Figure. 1.2: Generalized workflow of in vitro selection (SELEX) of RNA aptamers and post-selection analysis. After several rounds of selection, the enriched pool is amplified and submitted for high-throughput sequencing. After sequencing, the reads are counted and clustered. Clustered sequences provide the basis for downstream analysis, such as mutagenesis, alignment and sequence-motif discovery, and biochemistry.

Post-selection Sequence Analysis

Post-selection analysis of deep sequenced pools can become overwhelming due to the sheer quantity of available sequences, presumably enriched, for active molecules. Software development aimed at tackling this issue has progressed significantly over the last decade with tools intended to reduce the burden of deep sequencing data by providing analysis to determine enrichment and sort sequences based on genotypic familiarity. Early high-throughput sequencing adopters utilized analysis packages designed for genomic sequencing data manipulation in addition to homebrewed UNIX scripts utilizing the *sed*, *awk*, and *grep* commands[60]. The availability of these custom-built scripts was limited through networking connections and typically scripts were built to answer specific experimental questions and were not intended for general enriched pool workup. *In vitro* selections require a degree of bioinformatic wherewithal to fully benefit from the amount of sequencing data that is produced and due to an increased interest in functional RNA discovery, analysis of HTS data for discovery has improved[61]. This section aims to provide a general outline for post-selection analysis and provides resources for enriched pool workup.

In silico parsing of selected sequences provides the foundation of biochemical analysis by identifying sequence families enriched through selective pressure. Analysis traditionally begins by minimizing and sorting redundant sequencing data by identifying unique sequences and sorting them based on shared sequence characteristics. Pitt et al. were among the first groups to develop high-throughput sequencing analysis packages aimed at *in vitro* selection data workup. Their package, Sequence Evolution With Adaptive Landscape (SEWAL), assesses the overall fitness landscape of *in vitro* selection pools for the purpose of performing functional genotypic analysis[62]. SEWAL produces 3D frequency plots using a sorting algorithm to determine sequence patterns and visualizes changes in the sequencing space due to selective pressure across consecutive rounds. Identification of changes due to selective pressure can be used to predict evolutionary paths sequences follow towards improved fitness[57]. Sorting tools FASTAptamer and

AptaCluster use similar algorithms to compress pool data into non-redundant sequence sets (count) and sort sequences based on similarities in genotype (clustering) [63, 64]. Counting, ranks sequences based upon frequency and aids in identification of sequences which were enriched throughout the selection and assists in determining the complexity of the enriched pool. As mentioned, clustering sorts sequences into families of sequences which share a similar genotype. Clustering begins by separating abundant unique sequences into individual clusters. Less abundant sequences which identify with the seed sequence of a given cluster but vary by mutations, insertions, and deletions (indels) are then sorted. Additionally, alignments of individual clusters with tools such as MUSCLE or MAFFT provide visualization of allowed mutations and indels and the output of these tools are useful in determining sequence consensus in addition to potential evolutionary relationships [65-67]. While clustering leads to increased understanding of the genotypic variation that exists within families of functional RNAs, confirmation of selected activity relies on further biochemical analysis.

Identification of motif sequences from enriched pools provides insight into potential domains involved in a selected binding or catalytic activity. Methods aimed at identifying transcription factor binding sites, such as ChIP-seq, contribute to the development of DNA motif analysis software packages, which proves beneficial for the functional RNA discovery field. Motif identification software such as the MEME Suite family of tools, identify reoccurring sequence motifs which are enriched throughout genomic sequencing data[68]. However, identification of motifs based off sequence alone does not coincide with target binding or catalysis. BEEML, and a method used by Jolma et al., account for aptamer binding by utilizing a position weight matrix (PWM) to determine motifs based off binding energy models[51, 52]. Similarly, functional RNA motif identification relies heavily on secondary structure prediction due to RNA's complex tertiary structures. Backofen group's MEME in RNAs Including secondary Structure (MEMERIS) built on MEME by incorporating high throughput analysis capabilities in conjunction with secondary

structure prediction[69]. However, MEMERIS is restricted to motifs within predicted single-stranded loop and bulge regions of RNA sequences and does not consider changes upon interaction with a ligand. RNAcontext improved upon the motif analysis landscape by including loops, bulges, and stems into its predictions[70]. Because RNAcontext distinguishes between these secondary structures, it can also predict the preferred conformation of the identified motif within an aptamer domain. AptaMotif and APTANI use a minimum free energy (MFE) approach for structural prediction in conjunction with iterative sampling and sequence alignments, can determine potential motifs across multiple input files (i.e. rounds) [71, 72]. However, AptaMotif and APTANI are limited to identifying primary motifs which presents an issue in deeper data sets for complex targets that may contain multiple motif regions. AptaTrace overcomes this problem by tracing motifs based on sequence and binding energies through multiple rounds of selection data. AptaTrace identifies potential motifs by identifying whether these regions undergo selection towards secondary structure such as hairpin loops or bulges[73]. Recently, Hoinka et al. developed an all-encompassing software package, AptaSuite which includes AptaTrace and AptaCluster, and is the first open-source package designed for selection schemes to feature a graphical user interface[74].

RNA motifs identified from genomic selection are biological relevant and as a result homology is typically determined. Because RNA motif identification is not wholly based on sequence alone, secondary structure prediction must be incorporated within genome-wide searching tools. Packages such as Infernal and CMFinder, apply covariance-probabilistic models to identify related functional RNAs in genomic searches[75, 76]. However, their functionality is limited to identifying motifs across previously identified families of RNAs. RNArobo uses a contextual based motif searching algorithm for identification of novel motifs when no known homology has been previously annotated[77]. RNArobo parses genome-wide searches based upon an input motif descriptor which describes a simplified motif map and individual structural elements; for example,

nucleotide content. The methods employed by RNArobo are expedient in determining sequences containing difficult to predict structural elements, for example, pseudoknots, adding to its effectiveness in discovering unique genomic aptamers from *in vitro* selections.

Candidate aptamer sequences which dominate a selected pool bias towards enriched genotypes and this genotypic bias restricts possible nucleotide variation across the enriched pool, which limits structurally relevant information. Aptamer sequences are constrained within their given sequence space upon initial selection while mutagenesis of an enriched pool, introduces new, different constraints upon the sequence space and subsequent selection provides alternative genotypic outcomes[58]. As a result, mutagenesis is a useful strategy to overcome the lack of nucleotide diversity when followed by subsequent reselection. Mutagenesis of an enriched pool can be accomplished through mutagenic PCR as well as through error-prone PCR methods (EP-PCR). Mutagenic PCR is facilitated by nucleoside analogues, such as 8-oxo-dG and dPTP, to introduce templated mismatches. 8-oxo-dG and dPTP can be utilized by *Taq* polymerase allowing them to be simply introduced to standard PCR reaction mixtures in excess[78]. EP-PCR takes advantage of the low-fidelity of *Taq* polymerase to introduce mutations through doubling of the pool[79]. PCR is inherently error-prone and contributes to the genotypic variation of *in vitro* selection. However, by increasing the Mg^{2+} and introducing Mn^{2+} metal ions, the error rate of *Taq* polymerase increases from 0.02% to 0.066% per nucleotide position when coupled with disproportionate dNTP concentrations[80]. Covariation is a potential and sought-after outcome of mutagenic selection and secondary structure prediction that benefits from covariation can be accomplished with software such as ViennaRNA's RNAalifold[81].

High-throughput Characterization Methods

While deep sequencing has increased the depth of potential functional RNAs from selections, validation relies on biochemical confirmation of ligand binding or catalysis. Validation of selected activity can prove tedious with increased sequencing depth, as the number of

potentially identified functional RNAs becomes overwhelming to test *in vitro*. Moreover, traditional assays do not provide insights into conformational dynamics upon ligand binding or catalytic activity, and do not provide insight into functionally relevant domains. High-throughput methods of functional RNA discovery aim to address these issues through combinatorial methods of kinetic and structural analysis followed by deep sequencing. These methods intend to alleviate the strain of *in vitro* and *in silico* selection workup by identifying ligand associated motifs and determining rate constants.

Sequence enrichment does not necessarily equate to stronger or more specific binding/catalysis, therefore identification of high-affinity binders or ribozymes with higher turnover in a pool traditionally requires kinetic rate determination for individual sequences. Consequently, Abdelsayed et al. developed a workflow for the characterization of functional RNA aptamers by probing their dynamic properties against a titration of ligand. Their technique, Apta-Seq, utilizes a combinatorial, multiplexed approach focused on RNA chemical modification and deep sequencing to evaluate structural interaction of the aptamers to their target ligand[30]. Apta-Seq utilizes selective 2′-hydroxyl acylation with primer extension (SHAPE) to determine whether a 2′-OH is solvent accessible to allow acylation by a SHAPE reagent, such as 2-(azidomethyl)nicotinic acid acyl imidazole (NAI-N_3), under variable ligand concentrations[82, 83]. This method was applied to a human-genomic *in vitro* selection for ATP-binders followed by SHAPE combined with high-throughput sequencing (SHAPE-seq), an established pipeline for the mapping of acylation activities of diverse RNA pools[84]. Applying Apta-Seq, the group identified three novel human-genome derived ATP-binding RNA aptamers and confirmed the existence of two previously identified aptamers which demonstrated the robustness of the method to perform structural analysis in a high-throughput pipeline. Apta-Seq allows for high-throughput determination of dissociation rate constants, K_Ds, for selected pools by eliminating single-clone biochemical assessment. This technique is currently limited to RNA aptamer reactivity probing; however, the general workflow provides a unique framework which can be modified for DNA aptamers if base-

specific modifications are introduced with potential to be incorporated into workflows for ribozyme and riboswitch characterization.

Identification of structural motifs is critical in the determination of functionally relevant domains which assists in the minimization of selected sequences derived from a given pool. Previously, natural RNA aptamers and ribozymes have been discovered via computational investigation for sequence and structural homology[40, 85]. However, this approach limits the discovery of novel, naturally occuring aptamers and ribozymes to those which share consensus with characterized sequences. Consequently, Parallel Analysis of RNA Confirmations Exposed to Ligand Binding (PARCEL) was developed as a high-throughput method of identifying RNA-ligand interactions in whole transcriptomes[86]. The PARCEL workflow uses three methods of RNA-ligand interaction assessment in parallel, followed by next-generation sequencing. Tapsin et al. performed RNA-footprinting with double-strand specific RNase V1 and single-strand specific S1 nuclease assays in parallel in the presence or absence of a target metabolite. Additionally, the protocol applies partial hydrolysis of transcripts, in-line probing, to determine transcript flexibility[87, 88]. The methodology relies on the assumption that the target ligand provides protection to the RNA when bound. Transcripts treated in the presence or absence of metabolite were then deep sequenced and mapped to a reference genome. Reads are then normalized to ligand conditions and while PARCEL utilizes nuclease cleavage and in-line probing, both methods could be applied in parallel with acylation in Apta-Seq to provide additional structural insight and K_Ds.

Jalali-Yazdi et al. developed a high-throughput technique for the determination of k_{on}/k_{off} rate constants for pools of mRNA-peptide fusions[89, 90]. Their method, high-throughput sequencing kinetics (HTSK), provides k_{on}/k_{off} rates for most sequences in a given pool through a time-point analysis protocol. To determine the association rate constants, k_{on}, an enriched pool is added to ligands immobilized on magnetic beads and at various time points, bead fractions are removed, washed, and sequences are amplified and prepared for next-generation sequencing. The

fractional composition of a given sequence can be determined by the frequency of that sequence across all sequences at that time point. The amount of bound aptamer can be determined by multiplying the fractional composition to the total counts of the pool at each time point. Dissociation rate constants, k_{off}, are determined following the on-rate time points. Beads are washed in the presence of excess ligand, and time points collected. Fractional composition of sequences is determined similarly to the on-rates and K_Ds can be calculated from the rate constants. RNA Bind-n-Seq is a similar approach to HTSK and was used to identify RNA-protein interactions by combining high-throughput *in vitro* selection of RNA with a method of identifying DNA-protein interactions[91]. A randomized pool of RNA is incubated with streptavidin-tagged RNA binding proteins (RBPs) in varying concentrations which can be pulled down allowing bound RNA to be eluted and sequenced. Association and dissociation rate constants can be determined from these sequenced data assuming completed pull-down of the tagged RBPs, however, this method is limited to proteins or target molecules that can be tagged for pull-down assays.

The Illumina sequencing platform has been utilized for deep sequencing of whole genomes, single-cell RNA-seq, and for understanding epigenetics through bisulphite sequencing[92-94]. Nutiu et al. first illustrated the Illumina platform's versatility for aptamer discovery with their method, high-throughput sequencing fluorescent ligand interaction profiling (HITS-FLIP) [95]. HITS-FLIP quantitatively measures DNA-protein affinity through changes in fluorescent ligand associations of millions of DNA clusters bound to the Illumina flowcell. On/off rates can be determined by varying the protein concentrations introduced across the flowcell and measuring subsequent changes in fluorescence. Additionally, Jung et al.'s chip-hybridized associated-mapping platform (CHAMP) takes advantage of used MiSeq chips with linked DNA clusters to determine DNA-protein interactions[96]. The DNA clusters are treated to remove associated fluorescent nucleotides and a fluorescent-oligonucleotide probe is then hybridized to the clusters and is used as a reference marker for the DNA sequence. Fluorescent proteins are then incubated

in varying concentrations, imaged, and associations are made to the DNA sequence through imaging and software analysis. Similarly, the Buenrosto et al. and Tome et al. built on HITS-FLIP for high-throughput RNA aptamer discovery[97, 98]. Each method uses diverse single-stranded DNA libraries which contain T7 promoter regions to initiate transcription on the sequencer flow cell. Each method employs an approach to halt transcription with the intent on maintaining the association of the transcript with the DNA cluster. Fluorescently-labeled proteins which associate to halted transcripts, can be determined through fluorescent imaging of the flowcell. This allows for the identification of RNA aptamers that bind the protein ligand through the sequence of the associated DNA. Additionally, genotypic variations can be directly compared, decreasing the need for downstream mutagenic analysis and thus providing covariaton information for structure prediction. However, these methods are currently limited to ligands which can be fluorescently labeled, effectively narrowing potential small molecule targets.

1.3 References

1. Palade, G.E. Microsomes and ribonucleoprotein particles. *Microsomal particles and protein synthesis*, 36-61 (1958).

2. Crick, F.H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138-163 (1958).

3. Crick, F.H.C. & Brenner, S. Some footnotes on protein synthesis: a note for the RNA Tie Club. *Profiles in Medicine https://profiles.nlm.nih.gov/ps/access/SCBBFV.pdf* (1959).

4. Brenner, S. & Crick, F.H.C. What are the properties of genetic RNA. *A note for the RNA Tie Club. Profiles in Medicine https://profiles.nlm.nih.gov/ps/access/SCBBFZ.pdf* (1960).

5. Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**, 576-581 (1961).

6. Nirenberg, M.W. & Matthaei, J.H. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences* **47**, 1588-1602 (1961).

7. Holley, R.W. et al. Structure of a Ribonucleic Acid. *Science* **147**, 1462-1465 (1965).

8. Qu, Z. & Adelson, D.L. Evolutionary conservation and functional roles of ncRNA. *Front Genet* **3**, 205 (2012).

9. López-Jiménez, E. & Andrés-León, E. The Implications of ncRNAs in the Development of Human Diseases. *Noncoding RNA* **7** (2021).

10. Decoding noncoding RNAs. *Nature Methods* **19**, 1147-1148 (2022).

11. Eddy, S.R. Non–coding RNA genes and the modern RNA world. *Nature Reviews Genetics* **2**, 919-929 (2001).

12. Mattick, J.S. et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology* **24**, 430-447 (2023).

13. Garst, A.D., Edwards, A.L. & Batey, R.T. Riboswitches: structures and mechanisms. *Csh Perspect Biol* **3**, a003533 (2011).

14. Roth, A. & Breaker, R.R. The Structural and Functional Diversity of Metabolite-Binding Riboswitches. *Annual Review of Biochemistry* **78**, 305-334 (2009).

15. Peselis, A. & Serganov, A. Themes and variations in riboswitch structure and function. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1839**, 908-918 (2014).

16. Peselis, A. & Serganov, A. Structural insights into ligand binding and gene expression control by an adenosylcobalamin riboswitch. *Nat Struct Mol Biol* **19**, 1182-1184 (2012).

17. Wang, J.X. & Breaker, R.R. Riboswitches that sense S-adenosylmethionine and S-adenosylhomocysteine. *Biochem Cell Biol* **86**, 157-168 (2008).

18. Mironov, A.S. et al. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *cell* **111**, 747-756 (2002).

19. Nahvi, A. et al. Genetic control by a metabolite binding mRNA. *Chemistry & biology* **9**, 1043-1049 (2002).

20. McCown, P.J., Corbino, K.A., Stav, S., Sherlock, M.E. & Breaker, R.R. Riboswitch diversity and distribution. *Rna* **23**, 995-1011 (2017).

21. Benner, S.A., Ellington, A.D. & Tauer, A. Modern metabolism as a palimpsest of the RNA world. *Proceedings of the National Academy of Sciences* **86**, 7054-7058 (1989).

22. Chen, X., Li, N. & Ellington, A.D. Ribozyme catalysis of metabolism in the RNA world. *Chemistry & biodiversity* **4**, 633-655 (2007).

23. Higgs, P.G. & Lehman, N. The RNA World: molecular cooperation at the origins of life. *Nature Reviews Genetics* **16**, 7-17 (2015).

24. Breaker, R.R. The Biochemical Landscape of Riboswitch Ligands. *Biochemistry* **61**, 137-149 (2022).

25. Sudarsan, N., Barrick, J.E. & Breaker, R.R. Metabolite-binding RNA domains are present in the genes of eukaryotes. *Rna* **9**, 644-647 (2003).

26. Li, S. & Breaker, R.R. Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucleic acids research* **41**, 3022-3031 (2013).

27. Vu, M.M.K. et al. Convergent evolution of adenosine aptamers spanning bacterial, human, and random sequences revealed by structure-based bioinformatics and genomic SELEX. *Chemistry & biology* **19**, 1247-1254 (2012).

28. Curtis, E.A. & Liu, D.R. A naturally occurring, noncanonical GTP aptamer made of simple tandem repeats. *RNA Biology* **11**, 682-692 (2014).

29. Curtis, E.A. & Liu, D.R. Discovery of widespread GTP-binding motifs in genomic DNA and RNA. *Chemistry & biology* **20**, 521-532 (2013).

30. Abdelsayed, M.M. et al. Multiplex Aptamer Discovery through Apta-Seq and Its Application to ATP Aptamers Derived from Human-Genomic SELEX. *ACS chemical biology* **12**, 2149-2156 (2017).

31. Batey, R.T. Riboswitches: still a lot of undiscovered country. *Rna* **21**, 560-563 (2015).

32. Kruger, K. et al. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147-157 (1982).

33. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849-857 (1983).

34. Ban, N., Nissen, P., Hansen, J., Moore, P.B. & Steitz, T.A. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **289**, 905-920 (2000).

35. Peng, H., Latifi, B., Muller, S., Luptak, A. & Chen, I.A. Self-cleaving ribozymes: substrate specificity and synthetic biology applications. *Rsc Chem Biol* **2**, 1370-1383 (2021).

36. Reymond, C., Beaudoin, J.D. & Perreault, J.P. Modulating RNA structure and catalysis: lessons from small cleaving ribozymes. *Cell Mol Life Sci* **66**, 3937-3950 (2009).

37. Salehi-Ashtiani, K., Luptak, A., Litovchick, A. & Szostak, J.W. A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science* **313**, 1788-1792 (2006).

38. Chen, Y. et al. Hovlinc is a recently evolved class of ribozyme found in human lncRNA. *Nature Chemical Biology* (2021).

39. Claire, C.C. et al. Inhibition of CPEB3 ribozyme elevates CPEB3 protein expression and polyadenylation of its target mRNAs, and enhances object location memory. *bioRxiv*, 2023.2006.2007.543953 (2023).

40. Webb, C.H., Riccitelli, N.J., Ruminski, D.J. & Luptak, A. Widespread occurrence of self-cleaving ribozymes. *Science* **326**, 953 (2009).

41. Webb, C.H. & Luptak, A. HDV-like self-cleaving ribozymes. *RNA Biol* **8**, 719-727 (2011).

42. Ruminski, D.J., Webb, C.H., Riccitelli, N.J. & Luptak, A. Processing and translation initiation of non-long terminal repeat retrotransposons by hepatitis delta virus (HDV)-like self-cleaving ribozymes. *J Biol Chem* **286**, 41286-41295 (2011).

43. Watters, K.E., Strobel, E.J., Yu, A.M., Lis, J.T. & Lucks, J.B. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nature Structural & Molecular Biology* **23**, 1124-1131 (2016).

44. Tuerk, C. & Gold, L. Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science* **249**, 505-510 (1990).

45. Ellington, A.D. & Szostak, J.W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818-822 (1990).

46. Jijakli, K. et al. The in vitro selection world. *Methods* **106**, 3-13 (2016).

47. Griffin, L.C., Toole, J.J. & Leung, L.L.K. The discovery and characterization of a novel nucleotide-based thrombin inhibitor. *Gene* **137**, 25-31 (1993).

48. Morris, K.N., Jensen, K.B., Julin, C.M., Weil, M. & Gold, L. High affinity ligands from in vitro selection: complex targets. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 2902-2907 (1998).

49. Schütze, T. et al. Probing the SELEX process with next-generation sequencing. *PloS one* **6**, e29604 (2011).

50. Zykovich, A., Korf, I. & Segal, D.J. Bind-n-Seq: high-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing. *Nucleic Acids Research* **37**, e151-e151 (2009).

51. Jolma, A. et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* **20**, 861-873 (2010).

52. Zhao, Y. & Stormo, G.D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature biotechnology* **29**, 480-483 (2011).

53. Slattery, M. et al. Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. *Cell* **147**, 1270-1282 (2011).

54. Zimmermann, B., Gesell, T., Chen, D., Lorenz, C. & Schroeder, R. Monitoring genomic sequences during SELEX using high-throughput sequencing: neutral SELEX. *PloS one* **5**, e9169-e9169 (2010).

55. Ditzler, M.A. et al. High-throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase. *Nucleic Acids Research* **41**, 1873-1884 (2013).

56. Cho, M. et al. Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 15373-15378 (2010).

57. Pitt, J.N. & Ferré-D'Amaré, A.R. Rapid construction of empirical RNA fitness landscapes. *Science (New York, N.Y.)* **330**, 376-379 (2010).

58. Jiménez, J.I., Xulvi-Brunet, R., Campbell, G.W., Turk-MacLeod, R. & Chen, I.A. Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 14984-14989 (2013).

59. Neveu, M., Kim, H.-J. & Benner, S.A. The “Strong” RNA World Hypothesis: Fifty Years Old. *Astrobiology* **13**, 391-403 (2013).

60. Hoon, S., Zhou, B., Janda, K.D., Brenner, S. & Scolnick, J. Aptamer selection by high-throughput sequencing and informatic analysis. *BioTechniques* **51**, 413-416 (2011).

61. Hamada, M. In silico approaches to RNA aptamer design. *Biochimie* **145**, 8-14 (2018).

62. Pitt, J.N., Rajapakse, I. & Ferré-D'Amaré, A.R. SEWAL: an open-source platform for next-generation sequence analysis and visualization. *Nucleic acids research* **38**, 7908-7915 (2010).

63. Alam, K.K., Chang, J.L. & Burke, D.H. FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Molecular Therapy—Nucleic Acids* **4**, e230 (2015).

64. Hoinka, J., Berezhnoy, A., Sauna, Z.E., Gilboa, E. & Przytycka, T.M. AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application. *Research in Computational Molecular Biology* **8394**, 115-128 (2014).

65. Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30**, 3059-3066 (2002).

66. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004).

67. Ameta, S., Winz, M.-L., Previti, C. & Jäschke, A. Next-generation sequencing reveals how RNA catalysts evolve from random space. *Nucleic Acids Research* **42**, 1303-1310 (2014).

68. Bailey, T.L. et al. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202-W208 (2009).

69. Hiller, M., Pudimat, R., Busch, A. & Backofen, R. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research* **34**, e117 (2006).

70. Kazan, H., Ray, D., Chan, E.T., Hughes, T.R. & Morris, Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Computational Biology* **6**, e1000832 (2010).

71. Caroli, J., Taccioli, C., De La Fuente, A., Serafini, P. & Bicciato, S. APTANI: a computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data. *Bioinformatics* **32**, 161-164 (2016).

72. Hoinka, J., Zotenko, E., Friedman, A., Sauna, Z.E. & Przytycka, T.M. Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers. *Bioinformatics* **28**, i215-i223 (2012).

73. Dao, P. et al. AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments. *Cell Systems* **3**, 62-70 (2016).

74. Hoinka, J., Backofen, R. & Przytycka, T.M. AptaSUITE: A Full-Featured Bioinformatics Framework for the Comprehensive Analysis of Aptamers from HT-SELEX Experiments. *Molecular Therapy—Nucleic Acids* **11**, 515-517 (2018).

75. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).

76. Yao, Z., Weinberg, Z. & Ruzzo, W.L. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445-452 (2006).

77. Rampášek, L., Jimenez, R.M., Lupták, A., Vinař, T. & Brejová, B. RNA motif search with data-driven element ordering. *BMC bioinformatics* **17**, 216 (2016).

78. Zaccolo, M., Williams, D.M., Brown, D.M. & Gherardi, E. An Approach to Random Mutagenesis of DNA Using Mixtures of Triphosphate Derivatives of Nucleoside Analogues. *Journal of Molecular Biology* **255**, 589-603 (1996).

79. Wilson, D.S. & Keefe, A.D. Random Mutagenesis by PCR. *Current Protocols in Molecular Biology* **51**, 8.3.1-8.3.9 (2001).

80. Cadwell, R.C. & Joyce, G.F. Randomization of genes by PCR mutagenesis. **2**, 28-33 (1992).

81. Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R. & Stadler, P.F.J.B.B. RNAalifold: improved consensus structure prediction for RNA alignments. **9**, 474 (2008).

82. Spitale, R.C. et al. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486-490 (2015).

83. Merino, E.J., Wilkinson, K.A., Coughlan, J.L. & Weeks, K.M. RNA Structure Analysis at Single Nucleotide Resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE). *Journal of the American Chemical Society* **127**, 4223-4231 (2005).

84. Lucks, J.B. et al. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America* **108**, 11063-11068 (2011).

85. Barrick, J.E. & Breaker, R.R. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biology* **8**, R239 (2007).

86. Tapsin, S. et al. Genome-wide identification of natural RNA aptamers in prokaryotes and eukaryotes. *Nature Communications* **9**, 1289 (2018).

87. Regulski, E.E. & Breaker, R.R. in Post-Transcriptional Gene Regulation. (ed. J. Wilusz) 53-67 (Humana Press, Totowa, NJ; 2008).

88. Soukup, G.A. & Breaker, R.R. Relationship between internucleotide linkage geometry and the stability of RNA. *RNA* **5**, 1308-1325 (1999).

89. Roberts, R.W. & Szostak, J.W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 12297-12302 (1997).

90. Jalali-Yazdi, F., Lai, L.H., Takahashi, T.T. & Roberts, R.W. High-Throughput Measurement of Binding Kinetics by mRNA Display and Next-Generation Sequencing. *Angewandte Chemie* **55**, 4007-4010 (2016).

91. Lambert, N. et al. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Molecular Cell* **54**, 887-900 (2014).

92. Lippert, C. et al. Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 10166-10171 (2017).

93. Raine, A., Manlig, E., Wahlberg, P., Syvänen, A.-C. & Nordlund, J. SPlinted Ligation Adapter Tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. *Nucleic acids research* **45**, e36 (2017).

94. Ziegenhain, C. et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65**, 631-643.e634 (2017).

95. Nutiu, R. et al. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature Biotechnology* **29**, 659 (2011).

96. Jung, C. et al. Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. *Cell* **170**, 35-47.e13 (2017).

97. Buenrostro, J.D. et al. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nature Biotechnology* **32**, 562 (2014).

98. Tome, J.M. et al. Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nature Methods* **11**, 683-688 (2014).

Chapter 2: Characterization of a human RNA aptamer for cGMP identified through genomic SELEX

Publication note:

This chapter is a written draft of a manuscript in preparation.

Cole, K.H.*, Ho, B.T.*, Passalacqua, L.F.M., Nakata, M. Pham, J. Hua, C. Nguyen, P. and Lupták, A. Characterization of a human RNA aptamer for cGMP identified through genomic SELEX. *Manuscript in preparation*. (2024).

*These authors contributed equally

Contribution statement:

Conceptualization: A.L., B.T.H. Methodology: A.L., B.T.H., **K.H.C.**, L.F.M.P. Investigation: B.T.H., **K.H.C.**, L.F.M.P., M.N., J.P., C.H., P.N. Visualization: B.T.H., **K.H.C.** Writing – original draft: A.L., **K.H.C.** Writing – review and editing: all authors. Funding acquisition: A.L.

2.1 Abstract

Naturally occurring RNA aptamers make up the ligand sensing domain of the widespread riboswitch class of non-coding RNAs (ncRNAs). There are few examples of RNA-based regulators in eukaryotes which sense metabolites and determine gene product fates. In the past decade, several examples of eukaryotic RNA aptamers have been discovered to bind critical metabolites and secondary messenger molecules. These aptamers support the notion that eukaryotic transcriptomes contain small-molecular binders (aptamers) and the need to identify their potential biological roles. Here, genomic SELEX was used to identify a conserved-primate aptamer for the secondary messenger 3′,5′ cyclic-guanosine monophosphate (cGMP). This aptamer binds to cGMP at relevant concentrations and is highly selective for cGMP in comparison to other nucleotide analogues. This work highlights the first identification of an RNA aptamer with specificity towards cGMP, and more importantly one identified in a conserved region of the human genome.

2.2 Introduction

RNA-dependent regulation of gene expression by metabolite sensing is a common process in bacteria. Riboswitches exist within the 5′- or 3′-untranslated regions (UTR) of mRNAs and control gene expression through co-transcriptional binding of small-molecule metabolites. Upon binding of their substrate metabolite ligands within their aptamer domains, a kinetic and conformational switch determines the occlusion of key nucleotide positions that enable binding of the ribosome to the Shine-Dalgarno sequence (translation) or the formation of terminator helices (transcriptional), ultimately controlling the fate of the downstream gene product[1]. While riboswitches are widespread across bacteria and archaea[2], only one class of eukaryotic riboswitch has been discovered to-date: the thiamine pyrophosphate (TPP) riboswitch found in fungi, plants, and algae[3-6]. It has been postulated that riboswitches are ancestors of early RNA regulators that controlled gene expression prior to the involvement of protein enzymes[7]. The

discovery of human RNA aptamers for adenosine, GTP, and folic acid suggests that the human transcriptome has the structural potential to bind key messenger molecules[8-10] and that there is a need to mine for riboswitch-like regulators in the animal kingdom.

To this end, we performed human genomic SELEX in search of human aptamers for the secondary-messenger molecule 3′,5′ cyclic-guanosine monophosphate (cGMP). cGMP is a key secondary messenger in the regulation of receptor protein kinase activation in processes such as smooth muscle relaxation[11], neurotransmission[12], and phototransduction[13]. Additionally, intracellular concentrations of cGMP are tightly regulated through synthesis by guanylate cyclases and degradation by cyclic-nucleotide phosphodiesterases[14]. After nine rounds of selection, we identified 35 candidate cGMP aptamers. One aptamer, G62, is a primate-conserved aptamer which specifically recognizes cGMP with a physiologically relevant dissociation constant. G62 resides within an intergenic region near an expressed sequence tag (EST) for the optic nerve, suggesting that this aptamer may be involved in a phototransduction-associated pathway[15, 16].

2.3 Results

A human genomic DNA library was prepared, and *in vitro* selection was performed as previously described[8, 17]. To mimic co-transcriptionally binding of cGMP to putative aptamers, denaturing PAGE-purified *in vitro* transcription reactions were heat denatured and cooled to room temperature in a physiological buffer in the presence of cGMP-agarose beads. This allowed the RNA to fold around cGMP as if it were being bound to the secondary messenger co-transcriptionally. The selection was performed for six rounds with competitive elution steps containing 5 mM cGMP following washes with physiological buffer. Enrichment was observed after six rounds of selection with ~21% of the RNA population being competitively eluted from the cGMP-agarose (**Figure S2.1)**. A counter-elution step, containing 0.5 mM GMP, was introduced in the seventh round to increase stringency and to increase selectivity towards cGMP. After three additional rounds of selection, the enriched pool (Round 9, ~20% RNA eluted; Figure S2.1) was

cloned, and sequences were amplified from isolated colonies with selection-specific primers. Gel-purified amplicons were sequenced by Sanger and sequences were queried on the GRCh37/hg19 assembly on the UCSC BLAT server[18].

All isolated sequences mapped to an intergenic region within chromosome 20p12.3 and overlapped with an expressed-sequence tag (EST) of an adult-optic nerve (EST: EB385442; **Figure 2.1A**) [19]. Sequence conservation was observed amongst higher-order simians, with few nucleotide variations differences between human, chimpanzee, and gorilla (**Figure 2.1B**). The isolated aptamer (G62) was tested *in vitro* to confirm its affinity for cGMP. Purified aptamer was denatured and allowed to refold on cGMP-agarose, following a similar process to the selection scheme. G62 could be eluted off the column with competitive concentrations of free cGMP free cGMP (~41.6%), but no significant fraction of G62 could be eluted with guanosine 5′-monophosphate (GMP; **Figure 2.1C**). The aptamer also shows regioselectivity for binding to cGMP, as it does not have affinity for 2′-OH linked-agarose beads (~3.8%), but the aptamer competitively elutes from C8-AET linked-agarose beads (45.4%; **Figure 2.1D**).

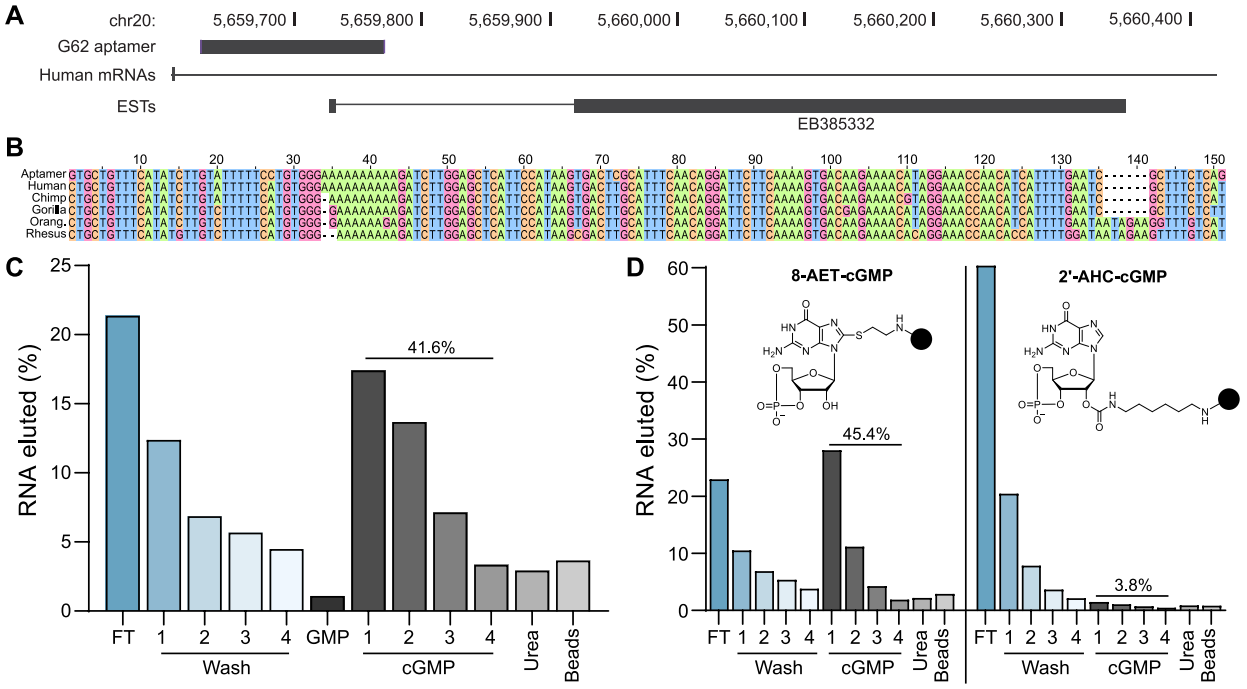


Figure 2.1: G62 is a functional, primate conserved cGMP aptamer. (**A**) Location of the G62 aptamer in an intergenic region of chromosome 20 (p12.3) in the human genome. G62's 3′-end overlaps with the expressed-sequence tag (EST), EB385332[19] of human optical nerve. (**B**) Alignment of G62 with the human, chimp, gorilla, orangutan (orang.), and rhesus genomic sequences, illustrating high conservation of sequences. (**C**) Affinity-column binding of the G62 aptamer to 8-AET-cGMP agarose beads with counter-elution with GMP and selective elution with cGMP (**D**) Affinity-column binding of G62 to 8-AET-cGMP agarose beads and 2′-AHC-cGMP agarose beads (depicted as black, filled circles). Binding and elution from 8-AET-cGMP is a replicate experiment of Figure 2.1C without a counter-elution with GMP. FT: flowthrough; GMP: guanosine 5′-monophosphate.

Biochemical validation of cGMP binding by G62

We applied our Apta-Seq pipeline[17] to round six of the enriched library. Apta-Seq revealed an additional 34 potential human-cGMP aptamers in round 9 of the selected pool (**Table S2.1**). As expected, G62 was the most enriched of the selected sequences at ~94.2%, while G66 (~3.6%) and G40 (~0.5%) were the next most abundant. The Apta-Seq trace identified positions U90 – A121 as being the most reactive to NAI-N_3 at varying concentrations of cGMP (**Figure 2.2A**). A major peak at A102 shows an increase in acylation frequency (i.e., RT-stop frequency) as the amount of cGMP is increased from zero to 10 mM. In contrast, A110 decreases in acylation frequency as cGMP concentrations are increased suggesting that cGMP induces a rearrangement of the binding pocket within this region. A102 and A110 had apparent dissociation constants (K_D) of ~7.2 μM and ~220 μM, respectively, which falls within the realm of reported intracellular concentrations in smooth muscle[20, 21]. A dissociation constant could not be determined for A112.

For further confirmation of acylation reactivity by NAI-N_3, classical SHAPE analysis was performed on G62 at various cGMP concentrations and observed by denaturing PAGE (**Figure 2.2B**) [22, 23]. Strikingly, we observed consistent acylation patterns for both A102 and A110. A102 had an apparent K_D of approximately ~8.3 μM not too dissimilar to that of the Apta-Seq, with an observable increase in band intensity at higher concentrations of cGMP. A110 behaved similarly to that of the Apta-Seq analysis. A110 reactivity can be observed decreasing, in contrast to A102, with an apparent K_D of ~108 μM (a 2-fold reduction from Apta-Seq). While we observed strong peaks for position A102, A110, and A112 in Apta-Seq, denaturing PAGE could resolve additional positions. U98 has a striking increase in reactivity at low concentrations of cGMP in comparison to other reactive positions (K_D ~0.1 μM). Positions A102 – A105 portray large rearrangement potential of the predicted aptamer-binding domain, further supporting these positions, as well as A103 – A105 as being part of the ligand-binding pocket. The observed reactivity of positions A112

(K_D ~117 μM) and G113 (K_D ~108 μM) further reflects the reactivity observed in Apta-Seq with dissociation constants resembling that of A110 (Figure 2.2A).

Next, in-line probing was performed as described elsewhere[24]. G62 was incubated for ~40 hours under basic conditions at room temperature with varying concentrations of cGMP. Most strikingly, two major positions illustrate the most in-line activity. Position A102 which had been observed to have strong SHAPE-reactivity illustrated a cleavage reactivity with an apparent K_D of ~53.1 μM. While this determined value is six-fold higher than observed via SHAPE profiling, the half-maximum value still falls within the physiologically relevant regime. A105 illustrated a strong cleavage pattern in comparison to A102 with a K_D of ~137 μM. Strikingly, while A110 was observed to have strong SHAPE reactivity, it remained uncleaved across all cGMP concentrations. This suggests that A110 may be forming a stable Watson-Crick base pair. **Figure 2.2D** includes all apparent dissociation constants for A102 across Apta-Seq, SHAPE, and in-line probing experiments.

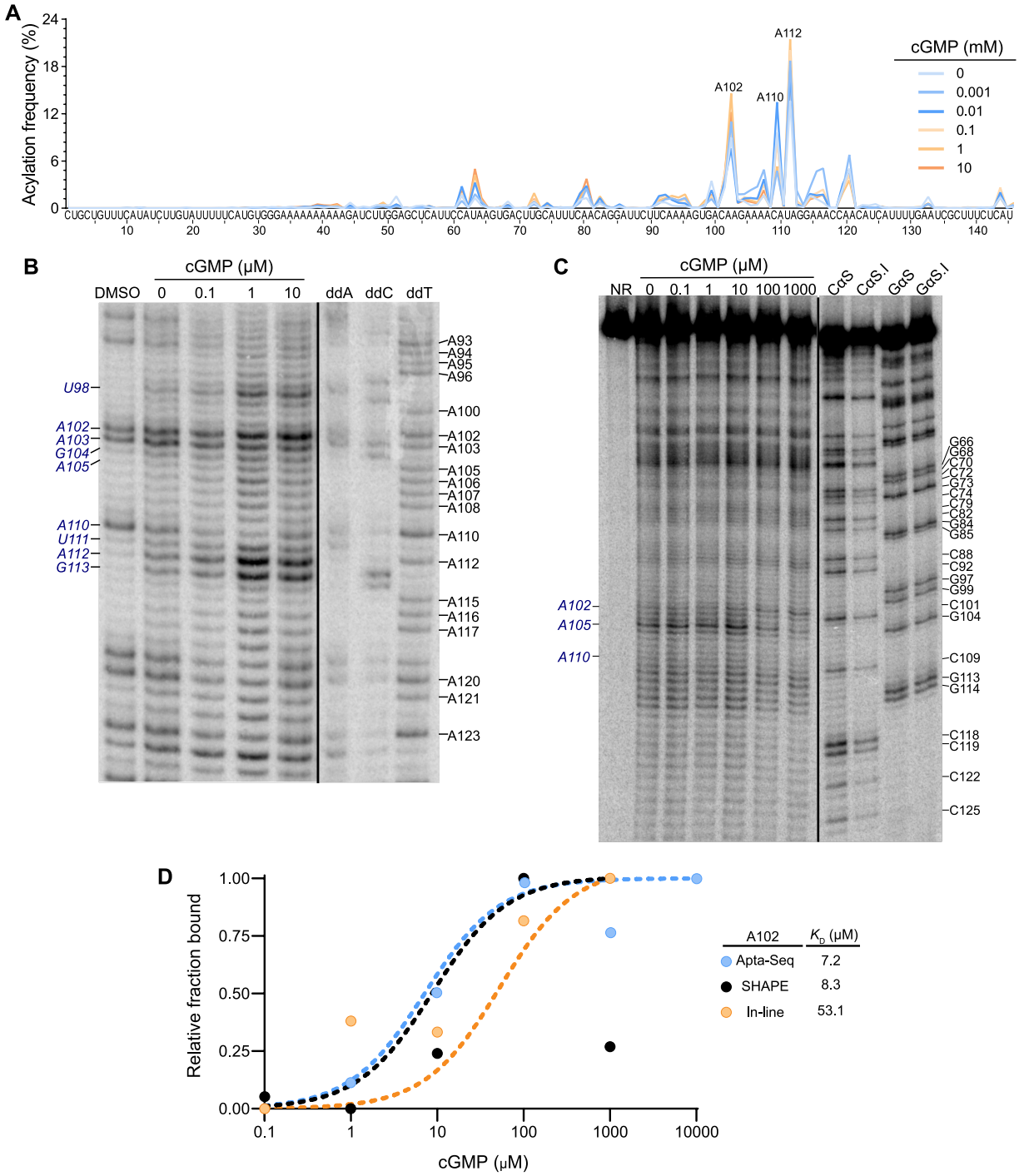


Figure 2.2: *In vitro* activity of G62 to cGMP. (**A**) Apta-Seq[17] trace of selective 2′-OH acylation (SHAPE) by 50 mM 2-(azidomethyl)nicotinic acid imidazolide (NAI-N_3) at various concentrations of cGMP (0 – 10 mM). Peaks are representative of RT stops that occur at the neighboring 3′ nucleotide. (**B**) SHAPE-gel of G62 at varying concentrations of cGMP (0 – 10 μM). Strong RT stops can be observed for U98, A102, A110, A112, and G113 as cGMP is titrated. DMSO: no NAI-N3; ddA, ddC, ddT: dideoxy sequencing lanes. (**C**) In-line probing of G62 with varying concentration of cGMP (0 – 1000 μM). NR: no reaction; CαS: cytosine-5'-(α-thio)-triphosphate; CαS.I: CαS cleaved with additional iodoethanol. GαS: guanosine-5'-(α-thio)-triphosphate; GαS.I: GαS cleaved with additional iodoethanol. (**D**) Binding isotherm of A102 from Apta-Seq (A; blue),

SHAPE gel (B; black), and in-line probing with cGMP (C; orange). Model curve fits and dissociation constants, K_D, were determined by linear least-squares analysis.

Truncation and mutational analysis of G62

Based on biochemical analysis through Apta-Seq, SHAPE, and in-line probing, we determined that the binding domain of G62 was most likely between U98 and A115. To further define the minimum necessary sequence for cGMP binding, we logically truncated the G62 aptamer at both the 5′-and-3′ ends (**Figure S2.2**). All constructs, apart from Construct 2, included positions U98 – A115. Constructs 1, 3, 5 – 7 included 3′-upstream extensions guided by comparative conservation with other non-human primates and secondary structure prediction. Radiolabeled, PAGE-purified transcripts were tested for affinity with cGMP using previously described refolding approaches and compared to the full-length G62 aptamer sequence (**Figure 2.3A; Figure S2.3**). Of the nine tested constructs, only one construct, Construct 4, bound with comparable affinity to the full-length construct. Construct 4 (50 – 138) is an 89 nt transcript which includes the predicted binding domain and ~59.3% of transcripts eluted with excess cGMP. Interestingly, constructs 8 and 9, which do not include nucleotides in at 5′ end of Construct 4, bound with some of the lowest affinity. This suggests that positions 50 – 67 contribute significantly to structure for stabilization of the binding domain.

We next wanted to assess the specificity of Construct 4 for cGMP over other guanosine and cyclic nucleotide analogues (**Figure 2.3B**). For each assay, G62 Construct 4 was refolded and bound to cGMP-agarose beads. After washing, the RNA was eluted with varying concentrations of nucleotides and the fraction of RNA eluted was measured by liquid scintillation counter. Expectedly, cGMP eluted G62 Construct 4 from cGMP-agarose beads at lower concentrations than assessed analogues. Construct 4 begins eluting from the column at 0.1 μM cGMP and peaks at 1 μM cGMP. Comparatively, GDP and GTP do not elute Construct 4 until 10 μM with 6% and 8% RNA eluted, respectively. Additionally, six-point mutations were chosen to determine their impact on cGMP binding by G62 Construct 4 (**Figure 2.3C; Figure S2.4**). For each mutation, a purine-to-pyrimidine mutation was chosen to disrupt potential Watson-Crick

base pairing interactions. Notably, A100C, A102C, G104U, and A105C abrogated binding to cGMP entirely, supporting probing evidence from G62 FL constructs that these positions form the binding pocket. A115C and A116C did not negatively impact binding to cGMP, suggesting that disruption of potential base pairing at these positions are distal to the binding pocket and have little influence on overall affinity to cGMP.

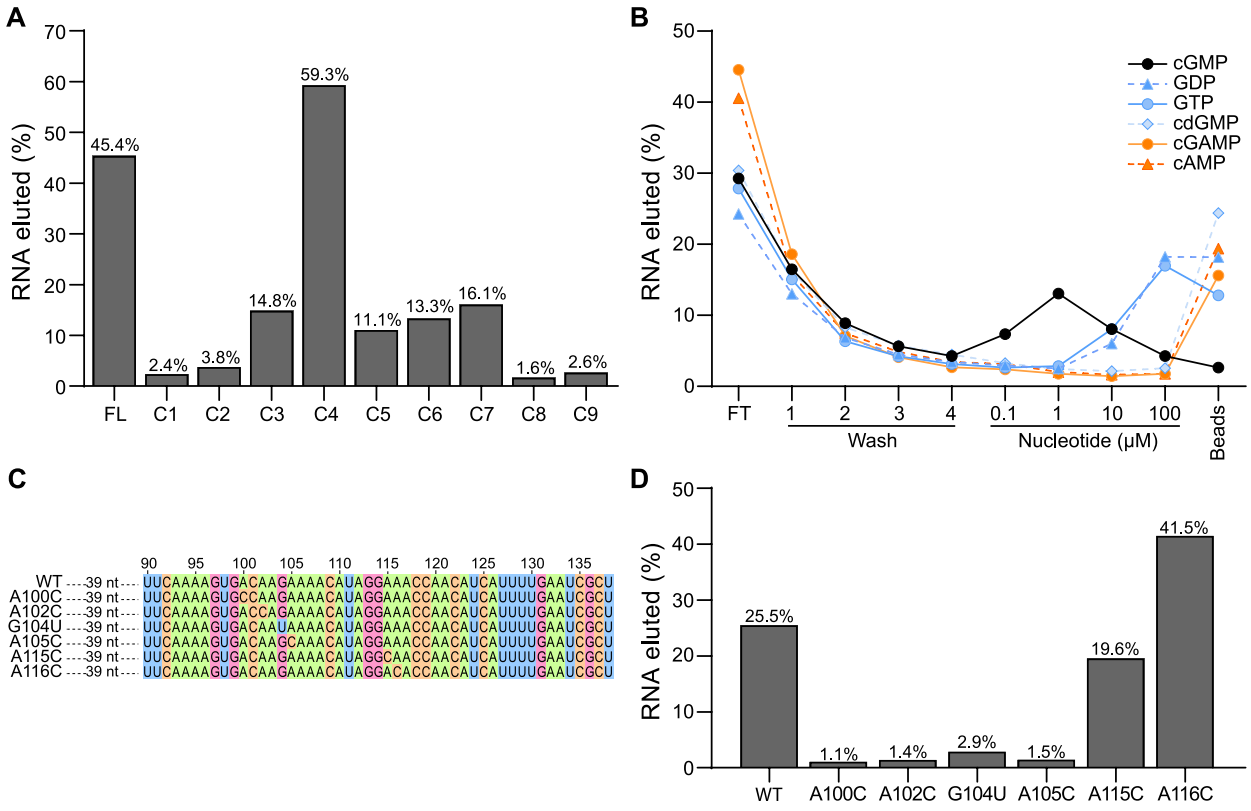


Figure 2.3: Mutational analysis of truncated G62 reveals positions critical for cGMP binding. (**A**) Affinity-column binding of full-length (FL) G62 and varied truncated constructs (C1-9). Alignment of all constructs can be found in **Figure S2.2** and complete binding profiles in **Figure S2.3**. (**B**) Affinity-column binding of G62 C4 with cGMP (black) and other nucleotide derivatives. (**C**) Alignment of G62 C4 mutant variants. The first 39 nt are not shown. WT: wild-type. (**D**) Affinity-column binding results of G62 C4 mutants and the complete binding profile can be seen in **Figure S2.4.**

2.4 Discussion

In summary, we have identified 35 potential cGMP aptamers from a human-genomic SELEX protocol. The most abundant aptamer from the selection was G62 found in an intergenic region of 20.p12.3 with sequence overlap with EST EB385332 (**Figure 2.1**). Biochemical analysis of G62 by SHAPE and in-line probing suggests that positions A102 – A115 interact with cGMP (**Figure 2.2**). Mutational analysis supports probing analysis of Figure 2, as A102C and A105C mutations completely abrogate the ability of G62 to bind with cGMP (**Figure 2.3D**). Additionally, G62 illustrates selectivity for cGMP over other phosphorylated guanosines, in addition to other cyclic nucleotide analogues such as cAMP (**Figure 2.3B**). While this data reports the ability of G62 to bind to cGMP, this data does not confirm the expression of the transcript, nor does it provide any evidence of its involvement in biological regulation.

Regulation of eukaryotic cellular processes has become increasingly more complex with the discovery of new and unique roles of functional, non-coding RNAs. For example, long non-coding RNAs (lncRNAs) are involved in diverse processes such as transcriptional[25] and translational[26] regulation, proliferation[27], differentiation[28], development[29], and more[30]. Recent evidence has suggested that a particular subclass of lncRNAs act as the animal equivalent of riboswitches through protein-protein-RNA interactions[31]. While no direct evidence has been able to uncover biological roles of human RNA aptamers, further development of methodologies to identify and uncover aptamer motifs will further aid our understanding[32].

2.5 Materials and Methods

In vitro transcription

Transcriptions were performed in a buffer containing 40 mM Tris-HCl pH 8, and 2 mM spermidine and supplemented with 16 mM $MgCl_2$, 10 mM DTT, 2 mM GTP, 2 mM CTP, 2 mM UTP, 2 mM ATP, and recombinant T7 RNA polymerase, and incubated at 37 °C for at least 2 hours. For radiolabeled transcripts, 0.2 mM ATP and ~2 µCi [α-^{32}P] ATP (250 µCi; PerkinElmer) was used in place of 2 mM ATP. Transcriptions were quenched by buffer exchange through G-25 Sephadex (Sigma-Aldrich, G2580) equilibrated with 7 M urea and 40 mM EDTA with xylene cyanol and bromophenol blue. Unlabeled RNAs were visualized by UV shadowing. Radiolabeled RNAs were purified via denaturing PAGE and visualized by phosphor imaging (GE Typhoon Trio). RNAs were excised from the gel and eluted into 300 mM KCl over three hours at room temperature and precipitated with the addition of 1 µL GlycoBlue (Invitrogen, AM9515) and 2.5 volumes of cold 100% ethanol and incubated at -80 °C for 2 hours. RNA was pelleted via centrifugation (20,000 *g*) for 20 minutes, washed twice in cold 70% ethanol, and resuspended TET (10 mM Tris-HCl pH 7.5, 0.1 mM EDTA, and 0.001% Triton X-100).

In vitro selection

The human genomic DNA pool used for *in vitro* selection was prepared and described previously [8, 33]. The transcription was purified via denaturing PAGE, eluted, precipitated, and subsequently resuspended in 200 µL binding buffer (140 mM KCl, 10 mM NaCl, 20 mM Tris-HCl, 5 mM $MgCl_2$, pH 7.5). RNA was heated to 70 °C for 2 min and refolded on pre-equilibrated C8-linked 3', 5'-cGMP agarose beads (Biolog, A019) in a 0.2 µm spin filter (Corning, CLS8162). The RNA pool was incubated on the beads at room temperature for 30 min with agitation and subsequently collected via centrifugation (3000 *g*). The beads were subsequently washed four times in 200 µL binding buffer and the fractions collected. Potential aptamers were eluted in 5 mM cGMP (3', 5'-cGMP; Sigma-Aldrich, G6129) supplemented binding buffer for 30 min at room temperature with

agitation and collected by centrifugation. Fractions were analyzed for radioactivity via liquid scintillation counter (Beckman LS6500). The elution fractions were pooled, and buffer exchanged in G-25 Sephadex equilibrated in TET. RNA was precipitated by addition of KCl to a final concentration of 300 mM KCl, 1 µL GlycoBlue, and 2.5 volumes of cold 100% ethanol and incubated at -80 °C overnight. Pellets were washed twice in cold 70% ethanol and resuspended in TET. The selection was carried out over nine rounds, with the introduction of a 30 min counter-selection step (0.5 mM GMP in binding buffer) prior to cGMP elutions, was introduced in the final two rounds.

In vitro selection primers

Selection forward:

5'-GATCTGTAATACGACTCACTATAGGGCAGACGTGCCTCACTAC-3'

Selection reverse:

5'-CTGAGCTTGACGCATTG-3'

Reverse transcription (RT)

Reverse selection primer was annealed to RNA by addition of 1X M-MLV reverse transcription reaction buffer, 0.5 mM dNTPs, 1 mM $MgCl_2$, and 2 µM reverse selection primer by heating to 72 °C for 1 min and snap cooling on ice for 5 min. After cooling, 100 units M-MLV reverse transcriptase (New England Biolabs, M0253) was added and dH_2O to 20 µL. Reverse transcription was carried out by first incubating at 25 °C 5 min, then 15 min at 42 °C, 50 °C, 55 °C, and 65 °C. The reaction was deactivated at 85 °C for 10 min.

Amplification

DNA was amplified in 1X *Taq* reaction buffer, 0.5 μM forward and reverse primers, 0.2 mM dNTPs, cDNA from reverse transcription, and *Taq* DNA polymerase (New England Biolabs, M0273). DNA was amplified with denaturation at 95 °C for 1 min (30 s for subsequent cycles), annealed at 55 °C for 30 s, and extended at 72 °C for 30 s per cycle. Sufficient amplification was verified by 3% agarose gel.

TOPO TA cloning

After nine rounds of selection, TOPO TA (Invitrogen, K4575J10) cloned the enriched pool. The pool was plated on LB + amp, agar plates and individual colonies were amplified by colony PCR using *in vitro* selection primers. Amplicons were analyzed by 3% agarose gel and subsequently purified by 2% agarose gel and isolated by Zymoclean Gel DNA Recovery kit (Zymo Research, D4007). Purified DNAs were Sanger sequenced by Genewiz (Azenta Life Sciences), using *in vitro* selection primers.

Affinity-column binding

Purified, radiolabeled RNAs were resuspended in binding buffer and denatured at 70 °C for 2 min, before being refolded on pre-equilibrated 8-AET-cGMP agarose beads for 30 min at room temperature. The flowthrough was collected after incubation via centrifugation at 3000 *g* for 1 min. The column was quickly washed four times with a column volume of binding buffer, and each wash collected via centrifugation. Columns were eluted four times (30 min each) with a column volume of cGMP (or other nucleotide; varying concentrations) supplemented binding buffer and collected via centrifugation. Beads were washed in 7 M urea and 20 mM EDTA to remove remaining RNAs. The fractions and beads were measured for radioactivity by liquid scintillation counter.

Apta-Seq

Apta-Seq was performed as previously described with modifications [17]. Round 9 of the selected pool was transcribed, purified by denaturing PAGE, eluted, and precipitated. RNA was resuspended in binding buffer to a final concentration of 1 µM. The RNA was denatured at 70 °C for 2 min, then aliquoted into new reaction tubes containing varying concentrations of 3', 5'-cGMP in binding buffer (0 – 10 mM) at a final concentration of 100 nM in 10 µL. The RNA was allowed to equilibrate at room temperature for 15 min to encourage refolding in the presence of 3', 5'-cGMP. After incubation, NAI-N_3, was added to a final concentration of 100 mM. The acylation reaction proceeded for 15 min at room temperature, and was quenched by addition of 90 µL dH_2O, and precipitated with 1 µL GlycoBlue and 2.5 volumes of 100% ethanol at -80 °C for 2 hours. Precipitated RNA was resuspended in dH_2O, and buffer exchanged through G-25 Sephadex pre-equilibrated with TET to remove excess NAI-N_3. RNA was captured on 20 µL DBCO-Sepharose for 15 min at room temperature to remove non-acylated transcripts with six column-volume washes in dH_2O.

Reverse transcription was performed, as previously described, on beads with the Apta-Seq-RT primer. Following reverse transcription, RNA was degraded by addition of 1 µL of 4 M NaOH to the suspension, and heated to 80 °C for 10 min. The cDNA was eluted via centrifugation (3000 *g*) for 5 min and precipitated by edition of 1 µL GlycoBlue and 2.5 volumes of 100% ethanol followed by incubation at -80 °C overnight. The cDNA was precipitated by centrifugation (20,000 *g*) for 20 min and washed three times in 70% cold ethanol. The cDNA was resuspended in CircLigase reaction buffer, 50 µM ATP, 2.5 mM $MnCl_2$, 100 units CircLigase (Lucigen, CL4111K), and dH_2O to 20 µL. The ligation reaction was incubated at 60 °C for 60 min, and heat inactivated at 80 °C for 10 min. The circularized cDNA was subsequently amplified with the Illumina universal forward primer and the CircLigation reverse primer and verified via 3% agarose gel. Non-denaturing PAGE purified amplicons were excised from the gel and eluted into 300 mM KCl over three hours at room temperature and precipitated with the addition of 1 µL GlycoBlue and 2.5

volumes of cold 100% ethanol and incubated at -80 °C for 2 hours. DNA was pelleted via centrifugation (20,000 *g*) for 20 min, and washed twice in cold 70% ethanol, and resuspended TET. Purified DNA was subsequently barcoded for sequencing on Illumina HiSeq 2500 at the UCI Genomics Research and Technology Hub. SHAPE reactivity was determined as previously described[17].

Apta-Seq primers and SHAPE RT primers

Apta-Seq-RT:

5'-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTGCGGCCGCGTGACTGGAGTTCAGACGTGTG

CTCTTCCGATCCTGAGCTTGACGCA-3'

Illumina universal forward:

5´-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3´

CircLigation reverse:

5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCTGAGCTTGACGCATTG-3'

Barcoding reverse:

5'-CAAGCAGAAGACGGCATACGAGAT*NNNNNN*GTGACTGGAGTTCAGACGTGTGCTCTTCCG-3'

G62 FL SHAPE RT:

5'-AGCGATTCAAAATGATGTTGGTT-3'

In-line probing

RNA was transcribed and purified as described previously. Approximately 5 pmol of RNA was 3′ end-labelled in a 10 µL reaction containing RNA ligase buffer, 2.5 µCi [5′-^{32}P] cytidine 3′, 5′-

bisphosphate (Perkin Elmer), and 10 units RNA ligase (New England Biolabs, M0204), incubated at 37 °C for 3 hours. End-labelled RNA was purified by denaturing PAGE. Labeled RNA was denatured at 70 °C for 2 min, then aliquoted (~50 nM) into binding buffer, pH 8, containing 3′, 5′-cGMP at varying concentrations and incubated at room temperature for 15 min. Reactions were subsequently incubated at 37 °C for 40 – 45 hours. Undigested RNA containing no ligand was stored at -80 °C as a no-cleavage control. Sequencing ladders were transcribed, as described previously, with the addition of phosphorothioate nucleotide triphosphate analogues (CαS, GαS, and UαS; TriLink Biotechnologies, N-8006/N-8007/N-8008) at a final concentration of 0.2 mM for each respective nucleotide. Phosphorothioate containing transcripts were cleaved by addition of iodoethanol to a final concentration of 10 mM and incubated at 90 °C for 2 min[34]. In-line probing and sequencing reactions were quenched by buffer exchange through G-10 Sephadex pre-equilibrated in 7 M urea and 20 mM EDTA containing xylene cyanol and bromophenol blue. Reactions were resolved by 10% denaturing PAGE and gels were dried (under vacuum at 80 °C) and exposed to a phosphor imaging screen for at least 24 hours prior to imaging.

SHAPE probing

G62 FL SHAPE RT primer (2 μM) was 5′-end labeled in T4 polynucleotide kinase reaction buffer with ~2 μCi [γ-^{32}P] ATP, 10 units T4 polynucleotide kinase, and dH_2O to 10 μL [35]. The labeling reaction was incubated at 37 °C for 60 min and subsequently purified by denaturing PAGE. Purified primer was resuspended in 10 μL of TET. Purified, unlabeled G62 RNA was resuspended in binding buffer to a final concentration of 1 μM. The RNA was denatured at 70 °C for 2 min, then aliquoted into new reaction tubes containing varying concentrations of 3', 5'-cGMP (or other nucleotide) in binding buffer (0 – 10 mM) at a final concentration of 100 nM in 10 μL. The RNA was allowed to equilibrate at room temperature for 15 min to encourage refolding in the presence of 3', 5'-cGMP. After incubation, NAI-N_3, was added to a final concentration of 100 mM. The acylation reaction proceeded for 15 min at room temperature, and was quenched by addition of

90 µL dH_2O, and precipitated with 1 µL GlycoBlue and 2.5 volumes of 100% ethanol at -80 °C for 2 hours. Precipitated RNA was resuspended in1X M-MLV reverse transcription reaction buffer, 1 µL of end-labeled G62 FL SHAPE RT primer and 0.2 mM dNTPs, and the primer was annealed to RNA by heating to 72 °C for 1 min and snap cooling on ice. After cooling, 100 units M-MLV reverse transcriptase was added and dH_2O to 10 µL. Reverse transcription was carried out by first incubating at 25 °C 5 min, then 15 min at 42 °C, 50 °C, 55 °C, and 65 °C. RNA was subsequently degraded by addition of 1 µL 4 M NaOH and incubated at 90 °C for 2 min. The reaction was then precipitated with addition of 1 µL GlycoBlue and 2.5 volumes of 100% ethanol and incubated at -80 °C for 2 hours. Sequencing lanes were produced by performing reverse transcription on unreacted G62 FL RNA, supplemented with respective 20 µM ddNTP (ddATP, ddCTP, ddTTP; Jena Biosciences, NU-1015/NU-1016/NU-1018). The cDNA was pelleted by centrifugation (20,000 *g* for min) and resuspended in 10 µL 7 M urea and 20 mM EDTA with xylene cyanol and bromophenol blue. Reactions were resolved by 10% denaturing PAGE and gels were dried (under vacuum at 80 °C) and exposed to a phosphor imaging screen for at least 24 hours prior to imaging.

K_D analysis

Dissociation constants (K_D) were determined by using linear least-squares modeling of binding isotherms using the following equation:

$$\text{Fraction bound} = \frac{\left(\frac{[\text{ligand}]}{[\text{ligand}] + K_D} - \text{Baseline}\right)}{\text{Range}}$$

Where “Range” corresponds to a range of values that Apta-Seq peaks, or SHAPE/In-line probing gel band intensities can assume. The Solver add-on for Microsoft Excel was used to create a model fit.

2.6 Supplementary Data

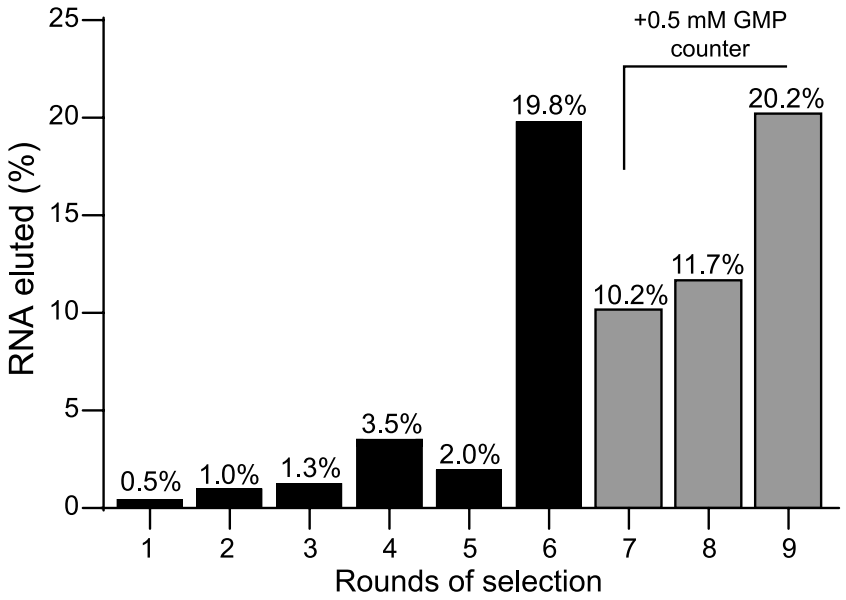


Figure S2.1: Enrichment plots of cGMP selection. Representative enrichment plot for total RNA eluted (%) for each round of genomic SELEX for cGMP. 0.5 mM guanosine 5′-monophosphate (GMP) was introduced as a counter-selection step (prior to elutions) in the seventh round and carry through to the ninth round.

Table S2.1: Reference positions and abundance (%) of each putative cGMP aptamer. Abundance is plotted as a percentage of sequences which plot at a specific genomic position over the total reads of the selection run. Location determined by alignment with bowtie2[36].

reference genome: GRCh37/hg19

Aptamer	Location	Abundance (%)
G2	Chr1:114474090-114474160	0.009
G3	Chr1:121484860-121484970	0.014
G12	Chr2:201982450-201982490	0.002
G13	Chr2:242287530-242287700	0.013
G16	Chr3:143261380-143261470	0.002
G17	Chr4:37279990-37280040	0.002
G18	Chr4:38910640-38910700	0.002
G20	Chr4:77529810-77529890	0.005
G21	Chr5:13780910-13780960	0.002
G22	Chr5:66308640-66308695	0.002
G23	Chr5:92821380-92821430	0.002
G24	Chr5:145442640-145442690	0.002
G27	Chr6:80199790-80199875	0.002
G28	Chr7:151946040-151946240	0.013
G31	Chr8:69075650-69075780	0.059
G32	Chr8:126033340-126033395	0.002
G34	Chr8:133250170-133250230	0.002
G35	Chr9:11283800-11283960	0.028
G36	Chr9:43314700-43314805	0.003
G37	Chr9:80768010-80768150	0.260
G38	Chr9:113703430-113703480	0.002
G39	Chr9:127939965-127940035	0.002
G40	Chr9:140567040-140567200	0.494
G50	chr14:82,619,315-82,619,444	0.070
G51	Chr15:55005050-55005180	0.228
G53	Chr15:60034970-60035120	0.048
G54	Chr15:72945430-72945510	0.022
G55	Chr15:74376410-74376485	0.028
G56	Chr15:75549290-75549370	0.025
G58	Chr16:48078428-48078548	0.014
G59	Chr16:87945550-87945610	0.002
G60	Chr17:71458065-71458115	0.002
G61	CHr19:27738630-27738710	0.007
G62	Chr20:5659560-5659770	94.247
G65	Chr21:11039790-11039990	0.824
G66	ChrX:9251420-9251585	3.563

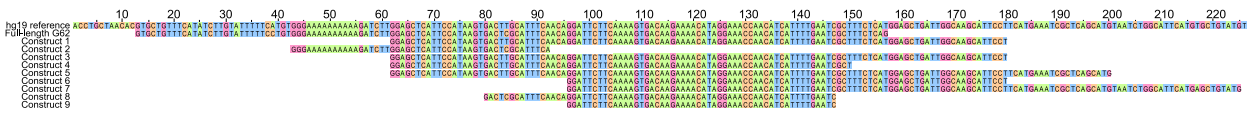


Figure S2.2: Alignment of truncated G62 constructs. Truncations were designed to include predicted binding sites at positions 95 – 145. Construct 2 is the exception and was used as a control for non-specific binding of cGMP. Constructs 1, 3, 5-7 include 3′-upstream extensions determined by comparative conservation with other non-human primates.

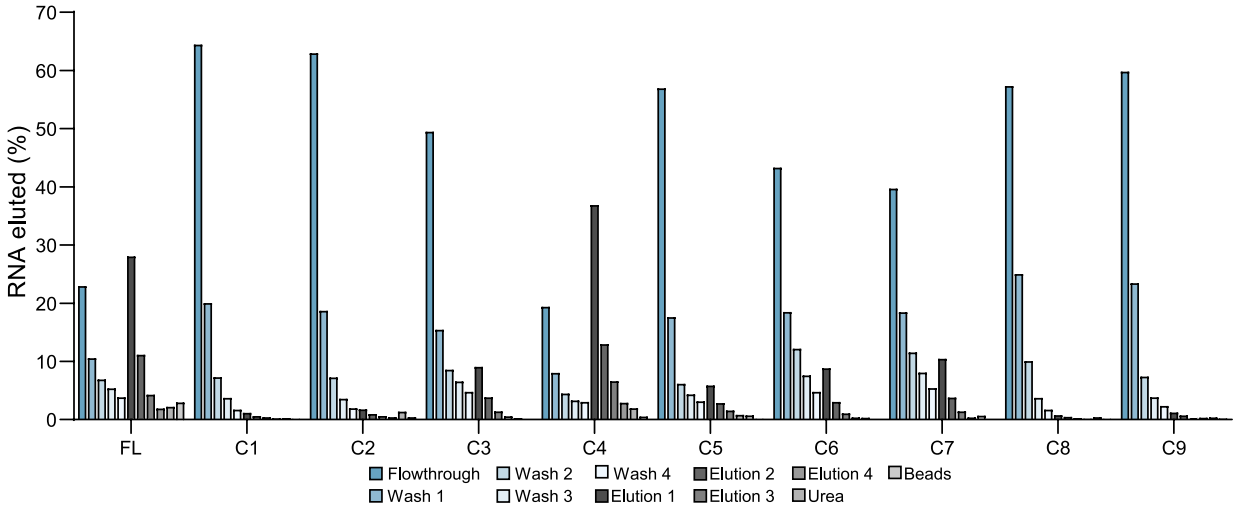


Figure S2.3: Complete binding profiles of G62 constructs. Affinity-column binding profiles of full-length (FL) and other constructs (C1-9). Total elution percentages are reported in Figure 3A.

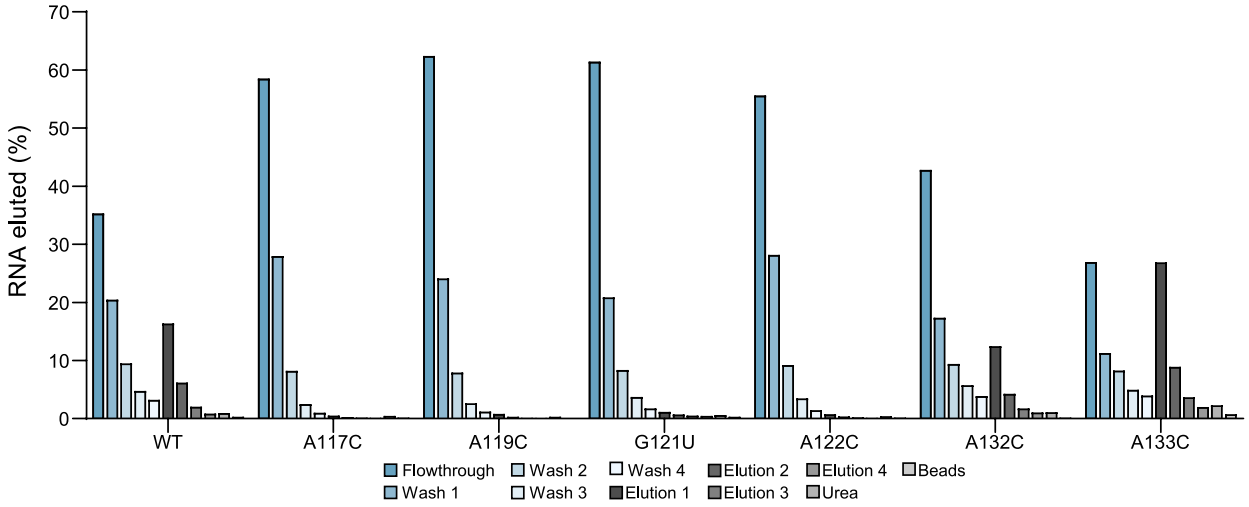


Figure S2.4: Complete binding profiles of G62 Construct 4 mutants. Affinity-column binding profiles of wild-type (WT) and various mutants (listed above). Total elution percentages are reported in Figure 3D.

2.7 References

1. Chauvier, A. & Walter, N.G. Regulation of bacterial gene expression by non-coding RNA: It is all about time! *Cell Chemical Biology* **31**, 71-85 (2024).

2. Breaker, R.R. Riboswitches and Translation Control. *Cold Spring Harb Perspect Biol* **10** (2018).

3. Li, S. & Breaker, R.R. Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucleic acids research* **41**, 3022-3031 (2013).

4. Kubodera, T. et al. Thiamine-regulated gene expression of Aspergillus oryzae thiA requires splicing of the intron containing a riboswitch-like domain in the 5′-UTR. *FEBS letters* **555**, 516-520 (2003).

5. Cheah, M.T., Wachter, A., Sudarsan, N. & Breaker, R.R. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature* **447**, 497-500 (2007).

6. Wachter, A. et al. Riboswitch control of gene expression in plants by splicing and alternative 3' end processing of mRNAs. *Plant Cell* **19**, 3437-3450 (2007).

7. Breaker, R.R. Riboswitches and the RNA world. *Cold Spring Harb Perspect Biol* **4** (2012).

8. Vu, M.M.K. et al. Convergent evolution of adenosine aptamers spanning bacterial, human, and random sequences revealed by structure-based bioinformatics and genomic SELEX. *Chemistry & biology* **19**, 1247-1254 (2012).

9. Curtis, E.A. & Liu, D.R. Discovery of widespread GTP-binding motifs in genomic DNA and RNA. *Chemistry & biology* **20**, 521-532 (2013).

10. Terasaka, N., Futai, K., Katoh, T. & Suga, H. A human microRNA precursor binding to folic acid discovered by small RNA transcriptomic SELEX. *Rna* **22**, 1918-1928 (2016).

11. Carvajal, J.A., Germain, A.M., Huidobro-Toro, J.P. & Weiner, C.P. Molecular mechanism of cGMP-mediated smooth muscle relaxation. *Journal of Cellular Physiology* **184**, 409-420 (2000).

12. Snider, R.M., McKinney, M., Forray, C. & Richelson, E. Neurotransmitter receptors mediate cyclic GMP formation by involvement of arachidonic acid and lipoxygenase. *Proceedings of the National Academy of Sciences* **81**, 3905-3909 (1984).

13. Torre, V., Matthews, H.R. & Lamb, T.D. Role of calcium in regulating the cyclic GMP cascade of phototransduction in retinal rods. *Proceedings of the National Academy of Sciences* **83**, 7109-7113 (1986).

14. Hofmann, F. The cGMP system: components and function. *Biological Chemistry* **401**, 447-469 (2020).

15. Hsu, Y.-T. & Molday, R.S. The cGMP-gated channel of photoreceptor cells: Its structural properties and role in phototransduction. *Behavioral and Brain Sciences* **18**, 441-451 (1995).

16. Yau, K.-W. & Hardie, R.C. Phototransduction Motifs and Variations. *Cell* **139**, 246-264 (2009).

17. Abdelsayed, M.M. et al. Multiplex Aptamer Discovery through Apta-Seq and Its Application to ATP Aptamers Derived from Human-Genomic SELEX. *ACS chemical biology* **12**, 2149-2156 (2017).

18. Kent, W.J. BLAT—The BLAST-Like Alignment Tool. *Genome Research* **12**, 656-664 (2002).

19. Bernstein, S.L., Guo, Y., Peterson, K. & Wistow, G. Expressed sequence tag analysis of adult human optic nerve for NEIBank: identification of cell type and tissue markers. *BMC Neurosci* **10**, 121 (2009).

20. Turunen, T.T. & Koskelainen, A. Determination of basal phosphodiesterase activity in mouse rod photoreceptors with cGMP clamp. *Scientific Reports* **9**, 1183 (2019).

21. Honda, A., Sawyer, C.L., Cawley, S.M. & Dostmann, W.R.G. in Phosphodiesterase Methods and Protocols. (ed. C. Lugnier) 27-43 (Humana Press, Totowa, NJ; 2005).

22. Merino, E.J., Wilkinson, K.A., Coughlan, J.L. & Weeks, K.M. RNA Structure Analysis at Single Nucleotide Resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE). *Journal of the American Chemical Society* **127**, 4223-4231 (2005).

23. Spitale, R.C. et al. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486-490 (2015).

24. Regulski, E.E. & Breaker, R.R. in Post-Transcriptional Gene Regulation. (ed. J. Wilusz) 53-67 (Humana Press, Totowa, NJ; 2008).

25. Fanucchi, S. et al. Immune genes are primed for robust transcription by proximal long noncoding RNAs located in nuclear compartments. *Nature genetics* **51**, 138-150 (2019).

26. Carrieri, C. et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491**, 454-457 (2012).

27. Hu, W.L. et al. GUARDIN is a p53-responsive long non-coding RNA that is essential for genomic stability. *Nature cell biology* **20**, 492-502 (2018).

28. Jain, A.K. et al. LncPRESS1 is a p53-regulated LncRNA that safeguards pluripotency by disrupting SIRT6-mediated de-acetylation of histone H3K56. *Molecular cell* **64**, 967-981 (2016).

29. Marcho, C., Bevilacqua, A., Tremblay, K.D. & Mager, J. Tissue-specific regulation of Igf2r/Airn imprinting during gastrulation. *Epigenetics & chromatin* **8**, 1-10 (2015).

30. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology* **22**, 96-118 (2021).

31. Venkata Subbaiah, K.C., Hedaya, O., Wu, J., Jiang, F. & Yao, P. Mammalian RNA switches: Molecular rheostats in gene regulation, disease, and medicine. *Computational and Structural Biotechnology Journal* **17**, 1326-1338 (2019).

32. Tapsin, S. et al. Genome-wide identification of natural RNA aptamers in prokaryotes and eukaryotes. *Nature Communications* **9**, 1289 (2018).

33. Salehi-Ashtiani, K., Lupták, A., Litovchick, A. & Szostak, J.W. A Genomewide Search for Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene. *Science* **313**, 1788-1792 (2006).

34. Gish, G. & Eckstein, F. DNA and RNA sequence determination based on phosphorothioate chemistry. *Science* **240**, 1520-1522 (1988).

35. Garfio, C.M., Gupta, M. & Spitale, R.C. Using 5′ 32P-labeled Primer and Reverse Transcription to Probe RNA Structure. *Current Protocols* **3**, e830 (2023).

36. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357 (2012).

Chapter 3: High-throughput demarcation of functional RNA domains

Publication note:

This chapter is a written draft of a manuscript in preparation.

Latifi, B.*, **Cole, K.H.***, Vu, M.M.K.*, Lupták, A. High-throughput demarcation of functional RNA domains. *Manuscript in preparation*. (2024).

*These authors contributed equally

Contribution statement:

Conceptualization: A.L., M.M.K.V, **K.H.C.** Methodology: A.L., M.M.K.V, **K.H.C.**, B.L. Investigation: M.M.K.V., **K.H.C.**, B.L. Visualization: B.L. Writing – original draft: A.L., B.L., **K.H.C.** Writing – review and editing: all authors. Funding acquisition: A.L.

3.1 Abstract

RNAs are versatile biomolecules with many functions, such as regulation of ligand-dependent gene expression and catalysis, making them ideal for *in vitro* selection experiments. Many strategies have been implemented to enrich an RNA population of varying length or origin for a selectable function. While successful, defining the ideal motif length or minimal structure within longer libraries is tedious, unless a common sequence or secondary-structure motif is uncovered. To increase the throughput of discovery of functional RNA domains, we have developed a technique that isolates minimal functional RNA domains from complex heterogeneous pools of RNA. This method allows for truncations to occur at the 5′ and 3′ ends of functional domains and replaces primer-binding sequences, thus removing sequence and structure bias introduced by the constant-sequence regions. The removal of primer-binding sequences is of particular importance in selections identifying genomic functional RNAs. We show several examples of genomic and synthetic functional RNAs minimizations. We also demonstrate that the method directly yields information about functional RNA positions where backbone breaks can be introduced to assemble the active RNA from multiple strands, facilitating the development of heterodimeric domains often used in RNA-based sensors. This approach provides a high-throughput pipeline to identify the boundaries of functional RNAs domains and should greatly enhance the rate of discovery of new aptamers, ribozymes, and other functional RNAs.

3.2 Introduction

The versatility of RNAs as functional macromolecules gives them the ability to contain a broad range of functions, such as ligand binding and catalysis. An approach for identifying and engineering RNAs for a specific modality often uses an *in vitro* selection step to evolve RNAs for a specific function and has revealed many diverse functional RNAs (aptamers, ribozymes, and more complex RNAs, such as riboswitches)[1–5]. However, the entire sequence of a functional RNA isolated from *in vitro* selection may not participate in the desired function. This is especially true

of selections using longer libraries, which are more likely to yield RNAs with complex folds[6]. In addition, demarcation of the boundaries of active domains becomes a significant challenge during the discovery of novel functional motifs. Identification of the minimal motif necessary for function can typically be achieved by manual removal of various segments[7] and biochemical confirmation of function. Rational truncations of an RNA, however, typically require a model of the secondary and potentially tertiary structure, which can be informed by sequence-motif discovery, phylogenetic and sequence co-variation analysis, or chemical probing. Identification of the minimal motif by this process can be tedious and especially elusive when secondary structure data are unavailable or difficult to decipher, such as in the case of tertiary structures that dominate the overall fold (e.g., non-canonical quadruplex structures), and sometimes difficult to apply to an entire *in vitro* selected pool.

Previous strategies to identify the minimum functional domain of an RNA sequence have relied on recombination events of its DNA template or by biochemical assessment of single sequences[7–10]. Recently, a combinatorial approach to identify truncated ribozymes was developed[11]; however, this method only allowed for 3′ truncations to occur while retaining the 5′ constant region from the original selection pool. DNA recombination strategies have also been limited to pools derived from single sequences, which can limit the output of discovering functional sequences. However, this limitation can be mitigated using the *in vitro* selection platform, which can accommodate much higher diversity with a practical limit of about 10^{16} molecules at a time[12,13]. Furthermore, a pool enriched for functional RNAs can be further diversified and re-selected to characterize multiple sequences at once without going near the threshold of the *in vitro* selection. Furthermore, genomic selections are not amenable to demarcation of minimal motifs by recombination because recombined sequences would likely lose their biological relevance.

To mitigate these issues, we provide a high-throughput pipeline that experimentally demarcates functional RNA domains within a diverse heterogenous pool and eliminates the need to individually test single truncated RNAs. This method is particularly powerful for revealing functional domains in genomic systematic evolution of ligands by exponential enrichment (SELEX), because it removes artificially introduced primer-binding sequences that are typically required to perform the selection. By using this pipeline, we can not only identify the minimum functional domain(s) necessary for binding but can also determine if the constant-sequence regions are necessary for secondary or tertiary structure elements[14].

This approach features the discovery of functional domains in aptamers and a riboswitch. We experimentally minimized the human *FGD3* adenosine aptamer by not only identifying its minimized functional domain, but also directly revealed how it can be assembled into a functional heterodimer from non-functional strands. We also identified a minimized domain of the Werewolf-1 (Were-1) photoriboswitch and discovered a new Ni-NTA aptamer. We anticipate that this method will increase the throughput of discovery of new functional RNAs domains but also facilitate their engineering into downstream applications (e.g., aptazymes, riboswitches, biosensors, etc.)[15–20].

3.3 Results

We developed a high-throughput method to delineate functional RNA domains in heterogenous populations (synthetic or transcriptomic) by hydrolyzing the RNA and subjecting it to a functional selection (**Figure 3.1**). First, the RNA is partially hydrolyzed under alkaline conditions to promote scission of the phosphodiester backbone (step 1). This approach enables truncations at either 5′, 3′, or both termini of the RNA, resulting in a population of RNAs of varying length. This process is monitored via agarose gel electrophoresis (**Figure S3.1**) to determine the ideal distribution of RNA lengths in the hydrolyzed pooled RNA. This step preserves a small fraction of full-length RNAs, but primarily yields truncated sequences, such as those lacking the

primer-binding regions introduced into *in vitro* selection pools. Deletion of constant-sequence segments can be important when performing *in vitro* selections from genomic pools (genomic SELEX[21]) in which the primer regions do not originate from the genome and can introduce structural bias. In contrast, for *in vitro* selection pools containing short, randomized segments, the primer-binding regions may be essential for the formation of the functional RNA secondary structure[22] and the deletion of these constant region may yield few, if any, functional RNAs. A method for experimental demarcation of active domains in such short pools would likely reveal that the primer-binding sequences are necessary for a function.

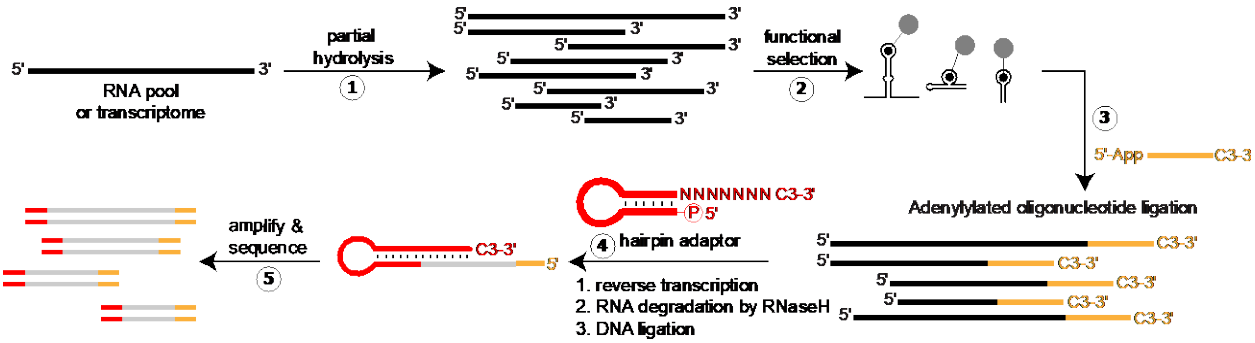


Figure 3.1. A schematic of high-throughput identification of minimized RNA domains. RNA (black) is first hydrolyzed under alkaline conditions to produce truncated transcripts of the RNA population (step 1). An *in vitro* selection is performed to isolate a function of interest (step 2) and a DNA adaptor (yellow) is ligated to the 3′ end of the RNA (step 3) to serve as a primer-binding region for reverse transcription (RT) (step 4). After RT, a DNA hairpin adaptor (red) is ligated to the 3′ end of the newly generated cDNA (gray; step 4) to serve as a primer-binding site for PCR. Finally, the cDNA is amplified with the newly appended primer-binding sites (step 5) and subsequently analyzed by high-throughput sequencing (HTS).

We next performed an *in vitro* selection step to enrich functional RNAs that specifically bind our target molecule of interest (step 2); this selection step enriches for RNAs that remain functional after truncations. RNAs that were captured on the target beads may not contain a consensus sequence for primer binding and therefore a new constant region was introduced through use of an adenylylated oligo (5′-App)[23–26] to facilitate reverse transcription from the 3′ terminus (step 3) (**Figure S3.3A**). Reverse transcription (RT) was performed with the addition of biotinylated dNTPs to help capture newly synthesized cDNA. This step was particularly important to prevent amplification of undesired ligation products, such as concatemerized ligation primers resulting from an excess of RT primer remaining during the RT reaction[27]. Subsequently, RNaseH was added to degrade RNA prior to the ligation of adapters to the 3′ termini of the cDNA (corresponding to the 5′ termini of the selected RNAs). After, biotinylated cDNA was captured on streptavidin magnetic beads to remove excess RT primer prior to ligation. The DNA ligation was then performed to append on a 5′-phosphorylated splint DNA hairpin on the 3′ end of the cDNA with the use of T4 DNA ligase (step 4)[27,28] (**Figure S3.3B**). This hairpin serves as a new primer-binding sequence for the DNA pool regardless of whether the previous 5′ primer region is present or not. Once the ligation was completed, the reaction was washed on streptavidin beads to remove excess ligation hairpin prior to amplification of the newly selected pool. Enriched sequences are then barcoded and submitted for high-throughput sequencing (HTS) to determine the minimum domain(s) of functional RNAs (step 5) or used in another round of selection.

Human *FGD3* adenosine aptamer

The first functional RNA we analyzed is the adenosine-binding aptamer identified in the intron of the human *FGD3* gene[29]. A genomic aptamer pool is an ideal model for validation of our approach (demarcating functional domains) because it requires non-genomic (artificial) primer-binding sites at the 5′ and 3′ ends; these sequences provide constant regions for amplification[21] and serve as promoters for *in vitro* transcription[30–33]. While these regions are necessary for *in vitro*

selections, they do not originate from within the genome and can modulate the structure of the genomic sequences[34]. Therefore, a method that assists in mapping functional genomic RNAs lacking the synthetic primer-binding sequences is desirable. The human *FGD3* aptamer structure was previously discovered using a genomic SELEX experiment, but the functional domain was defined based on the similarity of the sequence to previously-described adenosine aptamers[29,35–40]. A secondary structure-based truncation of the aptamer was attempted, but the presumed minimized sequence only bound ATP beads cotranscriptionally, suggesting that it misfolded after purification under denaturing conditions. The experimental minimization was therefore applied to the genomic pool containing the *FGD3* aptamer, and the HTS sequences were mapped to the reference aptamer sequence (**Figure 3.2A**).

The aligned sequences were first analyzed using bedtools genomecov[41] to assess both the 5′ and 3′ termini of the aptamer. Three major populations were identified with the 5′ termini at positions 1, 57, and 69–76. A smaller number of sequences had 5′ termini around positions 110–140. The 3′ termini were clustered around positions 165, 173, and 181 (**Figure 3.2B**). After, manually aligning the top 40 reads, we noticed a distinct truncation at the 5′ end of the previously defined functional aptamer (**Figure 3.2C**)[29,40]. These data together suggested that the first ~60 nucleotides of the *in vitro* selected sequence are not essential for binding. The 3′ ends of the selected sequences were more heterogeneous, but a clear minimal sequence could be discerned outside of the artificial primer-binding sequences. A construct, *FGD3* (76–165), from the beginning of one of the most prevalent 5′ truncation to the end of the innermost major 3′ truncation was tested in a typical affinity-based experiment (**Figure 3.2D**). It was observed that ~60% of RNA was eluted off the beads in the presence of excess ATP, suggesting that a robust adenosine-binding aptamer was identified and minimized.

We also tested a second approach for ligating the cDNA termini using CircLigase, rather than random splint ligation using T4 DNA ligase (**Figure S3.4**). In this selection scheme, the *FGD3*

aptamer was hydrolyzed as described above, but the RT step was performed with a primer containing the constant region from the original aptamer pool with both the Illumina forward and reverse primers. Subsequently, CircLigase was used to circularize the cDNA and then amplification was performed to further enrich for new aptamers. After 4 rounds, random priming with the Illumina reverse sequence was performed in the RT step to increase the sequence coverage and one more round of *in vitro* selection was ran prior to submission to HTS. Afterwards, the HTS sequences were again mapped to the full-length aptamer sequence and analyzed for minimal and non-overlapping (split-aptamer) sequences. While a sequence like the minimized *FGD3* aptamer was identified, two other striking results were revealed by the analysis of the sequencing data. First, an internal deletion [*FGD3* (Del)] was observed within the *FGD3* aptamer. This internal deletion exhibited high affinity during column binding and ~60% of the aptamer eluted in the presence of ATP; this was similarly observed for the *FGD3* (76–165) aptamer (**Figure 3.2E**). Second, a split aptamer in the form of a heterodimer consisting of two strands of the minimized *FGD3* aptamer was identified, while the individual strands did not bind the ATP column (**Figure S3.5**). When both sequences were transcribed simultaneously, they formed a functional heterodimer and ~3% of RNA bound and was eluted using free ATP (**Figure 3.2F**). These results show that the method not only demarcates the functional domain, but also provides a potential approach for directly discovering split aptamers. While the results obtained from the CircLigase minimization scheme yielded information about the FGD3 aptamer, the high-throughput pipeline using the 5′-App oligo ligation (**Figure 3.1**) was pursued due to limited sequence bias, ease of workflow and cost-effectiveness.

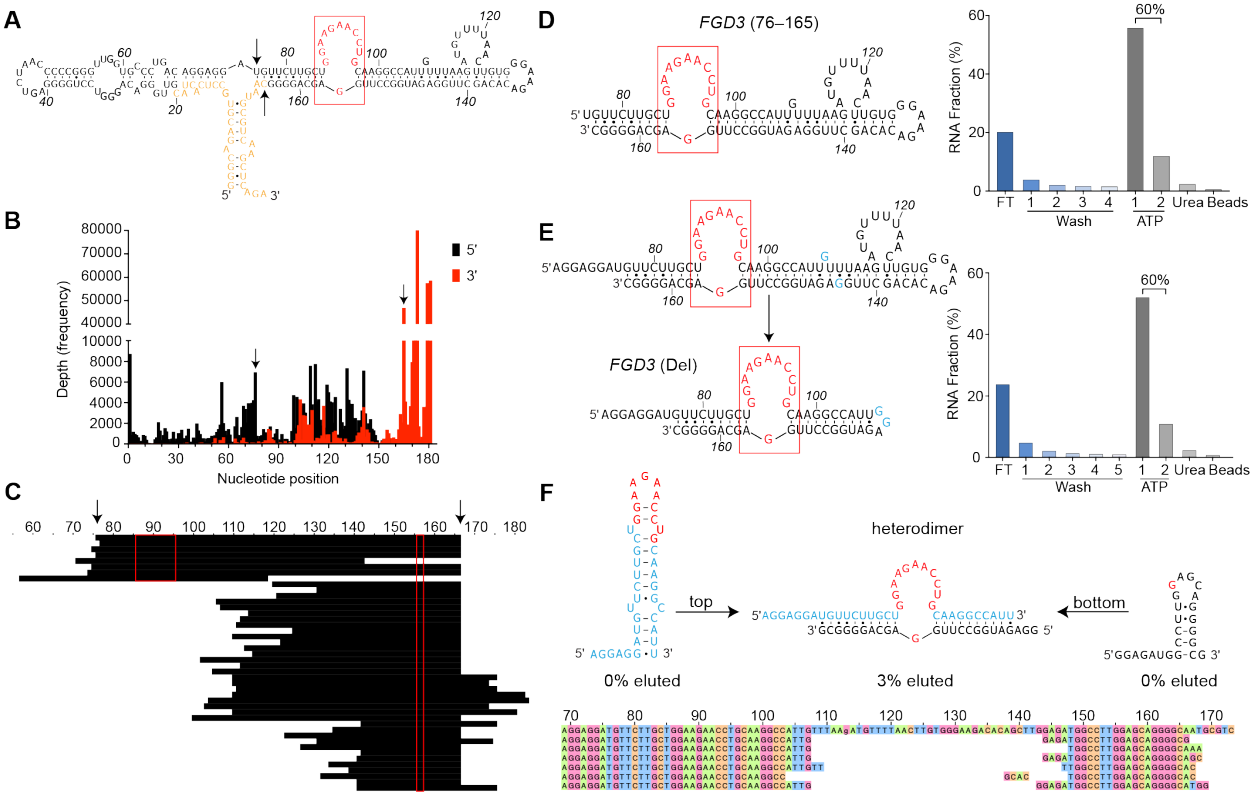


Figure 3.2. Experimental demarcation of the human *FGD3* adenosine aptamer domain. A. The secondary structure model of an aptamer revealed by genomic SELEX[29]. The adenosine binding loop is outlined with a red box; yellow nucleotides represent the primer-binding regions from the *in vitro* selection pool. The arrows mark frequent truncations that commonly occurred at the 5′ and 3′ ends of the reselected truncated pool. **B.** Mapping of minimized sequences to the full-length genomic *FGD3* aptamer. The graph displays the distribution of the 5′ (black) and 3′ (red) termini. **C.** Alignment of the top 40 most common sequences from the minimization experiment. Red boxes show the location of the adenosine-binding loop. **D.** *FGD3* (76–165) aptamer secondary structure and its binding to ATP beads. FT: flow-through; ATP: competitive elution with free ATP (5 mM) **E.** An internal deletion observed in the *FGD3* aptamer. The nucleotides highlighted in blue (107 and 144) show the positions of the deletion within the sequence. ATP column binding profile of the *FGD3* (Del) shows ATP-beads binding and elution profile almost identical to the truncated sequence shown in D. **F.** A heterodimeric minimized domain revealed by modeling of the co-folding of non-overlapping sequences shown in the alignment in C. The binding of individual strands (shown as % competitively eluted) was compared with ATP column binding by the heterodimer (see Figure S3.6 for full plots). The blue (top strand) and black (bottom strand) nucleotides within the split heterodimer represent the sequences from the 5′ and 3′ segments, respectively. The alignment shows the shortest non-overlapping sequences predicted to form the minimal split-aptamer domain.

Identification of Were-1 photoriboswitch aptamer domain

Next, we applied this minimization approach to a heterogenous pool of synthetic riboswitches. A photoriboswitch, Werewolf-1 (Were-1), was previously found to bind a photoactive molecular switch [amino *trans* stiff-stilbene (a-*t*SS)] with an approximate K_D value of 1.5 μM[42]. However, efforts aimed at identifying the minimal aptamer domain of Were-1 through chemical and enzymatic probing have been confounded by the conformational changes of the riboswitch upon either binding of the ligand or subsequent rearrangements of the expression platform.

To demarcate the minimal ligand-binding domain of Were-1, the full-length construct (**Figure 3.3A)** was doped (10%) into a selected riboswitch pool and hydrolyzed under alkaline conditions prior to selection on magnetic beads conjugated with a-*t*SS. The RNA captured on beads was eluted in the presence of a-*t*SS and sequenced after one round of *in vitro* selection. Sequences obtained from HTS were mapped to the Were-1 parent sequence and analyzed for minimal and non-overlapping (split-aptamer) sequences as described above. First, we measured the overall depth of reads at each position to discern which segments were enriched by the binding assay (**Figure 3.3B**). Positions 25–67 of the riboswitch were heavily represented and further decreased in frequency approaching the 5′ and 3′ termini of the riboswitch. Three positions, nucleotides 1, 25, and 109, were identified to be the most common 5′ termini, whereas the most frequent 3′ truncations occurred at nucleotides 67, 103, and 147 (**Figure 3.3C**). Manual alignment of the 20 most common sequences represented by these truncations could be categorized into three populations (**Figure 3.3D**): the first population aligned to the middle of the riboswitch, roughly overlapping with the random-sequence region of the starting pool, while the other two populations aligned to either end of the riboswitch, suggesting that they were parts of a of functional split-aptamer heterodimer. A sequence spanning positions 25–103 was among the most abundant presumed *cis* constructs. This sequence, Were-1 (25-103), was not only highly abundant, but also represented the consensus region observed in Figure 3B (25–67) and

supported reoccurring 5′ and 3′ truncations shown in Figure 3C. The open regions of Were-1 (25–103) were probed to confirm its binding activity to a-*t*SS using RNase T1 footprinting[43,44]. We observed a ligand-dependent reduction in RNase T1 digestion within Were-1 (25–103) as a-*t*SS concentration increased. An apparent dissociation constant (K_D) value was determined from four distinct sites (G42, G46, G49, and G77). An apparent K_D of 5.3 μM a-*t*SS was derived from these four sites, revealing an affinity similar to the full-length Were-1 photoriboswitch. The combination of ligand binding and RNA structure probing revealed a minimized aptamer domain for the Were-1 riboswitch (**Figure 3.3F**). A secondary structure model based on the open segments suggested that the minimized Were-1 (25-103) requires a flanking structured region to form a structure that binds a-*t*SS. In addition, positional frequency of the Were-1 riboswitch suggested that using the high-throughput platform on the single construct versus within a heterogenous pool had similar distribution of both the 5′ and 3′ termini (**Figure S3.7**).

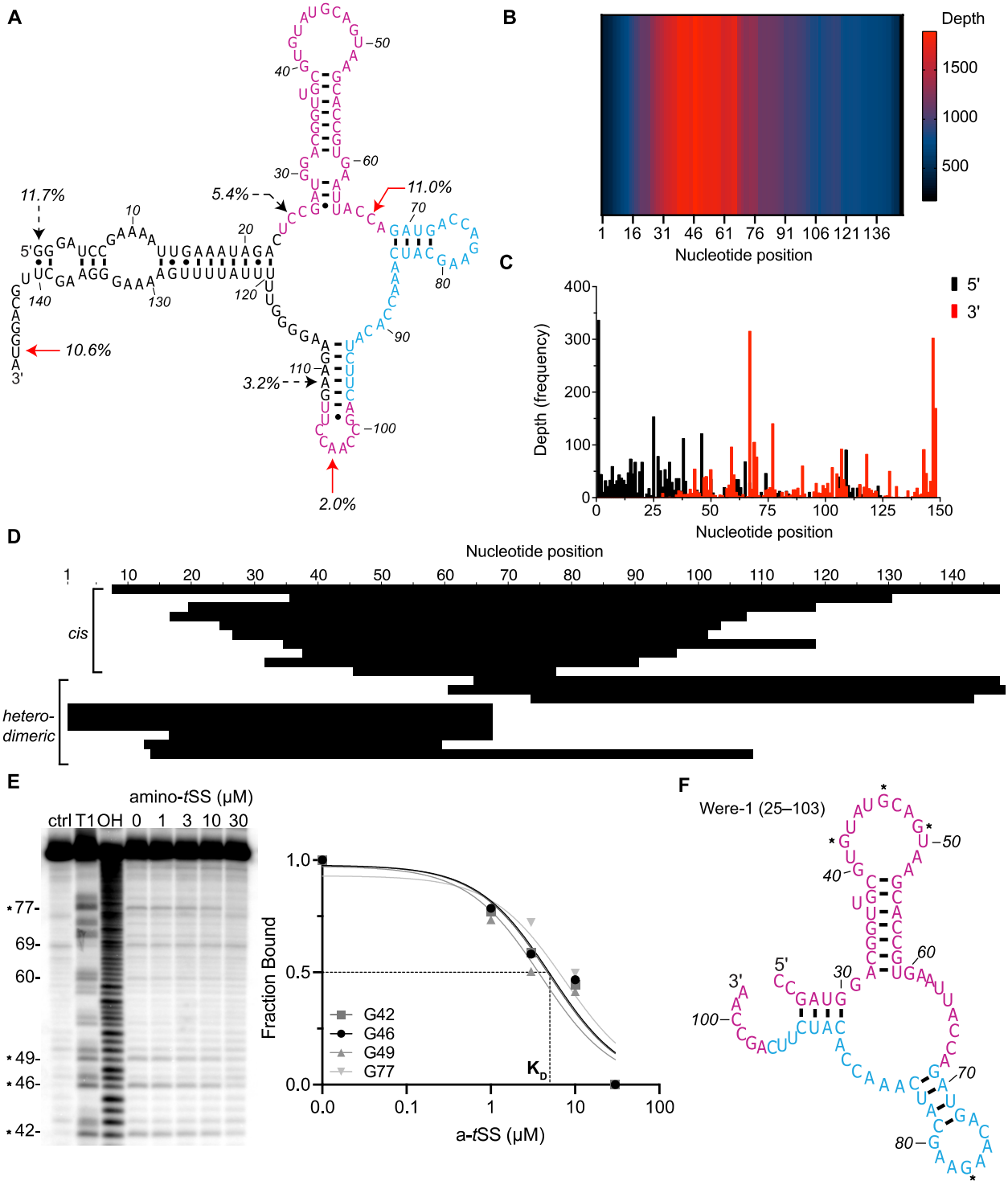


Figure 3.3. Demarcation of the minimized Were-1 photoriboswitch aptamer domain. A. Mapping of minimized sequences to the Were-1 riboswitch. The original model of the riboswitch secondary structure, with the pink letters representing positions derived from random regions of the starting *in vitro* selection pool and blue representing partially randomized positions. Nucleotides in black represent primer-binding regions of the pool. **B.** Heatmap depicting nucleotide positions with the highest representation among the truncated and re-selected sequences (red) and regions with lower frequency (blue) mapped to the Were-1 riboswitch. **C.**

Most common 5′ and 3′ termini amongst truncated sequences after selection for a-*t*SS binding. The three most frequent start and end positions, labeled in red and blue respectively, are mapped to the secondary structure model in A (red arrows). **D.** Top 20 most abundant truncated sequences aligned to full-length Were-1. Three populations were observed, sequences aligned to the middle (*cis*) and two populations on either end of the riboswitch (*heterodimeric*). **E.** RNase T1 probing of Were-1 (25–103). Lanes from left to right: undigested RNA (ctrl), digested G-sequencing control (T1), partial hydrolysis ladder (OH), and partial T1 digestion in the presence of increasing a-*t*SS concentrations, with ligand concentrations indicated above the gel image. Plot of the estimated fraction of RNA bound versus a-*t*SS concentration. Fraction bound values were estimated by measuring the band intensities at four sites (G42, G46, G49, and G77). Band intensities were normalized to positions G60 and G69. A lower contrast version of the gel image is shown in **Figure S3.6**. Secondary structure model of the Were-1 (25–103) aptamer**.** Asterisks indicate positions that were less susceptible to RNase T1 cleavage in the presence of ligand.

Identification of a Ni-NTA aptamer

To test our method of functional RNA minimization without *a priori* knowledge of the sequence identity or biochemical information, we used an *in vitro* selected pool containing previously identified Ni-NTA aptamers [45]. Analysis of the HTS data revealed an abundant sequence, which was further mapped to identify minimal functional variants (**Figure 3.4A**). To assess the subpopulations within the pool, we used FastAptamer[46] to count and cluster the sequences and compared the top 10 clusters; the top 4 clusters represented ~92% of the sequences within the pool (**Figure 3.4B**). Mapping of the 5′ termini showed nucleotide positions 1, 4, 7 and 11 had the greatest abundance, while a dominant 3′ truncation was apparent at nucleotide position 36 with an abundance of 89.2% (**Figure 3.4C**). Sequence alignment of the top 10 clusters revealed the putative minimal domain (**Figure 3.4D**). The 10 most abundant clusters of the novel Ni-NTA aptamer contain a purine-rich segment (position 21-30), as have previously identified Ni-NTA binding motifs[47]. The predicted secondary structure of the suggested Ni-NTA aptamer indicated that a portion of the 5′ constant region (blue) played an important role in the structure formation. A construct that started from the beginning of one of the most prevalent 5′ truncations (position 11) that was also predicted to form the same secondary structure, combined with the most common 3′ truncation, was tested for Ni-NTA affinity (**Figure 3.4E**). We observed that ~33% of RNA was eluted in the presence of imidazole at concentrations higher than 25 mM, supporting the model that the sequence represents a new Ni-NTA aptamer.

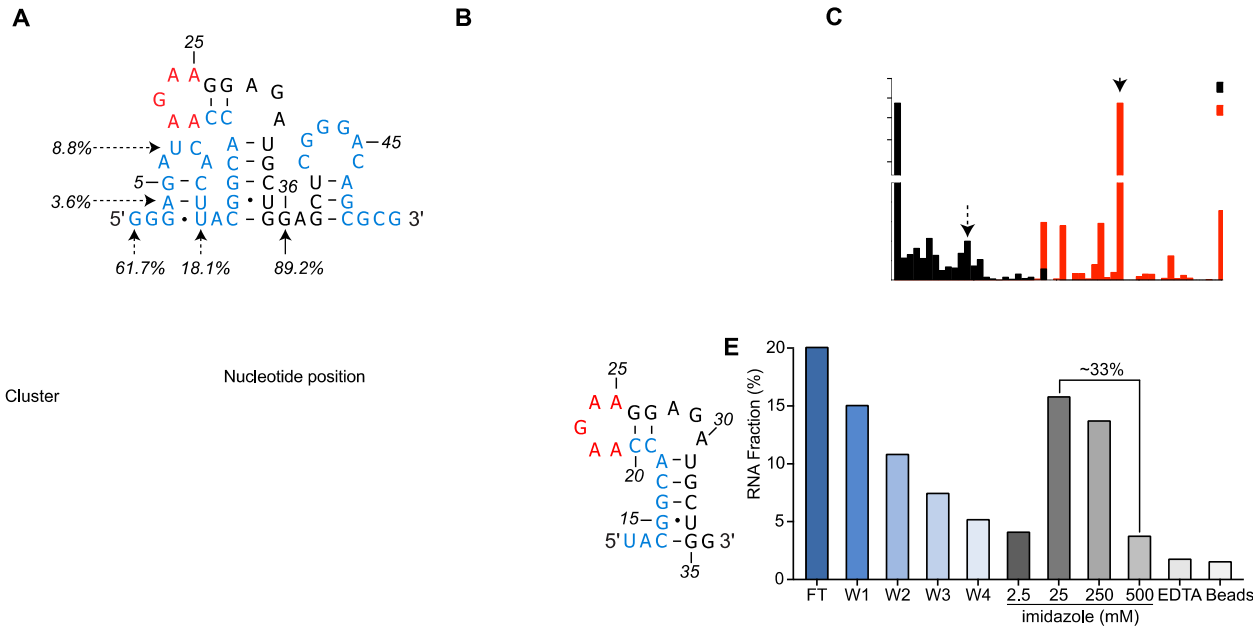


Figure 3.4. Identification and minimization of a novel Ni-NTA aptamer. A. Full-length sequence and proposed secondary structure of a novel Ni-NTA–binding aptamer. Regions in blue represent the primer-binding regions from the *in vitro* selection pool. Nucleotides in black and red are from the random regions of the pool. Arrows indicate either 5′ (dashed arrows) or 3′ truncations (solid arrows). **B**. The top ten clusters were plotted to show the distribution of abundance within the mapped populations. **C**. The read depth of the 5′ start sites (black) and 3′ truncation sites (red) mapped to the parent sequence. **D**. The top 10 clusters aligned to the parent sequence (A). The red outlines the common region among the sequences and the secondary structure model of the minimized Ni-NTA aptamer. **E**. The minimized Ni-NTA aptamer was immobilized on Ni-NTA resin and eluted in the presence of imidazole at indicated concentrations to assess the binding affinity of the aptamer.

3.4 Discussion

The discovery of functional RNAs has led to many advances in RNA biology and therapeutics. One limitation in increasing the output of these innovations is the identification of the minimal motifs necessary for the given function. *In vitro* selection is one method that is commonly used to discover new functional RNAs. Typically, an RNA identified by this method does not require the entire sequence to fulfill its biochemical role and the demarcation of the functional domain boundary is necessary for downstream applications. Conversely, functional RNAs may also be isolated from pools that are too short to yield functional RNAs solely from their heterogenous (random) regions[22] or are structurally and informationally complex, requiring primer-binding regions to form the active motifs[48]. This situation can be particularly difficult to interpret if the 3' constant region participates in the functional motif and the RNA is characterized using chemical probing methods that require reverse transcription for readout, such as approaches that analyze solvent accessible segments of the RNA using alkylating and acylating reagents[49–51]. All these scenarios would greatly benefit from a high-throughput approach that yields information about the minimal sequence necessary for the chosen biochemical function.

Here we developed a high-throughput platform that identified functional RNA domains within heterogenous RNA pools. This system identified multiple functional RNAs from both synthetic and genomic RNAs, but also led to the first direct discovery of split fragments of a functional aptamer. The development of this tool will enable researchers to remove sequence bias that can be often found with the presence of the constant regions introduced within the *in vitro* selection platform. By creating a system to interchange these sequences, we increase the opportunity to correctly identify functional regions within an RNA without losing their biological relevance. As an added benefit, the installation of these new primer regions can allow for the development of new workflows to be adapted in many biological systems, where these sequences serve as promoters for transcription and/or translation. Furthermore, the use of this system in

combination with the *in vitro* selection will rapidly increase the discovery of new functional RNAs. This optimized design significantly reduces the tedious workflow of biochemically assessing each individual construct identified from *in vitro* selection pools by demarcating the functional domains at a high-throughput scale.

Our hope is by leveraging this system, the ability to directly discover split aptamers can further expand the opportunity for functional RNAs to be applied in multiple settings such as electrochemical or biosensors to improve upon false positive or nonspecific signals and split ribozymes in biological settings to downregulate specific genes of interest. One exciting approach would be to identify novel split light up aptamers to use in RNA dynamics and localization that is often seen in split-protein based systems. We anticipate that the creation of this system will enable fast identification of diverse functional RNAs, which will not only increase discovery of novel targets, but help to engineer RNAs for multiple modalities. While we focused on the discovery of demarcating functional RNAs within synthetic and genomic RNA pools, we envision this platform accommodating the discovery of functional RNAs within a transcriptome. Extracting the transcriptome from eukaryotic or prokaryotic systems and hydrolyzing the RNA to discover new targets will expand the capabilities in sensor design, biomarker discovery, and RNA therapeutics

3.5 Materials and Methods

Synthesis of DNA oligonucleotides

All DNA sequences with or without modification(s) were purchased from Integrated DNA technologies (IDT).

Generation of RNA fragments using alkaline hydrolysis

Transcription reactions were performed at 37 °C for 2 hrs in 1X transcription buffer [40 mM Tris-HCl (pH 7.5) and 2 mM spermidine], 16 mM $MgCl_2$, 2 mM rNTPs each, 10 mM dithiothreitol (DTT),

and one unit of T7 RNA polymerase (house preparation). After, 4 units of DNase I (New England Biolabs) and a 1X final concentration of DNase I reaction buffer (New England Biolabs) was added to each reaction, incubated for 1 hr at 37 °C and purified twice using the RNA Clean and Concentrator kit (Zymo Research). RNA was then hydrolyzed under alkaline conditions (5 mM KOH, pH 10) at 90 °C. Time-points were taken during hydrolysis and then quenched in 100 mM Tris-HCl pH 7.5. Aliquots were assessed using a 3% agarose gel electrophoresis to determine RNA hydrolysis progression (**Figure S3.1A**). Hydrolyzed RNA was then buffer-exchanged using the RNA Clean and Concentrator kit and eluted in TET buffer [10 mM Tris-HCl (pH 7.5), 0.1 mM EDTA, and 0.001% Triton X-100].

***In vitro* selection**

Kit-purified RNAs were resuspended in a 1X binding buffer [140 mM KCl, 10 mM NaCl, 10 mM Tris-HCl (pH 7.5), and 5 mM $MgCl_2$] and subsequently refolded by heating to 60 °C for 5 mins and then cooled to 25 °C over the next 10 mins. The folded RNA was then incubated with ligand-conjugated beads in a Spin-X tube (Corning, NY, USA) for 20 mins. RNA was then centrifuged for 1 min at 3000 *g* to remove unbound RNAs and beads were washed four times with 1X binding buffer by incubating for 5 mins and centrifuging to remove non-specific binders. Finally, the RNA was eluted with free ligand [e.g., 100 µM of amino *trans*-stiff stilbene (a-*t*SS) or 5 mM ATP-$MgCl_2$ in 1X binding buffer for the selections of a-*t*SS or adenosine aptamers, respectively] by incubating for 30 mins at room temperature. The Ni-NTA aptamer was eluted in 7 M urea and 40 mM EDTA for 5 mins. Elutions were collected and either extracted with phenol and chloroform (a-*t*SS aptamers) or kit purified prior to precipitation with 300 mM KCl, 1 µl of GlycoBlue (Invitrogen, MA, USA) and 2.5 volumes of cold 100% ethanol at -80 °C. RNA was pelleted and then resuspended in TET buffer.

Addition of universal 3′ adapter to RNAs

To ligate a 3′ adapter to the truncated RNAs, the adapter oligo was first adenylylated in a reaction containing 1X T4 RNA ligase reaction buffer (New England Biolabs), 25 μM rATP, 20% PEG8000, 10 μM primer containing a 5′ monophosphate, and 20 U of T4 RNA ligase 1 (New England Biolabs) for 12 hrs at 25 °C[23,24,26]. The reaction was buffer exchanged into TET buffer using 10-times the reaction volume of G10 Sephadex beads (Sigma-Aldrich) with centrifugation at 2000 *g* and stored at -80 °C. The adenylylated oligo had the following primary structure, in which 5′-App-A and 3′SpC3 represent the 5′-5′ diphosphate-linked adenosines and a 3′ carbon spacer designed to prevent concatemerization of the oligo, respectively:

3′ adapter: 5′ -App-AGATCGGAAGAGCACAC- 3′SpC3

The hydrolyzed RNA was resuspended in TET buffer and added to a reaction containing 5 nM of the adenylylated 3′ adapter, 15% PEG8000, 1X RNA T4 ligase reaction buffer (New England Biolabs), and T4 RNA ligase 2, truncated KQ (New England Biolabs). The reaction was incubated at 25 °C for 18 hrs and 4 °C for 2 hrs prior to purification using the RNA Clean and Concentrator kit. RNA was then precipitated in 300 mM KCl, 1 μl of GlycoBlue (Invitrogen), and 2.5 volumes of cold 100% ethanol and pelleted by centrifugation.

Reverse transcription (RT)

RNA pellets were resuspended in TET buffer. Subsequently, 0.18 mM dNTPs, 0.02 mM biotinylated dUTP and dATP (Jena Bioscience) and 5 nM of the reverse primer were added to each reaction. The RNA and primer were annealed by heating at 90 °C for 1 min and cooled on ice for 3 mins before the addition of 10 mM DTT, 1X Protoscript buffer (New England Biolabs), 150 U of Protoscript II reverse transcriptase (New England Biolabs), 15 U of WarmStart RTx reverse transcriptase (New England Biolabs), and 4 U of *Bst* 3.0 DNA polymerase (New England Biolabs) to each reaction. The reaction was initiated at 25 °C for 10 mins and then the temperature

was ramped up to 35 °C for 10 mins, 42 °C for 30 mins, 55 °C for 15 mins and 65 °C for 30 mins. Finally, the enzymes were inactivated at 70 °C for 10 mins.

RT primer: 5′ -GTGTGCTCTTCCGATCT- 3′

Binding to streptavidin beads

5 U of RNaseH (New England Biolabs) was added to each RT reaction and incubated at 37 °C for 1 hr. After, cDNA was buffer-exchanged into G10 Sephadex beads with TET buffer by centrifugation at 2000 *g* to remove free biotinylated dATP and dUTP. The cDNA was then incubated with magnetic streptavidin beads (Dynabeads™ MyOne™ C1, Invitrogen) that were pre-equilibrated with streptavidin wash buffer (SWB; 60 mM Tris, 50 mM KCl and 0.001% Triton X-100) and incubated for 20 mins at room temperature. After, five washes were performed with SWB to remove excess RT primer.

Addition of universal 5′ PCR adapter to 3′ of cDNA

Ligation was performed on the magnetic beads following the RT step. The 3′ end of the cDNA was ligated with a 5′ phosphorylated splint hairpin containing a 3′ carbon spacer to attach a constant-sequence region (T7 RNA polymerase A1 promoter sequence) to minimized sequences. The ligation was performed with a final concentration of 1X T4 DNA ligase reaction buffer (New England Biolabs) and 400 U of T4 DNA ligase (New England Biolabs), 1 mM rATP, 0.5 M betaine, 20% PEG8000 and 5 nM of splint hairpin. The reaction was incubated at 16 °C for 6 hrs, 30 °C for 6 hrs and inactivated at 65 °C for 15 mins.

Splint hairpin for 1st round of *in vitro* selection:

5′ -P-CCTATAGTGAGTCGTATTAAAAAAATATAGGNNNNNNN- 3′SpC3

Splint hairpin for 2+ rounds:

5′ -P-TATAGTGAGTCGTATTAAAAACTATANNNNNNN- 3′SpC3

The DNA oligos contained a 5′ phosphate to facilitate ligation by T4 DNA ligase and a 3′ carbon spacer to prevent self-ligation. Both oligos were predicted to fold into a hairpin secondary structure with a random sequence (NNNNNNN) acting as a non-sequence-specific splint.

The ligated cDNA was washed five times with SWB and incubated for 5 mins on an orbital shaker to remove excess splint hairpin oligo. Two washes with 200 mM KOH were performed to remove cDNA from the magnetic streptavidin beads. After, elutions were pooled and precipitated with 1 µl of GlycoBlue (Invitrogen, MA, USA) and 2.5 volumes of cold 100% ethanol, and pelleted by centrifugation. The pellet was resuspended in TET buffer.

Library amplification by PCR

PCR amplification was performed on cDNA using a final concentration of a 1X Standard *Taq* reaction buffer (New England Biolabs), 0.2 mM dNTPs, 0.5 µM forward and reverse primer, and 1.25 U of DNA *Taq* polymerase (New England Biolabs). The reaction was amplified for 16 to 28 cycles (95 °C for 30 s, 50 °C for 30 s, and 72 °C for 30 s) to determine the optimal amplification (**Figure S3.1B**).

Forward primer (1st selection round):

5′ - CCTATATTTTTTTAATACGACTCACTATAGG - 3′

Forward primer (2nd round+): 5′ - TATAGTTTTTAATACGACTCACTATAGG - 3′

Reverse primer: 5′ - GTGTGCTCTTCCGATCT - 3′

HTS sequencing, processing, and mapping

DNA was then prepared for sequencing by first amplifying the DNA library with an Illumina forward adapter and reverse primer (above) to incorporate the Illumina adapter during PCR. The reaction

was amplified for 4 to 8 cycles (95 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s). After, the Illumina barcodes were added in a second PCR reaction and amplified for 1 cycle (95 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s) and then 7 cycles (95 °C for 30 s, 65 °C for 30 s, and 72 °C for 30 s). Both PCR reactions contained a final concentration of a 1X Standard *Taq* reaction buffer (New England Biolabs), 0.2 mM dNTPs, 0.5 µM forward and reverse primer, and 1.25 U of DNA *Taq* polymerase (New England Biolabs). All PCR products were visualized on a 3% agarose gel prior to purification with a DNA Clean and Concentrator kit (Zymo Research). The DNA libraries were then submitted for Illumina sequencing.

Illumina forward adapter:

5′ -ACGACGCTCTTCCGATCTCCTATATTTTTTTAATACGACTCACTATAGG- 3′

Sequenced reads were obtained using an Illumina MiSeq. The following discusses how each target of interest was processed to map sequences.

Mapping of reads

***FGD3* adenosine aptamer: Analysis of sequences generated using CircLigase**

Reads were merged with PEAR[52] using default settings. Adapters were clipped with cutadapt:

-a CAATGCGTCAAGCTCAG -g
GATCTGTAATACGACTCACTATAGGGCAGACGTGCCTCACTAC -m 40

Bowtie2[53] was used to map the reads to reference sequences, which corresponded to the human *FGD3* adenosine aptamer genomic sequence (Genbank JX863074). The parameters -N 1 and –local were used. For more permissive mapping, no minimum length was used during clipping and Bowtie1 scoring options were adjusted to -min_score G, 40, 8

Identification of unmapped reads

Sequences from high-throughput sequencing were analyzed using FastAptamer[46,54] count and cluster options. Clustering was performed with an edit distance of 10 with no filter flag. The top read per cluster was retrieved via grep -A 1 '[0-9]-[0-9]\{1,\}-1-0' and aligned manually in Jalview.

***FGD3* adenosine aptamer: Analysis of sequences generated using T4 DNA ligase**

Mapping of reads

Reads were merged with PEAR using default settings. Adapters were clipped with cutadapt[55]:

-g TATAGTTTTTAATACGACTCACTATA

Bowtie2:

Alignment to FGD3: bowtie2 -x FGD3 -N 1 --local -q filename.fastq -S filename.sam

Samtools:

samtools view -b -S filename.sam > filename.bam

samtools sort filename.bam -o filename.sorted.bam

samtools index filename.sorted.bam

samtools fasta -F4 filename.sorted.bam > filename.sorted.fasta

Bedtools genomecov:

Depth: bedtools genomecov -d -ibam filename.sorted.bam

5’ truncations: bedtools genomecov -d -strand + -5 -ibam filename.sorted.bam

Reverse reads for 5’ truncations: bedtools genomecov -d -strand - -3 -ibam filename.sorted.bam

3’ truncations: bedtools genomecov -d -strand + -3 -ibam filename.sorted.bam

Reverse reads for 3’ truncations: bedtools genomecov -d -strand - -5 -ibam filename.sorted.bam

Were-1 riboswitch:

Mapping of reads

Reads were merged with PEAR using default settings. Adapters were clipped with cutadapt[55]:

-g TATAGTTTTTAATACGACTCACTATA

Bowtie2:

Alignment to Were-1: bowtie2 -x Were-1 --end-to-end -q filename.fastq -S filename.sam

Samtools:

Sam files were converted to bam files as previously described.

RStudio:

Library(GenomicAlignments)

Filename <- system.file(“extdata,” “filename.sorted.bam,” package = “Rsamtools”)

Filename_R1 <- readGAlignments(filename)

Write.csv(filename_R1, “location for csv”)

Bedtools genomecov:

Bedtools genomecov files were processed as previously described.

Ni-NTA aptamer:

Mapping of reads

Reads were merged with PEAR using default settings. Adapters were clipped with cutadapt[55]:

-g TATAGTTTTTAATACGACTCACTATA

Bowtie2:

Alignment to Ni-NTA: bowtie2 -x Ni-NTA -N 1 --local -q filename.fastq -S filename.sam

Samtools:

Sam files were converted to bam files as previously described.

Bedtools genomecov:

Depth: bedtools genomecov -d -strand + -ibam filename.sorted.bam

5’ truncations: bedtools genomecov -d -strand + -5 -ibam filename.sorted.bam

3’ truncations: bedtools genomecov -d -strand + -3 -ibam filename.sorted.bam

Bowtie2 was used first to remove any reads matching the phiX genome and then to map sequences to the desired reference sequence. The parameter – N 1 and --local or --End-to-end were used when aligning aptamer to their respective sequences. After, samtools was used to sort and index mapped files and sequence alignments were visualized either using Jalview or in PRISM. Bedtools genomecov was used to determine the most common 5′ start sites and 3′ truncations and this helped to dictate design of minimized functional domains.

Column affinity chromatography of minimal RNAs

The transcription reactions of individual sequences were performed at 37 °C for 2 hrs in 1X transcription buffer, 16 mM $MgCl_2$, 2 mM each rGTP, rUTP, rCTP and 0.2 mM rATP, 3 µCi [α-^{32}P] ATP (Perkin Elmer), 10 mM dithiothreitol (DTT), and one unit of T7 RNA polymerase. After, 4 U of DNase I (New England Biolabs) and a 1X final concentration of DNase I reaction buffer (New England Biolabs) were added to each reaction; the reaction was incubated for 1 hr at 37 °C. RNAs were kit-purified using an RNA Clean and Concentrator and then refolded at 60 °C for 5 mins in

1X binding buffer. The RNA was cooled to 25 °C for 10 mins before binding to ligand-conjugated beads. The beads were subsequently washed for (5 mins per wash) using 1X binding buffer and collected by centrifugation at 2000 *g* for 30 s. Ligand of interest (respective to RNA) was supplemented into 1X binding buffer and incubated for 20 mins prior to collecting and each fraction was analyzed using a liquid scintillation counter (Beckman Coulter LS6500).

T1 nuclease probing

The HH-Were-1 minimized RNA construct (25-103; **Figure S2**) was transcribed using conditions described above and the reaction was buffer-exchanged using G25 Sephadex beads with 7 M urea, 40 mM EDTA and then purified using a 10% denaturing PAGE. RNA was excised and eluted from the gel into 300 µL of 300 mM KCl and precipitated by adding 700 µl of cold 100% ethanol.

5′ - labeled RNA was prepared in a reaction buffer (70 mM Tris-HCl, pH 7.6, 10 mM $MgCl_2$, 5 mM DTT) using 3 pmol of 5′-dephosphorylated RNA, 0.9 µM [γ-^{32}P] ATP (Perkin Elmer), and 20 units of T4 polynucleotide kinase (New England Biolabs). The reaction was incubated at 37 °C for 1 hour prior to buffer exchange using G25 Sephadex beads with 7 M urea, 40 mM EDTA and subsequently purified by 10% denaturing PAGE.

Purified 5′ radiolabeled RNA was divided into multiple reactions containing 1X binding buffer and varying concentrations of amino-*t*SS or no ligand (as a control). Reactions were then refolded at 60 °C for 5 mins and subsequently cooled at room temperature for 1 hr. Next, T1 nuclease was added to each reaction (0.0007 U; Thermo Fisher Scientific) and samples were incubated at room temperature for 1 min and quenched with 7 M urea and 40 mm EDTA. The G-sequencing ladder was prepared using the same labeled RNA, in 250 mM sodium citrate, pH 7 and 0.013 U of RNase T1, incubated for 5 mins at 55 °C and quenched with 7 M urea and 40 mM EDTA. The partial hydrolysis ladder (OH) was prepared in 50 mM $NaHCO_3$, 1 mM EDTA, pH 10 and incubated at 90 °C for 5 mins and quenched in 7 M urea and 40 mM EDTA. RNA was added to an equal

volume of phenol:chloroform:isoamyl alcohol (25:24:1) and mixed. Samples were centrifuged for 3 mins at 8,000 RPM and the aqueous phase was collected and transferred to a new tube. A second extraction with chloroform:isoamyl alcohol was repeated and transferred to a new tube. Samples were directly loaded onto a 12% denaturing PAGE gel. A phosphor imaging screen was exposed for a minimum of 24 hrs prior to imaging on a GE Typhoon 9410.

3.6 Acknowledgements

We would like to thank the staff at the Genomic High Throughput Facility (GHTF) Shared Resource of the Cancer Center Support Grant (P30CA-062203) at the University of California, Irvine, and NIH shared instrumentation grants 1S10OD010794-01, and 1S10OD021718-01. NASA Exobiology [80NSSC21K0488 to A.L.]; W.M. Keck Foundation to A.L.; NASA ICAR [80NSSC21K0596 to A.L.]; NSF CBET Biophotonics [1804220 to A.L.]

3.7 Supplemental Data

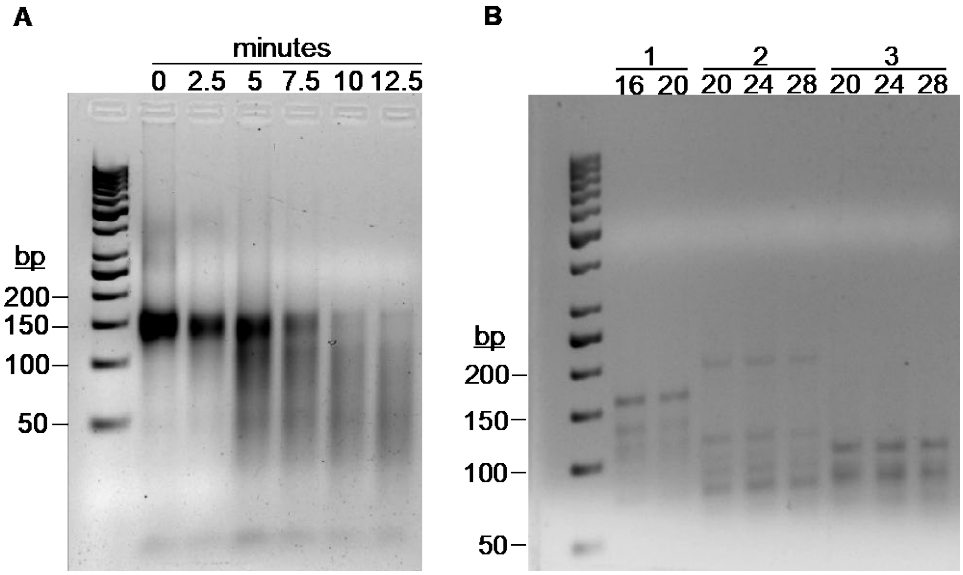
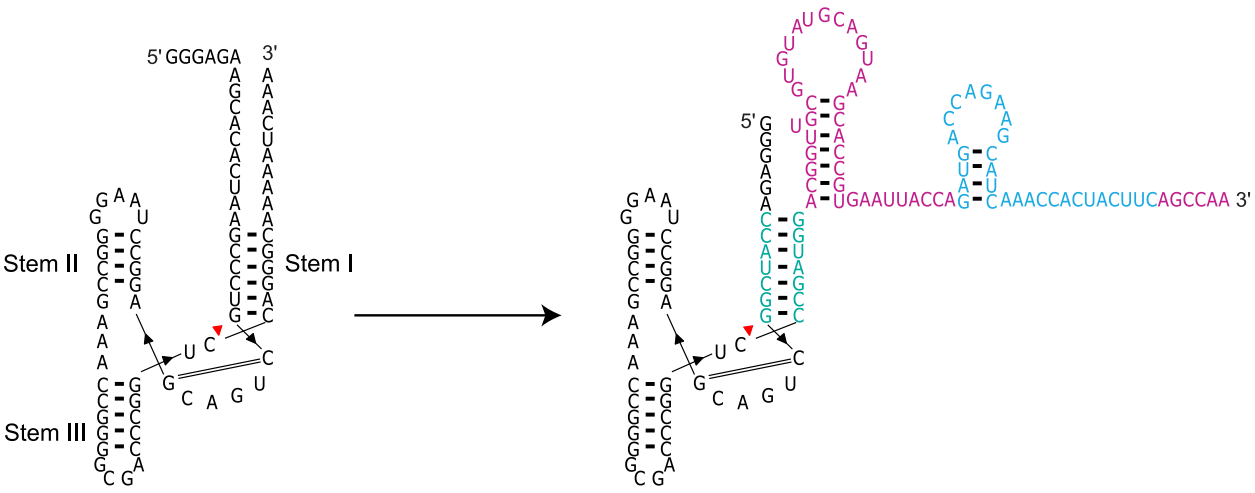


Figure S3.1. Identifying optimal timepoints for alkaline hydrolysis and PCR cycle amplification for the minimized RNA pool. A. heterogenous pool was hydrolyzed for 2, 5, 7.5, 10, and 12.5 minutes in 50 mM $NaHCO_3$ and 1 mM EDTA at 72°C, pH 10. Aliquots of each fraction were fractionated using a 3% agarose gel electrophoresis and the 10-minute timepoint was chosen for the demarcation of the functional domains, because this sample RNA had the broadest distribution of lengths. **B.** Three different aptamer pools were amplified by PCR after each aptamer pool was minimized to determine the optimal amplification cycle prior to submission for sequencing.



hhr1-SewB1_01643s-1 hhr1 + Were-1 (25–103) construct

Figure S3.2. Design of the HHr–Were-1 (25–103) construct. A type I hammerhead ribozyme (hhr-SewB1)[56] was used to derive a sequence that encoded the first 7 nucleotides of the Were-1 (25–103) construct in Stem I to allow efficient transcription of the a-*t*SS aptamer construct. Nucleotides in green represent the beginning of the aptamer construct. The ribozyme cleavage site is indicated by the red arrowhead.

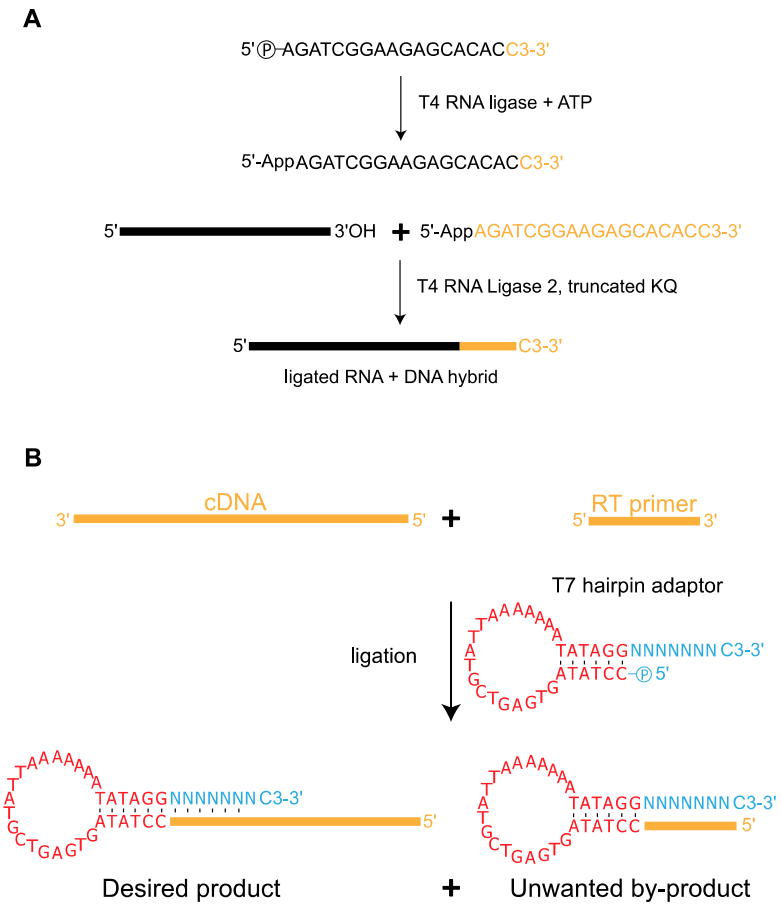


Figure S3.3. Addition of adapters for minimized functional RNAs. A. A 5′ phosphorylated primer with a 3′ carbon spacer was used with T4 RNA ligase to generate an adenylylated oligo intermediate which could be attached to the RNA captured from the *in vitro* selection. This newly added adaptor served as the new constant region for RT. **B.** A 5′ phosphorylated T7 hairpin adaptor with a 3′ carbon spacer was added onto the 3′ end of the generated cDNA to act as an adaptor for subsequent sequencing. Any remaining RT primer could also be ligated to the hairpin adaptor and generate the unwanted-by-product; therefore, washes were performed on streptavidin magnetic beads prior to the ligation to remove excess RT primer. The hairpin adaptor was synthesized with 3′ carbon spacer to prevent self-ligation.

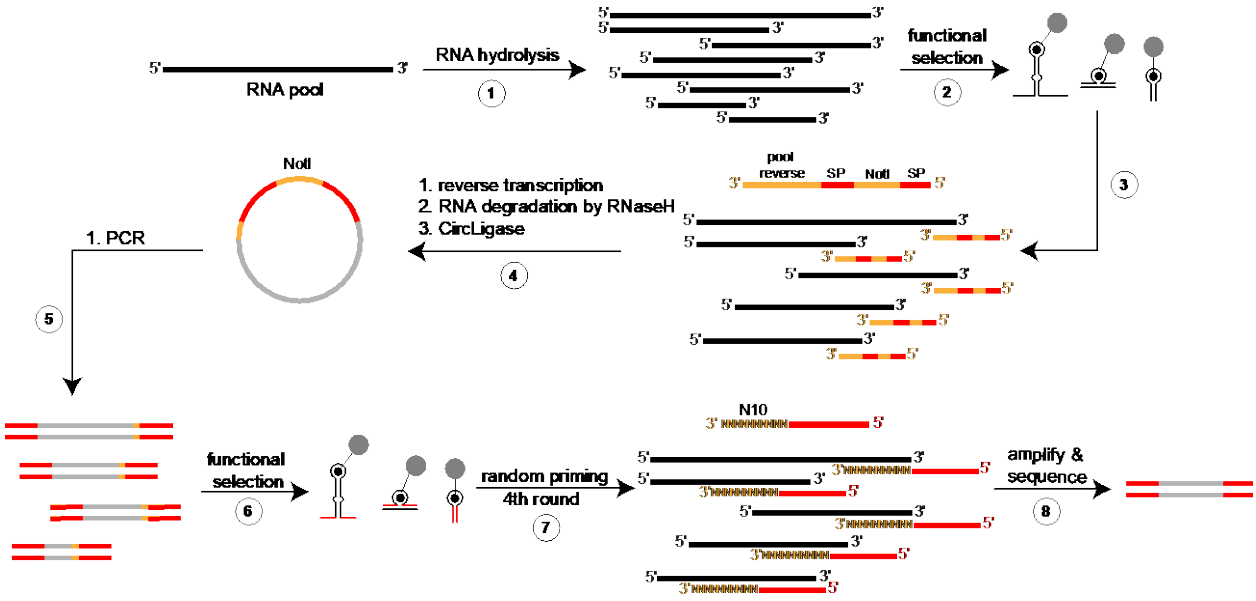


Figure S3.4. High-throughput Identification of minimized RNA domains using CircLigase. RNA (black) is first hydrolyzed to vary the length of the RNA population (**1**). A functional selection was performed to isolate binders of interest and then a reverse primer from the heterogenous pool (yellow) served as a constant region for RT with the addition of both sequencing primers (SP) (**2 & 3**). After, CircLigase was used to circularize the cDNA (gray) and then amplify sequences into duplexed DNA (**4 & 5**). Sequences were later transcribed and used for additional rounds of selection (**6**). After, random priming with the Illumina reverse sequencing primer was used and an additional round of selection was performed prior to sequencing for analysis of minimized functional RNAs (**7 & 8**).

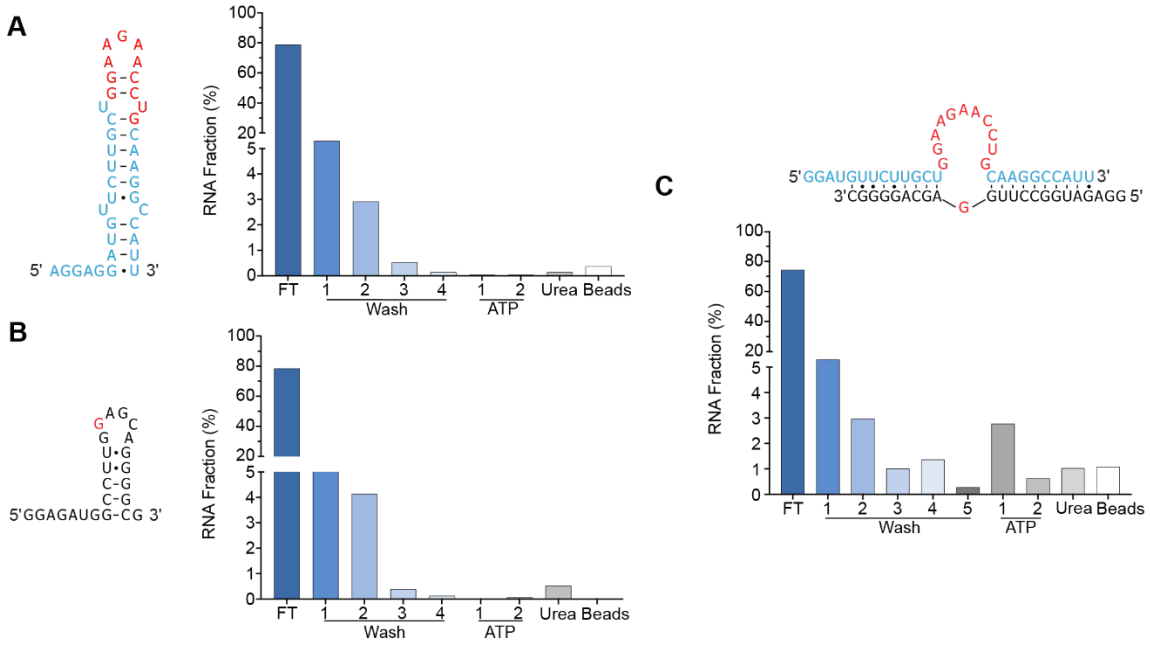


Figure S3.5. Column binding of the split ATP heterodimer. A. Strand of the split aptamer from the selection containing the adenosine binding loop with consensus sequence displayed in red. Nucleotides in blue represent sequences from the top strand of the heterodimer. **B**. Strand of the split aptamer from selection with opposing G in red. Nucleotides in black represent sequences from the bottom strand of the heterodimer. Graph of column binding fraction for this RNA showed no binding. **C**. Proposed structure of the split heterodimeric ATP aptamer. Graph of column binding fractions show ~3 % of binding to ATP.

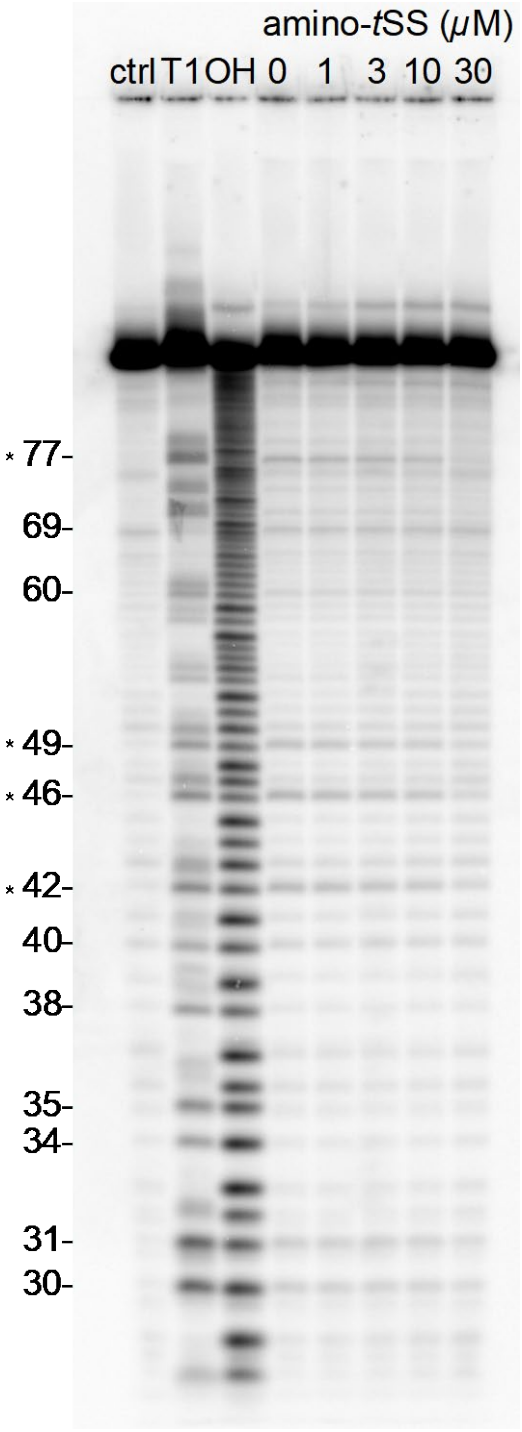


Figure S3.6. RNase T1 probing of the minimized Were-1 aptamer.

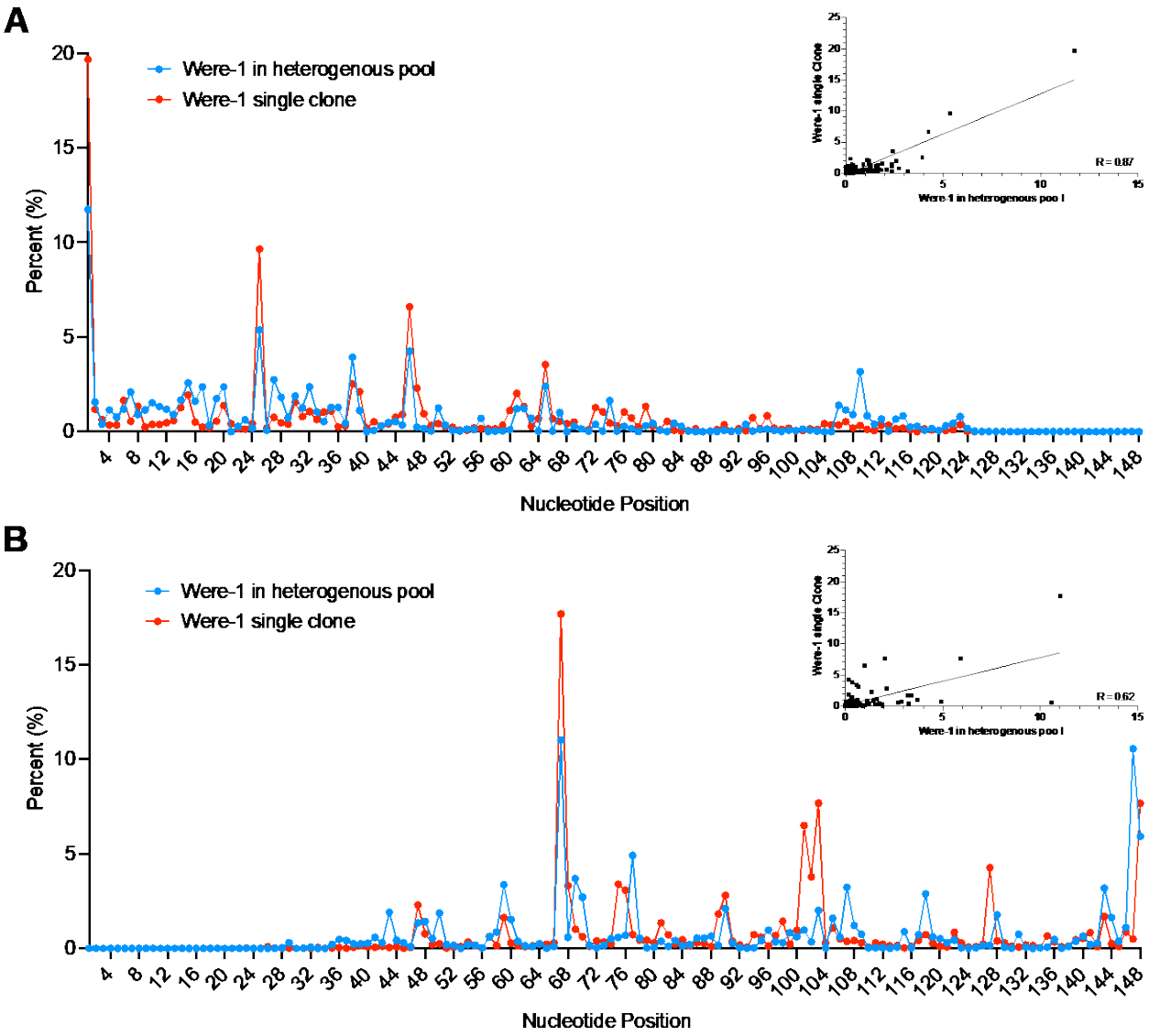


Figure S3.7. Positional frequency of the minimized Were-1 riboswitch. (**A**) 5′ positional frequency of Were-1. (**B**) 3′ positional frequency of Were-1. Data are plotted to compare positional changes in abundance within the 5′-App minimization platform either as a single construct or within a heterogenous pool. Pearson correlation coefficient (R) = 0.87 and 0.61 for 5′ and 3′ position, respectively.

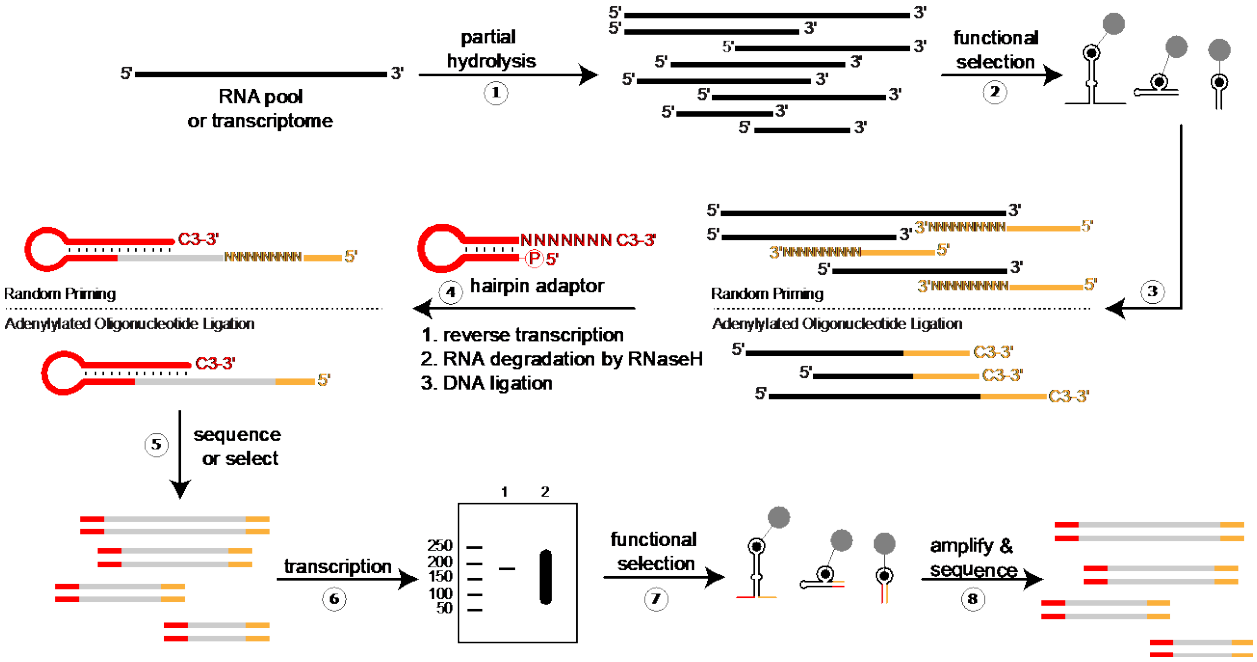


Figure S3.8. High-throughput Identification of minimized RNA domains for multiple rounds of selection. RNA (black) is first hydrolyzed to vary the length of the RNA population (**1**). A functional selection was performed to isolate binders of interest and then a 3′ DNA adaptor (yellow) was added to the RNA to serve as a constant region for RT (**2 & 3**). After, the DNA hairpin adaptor (red) was ligated to the 3′ end of the newly generated cDNA (gray) to serve as adaptor for PCR (**4**). DNA is amplified with the newly attached constant regions and either submit for sequencing or prepared for selection by transcribing the RNA (**5 & 6**). Newly synthetized RNA ranges in size with the additional of the 5′ and 3′ adaptor (red and yellow, respectively) and is further used for a second round of functional selection (**7**). After, sequences obtained from the selection are submitted for Illumina sequencing to analyze minimized RNA domains (**8**).

3.8 References

1. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–10 (1990).

2. Ellington, A. D. & Szostak J W. Selection in vitro of single stranded DNA molecules that fold into specific ligand binding structures. *Nature* **355**, 850–852 (1992).

3. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).

4. Martini, L. et al. In Vitro Selection for Small-Molecule-Triggered Strand Displacement and Riboswitch Activity. *ACS Synthetic Biology* **4**, 1144–1150 (2015).

5. Harbaugh, S. V., Martin, J. A., Weinstein, J., Ingram, G. & Kelley-Loughnane, N. Screening and selection of artificial riboswitches. *Methods* **143**, 77–89 (2018).

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

7. He, X., Guo, L., He, J., Xu, H. & Xie, J. Stepping Library-Based Post-SELEX Strategy Approaching to the Minimized Aptamer in SPR. *Analytical Chemistry* **89**, 6559–6566 (2017).

8. Bittker, J. A., Le, B. V. & Liu, D. R. Nucleic acid evolution and minimization by nonhomologous random recombination. *Nature Biotechnology* **20**, 1024–1029 (2002).

9. Liu, J. M. & Liu, D. R. Discovery of a mRNA mitochondrial localization element in Saccharomyces cerevisiae by nonhomologous random recombination and in vivo selection. *Nucleic Acids Research* **35**, 6750–6761 (2007).

10. Wang, Q. S. & Unrau, P. J. Ribozyme motif structure mapped using random recombination and selection. *RNA* **11**, 404–411 (2005).

11. Arriola, J. T. & Müller, U. F. A combinatorial method to isolate short ribozymes from complex ribozyme libraries. *Nucleic Acids Research* **48**, (2020).

12. Joyce, G. F. In vitro evolution of nucleic acids. *Current Opinions Structural Biology* **4**, 331–336 (1994).

13. Wilson, D. S. & Szostak, J. W. IN VITRO SELECTION OF FUNCTIONAL NUCLEIC ACIDS. *Annual Review Biochemistry* **68**, 611–647 (1999).

13. Shtatland, T. et al. Interactions of Escherichia Coli RNA with Bacteriophage MS2 Coat Protein: Genomic SELEX. *Nucleic Acids Research* vol. **28** (2000).

14. Peng, H., Latifi, B., Müller, S., Lupták, A. & Chen, I. A. Self-cleaving ribozymes: Substrate specificity and synthetic biology applications. *RSC Chemical Biology* **2**, 1370–1383 (2021).

15. Dey, S. K. et al. Repurposing an adenine riboswitch into a fluorogenic imaging and sensing tag. *Nature Chemical Biology* **18**, 180–190 (2022).

16. Mohsen, M. G., Midy, M. K., Balaji, A. & Breaker, R. R. Exploiting natural riboswitches for aptamer engineering and validation. *Nucleic Acids Research* **51**, 966–981 (2023).

17. Porter, E. B., Polaski, J. T., Morck, M. M. & Batey, R. T. Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors. *Nature Chemical Biology* **13**, 295–301 (2017).

18. Rehm, C., Klauser, B., Finke, M. & Hartig, J. S. Engineering Aptazyme Switches for Conditional Gene Expression in Mammalian Cells Utilizing an In Vivo Screening Approach. in RNA Scaffolds: Methods and Protocols (ed. Ponchon, L.) 199–212 (Springer US, New York, NY, 2021). doi:10.1007/978-1-0716-1499-0_14.

19. Yokobayashi, Y. Applications of high-throughput sequencing to analyze and engineer ribozymes. *Methods* **161**, 41–45 (2019).

20. Singerl, B. S., Shtatiandl, T., Brown, D. & Gold1, L. Libraries for genomic SELEX. *Nucleic Acids Research* **25**, 781–786 (1997).

21. Pressman, A. D. et al. Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA. *Journal of the American Chemical Society* **141**, 6213–6223 (2019).

22. Silverman, S. K. Practical, and general synthesis of 5′-adenylated RNA (5′-AppRNA). *RNA* **10**, 731–746 (2004).

23. Wang, Y. & Silverman, S. K. Efficient RNA 5′-adenylation by T4 DNA ligase to facilitate practical applications. *RNA* **12**, 1142–1146 (2006).

24. Chen, Y. R. et al. A cost-effective method for Illumina small RNA-Seq library preparation using T4 RNA ligase 1 adenylated adapters. *Plant Methods* **8**, (2012).

25. Duzdevich, D., Carr, C. E. & Szostak, J. W. Deep sequencing of non-enzymatic RNA primer extension. *Nucleic Acids Research* **48**, E70–E70 (2020).

26. Ritchey, L. E. et al. Structure-seq2: sensitive and accurate genome-wide profiling of RNA structure in vivo. *Nucleic Acids Research* **45**, E135 (2017).

27. Kwok, C. K., Ding, Y., Sherlock, M. E., Assmann, S. M. & Bevilacqua, P. C. A hybridization-based approach for quantitative and low-bias single-stranded DNA ligation. *Analytical Biochemistry* **435**, 181–186 (2013).

28. Vu, M. M. K. et al. Convergent evolution of adenosine aptamers spanning bacterial, human, and random sequences revealed by structure-based bioinformatics and genomic SELEX. *Chemical Biology* **19**, 1247–1254 (2012).

30. Salehi-Ashtiani, K., Lupták, A., Litovchick, A. & Szostak, J. W. A Genomewide Search for Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene. *Science* **313**, 1788–1792 (2006).

31. Abdelsayed, M. M. et al. Multiplex Aptamer Discovery through Apta-Seq and Its Application to ATP Aptamers Derived from Human-Genomic SELEX. *ACS Chemical Biology* **12**, 2149–2156 (2017).

32. Sedlyarova, N. et al. Natural RNA Polymerase Aptamers Regulate Transcription in *E. coli*. *Molecular Cell* **67**, 30-43.e6 (2017).

33. Gold, L. et al. From oligonucleotide shapes to genomic SELEX: Novel biological regulatory loops. *Proceedings of the National Academy of Sciences* **94**, 59–64 (1997).

34. Shtatland, T. et al. Interactions of Escherichia Coli RNA with Bacteriophage MS2 Coat Protein: Genomic SELEX. *Nucleic Acids Research* vol. **28** (2000).

35. Burke, D. H. & Gold, L. RNA aptamers to the adenosine moiety of S-adenosyl methionine: structural inferences from variations on a theme and the reproducibility of SELEX. *Nucleic Acids Research* **25**, 2020–2024 (1997).

36. Burgstaller, P. & Famulok, M. Isolation of RNA Aptamers for Biological Cofactors by In Vitro Selection. *Angewandte Chemie International Edition in English* **33**, 1084–1087 (1994).

37. Gebhardt, K., Shokraei, A., Babaie, E. & Lindqvist, B. H. RNA Aptamers to S-Adenosylhomocysteine: Kinetic Properties, Divalent Cation Dependency, and Comparison with Anti-S-adenosylhomocysteine Antibody. *Biochemistry* **39**, 7255–7265 (2000).

38. Jiang, F., Kumar, R. A., Jones, R. A. & Patel, D. J. Structural basis of RNA folding and recognition in an AMP–RNA aptamer complex. *Nature* **382**, 183–186 (1996).

39. Dieckmann, T., Suzuki, E., Nakamura, G. K. & Feigon, J. Solution structure of an ATP-binding RNA aptamer reveals a novel fold. *RNA* **2**, 628–640 (1996).

40. Sassanfar, M. & Szostak, J. W. An RNA motif that binds ATP. *Nature* **364**, 550–553 (1993).

41. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

42. Rotstan, K. A. et al. Regulation of mRNA translation by a photoriboswitch. *Elife* **9**, e51737 (2020).

43. Peng, Y., Soper, T. J. & Woodson, S. A. Rnase footprinting of protein binding sites on an mRNA target of small RNAs. *Methods in Molecular Biology* **905**, 213–224 (2012).

44. Kieft, J. S., Costantino, D. A., Filbin, M. E., Hammond, J. & Pfingsten, J. S. Structural Methods for Studying IRES Function. in *Methods in Enzymology* vol. **430** 333–371 (Academic Press Inc., 2007).

45. Chizzolini, F. et al. Large Phenotypic Enhancement of Structured Random RNA Pools. *Journal of the American Chemical Society* **142**, 1941–1951 (2020).

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

47. Hofmann, H. P., Limmer S., Hornung V. & Sprinzl, M. Ni2+-binding RNA motifs with an asymmetric purine-rich internal loop and a G-A base pair. *RNA* **3**, 1289–1300 (1998).

48. Carothers, J. M., Oestreich, S. C., Davis, J. H. & Szostak, J. W. Informational Complexity and Functional Activity of RNA Structures. *Journal of the American Chemical Society* **126**, 5130–5137 (2004).

49. Wells, S. E., Hughes, J. M. X., Haller Igel, A. & Ares, M. Use of dimethyl sulfate to probe RNA structure in vivo. in *Methods in Enzymology* vol. **318** 479–493 (Academic Press, 2000).

50. Spitale, R. C. & Incarnato, D. Probing the dynamic RNA structurome and its functions. *Nature Reviews Genetics* **24**, 178–196 (2023).

51. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA Structure Analysis at Single Nucleotide Resolution by Selective 2‘-Hydroxyl Acylation and Primer Extension (SHAPE). *Journal of the American Chemical Society* **127**, 4223–4231 (2005).

52. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

53. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).

54. Alam, K. K., Chang, J. L. & Burke, D. H. FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Molecular Therapeutics Nucleic Acids* **4**, e230 (2015).

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

56. Jimenez, R. M., Delwart, E. & Lupták, A. Structure-based search reveals hammerhead ribozymes in the human microbiome. *Journal of Biological Chemistry* **286**, 7737–7743 (2011).

Chapter 4: A modular platform for bioluminescent RNA tracking

Publication note:

This paper was originally published on the preprint server *bioRxiv* (6 November 2023).

Ng, K.K.*, **Cole, K.H.***, Halbers, L.P.*, Chan, C.E.T., Fuller, E.B., Callicoatte, C., Metcalfe, M., Chen, C.C., Barhoosh, A.A., Reid-McLaughlin, E., Kent, A.D., Steward, O., Lupták, A., Prescher, J.A. A modular platform for bioluminescent RNA tracking. *bioRxiv* (2023). https://doi.org/10.1101/2022.07.02.498144

*These authors contributed equally

Contribution statement:

Conceptualization: A.L., J.A.P., O.S., C.C.C. Methodology: K.K.N., **K.H.C.**, L.P.H., A.L., J.A.P. Investigation: K.K.N., **K.H.C.**, L.P.H., C.E.T.C., E.B.F., C.C., M.M., C.C.C., A.A.B., E.R.-M., A.D.K. Visualization: K.K.N., L.P.H., **K.H.C.** Writing – original draft: K.K.N., L.P.H., **K.H.C.**, A.L., J.A.P. Writing – review and editing: all authors. Funding acquisition: A.L., J.A.P., O.S.

4.1 Abstract

A complete understanding of RNA biology requires methods for tracking transcripts *in vivo.* Common strategies rely on fluorogenic probes that are limited in sensitivity, dynamic range, and depth of interrogation, owing to their need for excitation light and tissue autofluorescence. To overcome these challenges, we developed a bioluminescent platform for serial imaging of RNAs. Small RNA tags were engineered to recruit light-emitting luciferase fragments (termed RNA lanterns) upon transcription. Robust photon production was observed for RNA targets both in cells and in live animals. Importantly, only a single copy of the tag was necessary for sensitive detection, in sharp contrast to fluorescent platforms requiring multiple repeats. Overall, this work provides a foundational platform for visualizing RNA dynamics from the micro to the macro scale.

4.2 Introduction

RNA dynamics play pivotal roles in a multitude of cellular processes[1]. While we have a deep, molecular-level understanding of many facets of RNA biology *in vitro*, the picture in physiologically authentic environments—live animals—remains incomplete. This is due, in part, to a lack of methods for noninvasive tracking of RNAs *in vivo*. Conventional approaches rely on RNA tags coupled with fluorescent probes[2-4]. Such platforms require excitation light, which can be difficult to deliver in whole organisms without invasive procedures, excision of tissues, or delivery of fluorogenic dyes[5-7]. Furthermore, external light can induce autofluorescence, precluding sensitive detection of low abundance targets. Short imaging times are also necessary to avoid light-induced damage. Consequently, tracing the lifecycle of key RNAs in real time, in live mammals, has been elusive.

We reasoned that a potentially more suitable platform for RNA imaging in live animals could be achieved with bioluminescence. This modality relies on photon production from luciferase enzymes and luciferin small molecules. Since no excitation light is required,

bioluminescence can provide superior signal-to-noise ratios *in vivo* for visualization of low-copy transcripts. Additionally, serial imaging is possible without concern for phototoxicity or tissue damage. The development of luciferases with higher photon outputs and improved thermal stability (e.g., NanoLuc and related variants) for imaging *in vivo* has enabled facile visualization of cells, biomolecules, and other features both on the micro and macro level[8-11]. Recently, a split variant of this luciferase was applied to RNA targets, setting the stage for precise detection of cellular transcripts[12].

Here we report a general method that leverages advances in bioluminescence technology for multi-scale RNA detection. The approach features split fragments of NanoLuc (NanoBiT) fused to MS2 and PP7 bacteriophage coat proteins (MCP and PCP), fusions that we have termed RNA lanterns. MCP and PCP bind distinct RNA aptamers (MS2 and PP7, respectively) that can be appended to transcripts of interest[13,14]. Upon transcription, MCP and PCP bind the MS2-PP7 containing RNA bait, bringing the luciferase fragments into proximity to assemble a functional, light-emitting enzyme. We extensively optimized both the lanterns and RNA bait to maximize signal turn-on and minimize the size of the protein-RNA complex. Notably, a single rigidified RNA was sufficient for sensitive imaging both in cells and *in vivo*, rendering much larger RNA tags unnecessary. Additionally, the RNA bait is modular and can be used in conjunction with other split luciferases and for multi-scale imaging. The tools reported here are thus immediately useful to studies of RNA dynamics.

4.3 Results

General strategy for visualizing RNAs with bioluminescent light

The overall strategy was dependent on three key steps: RNA bait formation, MCP/PCP binding, and NanoBiT complementation. We took inspiration from MCP- and PCP-based reporters comprising split fluorescent proteins to visualize transcripts[15]. We envisioned that the RNA-binding proteins could be fused to the NanoBiT system[16], which has been used extensively to

examine protein-protein interactions and other biomolecular networks[17-19]. We appended each part of NanoBiT (the short peptide—SmBiT—and the larger protein fragment—LgBiT, **Fig. 4.1A**) to MCP and PCP, respectively. The resulting fusions (termed RNA lanterns) were hypothesized to bind their cognate aptamers on a single transcript, inducing NanoBiT complementation and generating photons in the presence of luciferin. The requisite components had not been used together prior to this work, necessitating optimization of each step.

As a starting point, we designed a bicistronic construct with an internal ribosome entry site (IRES) to mediate co-expression of the RNA lantern components (**Fig. 4.1A**). MCP was fused to SmBiT and PCP was fused to the larger protein fragment (LgBiT), building on a previously published split fluorescent protein platform, in which MCP and PCP were fused to the N- and C-terminal fragments of the Venus fluorescent protein, respectively[15]. MCP was further tagged with a nuclear localization signal (NLS) to reduce background complementation. This same strategy was used previously for the split Venus system[15] to separate the lantern fragments in the absence of the RNA bait. Upon expression, transcripts fused to the RNA bait could transport an MCP-fusion from the nucleus into the cytoplasm (or capture *de novo* translated MCP in the cytoplasm). Eventual binding of the PCP fusion would co-localize both halves of the RNA lantern on the target transcript, enabling NanoBiT complementation and thus light production. Additionally, the NLS reduces the likelihood of non-specific SmBiT/LgBiT complementation, a problem encountered in previous studies that reduces sensitivity[12].

We modeled the RNA lanterns using ChimeraX[20] to assess the design of the fusions (**Fig. 4.1B**). No obvious steric clashes or unfavorable orientations with a 5′-MS2-PP7-3′ RNA bait were observed, suggesting that lantern assembly was possible. NanoBiT complementation appeared feasible even with juxtaposed MS2 and PP7 aptamers. Small, compact baits are attractive tags to avoid disrupting the structures or functions of target RNAs. Encouraged by the modeling results, we moved forward with the lantern design and anticipated that additional engineering of

the lantern components (linkers and orientation) would be necessary to maximize luciferase complementation and signal output.

Biochemical optimization of the RNA tag and lanterns

The RNA detection platform relies on efficient NanoBiT formation, which can be tuned based on the SmBiT peptide sequence[16]. We examined two SmBiT peptides (SmBiT^{high}, K_D = 180 nM; SmBiT^{low}, K_D = 190 μM) to determine which would provide the best dynamic range: minimal signal in the absence of RNA bait and robust signal in its presence (**fig. S4.1A**). The designer probes were first expressed using an *in vitro* transcription/translation system (IVTT) featuring transcription by T7 RNA polymerase and translation in rabbit reticulocyte lysate (RRL). Background signal was determined in the absence of RNA bait. Full complementation of translated NanoBiT was achieved by adding exogenous SmBiT^{ultra} (K_D = 0.7 nM) or recombinant LgBiT at saturating concentrations, establishing the maximum potential signal (**fig. S4.1B**). All protein designs exhibited robust signal enhancement when SmBiT^{ultra} or recombinant LgBiT was added, but high background luminescence was observed in samples with SmBiT^{high}. The largest signal enhancements were achieved with SmBiT^{low}, due to the reduced background complementation observed with this peptide (**fig. S4.1C**). Additional repeats of SmBiT did not yield greater light output, possibly due to insufficient spacing between sequential peptides to support LgBiT binding. We therefore moved forward with MCP-SmBiT^{low}/PCP-LgBiT, the lantern combination that provided highest dynamic range.

We first asked whether the MCP-SmBiT^{low}/PCP-LgBiT lantern could detect the previously reported MS2-PP7 bait, comprising an unstructured 19-nucleotide strand joining the two aptamers (**Fig. 4.1C**)[15]. Upon RNA bait transcription, an approximate 17-fold signal increase was observed (**Fig. 4.1D**). We attributed the relatively low signal enhancement to the RNA bait structure. Secondary structure modeling with Vienna RNAFold[21] and Forna[22], revealed that the 19-nucleotide linker was likely flexible with the potential to sample non-productive conformations.

While an unstructured RNA bait could aid lantern binding, we surmised that extensive flexibility could potentially hinder NanoBiT complementation and thus diminish signal-to-noise ratios.

We hypothesized that increasing the rigidity of the RNA bait would favor lantern assembly and photon production. We thus redesigned the RNA bait to fix the spacing and orientation of the MS2 and PP7 sequences. We locked the aptamers into desired conformations, as part of a four-way junction with an established Ni-NTA-binding aptamer (**Fig. 4.1E and fig. S4.2A**)[23]. The RNA baits were further engineered to contain varying numbers of nucleotides between the MS2 and PP7 aptamer domains (M-X-P; where X is the number of nucleotides in the spacer). This panel enabled us to not only sample the spacing between the aptamer domains (~3 nm, according to secondary structure predictions) but also the helical phase of the aptamers respective to one another (up to ~360º). When the RNAs were present with the lantern fragments, we observed that a relatively short linker (M-3-P) could yield significantly higher luminescence, with up to a 330-fold increase over a no-RNA control (**Fig. 4.1E**). In the case of M-11-P, signal was abolished. We attributed this result to the phase angle of approximately 180º (one half of a helical turn) from M-3-P, preventing luciferase complementation. Importantly, signal was restored with the MS2 and PP7 aptamers brought back into phase (~360º) and a full helical turn apart (M-21-P). Signal still decreased because the overall length increased and likely prevented effective assembly of the luciferase parts. Photon production was also highly dependent on the concentrations of the RNA bait and the RNA lantern (**fig. S4.2B**) and RNA bait integrity. In experiments using singular MS2 and PP7 aptamers, or M-3-P^{mut} that does not bind the PCP protein (**fig. S4.3**) [24], no appreciable signal over background was observed (**Fig. 4.1E and fig. S4.3C**).

Due to the improved luminescence achieved through modulating the distance and phase angle, we also examined the length and relative orientation of the individual MS2 and PP7 aptamers. We took the optimal M-3-P RNA bait and added or removed base-pairs (bp) to either the MS2 or PP7 stem (**Fig. 4.1F and fig. S4.4A**). No significant improvements in photon output were observed among the suite of RNA baits, suggesting that the length and orientation of the

two aptamers were already optimal. We further confirmed that the lack of improvement was not a result of template DNA ratios (**fig. S4.4B**). Although the orientation of the aptamers had minimal effect on complementation, we found that the positioning of the aptamers was critical. When the placement of MS2 and PP7 were inverted, luciferase complementation was not observed (**fig. S4.5**).

We further examined the effect of protein linker length on RNA lantern assembly. Constructs were designed with additional glycine-serine (G_4S) units inserted between the RNA-binding proteins and NanoBiT segments (**fig. S4.6A**). These constructs were then tested with the M-3-P, M-7-P, and M-11-P RNA baits (**fig. S4.6B**). The largest light outputs were achieved with the M-3-P RNA bait, for all the lanterns tested. Interestingly, only minimal signal enhancements (~1.5 fold) were observed with the additional G_4S units compared to the original lanterns both in vitro and (**fig. S4.6, C, D and E**).

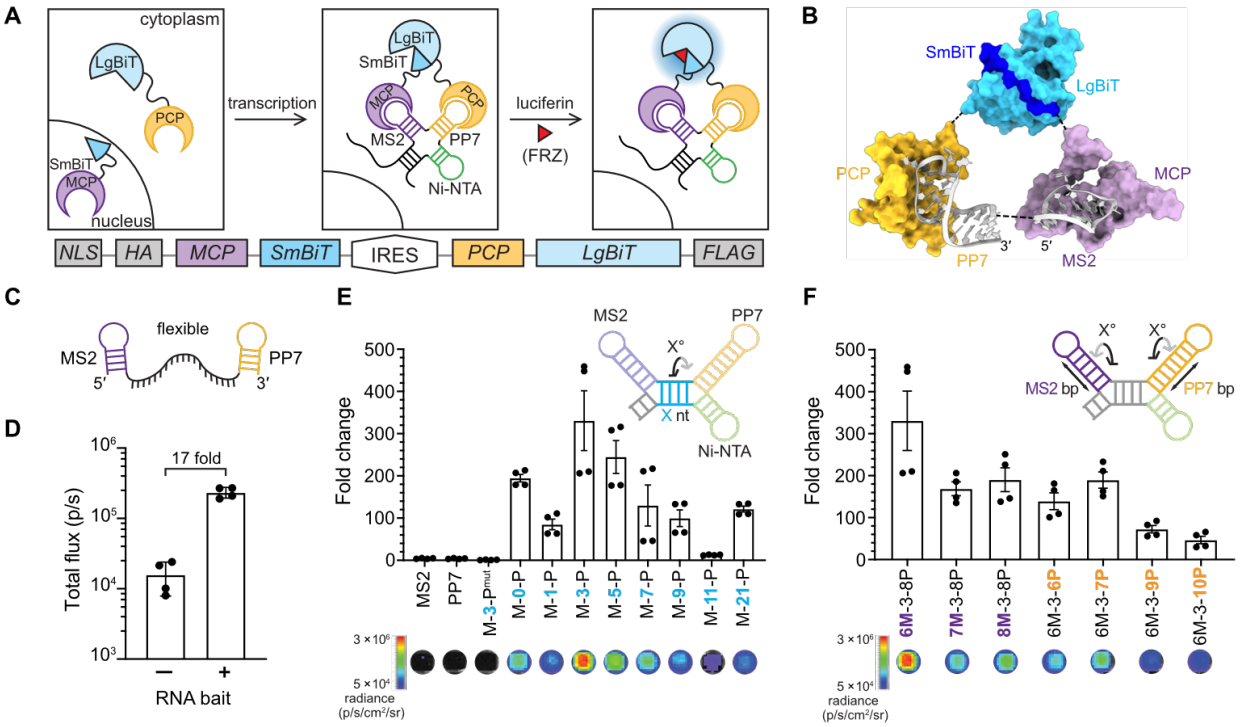


Figure 4.1. An optimized platform for tracking RNA dynamics. (**A**) Overall strategy to visualize transcripts using RNA lanterns. The lanterns comprise fusions of MS2 coat protein (MCP) and PP7 coat protein (PCP) with NanoBiT fragments (SmBiT and LgBiT, respectively). Transcription of bait RNA (comprising MS2 and PP7 aptamers) drives NanoBiT heterodimerization. In the presence of a luciferin substrate (furimazine, FRZ), light is produced (blue glow). The bicistronic construct encoding the RNA lantern is shown below the scheme. MCP and PCP were fused with HA and FLAG tags, respectively, for expression analyses. (**B**) Modeling of the RNA lantern complex. Crystal structures of MS2 (purple, 1ZDI)[32], PP7 (orange, 2QUX)[33] and NanoLuc (5IBO), highlighting LgBiT (dark blue) and SmBiT (cyan), were modeled in ChimeraX[20]. The MCP-SmBiT/PCP-LgBiT complex was modeled by aligning the corresponding N- and C- termini of each protein and aligning the 5' and 3' ends of the aptamers. (**C**) Predicted secondary structure of flexible RNA bait, as calculated by RNAfold[21]. (**D**) Initial tests of RNA lantern with a flexible RNA bait via an *in vitro* transcription and translation (IVTT) assay. The total flux observed in the absence or presence of RNA bait is plotted. (**E**) Engineered rigid RNA baits provided robust photon output in IVTT assay. RNA baits containing varying spacers between the MS2 and PP7 aptamers were constructed, with spacer length (X nt, blue, denoted below each bar graph), modulating the distance and helical phase angle between the aptamers. Fold change in signal over RNA lantern alone is plotted. MS2, PP7, and M-3-P^{mut} denote baits comprising isolated aptamers or a mutated PP7 aptamer (fig. S4.3, A and B), respectively, none of which were expected to result in RNA lantern assembly. Representative luminescence images are shown below the graph. (**F**) Rigid RNA baits comprising various MS2 and PP7 stem lengths (top right, purple and orange for the MS2 and PP7 aptamers, respectively). Luminescence readouts were acquired following IVTT.

The RNA sensing capabilities of the designer bait and lanterns were biochemically validated *in vitro* (**Fig. 4.2A**). Cells were engineered to stably express the lantern components, and lysates were then titrated with purified bait. As shown in Figures 4.2B-C, an RNA-dependent "hook effect" was observed (**Fig. 4.2, B and C**). Lantern complex formation was limited at low RNA concentrations, as expected. More lantern complementation was observed with increasing RNA bait, but the amount falls off at high RNA concentrations. This latter outcome is explained by MCP and PCP proteins binding distinct RNA bait transcripts, so that the luciferase bits are not brought together to assemble active lantern complexes.

The unique design of the RNA bait enabled direct interrogation of lantern binding and assembly. The Ni-NTA aptamer was used to retrieve the RNA bait-RNA lantern complex on resin: as shown in Figure 4.2D, the assembled NanoBiT enzyme was exclusively associated with the RNA bait (**Fig. 4.2D**). Pulldowns of active complexes using the HA and FLAG tags built into the RNA lantern components further confirmed the RNA-dependent assembly of the active luciferase (**Fig. 4.2, E and F**). We anticipate that the retrieval of targeted transcripts and associated biomolecules post-imaging will facilitate the often-critical follow-up analyses of RNA interactions.

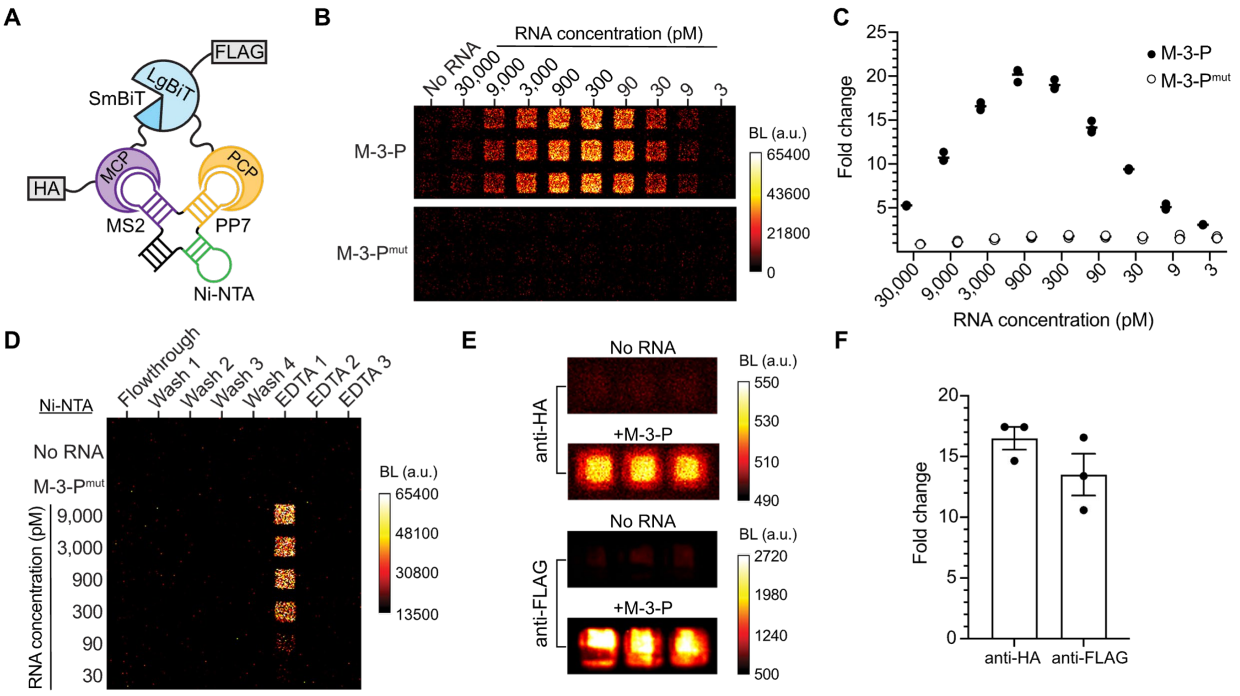


Figure 4.2. Biochemical validation of RNA lanterns. (**A**) The designer RNA bait comprises a 4-way-junction of MS2, PP7, and Ni-NTA aptamers. This unit assembles the RNA lantern components (MCP-SmBiT and PCP-LgBiT, fused to HA and FLAG epitopes, respectively). (**B**) Bioluminescent output from RNA bait (M-3-P) or inactive mutant (M-3-P^{mut}) combined with lysate from cells stably expressing the RNA lantern. (**C**) Fold change in bioluminescent signal over no-RNA controls. Each bar represents the median of *n* = 3 replicates with dots showing data from individual replicates. (**D**) Affinity purification of the RNA lantern complex using the Ni-NTA aptamer. Various concentrations of RNA bait were used, and captured complexes were eluted using metal chelator (EDTA). (**E**) Bioluminescent output of lantern complexes captured using anti-HA or anti-FLAG conjugate-agarose beads. (**F**) Fold change in signal from lantern complexes retrieved in the presence of M-3-P bait versus no RNA from (E). Error bars represent the standard error mean (SEM) of *n* = 3 replicates. BL, bioluminescence. a.u., arbitrary units.

RNA imaging platform is ultrasensitive and modular

Previous applications of MS2-PP7 for fluorescence imaging have required at least 12 copies of the aptamers to achieve adequate signal-to noise ratios ($MP_{12X,}$ **Fig. 4.3A**)[2,15,25,26]. In many cases, these constructs comprise an RNA bait that is at least 780 nucleotides long, a size that may impede the natural localization and behavior of the RNA under study[27-29]. Our M-3-P bait (69 nucleotides) requires only a single copy of each aptamer to produce detectable signal (**Fig. 4.3B**). When compared head-to-head with bait aptamers at equimolar concentration (M-3-P versus one twelfth RNA concentration of MP_{12X}), the structured bait produced 10-times more signal. Even when MP_{12X} was introduced at the same RNA concentration and the aptamers were 12-times more concentrated, the structured M-3-P bait outperformed the flexible design (**Fig. 4.3B**). These results suggest that the compact M-3-P is a more effective RNA bait, providing higher signal compared to the multimerized and extensively used MS2-PP7 tag.

The designer M-3-P bait can also facilitate complementation of other split reporters. When MCP and PCP were fused to split fragments of firefly luciferase (Fluc; **Fig. 4.3C**), the resulting RNA lantern yielded a large increase in signal over background (**Fig. 4.3, D and E**). The photon flux of the assembled Fluc is lower than NanoBiT, but the undetectable background in absence of RNA may be desirable in specific applications. M-3-P also enabled the assembly of bioluminescence resonance energy transfer (BRET) variants of NanoLuc. These probes comprise luciferase fluorescent protein fusions that produce red-shifted light upon complementation. Such wavelengths can provide more sensitive readouts in vivo because they are less absorbed and scattered by tissue. Two BRET-based RNA lanterns were developed using yellow (VenusΔC12)[9] and red (mScarlet-I)[30] fluorescent proteins fused to the C-terminus of PP7 and the N-terminus of LgBiT (**Fig. 4.3F**). Assembly of the yellow and red RNA lanterns was assessed by measuring photon production in the presence of rigidified RNA baits. As shown in Figure 4.3G, M-5-P provided the highest degree of signal turn-on (**Fig. 4.3G**). The larger bait provides more space between the bulkier lantern fragments in these cases, likely enabling more effective

complementation. The BRET-bases probes also exhibited different colors of light output (**Fig. 4.3H**). BRET efficiency was higher for the yellow than the red lantern due to the larger spectral overlap between NanoBiT and VenusΔC12 compared to mScarlet-I[9,31]. Collectively, these demonstrate that structured RNA baits can productively assemble diverse split proteins.

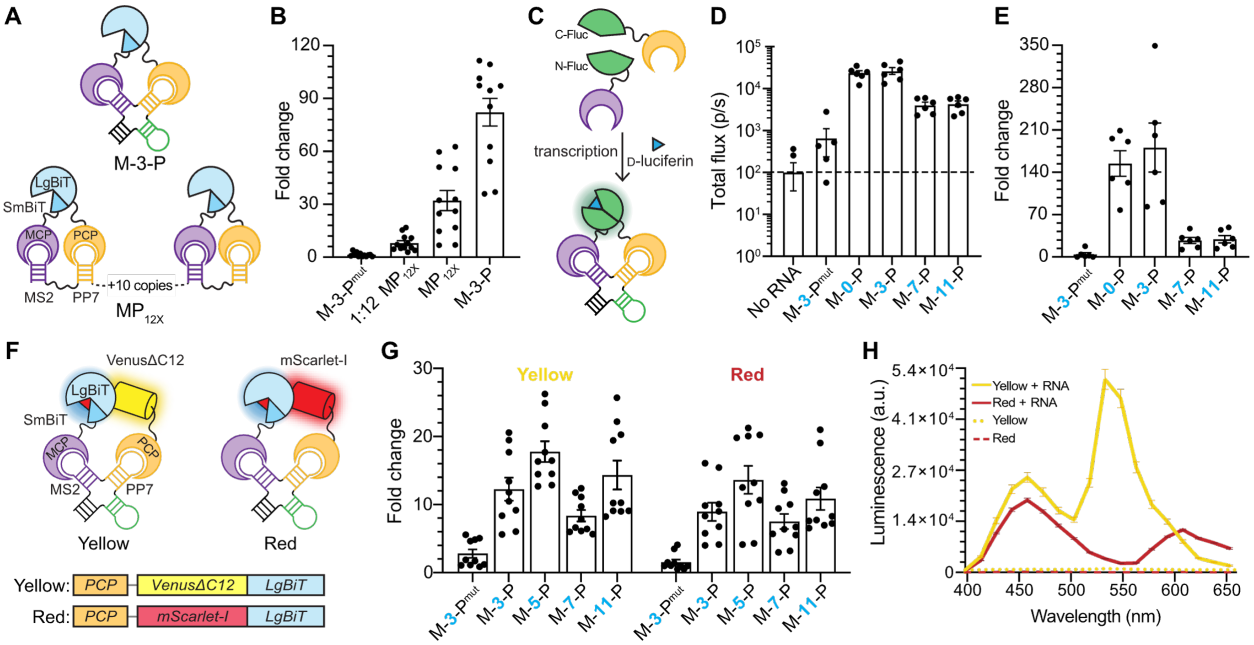


Figure 4.3. Robustness and modularity of structured RNA baits. (**A**) Comparison of M-3-P to a 12-copy unstructured RNA bait. Top: Schematic of the structured M-3-P RNA bait (single copy). Bottom: Schematic of a flexible RNA bait comprising 12 copies of the MS2 and PP7 aptamers (MP_{12X}). (**B**) MP_{12x} was evaluated against M-3-P at equimolar RNA concentrations (1 nM of template DNA) or at 1/12 the concentration (83 pM of template DNA). Fold change over no-RNA bait samples is plotted. (**C**) RNA bait assembles other lanterns. Schematic of MCP-PCP probes fused to split firefly luciferase (Fluc). RNA transcription followed by D-luciferin treatment enables photon production and visualization. (**D**) The Fluc lantern was assessed with a panel of RNA baits. Total light output for *n* = 6 replicates is shown. Only 2 replicates are shown for the no-RNA bait sample; the others were below the limit of detection represented by the electronic noise of the EMCCD camera (dashed line). (**E**) Graph of fold change in signal over no-RNA bait samples from (D). (**F**) Schematic of BRET-based RNA lanterns. (**G**) Luminescence fold change measurements for yellow and red RNA lanterns across a panel of RNA baits. Fold change calculated over no-RNA controls. (**H**) Emission spectra for BRET-based RNA lanterns. Error bars represent the standard error of the mean (SEM) for *n* = 12 replicates (**Fig. 4.3B**), *n* = 6 replicates (**Figs. 4.3D** and **4.3E**), *n* = 10 replicates (**Fig. 4.3G**), and *n* = 3 replicates (**Fig. 4.3H**). a.u., arbitrary units.

Imaging RNAs from the micro-to-macro scale

The robustness of the bait design enabled facile imaging of RNA *in cellulo* and *in vivo*. As an initial test, we developed a model system using an mRNA encoding green fluorescent protein (GFP) with the M-3-P RNA bait placed in the 3′ untranslated region (UTR) (**Fig. 4.4A**). GFP fluorescence would thus report on bait expression. Cell lines stably expressing RNA lanterns were generated and transfected with constructs encoding M-3-P–tagged *GFP* or *GFP* only. GFP fluorescence was observed in both cases, but bioluminescence (from RNA lantern assembly) was only observed in cells transfected with M-3-P–tagged *GFP* (**Fig. 4.4, B and C, and fig. S4.7**). Similar bioluminescence turn-on was observed using an analogous model transcript (**fig. S4.6, D and E**). Both studies illustrate the utility of the structured bait for RNA lantern assembly and transcript visualization.

Finally, as a proof of concept, we imaged RNAs in subcutaneous models *in vivo*. HEK293T cells were engineered to express RNA lanterns and mRNAs encoding either GFP or BFP with variable 3′ UTRs: M-3-P, M-3-P^{mut}, or no bait (**Fig. 4.4A**). The cells were incubated with luciferin, implanted in mice dorsal flanks, and imaged. Increased luminescence was observed from cells expressing RNA lanterns and *GFP-M-3-P* (right flank) compared to cells lacking the M-3-P bait (left flank (**Fig. 4.4, D and E**). Similar increases in luminescence were observed from cells expressing *BFP-M-3-P*, compared to control transcripts in the presence of the lantern components (**Fig. 4.4, F and G**). These data demonstrate that lanterns can be used for RNA-dependent imaging in tissue, setting the stage for real-time detection of gene expression in live animals.

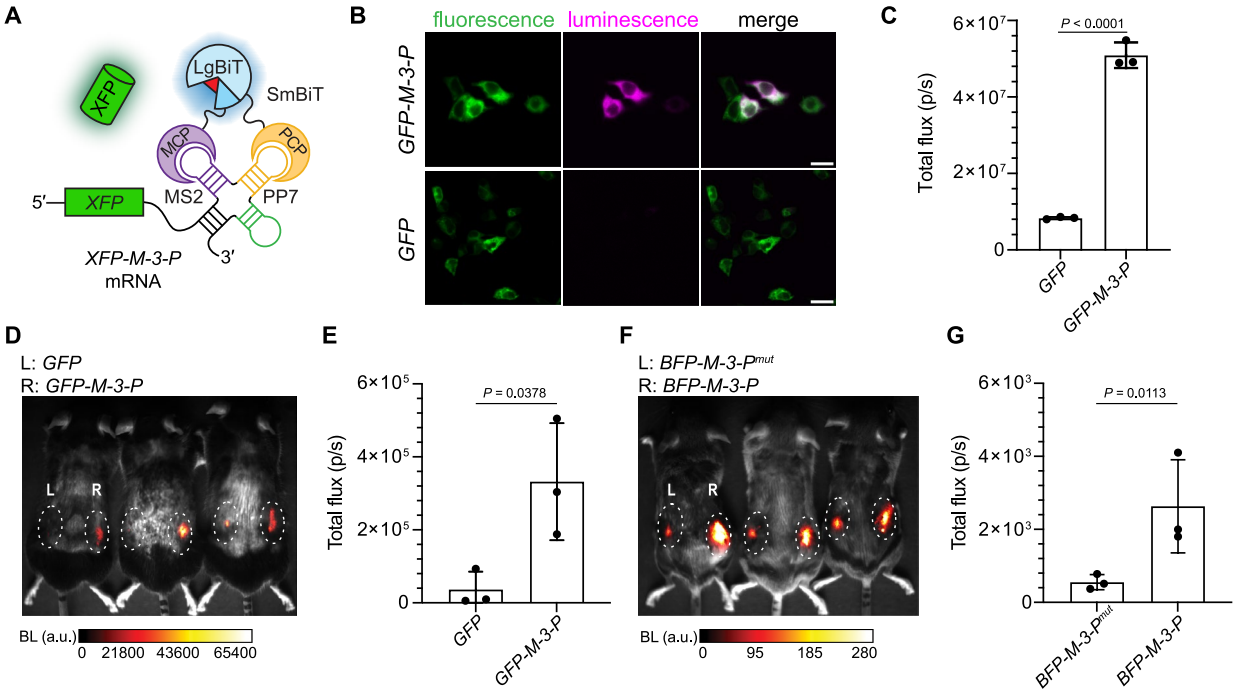


Figure 4.4. Imaging in mammalian cells and live mice with RNA lanterns. (**A**) Schematic of model mRNA encoding a fluorescent protein. The sequence was engineered with M-3-P in the 3′ UTR. XFP production can be further analyzed via fluorescence (green glow). (**B**) HEK293T cells expressing the RNA lantern were transiently transfected with DNA encoding *GFP–M-3-P* (Pearson's r = 0.84) or *GFP* alone (Pearson's r = 0.03). Luminescence was observed exclusively in cells containing mRNAs with the M-3-P bait. Scale bar = 20 μm. (**C**) Bulk measurement of photon flux from lantern-expressing cells transfected with DNA encoding *GFP* or *GFP–M-3-P*. (**D**) Mice implanted with cells expressing RNA lanterns and *GFP* alone (left flank) or *GFP-M-3-P* (right flank) and imaged with luciferin. (**E**) Photon flux from mice with implanted cells expressing *GFP(±)M-3-P* and RNA lanterns shown in (D). (**F**) Mice implanted with cells expressing RNA lanterns and *BFP-M-3-P*^{mut} (left flank) or *BFP-M-3-P* (right flank). (**G**) Photon flux from mice with implanted cells expressing either *BFP-M-3-P* or *BFP-M-3-P*^{mut} and RNA lanterns. Error bars represent standard deviation (SD) for $n = 3$ replicates (Fig. 4.4C, 4.4E, and 4.4G). *P*-values (95% confidence interval) were determined by unpaired, two-tailed *t*-test. BL, bioluminescence. a.u., arbitrary units.

4.4 Discussion

RNA dynamics have been historically visualized in living systems with fluorescent probes. Such methods require optically permissive platforms such as transparent model organisms or surgically implanted windows. The required excitation light can further induce high background signals in tissue and is often limited in depth. Bioluminescent probes (luciferases) overcome some of these limitations. Luciferase reporters use enzymatic reactions—instead of excitation light—to generate photons, which produces far less background signal and can be advantageous for serial imaging in tissue and whole animals.

To capitalize on the sensitivity and dynamic range of bioluminescence, we developed a genetically encoded split luciferase-based platform for visualizing RNA transcripts. The RNA lanterns combine the well-known NanoBiT system and MS2/PP7 platform for transcript tracking. We systematically optimized the components to achieve sensitive imaging. Substantial improvements in signal production were achieved by modulating the spacing and phase angle of the aptamer components. The optimized design significantly decreases the size of previously reported RNA-protein complexes for imaging. Only a single copy of the RNA bait was necessary for imaging target transcripts in cells and whole organisms.

The RNA bait and lanterns also comprise unique features for a range of applications. Both the proteins and bait are equipped with affinity tags for retrieval and downstream analyses of target transcripts and interacting biomolecules. Structured RNA baits can productively assemble diverse split proteins, including Fluc and various BRET reporters. Such modularity expands the color palette of RNA lanterns and sets the stage for even deeper tissue imaging and multiplexed assays. Given our finding that the BRET probes were more efficiently assembled with longer RNA baits, though, care should be taken to optimize each lantern and tag combination.

We anticipate that RNA lanterns will enable RNA dynamics to be visualized in vitro and in vivo. The probes will be particularly useful for serial imaging of transcripts, where localization and expression are difficult to examine over time. While we focused on visualizing model transcripts

with RNA bait incorporated into the 3′ UTR, the short tag can easily be applied to other transcripts through genetic manipulation. Further tuning of both the RNA lantern and tag will expand the number of transcripts that can be visualized in tandem. Such multiplexed studies will paint a more complete picture of RNA biology *in vivo*.

4.5 Materials and Methods

General information

Q5 DNA polymerase, restriction enzymes, and all buffers were purchased from New England Biolabs. dNTPs were purchased from Thermo Fisher Scientific. Luria-Bertani medium (LB) was purchased from Genesee Scientific. All plasmids and primer stocks were stored at -20 °C unless otherwise noted. Primers were purchased from Integrated DNA Technologies and plasmids were sequenced by Azenta Life Sciences. Sequencing traces were analyzed using Benchling.

General cloning methods

Polymerase chain reaction (PCR) was used to prepare genes of interest, and products were analyzed by gel electrophoresis. Products were excised and purified. Amplified genes were ligated into destination vectors via Gibson assembly. All PCR reactions were performed in a BioRad C3000 thermocycler using the following conditions: 1) 95 °C for 3 min, 2) 95 °C for 30 s, 3) –1.2°C per cycle starting at 72 °C for 30 s, 4) 72 °C for 30 sec, repeat steps 2–4 ten times, 5) 95 °C for 3 min, 6) 95 °C for 30 s, 7) 60 °C for 30 s, 8) 72 °C for 2 min repeat steps 6–8 twenty times, then 72 °C for 5 min, and hold at 12 °C until retrieval from the thermocycler. Gibson assembly conditions were: 50 °C for 60 min and held at 12 °C until retrieval from the thermocycler. Ligated plasmids were transformed into TOP10 *E. coli* cells using the heat shock method. After incubation at 37 °C for 18–24 h, colonies were picked and expanded overnight in 5 mL LB broth supplemented with ampicillin (100 µg/mL) or kanamycin (100 µg/mL). DNA was extracted from

colonies using a Zymo Research Plasmid Mini-prep Kit. DNA was subjected to restriction enzyme digestion to confirm gene insertion. Positive hits were further sequenced.

General cell culture methods

HEK293T cells (HEK, ATCC) and stable cells lines derived from HEK293T cells were cultured in complete media: DMEM (Corning) containing 10% (v/v) fetal bovine serum (FBS, Life Technologies), penicillin (100 U/mL), and streptomycin (100 µg/mL, Gibco). Cell lines stably expressing RNA lantern and linker designs were generated via lentiviral transduction. Transduced cells were further cultured with puromycin (20 µg/mL) to preserve gene incorporation. Cells were incubated at 37 °C in a 5% CO_2 humidified chamber. Cells were serially passaged using trypsin (0.25 % in HBSS, Gibco).

Protein expression and purification

LgBiT was encoded in a pCold vector. LgBiT was expressed in *E. coli* BL21 cells grown in LB medium (1 L). Expression was induced at an optical density (OD600) of ~0.6 by addition of 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG), followed by incubation at 37 °C for 4 h. Cells were harvested by centrifugation (4,000 x*g*, 10 min, 4 °C). Cells were then resuspended in lysis buffer (30 mL, 50 mM Tris HCl, 150 mM NaCl, 0.5% Tween-20, 1 mM phenylmethylsulfonyl fluoride [PMSF], pH 7.4). Cells were sonicated (Qsonica) at 40% amplitude, at 2 sec on 2 sec off intervals for 15 min. Cell debris was removed through centrifugation (10,000 x*g*, 1 h, 4 °C). Proteins were purified by Ni-NTA affinity chromatography. The column was washed with wash buffer (20 mM imidazole, 50 mM $NaPO_4$, pH 7.4). Protein was eluted from column with elution buffer (200 mM imidazole, 50 mM $NaPO_4$, pH 7.4). Protein was dialyzed overnight at 4 °C into phosphate buffer (50 mM $NaPO_4$, pH 7.4). Protein was concentrated to ~500 µL using Amicon Ultra-15 Centrifugal Filter Units (Merck Millipore MWCO 3 kDa). The concentration of proteins was determined using JASCO V730 UV-vis spectrophotometer at 280 nm. SDS-PAGE was also

performed to verify purity, and gels were stained with Coomassie R-250.

In vitro transcription

RNA was transcribed *in vitro* in a buffer containing 50 mM Tris-HCl pH 8, 2 mM spermidine, 0.01% Triton X-100, 2 mM rNTPs (each), 20 mM $MgCl_2$, 10 mM dithiothreitol (DTT), and recombinant T7 RNA polymerase. Transcription reactions were quenched with 30 mM EDTA and subsequently ethanol-precipitated with coprecipitant (Invitrogen GlycoBlue, AM9516) prior to purification via denaturing PAGE. Purified RNAs were eluted from gel pieces into 300 mM KCl and 0.1 mM EDTA for 3 h. Eluted RNAs were ethanol-precipitated with coprecipitant, and pellets were washed with 70% ethanol, dried, and resuspended in 10 mM Tris-HCl pH 7.5, 0.1 mM EDTA, and 0.001% Triton X-100 (TET). RNA concentrations were determined by UV-vis absorption spectroscopy (Thermo Scientific NanoDrop 2000, ND-2000).

In vitro transcription/translation (IVTT) in rabbit reticulocyte lysate

A rabbit reticulocyte lysate *in vitro* translation kit (Promega, L4960) was coupled to *in vitro* transcription with the supplementation of $MgCl_2$, rNTPs, and T7 RNA polymerase. Plasmid DNAs were 3′ linearized for run-off transcription. RNA scaffolds were prepared via primer elongation of synthetic DNA oligos (Integrated DNA Technologies). Linearized plasmid DNAs and RNA bait DNAs were kit purified prior to IVTT (DNA Clean and Concentrator; Zymo Research, D4004). All DNAs were eluted into Tris-EDTA (TE; 10 mM Tris-HCl and 0.1 mM EDTA, pH 7.5) and controls were supplemented with an equal volume of TE buffer.

IVTT reactions were prepared on ice with 70% v/v lysate, 0.02 mM amino acid mix, rNTPs (0.35 mM GTP and 0.15 mM each ATP/CTP/UTP), 1 mM $MgCl_2$, linearized plasmid DNA, RNA bait template DNA, 2 U/10 µL RNase inhibitor (Invitrogen, AM2694), and 0.5 µL/10 µL reaction of recombinant T7 RNA polymerase (house preparation). NanoBiT experiments were set up with 10

μL volumes and 1 nM of linearized plasmid DNA. Split firefly luciferase experiments were performed with 25 μL volumes and 3 nM of linearized plasmid DNA. RNA bait DNA templates were varied at specified ratios. IVTT reactions were incubated at 34 °C for 10 min and 30 °C for 140 min.

Luciferin substrates were supplied prior to imaging. For split NanoBiT RNA lantern experiments, furimazine was added to each reaction at 20 μM, and samples were incubated at room temperature for 5 min before imaging. For split firefly luciferase RNA lantern experiments, D-luciferin (100 μM) and ATP (1 mM) were added to each reaction. Samples were then incubated at room temperature for 5 min before image acquisition. Plates were imaged in a dark, light-proof chamber using an IVIS Lumina (PerkinElmer) CCD camera chilled to −90 °C. The stage was kept at 37 °C during imaging and the camera was controlled using Living Image software. Exposure times were set to 1 min and binning levels were set to medium. Regions of interest were selected for quantification and total flux values were analyzed using Living Image software. All data was exported to Microsoft Excel or Prism (GraphPad) for further analysis.

In cellulo linker optimization

Stable cell lines expressing variant RNA lanterns were plated in clear 12-well plates and incubated overnight. Cell lines were then transfected with 1 μL Lipofectamine 3000 (Invitrogen, L3000001), 2 μL P3000 Reagent (Invitrogen, L3000001), and 500 ng of the BFP-M-3-P plasmid. After 24 h incubation, cells were lifted and counted with a Countess II (Invitrogen). 50,000 transfected or non-transfected cells were plated in triplicate into black 96-well plates (Greiner Bio-One). Furimazine (20 μM) was added to each sample. Plates were imaged in a dark, light-proof chamber using an IVIS Lumina (PerkinElmer) CCD camera chilled to −90 °C. The stage was kept at 37 °C during imaging and the camera was controlled using Living Image software. Exposure times were set to 1 min and binning levels were set to medium. Regions of interest were selected for

quantification and total flux values were analyzed using Living Image software. All data was exported to Microsoft Excel or Prism (GraphPad) for further analysis. Remaining cells were analyzed on a Novocyte flow cytometer (ACEA BioSciences) for BFP expression.

General flow cytometry methods

Cells were treated with trypsin for 5 min at 37 °C and then neutralized with complete media. Cells were transferred to Eppendorf tubes and pelleted (500 x*g*, 5 min) using a tabletop centrifuge (Thermo Fisher Sorvall Legend Micro 17). The resulting supernatants were discarded, and cells were washed with PBS (2 × 400 µL). Cells were analyzed for XFP expression on a Novocyte flow cytometer. Live cells were gated, and singlet cells were further gated. For each sample, 10,000 events were collected on the “singlet cell” gate.

HEK293T lysate preparation

HEK293T cells stably expressing RNA lanterns (~10^6) were pelleted at 500 x*g* for 10 min at 4 °C and washed three times with binding buffer (140 mM KCl, 10 mM NaCl, 20 mM Tris-HCl pH 7.5, and 5 mM $MgCl_2$). Cells were pelleted during each washing step at 500 x*g* (1 min, 4 °C) After the final wash, the cell pellet was resuspended in 600 µL binding buffer supplemented with 1% Tween-80, 1X protease inhibitor cocktail (Roche, 11697498001), and 5 µg RNase inhibitor. Cells were sonicated on ice (5 cycles of 10 s on, 10 s off; Branson CPX2800). Cell debris was then removed by centrifugation (12,000 x*g*, 4 °C, 15 min). The resulting supernatant was aliquoted and stored at –80 °C for future use as 2X lysate stocks.

RNA titration assay

Purified RNAs (10X stocks in TET buffer) were serially diluted. Binding reactions contained 12.5 µL of 2X cell lysate, 10 µL of 1X binding buffer, and 2.5 µL of 10X RNA. Reactions were incubated at room temperature on an orbital shaker for 30 minutes. After incubation, 1 µL of a 200 µM

furimazine stock (Promega, Nano-Glo substrate) in 1X binding buffer was added to each reaction and then plated in a black, clear bottom 384-well plate and imaged with an Andor iXon Ultra 888 EMCCD camera (Oxford Instruments) equipped with an F 0.95 lens (Schneider) with an EM gain of 500. Images were acquired (5 s acquisition) over 10 min. Images were Z-stacked in ImageJ and densitometry was used to measure peak intensity values.

Ni-NTA pull-down

Triplicate reactions of selected RNA concentrations above, within, and below the curve (as determined by RNA titration assay) were combined (45 µL/each). Charged Ni-NTA agarose beads (~45 µL; Thermo Fisher, 88221) were washed three times in 90 µL 1X binding buffer in a spin-column (Corning, 8160) at 3000 x*g* for 30 s. The combined binding reactions were then incubated on the Ni-NTA agarose at room temperature for 30 minutes on an orbital shaker. After incubation, the columns were spun at 3000 x*g* for 30 s and the flowthrough was collected. Four 5-min washes were performed with 45 µL 1X binding buffer at room temperature on an orbital shaker. After each incubation, the columns were spun, and the flowthroughs were collected (washes 1-4). Columns were eluted three times with 1X binding buffer supplemented with 25 mM EDTA (20 mM final) and incubated for 5 min on an orbital shaker. Elution fractions were collected. Furimazine was added to the fractions (20 µM final) and the solutions were placed in a black, clear bottom 384-well plate. Images were acquired using an Andor iXon Ultra 888 EMCCD camera equipped with an F 0.95 lens, with an EM gain of 500. Images were acquired (10 s acquisitions) over 15 min. Final images were Z-stacked, and densitometry (ImageJ) was used to determine fold change over background and no RNA controls. Final images were overlaid with brightfield photos. Plots were produced using GraphPad Prism.

HA and FLAG pull-down

Reticulocyte lysate IVTT reactions (67.5 µL) were set up as previously described. To each reaction 100 nM M-3-P RNA (7.5 µL) or 1X binding buffer (7.5 µL) was added prior to incubation. Reactions were incubated at 34 °C for 10 min and 30 °C for 140 min. Pierce magnetic anti-HA beads (45 µL) and anti-FLAG beads (1.5 µL) were normalized for binding capacity. Two sets of each bead were washed three times in 1X binding buffer ten-times their initial volume (450 µL and 15 µL, respectively; 5 min with agitation). To each tube, 75 µL of IVTT reactions were added and incubated with agitation for 60 min at room temperature. The supernatant was removed and stored separately. After incubation, beads were washed three times in 1X binding buffer (75 µL, 5 min with agitation) and each wash was stored. Beads were resuspended in 30 µL 1X binding buffer and 20 µM furimazine, then plated in triplicate in a black, clear bottom 384-well plate and imaged with an Andor iXon Ultra 888 EMCCD camera equipped with an F 0.95 lens, with an EM gain of 100. Images were acquired (5 s acquisition) over 15 min. Images were Z-stacked in ImageJ and densitometry was used to measure peak intensity values.

Bioluminescence microscopy

HEK293T cells (5×10^5) stably expressing RNA lanterns were plated in 8-well Ibidi µ-Slides. After 24 h, the cells were transiently transfected with 100 ng of *GFP*-M-3-P plasmid or GFP plasmid using 0.3 µL Lipofectamine 3000 and 0.3 µL P3000 reagent. After 18 h, cell media was exchanged for phenol red-free DMEM (Gibco FluoroBrite DMEM, A1896701) supplemented with 20 µM furimazine. Live cell images were captured on an Olympus IX71 microscope equipped with an Andor iXon Ultra 888 EMCCD camera and a 40X oil objective (Olympus UPlanApo 40X/1.00 oil iris). Images were captured with an EM gain of 1000, 10 MHz horizontal read-out rate, 4.33 µs vertical clock speed, and an acquisition time of 180 s. The microscope stage was kept warm with a heating pad to maintain cell viability and to encourage enzyme turnover. Fluorescence images were captured immediately following luminescence imaging, using a blue LED light source (ThorLabs, Solis-470C). Fluorescent images were captured using the EMCCD's conventional

mode with an acquisition time of 0.5 s. Images were processed (removed outliers and Z-stacked) with ImageJ and colocalization was determined with the JaCoP plug-in[35].

In vivo cell transplants

Mouse experiments were approved by the UC Irvine Animal Care and Use Committee (IACUC) in compliance with the National Institute of Health guidelines. Mice were maintained on a 12 h light/dark cycle at 25 °C. All procedures were done during the light portion of the cycle. Our studies were done using Rosa26^{tdTomato} male mice obtained from the Jackson Laboratory (JAX: 007905).

Mice were anesthetized using isoflurane (1-2%) and were given subcutaneous dorsal injections of HEK293T cells (1×10^6) expressing RNA lanterns and *GFP*, *GFP-M-3-P*, *BFP-M-3-P^{mut}*, or *BFP-M-3-P* suspended in sterile PBS solution. Cells were normalized for mean fluorescence intensity (MFI) of *XFP* using flow cytometry. Following normalization, cells were implanted into the dorsal posterior subcutaneous flank of each mouse. Cells expressing RNA lantern and controls, *GFP* or *BFP-M-3-P^{mut}*, were implanted on the left side, while cells expressing the RNA lantern and *GFP-M-3-P* or *BFP-M-3-P* were implanted on the right side, allowing each mouse to serve as its own control. Each animal received a 20 μM furimazine injection in 200 μL of implant.

Animals were imaged immediately following transplantation. For imaging animals were anesthetized with an i.p. injection of ketamine (6.6 mg/mL) and xylazine (1.65 mg/mL) suspended in sterile PBS and placed on a warmed (37 °C) stage. Images were captured on an Andor iXon Ultra 888 EMCCD camera equipped with an F 0.95 lens. Images were acquired with an EM gain of 1000, an acquisition time of 300 s, and signal was captured as photons. Images were processed (removing outliers and integrated density) using ImageJ. Total flux (p/s) was determined based on a given region of interest (ROI) for each implantation. Background correction was accomplished by averaging ROIs containing no luminescence. Prism (GraphPad) was used to determine significant differences (unpaired, two-tailed *t*-test) between groups.

4.6 Acknowledgments

We thank Dr. Lorenzo Scipioni and Prof. Michelle Digman (UCI) for help with optical imaging experiments. We also thank members of the Luptak, Steward, and Prescher laboratories for helpful discussions. We also thank our funding sources. W.M. Keck Foundation: A.L., J.A.P., O.S. NASA ICAR 80NSSC21K0596: A.L. The Paul Allen Foundation: J.A.P. National Science Foundation 1804220: A.L. National Institute of Health grant R01CA229696: C.C.C.

4.7 Supplementary Data

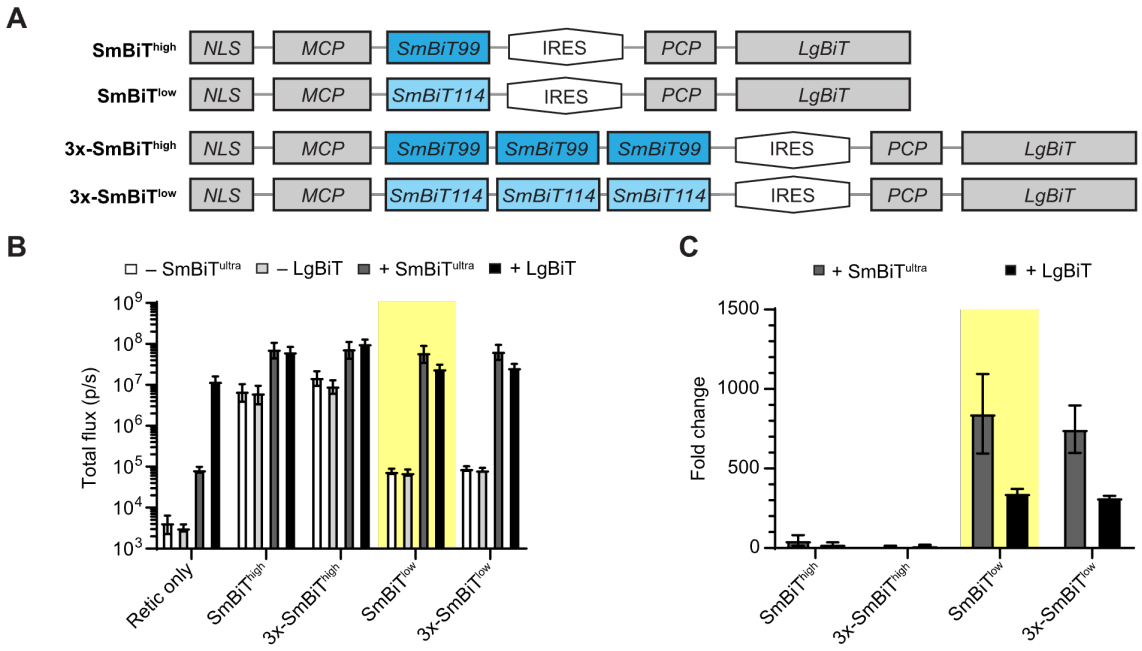


Figure S4.1: Optimization of the split Nluc RNA lantern. A. SmBiT99 (SmBiT^{high}, dark blue) and SmBiT114 (SmBiT^{low}, light blue) and their serial trimers were constructed as NLS-MCP fusions and co-expressed with IRES-driven PCP-LgBiT. **B.** The lantern constructs were expressed in IVTT (rabbit reticulocyte lysate, RRL) and evaluated for RNA-independent photon production alone or supplemented with synthetic SmBiT86 (SmBiT^{ultra} ,10 μM) or recombinant LgBiT (10 μM), as noted. The IVTT-expressed lanterns provide background photon measurements, whereas the supplemented experiments (+SmBiT^{ultra} or + LgBiT) represent the saturated complexes, reporting the maximum potential photon output from each construct. **C.** Fold change in light output from (**B**), calculated for pairs of (saturated/IVTT-expressed) constructs. Error bars represent the standard error of the mean (SEM) for *n* = 3 replicates experiments.

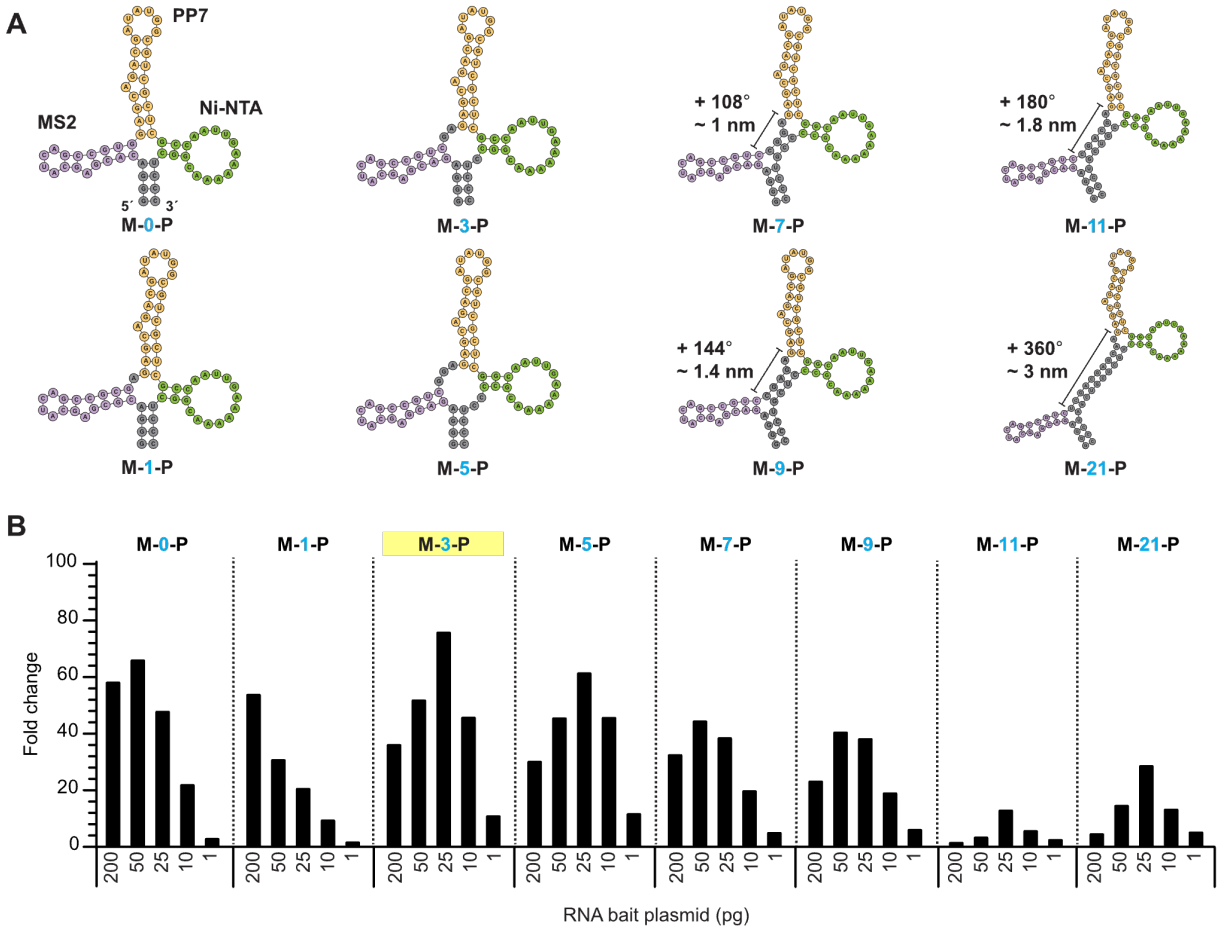


Figure S4.2: Design of novel MCP/PCP baits. A. RNA baits were designed with varying number of nucleotides between the MS2 (purple) and PP7 (orange) aptamers. A Ni-NTA aptamer (green) was also added for purification purposes and improved structural rigidity. **B.** Luminescence readouts obtained with each RNA bait. Data are plotted as the fold change in signal from each RNA bait over a no-RNA control sample. Decreasing concentrations of RNA bait template were assessed with 1 ng of the lantern DNA plasmid. The mean fold change for each sample (n = 2) is shown.

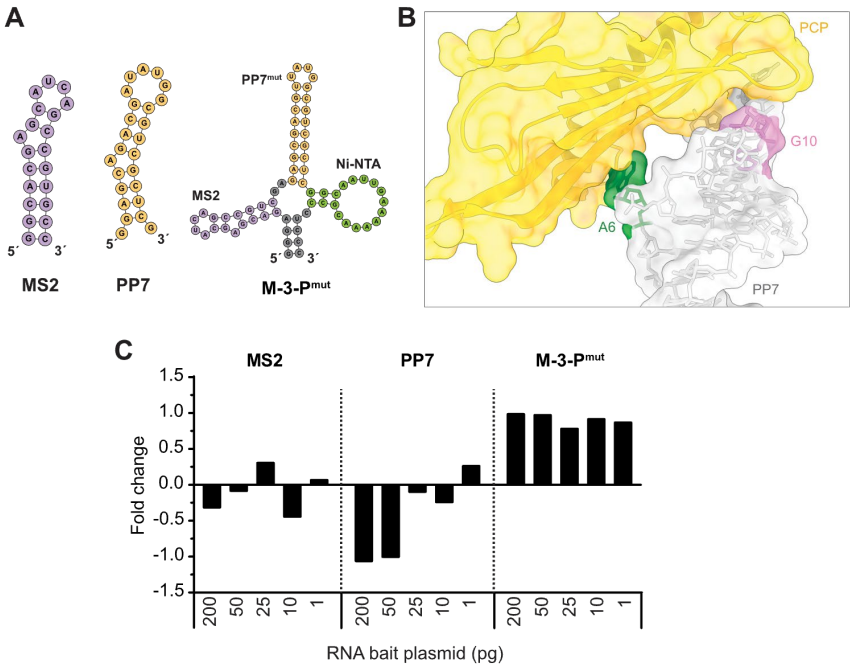


Figure S4.3: Lack of luminescence observed with control RNAs. A. Predicted secondary structures of control RNA baits comprising individual MS2 or PP7 aptamers or the M-3-P^{mut} construct containing two mutations (deletion of A6 and G10). **B.** Location of the deleterious PP7 mutations in the structure of the unmutated aptamer-protein complex (PDB: 2QUX). **C.** Fold change in signal observed with each control RNA bait over a no-RNA sample. Decreasing concentrations of RNA bait template were tested with 1 nM of the lantern DNA plasmid. The mean fold change for each sample (n = 2) is shown.

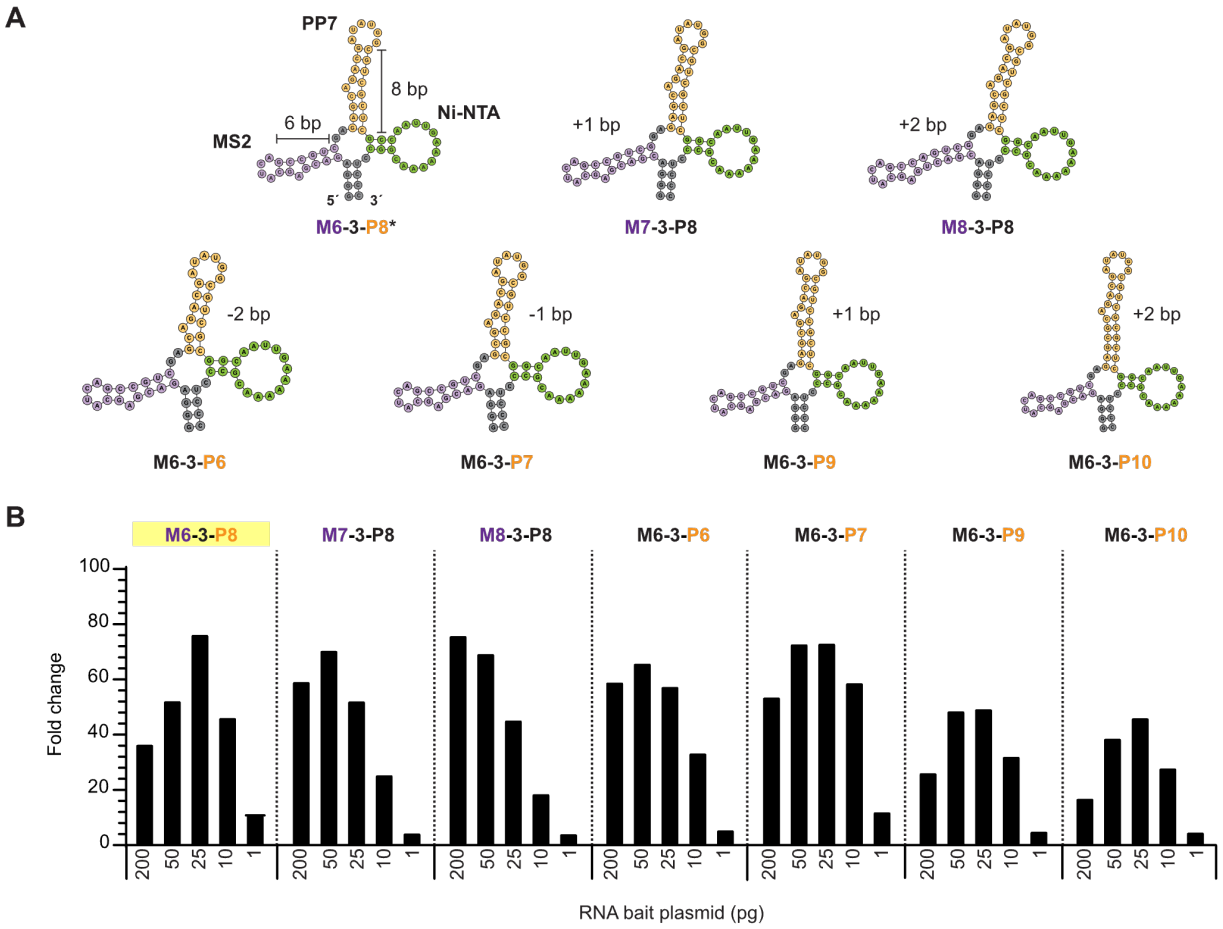


Figure S4.4: Modulation of aptamer stem lengths. A. The stem regions of the MS2 (purple) and PP7 (orange) aptamers were varied by changing the number of base pairs. The asterisk indicates RNA bait with unmodified MS2 and PP7 stems (Fig. 4E). **B**. Fold change in signal observed with each RNA bait design over a no-RNA control. Decreasing concentrations of each RNA bait template were tested with 1 ng of the lantern DNA plasmid. The mean fold change for each sample (n = 2) is shown.

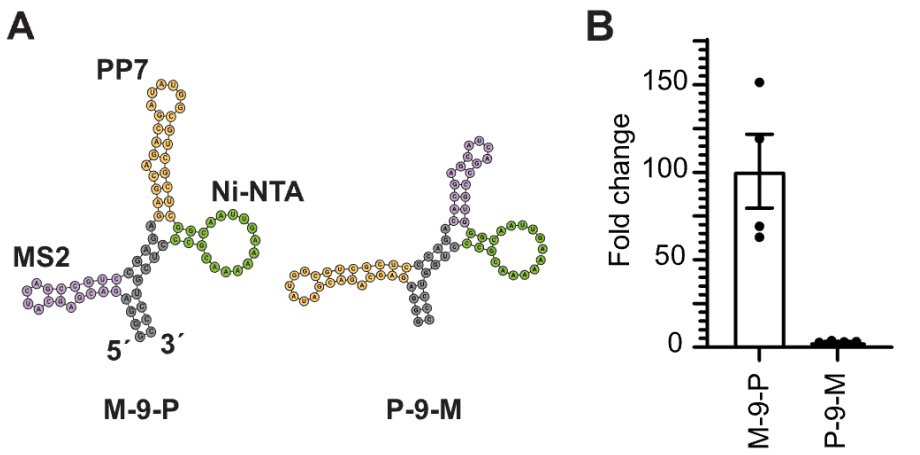


Figure S4.5: Aptamer order is critical for lantern assembly. **A.** The PP7 aptamer (orange) was placed either 5′ (right) or 3′ (left) of the MS2 aptamer (purple) to evaluate the effect of binding site orientation on lantern function. **B.** RNA baits were assessed using IVTT and the fold changes in light output over samples without RNA bait are plotted. Error bars represent the standard error of the mean (SEM) for n = 4 replicates.

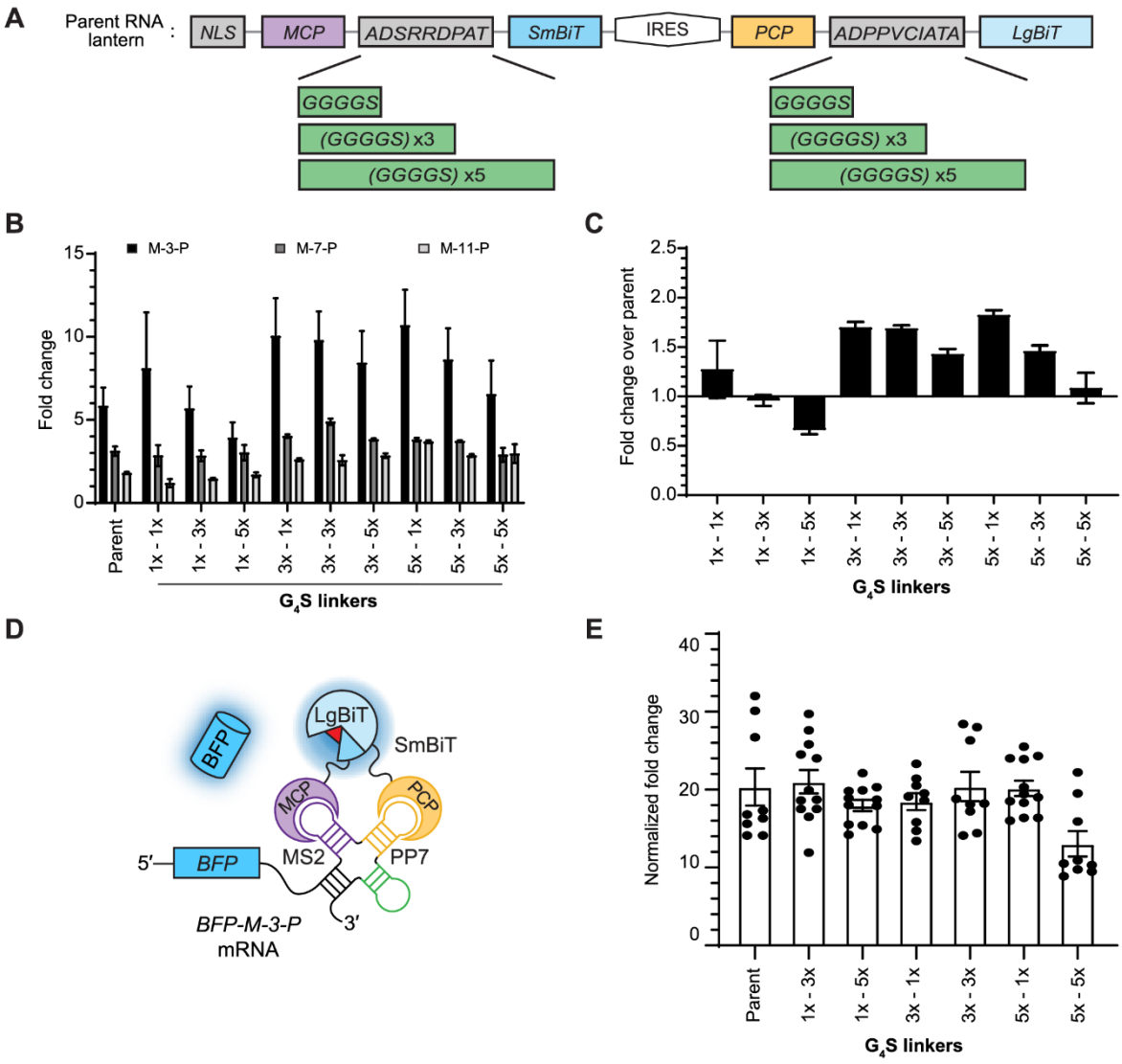


Figure S4.6: Lantern linker lengths have a smaller effect on photon output than RNA bait structure. A. Schematic of G_4S linker changes in the RNA lantern. **B.** Protein linkers comprising 1–5 copies of a glycine-serine repeating unit (G_4S) were examined in IVTT. Luminescence outputs were recorded in the presence of various RNA baits. Data are plotted as the fold change in signal over no-RNA bait controls. **C**. Fold change in luminescence observed with RNA lanterns comprising altered linkers and M-3-P versus the original lantern and M-3-P. No substantial increase in signal was observed with probes comprising G_4S units. For **A-B**, error bars represent the standard error of the mean (SEM) for n = 4 replicates. **D**. Schematic of an mRNA encoding blue fluorescent protein (BFP) and M-3-P in the 3′ UTR. Transcription of *BFP-M-3-P* mRNA recruits the RNA lantern, resulting in light production. BFP fluorescence enables confirmation of expression. **E.** RNA lanterns comprising various linker lengths were examined. Cells stably expressing each lantern were transfected with the *BFP-M-3-P* construct. Data are plotted as the fold-change in signal over non-transfected cells. Luminescence measurements were normalized to BFP expression (assessed via flow cytometry). Nine replicates are shown, and the error bars represent the standard error of the mean (SEM) for all replicates. Minimal improvement over the initial RNA lantern design was observed with varying glycine-serine linkers. The 5x-5x design (five

G_4S linkers in both protein fusions) exhibited reduced signal turn-on both *in vitro* and *in cellulo*, suggesting disfavored complementation with increasing linker length.

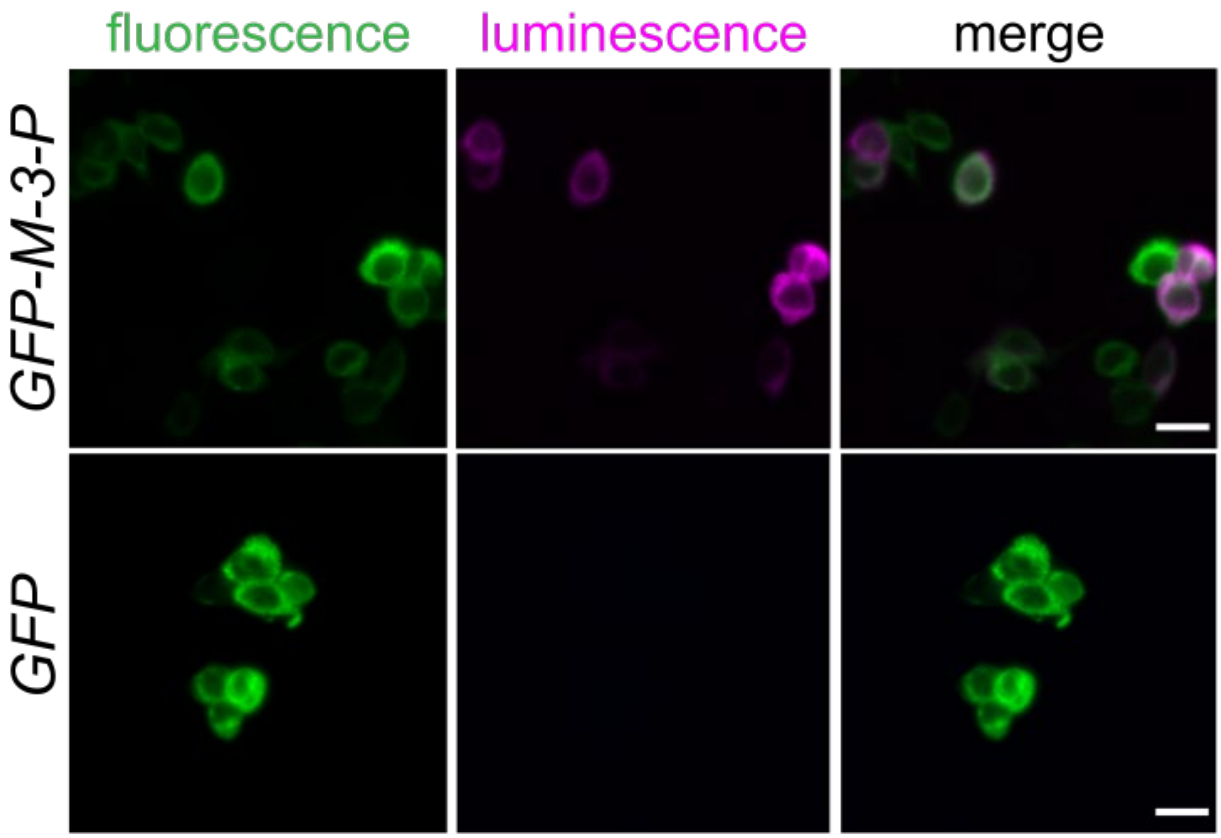


Figure S4.7 RNA imaging *in cellulo* with bioluminescent lantern. HEK293T cells expressing RNA lanterns were transfected with DNA encoding *GFP–M-3-P* (Pearson's r = 0.57) or *GFP* alone (Pearson's r = 0.15). The images shown are replicates of the experiment described in Fig. 4.4B. Scale bar = 20 μm.

Table S4.1

Features	Plasmid Vector	Promoter	Expression in	Used in
NLS - HA - MS2 - SmBiT114 - IRES - PP7 - LgBiT - FLAG	pCDNA 3.1/ pLenti 3rd generation	CMV	mammalian	Fig. 1D-G 2B-F Fig. 4B-G S1 S2 S3 S4 S5 S6 S7
NLS - HA - MS2 - SmBiT99 - IRES - PP7 - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S1
NLS - HA - MS2 - 3x SmBiT99 - IRES - PP7 - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S1
SV40 NLS - HA - MS2 - 3x SmBiT114 - IRES - PP7 - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S1
NLS - HA - MS2 - G4S - SmBiT114 - IRES - PP7 - G4S - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S6
NLS - HA - MS2 - G4S - SmBiT114 - IRES - PP7 - G4Sx3 - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S6
NLS - HA - MS2 - G4S - SmBiT114 - IRES - PP7 - G4Sx5 - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S6
NLS - HA - MS2 - G4Sx3 - SmBiT114 - IRES - PP7 - G4S - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S6
NLS - HA - MS2 - G4Sx3 - SmBiT114 - IRES - PP7 - G4Sx3 - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S6
NLS - HA - MS2 - G4Sx3 - SmBiT114 - IRES - PP7 - G4Sx5 - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S6
NLS - HA - MS2 - G4Sx5 - SmBiT114 - IRES - PP7 - G4S - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S6
NLS - HA - MS2 - G4Sx5 - SmBiT114 - IRES - PP7 - G4Sx3 - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S6
NLS - HA - MS2 - G4Sx5 - SmBiT114 - IRES - PP7 - G4Sx5 - LgBiT - FLAG	pCDNA 3.1	CMV	mammalian	S6
NLS - HA - MS2 - linker - SmBiT114 - IRES - PP7 - linker - YeLgBiT - FLAG	pCDNA 3.1	CMV	mammalian	Fig. 3G-H
NLS - HA - MS2 - linker - SmBiT114 - IRES - PP7 - linker - LumiLgBiT - FLAG	pCDNA 3.1	CMV	mammalian	Fig. 3G-H
NLS - HA - MS2 - FlucN - IRES - PP7 - FlucC - FLAG	pCDNA 3.1	CMV	mammalian	Fig. 3C
BFP - MS2-3-PP7	pCDNA 3.1	CMV	mammalian	Fig. 4F-G S6
BFP - MS2-3-PP7 (mut)	pCDNA 3.1	CMV	mammalian	Fig. 4F-G
Staygold - MS2-3-PP7	pCDNA 3.1	CMV	mammalian	Fig. 3D-E S7
Staygold	pCDNA 3.1	CMV	mammalian	Fig. 3D-E S7

Table S4.2

Plasmid and RNA sequences:

CMV enhancer/ promoter

NLS - HA - MCP - SmBiT114 - IRES - PCP - LgBiT - FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccatag
taacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcatatgcc
aagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttcctacttggca
gtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactcacggggat
ttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaacaactccgccc
cattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaacccactgcttactg
gcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttggctacccctacgac
gtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctcgtcgacaatggcgga
actggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgcgttcacaggcttacaaag
taacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaaaggcgcctggcgttcgtactt
aaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgcaaggtctcctaaaagatggaa
acccgattccctcagcaatcgcagcaaactccggcatctacgcggattctagacgggatcccgccaccgtgaccggctaccgcctgt
ttgaggagattctgtaacgtacgacgcgtggatccctcccccccccctaacgttactggccgaagccgcttggaataaggccggtgtg
cgtttgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcccggaaacctggccctgtcttcttgacgagcattcctagg
ggtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcagttcctctggaagcttcttgaagacaaacaac
gtctgtagcgacccttgcaggcagcggaaccccccacctggcgacaggtgcctctgcggccaaaagccacgtgtataagatacac
ctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaaagagtcaaatggctctcctcaagcgtattcaacaa
ggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctggggcctcggtgcacatgctttacatgtgtttagtcgaggtt
aaaaaaacgtctaggccccccgaaccacggggacgtggttttcctttgaaaaacacgatgataatatggccacagcctgcaggatg
ggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctgactgagatccagtccaccgcagaccgtcagatcttcgaag
agaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctccgtcaaaacggagccaagaccgcgtatcgcgtcaaccta
aaactggatcaggcggacgtcgttgattccggacttccgaaagtgcgctacactcaggtatggtcgcacgacgtgacaatcgttgcga
atagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaagtccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaacctt
gtgccgctgggccgtgcggatccaccggtatgcatcgccaccgccttcacactcgaagatttcgttggggactgggaacagacagcc
gcctacaacctggaccaagtccttgaacagggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccgatccaaaggat
tgtccggagcggtgaaaatgccctgaagatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgaccaaatggcccagat
cgaagaggtgtttaaggtggtgtaccctgtggatgatcatcactttaaggtgatcctgccctatggcacactggtaatcgacggggttac
gccgaacatgctgaactatttcggacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtg
gaacggcaacaaaattatcgacgagcgcctgatcacccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggact
acaaggacgacgatgacaag

NLS - HA - MCP - SmBiT99 - IRES - PCP - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg

gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgcggattctagacgggatcccg
ccaccgtgaccggctaccgcctgtttgagaaaattagctaacgtacgacgcgtggatccctccccccccctaacgttactggccg
aagccgcttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcccggaaacct
ggccctgtcttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcag
ttcctctggaagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaacccccacctggcgacaggtgcctc
tgcggccaaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaa
agagtcaaatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctggg
gcctcggtgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggcccccgaaccacggggacgtggttttcctttga
aaaacacgatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctga
ctgagatccagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctc
cgtcaaaacggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttccgaaagtg
cgctacactcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaa
gtccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgtgcggatccaccggtatgcatcgccac
cgccttcacactcgaagatttcgttggggactgggaacagacagccgcctacaacctggaccaagtccttgaacagggaggtgtg
tccagtttgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtgaaaatgccctgaagatcgacatc
catgtcatcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgtttaaggtggtgtaccctgtggatgat
catcactttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacatgctgaactatttcggacggccgtatg
aaggcatcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacggcaacaaaattatcgacgagcgcctg
atcacccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaaggacgacgatgacaa

NLS - HA - MCP - 3X SmBiT99 - IRES - PCP - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgcggattctagacgggatcccg
ccaccgtgaccggctaccgcctgtttgagaaaattagcgggagcggaagtggggttaccggttatcgtctgttcgaaaaaatcagc
gggtccggatcgggtgtgacaggctatcgtctgtttgagaaaatttcctaacgtacgacgcgtggatccctcccccccccctaacgtt
actggccgaagccgcttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggccc
ggaaacctggccctgtcttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaag
gaagcagttcctctggaagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaacccccacctggcgac
aggtgcctctgcggccaaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgccacgttgtgagttggata
gttgtggaaagagtcaaatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggtaccccattgtatgggatc
tgatctggggcctcggtgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggcccccgaaccacggggacgtggt
tttcctttgaaaaacacgatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcg
cactctgactgagatccagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacg

gcttcgctccgtcaaaacggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttc
cgaaagtgcgctacactcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacga
tttgaccaagtccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgtgcggatccaccggtatgc
atcgccaccgccttcacactcgaagatttcgttggggactgggaacagacagccgcctacaacctggaccaagtccttgaacag
ggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtgaaaatgccctgaa
gatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgtttaaggtggtgtacc
ctgtggatgatcatcactttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacatgctgaactatttcgg
acggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacggcaacaaaattatcg
acgagcgcctgatcacccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaaggacgacgatgac
aag

SV40 NLS - HA - MCP - 3x SmBiT114 - IRES - PCP - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgcccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgcggattctagacgggatcccg
ccaccgtgaccggctaccgcctgtttgaggagattctgggagcggaagtggggtgaccggatatcggctgtttgaggagattctg
gggtccggatcgggtgtgacagggtatcgcctgttcgaggaaatcctgtaacgtacgacgcgtggatccctcccccccccctaacg
ttactggccgaagccgcttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcc
cggaaacctggccctgtcttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaa
ggaagcagttcctctggaagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaacccccccacctggcga
caggtgcctctgcggccaaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgccacgttgtgagttggat
agttgtggaaagagtcaaatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggtaccccattgtatggga
tctgatctggggcctcggtgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggcccccgaaccacggggacgtg
gttttcctttgaaaaacacgatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttcggtcggcgaggctact
cgcactctgactgagatccagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctca
cggcttcgctccgtcaaaacggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggacgtcgttgattccggact
tccgaaagtgcgctacactcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtac
gatttgaccaagtccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgtgcggatccaccggtat
gcatcgccaccgccttcacactcgaagatttcgttggggactgggaacagacagccgcctacaacctggaccaagtccttgaaca
gggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtgaaaatgccctga
agatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgtttaaggtggtgtac
cctgtggatgatcatcactttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacatgctgaactatttcg
gacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacggcaacaaaattatc
gacgagcgcctgatcacccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaaggacgacgatga
caag

NLS - HA - MCP - G4S - SmBiT114 - IRES - PCP - G4S - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgatatcggcggcggaggctctgt
gaccggctaccgcctgtttgaggagattctgtaacgtacgacgcgtggatccctccccccccctaacgttactggccgaagccgc
ttggaataaggccggtgtgcgtttgtctatatgttatttccaccatattgccgtcttttggcaatgtgagggcccggaaacctggccctgt
cttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcagttcctctgg
aagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaacccccccacctggcgacaggtgcctctgcggcc
aaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaaagagtca
aatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctggggcctcgg
tgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggccccccgaaccacggggacgtggttttcctttgaaaaacac
gatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctgactgagatc
cagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctccgtcaaa
acggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttccgaaagtgcgctaca
ctcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaagtccctc
gtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgtatcgatggcggaggcgggtcttcacactcgaa
gatttcgttggggactgggaacagacagccgcctacaacctggaccaagtccttgaacagggaggtgtgtccagtttgctgcaga
atctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtgaaaatgccctgaagatcgacatccatgtcatcatcccgt
atgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgtttaaggtggtgtaccctgtggatgatcatcactttaaggtg
atcctgccctatggcacactggtaatcgacggggttacgccgaacatgctgaactatttcggacggccgtatgaaggcatcgccgt
gttcgacggcaaaaagatcactgtaacagggaccctgtggaacggcaacaaaattatcgacgagcgcctgatcacccccgacg
gctccatgctgttccgagtaaccatcaacagcctcgaggactacaaggacgacgatgacaag

NLS - HA - MCP - G4S - SmBiT114 - IRES - PCP - G4Sx3 - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgatatcggcggcggaggctctgt

gaccggctaccgcctgtttgaggagattctgtaacgtacgacgcgtggatccctcccccccccctaacgttactggccgaagccgc
ttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcccggaaacctggccctgt
cttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcagttcctctgg
aagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaacccccacctggcgacaggtgcctctgcggcc
aaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaaagagtca
aatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctggggcctcgg
tgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggcccccgaaccacggggacgtggttttcctttgaaaaacac
gatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctgactgagatc
cagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctccgtcaaa
acggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttccgaaagtgcgctaca
ctcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaagtccctc
gtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgtatcgatggcggaggcgggtctggcggcggcg
gcagcggtggaggaggttcttcacactcgaagatttcgttggggactgggaacagacagccgcctacaacctggaccaagtcct
tgaacagggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtgaaaatg
ccctgaagatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgtttaaggt
ggtgtaccctgtggatgatcatcactttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacatgctgaa
ctatttcggacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacggcaacaa
aattatcgacgagcgcctgatcacccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaaggacga
cgatgacaag

NLS - HA - MCP - G4S - SmBiT114 - IRES - PCP - G4Sx5 - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgatatcggcggcggaggctctgt
gaccggctaccgcctgtttgaggagattctgtaacgtacgacgcgtggatccctcccccccccctaacgttactggccgaagccgc
ttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcccggaaacctggccctgt
cttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcagttcctctgg
aagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaacccccacctggcgacaggtgcctctgcggcc
aaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaaagagtca
aatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctggggcctcgg
tgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggcccccgaaccacggggacgtggttttcctttgaaaaacac
gatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctgactgagatc
cagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctccgtcaaa
acggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttccgaaagtgcgctaca
ctcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaagtccctc
gtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgtatcgatggcggaggcgggtctggtggagggg
gatctggaggtggtgggtctggcggcggcggcagcggtggaggaggttcttcacactcgaagatttcgttggggactgggaaca
gacagccgcctacaacctggaccaagtccttgaacagggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccga

tccaaaggattgtccggagcggtgaaaatgccctgaagatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgacca
aatggcccagatcgaagaggtgtttaaggtggtgtaccctgtggatgatcatcactttaaggtgatcctgccctatggcacactggta
atcgacggggttacgccgaacatgctgaactatttcggacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactg
taacagggaccctgtggaacggcaacaaaattatcgacgagcgcctgatcacccccgacggctccatgctgttccgagtaacca
tcaacagcctcgaggactacaaggacgacgatgacaag

NLS - HA - MCP - G4SX3 - SmBiT114 - IRES – PCP - G4S - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgatatcggcggcggaggctctg
gaggcggcggaagcggtgggggtgggagcgtgaccggctaccgcctgtttgaggagattctgtaacgtacgacgcgtggatccc
tccccccccccctaacgttactggccgaagccgcttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtctttt
ggcaatgtgagggcccggaaacctggccctgtcttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggt
ctgttgaatgtcgtgaaggaagcagttcctctggaagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaa
ccccccacctggcgacaggtgcctctgcggccaaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgc
cacgttgtgagttggatagttgtggaaagagtcaaatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggt
accccattgtatgggatctgatctggggcctcggtgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggccccccg
aaccacggggacgtggttttcctttgaaaaacacgatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttc
ggtcggcgaggctactcgcactctgactgagatccagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtg
ggtcggctgcgcctcacggcttcgctccgtcaaaacggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggac
gtcgttgattccggacttccgaaagtgcgctacactcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctc
gcgcaaatcgttgtacgatttgaccaagtccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgt
atcgatggcggaggcgggtcttcacactcgaagatttcgttggggactgggaacagacagccgcctacaacctggaccaagtcc
ttgaacagggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtgaaaatg
ccctgaagatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgtttaaggt
ggtgtaccctgtggatgatcatcactttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacatgctgaa
ctatttcggacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacggcaacaa
aattatcgacgagcgcctgatcacccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaaggacga
cgatgacaag

NLS - HA - MCP - G4Sx3 - SmBiT114 - IRES – PCP - G4Sx3 - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg

gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgatatcggcggcggaggctctg
gaggcggcggaagcggtgggggtgggagcgtgaccggctaccgcctgtttgaggagattctgtaacgtacgacgcgtggatccc
tccccccccccctaacgttactggccgaagccgcttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtctttt
ggcaatgtgagggcccggaaacctggccctgtcttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggt
ctgttgaatgtcgtgaaggaagcagttcctctggaagcttcttgaagacaaacaacgtctgtagcgacccctttgcaggcagcggaa
ccccccacctggcgacaggtgcctctgcggccaaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgc
cacgttgtgagttggatagttgtggaaagagtcaaatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggt
accccattgtatgggatctgatctggggcctcggtgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggcccccg
aaccacggggacgtggttttcctttgaaaaacacgatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttc
ggtcggcgaggctactcgcactctgactgagatccagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtg
ggtcggctgcgcctcacggcttcgctccgtcaaaacggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggac
gtcgttgattccggacttccgaaagtgcgctacactcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctc
gcgcaaatcgttgtacgatttgaccaagtccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgt
atcgatggcggaggcgggtctggcggcggcggcagcggtggaggaggttcttttcacactcgaagatttcgttggggactgggaa
cagacagccgcctacaacctggaccaagtccttgaacagggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactcc
gatccaaaggattgtccggagcggtgaaaatgccctgaagatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgac
caaatggcccagatcgaagaggtgtttaaggtggtgtaccctgtggatgatcatcactttaaggtgatcctgccctatggcacactgg
taatcgacggggttacgccgaacatgctgaactatttcggacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcac
tgtaacagggaccctgtggaacggcaacaaaattatcgacgagcgcctgatcacccccgacggctccatgctgttccgagtaac
catcaacagcctcgaggactacaaggacgacgatgacaag

NLS - HA - MCP - G4SX3 - SmBiT114 - IRES - PCP - G4SX5 - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgatatcggcggcggaggctctg
gaggcggcggaagcggtgggggtgggagcgtgaccggctaccgcctgtttgaggagattctgtaacgtacgacgcgtggatccc
tccccccccccctaacgttactggccgaagccgcttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtctttt
ggcaatgtgagggcccggaaacctggccctgtcttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggt
ctgttgaatgtcgtgaaggaagcagttcctctggaagcttcttgaagacaaacaacgtctgtagcgacccctttgcaggcagcggaa
ccccccacctggcgacaggtgcctctgcggccaaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgc
cacgttgtgagttggatagttgtggaaagagtcaaatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggt
accccattgtatgggatctgatctggggcctcggtgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggcccccg
aaccacggggacgtggttttcctttgaaaaacacgatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttc
ggtcggcgaggctactcgcactctgactgagatccagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtg

ggtcggctgcgcctcacggcttcgctccgtcaaaacggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggac
gtcgttgattccggacttccgaaagtgcgctacactcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctc
gcgcaaatcgttgtacgatttgaccaagtccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgt
atcgatggcggaggcgggtctggtggaggggatctggaggtggtgggtctggcggcggcggcagcggtggaggaggttcttc
acactcgaagatttcgttggggactgggaacagacagccgcctacaacctggaccaagtccttgaacagggaggtgtgtccagtt
tgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtgaaaatgccctgaagatcgacatccatgtc
atcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgtttaaggtggtgtaccctgtggatgatcatca
ctttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacatgctgaactatttcggacggccgtatgaagg
catcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacggcaacaaaattatcgacgagcgcctgatcac
ccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaaggacgacgatgacaag

NLS - HA - MCP - G4SX5 - SmBiT114 - IRES - PCP - G4S – LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgatatcggcggcggaggctctg
gtggaggaggttctggcggcggtggttccggaggcggcggaagcggtgggggtgggagcgtgaccggctaccgcctgtttgag
gagattctgtaacgtacgacgcgtggatccctcccccccccctaacgttactggccgaagccgcttggaataaggccggtgtgcgtt
tgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcccggaaacctggccctgtcttcttgacgagcattcctaggg
gtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcagttcctctggaagcttcttgaagacaaacaa
cgtctgtagcgacccctttgcaggcagcggaacccccccacctggcgacaggtgcctctgcggccaaaagccacgtgtataagata
cacctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaaagagtcaaatggctctcctcaagcgtattc
aacaaggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctggggcctcggtgcacatgctttacatgtgtttagt
cgaggttaaaaaaacgtctaggcccccgaaccacggggacgtggttttcctttgaaaaacacgatgataatatggccacagcct
gcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctgactgagatccagtccaccgcagaccgtcag
atcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctccgtcaaaacggagccaagaccgcgtatc
gcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttccgaaagtgcgctacactcaggtatggtcgcacgacgtg
acaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaagtccctcgtcgcgacctcgcaggtcgaag
atcttgtcgtcaaccttgtgccgctgggccgtatcgatggcggaggcgggtcttcacactcgaagatttcgttggggactgggaaca
gacagccgcctacaacctggaccaagtccttgaacagggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccga
tccaaaggattgtccggagcggtgaaaatgccctgaagatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgacca
aatggcccagatcgaagaggtgtttaaggtggtgtaccctgtggatgatcatcactttaaggtgatcctgccctatggcacactggta
atcgacggggttacgccgaacatgctgaactatttcggacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactg
taacagggaccctgtggaacggcaacaaaattatcgacgagcgcctgatcacccccgacggctccatgctgttccgagtaacca
tcaacagcctcgaggactacaaggacgacgatgacaag

NLS - HA – MCP - G4Sx5 - SmBiT114 - IRES - PCP - G4Sx3 - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat

atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgatatcggcggcggaggctctg
gtggaggaggttctggcggcggtggttccggaggcggcggaagcggtgggggtgggagcgtgaccggctaccgcctgtttgag
gagattctgtaacgtacgacgcgtggatccctcccccccccctaacgttactggccgaagccgcttggaataaggccggtgtgcgtt
tgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcccggaaacctggccctgtcttcttgacgagcattcctaggg
gtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcagttcctctggaagcttcttgaagacaaacaa
cgtctgtagcgacccttgcaggcagcggaacccccacctggcgacaggtgcctctgcggccaaaagccacgtgtataagata
cacctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaaagagtcaaatggctctcctcaagcgtattc
aacaaggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctggggcctcggtgcacatgctttacatgtgtttagt
cgaggttaaaaaaacgtctaggccccccgaaccacggggacgtggttttcctttgaaaaacacgatgataatatggccacagcct
gcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctgactgagatccagtccaccgcagaccgtcag
atcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctccgtcaaaacggagccaagaccgcgtatc
gcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttccgaaagtgcgctacactcaggtatggtcgcacgacgtg
acaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaagtccctcgtcgcgacctcgcaggtcgaag
atcttgtcgtcaaccttgtgccgctgggccgtatcgatggcggaggcgggtctggcggcggcggcagcggtggaggaggttcttc
acactcgaagatttcgttggggactgggaacagacagccgcctacaacctggaccaagtccttgaacagggaggtgtgtccagtt
tgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtgaaaatgccctgaagatcgacatccatgtc
atcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgtttaaggtggtgtaccctgtggatgatcatca
ctttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacatgctgaactatttcggacggccgtatgaagg
catcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacggcaacaaaattatcgacgagcgcctgatcac
ccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaaggacgacgatgacaag

NLS - HA – MCP - G4Sx5 - SmBiT114 - IRES - PCP - G4Sx5 - LgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgatatcggcggcggaggctctg
gtggaggaggttctggcggcggtggttccggaggcggcggaagcggtgggggtgggagcgtgaccggctaccgcctgtttgag
gagattctgtaacgtacgacgcgtggatccctcccccccccctaacgttactggccgaagccgcttggaataaggccggtgtgcgtt
tgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcccggaaacctggccctgtcttcttgacgagcattcctaggg
gtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcagttcctctggaagcttcttgaagacaaacaa

cgtctgtagcgacccttgcaggcagcggaacccccacctggcgacaggtgcctctgcggccaaaagccacgtgtataagata
cacctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaaagagtcaaatggctctcctcaagcgtattc
aacaaggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctggggcctcggtgcacatgctttacatgtgtttagt
cgaggttaaaaaaacgtctaggccccccgaaccacggggacgtggttttcctttgaaaaacacgatgataatatggccacagcct
gcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctgactgagatccagtccaccgcagaccgtcag
atcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctccgtcaaaacggagccaagaccgcgtatc
gcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttccgaaagtgcgctacactcaggtatggtcgcacgacgtg
acaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaagtccctcgtcgcgacctcgcaggtcgaag
atcttgtcgtcaaccttgtgccgctgggccgtatcgatggcggaggcgggtctggtggaggggatctggaggtggtgggtctggc
ggcggcggcagcggtggaggaggttcttcacactcgaagatttcgttggggactgggaacagacagccgcctacaacctggac
caagtccttgaacagggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggt
gaaaatgccctgaagatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtg
tttaaggtggtgtaccctgtggatgatcatcactttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacat
gctgaactatttcggacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacgg
caacaaaattatcgacgagcgcctgatcacccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaa
ggacgacgatgacaag

NLS - HA – MCP - linker - SmBi114 - IRES - PCP -linker - YeLgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgcggattctagacgggatcccg
ccaccgtgaccggctaccgcctgtttgaggagattctgtaacgtacgacgcgtggatccctcccccccccctaacgttactggccga
agccgcttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcccggaaacctg
gccctgtcttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcagtt
cctctggaagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaacccccccacctggcgacaggtgcctct
gcggccaaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaaa
gagtcaaatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctgggg
cctcggtgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggccccccgaaccacggggacgtggttttcctttgaa
aaacacgatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctgac
tgagatccagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctcc
gtcaaaacggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttccgaaagtgc
gctacactcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaag
tccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgtgcggatccaccggtatgcatcgccacc
gccgtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccacaagtt
cagcgtgtccggcgagggcgagggcgatgccacctacggcaagctgaccctgaagctgatctgcaccaccggcaagctgccc
gtgccctggcccaccctcgtgaccaccctgggctacggcctgcagtgcttcgcccgctaccccgaccacatgaagcagcacgac
ttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaactacaagacccgcgcc
gaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctgg

ggcacaagctggagtacaactacaacagccacaacgtctatatcaccgccgacaagcagaagaacggcatcaaggccaactt
caagatccgccacaacatcgaggacggcggcgtgcagctcgccgaccactaccagcagaacaccccatcggcgacggccc
cgtgctgctgcccgacaaccactacctgagctaccagtccgccctgagcaaagaccccaacgagaagcgcgatcacatggtcc
tgctggagttcgtgaccgccgcgatgttcacactcgaagatttcgttggggactgggaacagacagccgcctacaacctggacca
agtccttgaacagggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtga
aaatgccctgaagatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgttt
aaggtggtgtaccctgtggatgatcatcactttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacatg
ctgaactatttcggacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacggc
aacaaaattatcgacgagcgcctgatcacccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaag
gacgacgatgacaag

NLS - HA – MCP - linker - SmBiT114 - IRES - PCP - linker - LumiLgBiT – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatgggcccaaaaaagaaaagaaaagttg
gctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcagttcgttctc
gtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctctaactcgc
gttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggtgcctaa
aggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaaggcaatgc
aaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgcggattctagacgggatcccg
ccaccgtgaccggctaccgcctgtttgaggagattctgtaacgtacgacgcgtggatccctccccccccctaacgttactggccga
agccgcttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtcttttggcaatgtgagggcccggaaacctg
gccctgtcttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggtctgttgaatgtcgtgaaggaagcagtt
cctctggaagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaacccccccacctggcgacaggtgcctct
gcggccaaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgccacgttgtgagttggatagttgtggaaa
gagtcaaatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggtaccccattgtatgggatctgatctgggg
cctcggtgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggccccccgaaccacggggacgtggttttcctttgaa
aaacacgatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttcggtcggcgaggctactcgcactctgac
tgagatccagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtgggtcggctgcgcctcacggcttcgctcc
gtcaaaacggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggacgtcgttgattccggacttccgaaagtgc
gctacactcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctcgcgcaaatcgttgtacgatttgaccaag
tccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgtgcggatccaccggtatgcatcgccacc
gccgtgagcaagggcgaggcagtgatcaaggagttcatgcggttcaaggtgcacatggagggctccatgaacggccacgagtt
cgagatcgagggcgagggcgagggccgcccctacgagggcacccagaccgccaagctgaaggtgaccaagggtggcccc
ctgcccttctcctgggacatcctgtcccctcagttcatgtacggctccagggccttcatcaagcaccccgccgacatccccgactact
ataagcagtccttccccgagggcttcaagtgggagcgcgtgatgaacttcgaggacggcggcgccgtgaccgtgacccaggac
acctccctggaggacggcaccctgatctacaaggtgaagctccgcggcaccaacttccctcctgacggccccgtaatgcagaag
aagacaatgggctgggaagcgtccaccgagcggttgtaccccgaggacggcgtgctgaagggcgacattaagatggccctgc
gcctgaaggacggcggccgctacctggcggacttcaagaccacctacaaggccaagaagcccgtgcagatgcccggcgccta
caacgtcgaccgcaagttggacatcacctcccacaacgaggactacaccgtggtggaacagtacgaacgctccgagggccgc
cactccaccggaaagttcacactcgaagatttcgttggggactgggaacagacagccgcctacaacctggaccaagtccttgaa
cagggaggtgtgtccagtttgctgcagaatctcgccgtgtccgtaactccgatccaaaggattgtccggagcggtgaaaatgccct
gaagatcgacatccatgtcatcatcccgtatgaaggtctgagcgccgaccaaatggcccagatcgaagaggtgtttaaggtggtgt

accctgtggatgatcatcactttaaggtgatcctgccctatggcacactggtaatcgacggggttacgccgaacatgctgaactatttc
ggacggccgtatgaaggcatcgccgtgttcgacggcaaaaagatcactgtaacagggaccctgtggaacggcaacaaaattat
cgacgagcgcctgatcacccccgacggctccatgctgttccgagtaaccatcaacagcctcgaggactacaaggacgacgatg
acaag

NLS - HA - MCP - FlucN - IRES - PCP - FlucC – FLAG

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctcgtttagtgaaccgtcagat
cgcctggagacgccatccacgctgttttgacctccatagaagacaccgactctagagccaccatgggcccaaaaaagaaaaga
aaagttggctacccctacgacgtgcccgactacgccatcgaaggccgccatatgctagccgttaaaatggcttctaactttactcag
ttcgttctcgtcgacaatggcggaactggcgacgtgactgtcgccccaagcaacttcgctaacgggatcgctgaatggatcagctct
aactcgcgttcacaggcttacaaagtaacctgtagcgttcgtcagagctctgcgcagaatcgcaaatacaccatcaaagtcgaggt
gcctaaaggcgcctggcgttcgtacttaaatatggaactaaccattccaattttcgccacgaattccgactgcgagcttattgttaagg
caatgcaaggtctcctaaaagatggaaacccgattccctcagcaatcgcagcaaactccggcatctacgcggattctagacggg
atcccgccaccgaagacgccaaaaacataaagaaaggcccggcgccattctatccgctggaagatggaaccgctggagagc
aactgcataaggctatgaagagatacgccctggttcctggaacaattgcttttacagatgcacatatcgaggtggacatcacttacg
ctgagtacttcgaaatgtccgttcggttggcagaagctatgaaacgatatgggctgaatacaaatcacagaatcgtcgtatgcagtg
aaaactctcttcaattctttatgccggtgttgggcgcgttatttatcggagttgcagttgcgcccgcgaacgacatttataatgaacgtga
attgctcaacagtatgggcatttcgcagcctaccgtggtgttcgtttccaaaaaggggttgcaaaaaattttgaacgtgcaaaaaaa
gctcccaatcatccaaaaaattattatcatggattctaaaacggattaccagggatttcagtcgatgtacacgttcgtcacatctcatct
acctcccggttttaatgaatacgattttgtgccagagtccttcgatagggacaagacaattgcactgatcatgaactcctctggatcta
ctggtctgcctaaaggtgtcgctctgcctcatagaactgcctgcgtgagattctcgcatgccagagatcctatttttggcaatcaaatca
ttccggatactgcgattttaagtgttgttccattccatcacggttttggaatgtttactacactcggatatttgatatgtggatttcgagtcgtc
ttaatgtatagatttgaagaagagctgtttctgaggagccttcaggattacaagattcaaagtgcgctgctggtgccaaccctattctcc
ttcttcgccaaaagcactctgattgacaaatacgatttatctaatttacacgaaattgcttctggtggcgctcccctctctaaggaagtcg
gggaagcggttgccaagaggttccatctgccaggtatcaggcaaggatatgggctcactgagactacatcagctattctgattaca
cccgaggggatgataaaccgggcgcggtcggtaaagttgttccatttttgaagcgaaggttgtggatctggataccgggaaaac
gctgggcgttaatcaaagaggcgaactgtgtgtgagaggtcctatgattatgtcctaacgtacgacgcgtggatccctcccccccc
ctaacgttactggccgaagccgcttggaataaggccggtgtgcgtttgtctatatgttattttccaccatattgccgtcttttggcaatgtg
agggcccggaaacctggccctgtcttcttgacgagcattcctaggggtctttcccctctcgccaaaggaatgcaaggtctgttgaatg
tcgtgaaggaagcagttcctctggaagcttcttgaagacaaacaacgtctgtagcgacccttgcaggcagcggaaccccccacc
tggcgacaggtgcctctgcggccaaaagccacgtgtataagatacacctgcaaaggcggcacaaccccagtgccacgttgtga
gttggatagttgtggaaagagtcaaatggctctcctcaagcgtattcaacaaggggctgaaggatgcccagaaggtaccccattgt
atgggatctgatctggggcctcggtgcacatgctttacatgtgtttagtcgaggttaaaaaaacgtctaggccccccgaaccacggg
gacgtggttttcctttgaaaaacacgatgataatatggccacagcctgcaggatgggctccaaaaccatcgttctttcggtcggcga
ggctactcgcactctgactgagatccagtccaccgcagaccgtcagatcttcgaagagaaggtcgggcctctggtgggtcggctg
cgcctcacggcttcgctccgtcaaaacggagccaagaccgcgtatcgcgtcaacctaaaactggatcaggcggacgtcgttgatt
ccggacttccgaaagtgcgctacactcaggtatggtcgcacgacgtgacaatcgttgcgaatagcaccgaggcctcgcgcaaat
cgttgtacgatttgaccaagtccctcgtcgcgacctcgcaggtcgaagatcttgtcgtcaaccttgtgccgctgggccgtgcggatcc
accggtatgcatcgccaccgccctatgattatgtccggttatgtaaacaatccggaagcgaccaacgccttgattgacaaggatg
gatggctacattctggagacatagcttactgggacgaagacgaacacttcttcatcgttgaccgcctgaagtctctgattaagtacaa

aggctatcaggtggctcccgctgaattggaatccatcttgctccaacaccccaacatcttcgacgcaggtgtcgcaggtcttcccga
cgatgacgccggtgaacttcccgccgccgttgttgttttggagcacggaaagacgatgacggaaaaagagatcgtggattacgtc
gccagtcaagtaacaaccgcgaaaaagttgcgcggaggagttgtgtttgtggacgaagtaccgaaaggtcttaccggaaaactc
gacgcaagaaaaatcagagagatcctcataaaggccaagaagggcggaaagatcgccgtgctcgaggactacaaggacga
cgatgacaag

BFP - MS2-3-PP7

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgcccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttcgccaccatgagcgagctgattaaggaga
acatgcacatgaagctgtacatggagggcaccgtggacaaccatcacttcaagtgcacatccgagggcgaaggcaagccctac
gagggcacccagaccatgagaatcaaggtggtcgagggcggccctctccccttcgccttcgacatcctggctactagcttcctcta
cggcagcaagaccttcatcaaccacacccagggcatccccgacttcttcaagcagtccttccctgagggcttcacatgggagaga
gtcaccacatacgaagacgggggcgtgctgaccgctacccaggacaccagcctccaggacggctgcctcatctacaacgtcaa
gatcagaggggtgaacttcacatccaacggccctgtgatgcagaagaaaacactcggctgggaggccttcaccgagacgctgt
accccgctgacggcggcctggaaggcagaaacgacatggccctgaagctcgtgggcgggagccatctgatcgcaaacatcaa
gaccacatatagatccaagaaacccgctaagaacctcaagatgcctggcgtctactatgtggactacagactggaaagaatcaa
ggaggccaacaacgagacctacgtcgagcagcacgaggtggcagtggccagatactgcgacctccctagcaaactggggca
caagcttaattaagaattctgcagatatccatcacactggcggccgcgggagacgagcatcagccgtcgagagcagacgatatg
gcgtcgctcggcaattgaaaaaacgccctccc

***BFP - MS2-3-PP7* (mut)**

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgcccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttcgccaccatgagcgagctgattaaggaga
acatgcacatgaagctgtacatggagggcaccgtggacaaccatcacttcaagtgcacatccgagggcgaaggcaagccctac
gagggcacccagaccatgagaatcaaggtggtcgagggcggccctctccccttcgccttcgacatcctggctactagcttcctcta
cggcagcaagaccttcatcaaccacacccagggcatccccgacttcttcaagcagtccttccctgagggcttcacatgggagaga
gtcaccacatacgaagacgggggcgtgctgaccgctacccaggacaccagcctccaggacggctgcctcatctacaacgtcaa
gatcagaggggtgaacttcacatccaacggccctgtgatgcagaagaaaacactcggctgggaggccttcaccgagacgctgt
accccgctgacggcggcctggaaggcagaaacgacatggccctgaagctcgtgggcgggagccatctgatcgcaaacatcaa
gaccacatatagatccaagaaacccgctaagaacctcaagatgcctggcgtctactatgtggactacagactggaaagaatcaa
ggaggccaacaacgagacctacgtcgagcagcacgaggtggcagtggccagatactgcgacctccctagcaaactggggca
caagcttaattaagaattctgcagatatccatcacactggcggccgcgggagacgagcatcagccgtcgagagcgacgttatgg
cgtcgctcggcaattgaaaaaacgccctccc

Staygold – M-3-P

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttgcggccgccaccatgctgctgcccgtcccc
ctgctgctgggcctgctgggcgccgccgcggatccgatggtgagcaagggcgaggagctgtttacaggcgtgacccccttcaagt
tccagctgaagggcaccatcaacggcaagagcttcaccgtggaaggcgagggcgagggcaatagccacgagggcagccac
aaaggcaagtacgtgtgcaccagcggcaaactgccaatgtcttgggccgccctgggaactagcttcggctatggcatgaagtact
acaccaagtaccccagcggcctgaagaactggttccacgaggtgatgcccgagggcttcacctacgacagacacatccagtac
aagggcgacggcagcatccacgccaagcaccagcacttcatgaagaacggcacctaccacaacatcgtggagttcaccggcc
aggacttcaaggagaacagccccgtgctgaccggcgacatgaacgtgagcctgcccaacgaggtgcagcacatccccagag
atgacggcgtggagtgcccagtgaccctgctgtaccctctgctgagcgacaagagcaagtgcgtggaggcctaccagaacacc
atcatcaagcccctgcacaatcagccagcccccgatgtgccataccactggatcagaaagcagtacacccagagcaaggacg
acaccgaggagagagaccacatcatccagagcgagaccctggaggcccacctgccatggcacgagccttctgcttctgccgtg
aaggatgaactgtaagaattctgcagatatccatcacactggcggccgcgggagacgagcatcagccgtcgagagcagacgat
atggcgtcgctcggcaattgaaaaaacgccctccc

Staygold

cgttacataacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccat
agtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcat
atgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggactttccta
cttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgggcgtggatagcggtttgactc
acggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaac
aactccgccccattgacgcaaatgggcggtaggcgtgtacggtgggaggtctatataagcagagctctctggctaactagagaac
ccactgcttactggcttatcgaaattaatacgactcactatagggagacccaagcttatggcgtctactcccttcaagttccagctgaa
gggcaccatcaacggcaagagcttcaccgtggaaggcgagggcgagggcaatagccacgagggcagccacaaaggcaag
tacgtgtgcaccagcggcaaactgccaatgtcttgggccgccctgggaactagcttcggctatggcatgaagtactacaccaagta
ccccagcggcctgaagaactggttccacgaggtgatgcccgagggcttcacctacgacagacacatccagtacaagggcgac
ggcagcatccacgccaagcaccagcacttcatgaagaacggcacctaccacaacatcgtggagttcaccggccaggacttcaa
ggagaacagccccgtgctgaccggcgacatgaacgtgagcctgcccaacgaggtgcagcacatccccagagatgacggcgt
ggagtgcccagtgaccctgctgtaccctctgctgagcgacaagagcaagtgcgtggaggcctaccagaacaccatcatcaagc
ccctgcacaatcagccagcccccgatgtgccataccactggatcagaaagcagtacacccagagcaaggacgacaccgagg
agagagaccacatcatccagagcgagaccctggaggcccacctg

RNA bait DNA sequences (T7 promoter)

Flexible
gtaatacgactcactataGGCACGAGCATCAGCCGTGCCTCCAGGTCGAATCTTCAAACGAGCA
GACGATATGGCGTCGCTCG
MS2
gtaatacgactcactataGGCACGAGCATCAGCCGTGCC

PP7
gtaatacgactcactataGGAGCAGACGATATGGCGTCGCTCG
MS2-3-PP7^{Mut}-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCGAGAGCGACGTTATGGCGTCGCTCG
GCAATTGAAAAAACGCCCTCCC
MS2-0-PP7-Ni
gtaatacgactcactataGGGACACGAGCATCAGCCGTGGAGCAGACGATATGGCGTCGCTCGC
GAATTGAAAAAACCGCTCCC
MS2-1-PP7-Ni
gtaatacgactcactataGGGACGCGAGCATCAGCCGCGAGAGCAGACGATATGGCGTCGCTCG
CGAATTGAAAAAACCGCTCCC
MS2-3-PP7-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCGAGAGCAGACGATATGGCGTCGCTC
GGCAATTGAAAAAACGCCCTCCC
MS2-5-PP7-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCGGAGAGCAGACGATATGGCGTCGCT
CGGCAATTGAAAAAACGCCCCTCCC
MS2-7-PP7-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCCGGAGAGCAGACGATATGGCGTCGC
TCGGCAATTGAAAAAACGCCCCGTCCC
MS2-9-PP7-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCCGAGAGAGCAGACGATATGGCGTCG
CTCGGCAATTGAAAAAACGCCCTCGTCCC
MS2-11-PP7-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCCGACGAGAGCAGACGATATGGCGTC
GCTCGGCAATTGAAAAAACGCCCGTCGTCCC
MS2-13-Ni-PP7
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCCTGAGCGGCAATTGAAAAAACGCCG
AGCAGACGATATGGCGTCGCTCGCTCAGTCCC
MS2-21-PP7-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCTCGACGAGCAAGAGCAGACGATATG
GCGTCGCTCGGCAATTGAAAAAACGCCTGCTCGTCGATCCC
7-MS2-7-PP7-Ni
gtaatacgactcactataGGGACGACGAGCATCAGCCGTCGCGGAGAGCAGACGATATGGCGTC
GCTCGGCAATTGAAAAAACGCCCCGTCCC
8-MS2-7-PP7-Ni
gtaatacgactcactataGGGACGACTGAGCATCAGCCAGTCGCGGAGAGCAGACGATATGGCG
TCGCTCGGCAATTGAAAAAACGCCCCGTCCC
7-MS2-3-PP7-Ni
gtaatacgactcactataGGGACGACGAGCATCAGCCGTCGGAGAGCAGACGATATGGCGTCGC
TCGGCAATTGAAAAAACGCCCTCCC
8-MS2-3-PP7-Ni
gtaatacgactcactataGGGACGACTGAGCATCAGCCAGTCGGAGAGCAGACGATATGGCGTC
GCTCGGCAATTGAAAAAACGCCCTCCC
6-MS2-3-6-PP7-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCGAGCAGACGATATGGCGTCGCGGCA
ATTGAAAAAACGCCCTCCC

6-MS2-3-9-PP7-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCGAGAGCGAGACGATATGGCGTCCGC
TCGGCAATTGAAAAAACGCCCTCCC
6-MS2-3-10-PP7-Ni
gtaatacgactcactataGGGAGACGAGCATCAGCCGTCGAGAGCGCAGACGATATGGCGTCGC
GCTCGGCAATTGAAAAAACGCCCTCCC
PP7-9-MS2-Ni
gtaatacgactcactataGGGAGAGCAGACGATATGGCGTCGCTCCCAGACACGAGCATCAGCC
GTGGGCAATTGAAAAAACGCCCTGGTCCC

4.8 References

1. Sharp, P.A. The Centrality of RNA. *Cell* **136**, 577-580 (2009).

2. Larson, D.R., Zenklusen, D., Wu, B., Chao, J.A. & Singer, R.H. Real-Time Observation of Transcription Initiation and Elongation on an Endogenous Yeast Gene. *Science* **332**, 475-478 (2011).

3. Halstead, J.M. et al. An RNA biosensor for imaging the first round of translation from single cells to living animals. *Science* **347**, 1367-1671 (2015).

4. Paige, J.S., Wu, K.Y. & Jaffrey, S.R. RNA mimics of green fluorescent protein. *Science* **333**, 642-646 (2011).

5. Das, S., Moon, H.C., Singer, R.H. & Park, H.Y. A transgenic mouse for imaging activity-dependent dynamics of endogenous Arc mRNA in live neurons. *Sci Adv* **4**, eaar3448 (2018).

6. Lee, B.H. et al. Real-time visualization of mRNA synthesis during memory formation in live mice. *Proceedings of the National Academy of Sciences* **119**, e2117076119 (2022).

7. Jiang, L. et al. Large Stokes shift fluorescent RNAs for dual-emission fluorescence and bioluminescence imaging in live cells. *Nature Methods* **20**, 1563-1572 (2023).

8. Hall, M.P. et al. Engineered Luciferase Reporter from a Deep Sea Shrimp Utilizing a Novel Imidazopyrazinone Substrate. *ACS Chemical Biology* **7**, 1848-1857 (2012).

9. Suzuki, K. et al. Five colour variants of bright luminescent protein for real-time multicolour bioimaging. *Nature Communications* **7**, 13718 (2016).

10. Schaub, F.X. et al. Fluorophore-NanoLuc BRET Reporters Enable Sensitive In Vivo Optical Imaging and Flow Cytometry for Monitoring Tumorigenesis. *Cancer Res* **75**, 5023-5033 (2015).

11. Yeh, H.-W. et al. ATP-Independent Bioluminescent Reporter Variants To Improve in Vivo Imaging. *ACS Chemical Biology* **14**, 959-965 (2019).

12. Eguchi, M., Yoshimura, H., Ueda, Y. & Ozawa, T. Split Luciferase-Fragment Reconstitution for Unveiling RNA Localization and Dynamics in Live Cells. *ACS Sensors* (2023).

13. Lim, F., Downey, T.P. & Peabody, D.S. Translational repression and specific RNA binding by the coat protein of the Pseudomonas phage PP7. *J Biol Chem* **276**, 22507-22513 (2001).

14. Bernardi, A. & Spahr, P.-F. Nucleotide Sequence at the Binding Site for Coat Protein on RNA of Bacteriophage R17. *Proceedings of the National Academy of Sciences* **69**, 3033-3037 (1972).

15. Wu, B., Chen, J. & Singer, R.H. Background free imaging of single mRNAs in live cells using split fluorescent proteins. *Scientific Reports* **4**, 3615 (2014).

16. Dixon, A.S. et al. NanoLuc Complementation Reporter Optimized for Accurate Measurement of Protein Interactions in Cells. *ACS Chemical Biology* **11**, 400-408 (2016).

17. Yeh, H.W. & Ai, H.W. Development and Applications of Bioluminescent and Chemiluminescent Reporters and Biosensors. *Annu Rev Anal Chem (Palo Alto Calif)* **12**, 129-150 (2019).

18. Bodle, C.R., Hayes, M.P., O'Brien, J.B. & Roman, D.L. Development of a bimolecular luminescence complementation assay for RGS: G protein interactions in cells. *Anal Biochem* **522**, 10-17 (2017).

19. Machleidt, T. et al. NanoBRET—A Novel BRET Platform for the Analysis of Protein–Protein Interactions. *ACS Chemical Biology* **10**, 1797-1804 (2015).

20. Goddard, T.D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **27**, 14-25 (2018).

21. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 26 (2011).

22. Kerpedjiev, P., Hammer, S. & Hofacker, I.L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**, 3377-3379 (2015).

23. Chizzolini, F. et al. Large Phenotypic Enhancement of Structured Random RNA Pools. *Journal of the American Chemical Society* **142**, 1941-1951 (2020).

24. Lim, F. & Peabody, D.S. RNA recognition site of PP7 coat protein. *Nucleic acids research* **30**, 4138-4144 (2002).

25. Lionnet, T. et al. A transgenic mouse for in vivo detection of endogenous labeled mRNA. *Nat Methods* **8**, 165-170 (2011).

26. Tutucci, E. et al. An improved MS2 system for accurate reporting of the mRNA life cycle. *Nat Methods* **15**, 81-89 (2018).

27. Garcia, J.F. & Parker, R. MS2 coat proteins bound to yeast mRNAs block 5′ to 3′ degradation and trap mRNA decay products: implications for the localization of mRNAs by MS2-MCP system. *RNA* **21**, 1393-1395 (2015).

28. Heinrich, S., Sidler, C.L., Azzalin, C.M. & Weis, K. Stem–loop RNA labeling can affect nuclear and cytoplasmic mRNA processing. *RNA* **23**, 134-141 (2017).

29. Garcia, J.F. & Parker, R. Ubiquitous accumulation of 3′ mRNA decay fragments in Saccharomyces cerevisiae mRNAs with chromosomally integrated MS2 arrays. *RNA* **22**, 657-659 (2016).

30. Bindels, D.S. et al. mScarlet: a bright monomeric red fluorescent protein for cellular imaging. *Nature Methods* **14**, 53-56 (2017).

31. Ng, K.K. & Prescher, J.A. Generalized Bioluminescent Platform To Observe and Track Cellular Interactions. *Bioconjugate Chemistry* **33**, 1876-1884 (2022).

32. Valegârd, K. et al. The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. *J Mol Biol* **270**, 724-738 (1997).

33. Chao, J.A., Patskovsky, Y., Almo, S.C. & Singer, R.H. Structural basis for the coevolution of a viral RNA-protein complex. *Nat Struct Mol Biol* **15**, 103-105 (2008).

34. Gibson, D.G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6**, 343-345 (2009).

35. Bolte, S. & CordeliÈRes, F.P. A guided tour into subcellular colocalization analysis in light microscopy. *Journal of Microscopy* **224**, 213-232 (2006).

Chapter 5: Development of an *in vitro* selection scheme to improve an aminoacylating ribozyme with tRNA specificity

Note:

Originally submitted, in part, as the written proposal for the Advancement to Candidacy Examination (2021) for the Department of Molecular Biology and Biochemistry.

Contribution statement: The STARzyme project was done in collaboration with Dr. Barbara Golden's laboratory at Purdue University. Figures 5.2A and Figure 5.3 contain data generated by Gary Gan while a graduate student in the Golden lab.

5.1 Abstract

A ribozyme with selective tRNA aminoacylation activity would be a powerful tool to introduce noncanonical amino acids (ncAAs) site-specifically into proteins both *in vitro* and *in vivo*. To this end, we have developed a model aminoacylating ribozyme, known as STARzyme (Specific tRNA aminoacylating ribozyme) [1]. STARzyme has been designed as a fusion of two well-characterized functional RNAs: the T-box riboswitch and the flexizyme. The T-box riboswitch specifically recognizes its cognate tRNA through its specifier loop, which base pairs to the anticodon of the tRNA[3]. The flexizyme is an artificially evolved ribozyme that aminoacylates RNAs with 3′-CCA tails by utilizing activated amino acid substrates[4, 5]. The current version of STARzyme allows for programmability of the specifier loop, enabling it to recognize the amber-suppressor mutant of the noncognate *M. jannaschii* tRNA (*Mj*tRNA$^{AMB}_{CUA}$). The *Mj*tRNA$^{AMB}_{CUA}$ is orthogonal in *E.coli* expression systems[7], and we have demonstrated that STARzyme can selectively charge the amber-suppressor mutant tRNA in a cell-free *in vitro* transcription and translation system.

5.2 Background and Significance

The 20 essential amino acids and the genetic code that represents them are highly conserved across all kingdoms of life. Covalent modification of amino acids is not a novel practice as nature has evolved mechanisms to post-translationally modify proteins to accomplish biological roles unachievable without chemical alteration[8, 9]. Efforts have been made to artificially expand the genetic code to incorporate alternative or noncanonical amino acids (ncAAs) into translating proteins, thus further diversifying the chemical repertoire of the encoded proteins[10, 11]. The most common approach to incorporate ncAAs is through amber suppression, where the anticodon of an orthogonal tRNA is reassigned to recognize the tRNA-less amber (UAG) stop codon[12]. The aminoacylation of the orthogonal tRNA is accomplished through recognition by a paired aminoacyl-tRNA synthetase (aaRS). Pairs such as the tyrosyl-aaRS from the archaea bacteria *M. jannaschii (Mj),* TyrS/tRNA$_{CUA}$, are commonly used introduce ncAAs in *E. coli* as they are

recognized by the *E. coli* translational machinery (e.g., ribosome and EF-Tu), but are not charged by *E. coli* aminoacyl-tRNA synthetases[7]. The recognition of ncAAs by the orthogonal aaRS typically requires directed evolution for the selection of active-site mutations which can accommodate varying ncAA side chains[13]. Directed evolution efforts have led to the successful selection of *M. jannaschii* TyrS mutants that can accommodate many derivatives of tyrosine[14]. While there are many successful approaches used to evolve the active site of orthogonal aaRS enzymes, each approach requires the selection of a functional enzyme that recognizes a single tRNA and ncAA[10].

The development of RNA catalysts that aminoacylate target tRNAs has been proposed as an alternative approach for ncAA incorporation. Suga and colleagues have developed a class of ribozyme, known as the flexizymes, which can aminoacylate tRNAs without anticodon recognition[15]. Flexizymes have proven to be versatile catalysts because they recognize the 3′-CCA tail of tRNAs and aromatic leaving groups of modified amino acid substrates, enabling aminoacylation of nearly any RNA with a 3′-CCA tail[5]. This versatility limits flexizyme's usefulness as an aaRS-like ribozyme *in vivo* as it lacks the capability to differentiate between tRNA anticodons, an essential checkpoint for deciphering the genetic code. Recent efforts have been made to evolve ribozymes which can differentiate between tRNAs, but were evolved at the cost of substrate flexibility[16]. In collaboration with the Golden Lab at Purdue University, we have developed an aminoacylating ribozyme that can selectively aminoacylate its cognate tRNA while having amino-acid substrate flexibility. This ribozyme, known as the STARzyme, was developed as a rational fusion between the stem I of the T-box riboswitch and the dFx flexizyme (**Figure 5.1A**)[4, 17]. The T-box stem I contains a specifier loop consisting of three nucleotides, which base pairs the anticodon of a cognate tRNA[3, 18]. Programmability of this recognition element allows for the use of all STOP codons (amber, ochre, opal), with the potential of utilizing other unassigned codons (e.g., *E. coli* $tRNA^{Ser}_{UGA}$ and $tRNA^{Ser}_{CGA}$) for further genetic code expansion[19]. Furthermore, STARzyme's amino acid specificity is lower than that of aaRSs because its catalytic

subunit, the dFx flexizyme, recognizes most dinitrobenzyl ester (DBE) activated amino-acid substrates[4]. Flexizyme variants have been evolved to recognize alternative leaving groups (e.g., cyanomethyl ester [CME], chlorobenzyl thioester [CBT] or amino-derivatizing benzyl thioester [ABT]) of activated amino acids with only small nucleotide variations required for recognition[15]. We hypothesize that *in vitro* selection of this candidate ribozyme will result in an aaRS-like ribozyme that can function *in vivo* with an orthogonal tRNA, allowing for a flexible system for amino acid incorporation.

5.3 Results and Discussion

The development of an aaRS-like ribozyme with tRNA specificity and amino-acid substrate flexibility would be an innovative tool in the field of synthetic biology. A ribozyme with this functionality *in vivo* can directly compete in the niche currently occupied by aaRS protein enzymes. Simply changing the sequence of three base pairs in the DNA template of the STARzyme recodes the specifier loop of the T-box module and facilitates the recognition of practically any tRNA anticodon, as illustrated in **Figure 5.1B**. Recoding of the specifier loop is especially appealing as more genetically recoded organisms (GROs) with expanded unassigned codons are developed (e.g., *E. coli* Syn61.Δ3) [19]. Furthermore, the use of activated amino-acid substrates allows for an increase in the diversity of available ncAA side chains[20]. The specificity for activated amino-acid substrates, as opposed to side chains, means STARzyme does not need to be evolved to recognize individual ncAAs like that of aaRS protein enzymes. This adaptable platform enables STARzyme to be a robust synthetic biology tool with applications in the development of pharmaceuticals, biophysical probes, and cross-linkable peptides[21].

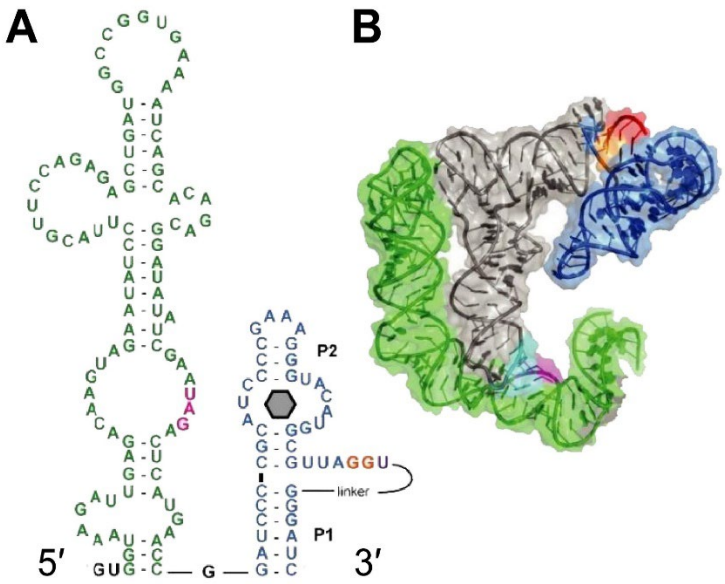


Figure 5.1: Predicted STARzyme structure. (**A**) Secondary structure of Second-generation STARzyme with *B. subtilis tyrS* T-box (green) and circular permuted dFx flexizyme (blue). Specifier loop is designated with pink nucleotides. (**B**) 3D model of a $tRNA^{gly}_{GCC}$ (grey) being recognized by the *glyQS* T-box stem I (green; PDB: 4MGN[2]). Specifier loop (purple) base pairs with anticodon of the tRNA (cyan). The CCA tail of the tRNA forms a helix with the acceptor stem (red) of the Fx3 flexizyme (blue; PDB: 3CUL[6]).

The first-generation STARzyme was designed as a rational fusion of the *G. kaustophilus glyQS* T-box to the dFx flexizyme[1]. The *G. kaustophilus glyQS* T-box was initially chosen for its characterization *in vitro* and previous crystallization, which provides useful insights into the characteristics of the stem I specifier-loop regions[2, 22]. Initial fusion efforts resulted in a ribozyme that could recognize its cognate tRNA and exhibited an ability to aminoacylate bound and unbound tRNAs indiscriminately. Examination of the *glyQS* T-box[2] and the Fx3 flexizyme[6] crystal structures suggested that the failure to differentiate between tRNAs was due to an inability to bind both the T-box specifier loop and the 3′-CCA tail recognition sequence of the flexizyme. To overcome this, a circular permutation was implemented to reorient the dFx flexizyme's position with respect to the T-box, allowing for discrimination of cognate and noncognate tRNAs (**Figure 5.1B**). An essential quality of an aaRS-like ribozyme is its ability to be easily "reprogrammed" to recognize the anticodon of orthogonal tRNAs. However, recoding of the specifier loop of the *glyQS* T-box resulted in misfolded RNAs, preventing its use as a model system.

The *B. subtilis tyrS* T-box was chosen as a replacement for the *G. kaustophilus glyQS* T-box because the specifier loop of stem I has been shown to be reprogrammable for recognition of orthogonal tRNAs, such as the *M. jannaschii* tRNA (*Mj*tRNA). Replacement of the cognate specifier sequence with the amber codon (UAC→UAG) was effective in discriminating between the *Mj*tRNA$^{AMB}_{CUA}$ and the *Mj*tRNA$^{Tyr}_{GUA.}$ as observed by the appearance of aminoacylated-tRNAs on acid PAGE (**Figure 5.2A**). After six hours, ~35% of *Mj*tRNA$^{AMB}_{CUA}$ is charged by STARzyme in comparison to ~94% of charged *Mj*tRNA$^{Tyr}_{GUA}$ by dFx flexizyme alone. This suggests that while the addition of the T-box adds tRNA discrimination, it also reduces the aminoacylation rate of the dFx component of STARzyme. We tested this by performing a temperature study and observing the initial rates of aminoacylation (**Figure 5.2B**). While the initial rate at 37 °C is greater than the other temperatures tested, the similarities between the initial rates at the other temperatures (4, 12, and 23 °C) implies that aminoacylation at the 3′-OH is not the limiting step. We hypothesize

that the affinity of the T-box for the tRNA is too strong, regardless of the aminoacylation status of the bound tRNA.

We observed that STARzyme has preference for its cognate tRNA as defined by the specifier loop sequence (**Figure 5.2A**); however, we wanted to assess the ribozyme's ability to specifically charge *Mj*tRNA$^{AMB}_{CUA}$ in the presence of noncognate tRNAs. We tested if STARzyme could selectively charge the *Mj*tRNA$^{AMB}_{CUA}$ within a PURE *in vitro* transcription and translation (IVTT) system containing *E. coli* tRNAs[23]. A mutant dihydrofolate reductase (DHFR) was produced containing a P105X substitution and *in situ* aminoacylation was performed with the addition of STARzyme, *Mj*tRNA$^{AMB}_{CUA}$, and DBE-alkynyl amino acid. SDS-PAGE confirmed that full-length DHFR^{P105X} was produced in comparison to DHFR^{WT} (not shown). Mass spectrometry (MS) analysis of the purified protein confirmed that the alkynyl amino acid was incorporated at the 105 position (**Figure 5.3**).

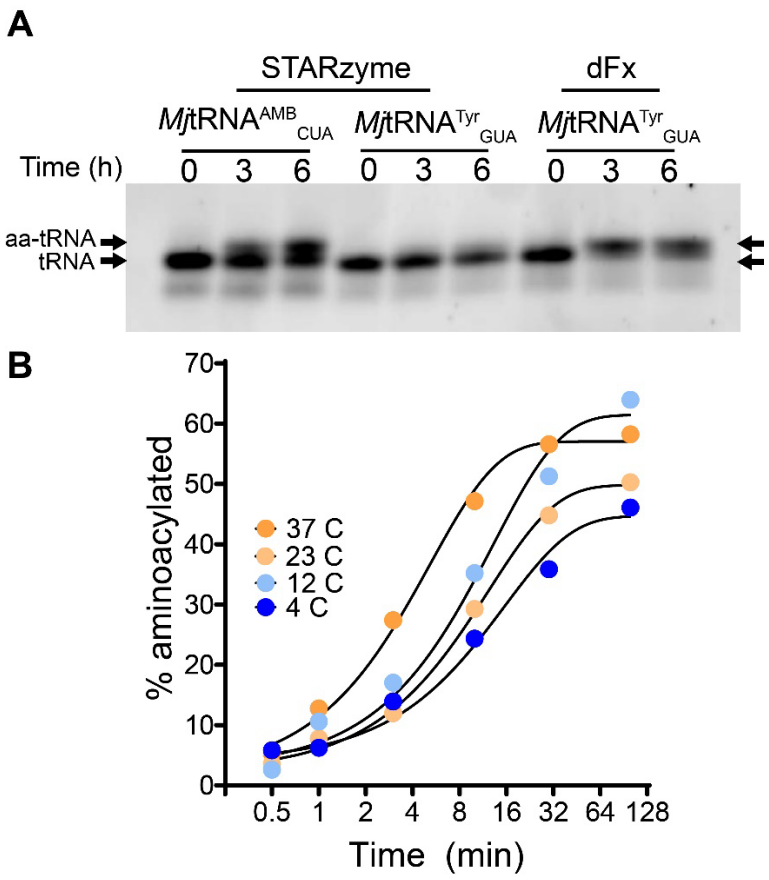


Figure 5.2: STARzyme specificity. (**A**) Amber specifying STARzyme recognizes the orthogonal *Mj*tRNA$^{AMB}_{CUA}$ and charges it (Gary Gan, Golden Lab, Purdue University). (**B**) STARzyme exhibits minimal temperature dependence for aminoacylation of *Mj*tRNA$^{AMB}_{CUA}$.

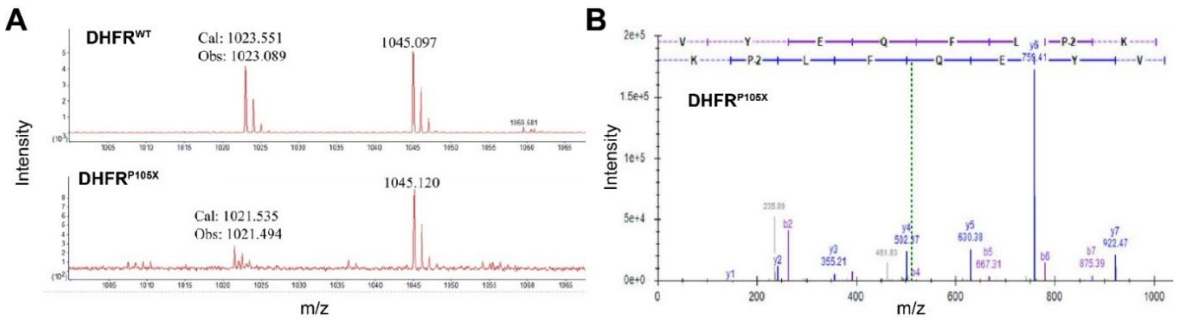


Figure 5.3 STARzyme functions in an IVTT reaction: Mass spectrometry of DHFR^{WT} and DHFR^{P105X}. (**A**) MALDI-TOF (**B**) LC-MS/MS of DHFR^{P105X} peak confirming incorporation of the alkynyl amino acid at position 105. Data generated by Gary Gan, Golden Lab, Purdue University.

The preliminary results suggest that the fusion and circular permutation of the *B. subtilis tyrS* T-box and dFx flexizyme provides adequate flexibility for both anticodon recognition and aminoacylation of a target tRNA. Moreover, the recoding of the specifier loop of the T-box modality allowed for use of the *Mj*tRNA$^{AMB}_{CUA}$. Aminoacylation assays demonstrate that STARzyme suffers from a lower-turnover rate compared to the dFx flexizyme alone, suggesting that further optimization is required for utility. In addition, STARzyme exhibited the ability to aminoacylate the *Mj*tRNA$^{AMB}_{CUA}$, leading to amber suppression and full-length protein production. These results provide the basis for further optimization of the STARzyme.

Using *in vitro* evolution to improve upon the current generation of STARzyme using amber suppression in firefly luciferase

The current generation of STARzyme was rationally designed as a fusion of two functional RNAs, so it is to be expected that the current rate of turnover of STARzyme is low. While the activity of STARzyme is low, we hypothesize that an *in vitro* selection using this generation of STARzyme will enrich for better ribozymes. The goal is to improve upon the turnover capacity of STARzyme by coupling ribozyme activity to translation, utilizing *in vitro* transcription and translation (IVTT) systems with *in vitro* selection. Traditional methods of ribozyme selection rely on affinity columns, filters, or gel supports for capturing catalytic events; however, these approaches prevent the detection of multi-turnover ribozymes because the RNAs are inherently evolved away due to selective pressure. These methods focus on capturing ribozymes by physical association or by covalent modification resulting from the activity, allowing for a direct pull-down of phenotype and genotype. Nevertheless, pull-down methods do not account for the release of the formed product, resulting in ribozymes with poor turnover. To account for the release of the product, a selective phenotype of the functional RNA must be co-localized with its DNA template. Gel immobilization techniques have been used to decrease the diffusion rate of individual DNA templates to aid in the visualization of co-localized PCR products[14, 24, 25], and have recently been

used in IVTT applications[26]. We believe that a gel-immobilization strategy for STARzyme selection will provide a robust approach to achieving turnover.

Our lab has used gel-immobilization to detect ATP production using luminescence output by firefly luciferase (Fluc)[27]. Using this approach, we have been able to detect single, light emitting Fluc enzymes immobilized within a gel matrix using a high-powered macroscope imaging system (**Figure 5.4A**). This imaging system consists of two macro lenses mounted in opposition of one another, coupled to an ultra-cooled electron-multiplying CCD (EMCCD) camera. Recently, our lab has been able to detect picomolar concentrations of ATP using this assay, quantifying its sensitivity[27].

Our selection strategy aims to utilize the EMCCD macroscope system alongside the bioluminescent properties of Fluc (**Figure 5.4B**). To select for aminoacylation activity, mutant Fluc genes containing amber (TAG) STOP codons at critical positions will be used to determine if successful aminoacylation, followed by incorporation of an amino acid, rescues enzymatic activity via the translation of a full-length enzyme product. The selection will seek to identify active STARzyme variants by bioluminescence within an IVTT system, immobilized by the addition of low-melting point agarose (or polyacrylamide). Due to the immobilization of the DNA template and RNA transcripts in the gel, functional STARzyme variants will be co-localized at the location of active Fluc. Long-exposure images taken with the macroscope will assist in pinpointing the location of STARzyme variants, facilitating their eventual excision from the gel. The excised DNA/RNA can then be amplified by RT-PCR to enrich for functional STARzyme variants which are carried on to successive rounds of selection (**Figure 5.4C**).

To couple aminoacylation activity and translation, we are utilizing the recombinant *E. coli* based PURE ΔRF1 *in vitro* transcription and translation system[28]. This system includes all the necessary factors for transcription and translation, including the essential initiation factors (IF1, IF2, and IF3), elongation factors (EF-G, EF-Tu, and EF-Ts) and two release factors (RF2 and RF3), as well as ribosomes, ribonucleotides, amino acids, tRNAs, and aaRSs. For our selection

to be successful, the nonessential release factor 1 (RF1) must be omitted as it recognizes the STOP codons UAG and UAA[29]. This allows us to utilize the orthogonal *Mj*tRNA$^{AMB}_{CUA}$ for amber codon suppression without premature mRNA release caused by RF1 competing for the A-site of the ribosome. Additionally, the PURE system takes advantage of T7 RNA polymerase, allowing for efficient RNA transcription from T7 promoter containing DNA.

Our pool will be designed such that the anticodon loop of the *B. subtilis tyrS* T-box, and the DBE and 3′-CCA tail recognition elements of dFx are maintained constant. These regions will be kept constant since the anticodon loop is required for discrimination of the cognate tRNA, while the DBE and 3′-CCA tail recognition elements of dFx are necessary for catalytic activity at the 3′ end of the tRNA. The rest of the pool will be synthesized with 9% degeneracy at each nucleotide position. By starting with a mutagenized STARzyme sequence, the diversity required to identify a functional ribozyme is reduced considerably, enabling us to start with ~10^{10} sequences[30, 31]. The selection pool will be flanked by constant regions necessary for transcription and RT-PCR. Additionally, the flanking regions will be used for next-generation sequencing for analysis of the pool. The constant regions will also contain restriction enzyme sites which will be used for cloning and the proposed *in vivo* analysis.

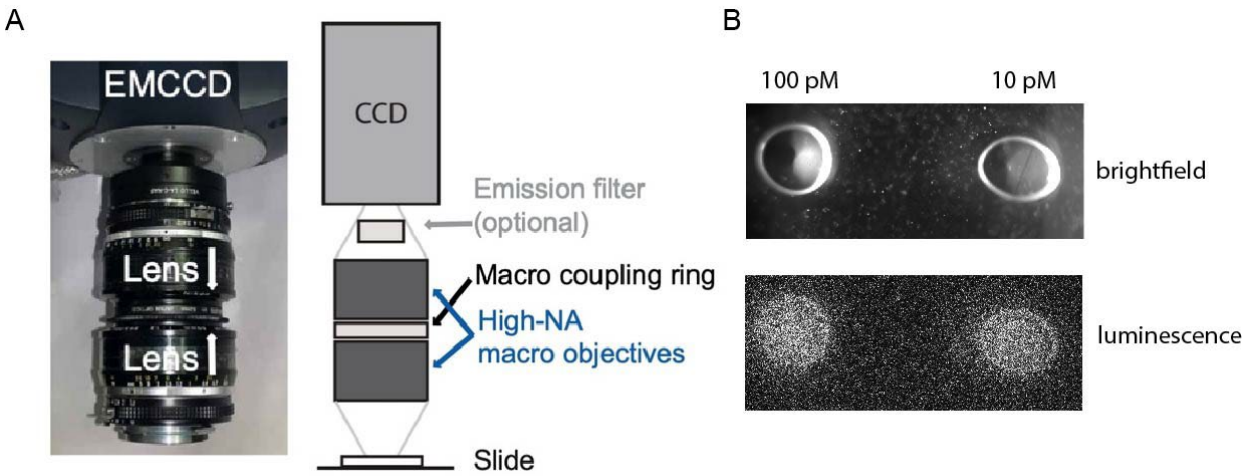


Figure 5.4: Macroscope imaging system (adapted from Chizzolini et al.[26]) and *in vitro* selection method (**A**) Electron-multiplying CCD (EMCCD) camera with opposing macro lenses allowing for 1:1 imaging of an area the size of the CCD sensor (15 mm x 15 mm). (**B**) Luminescence of PURE ΔRF1 reactions (1 µL) on glass cover slips and varying plasmid concentrations. (**C**) STARzyme DNA templates are transcribed and translated in gel immobilized IVTT on frame-sealed glass slides (blue border). Functional STARzyme will aminoacylate *Mj*tRNA$^{AMB}_{CUA}$. Charged tRNAs suppress amber STOP codons of mutant Fluc leading to functional proteins. Luminescence is used to locate regions containing functional Fluc enzymes and their co-localized DNA/RNA templates for excision and RT-PCR.

Previous works have suggested that G200X (**Figure 5.5A**), an essential glycine in the ATP binding center of Fluc, significantly reduces luminescence, and we have confirmed this within a PURE ΔRF1 system compared to $Fluc^{WT}$ (**Figure 5.5B**)[32]. We have also shown that this mutation can be suppressed when $tRNA^{AMB}_{CUA}$ is charged with glycine (**Figure 5.5C**). There is detectable luminescence of $Fluc^{G200X}$ in PURE ΔRF1 (not shown), which suggests that in the absence of RF1, the ribosome stalls at the amber STOP codon until a charged tRNA misincorporates a canonical amino acid[29, 33]. Others have shown that tyrosine, phenylalanine, tryptophan, glutamine, and glutamate are commonly incorporated at amber codons due to their tRNA anticodon similarities[29, 33-36]. One possible solution to reduce background luminescence resulting from misincorporation is to increase the number of amber codons present in Fluc. Hong et al. illustrated that increasing the number of amber codons in a cell-free translation system significantly reduces the full-length protein product formed[36]. We plan to use site-directed mutagenesis to introduce additional amber mutations at G203 and G207 within the ATP-binding pocket as a solution to the background luminescence stemming from misincorporation.

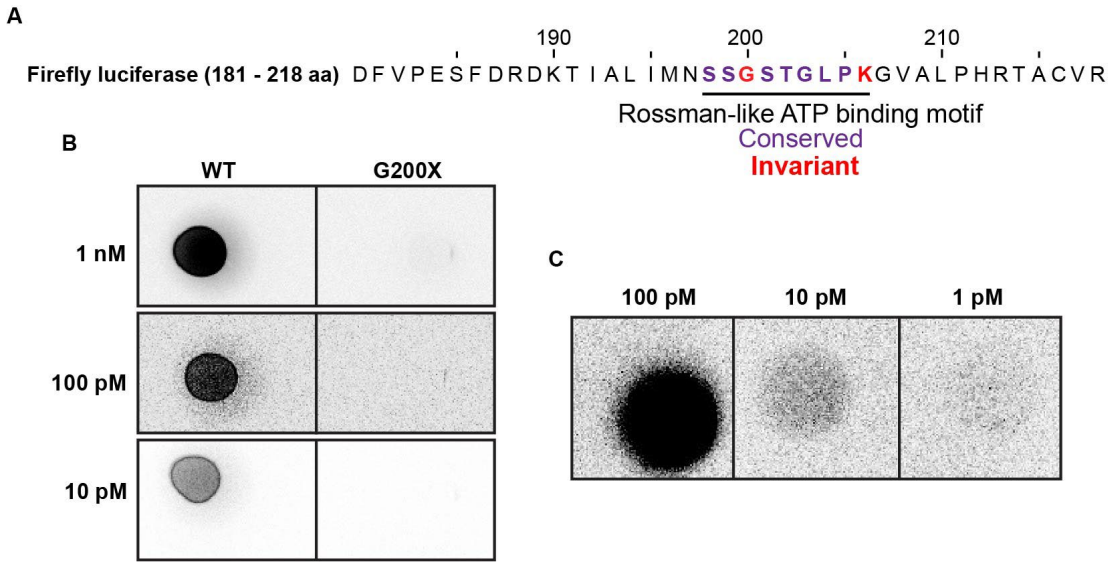


Figure 5.5 Fluc^{G200X} is a non-functional mutant: Fluc^{G200X} mutant significantly reduces background luminescence in a PURE ΔRF1 system. (**A**) G200 of the ATP-binding pocket of Fluc is an invariant position among adenylate-forming enzymes. (**B**) Fluc^{WT} and Fluc^{G200X} plasmids at varying concentrations were supplemented into a PURE ΔRF1 reaction and incubated for 1.5 hours at 37 °C. Reactions were sandwiched between glass-cover slips using a BioRad FrameSeal at 1 µL volumes. (**C**) Approximately 6.5 µM of dFx charged *Mj*tRNA$^{AMB}_{CUA}$ was incubated in a PURE ΔRF1 reaction with varying G200X plasmid concentrations for 1.5 hours at 37 °C.

Water-in-oil based selection as an alternative approach with amber suppressed DNA methyl transferase

As mentioned previously, *in situ* charging of *Mj*tRNA$^{AMB}_{CUA}$ in PURE ΔRF1 by STARzyme was confirmed via mass spectrometry (MS) by our collaborators in the Golden lab at Purdue University (Figure 5.3). However, recapitulation of *in situ* charging within the PURE ΔRF1 system utilizing Fluc^{G200X} and the macroscope system has not yet been confirmed. Our collaborators optimized their IVTT system and found that the addition of STARzyme may inhibit translation but could be rescued with Mg^{2+} and ATP. The addition of Mg^{2+} and ATP did not rescue translation in the Fluc system. One possible hypothesis for this effect is that using purified, charged tRNAs in IVTT can be inhibitory because of Mg^{2+} cation carryover, first reported by Chen and colleagues[37]. Their solution was to purify the aminoacylated tRNAs by anion-exchange chromatography for complete cation exchange of the purified RNAs with K^+. Alternatively, the STARzyme and *Mj*tRNA$^{AMB}_{CUA}$ can be transcribed *in vitro* by supplementing the system with DNA templates. Synthesis of orthogonal tRNAs *in vitro* has been shown to enhance the incorporation of ncAAs when utilizing amber suppression[36]. To address the issue of *in situ* charging, we plan to test anion-exchange purified STARzyme and *Mj*tRNA$^{AMB}_{CUA}$, as well as test *in vitro* transcribed STARzyme and *Mj*tRNA$^{AMB}_{CUA}$ within the PURE ΔRF1 system.

While mutagenesis of the current STARzyme provides an ideal basis for the selection of a higher activity variant, the information content may be too low to yield a ribozyme with the catalytic capacity we seek [30, 38]. This could require modifications to the pool design such as completely randomizing the linker region used for circular permutation of dFx to allow for higher flexibility. If the off-rate remains an issue, a molecular thermometer sequence could be introduced between the T-box Stem I and dFx to assist in releasing the charged tRNA product[39].

The main pitfall of this selection strategy is the potential for high-background luminescence of amber-containing Fluc enzymes in a ΔRF1 system. If the background luminescence is too high, then the identification of active STARzyme clusters may prove to be difficult, especially if the

number of active variants is low. Ideally, one functional STARzyme will yield more than one functional Fluc enzyme; however, high background luminescence may lead to false-positive extractions.

Our alternative methodology for the *in vitro* selection of STARzyme is to use a water-in-oil (WO) emulsion approach to select for aminoacylation activity. WO emulsions are a proven method of co-localization and have been used for successful *in vitro* selections of both RNA and protein enzymes, including orthogonal tRNA/aaRS pairs[40-45]. Tawfik and Griffiths classically demonstrated the utility of WO emulsions for the *in vitro* selection of a functional DNA methyltransferase, M.*Hae*III, from a pool of approximately 10^7 DNAs[41]. Using a similar approach, we aim to utilize a WO emulsion system where each droplet contains PURE ΔRF1, a single DNA template for STARzyme, DBE amino acids, *Mj*tRNA$^{AMB}_{CUA}$, and a M.*Hae*III gene containing amber mutations. Successful aminoacylation and amber suppression leads to a functional M.*Hae*III (**Figure 5.6A**). The DNA pool will be designed to contain multiple *Hae*III restriction enzyme sites (GGCC) in the 5′ constant region upstream from the T7 promoter. Successfully translated and functional M.*Hae*III enzymes can methylate these restriction sites, protecting the DNA template within the compartment from degradation digestion. This allows us to biotinylate the 5′ end of the pool for affinity chromatography with streptavidin and allows the use of anti-5-methylcytosine antibodies for pull-down purification. Thus far, we have illustrated that the amber-suppression of the G9X mutation of M.*Hae*III produces full-length, functional enzymes which methylate and protect the DNA template from degradation by *Hae*III (**Figure 5.6B**). Preliminary data suggest that this approach will be useful as an alternative method of selection.

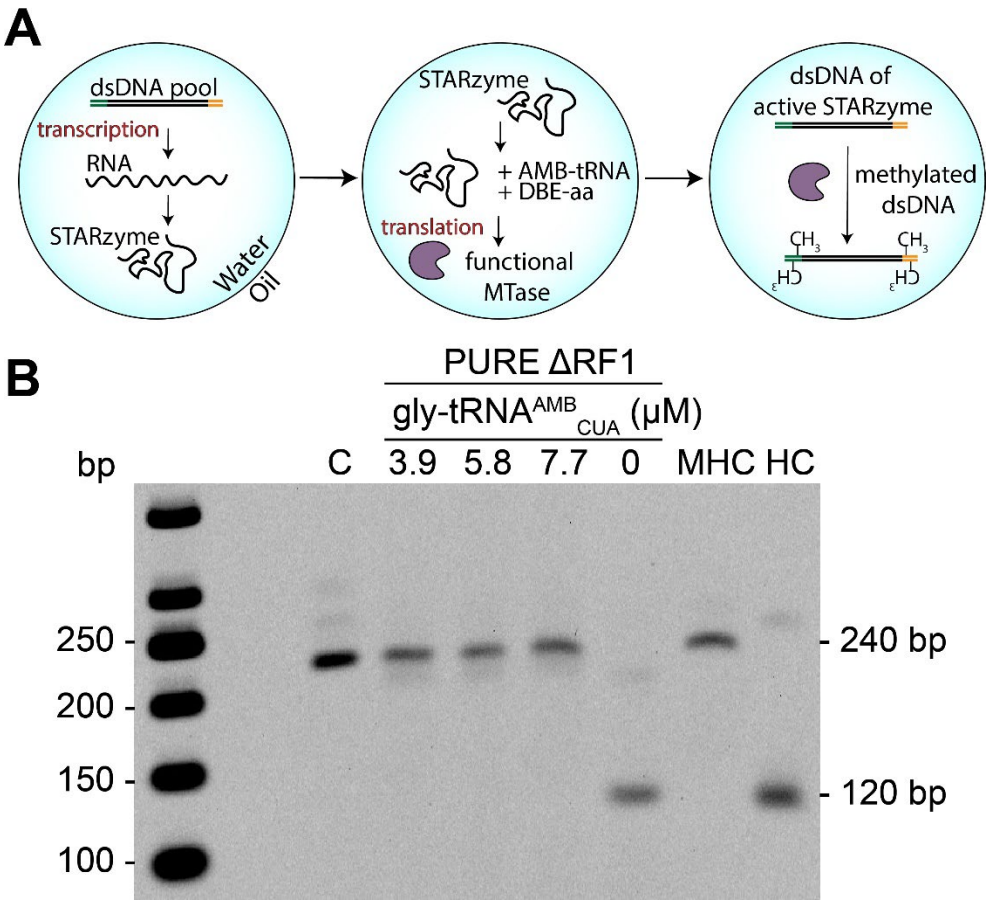


Figure 5.6: Emulsion selection scheme (**A**) WO selection scheme. The aqueous layer is made of the PURE ΔRF1 IVTT system. A mutagenized STARzyme pool is introduced into the emulsion alongside the *Mj*tRNA$^{AMB}_{CUA}$, dinitrobenzyl-amino acid substrate (DBE-aa), and the M.*Hae*III^{G9X}. Transcribed, active STARzyme will aminoacylate the orthogonal tRNA, which is then incorporated leading to a functional M.*Hae*III. The active M.*Hae*III recognizes *Hae*III restriction sites on the template DNA, protecting the DNA template from cleavage. (**B**) Amber-suppression of M.*Hae*III^{G9X} by introduction of varying concentrations of gly-tRNA$^{AMB}_{CUA}$ into a PURE ΔRF1 reaction (not emulsified). Reactions were incubated for 2 hours at 37 °C. Bands indicate protection/digestion of 5′ FAM-STARzyme on 2.5% agarose (C: control 5′ FAM-STARzyme, MHC: M.*Hae*III + *Hae*III treatment of STARzyme, HC : *Hae*III digestion of STARzyme).

Introduction of amino acids into proteins *in vivo* using an orthogonal tRNA aminoacylated by next-generation STARzymes

Potential candidate STARzyme variants, identified through successful *in vitro* selection, will be accessed for their ability to aminoacylate *Mj*tRNA$^{AMB}_{CUA}$ in *E. coli.* We hypothesize that *in vitro* selected STARzyme sequences will exhibit *in vivo* aminoacylation activity, leading to amino acid incorporation and full-length protein production. We will use a firefly luciferase (Fluc) reporter assay to measure luminescence as an output for suppression of amber STOP codons within the Fluc gene. Furthermore, successful amino acid incorporation will be validated using mass spectrometry (MS). The development of this workflow is critical to both determining the efficiency of individual STARzyme variants to incorporate canonical amino acids, but also to test the incorporation of ncAAs in future applications.

In vivo aminoacylation can be assessed by two methods: 1) time-course luminescence measurements and 2) amino acid incorporation analysis by MS. The Fluc reporter assay is a classical and proven method of measuring expression levels[46]. Our lab has used bioluminescence to measure translation kinetics by photoriboswitch regulation, providing us with knowledge and experience in the practical applications of the assay[47]. An MS pipeline will be used to assess the incorporation of target amino acids at defined positions in Fluc. MS analysis will be performed at the University of California Irvine, Mass Spectrometry Facility. We envision using this system to simultaneously screen multiple STARzyme clones identified in Specific Aim 1 by luminescence followed by MS analysis of the expressed luciferase (**Figure 5.7A**). This system will be used to test the aminoacylation of STARzyme with DBE canonical amino acids as well as ncAAs (e.g., cyclopropenone amino acid[48]), and to assess the amino acid incorporation success rate *in vivo* for each STARzyme clone.

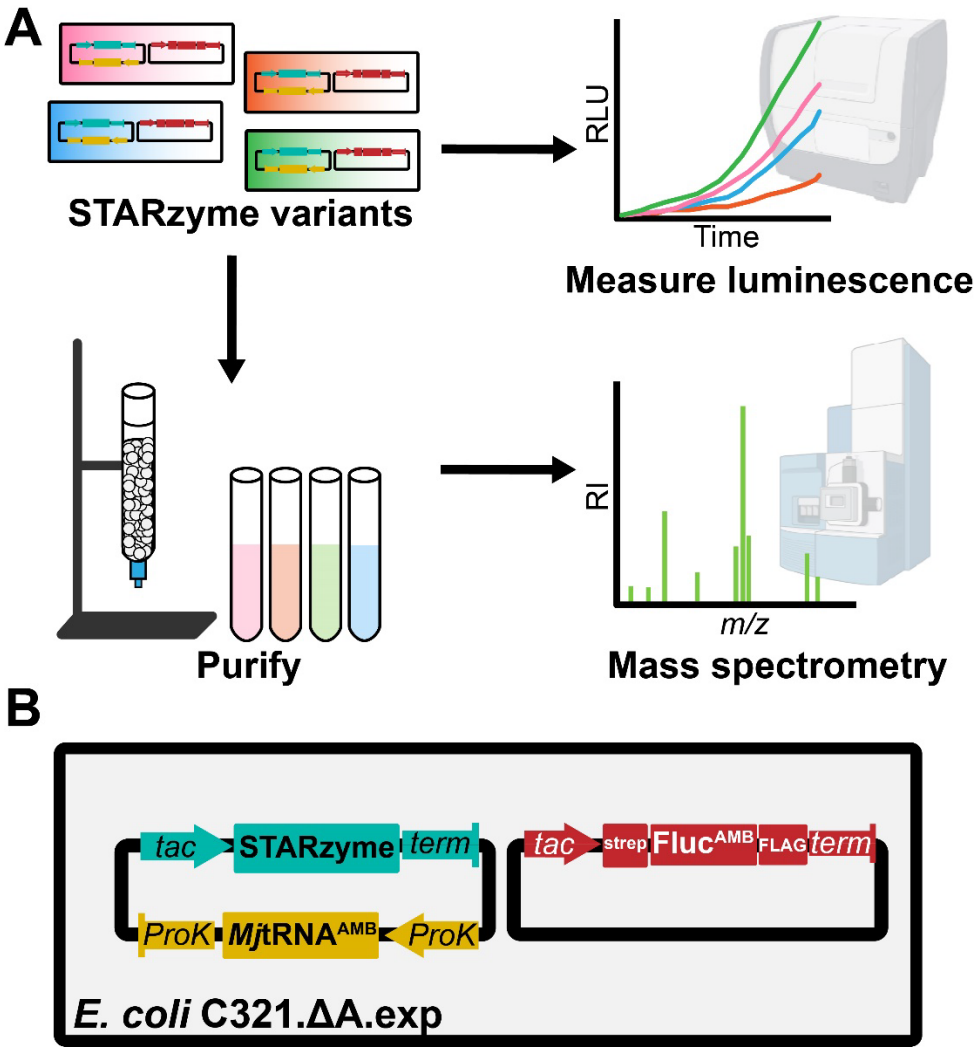


Figure 5.7: *In vivo* analysis approach. (**A**) Candidate STARzymes are cloned into an expression plasmid for luminescence measurements. Amino acid incorporation will be determined by purification and subsequent mass spectrometry analysis. (**B**) Proposed two plasmid system for testing STARzyme activity *in vivo*. Fluc will be used as an output for aminoacylation activity in the amber-less C321.ΔA.exp *E. coli* strain. Two epitope tags flank Fluc for protein purification.

Our approach will use the C321.ΔA.exp *E. coli* strain deposited on Addgene by the George Church lab at Harvard University[49]. This strain has been evolved to omit all UAG STOP codons by replacing all instances of UAG with UAA. In addition, the C321.ΔA.exp strain lacks release factor 1 (RF1) eliminating potential competition within the A-site of the ribosome. The utility in this strain lies in the fact that charged orthogonal tRNAs are less likely to misincorporate into actively translating endogenous proteins, which proves useful in measuring aminoacylation of *Mj*tRNA$^{AMB}_{CUA}$ by candidate STARzymes. A two-plasmid system will be used for the expression of the STARzyme/tRNA pair and Fluc (**Figure 5.7B**). A plasmid containing Fluc amber mutants will be transformed into the C321.ΔA.exp strain and maintained. A second plasmid carrying the *Mj*tRNA$^{AMB}_{CUA}$ and the cloned STARzyme variant to be analyzed will be transformed prior to performing individual experiments. Transcriptional regulation and processing of *Mj*tRNA$^{AMB}_{CUA}$ will be maintained under the control of the endogenous *E. coli* tRNA$^{Pro}_{CCG}$ codon (*proK*) which has been used successfully to transcribe and process orthogonal tRNAs *in vivo*[45, 50, 51]. The *proK* codon will be used as it regulates Rho-independent transcription and maintenance of the most common proline codon (CCG) in *E. coli,* ensuring efficient transcription of our orthogonal tRNA[52]. STARzyme will be maintained under the control of an inducible repressor such as *tac* or *araBAD* for tightly regulated transcription. Fluc will be under the same promoter system as STARzyme to allow for induction by the addition of a single metabolite. The induction of both STARzyme and Fluc allows for a direct readout of aminoacylation activity due to photon output by Fluc, providing a more quantitative assessment of STARzyme activity. However, this inducible system can be modified to utilize a dual-inducible codon system in the future to fine-tune both STARzyme and protein expression[53].

The Fluc expression system will be designed such that the N- and C-termini of Fluc will be flanked by epitope tags (Strep-tag II[54] and FLAG) for convenient purification prior to assessing amino acid incorporation by MS (**Figure 5.3**). The validation of amino acid incorporation will require high-resolution MS methods to better discriminate individual amino acid mass differences.

A bottom-up (trypsin digested) approach using liquid chromatography-tandem mass spectrometry (LC-MS/MS) will be used to discriminate mass peaks in two dimensions more efficiently[55]. Ionized peptides identified by abundance in the first dimension (MS1) will be fragmented via electron transfer dissociation (ETD) for further separation and identification of side chains (MS2) [56]. A bioinformatics approach will be used for the analysis and the deconvolution of the MS data[57]. Peptide sequence information will be obtained using conventional software packages such as SEQUEST, Mascot, and X!Tandem. The overall abundance of amino acid incorporation (correctly and incorrectly incorporated) will be assessed using the VIPER and MaxLFQ suites[58-62]. The expected outcomes of the MS data analysis are the quantification of amino acid incorporation at amber codons for both correctly incorporated and misincorporated amino acids. This information will be combined with Fluc assay information for the evaluation of candidate STARzyme performance *in vivo*.

The most notable pitfall of this assay is if STARzyme variants identified in from SELEX do not have activity *in vivo*. A lack of activity *in vivo* would suggest that IVTT conditions are not sufficient to select for aminoacylation that can be replicated in *E. coli.* Factors such as ionic strength or the presence of regulatory proteins may have inhibitory properties within the cell[63]. To enhance *in vivo* activity, an additional selection can be performed utilizing the two-plasmid system in C321.ΔA.exp cells. An established method of *in vivo* selection of activity is compartmentalized partnered replication (CPR), a WO emulsion approach previously used to select for tRNA/synthetase pairs and DNA polymerases[45, 64]. Selection pools based on active STARzyme sequences identified in Specific Aim 1 will be cloned into the STARzyme/tRNA expression plasmid, while Fluc will be replaced by KlenTaq DNA polymerase containing amber STOP codons[65]. Successful aminoacylation by active STARzymes allows for amber suppression of D610X and D785X mutations within the catalytic core of the polymerase leading to a full-length functional enzyme being expressed within *E. coli*[66]. After overnight culture, *E. coli* cells are compartmentalized in WO emulsion and subsequently broken open by heat cycling. The presence

of KlenTaq, dNTPs, and primers in the WO emulsion buffer allows for amplification of the functional STARzyme template. The CPR method provides a robust strategy for further evolution of *in vitro* functional STARzymes for *in vivo* activity.

Alternative approaches can be used for the measurement of *in vivo* activity. A dual reporter system using red fluorescent protein (RFP) and green fluorescent protein (GFP) driven by the same promoter can be used in the two-plasmid system approach described above. RFP emission is used as a measurement of the baseline of expression while GFP contains STOP codons to report functional amino acid incorporation. A more common approach is to use a selectable marker. A commonly used selectable marker for amino acid incorporation is the *β*-lactamase assay, where amber mutations are introduced into the *β*-lactamase gene, thus inhibiting resistance to lactam antibiotics such as ampicillin[45, 67, 68].

5.4 Conclusions and Perspectives

Overall, this project aimed to evolve a model ribozyme with selective aminoacylation activity for improved catalytic activity both *in vitro* and *in vivo*. The initial STARzyme demonstrated an ability to specifically aminoacylate a cognate tRNA. By coupling STARzyme's aminoacylation activity with translation, we are confident that *in vitro* selection can identify a ribozyme with improved catalytic turnover. The functionality of a ribozyme with these capabilities would prove useful in the fields of genetic code expansion which lends itself to being valuable in pharmaceuticals, macromolecular imaging, and general synthetic biology applications.

5.5 Materials and Methods

Transcription reactions

RNAs were transcribed from *Bsa*I linearized plasmids in a 1X transcription buffer (40 mM Tris-HCl [pH 7.5] and 2 mM spermidine), 30 mM $MgCl_2$, 3 mM rNTPs, 10 mM dithiothreitol (DTT), and recombinant T7 RNA polymerase for 3-5 hours. After transcription, RNAs were ethanol

precipitated by addition of 3 M KCl to a final concentration of 200 mM KCl, glycogen, and 2.5 times volume 100% cold ethanol at -80 °C for 1 hour. RNAs were pelleted by centrifugation (20,000 x*g* for 20 minutes) and washed in 70% cold ethanol. Pellets were resuspended in 10 mM Tris-HCl, pH 7.5, and 0.1 mM EDTA. Pellets were resuspended in 10 mM Tris-HCl, pH 7.5, and 0.1 mM EDTA and then kit purified by an RNA Clean and Concentrator kit (Zymo Research). RNA was quantified via UV-vis spectroscopy on a NanoDrop 2000 (ThermoScientific).

Preparation of radiolabeled *Mj*tRNA$^{AMB}_{CUA}$

Radiolabeled tRNAs were transcribed in 1X transcription buffer, 16 mM $MgCl_2$, 2 mM GTP, 2 mM CTP, 2 mM UTP, 0.2 mM ATP, [α-^{32}P] ATP, 10 mM DTT and recombinant T7 RNA polymerase. After 1 hour of transcription at 37 °C, the thermocycler was cycled between 55 °C and 37 °C in 5-minute intervals (over 30 minutes) to promote cleavage of the hammerhead ribozyme/tRNA transcripts. Transcriptions were purified via denaturing PAGE using phosphorimaging and a GE Typhoon imager. Bands corresponding to the cleaved tRNA product were excised and eluted into 300 mM KCl and precipitated in 2.5 times volume 100% cold ethanol at -80 °C for 1 hour. RNAs were pelleted by centrifugation (20,000 x*g* for 20 minutes) and washed in 70% cold ethanol. Pellets were resuspended in 10 mM Tris-HCl, pH 7.5, and 0.1 mM EDTA.

Measurement of aminoacylation kinetics by anion-exchange thin layer chromatography (TLC)

RNAs were refolded in a refolding buffer containing 30 mM HEPES, pH 8, and 30 mM KCl. RNAs were refolded by first heating and holding at 95 °C for 1 min, then slowly cooling to 37 °C over 30 minutes. Following incubation at 37 °C, $MgCl_2$ was added to a final concentration of 15 mM. RNA was then allowed to cool to room temperature over 10 minutes in the presence of $MgCl_2$. Radiolabeled *Mj*tRNA$^{AMB}_{CUA}$ was incubated with varying concentration of STARzyme RNA in the presence of 5 mM DBE-glycine amino acid in DMSO and incubated at varying time points and temperatures. Reactions were stopped via buffer exchange through G-10 Sephadex beads pre-

equilibrated in 50 mM NaOAc pH 5.5. Reactions were then incubated with 100 units of nuclease P1 (New England Biolabs) for 15 minutes at room temperature. Reactions were spotted (2 µL) on pre-washed (ddH_2O) and dried (air) PEI-cellulose TLC plates. Separation was done in a running buffer consisting of glacial acetic acid/1 M LiCl/ddH_2O (5:10:85) until the solvent front was near the end of the plate. The plate was then air dried and a phosphorimaging screen was exposed for at least 24 hours depending on the number of counts. The amount of aminoacylation was determined as a percentage of charged and uncharged using densitometry while considering the number of labeled adenosine that were possibly radiolabeled in the tRNA.

Amber suppression of M.*Hae*III^{G9X} and Fluc^{G200X} in IVTT

Charging of purified *Mj*tRNA$^{AMB}_{CUA}$ was done using either STARzyme or the dFx flexizyme. Charging reactions were performed in refolding buffer with 15 mM $MgCl_2$ (STARzyme) or 100 mM $MgCl_2$ (dFx flexizyme) and 5 mM DBE-glycine. Aminoacylation by STARzyme can be carried out at various temperatures, but for efficient charging and yield, charging for both STARzyme and dFx flexizyme were carried out at 4 °C for 10-16 hours. Charged tRNAs were precipitated in 300 mM NaOAc pH 5.5 and 100% cold ethanol to reduce hydrolysis of the aminoacylate. PureExpress ΔRF1 was used for IVTT. Reactions were carried out following manufacturer's recommendations but supplemented with varying amounts of gly-tRNA for amber suppression of the respective gene (M.*Hae*III^{G9X} or Fluc^{G200X}). Reactions were incubated at 37 °C for 2 hours.

For M.*Hae*III^{G9X} reactions, a 5′-fluorescein labeled DNA template was used to determine the methylation status of the respective DNA template, following digestion by the restriction enzyme *Hae*III. The IVTT reactions were desalted with G-10 Sephadex pre-equilibrated with 10 mM Tris-HCl, pH 7.5, and 0.1 mM EDTA and a dye-free loading dye containing 20% glycerol was added. Reactions were then run on a 3% agarose gel containing a 50 bp ladder (New England Biolabs) loaded with 0.1% SYBR Gold for visualization. The gel was imaged on a GE Typhoon imager using the 488/532 BP filter set.

Fluc^{G200X} reactions were visualized using an Oxford Instruments Andor iXon Ultra 888 EMCCD camera with 300 EM for 5 minutes, following the addition of 1 mM ATP and 100 μM D-luciferin.

5.6 References

1. Golden, B.L., Chen, J., Luptak, A. Gan, R., Edn. C12N 15/113 (20100101); C12P 21/00 (20060101) (Purdue Research Foundation, The Regents of the University of California, United States of America; 2018).

2. Grigg, J.C. & Ke, A. Structural determinants for geometry and information decoding of tRNA by T box leader RNA. *Structure* **21**, 2025-2032 (2013).

3. Zhang, J. & Ferré-D'Amaré, A.R. Structure and mechanism of the T-box riboswitches. *Wiley Interdiscip Rev RNA* **6**, 419-433 (2015).

4. Murakami, H., Ohta, A., Goto, Y., Sako, Y. & Suga, H. Flexizyme as a versatile tRNA acylation catalyst and the application for translation. **50**, 35-36 (2006).

5. Fujino, T., Kondo, T., Suga, H. & Murakami, H. Exploring the Minimal RNA Substrate of Flexizymes. *Chembiochem* **20**, 1959-1965 (2019).

6. Xiao, H., Murakami, H., Suga, H. & Ferré-D'Amaré, A.R. Structural basis of specific tRNA aminoacylation by a small in vitro selected ribozyme. *Nature* **454**, 358-361 (2008).

7. Wang, L., Brock, A., Herberich, B. & Schultz, P.G. Expanding the Genetic Code of Escherichia coli. **292**, 498-500 (2001).

8. Müller, M.M. Post-Translational Modifications of Protein Backbones: Unique Functions, Mechanisms, and Challenges. *Biochemistry* **57**, 177-185 (2018).

9. Millar, A.H. et al. The Scope, Functions, and Dynamics of Posttranslational Protein Modifications. *Annual Review of Plant Biology* **70**, 119-151 (2019).

10. Chin, J.W. Expanding and reprogramming the genetic code. *Nature* **550**, 53-60 (2017).

11. Liu, C.C. & Schultz, P.G. Adding New Chemistries to the Genetic Code. *Annual Review of Biochemistry* **79**, 413-444 (2010).

12. Noren, C.J., Anthony-Cahill, S.J., Griffith, M.C. & Schultz, P.G. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science* **244**, 182-188 (1989).

13. Chin, J.W. Expanding and Reprogramming the Genetic Code of Cells and Animals. *Annual Review of Biochemistry* **83**, 379-408 (2014).

14. Liu, Y. et al. Digital quantification of gene methylation in stool DNA by emulsion-PCR coupled with hydrogel immobilized bead-array. *Biosensors and Bioelectronics* **92**, 596-601 (2017).

15. Morimoto, J., Hayashi, Y., Iwasaki, K. & Suga, H. Flexizymes: Their Evolutionary History and the Origin of Catalytic Function. **44**, 1359-1368 (2011).

16. Ishida, S., Terasaka, N., Katoh, T. & Suga, H. An aminoacylation ribozyme evolved from a natural tRNA-sensing T-box riboswitch. *Nature Chemical Biology* **16**, 702-709 (2020).

17. Grundy, F.J. & Henkin, T.M. tRNA as a positive regulator of transcription antitermination in B. subtilis. *Cell* **74**, 475-482 (1993).

18. Zhang, J. & Ferré-D'Amaré, A.R. Co-crystal structure of a T-box riboswitch stem I domain in complex with its cognate tRNA. *Nature* **500**, 363-366 (2013).

19. Robertson, W.E. et al. Sense codon reassignment enables viral resistance and encoded polymer synthesis. *Science* **372**, 1057-1062 (2021).

20. Lee, J. et al. Expanding the limits of the second genetic code with ribozymes. *Nature Communications* **10**, 5097 (2019).

21. Arranz-Gibert, P., Vanderschuren, K. & Isaacs, F.J. Next-generation genetic code expansion. *Current Opinion in Chemical Biology* **46**, 203-211 (2018).

22. Grigg, J.C. et al. T box RNA decodes both the information content and geometry of tRNA to affect gene expression. **110**, 7240-7245 (2013).

23. Shimizu, Y. et al. Cell-free translation reconstituted with purified components. *Nature Biotechnology* **19**, 751-755 (2001).

24. Mitra, R.D. & Church, G.M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res* **27**, e34 (1999).

25. Chetverina, H.V., Samatov, T.R., Ugarov, V.I. & Chetverin, A.B. Molecular colony diagnostics: detection and quantitation of viral nucleic acids by in-gel PCR. *Biotechniques* **33**, 150-152, 154, 156 (2002).

26. Ouyang, X., Zhou, X., Lai, S.N., Liu, Q. & Zheng, B. Immobilization of Proteins of Cell Extract to Hydrogel Networks Enhances the Longevity of Cell-Free Protein Synthesis and Supports Gene Networks. *ACS Synthetic Biology* **10**, 749-755 (2021).

27. Chizzolini, F., Kent, A.D., Passalacqua, L.F.M. & Lupták, A. Enzymatic RNA Production from NTPs Synthesized from Nucleosides and Trimetaphosphate*. *Chembiochem* **22**, 2098-2101 (2021).

28. Shimizu, Y. et al. Cell-free translation reconstituted with purified components. *Nat Biotechnol* **19**, 751-755 (2001).

29. Johnson, D.B.F. et al. Release Factor One Is Nonessential in Escherichia coli. *ACS Chemical Biology* **7**, 1337-1344 (2012).

30. Wedel, A.B. Fishing the best pool for novel ribozymes. *Trends in Biotechnology* **14**, 459-465 (1996).

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

32. Conti, E., Franks, N.P. & Brick, P. Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure* **4**, 287-298 (1996).

33. Young, T.S., Ahmad, I., Yin, J.A. & Schultz, P.G. An Enhanced System for Unnatural Amino Acid Mutagenesis in E. coli. *Journal of Molecular Biology* **395**, 361-374 (2010).

34. Park, H.S. et al. Expanding the genetic code of Escherichia coli with phosphoserine. *Science* **333**, 1151-1154 (2011).

35. Ogawa, A., Hayami, M., Sando, S. & Aoyama, Y. A Concept for Selection of Codon-Suppressor tRNAs Based on Read-Through Ribosome Display in an In Vitro Compartmentalized Cell-Free Translation System. *J Nucleic Acids* **2012**, 538129 (2012).

36. Hong, S.H. et al. Cell-free protein synthesis from a release factor 1 deficient Escherichia coli activates efficient and multiple site-specific nonstandard amino acid incorporation. *ACS Synth Biol* **3**, 398-409 (2014).

37. Chen, J., Chen, M. & Zhu, T.F. Translating protein enzymes without aminoacyl-tRNA synthetases. *Chem* **7**, 786-798 (2021).

38. Carothers, J.M., Oestreich, S.C., Davis, J.H. & Szostak, J.W. Informational Complexity and Functional Activity of RNA Structures. *Journal of the American Chemical Society* **126**, 5130-5137 (2004).

39. Sadler, F.W., Dodevski, I. & Sarkar, C.A. RNA Thermometers for the PURExpress System. *ACS Synthetic Biology* **7**, 292-296 (2018).

40. Ghadessy, F.J., Ong, J.L. & Holliger, P. Directed evolution of polymerase function by compartmentalized self-replication. *Proceedings of the National Academy of Sciences* **98**, 4552-4557 (2001).

41. Tawfik, D.S. & Griffiths, A.D. Man-made cell-like compartments for molecular evolution. *Nature Biotechnology* **16**, 652-656 (1998).

42. Abil, Z., Ellefson, J.W., Gollihar, J.D., Watkins, E. & Ellington, A.D. Compartmentalized partnered replication for the directed evolution of genetic parts and circuits. *Nature Protocols* **12**, 2493-2512 (2017).

43. Griffiths, A.D. & Tawfik, D.S. Directed evolution of an extremely fast phosphotriesterase by in vitro compartmentalization. *EMBO J* **22**, 24-35 (2003).

44. Ellefson, J.W. et al. Directed evolution of genetic parts and circuits by compartmentalized partnered replication. *Nature Biotechnology* **32**, 97-101 (2014).

45. Maranhao, A.C. & Ellington, A.D. Evolving Orthogonal Suppressor tRNAs To Incorporate Modified Amino Acids. *ACS Synthetic Biology* **6**, 108-119 (2017).

46. Syed, A.J. & Anderson, J.C. Applications of bioluminescence in biotechnology and beyond. *Chem Soc Rev* **50**, 5668-5705 (2021).

47. Rotstan, K.A. et al. Regulation of mRNA translation by a photoriboswitch. *eLife* **9**, e51737 (2020).

48. Row, R.D., Nguyen, S.S., Ferreira, A.J. & Prescher, J.A. Chemically triggered crosslinking with bioorthogonal cyclopropenones. *Chemical Communications* **56**, 10883-10886 (2020).

49. Lajoie, M.J. et al. Genomically recoded organisms expand biological functions. *Science (New York, N.Y.)* **342**, 357-360 (2013).

50. Ryu, Y. & Schultz, P.G. Efficient incorporation of unnatural amino acids into proteins in Escherichia coli. *Nature Methods* **3**, 263-265 (2006).

51. Biddle, W., Schmitt, M.A. & Fisk, J.D. Evaluating Sense Codon Reassignment with a Simple Fluorescence Screen. *Biochemistry* **54**, 7355-7364 (2015).

52. Nakamura, Y., Gojobori, T. & Ikemura, T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic acids research* **28**, 292-292 (2000).

53. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic acids research* **25**, 1203-1210 (1997).

54. Schmidt, T.G.M. & Skerra, A. The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nature Protocols* **2**, 1528-1535 (2007).

55. Karpievitch, Y.V., Polpitiya, A.D., Anderson, G.A., Smith, R.D. & Dabney, A.R. Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects. *Ann Appl Stat* **4**, 1797-1823 (2010).

56. Alley, W.R., Jr., Mechref, Y. & Novotny, M.V. Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data. *Rapid Commun Mass Spectrom* **23**, 161-170 (2009).

57. Chen, C., Hou, J., Tanner, J.J. & Cheng, J. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *International journal of molecular sciences* **21**, 2873 (2020).

58. Brodbelt, J.S. & Russell, D.H. Focus on the 20-year anniversary of SEQUEST. *J Am Soc Mass Spectrom* **26**, 1797-1798 (2015).

59. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567 (1999).

60. Bjornson, R.D. et al. X!!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers. *J Proteome Res* **7**, 293-299 (2008).

61. Monroe, M.E. et al. VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics* **23**, 2021-2023 (2007).

62. Cox, J. et al. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ*. *Molecular & Cellular Proteomics* **13**, 2513-2526 (2014).

63. Cochella, L. & Green, R. Isolation of antibiotic resistance mutations in the rRNA by using an in vitro selection system. *Proc Natl Acad Sci U S A* **101**, 3786-3791 (2004).

64. Lu, W.-C. & Ellington, A.D. In vitro selection of proteins via emulsion compartments. *Methods* **60**, 75-80 (2013).

65. Barnes, W.M. The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene* **112**, 29-35 (1992).

66. Eom, S.H., Wang, J. & Steitz, T.A. Structure of Taq polymerase with DNA at the polymerase active site. *Nature* **382**, 278-281 (1996).

67. Thyer, R., Robotham, S.A., Brodbelt, J.S. & Ellington, A.D. Evolving tRNASec for Efficient Canonical Incorporation of Selenocysteine. *Journal of the American Chemical Society* **137**, 46-49 (2015).

68. Wang, L. & Schultz, P.G. A general approach for the generation of orthogonal tRNAs. *Chem Biol* **8**, 883-890 (2001).

Chapter 6: Single-tube collection and nucleic acid analysis of clinical samples for SARS-CoV-2 saliva testing

Publication note:

This paper was originally published in the journal *Scientific Reports.* A correction was made in Supplementary Table S6.1 to fix an incorrect oligonucleotide sequence for *ACTB*.

Cole, K.H., Bouin, A., Ruiz, C. Semler, B.L., Inlay, M.A., Lupták, A. Single-tube collection and nucleic acid analysis of clinical samples for SARS-CoV-2 saliva testing. *Scientific Reports* **12**, 3951 (2022). https://doi.org/10.1038/s41598-022-07871-4

Contribution statement:

A.L. and M.I. proposed the work. A.B. and B.L. performed all virus work. C.R. and **K.H.C.** performed qPCR work. **K.H.C.** performed all RT-LAMP work, prepared all figures, and wrote the manuscript. All authors reviewed the manuscript.

6.1 Abstract

The SARS-CoV-2 pandemic has brought to light the need for expedient diagnostic testing. Cost and availability of large-scale testing capacity has led to a lag in turnaround time and hindered contact tracing efforts, resulting in a further spread of SARS-CoV-2. To increase the speed and frequency of testing, we developed a cost-effective single-tube approach for collection, denaturation, and analysis of clinical samples. The approach utilizes 1 μL microbiological inoculation loops to collect saliva, sodium dodecyl sulfate (SDS) to inactivate and release viral genomic RNA, and a diagnostic reaction mix containing polysorbate 80 (Tween 80). In the same tube, the SDS-denatured clinical samples are introduced to the mixtures containing all components for nucleic acids detection and Tween 80 micelles to absorb the SDS and allow enzymatic reactions to proceed, obviating the need for further handling of the samples. The samples can be collected by the tested individuals, further decreasing the need for trained personnel to administer the test. We validated this single-tube sample-to-assay method with reverse transcription quantitative real-time polymerase chain reaction (RT-qPCR) and reverse transcription loop-mediated isothermal amplification (RT-LAMP) and discovered little-to-no difference between Tween- and SDS-containing reaction mixtures, compared to control reactions. This approach reduces the logistical burden of traditional large-scale testing and provides a method of deployable point-of-care diagnostics to increase testing frequency.

6.2 Introduction

The pandemic caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2)[1] is an ongoing health crisis that has put a spotlight on current methodologies of viral detection. Large scale testing has become a part of the weekly routine to help quell viral spread. In the early stages of the global pandemic, SARS-CoV-2 testing was mainly reserved for those who were symptomatic or had been exposed to other virus-positive individuals, limiting test availability for asymptomatic or pre-symptomatic individuals. Bottlenecks in testing assisted in the spread of

SARS-CoV-2 because asymptomatic individuals make up a large portion of the SARS-CoV-2 infected population[2]. While testing has expanded in recent months to include asymptomatic individuals, the length of time between sample collection and result determination has hindered our ability to identify positive individuals before they become infectious and capable of spreading the virus. This lag in test turnaround is due in part to the lengthy sample processing, RNA purification step, and diagnostic assay setup. Furthermore, the high demand for sterile swabs, buffers, tubes, pipette tips and PCR reagents has led to supply chain complications that have further impeded efforts to expand testing capability. Lastly, the number of steps required—from sample collection to result notification—has placed tremendous pressure on clinical labs and staff to avoid mix-ups that could lead to incorrect diagnoses.

Safety is also an important issue, because specimens are potentially infectious during transport until inactivated during processing. While newer methods involving saliva collection have eliminated many of the RNA purification steps, several processing steps remain, and these methods have not had the transformative impact on testing as envisioned. At-home testing solutions, while promising, are still in development and are expensive on a per-test basis. These issues highlight the need for improved testing methods. Current diagnostic-testing workflows for nucleic-acid based methods require long processing times, hindering the time between testing and result determination. Lag in test turnaround is in part due to the RNA purification process necessary to achieve optimal test sensitivity, which has been weighted as the diagnostic standard[3]. While sensitivity of each test is important, testing frequency is also a critical factor for early detection in households and communities, necessitating the development of simple, fast, and on-site testing technologies[3,4].

Standard methods of viral diagnostics rely on viral transport media (VTM) to preserve the specimen after collection from the patient. VTM does not inactivate viruses; therefore, patient specimens collected in VTM are potentially infectious. While inactivation is performed by sample

heating or by the addition of sodium dodecyl sulfate (SDS)[5], these steps are performed after transportation and delivery to a diagnostic facility. Recent methods have aimed to streamline this process of viral inactivation and genomic RNA extraction for RT-qPCR[6,7], RT-LAMP[8-10], and CRISPR/Cas9[11-13] by utilizing a heat inactivation step prior to assessment. However, these techniques require further handling of the inactivated samples, such as transfer of specific volumes of the sample material into the diagnostic solution for downstream analysis, thus prolonging the process and requiring dedicated personnel or robotic facilities to handle the samples.

We sought to expedite the inactivation of viruses, bacterium, and human cells by employing the common anionic surfactant, dodecyl sulfate (SDS), for the denaturation of viral and cellular structures and subsequent release of RNA with the addition of ethylenediaminetetraacetic acid (EDTA) to chelate divalent metal ions required for nuclease activity (Tris buffer is used to maintain a near neutral pH). Others have utilized similar approaches using surfactants (SDS, polysorbate [Tween], Triton X-100, and nonyl phenoxypolyethoxylethanol [NP-40])[5], reducing agents (e.g., tris(2-carboxyethyl)phosphine [TCEP], dithiothreitol [DTT] to inactivate RNase A family of enzymes) and EDTA, and/or heat inactivation[8,14]. While the inclusion of SDS at low millimolar concentrations into any reaction mixture will inactivate common pathogenic viruses, it also leads to protein denaturation[15] and is thus incompatible with enzymatic activity, such as subsequent RT and PCR steps. However, SDS can be sequestered by non-ionic surfactants, such as Tween, at or above their critical micelle concentration (CMC) via hydrophobic interactions[16]. The inclusion of Tween in an enzymatic reaction mixture therefore abrogates the inhibitory effects of SDS through adsorption by the Tween micelles, allowing for SDS-treated clinical samples to be added directly to an enzyme reaction mixture. Once adsorbed into the Tween micelles, SDS does not denture any proteins in the reaction mixture, allowing for workup of the RNAs (and DNAs) present in the sample using enzymes.

Current clinical sample procurement methods are accomplished using sterile swabs for the collection of nasopharyngeal or nasal samples, in addition to saliva collection, by utilizing specially fitted collection tubes for direct oral deposition. We sought a simple-to-use method of sample acquisition that could be easily mastered by untrained individuals for self-collection. Plastic inoculation loops used in microbiology filled this need, because they are inexpensive, disposable, and can reproducibly remove specific volumes (e.g., 1 μL) of saliva (**Figure 6.1A**). These loops provide a means of consistent, small-volume saliva retrieval in addition to their convenient size for introduction into a PCR test tube. Viral samples retrieved by loop can be directly inactivated by applying the sample directly to an SDS/TE mixture and briefly mixing.

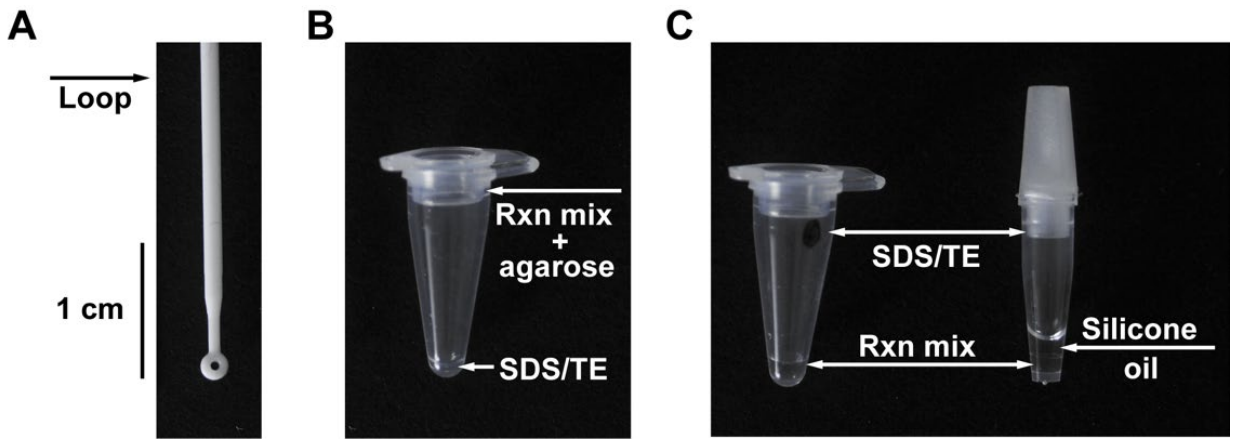


Figure 6.1: Inoculation loop for sample acquisition and configuration of reaction tubes. (**A**) A 1 µL inoculation loop is used for oral sample withdrawal and collection. (**B**) A 0.1 mL PCR tube with a Tween-containing reaction mixture immobilized by agarose is placed inside the cap of the tube and 1 µL of SDS/TE solution is placed at the bottom. A clinical sample withdrawn using the inoculation loop is transferred to the bottom of the tube and denatured by mixing it with the SDS/TE solution using the same loop. Subsequently, the nucleic acid detection mixture Tween 80 is transferred from the cap to the bottom of the tube by centrifugation or shaking. (**C**) Alternatively, 0.1 mL PCR tube with a dried 1 µL spot of SDS/TE on the sidewall of the tube and 5 µL of a Tween-containing reaction mixture at the bottom can be used (left tube). The inoculation loop is rubbed on the side of the tube to dissolve the SDS/TE mixture (marked with permanent marker) and denature the clinical sample and then used to drive the sample to the bottom of the tube. For systems with optical readout at the bottom of the sample tube (e.g., Biomolecular Systems Mic, Quantabio Q tube or Qiagen RotorGene Q) the SDS/TE solution can be dried on the sidewall and the Tween-containing RT-qPCR reaction mixture can be stored underneath 5 µL of oil (right tube).

To utilize this system within a single tube, the SDS/TE solution must be initially separated from the enzyme reaction mixture. This was accomplished by immobilizing the reaction mixture with agarose in the lid of the tube, while supplying the SDS/TE mixture at the bottom (**Figure 6.1B**). Alternatively, the SDS/TE could be heat-dried on the inner-side wall of the PCR tube with the reaction mix contained at the bottom of the PCR tube (**Figure 6.1C**). This configuration allows for streamlined sample retrieval, inactivation by SDS/TE, and subsequent introduction into the Tween-containing enzymatic reaction mixture. Utilizing this approach, we demonstrate both RT-qPCR and RT-LAMP analyses of viral particles harboring the SARS-CoV-2 spike protein. The method thus represents a practical sample-to-assay approach with the potential to accelerate the diagnostic process as well as providing a cost-effective solution for increasing testing frequency.

6.3 Results

Optimization of SDS and Tween for viral inactivation and qPCR

We first tested qPCR reactions to determine the concentration of Tween necessary to sequester SDS and prevent the inhibition of amplification. To test this, human RNase P gene (RPP30) was amplified from a control plasmid at varying concentrations of Tween 20 and Tween 80. We first established that 0.005% (w/v) final concentration of SDS is sufficient to inhibit the amplification reaction. We then increased the SDS concentration to 0.01 % (w/v) and tested Tween 20 and Tween 80 at 1%, 2%, and 5% (v/v) for the ability to amplify DNA (**Figure 6.2A**). At 5% (v/v) Tween 20 supported amplification in the presence of SDS but was ineffective at 1% and 2% (v/v). Tween 80 abolished the inhibitory effects of SDS at all three concentrations and supported amplification comparable to the control reaction. The results are consistent with the higher stability of the Tween 80 micelles, compared to Tween 20 (Tween 20 has a higher critical micelle concentration than Tween 80)[16,17]. While not directly tested, we chose to use 3% (v/v) Tween 80 for our subsequent applications as it was easier to consistently pipette in comparison to the more viscous 5% (v/v) Tween 80 solutions.

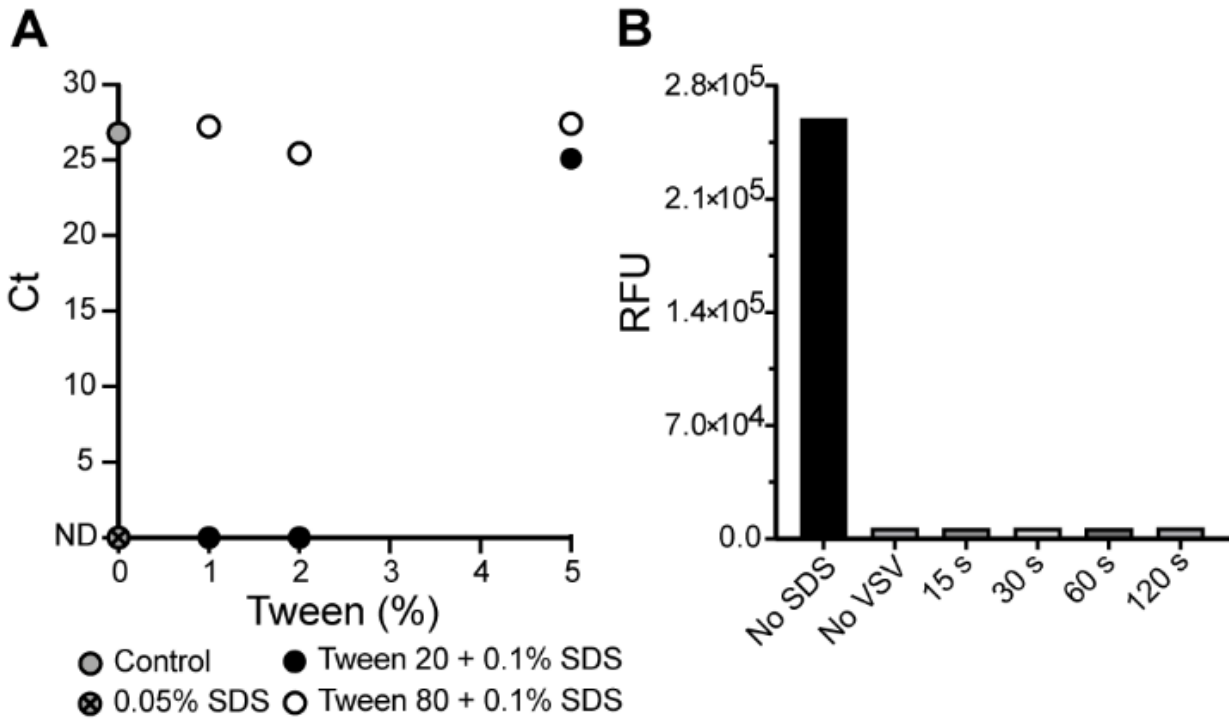


Figure 6.2: Effective SDS and Tween concentration determination for qPCR and virus inactivation. (**A**) Determination of the minimal Tween 20 and Tween 80 concentration necessary for SDS sequestration to enable amplification of 20,000 copies/µL of RNase P DNA. Ct: cycle threshold; ND: not detected. (**B**) Fluorescence assay of eGFP-expressing recombinant VSV-SARS-CoV-2 (10^7 pfu/mL) infecting Vero-E6 cells after incubation with 0.1% (w/v) SDS for 15–120 seconds. Only a sample that was not exposed to SDS showed infectivity. 15 second exposure to 0.1% (w/v) SDS proved sufficient to completely inactivate the virus, likely due to dissolution of its membrane and denaturation of the sample proteins.

A vesicular stomatitis virus (VSV) expressing enhanced green fluorescent protein (eGFP) and the SARS-CoV-2 spike protein (VSV-eGFP-SARS-CoV-2[18]), was used as a model system to test the SDS inactivation efficacy (**Figure 6.2B**). Like SARS-CoV-2, VSV is an enveloped RNA virus that loses infectivity when the envelope is disrupted by detergent treatment. To assess virus inhibition, 1-µL inoculation loops were dipped into a VSV-eGFP-SARS-CoV-2 viral suspension of 10^7 pfu/mL and introduced to a heat dried 1 µL spot containing 0.1% (w/v) SDS in a Tris-EDTA buffer on the inner-side wall of a PCR tube. The dried spot was indicated with a permanent marker to better identify the location of the SDS/TE. The 1 µL viral samples on the inoculation loops were mixed on the SDS spot for 15-120 seconds prior to introduction into a 5-µL RT-qPCR reaction mixture containing 3% (v/v) Tween 80 at the bottom of the PCR tube. To prevent cross contamination, a new 1-µL loop was used to remove 1 µL from the RT-qPCR reaction mixtures and inoculate Vero-E6 cell monolayers grown 24 hours prior to infection. Cells were grown for 5 days post-infection with daily fluorescent signal monitoring and until the untreated viral sample achieved complete infection. After incubation, eGFP expression in SDS-treated, no-SDS, and no-VSV samples was measured by fluorescence. Inhibition of VSV-eGFP-SARS-CoV-2 was observed at the minimum time of mixing on the 0.1% (w/v) SDS/TE spot at 15 seconds and was comparable to mixing for longer periods and to the no-VSV control. Greater eGFP expression was observed in the untreated VSV-eGFP-SARS-CoV-2 control (**Figure 6.2B**). The substantial fluorescence observed in the untreated sample demonstrates that SDS fully inactivates the virus, but the 3% (v/v) Tween 80 dissolved in the enzyme reaction mixture does not. These results illustrate that 0.1% (w/v) SDS with 15 seconds of mixing is sufficient to inactivate VSV-eGFP-SARS-CoV-2.

Comparison of immobilization methods

For point-of-care testing, immobilization of the master mix is necessary to prevent the mixing of the Tween-80-containing enzyme/primer master reaction with the SDS prior to virus

denaturation, particularly when SDS is supplied at the bottom of the reaction tube. Several methods can be used for the separation of the reagents within a single tube: 1) mineral oil or silicone oil can be overlaid above the enzyme mix[19] (Figure 6.1C), 2) low-melting temperature wax can be overlaid on top of the mixture and solidified to make a solid barrier at room temperature[20], or 3) low-melting-point agarose can be introduced into the mixture at a slightly elevated temperature (e.g., 42 ºC) and then stored at room temperature or lower to immobilize the mixture. We confirmed that both the oil overlay and the agarose inclusion supported immobilization of the enzyme mixture, and for most experiments, we used low-melting-point agarose for its ability to melt at the low reaction temperatures (≤50 °C), enzyme stabilization[21], and its optical transparency[22]. The addition of 1% (w/v) agarose in the presence of SDS (0.01% (w/v) final) and 3% (v/v) Tween 80 showed no discernable inhibition of the amplification of SARS-CoV-2 N gene from a SARS-CoV-2 plasmid when compared to reactions containing 3% (v/v) Tween 80, with and without 0.01% SDS (w/v, final) in 10 µL samples (**Figure 6.3A**).

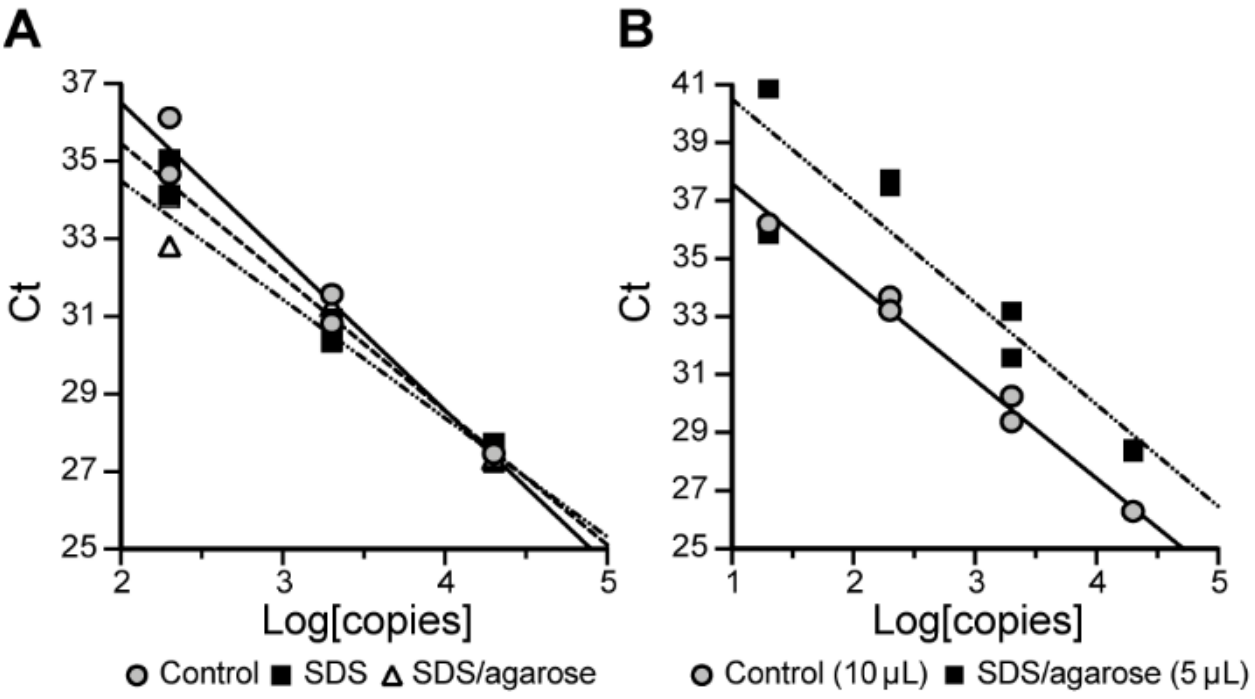


Figure 6.3: Comparison of standard qPCR to SDS/Tween and SDS/Tween agarose reactions. (**A**) Comparison of 10 µL qPCR reactions for SARS-CoV-2 N1 gene DNA amplification, containing 3% (v/v) of Tween 80 with (R^2 = 0.9781) and without 0.01% SDS (w/v, final; R^2 = 0.9818) and 1% (w/v) agarose (R^2 = 0.9260). (**B**) Comparison of a 10 µL qPCR reaction for SARS-CoV-2 N1 gene DNA containing 3% (v/v) Tween 80 (R^2 = 0.9908) with a 5 µL qPCR reaction (R^2 = 0.8451) for SARS-CoV-2 N1 DNA detection, containing 3% (v/v) Tween 80, 0.01% SDS (w/v, final) and 1% (w/v) agarose. In all conditions containing SDS, the DNA samples were initially incubated with 0.1% SDS (w/v) in Tris/EDTA buffer to simulate denaturing conditions for clinical samples. Ct: cycle threshold.

Next, we compared a 5 µL reaction containing the SDS/Tween/agarose to a standard 10-µL reaction mixture containing 3% (v/v) Tween 80 to test any inhibitory effects under more stringent conditions (**Figure 6.3B**). The smaller reaction volume with the addition of both 0.01% SDS (w/v, final) and 1% (w/v) agarose increased the number of cycles necessary for detection of SARS-CoV-2 N1 at each concentration of plasmid DNA, thus exhibiting a slight decrease in amplification efficiency (**Figure 6.3B**). Importantly, a 5 µL reaction with a sample introduced into a 1 µL drop of SDS solution and subsequently mixed with a master mix containing Tween 80 allowed quantitative detection of DNA plasmids.

RT-qPCR of SARS-CoV-2 patient RNA and viral particle

While the SDS/Tween assay was compatible with qPCR, understanding RT-qPCR functionality was necessary to establish a method for sample-to-assay assessment of infection. To assess RT-qPCR, 50 µL of two positive SARS-CoV-2 patient nasal swabs samples stored in VTM with 0.5% (w/v) SDS were purified using a Zymo Research Quick DNA/RNA-viral extraction kit and eluted in an equivalent volume of nuclease-free water. We diluted the unpurified RNA (VTM with 0.5% [w/v] SDS) five-fold in nuclease-free water to reach an effective SDS concentration of 0.1% (w/v) and compared that to equally dilute purified samples (**Figure 6.4A**). Reactions were performed at 10 µL with the RT-qPCR reaction mixture immobilized in the lid of the reaction tube (0.1 mL optically transparent qPCR tubes). The enzyme mixture also contained SARS-CoV-2 N1 primers and probe, 3% (v/v) Tween 80, and 0.5% agarose (w/v). Samples were prepared in triplicate with 1-µL of sample added by pipetting. Amplification of both purified and unpurified patient samples provided consistent Ct values with no observable differences between them.

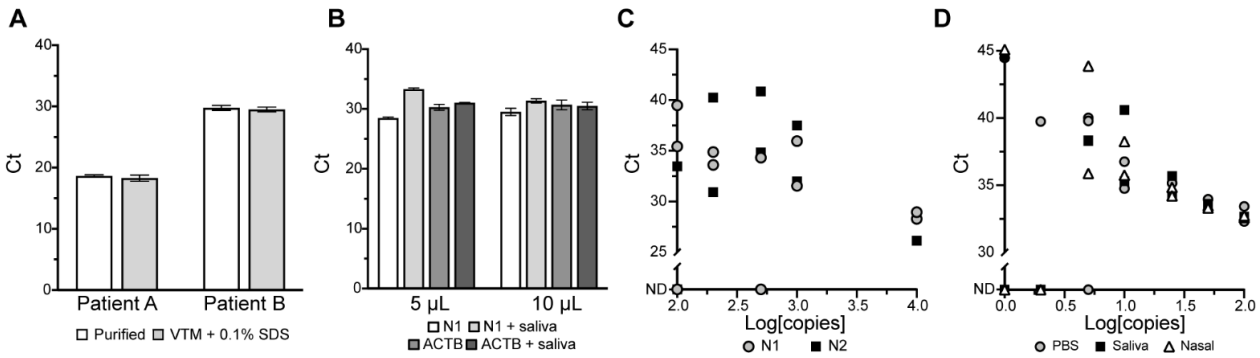


Figure 6.4: Single-tube sample-to-assay RT-qPCR of SARS-CoV-2 positive-patient RNA sample and inactivated viral particles. (**A**) Comparison of five-fold diluted purified and unpurified SARS-CoV-2 patient RNA samples (nasal swabs stored in VTM + 0.5% [w/v] SDS) in triplicate 10 µL RT-qPCR reactions containing SARS-CoV-2 N1 primers and probe, 3% (v/v) Tween 80, and 0.5% (w/v) agarose. Samples were diluted five-fold to achieve a 0.1% (w/v) SDS working concentration for unpurified samples in VTM. Data are presented as ± SEM. (**B**) Comparison of 5 µL and 10 µL RT-qPCR reactions containing 3% (v/v) Tween 80, and 0.5% (w/v) agarose with a purified SARS-CoV-2 patient RNA sample. Triplicate reactions were performed at each volume, comparing the patient sample with and without spiking into saliva. Reactions were performed with SARS-CoV-2 N1 and beta-actin (ACTB) primer-probe sets. Data are presented as ± SEM. (**C**) RT-qPCR results of SARS-CoV-2 synthetic RNA (n = 11) in water. RT-qPCR was performed for 45 cycles using SARS-CoV-2 N1- and N2-specific primers and probes on a Quantabio Q thermal cycler. Samples were tested blind with concentrations and storage media revealed thereafter (ND = not detected). (**D**) Scatter plot of Ct value (y-axis) versus log[copies] of SARS-CoV-2 viral particles (x-axis) of inactivated SARS-CoV-2 viral particle samples (n = 49) in PBS, saliva, and nasal media. RT-qPCR was performed for 45 cycles with SARS-CoV-2 N1-specific primers and fluorescent probe on a Bio-Rad CFX Connect thermal cycler. Ct: cycle threshold.

We then assessed one purified patient RNA sample that was spiked into saliva obtained from one of our authors to determine whether saliva could inhibit or interfere with detection (**Figure 6.4B**). Reactions were performed at two volumes (5 µL and 10 µL) and contained SARS-CoV-2 N1 and beta-actin (ACTB) primer-probe sets, 3% (v/v) Tween 80, and 0.5% agarose (w/v). Samples were prepared in triplicate, retrieved with 1-µL inoculation loops, and inactivated by mixing with 0.1% (w/v) SDS in TE and 5 mM DTT[8] stored at the bottom of the reaction tube. Reactions were performed using the previously mentioned CDC-recommended protocol. A small difference in amplification of SARS-CoV-2 N1 was observed at both reaction volumes when the RNA sample was spiked into saliva. However, this same difference was not observed for ACTB between either the 5 µL or 10 µL reaction volumes. We hypothesize that this difference is due to the purified-patient RNA degrading when incubated in saliva, prior to inactivation by SDS. The ACTB mRNA was likely protected within cells collected with saliva and released after SDS treatment, explaining the consistency of the Ct values in the ACTB reactions with and without saliva. Overall, these results showed that the reaction mixture containing both Tween 80 and agarose can detect SDS treated SARS-CoV-2 RNA.

Next, we tested our sample-to-assay method utilizing an RT-qPCR reaction mixture overlaid with silicone oil. We tested samples of synthetic SARS-CoV-2 RNA stored in water (n=11, provided by the XPRIZE Foundation; **Figure 6.4C**). The RT-qPCR reaction mixture was prepared and stored beneath a layer of silicone oil in a Quantabio Q tube produced specifically for the Q qPCR instrument, which measures fluorescence from the bottom of the tubes by rotating the samples above immobile excitation LEDs and photomultiplier detectors. This process also serves to centrifuge the samples to the bottom of the tubes. We assessed these samples by first collecting each sample using an inoculation loop and mixing the sample on a dried SDS/TE spot on the side of the tube for 15 seconds, which was marked with a permanent marker. After mixing, the sample was driven by the inoculation loop to the bottom of the test tube (past the silicone oil)

into the reaction mixture. Because the samples consisted of synthetic RNA stored in water, it was expected that some of the synthetic SARS-CoV-2 RNA had degraded. This experiment provides an example of this method when utilizing a system that reads fluorescence from the bottom of the tube and maintains partitioning of the reaction mixture with silicone oil. Additionally, this provides a useful method of detection of >200 copies/μL of viral RNA.

We next assessed 49 inactivated SARS-CoV-2 viral particle samples diluted in phosphate buffered saline (PBS), saliva, and nasal media, provided as blind samples by the XPRIZE Foundation (**Figure 6.4D**). Reaction tubes (0.1 mL optically transparent qPCR tubes) were prepared by drying 1 μL of 0.1% (w/v) SDS in TE buffer solution on the inner-side wall of the PCR tube. Samples were collected by dipping 1-μL inoculation loops in the stored samples, swirled on the dried SDS for 15 seconds and subsequently driven to the bottom of the tube containing a 5-μL RT-qPCR reaction mixture that included 3% (v/v) Tween 80 and N1-specific primers and probe. Reactions containing 100, 50, and 25 copies of SARS-CoV-2 viral particles appeared at similar cycles regardless of storage media (n = 18) while samples containing 10 copies varied between 35 and 40 cycles (n = 7) with greater variability and false-negative results at lower concentrations (**Figure 6.4D**). Based on these results, we estimate our limit-of-detection (LOD) to be 10 copies/μL.

Sample-to-assay RT-LAMP with the addition of Tween 80 and SDS

After verifying sensitive, reproducible RT-qPCR detection with SDS-denatured samples and stabilization of the enzymatic reactions with Tween, we then tested RT-LAMP for a single-tube sample collection and analysis. RT-LAMP has key advantages over RT-qPCR because it is performed isothermally and does not require additional equipment for readout, resulting in diagnostic simplicity. These advantages make RT-LAMP a promising tool for point-of-care detection of viral infection[8-10, 23-27]. Previous work has established that RT-LAMP can withstand

3% (v/v) Tween 20 in addition to 3% (v/v) Triton X-100 with little inhibition of amplification[8]. To validate RT-LAMP detection of RNA samples in the presence of SDS and Tween, real-time fluorescence measurements were collected on a real-time PCR thermal cycler, with the addition of the nucleic-acid intercalating dye SYTO 82[28-30]. Four concentrations of SARS-CoV-2 viral particles (100, 50, 25, and 10 copies/μL) were collected using 1 μL inoculation loops and incubated at 65 ºC for 50 minutes in the RT-LAMP reaction. Amplification of 100 to 10 copies of RNA was detected after 33 minutes using N1-specific primers designed by Huang et al. (**Figure 6.5A**) [31]. Non-specific amplification was not observed in no-reverse-transcriptase (NRT) and 0 copy controls, which was further confirmed by agarose gel electrophoresis (**Figure 6.5B**).

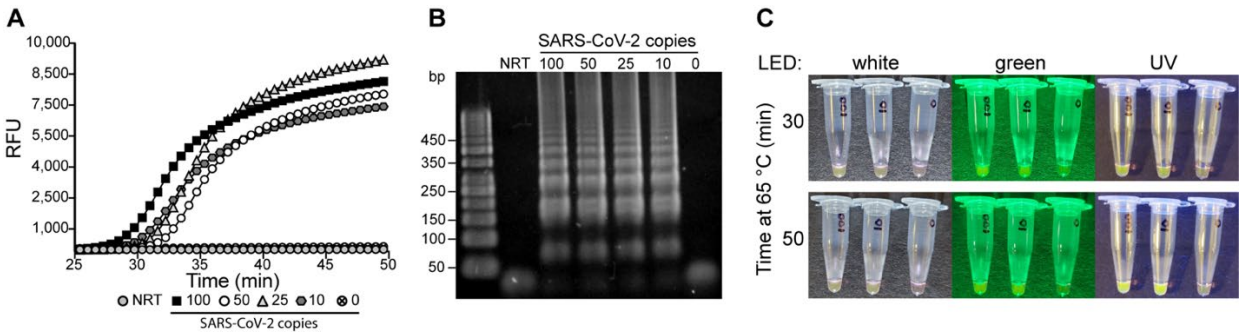


Figure 6.5: RT-LAMP amplification with SDS and Tween 80. (**A**) Real-time fluorescence of SARS-CoV-2 N gene amplification by RT-LAMP was performed at 65 °C for 50 minutes. Indicated concentrations (copies) of SARS-CoV-2 viral particles were assessed including a no-reverse-transcriptase (NRT) and no-template control (NTC; 0 copies), with all SARS-CoV-2 samples providing a positive result after ~30 minutes. (**B**) Amplification products (1 µL) from the real-time fluorescence RT-LAMP reaction were fractionated using agarose gel electrophoresis. The expected laddering pattern of LAMP amplification products was observed at all SARS-CoV-2 viral particle concentrations and not observed in NRT and NTC controls. The uncropped gel can be found in Supplementary Figure 1. (**C**) Fluorescence of RT-LAMP reactions of SARS-CoV-2 viral particles containing 20 µM SYTO 82 observed after 30 and 50 minutes of incubation at 65 °C. Reactions were analyzed with white, green, and UV (365 nm) handheld LED flashlights at each time point, with readily observable fluorescence of amplification reactions starting from 100 copies and 10 copies (labeled) after 30 minutes with each illumination. After 50 minutes, fluorescence in the 100-copy and 10-copy tubes increased with no observable amplification in the 0-copy control. RFU: relative fluorescence units.

As mentioned, the isothermal reaction of RT-LAMP eliminates the need for expensive detectors, and with additional indicators, positive results can be observed without costly equipment or gel electrophoresis. Previous methods have utilized pH (e.g., phenol red) or complexometric dyes (e.g., hydroxynaphthol blue), and nucleic acid intercalators (e.g., SYBR Green, EvaGreen, and SYTO 82) for the detection of target RNA[8,9,23,27,29,32,33]. After validating RT-LAMP using real-time fluorescence, we hypothesized that amplification can be observed without dedicated fluorescence readers. We tested this by increasing the concentration of SYTO 82 in the RT-LAMP reaction from 5 to 20 µM. Amplification of 100 copies and 10 copies of SARS-CoV-2 RNA was readily observed in 30 minutes at 65 °C, with no amplification observed in the NTC reaction after 50 minutes (**Figure 6.5C**). These results were obtained under consumer-grade white, green, and UV (365 nm) excitation light using handheld LEDs and imaged by a mobile phone.

6.4 Discussion

We present a method aimed at increasing the frequency of viral testing by decreasing the handling and time between sample collection and assessment for direct RNA detection in patient saliva within a single reaction tube (**Figure 6.1**). In previous studies, nucleic acid detection techniques have required extraction of the genetic material from the clinical samples[8,11,12,31,34,35], with some methods using heat denaturation to inactivate the pathogens and human cells in the clinical samples and release the genetic material for subsequent analysis[6-9,13,26,36,37]. In most cases, the sample collection and inactivation or extraction is performed in separate tubes from the analysis step and thus require additional handling by trained personnel. We sought to simplify the sample preparation to decrease the time and effort necessary to collect, denature, and analyze the sample. In effect, we have eliminated all pipetting and open-tube processing steps. If an individual collects their own specimen utilizing an inoculation loop (**Figure 6.1**), adds it into the

tube, and caps it, trained staff will only need to place the tube in the RT-qPCR instrument and analyze the results. After sample collection, the tube can be washed on the outside with a denaturing solution (e.g., soap or sanitizer) and placed in a tube rack or directly into the qPCR instrument by the tested individual, further eliminating handling by technical personnel.

To render the clinical samples noninfectious, we used laboratory detergent (SDS), which is commonly used to denature proteins and disintegrate lipid membranes. Detergent is the most active and widely available denaturant and, as expected, we observed that incubating a recombinant VSV-eGFP-SARS-CoV-2 virus sample for 15 seconds made the sample completely non-infectious (**Figure 6.2B**). However, SDS also denatures enzymes that are used for nucleic acid detection, and to avoid a purification step, we required a mechanism of SDS sequestration from the solution. One approach is to precipitate the dodecyl sulfate with potassium, which, unlike the sodium salt, is insoluble; however, precipitation would lead to flocculates that could potentially hamper the downstream analysis of the solution. Instead, we chose to adsorb dodecyl sulfate into micelles formed by very stable surfactants, such as the Triton and Tween series of non-ionic detergents. Of the commercially available non-ionic detergents, Tween 80 forms the most stable micelles with low micromolar critical micelle concentration over a wide range of temperatures. Both Tween and Triton surfactants have been widely used in enzymatic reactions, including PCR and droplet PCR, and are therefore highly suitable for SDS sequestration. With the addition of Tween to an RT-qPCR reaction, the denaturing effects of the SDS are abolished, enabling amplification of target DNA (**Figure 6.2A**).

The addition of agarose was tested to determine whether a stabilizing medium could be supplemented to better immobilize the reaction mixture away from the place where a small volume of SDS/TE was applied (and sometimes dried). We propose this approach as a means of immobilizing the reaction mixture during sample collection to prevent the loss of the reagents

crucial for diagnostics. We observed no inhibition of qPCR due to SDS/Tween and agarose at equal reaction volumes (10 µL), when compared with standard reaction conditions (**Figure 6.3B**), and the amplification efficiency was only slightly lower when 1 µL samples were incubated in 5 µL enzyme reactions. A similar result was observed for RT-qPCR utilizing a column-purified SARS-CoV-2 patient sample, even in the presence of saliva (**Figure 6.4B**). We believe that this slight decrease in amplification efficiency is acceptable because the mean cycle threshold of N1 primer sets used in standard RT-qPCR protocols for self-saliva collection is 32.8 for asymptomatic patients and those with early- and late-onset of disease[38].

Testing inactivated SARS-CoV-2 viral particles in PBS, saliva, and nasal media revealed variation at concentrations of 10 copies and below, giving us an estimated limit of detection of 10 copies/µL (**Figure 6.4D**). Because others have illustrated a detection limit of 5.6 copies/µL of SARS-CoV-2 following the Centers for Disease Control and Prevention's RT-qPCR guidelines for the N1-primer set, we believe our protocol falls within a meaningful detection range for this coronavirus[39]. This slight sensitivity decrease is compensated by the great reduction in diagnostic handling and turnaround time, thereby allowing for more frequent testing. Furthermore, 10 copies/µL is almost two orders of magnitude below the median viral load of sputum samples (752 copies/µL) collected from infected individuals[40].

RT-LAMP has the potential to fill a niche previously occupied by rapid-antigen tests. It has been reported that rapid-antigen tests have a high false-negative rate in saliva samples with Ct ≥25 while RT-LAMP had greater sensitivity for samples with Ct ≤35[38]. Our method applied to RT-LAMP allowed for the detection of 10 copies (mean Ct = 34.8) of SARS-CoV-2 N1 RNA in 30 minutes (**Figure 6.5**). By increasing the concentration of SYTO 82 in the RT-LAMP reaction, a positive result could be seen under white light and better observed with green and UV (365 nm) LEDs. Based on these results, RT-LAMP would require the same incubation time as commercial

rapid-antigen tests (30 minutes) with greater sensitivity, although requiring incubation at 65 °C. Additionally, our method of RT-LAMP does not require the addition of buffering reagents typically required of lateral-flow rapid-antigen tests. While it does not have the same accuracy of RT-qPCR and suffers from non-specific amplification activity, RT-LAMP paired with this sample-to-assay method could replace the market of rapid-antigen tests while maintaining the simplicity of result determination.

6.5 Materials and Methods

Fluorescent VSV inactivation assay

Vero-E6 cells were grown in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum and a 1X antibiotic-antimycotic solution and maintained at 37°C with 5% CO_2. Vero-E6 cells were seeded 24 hours prior to infection in 6-well plates. Infection was performed using VSV-eGFP-SARS-CoV-2 (generously provided by Paul W. Rothlauf and Sean P.J. Whelan at Washington University in St. Louis[18]) expressing SARS-CoV-2 spike protein (*S* gene). Following infection, cells were grown at 34 °C with 5% CO_2 in DMEM supplemented with 10% FBS.

SDS inactivation was performed by dipping a 1 µL inoculation loop (Globe Scientific, 2810) in a VSV-eGFP-SARS-CoV-2 suspension at 10^7 pfu/mL and applying the inoculation loop to a dried 1 µL spot of SDS/TE (0.1% w/v SDS, Tris-HCl, pH 8.0) on the side wall of a PCR reaction tube, followed by mixing for 15-120 seconds. The treated inoculation loop was then driven down into a RT-qPCR reaction mixture (see standard RT-qPCR methods) containing 3% (v/v) Tween 80 and disposed of. A new 1 µL inoculation loop was then used to sample the inoculated RT-qPCR reaction mixture and introduced into the culture media of Vero-E6 cells seeded in a 6-well culture plate. Cells were grown at 34 °C with 5% CO_2 in DMEM supplemented with 10% FBS for 5 days. Cells were fixed in PBS supplemented by 3.7% formaldehyde for 20 minutes at room temperature

and washed 3 times with phosphate buffered saline (PBS). Fluorescence was measured using a CLARIOstar Plus Microplate Reader.

DNA and RNA controls

Human RNase P gene (*RPP30*) was amplified from the Hs_RPP30_Positive control plasmid (Integrated DNA Technologies, IDT, 10006626). SARS-CoV-2 N1 DNA was amplified from the 2019-nCoV_N_Positive control plasmid (IDT, 10006625). SARS-CoV-2 viral particles and synthetic RNA were provided by the XPRIZE Foundation (Team ID: 3650) and produced by ZeptoMetrix and Twist Biosciences, respectively. Viral particles were stored at 4 °C per manufacturer recommendation. Synthetic RNA was stored at -80 °C per manufacturer recommendation.

Purification of SARS-CoV-2 patient samples

Human specimen samples were obtained under an institutional review board (IRB) approved protocol (HS# 2012-8716) at the University of California, Irvine (UCI). Specimens collected originally for diagnostic purposes were processed and stored in the University of California, Irvine Medical Center hospital. Samples were collected under the HS# 2012-8716 IRB approved protocol as post-diagnostic remnants from the Pathology Department at the University of California, Irvine, and processed for research use by the Chao Family Comprehensive Cancer Center Experimental Tissue Shared Resource. All experimental protocols were approved by the UCI IRB and the UCI Institutional Biosafety Committee (IBC). UCI IRB categorized our protocol as Non-Human Subject Research, since de-identified specimens, stripped of all personal health information (PHI) were received from the Experimental Tissue Resource (HS# 2012-8716) and the UCI IBC (BUA-R191) covered the usage of inactivated, denatured specimens.

Pharyngeal swabs were maintained at 4 ºC until initial diagnostic test was performed, followed by freezing in the original collection tubes at -80 ºC. At the time of releasing for research, the swabs

were thawed and aliquoted at desired volumes, inactivated by incubation at room temperature for 30 min in 0.5% SDS[5], and released to the investigators. A 50 µL volume of inactivated samples, in viral transport media (VTM) and 0.5% (w/v) SDS, were purified using the Zymo Research Quick-DNA/RNA viral extraction kit (D7020) following the manufacturer's protocol and eluted in 50 µL of nuclease-free water.

Standard RT-qPCR reactions

Standard RT-qPCR reactions were performed using either 2X qScript XLT 1-Step RT-qPCR ToughMix (Quantabio) or 4X TaqPath 1-Step Multiplex Master Mix (Invitrogen) at 1X final concentration. Forward and reverse primers (Supplementary Table 1) were used at 500 nM with hydrolysis probes at 125 nM final concentration. Reactions were performed at either 5 µL or 10 µL after the addition of dH_2O to a desired volume and 1 µL of DNA/RNA samples. Reactions were measured on a Bio-Rad CFX Connect for 25 °C for 2 minutes, 50 °C for 15 minutes, 95 °C for 2 minutes, and 45 cycles of 95 °C for 3 seconds, and 55 °C for 30 seconds.

RT-qPCR with SDS, Tween, and with/without agarose

RT-qPCR reactions containing SDS and Tween 80 were assembled similarly to the standard RT-qPCR reaction mixtures and contained 1X RT-qPCR master mix, 500 nM forward and reverse primers, and 125 nM of hydrolysis probes. Tween 80 was added to the desired final concentration of 3% (v/v). The SDS solution (0.1% [w/v] SDS in 10 mM Tris-HCl and 0.1 mM EDTA, pH 8.0) was added at a volume of 1 µL to the bottom of a 0.1 mL flat-top optically transparent PCR tube (Thomas Scientific) or added to the inner-side wall of the tube and allowed to dry at 70 °C for 5-10 min. The SDS spot was marked with permanent marker to identify its location.

For agarose containing reactions, ultra-low gelling temperature agarose (Sigma Aldrich, A2576) was prepared at a working concentration of 2.5-5% (w/v) in dH_2O by heating the mixture to 80 °C for 2 minutes and holding at 42 °C. Warm agarose was added to a RT-qPCR reaction mix (1X

RT-qPCR master mix, 500 nM of forward and reverse primers, 125 nM hydrolysis probes, and 3% [v/v] Tween 80) to a final concentration of 0.5-1% (w/v). The mixture was then added to the lid or to the bottom of a 0.1 mL optically clear flat-top PCR tube and allowed to solidify at room temperature.

For oil overlayed reactions, RT-qPCR reaction mixtures were added beneath the 5 µL silicone oil layer in a Q Tube (Quantabio, 95910-20). The SDS solution (0.1% [w/v] SDS in 10 mM Tris-HCl and 0.1 mM EDTA, pH 8.0) was added at a volume of 1 µL to the side of the tube and allowed to dry. Samples were added via 1 µL inoculation loop directly to the SDS spot and mixed for 15 seconds prior to being driven down into the silicone oil/reaction mixture at the bottom of the tube. Samples were incubated on a Quantabio Q thermal cycler for 37 °C for 30 seconds, 42 °C for 4 minutes, 50 °C for 5 minutes, 95 °C for 1 minute, and 45 cycles of 95 °C for 5 seconds, and 55 °C for 10 seconds.

RT-LAMP

5X RT-LAMP reaction buffer was prepared using 10X Isothermal Amplification Buffer (New England Biolabs; NEB) with the addition of 40 mM $MgSO_4$, 25 µM SYTO 82 (Invitrogen, S11363), 4.5% (v/v) Tween 80 and dH_2O to a preferred volume. RT-LAMP reactions contained 2 µL 5X RT-LAMP reaction buffer (1X reaction buffer: 20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 50 mM KCl, 10 mM $MgSO_4$, 0.1% [v/v] Tween 20, 0.9% [v/v] Tween 80, 5 µM SYTO 82, pH 8.8), 2 µL 7 mM dNTPs, 1 µL *Bst* 2.0 WarmStart DNA polymerase (8 U/µL) (NEB) 0.25 µL WarmStart RTx reverse transcriptase (150 U/µL) (NEB), and 1 µL 10X SARS-CoV-2 N1 LAMP primer mix (2 µM F3/B3, 16 µM FIP/BIP, 4 µM LF/LB; **Supplementary Table S6.1**) and dH_2O to 9 µL. Prior to the addition of the reaction mixture, 1 µL 0.1% (w/v) SDS in 10 mM Tris-HCl pH 8, and 0.1 mM EDTA was added to the bottom of a 0.1 mL flat-top optically transparent PCR tube. Samples were collected with 1 µL inoculation loops and mixed in the SDS solution for 15 seconds. After sample addition,

the reaction mixture was added for a final volume of 10 µL. Reactions were monitored using a Bio-Rad CFX Connect thermal cycler at 65 °C for 50 min.

End-point assays were prepared under identical conditions to real-time RT-LAMP reactions with a final concentration of 20 µM SYTO 82 for improved visualization. Reactions were incubated for 30 min and 50 min and visualized by white, green and UV (365 nm) LED lights. Images were taken by mobile phone.

6.6 Acknowledgements

We thank Paul W. Rothlauf and Sean P.J. Whelan for the recombinant VSV-eGFP-SARS-CoV-2 and the XPRIZE Foundation for providing inactivated SARS-CoV-2 viral particles through the XPRIZE Rapid Covid Testing competition. The authors wish to acknowledge the support of the Chao Family Comprehensive Cancer Center Experimental Tissue Shared Resource, supported by the National Cancer Institute of the National Institutes of Health under award number P30CA062203. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The project described was supported by the UC Irvine COVID-19 Basic, Translational and Clinical Research Funding Opportunity (to A.L. and B.L.S.), NSF CBET 1804220 (to A.L.), NIH grant AI145003 (to B.L.S.), and by numerous donors through a crowdsourcing campaign (ZotFunder "Simple and Rapid COVID-19 Diagnostic Test”), whose support is gratefully acknowledged. A.L. was a Fellow of the J. S. Guggenheim Memorial Foundation.

6.7 Supplementary Data

RT-qPCR primer design

Human RNase P (RPP30) primers and probe were synthesized by IDT (oligonucleotides: 25 nmol, standard desalting; hydrolysis probe: 25 nmol, HPLC purified). SARS-CoV-2 N1 and N2 primers and probes were synthesized by IDT (oligonucleotides: 25 nmol, standard desalting; hydrolysis probe: 25 nmol, HPLC purified). Beta-actin (ACTB) primers and probe (IDT, Hs.PT.39a.22214847) were synthesized by IDT (oligonucleotides: 25 nmol, standard desalting; hydrolysis probe: 25 nmol, HPLC purified). Oligonucleotides and probes were resuspended in 10 mM Tris-HCl, pH 7.5 and 0.1 mM EDTA to 100 µM. RNase P and SARS-CoV-2 N1 primers and probes were designed for research-use only by the Centers for Disease Control and Prevention[1]. Primer and probe sequences are listed in **Supplementary Table S6.1**.

RT-LAMP primer design

N1 RT-LAMP primers were designed by Huang *et al.* for the SARS-CoV-2 *N* gene[31]. Oligos were synthesized by IDT (25 nmol, standard desalting). Oligomers were resuspended in 10 mM Tris-HCl, pH 7.5 and 0.1 mM EDTA to 100 µM. RT-LAMP sequences can be found in **Supplementary Table S6.1**.

Table S6.1: RT-qPCR and RT-LAMP primers. Human RNase P and SARS-CoV-2 primers and probes were designed by the Centers for Disease Control and Prevention. SARS-CoV-2 N2 primers and probe were designed for this study. Human beta-actin (ACTB) primers and probe were designed by IDT (Hs.PT.39a.22214847). SARS-CoV-2 N1 RT-LAMP primers were designed by Huang *et* al[31]. All primers were purchased and synthesized by IDT.

	Target	Name	Sequence
RT-qPCR	Human RNase P[1] (RPP30; NM_006413.5)	RNase P Forward	AGATTTGGACCTGCGAGCG
		RNase P Reverse	GAGCGGCTGTCTCCACAAGT
		RNase P Probe	FAM-TTCTGACCTGAAGGCTCTGCGCG-BHQ1
RT-qPCR	SARS-CoV-2 N1 (N; NC_045512.2)	2019-nCoV_N1 Forward	GACCCCAAAATCAGCGAAAT
		2019-nCoV_N1 Reverse	TCTGGTTACTGCCAGTTGAATCTG
		2019-nCoV_N1 Probe	FAM-ACCCCGCATTACGTTTGGTGGACC-BHQ1
RT-qPCR	SARS-CoV-2 N2	N2 Forward	TTACAAACATTGGCCGCAA
		N2 Reverse	GTGACTTCCATGCCAATGC
		N2 Probe	SUN-ACAATTTGCCCCCAGCGCTTCAG-3IABkFQ
RT-qPCR	Human beta-actin (ACTB; NM_001101.5)	ACTB (Exons 1 – 2) Forward	see IDT: Hs.PT.39a.22214847
		ACTB Reverse	see IDT: Hs.PT.39a.22214847
		ACTB Probe	see IDT: Hs.PT.39a.22214847
RT-LAMP	SARS-CoV-2 N1[31]	F3	TGGACCCCAAAATCAGCG
		B3	GCCTTGTCCTCGAGGGAAT
		FIP	CCACTGCGTTCTCCATTCTGGTAAATGCACCCCGCATTACG
		BIP	CGCGATCAAAACAACGTCGGCCCTTGCCATGTTGAGTGAGA
		LF	TGAATCTGAGGGTCCACCAAA
		LB	GGTTTACCCAATAATACTGCGTCTT

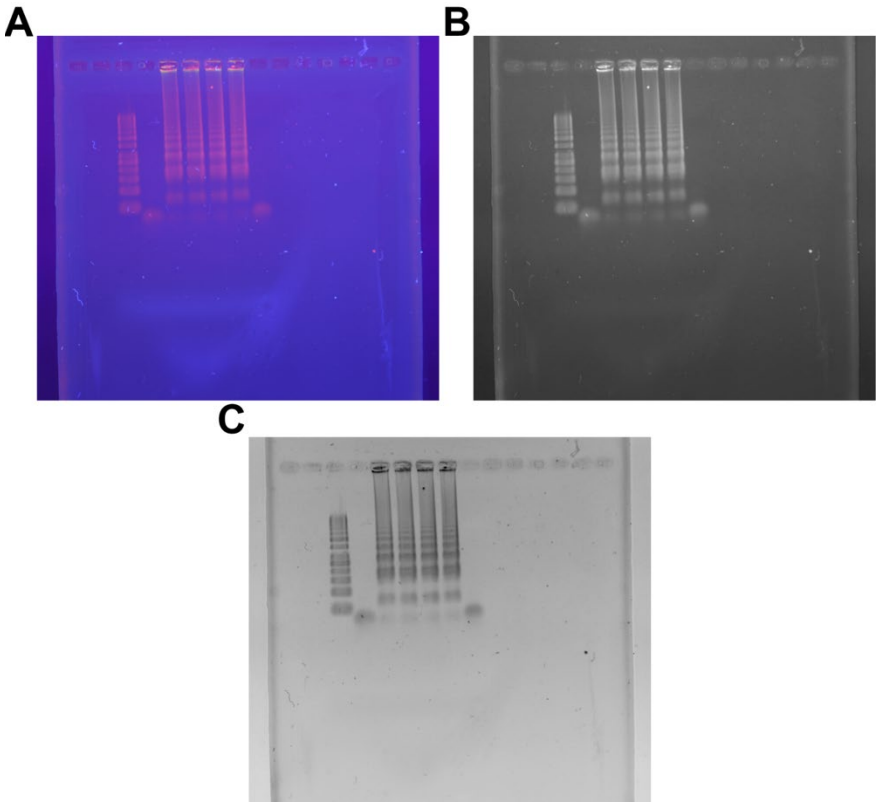


Figure S6.1: Full-length agarose gel from Figure 6.5B. (**A**) Original uncropped gel image. (**B**) Red-channel image of gel for ethidium bromide visualization and contrast. (**C**) Inverted red-channel image.

6.8 References

1. Zhu, N. et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727-733 (2020).

2. Sakurai, A. et al. Natural History of Asymptomatic SARS-CoV-2 Infection. *N Engl J Med* **383**, 885-886 (2020).

3. Woloshin, S., Patel, N. & Kesselheim, A.S. False Negative Tests for SARS-CoV-2 Infection - Challenges and Implications. *N Engl J Med* **383**, e38 (2020).

4. Larremore, D.B. et al. Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Sci Adv* **7**, eabd5393 (2021).

5. Patterson, E.I. et al. Methods of Inactivation of SARS-CoV-2 for Downstream Biological Assays. *J Infect Dis* **222**, 1462-1467 (2020).

6. Vogels, C.B.F. et al. SalivaDirect: A simplified and flexible platform to enhance SARS-CoV-2 testing capacity. *Med (N Y)* **2**, 263-280 e266 (2021).

7. Ranoa, D.R.E. et al. Saliva-Based Molecular Testing for SARS-CoV-2 that Bypasses RNA Extraction. *bioRxiv*, 2020.2006.2018.159434 (2020).

8. Rabe, B.A. & Cepko, C. SARS-CoV-2 detection using isothermal amplification and a rapid, inexpensive protocol for sample inactivation and purification. *Proc Natl Acad Sci U S A* **117**, 24450-24458 (2020).

9. Dao Thi, V.L. et al. A colorimetric RT-LAMP assay and LAMP-sequencing for detecting SARS-CoV-2 RNA in clinical samples. *Sci Transl Med* **12**, eabc7075 (2020).

10. Lee, J.Y.H. et al. Validation of a single-step, single-tube reverse transcription loop-mediated isothermal amplification assay for rapid detection of SARS-CoV-2 RNA. *Journal of Medical Microbiology* **69**, 1169-1178 (2020).

11. Joung, J. et al. Detection of SARS-CoV-2 with SHERLOCK One-Pot Testing. *N Engl J Med* **383**, 1492-1494 (2020).

12. Fozouni, P. et al. Amplification-free detection of SARS-CoV-2 with CRISPR-Cas13a and mobile phone microscopy. *Cell* **184**, 323-333 e329 (2021).

13. Myhrvold, C. et al. Field-deployable viral diagnostics using CRISPR-Cas13. *Science* **360**, 444-448 (2018).

14. Darnell, M.E., Subbarao, K., Feinstone, S.M. & Taylor, D.R. Inactivation of the coronavirus that induces severe acute respiratory syndrome, SARS-CoV. *J Virol Methods* **121**, 85-91 (2004).

15. Bhuyan, A.K. On the mechanism of SDS-induced protein denaturation. *Biopolymers* **93**, 186-199 (2010).

16. Cirin, D.M., Posa, M.M., Krstonosic, V.S. & Milanovic, M.L. Conductometric study of sodium dodecyl sulfate-nonionic surfactant (Triton X-100, Tween 20, Tween 60, Tween 80 or Tween 85) mixed micelles in aqueous solution. *Hemijska Industrija* **66**, 21-28 (2012).

17. Mandal, A.B., Nair, B.U. & Ramaswamy, D. Determination of the critical micelle concentration of surfactants and the partition coefficient of an electrochemical probe by using cyclic voltammetry. *Langmuir* **4**, 736-739 (2002).

18. Case, J.B. et al. Neutralizing Antibody and Soluble ACE2 Inhibition of a Replication-Competent VSV-SARS-CoV-2 and a Clinical Isolate of SARS-CoV-2. *Cell Host Microbe* **28**, 475-485 e475 (2020).

19. Kramer, M.F. & Coen, D.M. Enzymatic amplification of DNA by PCR: standard procedures and optimization. *Curr Protoc Mol Biol* **Chapter 15**, Unit 15 11 (2001).

20. Kaijalainen, S., Karhunen, P.J., Lalu, K. & Lindstrom, K. An alternative hot start technique for PCR in small volumes using beads of wax-embedded reaction components dried in trehalose. *Nucleic Acids Res* **21**, 2959-2960 (1993).

21. Kunkel, J. & Asuri, P. Function, structure, and stability of enzymes confined in agarose gels. *PLoS One* **9**, e86785 (2014).

22. Fujiwara, E., Cabral, T.D., Sato, M., Oku, H. & Cordeiro, C.M.B. Agarose-based structured optical fibre. *Sci Rep* **10**, 7035 (2020).

23. Jiang, M. et al. Development and Validation of a Rapid, Single-Step Reverse Transcriptase Loop-Mediated Isothermal Amplification (RT-LAMP) System Potentially to Be Used for Reliable and High-Throughput Screening of COVID-19. *Front Cell Infect Microbiol* **10**, 331 (2020).

24. Chow, F.W. et al. A Rapid, Simple, Inexpensive, and Mobile Colorimetric Assay COVID-19-LAMP for Mass On-Site Screening of COVID-19. *Int J Mol Sci* **21**, 5380 (2020).

25. Mohon, A.N. et al. Development and validation of direct RT-LAMP for SARS-CoV-2. *medRxiv*, 2020.2004.2029.20075747 (2020).

26. L'Helgouach, N. et al. EasyCOV : LAMP based rapid detection of SARS-CoV-2 in saliva. *medRxiv*, 2020.2005.2030.20117291 (2020).

27. Lee, S.H. et al. One-Pot Reverse Transcriptional Loop-Mediated Isothermal Amplification (RT-LAMP) for Detecting MERS-CoV. *Front Microbiol* **7**, 2166 (2016).

28. Gudnason, H., Dufva, M., Bang, D.D. & Wolff, A. Comparison of multiple DNA dyes for real-time PCR: effects of dye concentration and sequence composition on DNA amplification and melting temperature. *Nucleic Acids Res* **35**, e127 (2007).

29. Oscorbin, I.P., Belousova, E.A., Zakabunin, A.I., Boyarskikh, U.A. & Filipenko, M.L. Comparison of fluorescent intercalating dyes for quantitative loop-mediated isothermal amplification (qLAMP). *Biotechniques* **61**, 20-25 (2016).

30. Quyen, T.L., Ngo, T.A., Bang, D.D., Madsen, M. & Wolff, A. Classification of Multiple DNA Dyes Based on Inhibition Effects on Real-Time Loop-Mediated Isothermal Amplification (LAMP): Prospect for Point of Care Setting. *Front Microbiol* **10**, 2234 (2019).

31. Huang, W.E. et al. RT-LAMP for rapid diagnosis of coronavirus SARS-CoV-2. *Microb Biotechnol* **13**, 950-961 (2020).

32. Zhang, Y. et al. Rapid Molecular Detection of SARS-CoV-2 (COVID-19) Virus RNA Using Colorimetric LAMP. *medRxiv*, 2020.2002.2026.20028373 (2020).

33. Park, G.S. et al. Development of Reverse Transcription Loop-Mediated Isothermal Amplification Assays Targeting Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). *J Mol Diagn* **22**, 729-735 (2020).

34. Visseaux, B. et al. Evaluation of the QIAstat-Dx Respiratory SARS-CoV-2 Panel, the First Rapid Multiplex PCR Commercial Assay for SARS-CoV-2 Detection. *J Clin Microbiol* **58**, e00630-00620 (2020).

35. Ben-Ami, R. et al. Large-scale implementation of pooled RNA extraction and RT-PCR for SARS-CoV-2 detection. *Clin Microbiol Infect* **26**, 1248-1253 (2020).

36. Smyrlaki, I. et al. Massive and rapid COVID-19 testing is feasible by extraction-free SARS-CoV-2 RT-PCR. *Nat Commun* **11**, 4812 (2020).

37. Arizti-Sanz, J. et al. Streamlined inactivation, amplification, and Cas13-based detection of SARS-CoV-2. *Nat Commun* **11**, 5921 (2020).

38. Nagura-Ikeda, M. et al. Clinical Evaluation of Self-Collected Saliva by Quantitative Reverse Transcription-PCR (RT-qPCR), Direct RT-qPCR, Reverse Transcription-Loop-Mediated Isothermal Amplification, and a Rapid Antigen Test To Diagnose COVID-19. *J Clin Microbiol* **58**, e01438-01420 (2020).

39. Wyllie, A.L. et al. Saliva or Nasopharyngeal Swab Specimens for Detection of SARS-CoV-2. *N Engl J Med* **383**, 1283-1286 (2020).

40. Pan, Y., Zhang, D., Yang, P., Poon, L.L.M. & Wang, Q. Viral load of SARS-CoV-2 in clinical samples. *Lancet Infect Dis* **20**, 411-412 (2020).