UNIVERSITY OF CALIFORNIA,
IRVINE

Discovery and characterization of functional RNAs through *in vitro* selection and next generation sequencing

DISSERTATION

submitted in partial satisfaction of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in Pharmacological Sciences

by

Bao Tan Ho

Dissertation Committee:
Associate Professor Andrej Lupták, Chair
Assistant Professor Robert Spitale
Associate Professor Weian Zhao

2017

DEDICATION

to

my family and my loved ones

"There is nothing to writing. All you do is sit down at a typewriter and bleed."
- Ernest Hemingway

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

RNA	ribonucleic acid
DNA	deoxyribonucleic acid
cGMP	3´, 5´-cyclic guanosine monophosphate
cAMP	3´, 5´-cyclic adenosine monophosphate
GMP	guanosine monophosphate
GDP	guanosine diphosphate
GTP	guanosine triphosphate
IMP	inosine monophosphate
dPTP	2'-deoxy-P-nucleoside-5'-triphosphate
8-oxo-dGTP	8-oxo-2'-deoxyguanosine-5'-triphosphate
PAGE	polyacrylamide electrophoresis
RT	reverse transcription
PCR	polymerase chain reaction
UTR	untranslated region
FT	flow-through
WT	wildtype
SELEX	systematic evolution of ligands by exponential enrichment
SHAPE	selective 2´-OH acylation analyzed by primer extension
NMR	nuclear magnetic resonance
NAIM	nucleotide analog interference mapping
SPAAC	strain-promoted azide-alkyne cycloaddition

ACKNOWLEDGEMENTS

My graduate journey, such as it is, owes its completion to many extraordinary individuals.

I would like to acknowledge my graduate advisor Andrej Lupták for his guidance, support, and patience. It has been an incredible experience and I could not have asked for a better mentor.

I would like to thank my committee members: Drs. Robert Spitale and Weian Zhao.

My dissertation was supported by the National Science Foundation, the John Templeton Foundation, and the Department of Pharmaceutical Sciences at the University of California, Irvine.

Thank you to past and current members of the Lupták laboratory for their encouragements, joy and laughter, misery in company. I would like to highlight my gratitude towards Julio Polanco, Michael Wu, Dang Nguyen, Luiz-Fernando Passalacqua, Michael Vu, Michael Abdelsayed, Kelly Rotstan, Claire Chen, Drs. Randi Jimenez and Fabio Chizzolini. In addition, I would not be writing this thesis without standing upon the mountain of efforts from my undergraduate students: Mark Nakata, Jessica Pham, Carolyn Hua, and Eunice Oh. We have grown beyond mentor and mentees, and I would not have it any other way.

Special thanks to the friends I made in Irvine, Dr. Beverly Chou, Riann Egusquiza, Thomas Ford-Hutchinson, Camila Zanette, and Guilherme Puglia. I would like to thank my family: my parents and sisters for their unconditional love. Their kindness and encouragement have truly humbled me. Finally, my profound gratefulness towards Yunjin Cha, without whom my graduate journey, however still possible, would have been much less gentle and ineluctably more vacant.

CURRICULUM VITAE

BAO TAN HO

7316 Palo Verde Rd, Irvine, CA 92617 | 858-610-4039 | btanho@gmail.com

EDUCATION

University of California, Irvine 2011 – 2017
Department of Pharmaceutical Sciences, **Ph.D. Candidate**

University of California, San Diego 2007 – 2011
Department of Chemistry and Biochemistry, **Bachelor of Science**

SKILLS AND TECHNIQUES

- Technical Knowledge area: Directed Evolution, Biochemistry, Bioinformatics, Chemistry
- Expert in *in vitro* selection, transcription, reverse transcription, and amplification
- Biochemical characterization of nucleic acids: UV/Vis and fluorescent spectroscopy, polyacrylamide and agarose gel electrophoresis, isothermal titration calorimetry, high-performance liquid chromatography
- RNA structural probing assays: *i.e.* selective 2´-hydroxyl acylation analyzed by primer extension, in-line probing, and nucleotide analog interference mapping
- Experienced in high-throughput genomics and bioinformatics: next-generation sequencing, cluster computing, demultiplexing, mapping, variant calling, and genomic database navigation
- Synthesis, purification, and characterization of small molecules
- Tissue culture techniques (cell handling, MTT assays, flow cytometry, targeted delivery)
- Experience with animal: mouse subcutaneous and intravenous injection, blood collection, and ethical treatment of research subjects
- Data presentation and analysis: ChemDraw, PyMOL, Microsoft Office Suite, Adobe Creative Suite

EXPERIENCE

Graduate Student Researcher, Advisor: Professor Andrej Lupták 2012 – 2017
University of California, Irvine, Department of Pharmaceutical Sciences

- Identified and characterized genomic RNA aptamers for cyclic nucleotides using *in vitro* selection and structural probing
- Developed a high-throughput methodology for massively parallel characterization of selected aptamers, *i.e.* secondary structure and binding constants
- Synthesized chemical probes to investigate RNA structures
- Mentored four undergraduate researchers who performed independent selections for different targets, and trained for RNA structure probing
- Managed laboratory finances and inventories necessary for experiments
- Supervised laboratory safety protocols, revised and updated standard operating procedures, observed EH&S and OSHA safety requirements

Instructor of Record, Biochemistry Lecture 2016
University of California, Irvine, School of Biological Sciences
- Lectured undergraduate students on structural biochemistry, metabolism, macromolecules, and related cellular processes
- Prepared original course material, including lecture slides, syllabus, exams, and homework
- Adopted computer-assisted learning, and interactive study sessions

Teaching Assistant, Medicinal Chemistry Lecture and Laboratory 2013 – 2017
University of California, Irvine, Department of Pharmaceutical Sciences
- Instructed laboratory procedures, demonstrated experimental techniques, and graded reports
- Led discussion and review sessions on lecture material, served as liaison between course instructors and other teaching assistants
- Organized instructional meetings, revised laboratory protocol, maintained reagents and equipments, restocked and updated inventory for necessary experiments

Undergraduate Student Researcher, Advisor: Professor Thomas Hermann 2009 – 2011
University of California, San Diego, Department of Chemistry and Biochemistry
- Synthesized novel RNA binders targeting the Internal Ribosome Entry Site of Hepatitis C Virus RNA
- Performed small molecules derivatization, peptide coupling, purification and characterization

Undergraduate Math and Sciences Tutor 2007 – 2008
San Diego Mesa Community College, Tutoring Center
- Helped college students develop problem solving skills, explained mathematical, biological, chemical, and physical course materials

PUBLICATIONS

1. Abdelsayed, M.*, **Ho, B. T.***, Vu, M.*, Polanco, J. Spitale, R., and Lupták, A. (2017) Multiplex aptamer discovery through Apta-Seq and its application to human-genomic SELEX of ATP aptamers. ACS Chem. Biol. acschembio.7b00001

2. **Ho, B. T.**, Polanco, J., Jimenez, R., and Lupták, A. (2014) Discovering human RNA aptamers by structure-based bioinformatics and genome-based in vitro selection. Methods Enzymol. 549, 29–46.

3. Dutta, S., Dibrov, S. M., **Ho, B. T.**, Higginson, C. J., and Hermann, T. (2012) Synthesis and Crystal Structure of a Novel Heterocycle, 2-Oxa-4,7-Diazabicyclo[3.3.1]Non-3-Ene. J. Chem. Crystallogr. 42, 119–129.

4. Dutta, S., Higginson, C. J., **Ho, B. T.**, Rynearson, K. D., Dibrov, S. M., and Hermann, T. (2010) 1,3-Diazepanes of Natural Product-Like Complexity from Cyanamide-Induced Rearrangement of Epoxy-δ-lactams. Org. Lett. 12, 360–363.

** indicates co-first authorship*

HONORS AND AWARDS

RNA Society Travel Fellowship	2017
Amgen Scholar Research Fellowship	2010
Erion Foundation Scholarships	2009 – 2010
UC San Diego Provost's Honors	2008 – 2009

PRESENTATIONS

RNA 2017, Prague, Czech Republic 2017

Ho, B. T., Pham, J. V., Nakata, M. T., Luptak, A. *In vitro selection of cyclic nucleotide aptamers from a human genomic library* **(Poster)**

Vertex Research Day, Irvine, CA 2017

Ho, B. T., Hua, C., Pham, J. V., Nakata, M. T., Luptak, A. *In vitro selection of cyclic nucleotide aptamers from a human genomic library* **(Oral Presentation)**

UC Irvine Directed Evolution Meeting, Irvine, CA 2017

Ho, B. T., Hua, C., Pham, J. V., Nakata, M. T., Luptak, A. *In vitro selection of cyclic nucleotide monophosphate aptamers from genomic libraries* **(Oral Presentation)**

UC San Diego Chemistry Symposium, La Jolla, CA 2011

Ho, B. T., Dutta, S., Hermann, T. *Synthesis of HCV IRES ligands based on a diazepane scaffold* **(Poster)**

Amgen Scholar Symposium, Los Angeles, CA 2010

Ho, B. T., Dutta, S., Higginson, C. J., Hermann, T. *Synthesis of novel RNA binders based on diazepane scaffold* **(Oral Presentation)**

SIGNIFICANT EXTRACURRICULAR ACTIVITIES

California Community College Internship Program, Irvine, CA 2016

- Engaged in lesson and curriculum planning, participated in academic senate and self-governance, lectured community college students on algebraic and trigonometric course materials

Northern California Junior Science and Humanities Symposium, CA 2016 – 2017

- Volunteered to serve as judge on the biomedical panel for Northern California high school science oral presentations (9^{th} to 12^{th} grade)

Irvine Unified School District Science Fair, Irvine, CA 2013 – 2015

- Volunteered to serve as judge for local student science posters (6^{th} to 8^{th} grade)

UC San Diego Vietnamese Student Association, Academic Chair, La Jolla, CA 2009 – 2010

- Promoted scholarship among members, encouraged academic retentions, and organized student peers tutoring
- Outreached and engaged with local high school students of diverse, high-risk backgrounds, and raised awareness towards higher education

ABSTRACT OF THE DISSERTATION

Discovery and characterization of functional RNAs through *in vitro* selection and next generation sequencing

by
Bao Tan Ho
Doctor of Philosophy in Pharmacological Sciences
University of California, Irvine, 2017
Professor Andrej Lupták, Chair

Biochemical properties of ribonucleic acids enable their instrumentality in physiological functions. RNA sequence serves as the intermediary of the genetic code, while secondary and tertiary structures are responsible for RNA's involvement in regulation and catalysis. As such, discovery and biochemical characterization of RNAs promise to illuminate the relationship between an RNA's sequence, structure, and functions, thus provide insights into their biological significance. Efforts in functional RNA discovery often turn to *in vitro* selection or SELEX (systematic evolution of ligand by exponential amplification) to identify RNAs of a functional trait of interest from a diverse pool of sequences. Selected RNA species are subsequently sequenced and characterized for binding affinity or catalytic activity. Given the size and diversity of the enriched pool at the end of a selection, elucidation of RNA sequence and structure remains a labor-intensive process. The work presented here focuses on massive parallel identification and characterization of functional RNAs emerged from *in vitro* selections through a bioinformatics pipeline known as Apta-Seq.

Enriched sequences from multiple selections were examined via high-throughput sequencing coupled with selective 2´-hydroxyl acylation analyzed by primer extension (SHAPE).

These sequences are the output of *in vitro* selections of RNA aptamers for cyclic-nucleotide monophosphate from genomic DNA libraries. High-throughput analyses of sequences emerged from genomic *in vitro* selections leads to the discovery of an abundant aptamer with affinity and selectivity towards 3´, 5´-cyclic guanosine monophosphate. Independent structure probing experiments performed on this aptamer produced a binding constant at 5.9 μM and identified key nucleotide residues involved in binding of the ligand. These results established correlation between high-throughput SHAPE data and in-line probing experiments, thus demonstrating Apta-Seq as a sensitive method to massive parallel investigate RNA sequence identity and secondary structure.

Moreover, selection of multi-turnover ribozymes of strain-promoted azide-alkyne catalysis was explored in this work. However, the unsatisfactory output of the selection encourages revision of the selection strategy. Nonetheless, the progress of the selection has the potential to address the challenge of RNA *in vivo* imaging.

Chapter 1

Discovery and characterization of novel RNA aptamers and ribozymes

Importance of RNA aptamers in biology

Nucleic acids aptamers are single–stranded, oligomeric DNA or RNA molecules that bind to their target ligands at high affinity and specificity. While DNA aptamers exhibit a greater degree of chemical stability, RNA aptamers offer a wider range of structural complexity, with secondary elements such as non-canonical base-pairing, stem, hairpin loop, bulge, internal loop, and junction, as well as tertiary motifs, *e.g.* triple helices and helix-loop interactions[1]. The diversity in conformations of RNA aptamers translates into a distinguished affinity repertoire, where these molecules can bind a wide array of ligands, ranging from ions, proteins, to cells[2–4]. RNA aptamers naturally exist as part of the larger construct known as riboswitch, where changes in structure of the RNA aptamer domain upon ligand binding induce downstream expression modulation[5,6]. As RNA aptamers remain versatile in their affinity and specificity, riboswitches, by extension, can be categorized into different classes based on their substrates. These RNA sensor elements have been found mostly in bacteria, and can bind vitamins, purines, sugars, among many other metabolites. The investigation of RNA aptamers not only illuminates the identity of novel riboswitches, but also provides further insights on gene expression pathway in many organisms.

Aptamer Domain in Riboswitches

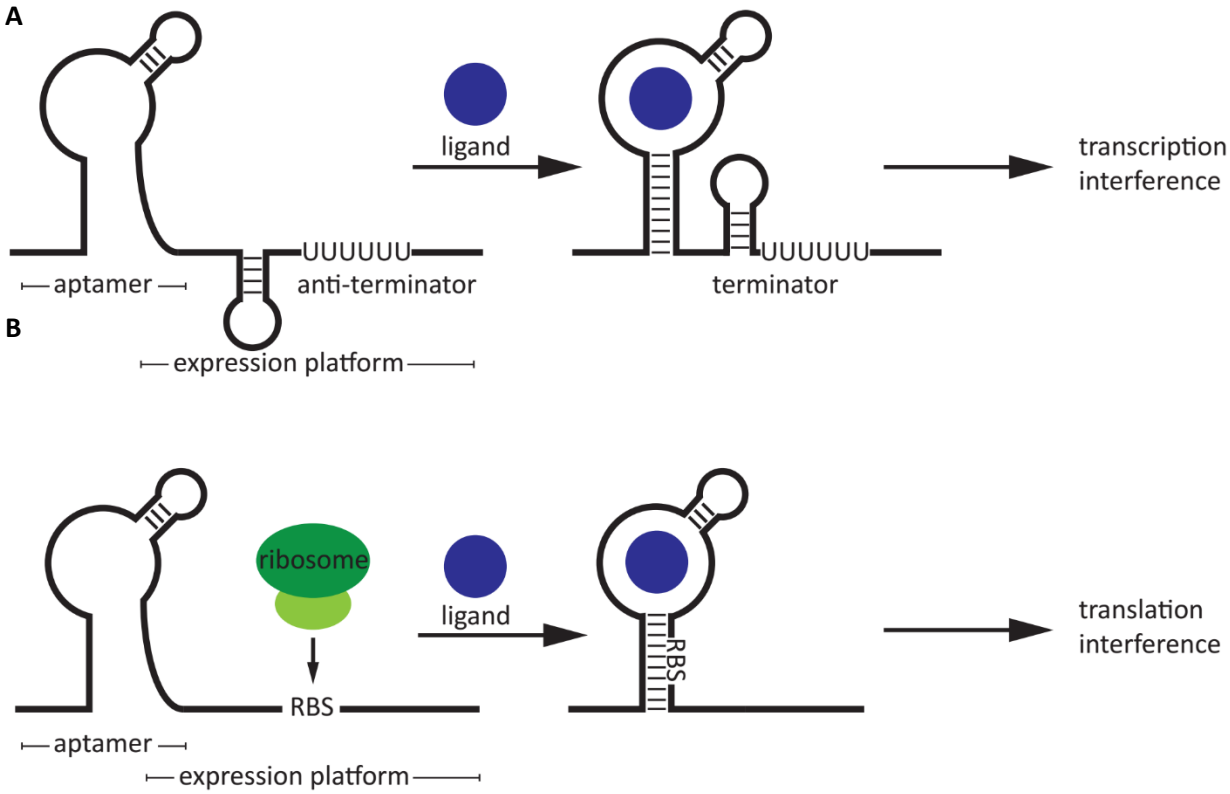


Figure 1-1: Riboswitch mechanism. (A) Effector molecule binds to the aptamer domain of the riboswitch, turning off transcription (B) Riboswitches' mechanism can also regulate translation.

Riboswitches, found in the 5´-untranslated region (UTR) of mRNAs, typically contain a ligand-specific aptamer domain, followed by an expression platform that enables downstream regulation. In order for the aptamer domain to remain specific, it must be able to distinguish between various metabolite derivatives, and maintain an appropriate affinity for the effector molecule. The expression platform contains a switching sequence, whose folding affects two mutually exclusive secondary structures, thereby influencing subsequent transcriptional or translational mechanism. The aptamer domain, that encompasses the switching sequence, is typically highly evolutionary conserved. For example, as previously demonstrated by Grundy and Henkin, the S-adenosylmethionine sensing S-box motif is conserved among *Staphylococcus*

arreus metK, *Clostridium acetobutylicum*, and *Bacillus subtilis*[7]. A conserved RNA motif was also reported to be involved in the regulation of bacterial riboflavin synthesis[8]. The requirements for high affinity and conservation have led to classification of different riboswitch classes, with effector molecules ranging widely, from the first riboswitch reported with coenzyme B_{12}, to multiple *S*-adenosylmethionine subclasses[5,9]. As these regulatory mechanisms depend on a stepwise process during which sequestered metabolite sensing motifs propagate subsequent expression modulation, the aptamer domains arguably serve as logic gates to these events. Thus aptamer discovery becomes a gateway to riboswitch identification, and investigation on gene expression.

***In vitro* selection as an approach to identify novel aptamers**

Aptamers were first identified and reported independently by the Gold and Szotak laboratories in 1990[10,11]. An *in vitro* selection for aptamers requires a diverse starting library, either synthetic or genomic DNA sequences. Species within the library are enriched for a property of interest by iterative rounds of partition and amplification. An initial high diversity DNA library is an essential prerequisite to a selection. While initial reports of the methodology by Gold and Szostak boasted a diversity of 65,536 and 10^{10} molecules respectively[10,11], contemporary *in vitro* selections have been shown to accommodate more diverse starting pools, starting with 10^{16} sequences[12]. This diversity can be achieved by large combinatorial synthesis of oligonucleotides. A typical RNA aptamer selection begins with a transcription of the DNA library into single-stranded RNAs. The RNAs can be tagged with radiolabels or fluorophores to allow for monitoring enrichment. These sequences are screened for affinity towards an effector molecule of interest, isolated, reverse transcribed, and amplified for the next round of selection.

Due to the iterative nature of this method, a selection can lead to significant enrichment of sequences until desired affinity and specificity of the aptamers are achieved.

While conceptually simple, *in vitro* selection for aptamers experiments differ tremendously depending on the target molecules, the initial libraries, and the compartmentalization of species during screening. Selections were performed with classical affinity chromatography experiments, water-in-oil apparatus, to more sophisticated approaches utilizing microarrays and microfluidics[13,14]. Some of the previously reported targets for aptamers include small molecules such as nucleotides and fluorophores. Synthetic aptamers, enriched from a random library prepared by phosphoramidite synthesis, have enjoyed immense applications in *in vivo* imaging, *e.g.* Spinach, Broccoli, and Mango aptamers[15–17]. Meanwhile, genomic aptamers, selected from fragmented genomes, have a broad spectrum of targets, including transcription factors, proteins, small molecules, and other target RNAs[18–20]. As genomic selections begin with libraries derived from genomic DNAs, this approach can be applied to investigate low abundant RNAs, such as transcripts from silenced domains, as well as transients RNA species that might elude transcriptomic recovery and analysis[21].

In vitro selections have been performed to elucidate aptamers for small molecules. The first selection conducted by Ellington and Szostak had isolated RNA aptamers of organic dyes[11]. Following their success, many other aptamers against small molecules were reported, as it was suggested that aptamers serve as ideal molecular recognition probes, due to their significantly high selectivity towards small molecules[22,23]. In addition, small molecules enjoy the ease of diffusion across cell membranes, hence playing important roles in cellular functions. As part of riboswitches, the aptamer domains typically, and strongly, bind to small molecules and ions,

with the guanine riboswitch having a K_D of 5 nM, and the thiamine pyrophosphate within the picomolar range[24,25]. As an alternative to small molecule aptamer selections, direct approach of selecting for riboswitches have also been performed, as Martini et. al. have demonstrated[26,27]. However these investigations adopted a synthetic starting library, as such, the resulting riboswitches are not naturally–occurring.

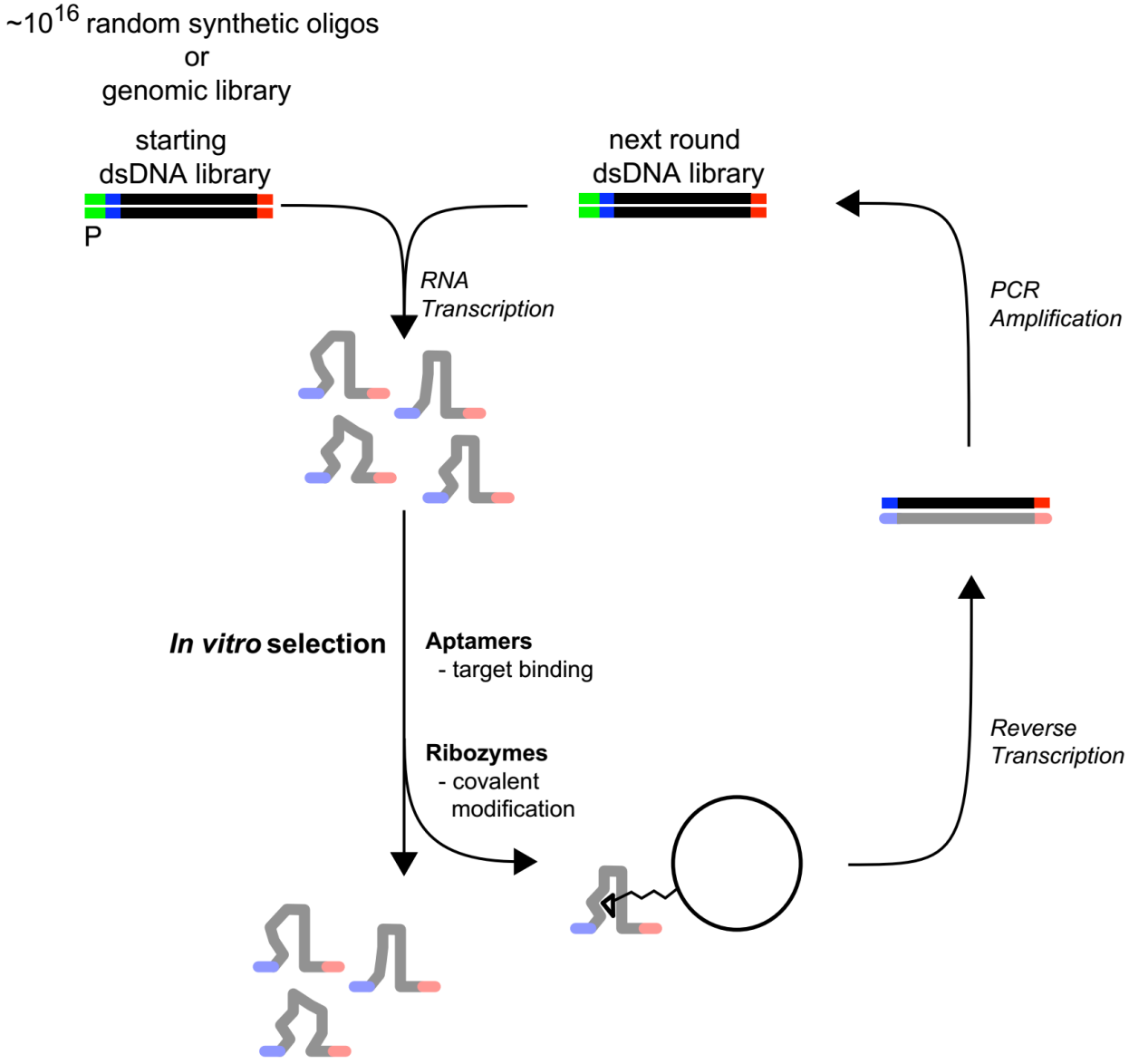


Figure 1-2: *In vitro* selection scheme

Genomic selections of small molecules binding aptamers, specifically nucleotides, have yielded adenosine and guanosine triphosphate aptamer, both from human genomic libraries[20,28]. In one of the earlier *in vitro* selections performed by Sassanfar et. al.[29], however started with a random library, an ATP-binding RNA motif was identified. Vu et. al., via computational and experimental approaches, have selected the same motif from a human genomic library, as well as demonstrated its convergent evolution based on its occurrences in other mammalian and bacterial genomes[20]. The aptamer was reported to bind ATP and AMP, while showing selectivity against dATP and GTP. While the biological function of the aptamer is still under investigation, the aptamer has been shown to span the intron of the FGD3 gene, suggesting a regulatory role in GTPase-dependent exocytosis. A GTP-binding motif was also reported to be widespread in eukaryotic genomes. The motif consists of G-quadruplexes and was suggested to play a role in transcription, RNA processing, and translation[28]. Due to the abundance of nucleotides and nucleotide derivatives involved in cellular processes, genomic selections for nucleotide-binding aptamers have a tremendous potential to elevate our understanding of biological mechanisms.

In vitro selection as a strategy to identify new ribozymes

The term 'ribozyme' was initially incepted as an RNA specie was demonstrated to perform enzymatic activity in *Tetrahymena thermophila*[30] of self-splicing and ligating activity. Subsequent examples of catalytic RNAs have propelled the RNA world hypothesis, where RNA was postulated to be the catalytic entity of biochemical processes before the dominance of peptide-based enzymes[31,32]. Over the years, more naturally-occurring ribozymes have become well validated and characterized, as well as categorized into different families based on the catalytic motifs. These motifs facilitate site-specific phosphodiester scission via either *cis* or

trans mechanism, by forming secondary and tertiary structures that allow a 2´-hydroxyl to perform a nucleophilic attack on the adjacent phosphodiester, giving rise to a 2´, 3´-cyclic phosphate and a 5´-hydroxyl cleaved products. This enzymatic singularity of naturally occurring ribozyme is responsible for protein synthesis, as well as RNA processing in the cell. For example, the glmS ribozyme has been shown to be involved in regulating the flux of glucoseamine-6-phosphate in Gram positive bacteria, while HDV-like ribozymes are distributed widely in nature and play many biological roles[33,34].

Artificial ribozymes, evolved with *in vitro* selection, are capable of a wide array of chemical transformations, in addition to phosphodiester scission. Earlier efforts in the field have yielded ribozymes possessing capabilities such as nucleotide synthesis from a sugar and nitrogenous base, amide bond synthesis, acyl transfer, as well as well-established chemistries such as Michael and Diels-Alder addition[35–39]. These selections began with a highly diverse random pool of RNA sequences, among which the species with desired catalytic activity were isolated, reversed transcribed, then amplified for the next round of selection. The designs of a ribozyme selection might fundamentally differ from that of an aptamer SELEX. For instance, while the criteria of success for an RNA aptamer selection depends on affinity of selected aptamers towards the target ligand, as demonstrated by the amount of RNAs present in the elution fraction, the chemical sophistication of an artificially evolved ribozyme might not require a high degree of affinity towards its substrates. The experiments were often set up to provide binary output to distinguish inactive from active sequences. In the selection for nucleotide synthesis ribozyme, the active sequences were expected to contain a covalently linked 4S-uracil, which was subsequently reacted with biotinylated iodoacetyl for downstream separation[35].

Similarly, the Diels-Alder ribozyme selection, performed in water-in-oil emulsion, relied on a biotinylated dienophile to pull down the catalytically active RNA sequences[39]. Due to the expanding chemical repertoire of artificial ribozymes, their selection strategies must also increase in sophistication, so as to accommodate novel catalytic RNAs' structures and mechanisms.

Emerging tools in chemical probing for RNA structure elucidation

In vitro selection of RNA aptamers and ribozymes are often accompanied by RNA structure characterization, as the function of an RNA molecules is very well connected to its secondary and tertiary conformation. As RNA structures are dynamic, RNA structure probing techniques are also expanding in their applicability. In essence, a chemical probing experiment of RNA consists of treating the RNA sample with a chemical reagent, upon which a readout of signals indicating modified sites is generated. These chemical or enzymatic modification sites provide insights towards the RNA of interest's secondary and tertiary structure, as each type of RNA topology is susceptible to certain interaction facilitated by the probe. Some of the earliest reports of RNA chemical probing include dimethyl sulfate, which modifies the N7 of guanosines and N3 positions of cytidines when these bases are not involved in RNA structural interactions, thus the sites available for modification are suggested not to participate in base pairing of the RNA structure[40]. RNAs can also be probed using enzymatic cleavage assays utilizing RNases. Due to the high specificity of enzymes, these technique offer precise modification, *i.e.* single-stranded or double-stranded break, onto the RNA of interest[41]. However, the conditions required for these enzymes to ensure their activities might not coincide with the condition at which the RNA under investigation is active.

Another aspect of RNA characterization is solvent accessibility. As a RNA molecule forms higher order structure, regions of the RNA becomes shielded, and less exposed to the surrounding aqueous environment. Hydroxyl radical footprinting was developed as an approach aimed to investigate solvent accessibility of each nucleotide in an RNA molecule[42]. The radical is most commonly generated by using hydrogen peroxide or iron (II)–EDTA complex. The mechanism of strand breakage relies on the abstraction of a hydrogen atom from a deoxy carbon of the ribose. Unlike the previously described enzymatic digestion by RNases, this breakage is not site specific due to the highly reactive nature of the hydroxyl radical, and thus solely depends on that digested nucleotide's exposure to the solvent. This technique still remains relevant in face of upcoming alternative in RNA structure probing, and is widely used to investigate RNA-protein three dimensional interactions[43,44].

More contemporary techniques in chemical mapping of RNA are in-line probing and selective 2´-hydroxyl acylation analyzed by primer extension (SHAPE)[45,46]. In-line probing takes advantage of RNAs' susceptibility to degradation[47]. The 2´-oxygen, once properly situated, can perform a nucleophilic attack on the adjacent phosphorus center of the phosphodiester bond, yielding a 2´, 3´-cyclic 5´ leading fragment, and a 5´-OH remainder. These fractions can be electrophoresed to provide single nucleotide resolution of the RNA's degradation pattern. The rate of degradation depends on the local structure surrounding a nucleotide. As a nucleotide is base paired to another, it becomes rigid, and consequently less likely to perform the nucleophilic substitution. Whereas a nucleotide within a single-stranded region of a folded RNA might sample a larger range of motions, and is more prone to cleavage. As such, in-line probing can be used to investigate the dynamics of folded RNA, and has been applied extensively to

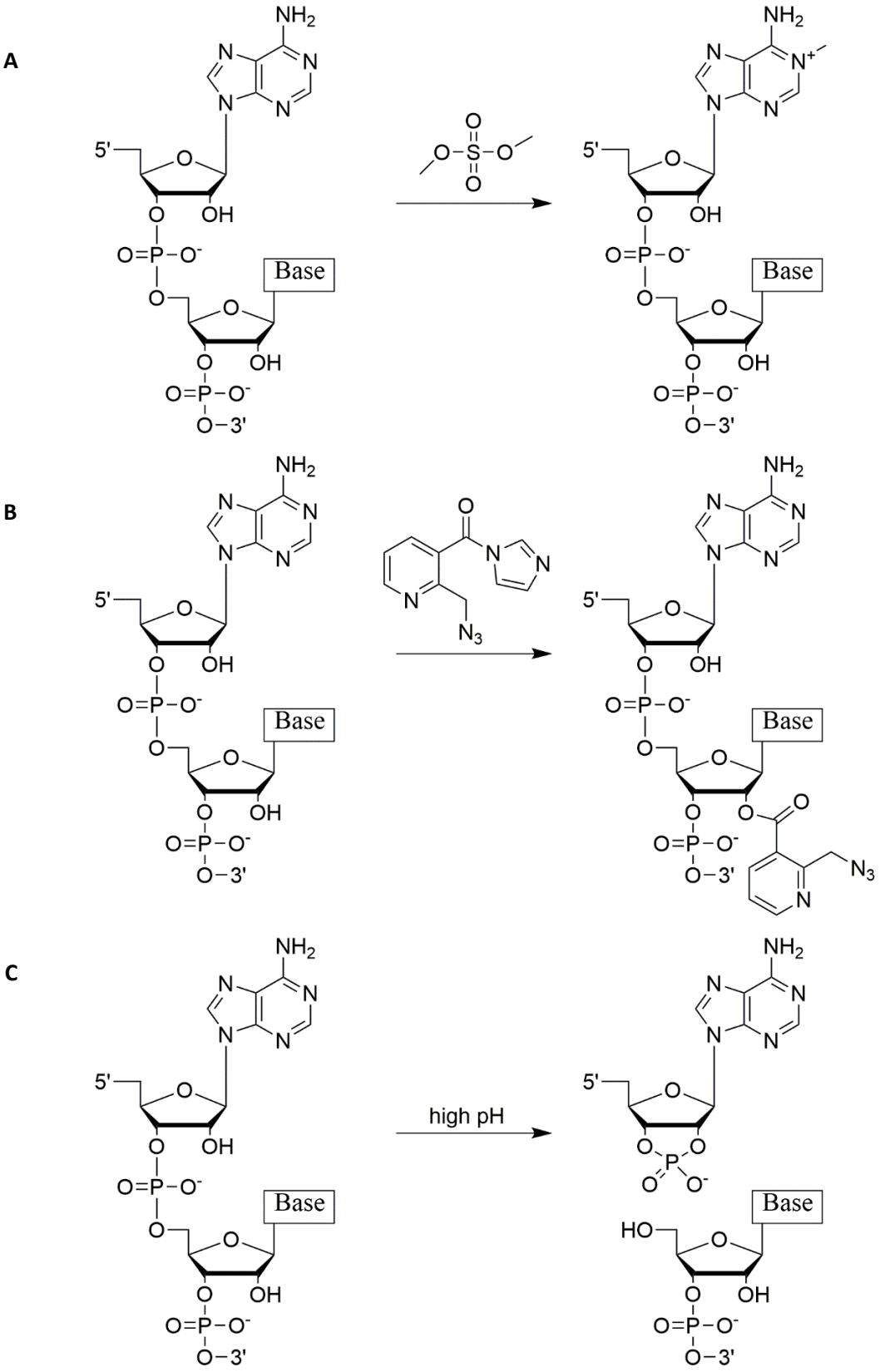


Figure 1-3: Chemical probing techniques for RNA structure characterization. (A) Dimethylsulfate. (B) Selective 2´-hydroxyl acylation analyzed by primer extension. (C) In-line probing.

elucidate riboswitches' structures, since riboswitches change their folded structures upon binding to their target ligands, causing their in-lining to be ligand-dependent

Selective 2´-hydroxyl acylation analyzed by primer extension is similar to in-line probing in that the technique also reacts with the backbone of the RNA, specifically the alcohol functionality at C2´ position (Fig 1-3). The functional group, depending on the local nucleotide flexibility or dynamics, would undergo acylation by a SHAPE reagent, at which the acylated site will terminate reverse transcription one nucleotide downstream, creating a single nucleotide resolution of RNA structural output[46,48]. Over the years, several SHAPE reagents have been developed to expand the reactivity and sensitivity of acylation in order to accommodate the timescale of RNA folding, which has proven to be particularly useful when analyzing *in vivo* transcriptome or long distance RNA interactions[49,50]. SHAPE reagents can be retrofitted by altering the electrophilicity of the carbonyl carbon, making the reagents more susceptible to the nucleophilic attack by the 2´-hydroxyl. Since the inception of the first SHAPE reagent NMIA, faster acting probes were reported to allow for shorter experiments[51]. Even though both rely on reverse transcription readout, SHAPE enjoys an incredible advantage over DMS probing by reacting broadly with all four nucleotides of the RNA. Furthermore, while in-line probing is an attractive option, it should only be used for small sized transcripts (~200 nucleotides), whereas a SHAPE reagent can provide readout for an RNA regardless of its length[48].

Bioinformatics as an approach to identify and characterize RNAs

In-line probing and SHAPE are powerful techniques to determine the structure of aptamers and ribozymes. In a broader sense, these techniques have been coupled to *in vitro* selection of such RNAs to probe the secondary structure of selected sequences[52,53].

Furthermore, the advances in next generation sequencing has enabled rapid identification of short DNA sequences, which in turn have exponentially improved the effectiveness of SHAPE[54,55]. These protocols can generate large datasets that provide parallel analysis of RNA sequence and structure identity. To corroborate the output of these protocols, structure prediction pipelines have increasingly incorporated SHAPE reactivities into their calculation[56,57]. These pipelines typically require the a basic toolkit for processing, sorting, and aligning DNA reads from a next generation sequencing (NGS) output, such as bowtie2, cutadapt, and samtools[58–60]. As the reads generated by NGS contain specific constant regions used for annealing[61], these cassettes usually have to be trimmed away, so that only the sequences of interest remain. Although the setups for these pipelines typically follow a rigid format, as NGS can only accommodate a limited set of primers, the experiment can be adopted to investigate a variety of research interests[62,63]. In addition, NGS technologies, however massively parallel, can only read short sequences, therefore consensus sequences for larger constructs need to be generated[64], and play a pivotal role in structural investigation of larger RNAs[65].

Aptamer and ribozyme SELEX are now employing these strategies to overcome the inherent difficulties in *in vitro* selection: size and diversity of final output. These various techniques are discussed in the approaches and results of the selections mentioned within this dissertation. Improvements and bypasses to these protocols, followed by their implications, are also elaborated in great details.

Chapter 2

In vitro selection of cyclic guanosine monophosphate aptamers from a human genomic library

Introduction

RNAs have been shown to demonstrate an active role in cellular processes. The discovery of riboswitches, a common element found predominantly in bacterial mRNAs, have tremendously reinforced the RNA World hypothesis, the proposal that life in the primordial state has relied on RNA as the initial regulatory and catalytic unit of biological transformations. Current efforts exploring riboswitches share the mutual understanding that a riboswitch is composed of an aptamer domain, and an expression platform. The aptamer domain, responsible for the riboswitch's ligand recognition, displays high affinity and specificity towards the effector molecule. Earlier reports have included target ligands for riboswitches ranging from purines, protein coenzymes, amino acids, phosphorylated sugars, to even inorganic chemical species[66]. Contemporary research in the aptamer field also resulted in artificial riboswitches, giving rise to synthetic elements regulating biological circuits[67]. Within these systems, natural or synthetic, the aptamer domain remains the central figure in the riboswitch's effectiveness, mainly due to their structure robustness and integrity in the presence, or absence, of the target ligand. Specifically, the sequence encompassing the aptamer and expression domain, known as the switching sequence, has to adopt two mutually exclusive direct folding that represent the binary affair of regulation. As such, aptamer discovery remains a direct approach to riboswitch identification and engineering.

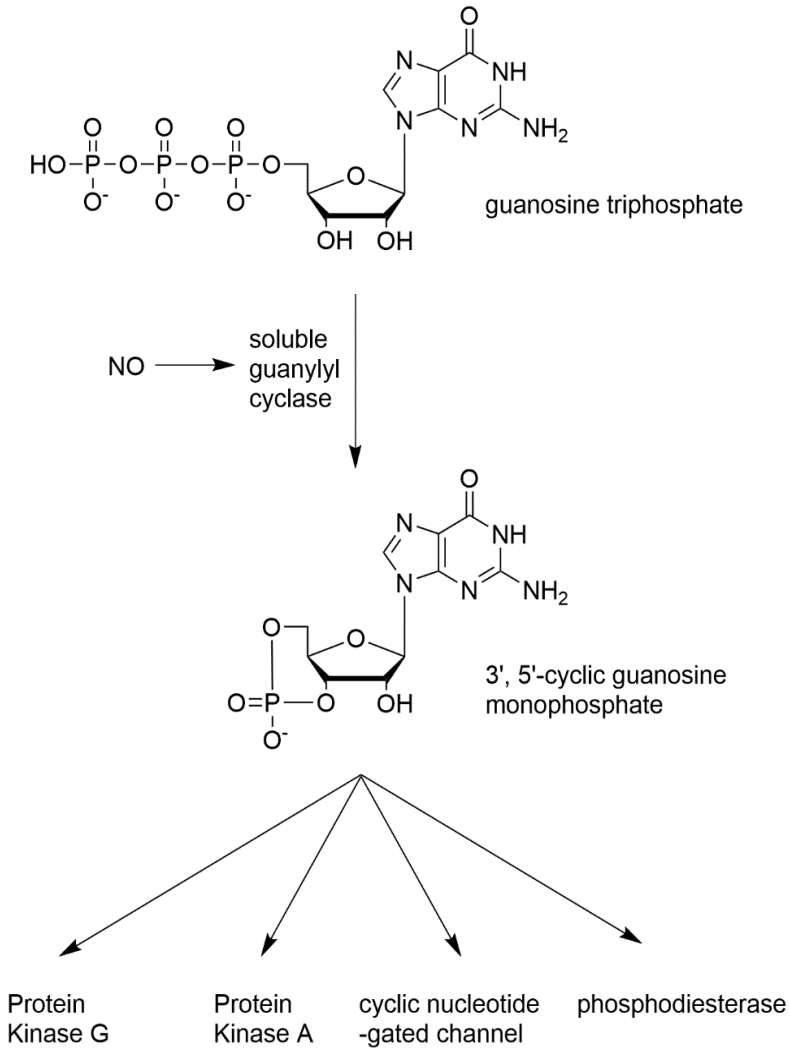


Figure 2-1: Biosynthesis and regulation of 3´, 5´-cyclic guanosine monophosphate

Among the most widely studied secondary messenger is 3´, 5´-cyclic guanosine monophosphate (cGMP). The nitric oxide and cGMP signal transduction pathway is involved in important physiological processes such as cell differentiation, proliferation, cell membrane permeability, and changes in smooth muscle tone[68,69]. cGMP is produced from guanosine triphosphate upon soluble guanylyl cyclase binding to nitric oxide, which is synthesized from nitric oxide synthase[70,71]. cGMPs can also be synthesized from the triphosphate precursor by transmembrane particulates, which are activated by natriuretic peptides[72]. The signaling molecule primarily targets cGMP-dependent protein kinase (PKG), cAMP-dependent protein kinase (PKA), cyclic nucleotide-gated channel, and phosphodiesterase. Decreased levels of cGMP have been linked to hypertension found in the rats, which are characterized by the lack of activity in one of the isoforms of cGMP-dependent protein kinase.

The effector ligand also interacts with cyclic nucleotide-gated channel, particularly for photoreceptor functions[73]. Rod photoreceptors are responsible for capturing light. The absorption of a photon by rhodopsin induces a conformational change of transducin, a guanosine nucleotide-binding protein, which subsequently activates phosphodiesterase (PDE) complex that hydrolyzes cyclic GMP. In the absence of light, the lack of PDE-meditated degradation of cyclic GMP leads to its high intracellular concentration, the signal molecule binds to cyclic nucleotide-gated channels and maintains them in the open configuration. Thus cGMP plays a direct role in depolarization of photoreceptors. Moreover, the cyclic nucleotide binding site of PKA and PKG are highly similar, cAMP and cGMP have been observed to cross–activate these proteins, specificity not taken into consideration[74].

In addition, another signaling molecule that shares many of cGMP's chemical identities is cyclic-di-guanosine monophosphate (cdGMP). Cyclic-di-GMP is a circular guanosine dinucleotide that serves as second messenger in bacteria. It is synthesized from two GTP molecules by diguanylate cyclase, and is involved in a wide array of functions, ranging from production of extracellular polysaccharides, adhesive proteins, to biofilm definition[75,76]. Cyclic-di-GMP riboswitches have been reported in multiple bacterial systems. Several aptamer domains have been found in the 5´ UTR of phosphodiesterase, whereas another class of the c-d-GMP has been located adjacent to group I intron[77,78]. These riboswitches were identified using bioinformatics, and have been shown to play a pivotal role in regulation of pilus formation and alternative splicing. Furthermore, aptamer domains recognizing purine derivatives, *e.g.* the guanine riboswitch, have been well studied and characterized[79]. These considerations compelled the hypothesis of the existence of an analogous regulatory model responsive to a

cyclic guanosine moiety at an RNA level within eukaryotes. Therefore, discoveries of riboswitches, and aptamers by extension, towards cGMP in eukaryotic genomes promise to expand on the biological functions of the second messenger, as well as provide insights on transcription-based regulation in higher organisms.

Previous efforts in genomic SELEX have isolated aptamers with affinity and specificity towards nucleotides, single-stranded RNA elements, and proteins[19,20,28,80]. These studies typically relied on *in vitro* selection as the primary method of aptamer identification, where a highly diverse genomic library is exposed to a target ligand, followed by a partition to collect the bound sequences and discard the non-binders. This process is repeated until the initially diverse library converges onto a smaller subset of sequences, where each aptamer can be individually characterized. Specifically, genomic SELEX performed with the human genomic libraries have identified RNA aptamers for adenosine (triphosphate), guanosine triphosphate, and a transactivating responsive RNA element[19,20,28]. The adenosine triphosphate aptamer in particular had its binding site reported previously by Sassanfar and Szostak, where the aptamer was demonstrated to bind adenosine monophosphate, triphosphate, and adenosine in the micromolar range, as well as displayed specificity against other nucleotides[29]. This binding loop has been confirmed in the human adenosine aptamer, suggesting that the motif is robust, as well as conserved among many organisms. Other genomic SELEX have isolated natural-occurring aptamers for different ligands as well. However, the challenge present within these selections persists: the size and diversity of sequences within a selected library. As biochemical characterization of aptamers remains a rigorous and precise process, elucidation of their secondary structures, binding sites, and K_Ds become tedious and time-consuming. With the

development of next generation sequencing, several laboratories have initiated high-throughput analyses of nucleic acids' sequences and affinities[62,81,82]. However these methods are limited in their applications towards structure predictions and determination of binding constants as they required covalent attachment of the target ligands to solid support, such that their analytical output was biased against sequences with high dissociation kinetics. Thus, driven by the possible additional biological novelty of 3´, 5´-cyclic guanosine monophosphate aptamers, and the tremendous potential of high-throughput characterization of RNAs, a protocol was developed to overcome the aforementioned obstacles in functional nucleic acids discovery. The design, procedures, results, and pitfalls of the identification and characterization of cGMP aptamers are discussed herein.

Designing an *in vitro* selection for cGMP aptamers from a human genomic library

An RNA aptamer selection, while elegant and simplistic in its construction, relies on a specific set of conditions for its success. When performing a genomic SELEX, the initial DNA library must be highly diverse and representative of the genome(s) of interest[83]. Genomic DNA library as a starting point of a selection reduces the bias against low-abundant and tissue-specific transcript, as opposed to a pool of sequences generated from total RNA[28]. The identification of the human adenosine aptamers completed previously in our laboratory utilized a diverse and robust pool composed of human genomic DNA[20,84]. This DNA library boasts a single–nucleotide coverage of the human genome, resulting in a starting diversity of 10^{10} sequences. The nucleic acid species present in this pool contain a genomic region flanked by fixed cassette required for (reverse-) transcription and amplification which total to a double

stranded DNA library of 200 base pairs. This library was used for the *in vitro* selection of cGMP human genomic aptamers.

In addition, the partitioning condition during which the RNA sequences bind to cGMP should be given much consideration, *i.e.* selection buffer, concentration of ligand immobilized onto the selection matrix, and counter selection for increased stringency. A physiological-like buffer not only promotes folding of the purified transcripts, but also encourages biologically relevant binding constants from selected sequences. As the partitioning process of a selection often involves annealing of RNA sequences upon immobilized ligands, after which non-binders are removed and bound sequences collected, the purified RNA must be heated to denaturing temperature, then allowed to fold onto cGMP-agarose. Competitive elution with the ligand in solution requires a concentration of cGMP equal or greater than the amount of molecules covalently attached to the solid support. The cGMP-agarose used in the selection was supplied at approximately 3 mM in suspension. Consequently, the elution buffer was prepared with 5 mM dissolved in physiological-like solution in order to drive the equilibrium of the competitive elution event towards the binding of RNA sequences to non-immobilized cGMPs. Furthermore, the selection matrix plays a critical role in the viability of cGMP aptamers. The agarose beads can be derivatized with the cyclic nucleotide via a carbamoyl linker to the 2´-OH, or aminoethylthiol linker to the C8 position (see figure). These matrices have been widely used in affinity pull-down approaches to elucidate cGMP-responsive peptides[85,86] to great success. As RNA aptamers can adopt highly flexible structures, resins with different linkers offer different electrostatic surface interaction with the RNAs, thus the population of aptamers selected on a specific resin might not exhibit the same affinity and specificity on a different matrix of the same

ligand. The C8-cGMP agarose was chosen, however arbitrarily, for the selection matrix. Evidence also suggested that the *syn* conformation of the cyclic nucleotide is responsible for selectivity in kinases[87].

A counter–selection helps increase the specificity of cyclic-GMP aptamers. While many guanosine derivatives naturally exist within the cell, the aptamers should be pressured to recognize the nitrogenous base moiety, and become selective against the phosphorylation states. Since the defining feature, responsible for cellular specificity[68], of cGMP is the 3′, 5′-cyclic phosphodiester linkage, a counter selection against a molecule without this functionality aims to enrich for aptamers that directly interact with the cyclic phosphodiester bond. Guanosine monophosphate (GMP) was chosen as the counter selection target. RNA sequences, once annealed onto C8-cGMP agarose, was exposed to GMP dissolved in physiological-like buffer, then partitioned into fractions as previously conducted. Due to the structural similarities between GMP and cGMP, the counter selection buffer was prepared with 0.5 mM guanosine monophosphate. Higher concentrations of the counter selection ligand have been shown to subtract a significant portion of sequences bound away from the matrix.

The RNA sequences were prepared using a T7 RNA polymerase system, as the fixed sequences within the library design included a T7 promoter. Furthermore, the transcriptions were performed in the presence of [^{32P}] α-ATP, where the adenosine triphosphate contains an isotopic phosphorus atom on the alpha phosphate. This resulting radioactive transcripts can be traced via a liquid scintillation counter to quantify enrichment after each round of selection. It should be noted that the RNA sequences were purified after transcription via denaturing polyacrylamide electrophoresis (PAGE) to remove unreacted nucleotides and aborted

transcripts. This procedure necessitated an annealing step before cGMP column binding, as the RNAs must be folded upon the immobilized ligands.

The selection should be stopped when enrichment has plateaued. Then the enriched pool can be sequenced for aptamer identities, followed by determination of secondary structures and binding constants.

Implementation of next generation sequencing towards high-throughput identification and biochemical characterization of cyclic-GMP aptamer from a human genomic library

Selected aptamers' identity, affinity, specificity, and structures must be rigorously investigated. Conventionally, aptamers have been cloned into bacterial vectors, resulting in chimeric constructs containing a reporter gene for screening, after which successful inserts would be subjected to Sanger sequencing[88]. Individual aptamers isolated from cloning can be interrogated for their binding characteristic by performing column binding of their transcripts to C8-cGMP agarose. Aptamers are also characterized by chemical probing techniques for their binding constants and secondary structures. Among the most applicable approaches are in-line probing, and selective 2´-hydroxyl acylation analyzed by primer extension. This approach, however enjoys the high quality score of sequencing, is neither suited for enriched pools containing a diverse set of aptamers, nor applicable in selection with an overwhelmingly dominant sequence. In addition, biochemical characterization of RNAs have proven to be a laborious pursuit. Given the size and diversity of a selected pool, complete identification and characterization of aptamers therefore become tremendously time-consuming.

Current advancements in aptamer discovery have begun to adopt massive parallel sequencing into their workflows[89]. As the goal of an *in vitro* selection is the enrichment of

sequences of interest, next generation sequencing (NGS) has been applied to develop chip-based automated selection, which allowed rapid identification of active species, to monitor diversity of the DNA library after each round of selection, or to obtain total sequence identity of the final selected products[90–92]. In addition, NGS has allowed high-throughput structure probing of RNAs, and as a result, exponentially expanded the realm of RNA biochemistry[55,93,94]. These innovative approaches have created a foundation for rapid elucidation of RNA aptamers sequences and structures, especially as aptamers, and riboswitches by extension, can adopt multiple structural conformations in the presence of their effector molecules.

In an effort aimed at high-throughput biochemical characterization of RNA aptamers, my colleagues and I have developed Apta-Seq, a single pipeline to combine these protocols, and applied on several human genomic aptamer selections[53]. The rationale, setups, and pitfalls of the pipeline are elaborated in Chapter 4 of this document, while the output regarding the cyclic-GMP RNA aptamers selection from a human genomic library is discussed herein.

Apta-Seq revealed an overwhelmingly abundant, highly primate-conserved sequence within the cyclic-GMP RNA aptamers selection among other candidates

The selection for cGMP aptamers was carried over 9 rounds. At the 7th round, a counter-selection step was introduced, where guanosine monophosphate at 500 μM in binding buffer was added before the washing procedure. The enrichment of sequences were evaluated via scintillation counter, and were shown to plateau, thus no further round of selection was performed. Initially the pool was subjected to TOPO-TA cloning, however the clones obtained

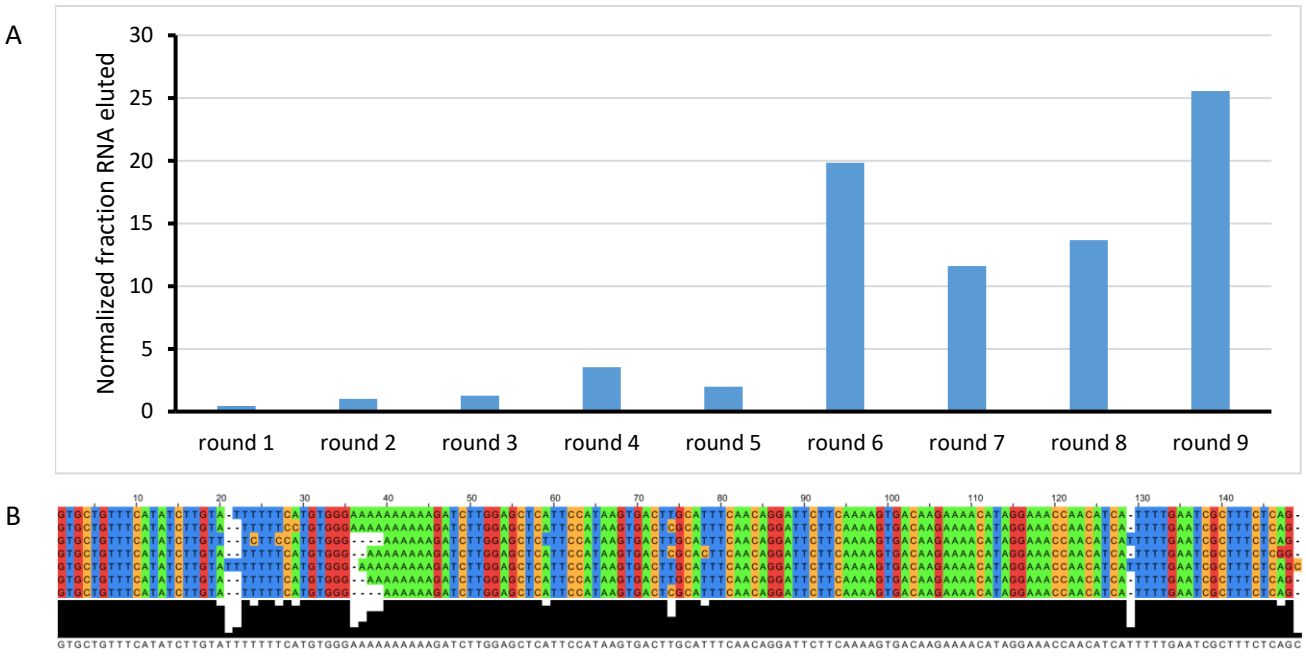


Figure 2-2: **Results of cGMP RNA aptamers selection and vector cloning** (A) Fraction RNA bound over the rounds of cGMP RNA aptamer selection. (B) Alignment of RNA aptamers obtained through vector cloning indicated extreme sequence similarity.

shared extreme sequence similarity, thus Apta-Seq was considered as an alternative approach for high-throughput analyses of sequences, abundance, and structural predictions of all selected aptamers.

Using Apta-Seq, seventy possible aptamers for 3´, 5´-cyclic guanosine monophosphate were identified from the human genome (Appendix Table 2-1). The most abundant aptamer to emerge from this analysis was G62. Several selected sequences were amplified from the enriched pool and tested for affinity. In addition, Apta-Seq also delivered SHAPE profiles of each individual sequence, provided enough sequencing depth for a particular genomic location. Apta-

Seq demonstrated the ability of identify low abundant aptamers, which might otherwise fail to express through conventional vector cloning.

In vitro activity of aptamer candidates

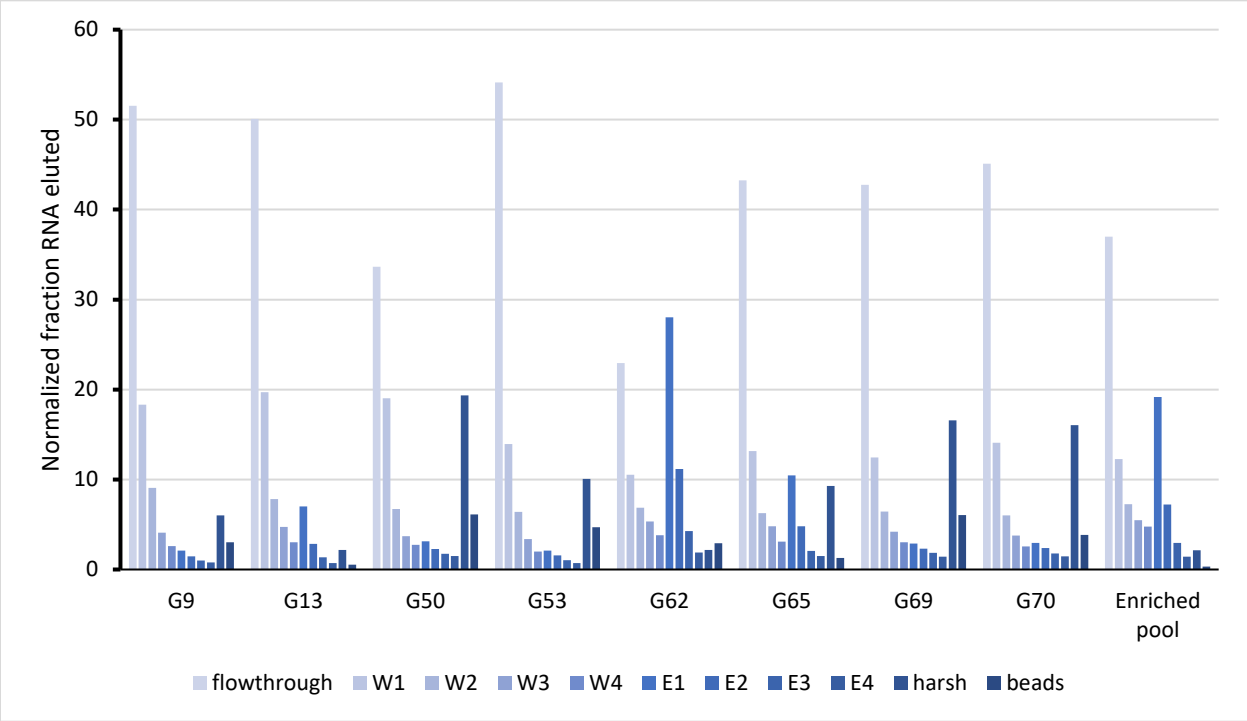


Figure 2-3: Elution profiles of selected aptamers towards cGMP

Among the few sequences amplified for affinity validation, three aptamers, G13, G62, and G65, were interrogated thoroughly for their specificity. Three different parameters of specificity were investigated through affinity chromatography. First and foremost, the sequences were tested in their ability to differentiate between 2´-OH-cyclic GMP and C8-cyclic GMP agarose. Secondly, the sequences should display selectivity against 3´, 5´-cyclic adenosine monophosphate. These cyclic purines have demonstrated cross-talk in cellular messaging[74,85], thus it is important that any potential aptamer should be able to distinguish the two molecules. Last but not least, as many guanosine phosphorylated derivatives exist within the cell, these aptamers should display different affinities towards these different molecules. G13, G62, and

G65 all preferred the C8-cGMP matrix to the 2´-OH variant (Figure 2-3). Furthermore, these sequences neither bound to 3´, 5´-cyclic adenosine monophosphate 2´-OH agarose, nor competitively eluted with cAMP after they had been annealed onto C8-cGMP column (Appendix Figure 2-4). Further column binding experiments revealed that these three aptamers are incredibly selective towards cyclic-GMP. In Figure A, the aptamers equilibrated on C8-cGMP agarose, followed by a titration of cAMP at 10^{-3} to 1 mM. As most of the bound sequences were removed from the matrix during the harsh elution, it could be concluded that the aptamers did not detach themselves from the immobilized ligands, *i.e.* competitive elution did not drive the equilibrium of the binding event towards the aptamer-cAMP complex. In contrast, Figure B denoted that the aptamers were annealed onto cAMP-agarose, and subsequently eluted with 5 mM cAMP in binding buffer. The RNA sequences in this procedure should not bind to the immobilized cAMP, thus were washed away in the earlier fractions. Consequently the harsh elution of these partitions were not comparable to those detailed in Figure A. However, it should be noted that the cAMP was covalently linked to the matrix through the 2´-OH position of the ribose. If the specificity of G13, 62, and 65 did not depend on the electrostatic interactions at the C2´ or the C8 functional groups, then the effectiveness of these selectivity investigations would become limited. An additional experiment utilizing C8-cAMP agarose matrix would

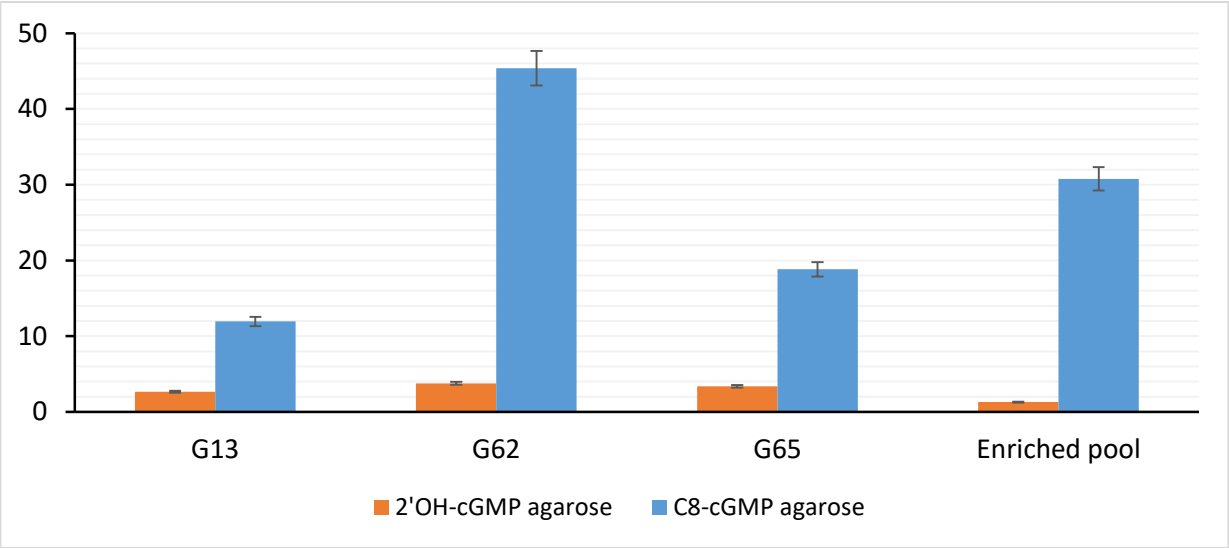


Figure 2-4: **Binding of G13, G62, and G62 on two different cGMP matrices.** The aptamer displayed remarkable regioselectivity towards the C8-linked cGMP agarose.

greatly complement the data previously shown. These rigorous column binding experiments have corroborated that aptamer G62, in addition to being the most abundance, displayed extraordinary selectivity and affinity towards cyclic-GMP.

Structure probing and expression analysis of G62 aptamer

Apta-Seq output on G62's frequency of reverse transcription stops revealed a highly conserved and ligand-sensitive domain. The aptamer is located on the negative strand of chromosome 20, and mapped onto an expressed sequence tag representing fragments of transcripts found within human optical nerve tissue (Figure 2-4). Curiously, the consensus sequence generated from Apta-Seq pipeline spans a 220-nucleotides domain, longer than the size-selected library. To identify the aptamer's genomic 3′-ends, raw sequencing reads containing 3′-fixed sequence were aligned. The alignment of these sequences implied that multiple aptamers were independently selected for cyclic-GMP affinity, and the coverage of the human genomic library maintained a high resolution.

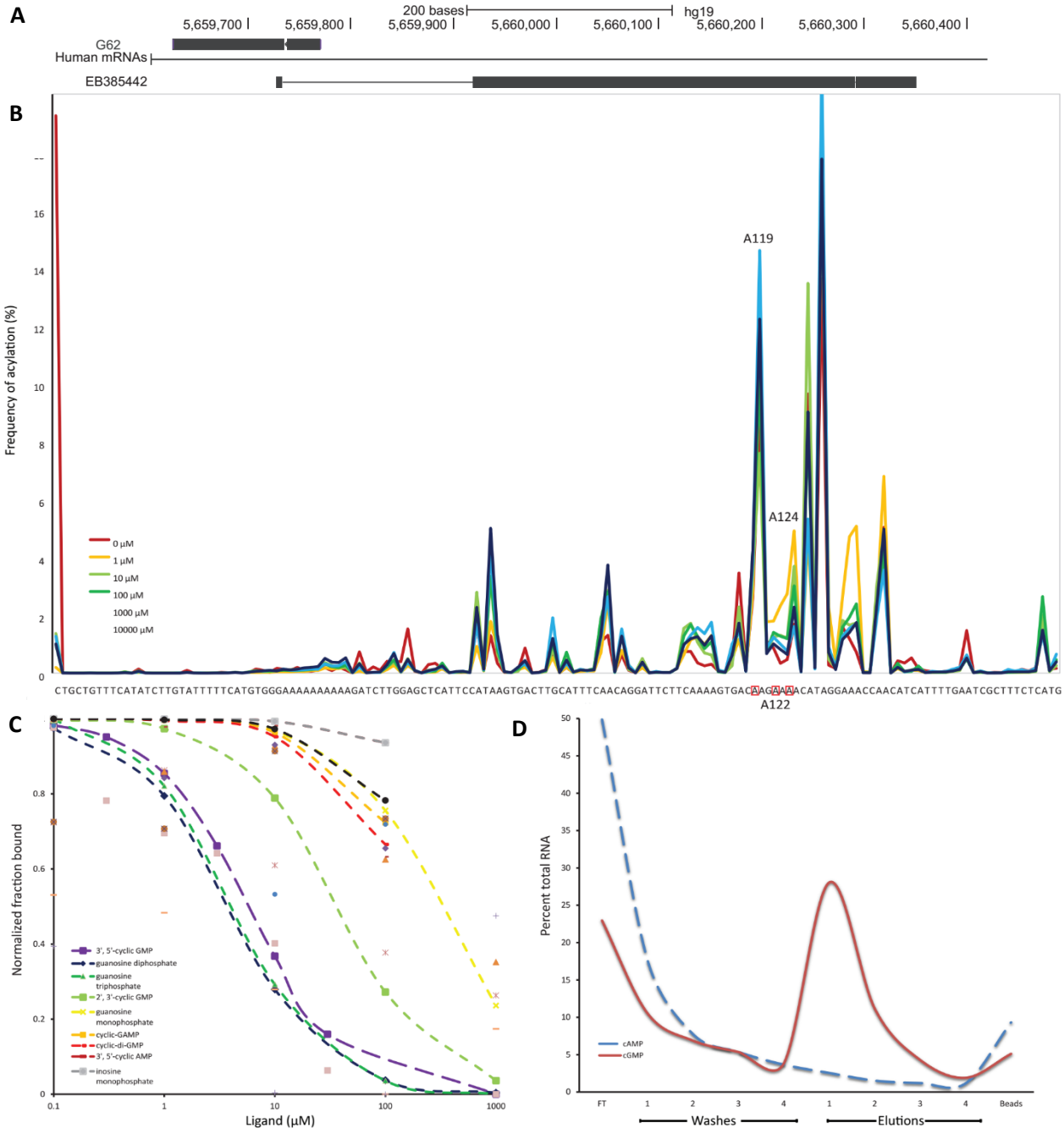


Figure 2-5: **Aptamer G62 in depth.** (A) Chromosomal location of aptamer G62. The aptamer is located near multiple human mRNAs and an expressed sequencing tag. (B) Apta-Seq output of G62's reverse transcription stops. Each peak represents the frequency of acylation on the nucleotide immediately upstream. Positions with cGMP dose response are annotated. (C) K_D graph of A122 of cGMP and its derivatives. G62 aptamer shares similar affinity towards cGMP, GTP, and GDP. (D) Column binding of G62 aptamer with cGMP and cAMP. The aptamer was annealed onto C8-cGMP agarose and eluted with the cyclic nucleotides.

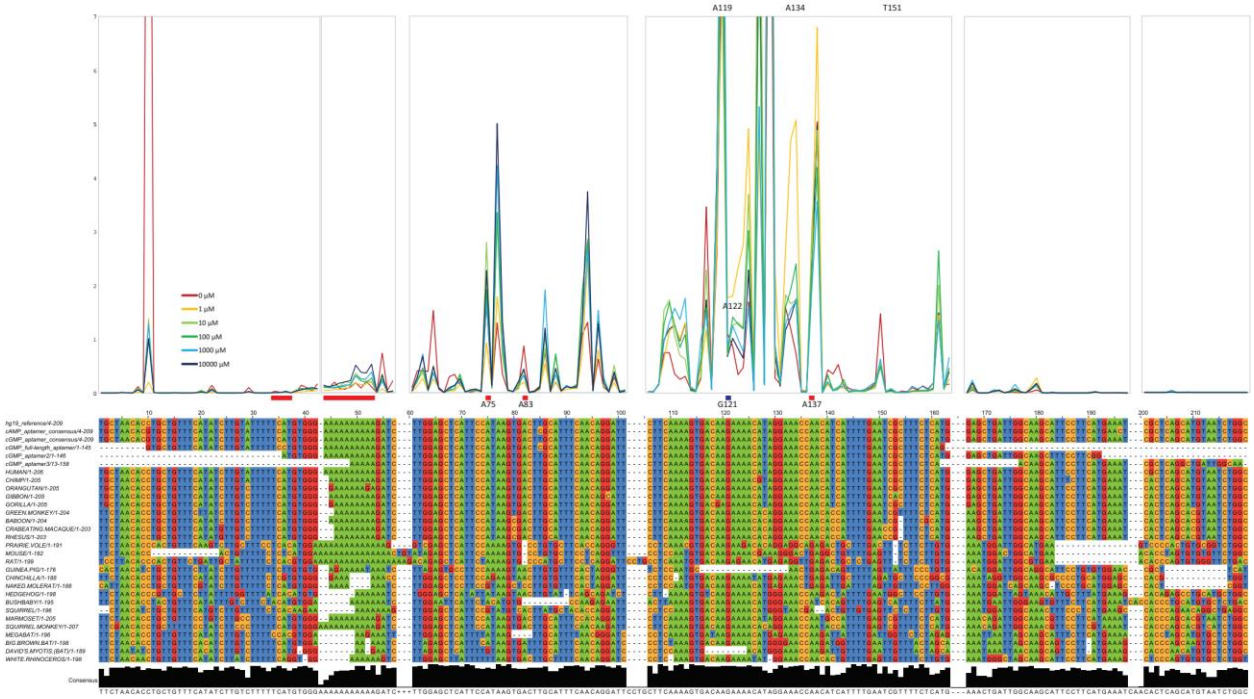


Figure 2-6: Alignment of G62's RT stops frequency to its human genomic sequence, selected aptamers with heterogeneous 3´-ends, and its occurrences in other mammalians

Furthermore, the ligand-responsive region of the aptamer is highly conserved among primates, and is of remarkable sequence similarity among other mammals (Figure 2-5). Analysis of RT stop counts revealed the nucleotides A119, A122, and A124 with highly variable SHAPE-reactivities in the presence of cGMP. Other bases in the vicinity appear to exhibit a certain degree of ligand-responsiveness, however no immediate dose-dependence can be concluded. This presents an inherent limitation of Apta-Seq: the inability to assign SHAPE reactivities to highly mutable or variable nucleotide positions, particularly instances pertaining to acylation of highly similar sequences. Thus, biochemical characterization of previously isolated G62 from vector cloning was conducted to confirm the aptamer's output from Apta-Seq. We predict that independent structure probing experiment should corroborate G62's traces of acylation. A SHAPE experiment was performed on G62's dose-response to the selection ligand. Furthermore,

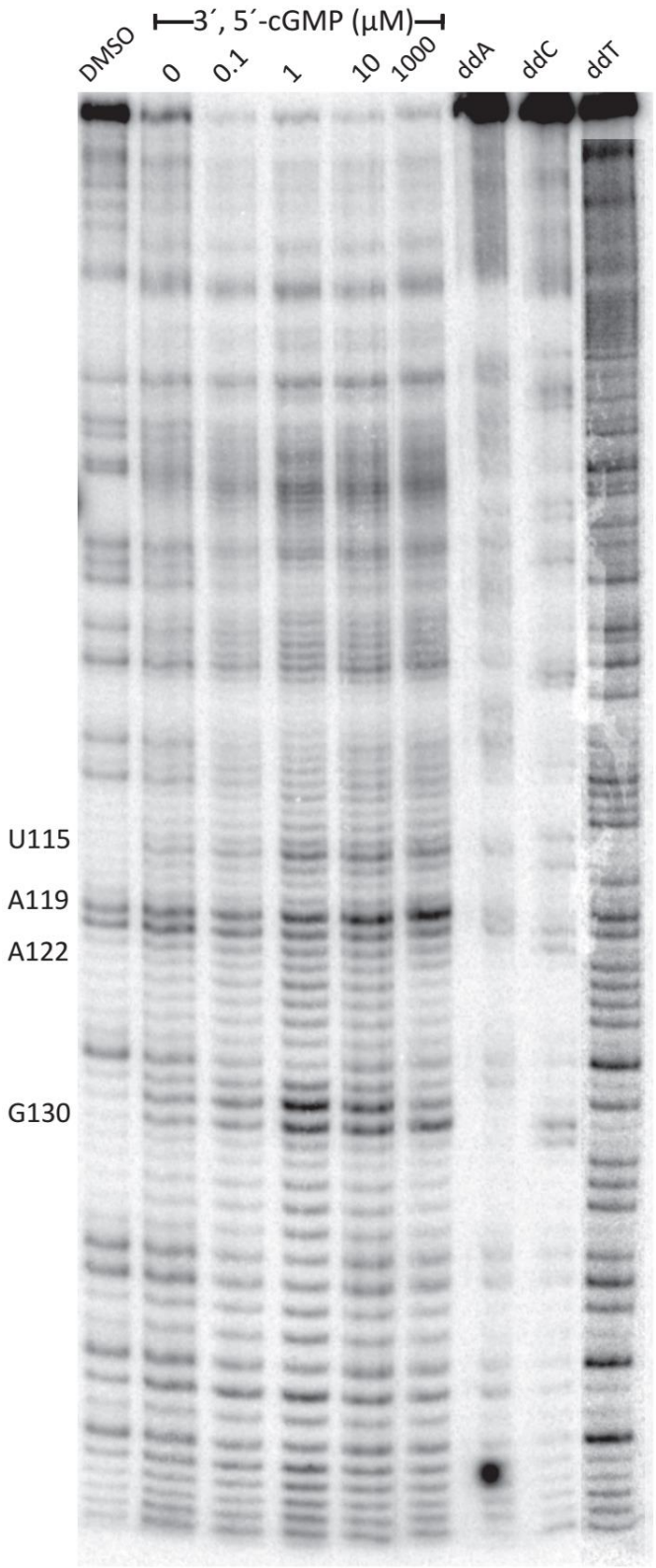


Figure 2-7: SHAPE gel of G62 aptamer. A119 shows increased SHAPE reactivities as concentration of cGMP elevated.

in-line probing of the aptamer towards 3´, 5´-cyclic guanosine monophosphate and its derivatives verified the robustness of Apta-Seq (Appendix Figure 2-1). Guanosine derivatives tested include molecules not only have similar chemical space to the selection ligand, but also are notable due to their biological significances. Cyclic-di-GMP and cyclic-GAMP have been widely documented as important signaling molecules in plants and bacteria[86,95,96], while inosine and 2-amino purine monophosphate have been typically utilized in chemical biology research as analogs to study *de novo* purine synthesis[97,98]. Guanosine di, and triphosphate were also examined for their binding towards the aptamer as these molecules are biological precursors of cyclic-GMP. The comparison between guanosine monophosphate to its cyclic phosphates and 2´-deoxy variants in-line

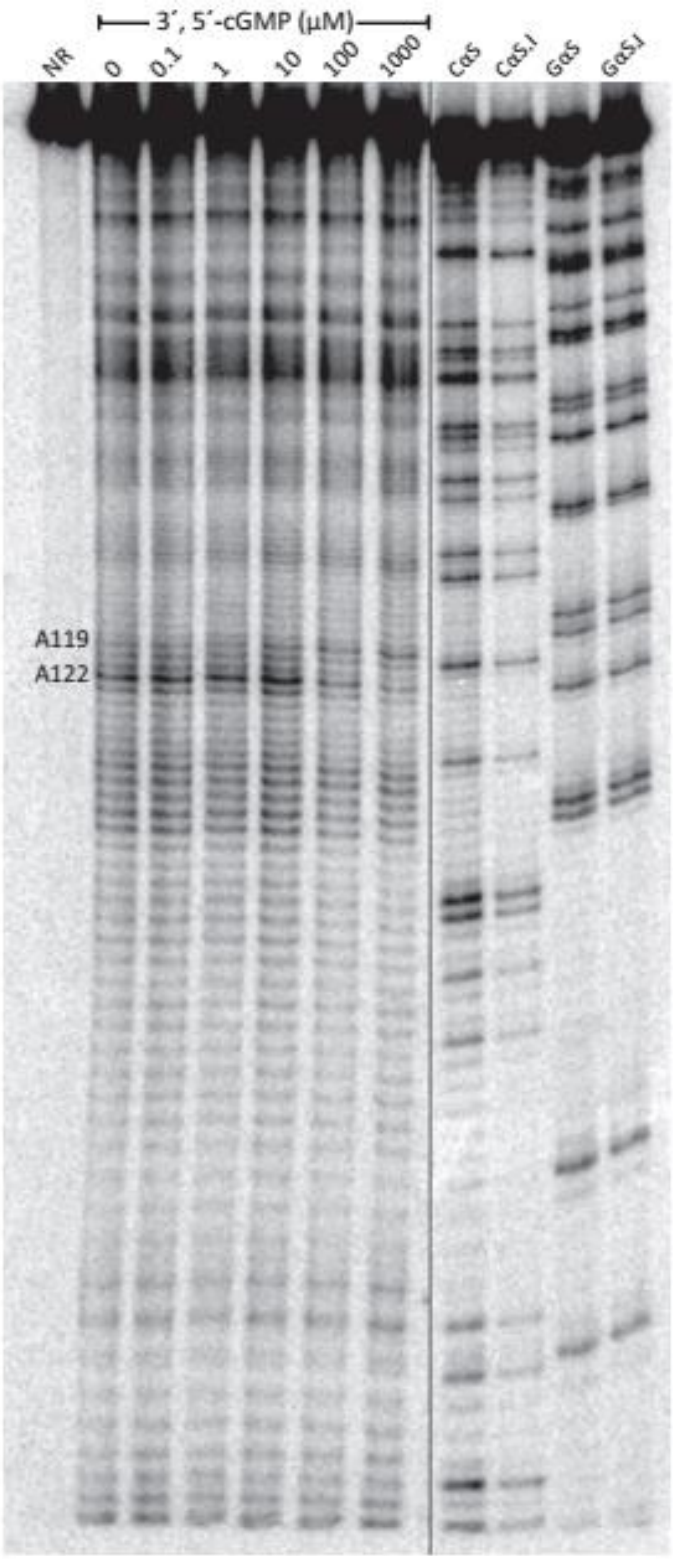


Figure 2-8: In-line probing gel of G62 aptamer with cGMP. A119 shows increased SHAPE reactivities as concentration of GMP elevated, while A122 becomes more structured.

probing provides biochemical insights for the aptamer specificity to its substrate. In addition, nucleotide triphosphate thiol analogs were used to examine interference mapping of the aptamer[99].

SHAPE analysis of G62 aptamer revealed additional dose-responsive nucleotides U115 and G130, in addition of the aforementioned A119 and A122, suggesting these positions became more susceptible to acylation as the concentration of cyclic-GMP increased. In-line probing of the aptamer also indicated the sensitivity of A119, A122, and G130 towards the target ligand. In Figure 2-8, A122 displayed a strong in-line signal at low concentration of cyclic-GMP that gradually diminished at higher concentrations, implying the position became more structured as

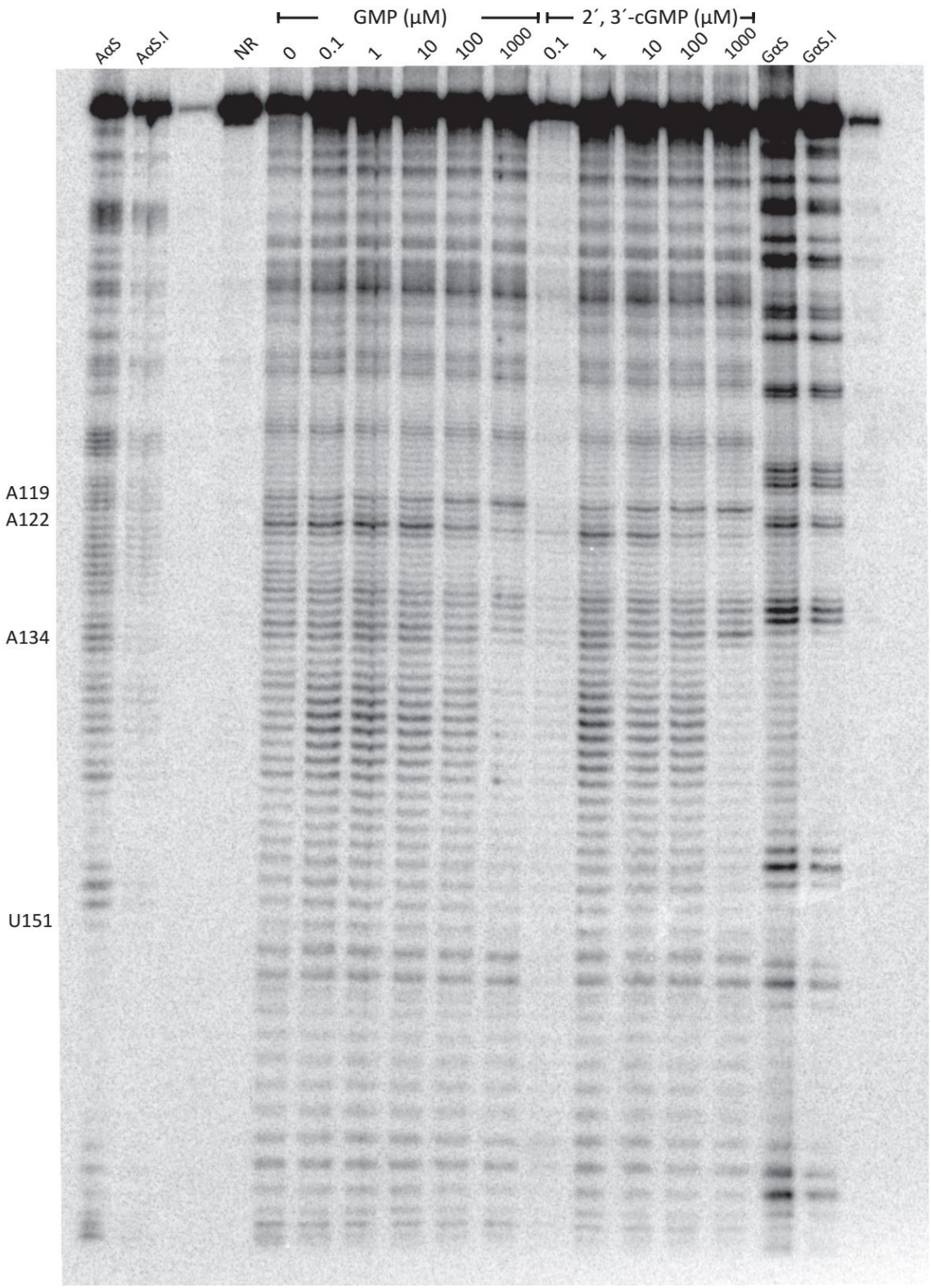


Figure 2-9: In-line probing gel of G62 aptamer with GMP and 2′, 3′-cGMP. In addition to A119 and A122's in-line activities, structure rearrangement was observed from A134 to U151.

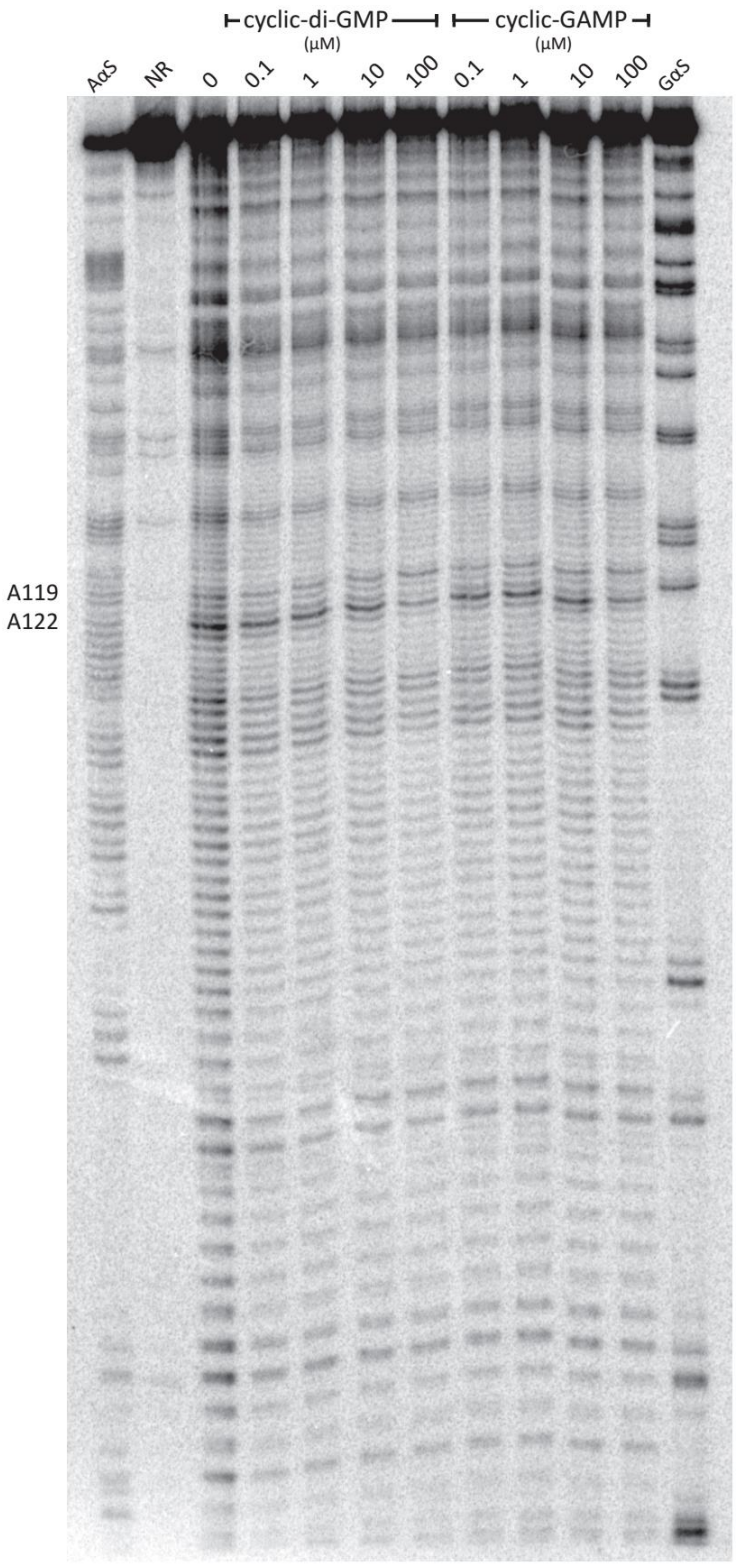


Figure 2-10: In-line probing gel of G62 aptamer with cdGMP and cGAMP. A122 shows weaker in-line activity with the cyclic-di-nucleotides compared to cGMP.

the target ligand bind to the aptamer. In contrast, A119 became less structured in response to the dose increase of cyclic-GMP, a phenomenon not observed in other guanosine derivatives (Figure 2-7, 9, 10, 11). The aptamer does not respond to varying concentrations of deoxy-guanosine, inosine, or cyclic 2-aminopurine monophosphate. We hypothesize that the affinity of the aptamer strongly depends on the guanosine moiety and 2´-hydroxyl of the ligand. In addition, in-line probing of the aptamer towards cyclic-di-GMP and cyclic-GAMP showed similar, albeit insignificant, sensitivity of A122 towards increasing dosage of these molecules. Figure 2-10 suggests that aptamer G62, however recognized the guanosine

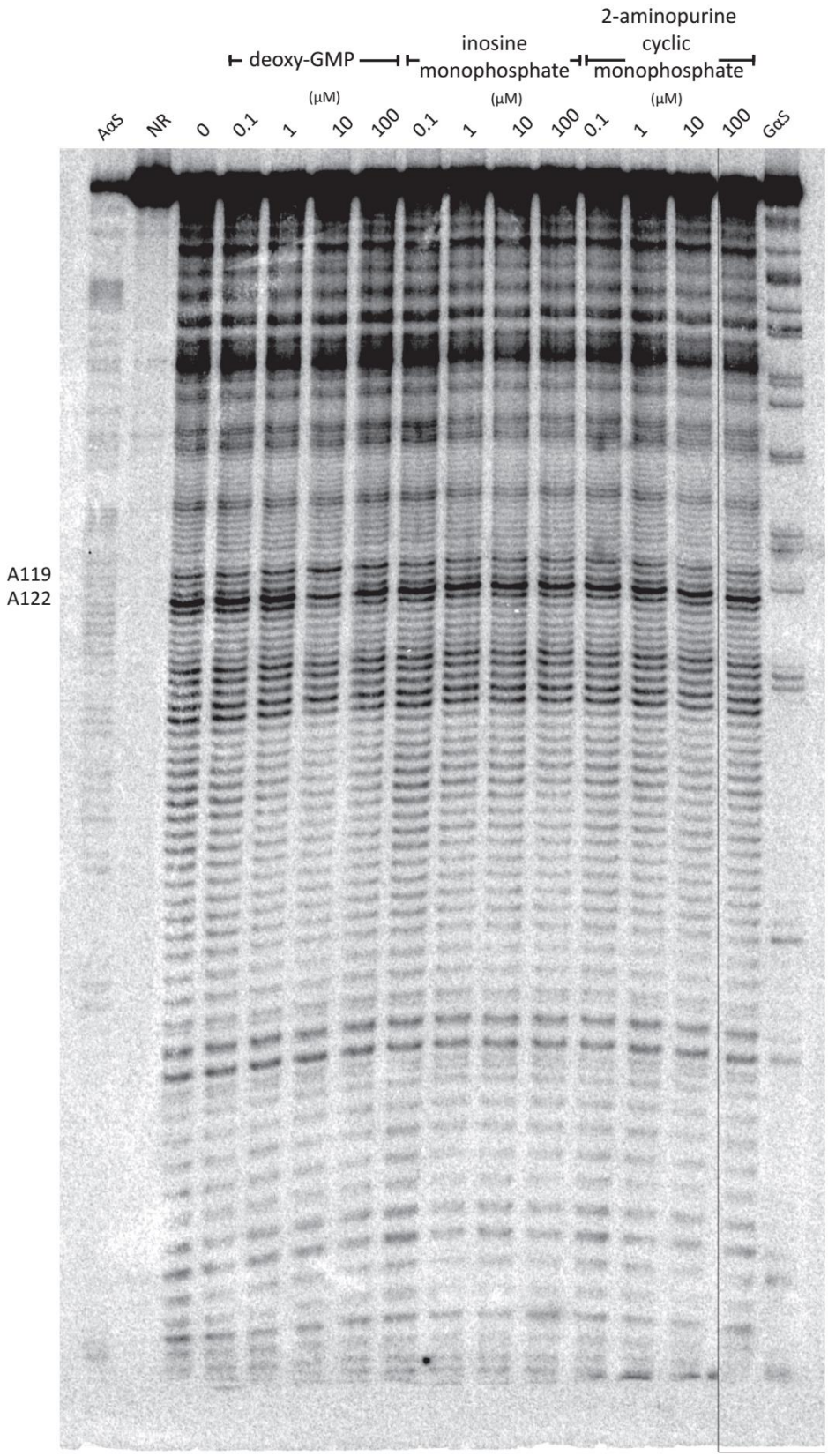


Figure 2-11: In-line probing gel of G62 aptamer with deoxy-GMP, IMP, and AMP. A122 shows no in-line activity towards these nucleotide monophosphate analogues.

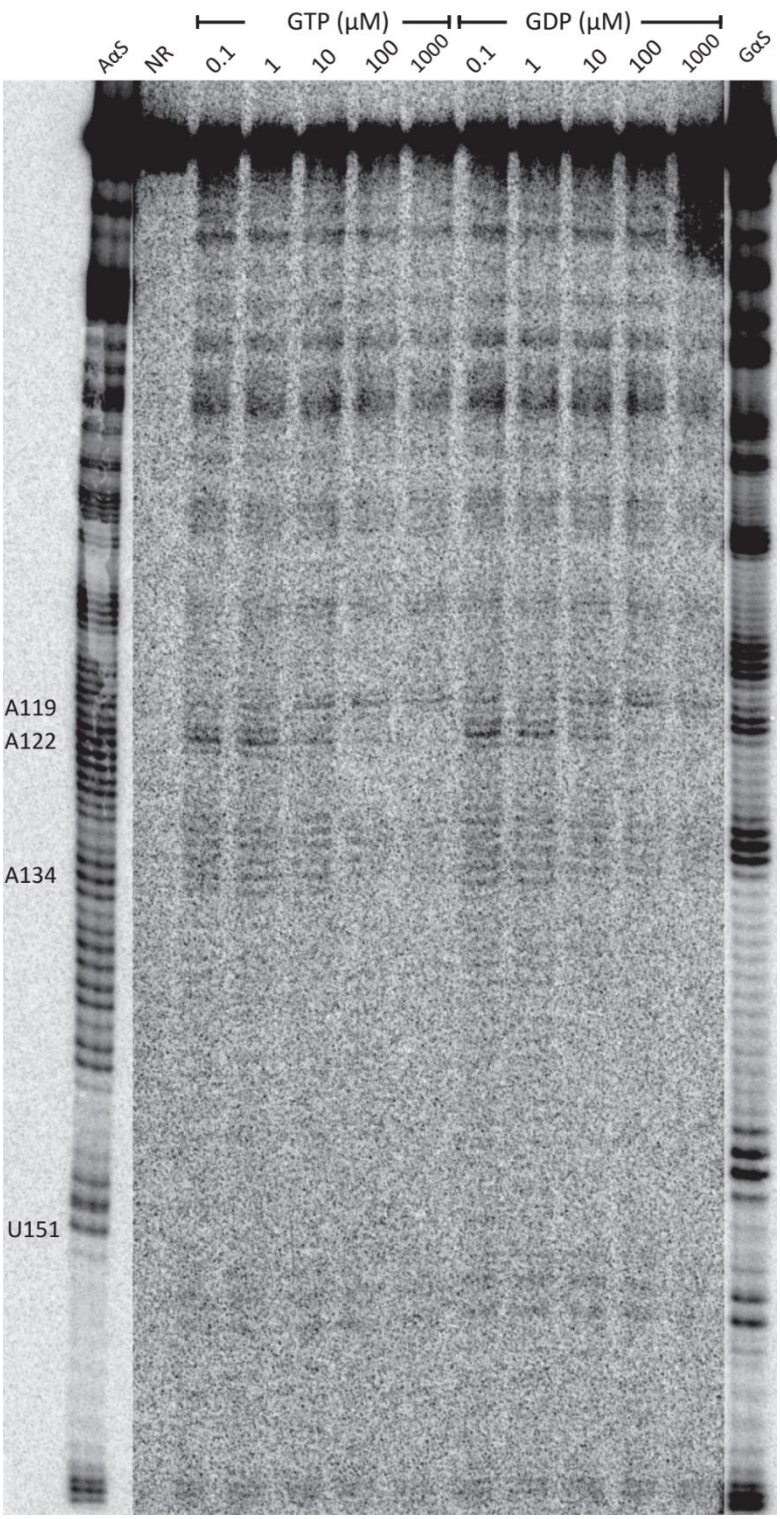


Figure 2-12: In-line probing gel of G62 aptamer with GTP and GDP. A122 shows strong in-line activity with poly-phosphorylated guanosines.

scaffolds of cdGMP and cGAMP, and not tolerate the dicyclic nature of these derivatives. Additional in-line probing experiments with guanosine monophosphate and 2′, 3′-cyclic guanosine monophosphate showed a region ranging from A134 to U151 of G62 becoming completely structured upon ligand binding. Similar structural shifts were observed as the aptamer binds to guanosine di and triphosphates. The constraints generated from these experiments were inputted into the Vienna RNA package for structure prediction, unfortunately the secondary structure of G62 remains elusive, as the program fails to generate an adequate model to explain the

conformational change involving A122 and A134-U151. In addition, these in-line experiments have provided binding constant G62 aptamer towards cyclic-GMP as 5.912 µM. It was suggested that the aptamer is promiscuous towards poly-phosphorylated states of guanine nucleoside. K_Ds of G62 aptamer measured through in-line probing for GTP and GDP are 4.05 and 3.87 µM, respectively. Comparing these values to average cellular concentration of cyclic-GMP, GTP, and GDP of 41, 17, and 500 µM in human cells[100,101], we conclude that the binding constants of G62 towards these phosphorylated guanosine are physiologically relevant, and further investigation is required to establish specificity of the aptamer.

We have also performed truncation of G62's sequence to generate a short genomic constructs lacking the synthetic primer-binding regions, with robust affinity and specificity towards cyclic-GMP. Selective amplifications of constructs using gene-specific primers were performed, followed by installation of T7 promoter onto the genomic constructs for *in vitro* transcriptions. We elucidated a construct of 89 nucleotides in length after extensively tested numerous different genomic sequences (Appendix Figure 2-2). Construct 4 displays remarkable affinity and specificity towards cyclic-GMP through competitive elution on chromatography. This construct was annealed on agarose matrix with immobilized cGMP, and eluted with an increasing dose of guanosine derivatives (Figure 2-12). The construct is more sensitive to cGMP than other tested guanosine nucleotides. Binding constant for the short genomic construct is yet to be measured. In-line probing of construct 4 demonstrated that the nucleotides correspond to A119 and A122 of the full-length aptamer experience the same activities towards cyclic-GMP (Figure 2-13).

As position A119 and A122 were identified as responsive to cyclic-GMP titration, two constructs containing mutations A119C and A122C were transcribed and tested for *in vitro* affinity. In addition, mutants A117C, A132C, and A133C were examined for binding as Apta-Seq have indicated these positions to have acylated activities in the presence of cyclic-GMP. Mutants

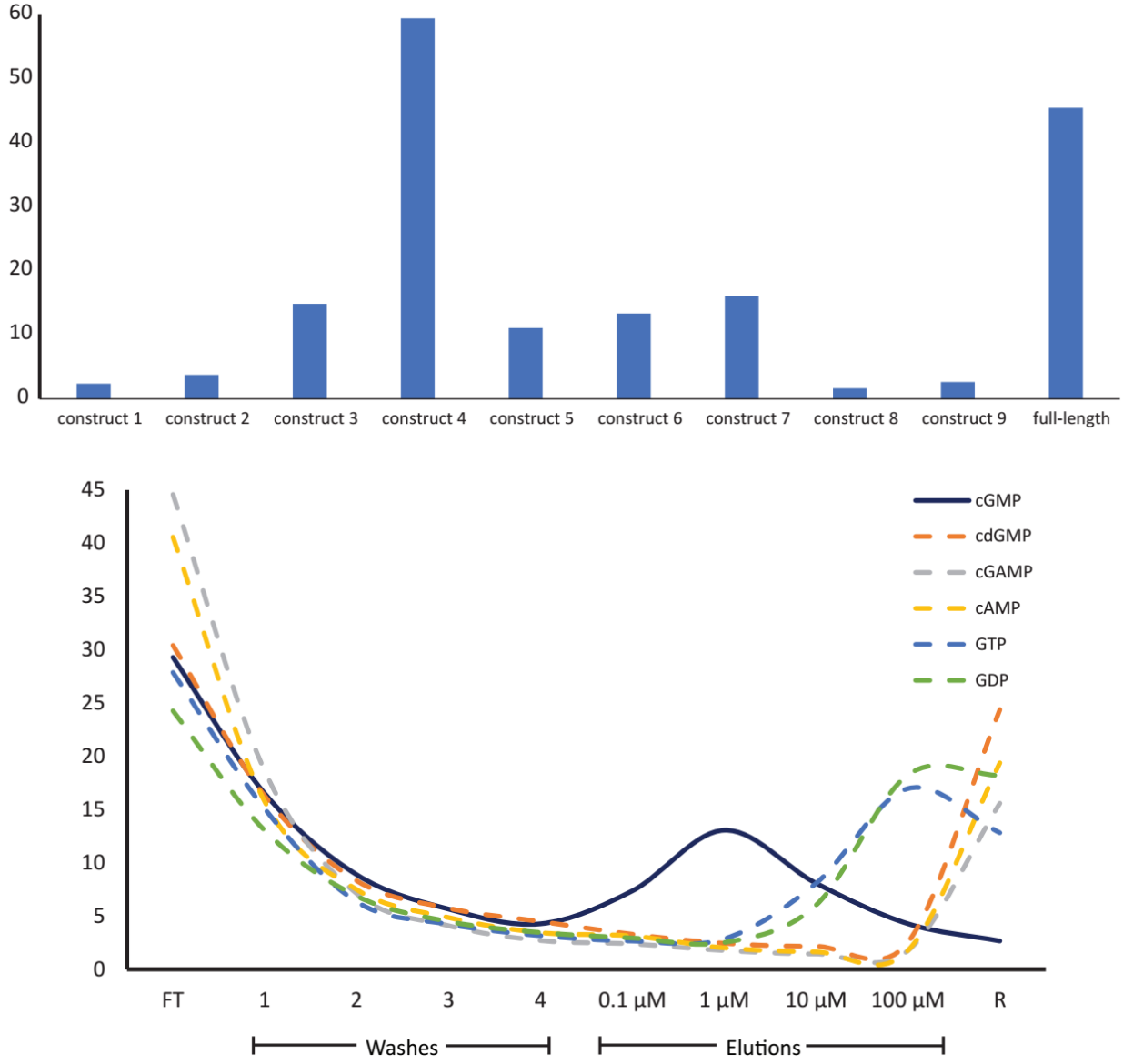


Figure 2-13: Column binding of truncated G62 aptamers. (*top)* Nine different genomic constructs of the G62 aptamer was tested for binding. Total RNA eluted for each construct was normalized and evaluated. Construct 4 was identified as the highest binder. (*bottom)* Selectivity of construct 4 towards guanosine-based molecules. The construct was annealed on C8-cGMP agarose and eluted with a gradient of ligand concentration. The construct dissociated from the matrix with lower concentration of cGMP compared to other guanosine derivatives.

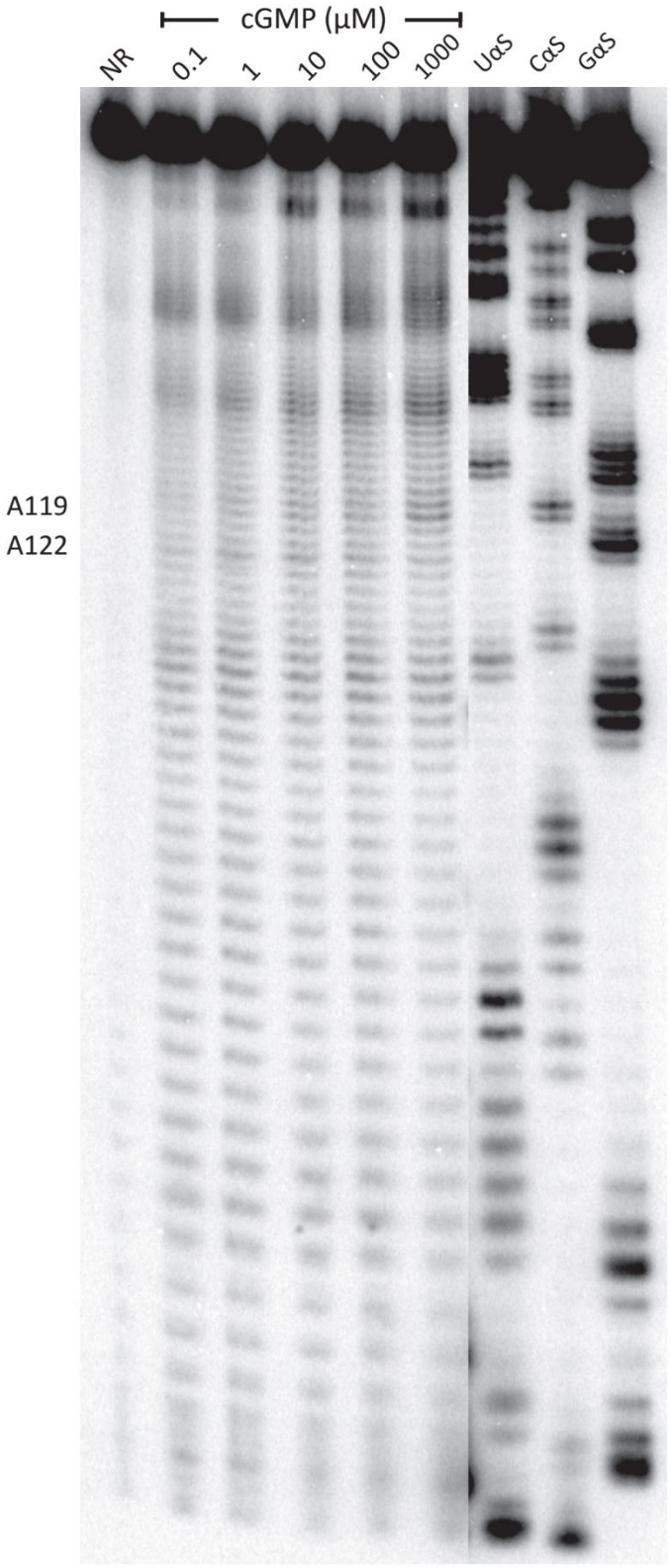


Figure 2-14: In-line probing gel of G62 genomic construct 4 with cGMP. A 119 and A122 show similar in-line activity with the ligand.

A117C, A119C, and A122C showed complete lack of binding to cGMP, while the mutations at position A132 and A133 did not have any influence on binding (Figure 2-14). The abrogated activity of mutants A117C, A119C, and A122C strongly suggested that these nucleotides are involved in binding of G62 to cyclic-GMP.

Other nucleotides have been identified as instrumental to G62's affinity towards cGMP through nucleotide analog interference mapping. Backbone substitution with phosphorothioate adenosine and guanosine analogues at A199, A132, and A149 have been observed to produce inactive variants (Appendix Figure 2-3). Thus, further investigation should include mutations at these positions, as well as positions in the A134 to U151 region.

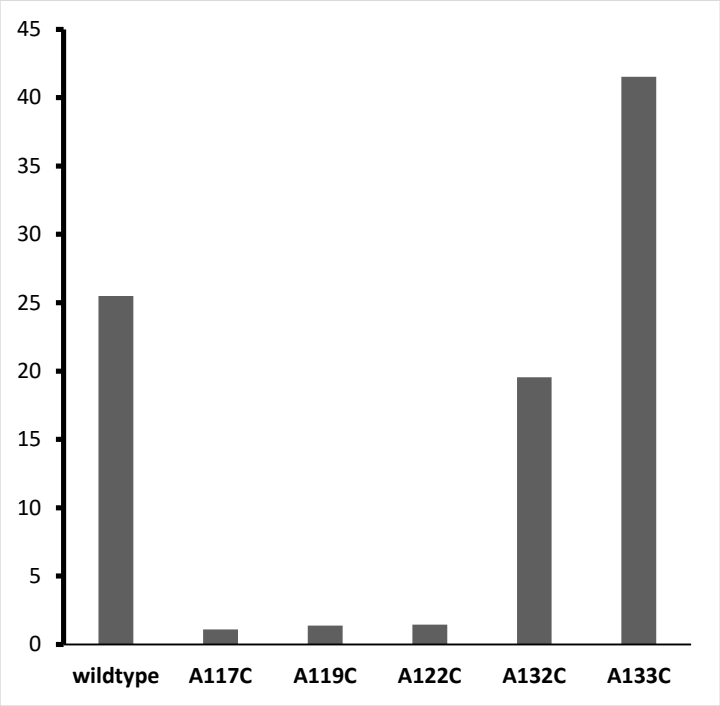


Figure 2-15: Normalized fraction RNA bound and eluted from cGMP agarose of wildtype and mutant aptamers.

Next, the G62 aptamer was subjected to random mutagenesis using a mixture of triphosphates of nucleoside analogues. Using dPTP and 8-oxodGTP in a *Taq*-based amplification[102], random mutations were introduced to a the G62 sequence. Four rounds of selections were performed on the mutagenized G62 sequences, followed by submission to high-throughput sequencing. During mutagenesis selections, the RNA sequences can mutate while preserving the secondary structure. As a mutation disrupts a base–pair, evolution favors a compensatory mutation that restores the RNA's secondary structure[103], and by extension, its affinity towards cyclic-GMP. We hypothesize that the aptamers correspond to G62 sequence exhibit a consensus structure that maintains sequence-independent double-stranded regions that can be revealed through random mutagenesis by determining the sites at which the degree of co-occurring is higher than random mutations and sequencing error. Unfortunately, the data obtained from G62's random mutagenesis did not yield immediate co-variation pattern. Nucleotides mutated without constraints were identified via bioinformatics variants calling (Figure 2-5) and corresponded to positions with low conservation in phylogenetic alignment. Nucleotides within extended regions of Ts and As have shown to contain stochastic point mutations, most likely

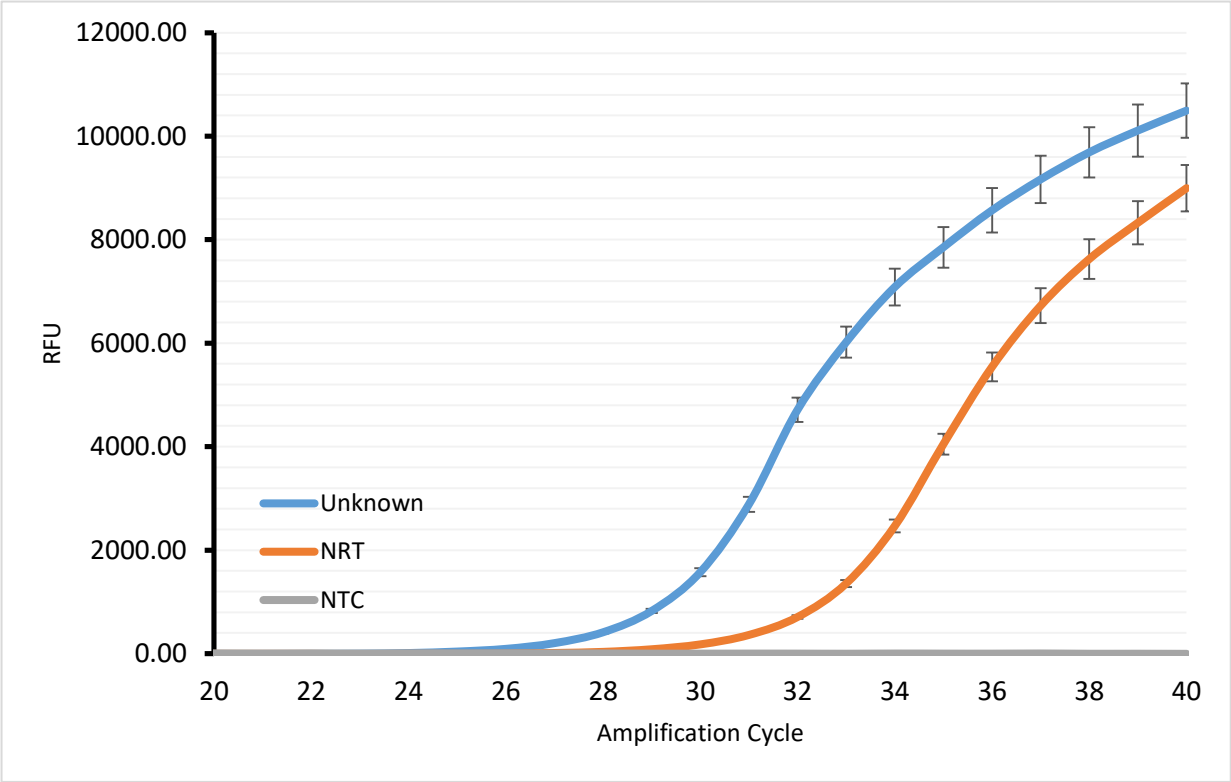


Figure 2-16: Quantification of G62 amplification from mouse embryonic brain cDNAs. No reverse transcription (NRT) and no template control (NTC) were included in the analysis.

due to a combination of sequencing errors and single-strandedness. Nucleotides A75 and A137 showed high rate of transition mutation, which is additionally reflected through phylogeny. A83, however indicated as conserved through comparative genomics, also displayed a high level of mutation. The previously identified nucleotides A117, A119, and A122 proved resilient against mutagenesis within evolved sequences. Last but not least, phylogenetic and mutagenesis data pointed towards an extraordinarily conserved position at G121. We predict that this guanosine is another key nucleotide responsible for the aptamer's affinity towards cyclic-GMP.

Next, we performed *in vivo* expression analysis of the aptamer in HeLa cells and mouse embryonic brain cells. Gene-specific primers for human and mice expression were used to reverse transcribe and amplify total RNA extract and monitor expression under quantitative PCR.

Our data proved that the aptamer is expressed in mouse embryonic brain (Figure 2-15). Further investigation should verify tissue-specific copy number of the aptamer, as well as exploration into the aptamer's *in vivo* structure elucidation.

Conclusion

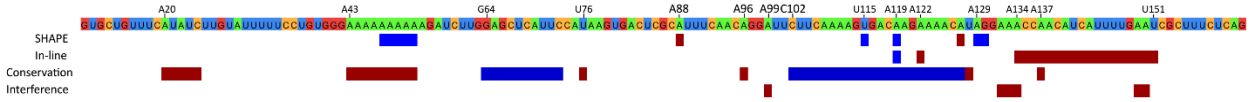


Figure 2-17: Reactivity profile of G62 aptamer. SHAPE, in-line, nucleotide analog, and conservation profile of G62 full-length aptamer. For SHAPE and in-line probing profile, blue bars indicate nucleotides at which reactivity increases proportionally to the cGMP dose, red bars indicate positions that experience less reactivity with increasing concentration of cGMP. For conservation, blue nucleotides are highly conserved, whereas red positions are partially conserved or highly mutagenic. The data from NAIM experiments indicated backbone substitution at A99, A132, and A149 are detrimental to G62's affinity towards cGMP.

RNA aptamers against 3′, 5′-cyclic guanosine monophosphate were enriched using *in vitro* selection from a human genomic library. Massive parallel biochemical characterization of these aptamers have yielded their sequence and structure identities. In-line probing experiments of the most abundant aptamer, G62, provided correlated structural probing data with Apta-Seq output, and have identified a highly conserved, ligand-responsive genomic region located on chromosome 20, near an expressed sequencing tag mapped onto a human optic nerve. The aptamer's K_D (5.912 µM) is within the physiologically concentration of cyclic-GMP, and its genomic construct demonstrates selectivity towards other guanosine derivatives. Residues responsible for cyclic-GMP binding of G62 were identified and confirmed via mutagenesis studies. The work presented in this chapter also included preliminary data for *in vivo* expression of G62 aptamer in mouse embryonic brain. These results should serve as a forerunner to the discovery of the first mammalian cyclic-GMP RNA aptamer.

Materials and Methods

In vitro transcription and purification of aptamers

All transcriptions for this work were performed using the T7 polymerase-promoter system. To promote isotope incorporation, the concentration of non-radioactive adenosine triphosphate was lowered ten-fold in comparison to the other ribonucleotides. Transcriptions were typically incubated at 37 ^{o}C for a minimum of 2 hours. The enzyme, along with its appropriate working buffer, were obtained from New England Biolabs. Radioactive nucleotide was purchased from Perkin Elmer (Table 2-1). Isotopic rATP was supplied at a specific activity of 250 µCi. Radioactive decay should be taken into consideration when using older stock of the reagent.

Isotope-labeled transcripts were purified using denaturing PAGE containing 7.5 percent acrylamide. The gel was scanned by a phosphoimager; the band containing the RNAs of interest was excised, eluted in aqueous potassium chloride (300 mM), and precipitated with glycerol in ethanol at - 20^{o}C.

Column binding of purified RNA aptamers

Purified RNAs were reconstituted in binding buffer, denatured at 70 ^{o}C for 5 minutes, and annealed for 25 minutes at room temperature on C8-cGMP agarose, which had been previously equilibrated to binding buffer (Table 2-2). Centrifugation at 3000 relative centrifugal force (rcf) was performed on the agarose column for 1 minute, followed by four short washes with binding buffer, and four 25-minute elutions with 5 mM cyclic-GMP. Any remaining RNAs were flushed from the column with harsh elution buffer (Table 2-2). Each fraction, including the C8-cGMP beads, was collected in individual tubes. These tubes were quantified for their

radioactive traces via a liquid scintillation counter to determine enrichment of aptamer across selection rounds.

Eluted fractions were combined, filtered through a packed G25 Sephadex column to remove excess cyclic-GMP and other small molecules, and then precipitated in glycerol and ethanol.

Reverse transcription and amplification of selected RNA sequences

Purified sequences were reverse transcribed and amplified for the next round of selection (Table 2-3, 2-4). M-MLV enzyme and buffer were procured from Thermo Fisher Scientific, while reagents for the amplification were obtained from New England Biolabs. The polymerase chain reaction was aliquoted over time to ensure the DNA sample did not contain unwanted amplification products. These aliquots were analyzed on a 2 percent agarose matrix embedded with ethidium bromide for ultra-violet visualization.

Primers used for reverse transcription and amplification of selected aptamers

Forward primer: 5′-GATCTGTAATACGACTCACTATAGGGCAGACGTGCCTCACTAC-3′

Reverse primer: 5′-CTGAGCTTGACGCATTG-3′

In-line probing and gel electrophoresis

3′-labelled RNA (0.1 µM) was suspended in binding buffer, pH ~8.0, in increasing concentration of ligand of interest. The reactions were incubated at 37 ^{o}C for 40 – 45 hours. A sample of end-labelled RNA was stored separately in binding buffer at -80^{o}C as a no-cleavage control.

Transcripts prepared with 1:10 phosphorothioate nucleotide triphosphate analogues to rNTPs were end-labelled and annealed onto cGMP agarose. Flowthrough and eluted fractions

were collected, cleaved with iodoethanol, and loaded onto a 10% polyacrylamide along with in-line probing aliquots. Gel was run at 50 W for any duration of interest.

Mutagenesis PCR

dPTP and 8-oxo-guanosine triphosphate was added to amplification reaction (Table 2-4) at a 1:25 analogues:dNTPs molar equivalence. The amplification was performed over 30 cycles and analyzed over agarose electrophoresis. A subsequent PCR was performed without the addition of mutagenic nucleotides, and the amplified templates were transcribed for *in vitro* evolution.

Table 2-1: Reagents and concentrations for transcription

Reagent	Final concentration
RNA polymerase buffer	1x
T7 polymerase	25 units
DNA template	0.5 µM
rATP	0.1 mM
rCTP	1 mM
rGTP	1 mM
rUTP	1 mM
[^{32}P] α-rATP*	0.5 mCi/mL
DTT	5 mM
DMSO	10%

Table 2-2: Selection buffers for cGMP genomic SELEX

Binding	140 mM KCl, 10 mM NaCl, 20 mM Tris, 5 mM $MgCl_2$, pH 7.5
Elution	5 mM cGMP, 140 mM KCl, 10 mM NaCl, 20 mM Tris, 5 mM $MgCl_2$, pH 7.5
Counter elution	0.5 mM GMP, 140 mM KCl, 10 mM NaCl, 20 mM Tris, 5 mM $MgCl_2$, pH 7.5
Harsh elution	8 M urea, 5 mM ethylenediaminetetraacetic acid

Table 2-3: Reagents and concentrations for reverse transcription	
Reagent	**Final Concentration**
M-MLV buffer	1x
M-MLV reverse transcriptase	100 units
dNTPs	1 mM each
Reverse primer	0.5 µM

Table 2-4: Reagents and concentrations for PCR	
Reagent	**Final concentration**
Standard Taq buffer	1x
dNTPs	200 µM
forward primer	500 nM
reverse primer	500 nM

Chapter 3

Metagenomic *in vitro* selections of RNA aptamers against cyclic-nucleotide monophosphates

Introduction

As the findings of the *in vitro* selection of 3′, 5′-cyclic guanosine monophosphate aptamers suggested, functional RNAs are often conserved across multiple organisms. This observation is reinforced by other occurrences of conservation of riboswitches and ribozymes[34,77,104,105]. Furthermore, the adenosine and the guanosine triphosphate-binding motifs have been reported as widespread within primates, along with other mammals and bacteria[20,28]. The identification of naturally occurring GTP aptamers by Curtis *et. al.* began with a DNA library derived from the genomes of eubacteria, archaebacteria, and eukaryotes, while the adenosine-binding loop has been reported in viral, bacterial, and mammalian genomes. In fact, the adenosine aptamer motif has been regarded as extremely robust, for its affinity permeates virtually all SELEX investigations concerning the nucleotide-derived cofactors[106]. These instances of phylogeny among nucleotide-specific RNA aptamers compel an *in vitro* selection on a metagenomic scale, spanning a diverse set of organisms ranging from bacteria, plants, to higher eukaryotes. A metagenomic SELEX provides access to functional annotation of aptamers across comprised genomes in the starting library, and consequently engenders novel hypotheses of convergent evolution among the included organisms. Furthermore, metadata generated from next generation sequencing of an enriched metagenomic pool has the potential to validate previously completed selections, thus delivering experimental corroboration of

phylogenetic analyses accomplished via bioinformatics approach on genomic databases. Current definition of 'metagenomics' is confined to direct genetic examination of microbial diversity within ecosystems collected from various human tissue or ecological points of interest[107–109]. As such, a metagenomic DNA library in the context of *in vitro* selection within this document should not be confused with the contemporary classification. Rather, the metagenomic SELEX was performed on a diverse and heterogeneous DNA library. Inspired by the completion of the human genomic SELEX of 3′, 5′-cyclic guanosine monophosphate, metagenomic selections were performed against cGMP and cAMP.

3′, 5′-cyclic adenosine monophosphate, along with the previously discussed guanosine variant, is a second messenger molecule involved in signal transduction in all domains of life. More specifically, cAMP has been shown to stimulate phosphorylase, protein kinase A, nucleotide-gated channels, and transcription factors[110]. In prokaryotes, cAMP directly plays a role in regulation of transcription by binding to a cAMP-responsive protein. This cAMP-protein complex binds to specific DNA sequences to enhance or inhibit transcription of the associated genes[111]. The dinucleotide variant of cAMP is broadly conserved and influences growth as well as infection by binding to cyclic-di-AMP protein receptors, and to key metabolic checkpoint enzymes such as pyruvate carboxylase and NADH-dehydrogenase[112]. In mammalian systems, intracellular cAMP-depedent protein kinase A phosphorylates many proteins, which in turn modulate transcription via activation and repression of the genes[113]. However, most of these regulation elements are peptide, and their mechanisms involve a binding event of the second messenger to a cAMP-dependent protein. The second messenger is intracellularly generated from adenosine triphosphate through adenylyl cyclase. Binding of cAMP to protein kinase A

dissociates the catalytic subunit, which subsequently phosphorylates nuclear cAMP-response-element binding protein (CREB). CREB recruits a co-activator that enhances transcription via its intrinsic interaction with RNA polymerase II[110]. Thus, a metagenomics *in vitro* selection for RNA aptamers against cAMP promises to illuminate direct involvement of the effector molecule. In addition, the presence of cross-talk between cGMP and cAMP tantalizes possible promiscuity, or lack thereof, of RNA aptamers' specificity between the two cyclic nucleotides[114]. Whereas cross-activation of protein kinases has been previously reported, protein kinase A and G prefer different phosphorylation site, suggesting specificity might not result in immediate effect in cellular communication. However non-parallel the comparison between peptide and nucleic acid substrate, functional characterization of RNA aptamers towards these cyclic nucleotide must consider downstream expression for complete annotation. The progress towards selections of cyclic-AMP and cyclic-GMP RNA aptamers from a metagenomics library is discussed in this chapter.

Construction of a metagenomic library and its initiation in *in vitro* selection of RNA aptamers

The construction of a metagenomic library was performed by Julio Polanco from the Lupták laboratory during his doctorate study. The library comprises of genomes collected from twenty-one different organisms and boasts a diversity of 10^{10} sequences. Chromosomal DNAs from these organisms were sheared under sonication, followed by adapter ligation for transcription and amplification purposes, and size-selected for ligated products spanning 200 to 700 nucleotides. The combined library resulted in a heterogeneous and diverse pool fit for metagenomic *in vitro* selection (Appendix Table 3-1).

Ergo, a selection aimed to take advantage of the heterogeneity and diversity offered by a metagenomic library must adapt its strategy to highlight these benefits, as well as monitor the reduction in these factors over the its iteration. As the DNA templates vary in sizes, the transcripts of the metagenomic pool, however should be purified over electroporation, must not contain any procedure that might suggest size selection. As such, the transcribed product can be excised from the polyacrylamide matrix as a vertical gel piece in order to remove only short, aborted constructs. In addition, over-amplification presents a concern for pool diversity. Byproducts, self-primed amplicons, and primers require attention during electrophoretic analysis of selection amplification. As more rounds of selection were performed, the reduction in heterogeneity and diversity visualized on ethidium bromide-stained agarose matrices must be interpreted with caution as to differentiate selected DNA sequences from unwanted PCR products.

Progress towards the *in vitro* selection of RNA aptamers against cyclic-AMP and cyclic-GMP in a metagenomic library

The metagenomic selection for cyclic-nucleotide aptamers were performed over fourteen rounds. The selections were halted when the amount of RNAs eluted plateaued after several rounds. As previously described in Chapter 2, the cAMP aptamer selection was pre-washed with guanosine monophosphate to limit the dominance of G62 binding motif. In addition, selections were performed on agarose matrices derivatized with cyclic-nucleotides covalently attached at the 2´-OH position as to ensure parallel comparison between the enriched pools. Furthermore, transcribed products at later rounds of selections were reconstituted in physiological-like buffer containing reverse primer for the library to eliminate

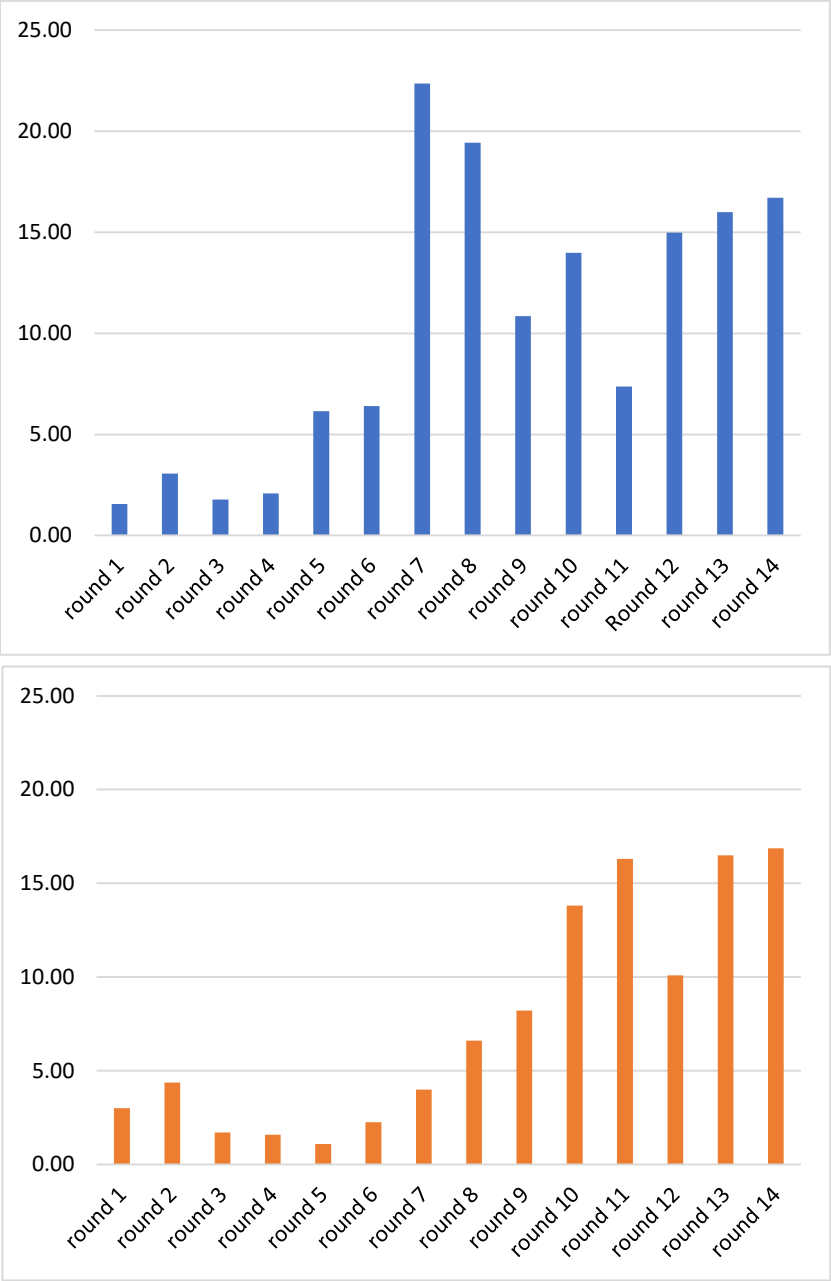


Figure 3-1: Normalized RNA bound and eluted in metagenomic *in vitro* selection of cGMP *(top)* and cAMP *(bottom)* aptamers. GMP and pool reverse primer was added to the selection to increase stringency.

aptamers forming binding sites for target ligand utilizing the constant regions. The reverse primer was introduced during the annealing step at which the aptamers were denatured at high temperature and subsequently transferred onto cyclic-nucleotide matrix. Transcripts with constant region-involved affinity are presumed to have their secondary structures saturated with the reverse primer, thus becoming excluded from the selection.

The heterogeneity of the starting metagenomic library witnessed a drastic decrease as the selection progressed. Whereas the initial library boasted DNA sequences ranging from 100 to 700 nucleotides, the enriched pools of both selections were consisted of a nucleic acids ranging from 100 to 300 bas–-pairs, with the

majority of sequences at approximately 150 nucleotides. We hypothesized that longer transcripts were more prone to aggregation, alternative secondary structures, and *trans* or *cis*-acting degradation.

The selected sequences were prepared for Apta-Seq for characterization. However the attempt was thwarted by low amplification yield as the sequences were annealed and cluster-amplified on the flow cell. The enriched pools were subsequently submitted for sequencing without acylation, and the reads were assembled onto consensus sequences. The assembly process proved to be computationally intensive, and thus was set to be conducted via the BIO Cluster of the University of California, Irvine. As the sequences originated from multitudes of genomes, the reads were trimmed, filtered, and aligned against well-annotated genomes such as human and mouse. When aligned to the reference genome hg19, bowtie2 determined over 50 percent of the reads were successfully mapped. The genomic locations and annotations of a few notable aptamers are included in Appendix Table 3-2 and 3-3.

The FGD3 and G62 aptamers made an appearance in the alignments of the selected sequences onto the human reference genome. The FGD3 aptamer was previously identified from the *in vitro* selection of RNA aptamers for adenosine triphosphate[20], while the G62 aptamer has been characterized and discussed in great detail in Chapter 2. The emergence of these sequences from the metagenomic library suggest the tyranny of their affinity towards the target ligand, especially the FGD3 aptamer[106]. However we must be skeptical of possible barcode misread that occurred during sequencing readout. Nonetheless, the high abundance of G62 aptamer in the metagenomic SELEX for cGMP aptamer affirms the robust affinity of the aptamer towards guanosine derivatives. In addition, aptamer MA25/MG33 has been a subject

of interest of our genomic SELEX efforts recently. Its presence in both selection dataset further encourages caution while interpreting high-throughput sequencing data. The possibility of cross-barcode contamination highlights the lack of quality control during the sequencing process. Such limitation compels a strategy to quantify amplification efficiency to determine the frequency of incorrect priming, which plays a pivotal role in the informational hygiene of resulting dataset.

The sequences emerged from the metagenomic cAMP aptamer selection aligned to the human genome were found in adenosine-related genes. MA17 was aligned onto an intronic region that expresses two different isoforms of protein kinase 2. Curiously, MA16 was mapped only a cAMP-specific phosphodiesterase (PubMed: PDE7B). This PDE also recognizes cyclic-GMP, yet the aptamer corresponding to its intronic region did not emerge from the metagenomic SELEX of cGMP aptamer. These results require complete biochemical and phylogenetic annotation of the listed aptamers, and needed to expand to alignment to other genomes included in the starting library.

Conclusion

In summary, two parallel *in vitro* selections for RNA aptamers against cyclic nucleotides, cGMP and cAMP, were performed. These selections began with a metagenomic library composed of chromosomal DNAs from multitudes of organisms. The selection utilized agarose matrix with the same covalent attachment, counter selection, and reverse primer pre-elution to ensure affinity, specificity, and structure integrity of selected sequences. The aptamers emerged from these selections, as identified via high-throughput sequencing, were aligned onto the well-annotated human reference genome hg19, and revealed many possible cyclic-nucleotide

responsive transcripts. Further investigation should include complete assembly of sequencing reads, massive parallel structural characterization of selected aptamers, and *in vivo* studies to determine the biological and conservational relevance of these aptamers. Last but not least, the annotation of aligned sequences should proceed with skepticism, as cross-barcode contamination was suggested during the processing of sequencing data.

Materials and Methods

Transcription, purification, and primer extensions were performed following the previously described protocols included in Chapter 2.

Oligonucleotides used for reverse transcription, amplification, and library preparation:

Forward: 5-´TAGATCTTAATACGACTCACTATAGGGAGACACTCTTTCCCTAC

ACGACGCTCTTCCGATCT-3´

Reverse: 5´-CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-3´

Reverse primer for high-throughput sequencing: 5´-GTGACTGGAGTTCAG

ACGTGTGCTCTTCCGATCT-3´

Reverse primer for Apta-Seq: 5´-AGATCGGAAGAGCGTCGTG TAGGGAAAGAGTGTGCG

GCCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCG-3´

Chapter 4

Development of Apta-Seq: A high-throughput pipeline for RNA aptamers biochemical characterization

Introduction

The advent of massive parallel sequencing has pushed the boundaries of genomic, transcriptomic, and proteomic investigations. Earlier sequencing-based analyses were coupled with a functional screen, typically an immunoprecipitation assay on a flow cell, to study interaction between proteins and nucleic acids[115–117]. These technologies were reliant on crosslinking events between peptide polymers and the targets of interest, followed by antibodies-based isolation of bound complexes. Other sequencing-based studies have deviated from immunoprecipitation for target acquisition, due to biases inherent in affinity and specificity within antibodies[118,119]. High-throughput sequencing-coupled investigations of nucleic acids interactions with other macromolecules performed on microarray have taken advantage of canonical base-pairing to anneal DNA sequences to the chip, after which a functional screen on annealed sequences was conducted[62,82]. However, the reliance on Watson-Crick base-paring between the nucleic acids and the flow cell became a maMor caveat of these procedures, as an immobilized RNA might sample a diverse range of molecular dynamics, leading to limited conformations. In addition, these methods measure the binding kinetics or only the rates of association when the fluorescently-labeled ligand dissociates from the complex. Other technologies relied on extraction of DNA or RNA sequences under the chemical probing conditions appropriate for the functional screen of interest, followed by library preparation for

deep sequencing[55,94,120]. These experiments are consisted of generation of chemically modified RNAs, followed by a massively parallel reverse transcription readout, and bioinformatics analyses of deep sequencing datasets. Inspired by the strategies, a direct implementation of SHAPE-seq onto *in vitro* selected aptamers was developed to couple selection with structural and binding characterization of emerged RNA sequences. The pipeline was initially applied to validate and further explore the results of a previously accomplished *in vitro* selection of ATP aptamers in a human genomic library, and was subsequently utilized broadly towards many ongoing selections, including the 3′. 5′-cyclic guanosine monophosphate RNA aptamer selection from a human genomic, and a metagenomics library. Using the Illumina sequencing platform, Apta-Seq can provide millions of reads, which in turn significantly aids the discovery, sort, and characterization of aptamers. As the structure probing step of Apta-Seq is performed in solution and prior to cDNA synthesis, the strategy is not impeded by the same limitations of HiTS-RAP/FLIP as discussed above, and deliver structure outputs and binding constants under equilibrium condition.

The assistance of the Spitale laboratory with the synthesis of SHAPE reagent, as well as insights on computational analysis of next generation sequencing data, were instrumental to the development of Apta-Seq, and are to be discussed herein.

Library preparation for high-throughput sequencing of acylated RNA aptamers

As the selected RNA aptamers became acylated, the sequences needed to be Illumina-compliant for high-through put sequencing. In addition, the acylated RNAs should be enriched over the non-acylated sequences, thus a pull-down mechanism was deemed necessary to separate the modified aptamers. A SHAPE reagent with azido capture handle has been

developed for high-throughput transcriptomic investigation, and was considered as the acylation reagent for Apta-Seq[121]. Furthermore, dibenzocyclooctyne coupled agarose has been widely used in the laboratory for copper-free click chemistry has made the separation of acylated aptamers rapid, robust, and convenient.

Reverse transcription of captured RNAs were performed on DBCO-agarose. Due to the stability of hybridization, RNA hydrolysis with NaOH was performed to isolate cDNA sequences.

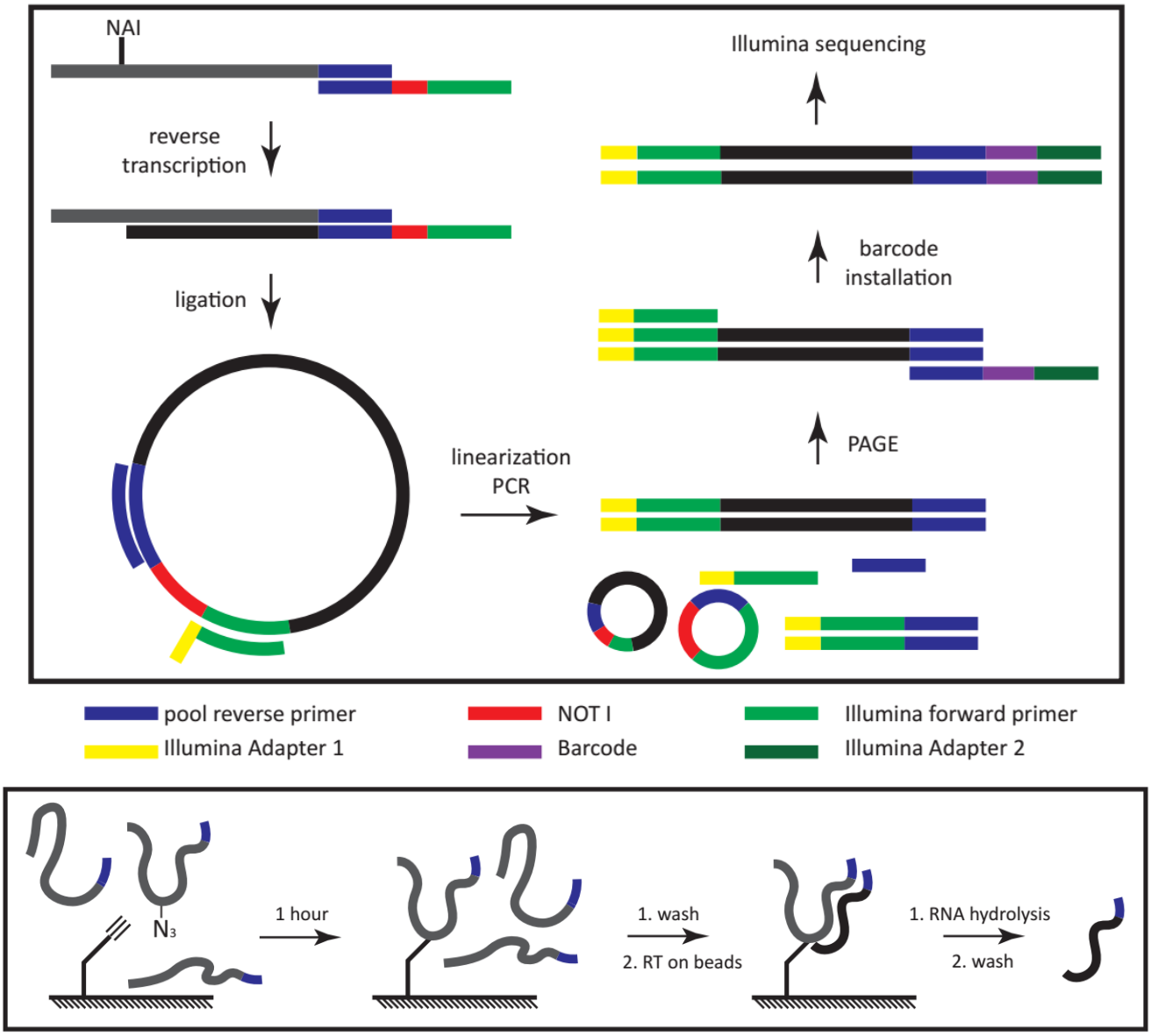


Figure 4-1: Apta-Seq library preparation scheme.

The reverse primer used in the extension contain a portion of the fixed Illumina sequence, as well as the reverse complement of the forward Illumina adapter. As the ligated cDNA were linearized, the amplified products should contain an Illumina cassette appropriate for sequencing (Figure 4-1).

Synthesis of 2-(azidomethyl) nicotinic acid acyl imidazole

The SHAPE reagent of choice was 2-(azidomethyl) nicotinic acid acyl imidazole[49]. The compound, NAI-N_3, boasted a lower rate of hydrolysis compared to the initially reported SHAPE reagent, and contained an azido functionality for click-based chemistries[122]. The synthesis of NAI-N_3 proceeded following a previously established protocol, and spectroscopic data of each intermediate was collected to ascertain the success of each synthesis step. A gel shift assay was performed at the end of the synthesis to evaluate the activity of the synthesized compound.

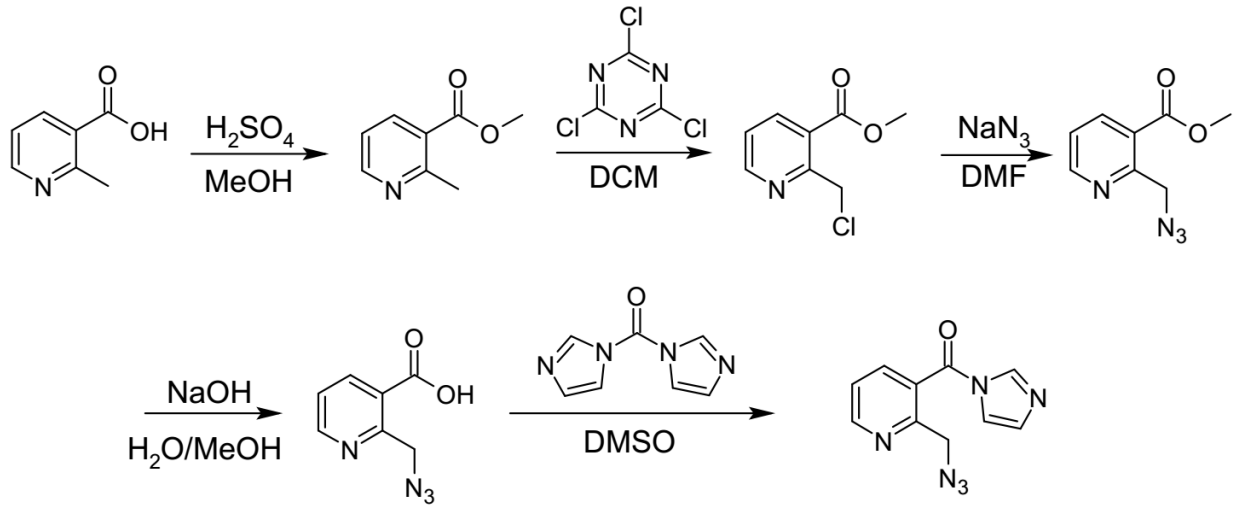


Figure 4-2: **Synthesis scheme of 2-(azidomethyl) nicotinic acid acyl imidazole.** As previously reported in Spitale *et. al.*, synthesis started with precursor methyl 2-methyl nicotinate[49]. However, the above first step of synthesis began with acid-catalyzed methyl esterification of methyl 2-methyl nicotinic acid.

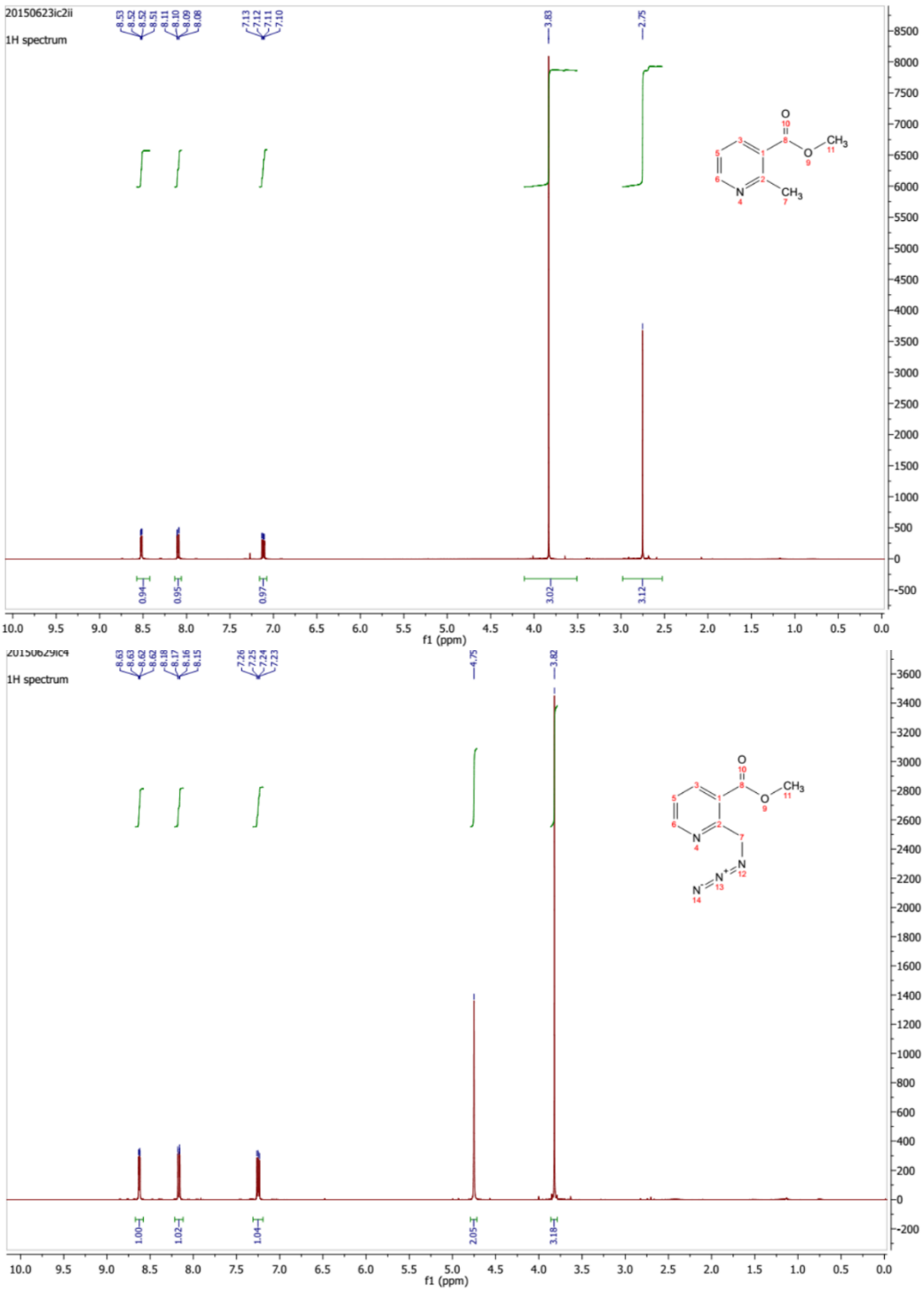


Figure 4-3: ^{1}H NMR detection of (*top*) methyl 2-methyl nicotinate and (*bottom*) methyl 2-(azidomethyl) nicotinate

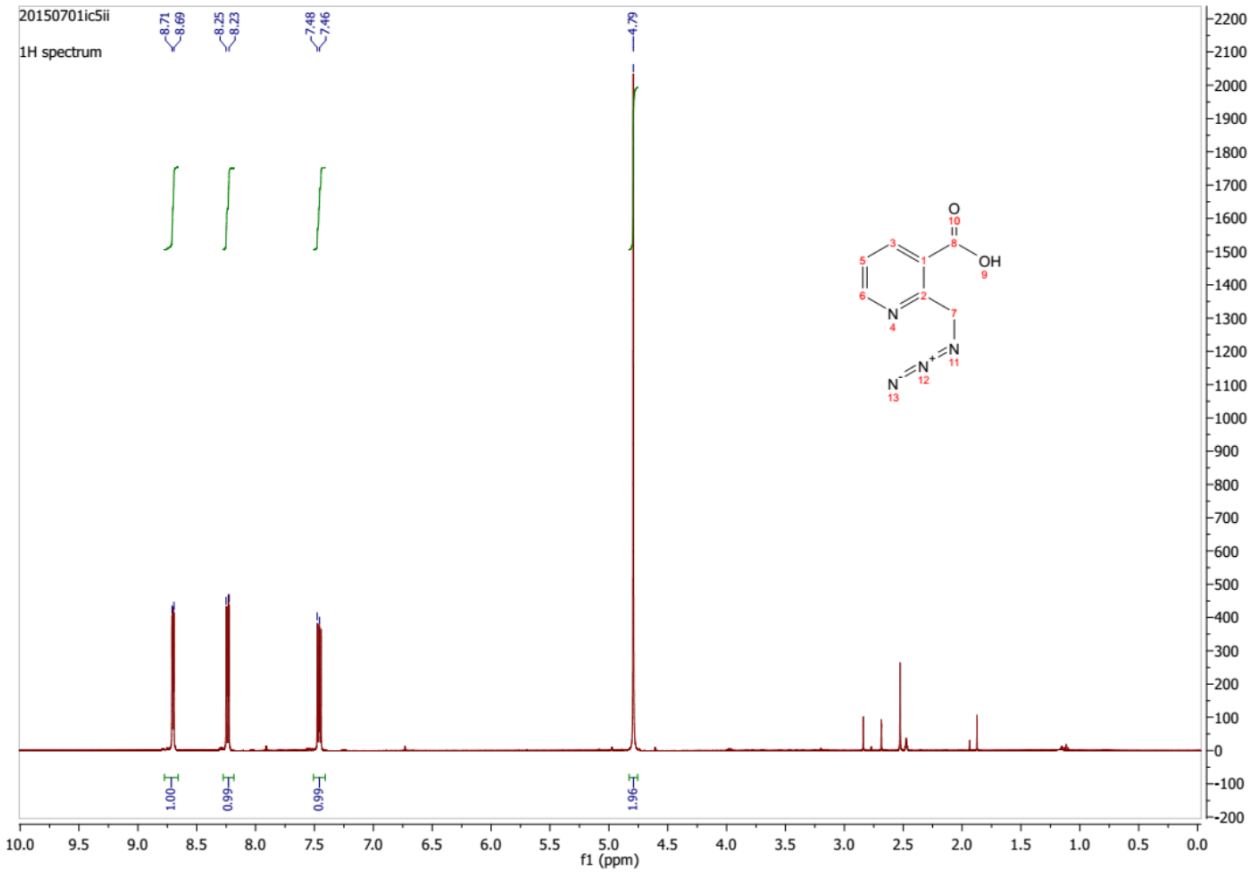


Figure 4-4: **^{1}H NMR detection of methyl 2-(azidomethyl) nicotinic acid.**

Designing a bioinformatics pipeline for high-throughput RNA aptamers characterization

Massive parallel sequencing of RNAs coupled with SHAPE has been reported by several investigators[49,54,55]. The stability towards hydrolysis of NAI-N_3, as well as the azido handle, appealed to the design of Apta-Seq. The azido functional group served as a chemical pulldown for acylated RNAs, leading to their capture on DBCO-agarose. RNA aptamers were acylated across a titration of their target ligands, and incubated on DBCO beads for copper-free click capture. Their subsequent reverse transcription presumably yielded and enriched for cDNAs with 3´-end representing one nucleotide downstream of the modification site, as non-acylated RNAs were removed from the aptamer population. However, as the enriched pool was not

equipped with Illumina-compliant fixed sequences, these adapter needed to be installed upon the cDNAs via a protocol detailed by Loughrey *et. al*[55]*.*

Next generation sequencing creates large batches of data. The Genomic Facility at the University of California, Irvine provides fastq output already demultiplexed by barcodes. Hence the analyses to identify and characterize RNA aptamers involved adapter trimming, quality score and length filtering, genomic alignment or assembly whether the *in vitro* selection begun with genomic or synthetic DNA library, and generation of consensus sequences, abundances, and annotations.

Initially, the module FASTaptamer was the primary resource for reads processing and aptamer discovery[123]. The application is capable of generating statistics such as ranks and frequencies of reads, as well as clusters and enrichment of aptamers among different libraries. However its shortcomings manifest as the discovery process requires structure prediction, consensus sequence, and abundance of aptamers written in more contemporary file extensions.

Raw reads stored in .fastq extensions must be inspected and filtered against adapter sequences, low confidence bases, amplification artifacts, duplication, and length[124]. More specifically, as the end of a sequencing read typically contains low quality bases, a trimming step is often required. The presence of such bases, in addition to adapter sequences, compels a text removal procedure from the raw .fastq files. Common applications to achieve this task includes cutadapt, Trimmomatic, and ConDeTri[59,125,126]. The trimming algorithm underlies these programs is either running sum or sliding window. The running sum approaches the summation of the difference between all base quality values is scored against a quality threshold, and the sequences are trimmed at the location that provided the lowest sum. Sliding window, a

trimming algorithm with a frame of defined size and quality threshold, slides against the raw reads and trims away the window that does not meet the defined standards[127]. Cutadapt was chosen as the trimming program of choice due to its accessibility, installation, and update, as well as the ability to customize the adapter sequences for trimming, and other capabilities for read processing. Trimmomatic, however powerful, needed to be invoked via Mava, and offers little in terms of customizability, especially in scenarios where reads containing adapters should be kept and imported to a separate output. In addition, cutadapt allowed for length-dependent error within the specified adapter sequences, as longer input of adapters are more likely to be subMected to sequencing errors.

The enormous amount of short, trimmed reads generated from Illumina sequencing and subsequent cutadapt calls for fast and accurate alignment against reference genomes. Depending on the experimental input, sequences are typically mapped using Bowtie(2), SOAP, BWA, or TopHat2[58,128–131]. As the input of Apta-Seq are RNA aptamers, generated from libraries of double-stranded DNA, as opposed to RNA sequences extracted from *in vivo* experiments, alignment programs such as TopHat and TopHat2 might not necessarily yield meaningful results, for these algorithms focus on the identification of splice Munctions and indels. Bowtie, BWA, and SOAP are considered more suited for reads less than 50 nucleotides. As several dataset were specified with longer reads, Bowtie2 became the program of choice for alignment. The program also supports alignment in local mode, which is more lenient towards mismatches and sequencing errors. In addition, bowtie2 allows alignment to overlap ambiguous character in the genomic reference, which is useful for consensus sequences of aptamer whose identity conveys

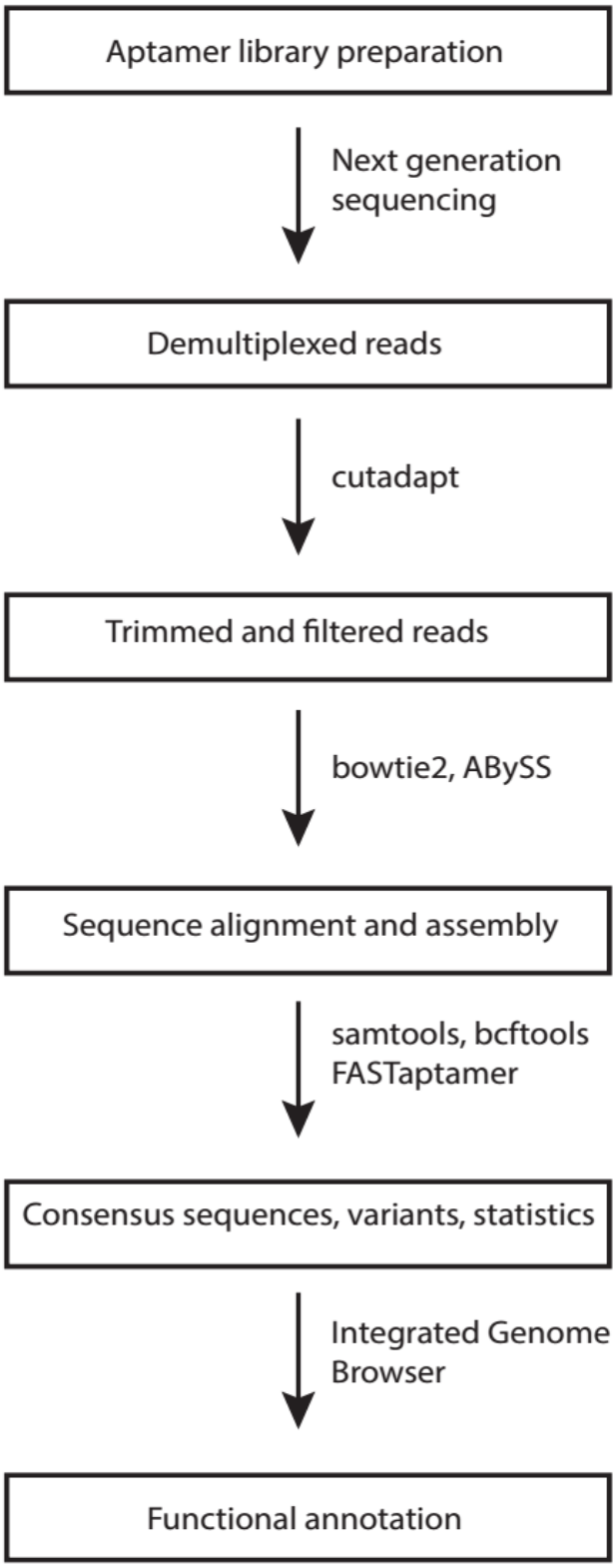


Figure 4-5: Bioinformatics pipeline of Apta-Seq. Raw sequencing reads are demultiplexed, trimmed, processed, and aligned/assembled to generate consensus sequences for RT stops analysis.

single-nucleotide polymorphisms. The gapped-read approach of bowtie2 is also of considerable advantage to identification of aptamer sequences, as multiple evolved sequences of an aptamer might emerge from the same selection. While variants calling can be accomplished downstream of the bioinformatics workflow, gapped-read alignment position of leniency towards mutations is important in the context of aptamer discovery. Bowtie, however allows mismatches, is more stringent as it was originally coded for short reads, as mismatches within shorter reads are more unforgiving to the alignment score of a sequence.

Alignment output of Bowtie2 can be sorted and indexed to provide chromosomal coordinates, most probable genotype, and relative abundance of aptamers. These post-alignment processing steps are universally performed with SAMtools and bcftools[60]. Finally, a graphic interface is used to visualize the sorted and indexed alignments. Both Integrated Genome Browser and Integrated Genome Viewer are popular tools for interactive genome nagivation[132,133]. Integrated Genome Browser includes quick load server and many useful analytics, in addition to ease of configurability, thus was the interface of choice for genomic annotation.

Should Apta-Seq apply to an *in vitro* selection performed on a synthetic DNA library, the reads must be assembled into consensus sequences. These sequences serve as a reference upon which alignment would be performed to provide SHAPE reactivities output.

As mapped reads of Apta-Seq are in essence reverse transcribed sequences, the aligned 5´-ends represent the nucleotides immediately downstream of the acylation sites. The galaxy module StructureFold computes frequencies of reverse transcription stops, and correlates such frequencies between acylated and DMSO dataset to generate SHAPE reactivities, followed by a

secondary structure prediction algorithm that utilizes established approaches such as ViennaFold and RNA Structure to predict the structural changes of selected aptamers in the presence of ligand[57,134–136].

Apta-Seq pipeline led to fast and reliable discovery of novel RNA molecules

Apta-Seq was performed on the final round of *in vitro* selections of adenosine triphosphate, cyclic-GMP, and cyclic-AMP RNA aptamers in a human genomic library. The results of the cyclic nucleotides aptamers were previously discussed in Chapter 2 and 3 of this document. Apta-Seq proved essential in ATP aptamers characterization, validating the previously reported aptamers[20], as well as providing additional information regarding their alignments and abundances.

Table 4-1: Abundance and affinity of reported ATP aptamers

Entry	Rel. abundance (%)	K_D (μM)
FGD3	0.3839	130, 700
ERV1	0.4893	476
PRR5	0.1742	950
L1PA15	0.0848	940
THE1B-ERVL	0.1191	980

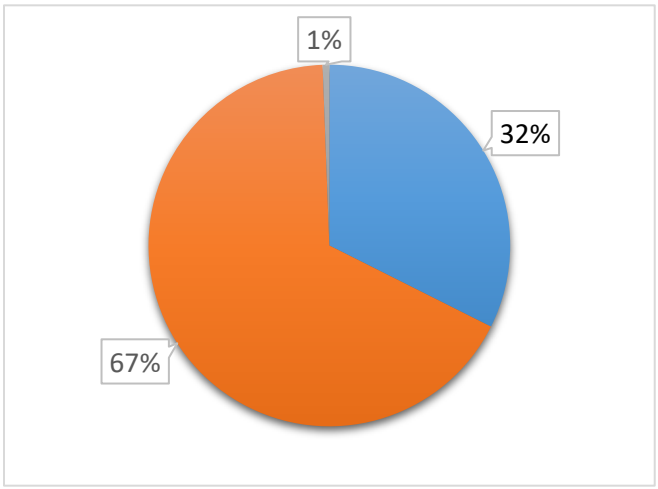
Relative abundance for each aptamer was calculated out of 1533526 reads mapped to the human genome.

While the protocols described in Loughrey *et. al.*[55] neglected to mention the critical influence of adapter population present in sequencing dataset, Apta-Seq has incorporated a polyacrylamide gel electrophoresis of the amplified products of linearized cDNAs prior to the barcode installation that concluded the sequences for submission. This purification procedure was expected to eliminate artifacts generated by various amplification steps throughout the library preparation process, such as self-ligated adapters, non-priming oligonucleotides, and

A

```
Time loading reference: 00:00:00
Time loading forward index: 00:00:00
Time loading mirror index: 00:00:00
Multiseed full-index search: 00:00:04
172686 reads; of these:
  172686 (100.00%) were unpaired; of these:
    44711 (25.89%) aligned 0 times
    126656 (73.34%) aligned exactly 1 time
    1319 (0.76%) aligned >1 times
74.11% overall alignment rate
Time searching: 00:00:05
Overall time: 00:00:05
```

B



C

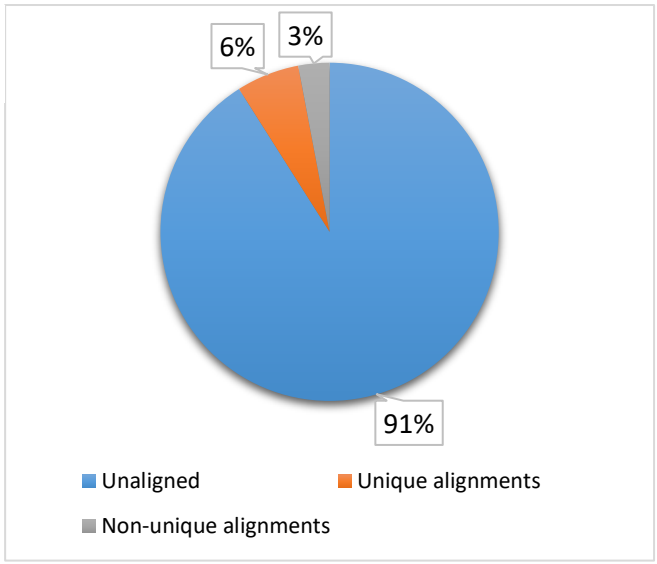


Figure 4-6: Alignment statistic using bowtie2. (A) Bowtie2 output statistics regarding the alignment of raw reads. A library prepared with purification (B) aligns better than a library prepared without removal of primers, primer-dimers, and self-ligated amplicons.

over-amplified constructs. As bowtie2 outputted statistics for each alignment (Figure 4-6), the dataset with the purification step included aligned significantly better than the dataset without.

Other variables were investigated for their potential influences on the processing of demultiplexed dataset. As bowtie2's end-to-end alignment proved too stringent for the fastq file of the ATP aptamer selection, --local setting was applied, in which the program allows gaps, mutations, and insertions in the mapped reads to give the best matches possible. This option has helped identified new ATP aptamers within the selection, as well as confirmed the presence of previously reported aptamers. In addition, --non-deterministic was examined as an alternative in bowtie2 alignment, due to

the advantages it offers for reporting identical reads[58]. Due to the nature of genomic SELEX, aptamers might have evolved from overlapping or identical sequences from different regions of the reference genomes, thus --non-deterministic helps address such possibility. Furthermore, previously established protocols have taken the initiative to remove duplicated amplicons from the dataset[121]. However, as the abundance of an aptamer directly correlates with the efficiency of a selection, genomic alignment performed on their resulting cDNA sequences should keep all occurrences of a read.

Moreover, consensus sequences of selected aptamers can be assembled with ABySS[64], upon which bowtie2 alignment was performed. This approach was extremely useful in identification of synthetic aptamers or ribozymes, where the DNA library was randomly generated through phosphoramidite synthesis. There are also merits in constraining Apta-Seq algorithm onto a specific region, however this approach allowed mismatches in alignment. Genome assembly with raw fastq reads might bias towards high frequency aptamers, as consensus sequences with more depth enMoy higher priority in assembly, which might not be beneficial for low abundance aptamers, whose frequency might not directly correlate to their affinity and specificity[137].

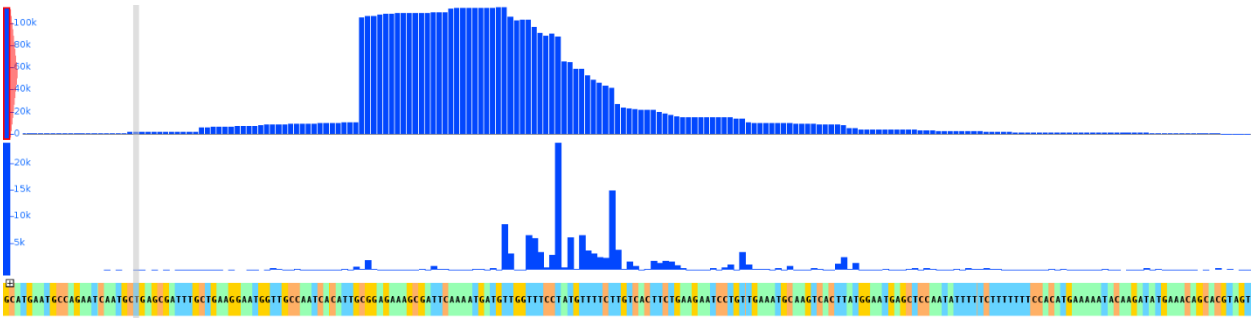


Figure 4-7: Integrated Genome Browser can generate consensus sequences and RT-stops counts, that directly output as .bed format for downstream analysis.

Output of Apta-Seq: Frequency of RT stops and structure predictions

Apta-Seq provides output as the amount of reverse transcription stops at any given nucleotide of an RNA sequence. This output can be obtained via the StructureFold module from Galaxy, or as a graph function of Integrated Genome Browser. The RT stop count module of StructureFold is based on iterative mapping of raw reads onto a reference sequence, where each 5′ end is counted towards a reverse transcription stop, followed by the stops binned together on top of the nucleotide. RT stops from an acylated samples are compared against those of the DMSO control, upon which SHAPE reactivities were calculated, thus leading to structure prediction by built-in Vienna or RNA Structure algorithm.

While most users of StructureFold are interested in the final structure prediction output, the RT stop counts were instrumental to the determination of binding constants for each aptamer discovered. The RT stops at a nucleotide are charted across all applicable titrations of the ligand, then normalized and compared to the concentration of RNA present. This data can also be obtained via Integrated Genome Browser, which supports BEDgraph output, where the RT stop refers to the chromosomal position of the nucleotide. However, this approach does not address possible mutation that a genomic aptamer might exhibit due to mutation introduced during the selection process, or single nucleotide polymorphism between the research subMect's genome and the database (Figure 4-7).

Conclusion

In summary, a high-throughput analysis was developed to provide sequence identity, binding constants, and structure prediction of selected RNA aptamers. This method can be

broadened to DNA and unnatural oligonucleic acids selections, given an assembly of their library is provided. The application of Apta-Seq not only validated the results of previously established selection, but also discovered new aptamers. This pipeline shall be a powerful tool in the discovery of potential riboswitches.

Materials and Methods

Synthesis of methyl 2-methyl nicotinate

In a dry 50-mL round bottom flask with a magnetic stir bar, 20 mL of dry MeOH was added, followed by addition of methyl 2-methyl nicotinic acid (0.4246 g). Reaction flask was heated to 70^{o}C under reflux, then concentrated sulfuric acid (4 mL) was added dropwise. The reaction was heated for 30 minutes, allowed to cool to room temperature, followed by sodium carbonate workup (50 mL, saturated). The solution was mixed, and transferred to a separatory funnel. The organic layer was extracted, washed with saturated sodium carbonate, brine, and purified through flash chromatography in 1:1 dichloromethane:ethylacetate to give an isolated yield of 74.9%. Purified product was confirmed via proton NMR and mass spectroscopy.

Synthesis of methyl 2-(azidomethyl) nicotinate

In a dry 50-mL round bottom flask with a stir bar, anhydrous dichloromethane (5 mL) was added, followed by addition of methyl 2-methyl nicotinate under inert gas. Subsequently, trichloroisocyanuric acid (1.5 molar equivalence) was added. The reaction was stirred at room temperature overnight, then transferred onto an ice bath. 40 mL of saturated sodium bicarbonate was added to the reaction mixture, followed by transfer to the separatory funnel. The organic layer was extracted twice with additional dichloromethane, washed with brine, dried with magnesium sulfate, and condensed over reduced pressure.

Crude methyl 2-chloromethyl nicotinate was transferred onto a dry 25-mL round bottom flask, followed by addition of anhydrous dimethylformamide (DMF). Sodium azide (2 molar equivalences) was added into the flask, and stirred at room temperature overnight. The reaction was quenched with sodium bicarbonate, extracted with ethylacetate, and dried with brine and magnesium sulfate. Product was confirmed with proton NMR and mass spectroscopy, and gave a crude yield of 55.5%.

Deprotection of methyl 2-(azidomethyl) nicotinate to 2-(azidomethyl) nicotinic acid

In a scintillation vial with a mini stir bar, methyl 2-(azidomethyl)nicotinate was dissolved in methanol (2.5 mL), followed by addition of aqueous sodium hydroxide (10%, 2.5 mL). Reaction progress was monitored via thin layer chromatography at 10% methanol: dichloromethane until the starting material spot disappeared. The reaction mixture was diluted with 12.5 mL of deionized water, and washed with 5 mL ethyl ether. Aqueous hydrochloric acid (10%) was added until the reaction solution reach pH 4, then extracted with 25 mL ethylacetate five times, dried with magnesium sulfate, and condensed under reduced pressure. Condensed mixture resulted in a yellow solid, which was further desiccated overnight under high vacuum. Crude yield of the reaction was 37.4%, suggested ineffectiveness in extraction of product.

Synthesis of 2-(azidomethyl) nicotinic acid acyl imidazole

In a small vial containing 2-(azidomethyl) nicotinic acid in anhydrous dimethyl sulfoxide, carbonyldiimidazole (1 molar equivalence) was added under inert gas. The reaction was kept away from light and stirred for 1 hour. Reaction mixture was transferred onto a 1.5 mL plastic vial to create a solution of NAI-N3 at 800 mM. The vial was aliquoted into multiple vials and

stored at – 80°C wrapped in aluminum foil for SHAPE experiments. Product formation was confirmed via mass spectroscopy.

Gel shift assay to determine reactivity of NAI-N_3

Three reactions were prepared: 1) a no reaction control containing [^{32}P] α-ATP and DMSO, 2) a reaction containing NAI-N_3 and [^{32}P] α-ATP, and 3) a reaction between [^{32}P] α-ATP and triazole adduct of NAI-N_3 and DBCO-amine (see Chapter 5). The reactions were prepared at 37°C for 15 minutes, followed by electrophoresis analysis on a 15% polyacrylamide matrix.

Table 4-2: Final concentration of NAI-N_3 gel shift reactions

Reagent	DMSO control	Reaction 1	Reaction 2
[^{32}P] α-ATP	0.33 x 10^{-3} mM	0.33 x 10^{-3} mM	0.33 x 10^{-3} mM
NAI-N_3	0 mM	40 mM	40 mM
DBCO-amine	0 mM	0 mM	40 mM

Oligonucleotides for SHAPE reverse transcription, Circ ligation PCR, and barcoding PCR

SHAPE-RT: 5´-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTGCGGCC

GCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCTGAGCTTGACGCA-3´

Illumina Universal-F: 5´-AATGATACGGCGACCACCGA

GATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3´

Circ-ligation-R: 5´-GTGACTGGAGTTCAGACGTGTGCTC

TTCCGATC CTGAGCTTGACGCATTG -3´

Barcoding-R: 5´-CAA GCA GAA GAC GGC ATA CGA GAT *NNNNNN* GTG ACT GGA GTT

CAG ACG TGT GCT CTT CCG-3´

*where NNNNNN is the reverse complement of the barcode as it was read by the sequencing instrument

Acylation of RNA aptamers with NAI-N_3 in increasing concentration of target ligand

Transcribed RNA aptamers were purified via polyacrylamide gel electrophoresis, and reconstituted to a stock concentration of 1 µM in binding buffer (see Chapter 2 for buffer conditions). The suspended RNA solution was heated at 70°C for 5 minutes, followed by division into aliquots of the target ligand (0, 1, 10, 10^2, 10^3, 10^4 µM) to a final volume of 10 µL. The reactions were incubated at room temperature for 15 minutes, followed by addition of NAI-N_3 to a final concentration of 100 mM. The acylation reaction proceeded for 15 minutes, after which it was quenched with 90 µL of distilled, deionized water, and precipitated in ethanol and glycerol.

Table 4-3: Reagents and concentrations for aptamers acylation

Reagent	Final concentration
Binding buffer	1x
Target ligand	1 - 10^4 µM
RNA	0.1 µM
NAI-N3	100 mM

Reverse transcription of acylated RNA aptamers

Acylated RNA aptamers were collected, reconstituted in water, and filtered on G25 Sephadex column to remove twice to remove excess NAI-N_3. The supernatant was subsequently transferred to 20 µL drained DBCO-agarose on SpinX filters (see Chapter 5 regarding preparation.) These beads were meticulously washed to remove unreacted DBCO-amine from the succinimide beads coupling reaction. After 15 minutes of incubation between acylated RNAs and DBCO-agarose, the beads were washed stringently to eliminate non-acylated sequences,

followed by addition of reverse transcription solution. Once the reverse transcription was completed, 1 µL of NaOH (4M) was added to the beads suspension and the filters were heated to 80°C for 10 minutes to degrade RNAs. The bead suspensions were subMected to centrifugation, and the flowthrough was precipitated in glycerol and ethanol.

Table 4-4: Reagents and concentrations for Apta-Seq RT

Reagent	Final concentration
MMLV buffer	1x
MMLV enzyme	25 units
dNTPs	1 mM

Circ-ligation of Apta-Seq cDNAs, and purification of circDNAs

CircLigase, reagents, and protocols were obtained from Epibio. Reactive condition for the ligation followed Table 4-5: the reagents were added to cDNAs suspended in water to a final volume of 20 µL. Reaction was incubated at 60°C for 1 hour, followed by 80°C for 10 minutes. The circularized DNA sequences were amplified with aforementioned Illumina Universal-F and Circ-ligation-R. After reaction was purified on a polyacrylamide gel, and visualized with SYBR stain, the band representing the heterogeneous products were excised, eluted in aqueous potassium chloride (300 mM), and precipitated in glycerol and ethanol.

Table 4-5: Reagents and concentrations for Circ ligation

Reagent	Final concentration
CircLigase buffer	1x
Circ ligase	5 units
ATP	1 mM
$MnCl_2$	50 mM

Illumina adapter installation and sequencing

A polymerase chain reaction was performed using Illumina Universal-F and Circ-ligation-R. Amplified products were barcoded via a subsequent PCR with a reverse primer containing the barcode of interest. Amplicons were verified using ethidium bromide-stained agarose gel electrophoresis, and sequenced on Illumina HiSeq 2500 at UCI Genomics Facility.

Text trimming and processing of sequencing reads using cutadapt

The following protocols were accomplished on Ubuntu 14.04. Programs and packages can be obtained and/or downloaded via Ubuntu Terminal, or the Application Manager of the operating system. More information can be found on the program's website.

Command line

cutadapt -a file:<3´-adapter> -g file:<5´-adapter> -o <output.fastq> <input.fastq> -m 25

The primers used for Illumina installations, as well as constant regions of the genomic library, were saved as a file and called directly to the command for trimming. The option –m was used to specified the minimum length of trimmed reads to 25 nucleotides, *i.e.* shorter reads were removed from the output file after trimming. Cutadapt can also be adMusted to keep the adapters in the output file, among many other options.

Mapping trimmed reads onto human reference genome

Indexed genome was obtained from National Center for Biotechnology Information (NCBI) or UCSC Genome Downloads.

Command line

bowtie2 –x /path/to/indexed/genome –U <input.fastq> -S <output.sam> --time --non-deterministic –k <integer>

Option –U allowed alignment of single-end reads to the reference genome. In the event of paired-end reads, options -1 <mate1.fastq> -2<mate2.fastq> were invoked to specify both inputs. Option --non-deterministic helped bowtie2 report different alignments for identical reads, while –k dictated the aligner to report the specific instances of alignments.

Sequence alignment file formatting, sorting, and indexing

Command line

samtools view –bS –o <output.bam> <input.sam>

samtools sort <output.bam> > <output.sorted.bam>

samtools index <output.sorted.bam>

The –bS indicated the output is .sam, while the input is .sam. The .bam is a binary output of the alignment, and needed to be sorted and indexed. Samtools could also provide an interactive map of the genome, as well as output specific sequences from a chromosomal space restricted by its coordinate.

samtools tview <output.sorted.bam> <reference.fasta> -p <coordianates>

samtools view –bo <output.sorted.bam> <coordinate> -o <regional.bam>

samtools bam2fq <regional.bam> > <regional.fastq>

Sequences pertaining to a specific location, *i.e.* mapped into an aptamer, can be extracted with the subcommand bam2fq, provided the input .bam were previously restricted to the aptamer's chromosomal location.

Genome assembly with ABySS

Short reads generated by Illumina sequencing can be assembled into contigs, as for the selection of cyclic nucleotide aptamers from a metagenomic DNA library (Chapter 2), or the

selection of multi-turnover ribozymes of strained promoted alkyne-azide cycloaddition (Chapter 5). Genome assembly is a computational intensive process and is recommended to perform remotely on a cluster.

Command line

abyss-pe k=<integer> name=<output> in='input1.fastq input2.fastq'

However designed to perform assembly of paired-end libraries, ABySS can be configured to read single-end with the option se.

Sequence analysis with FASTAptamer

Simple text processing were performed on FASTaptamer, instead of built-in terminal commands such as awk, sed, or grep. The binaries for fastaptamer were installed onto the local path variable.

Command line

fastaptamer_count -i <input.fastq> -o <output.fasta>

fastaptamer_cluster -i <input.fasta> -o <output.clustered.fasta>

fastaptamer_enrich –x <input1.fasta> -y <input2.fasta> -o <output.txt>

Fastaptamer were invoked to assess the abundance and frequency of a read within a population. It was also used to cluster sequences together to generate consensus sequences of aptamers. In addition, fastaptamer_enrich was used to calculate fold enrichment for each sequence across multiple dataset.

Short reads processing with FASTX-toolkit

Command line

fastx_collapser -i <input.fastq> -o <output.fastq>

Visualizing alignments with Integrated Genome Browser

Sorted .bam files were visualized with Integrated Genome Browser (IGB). The graphic interface provided tools for analyzing depth, chromosomal location, direction of reads, and quality score of each nucleotide.

Calculation of binding constants for aptamers

Reverse transcription stops were binned onto an aptamer's reference sequence, either via a fasta input or chromosomal coordinates. These values were normalized across all samples, and worked into the Excel Solver module to provide apparent K_D at each nucleotide.

Chapter 5

In vitro selection of multi-turnover ribozymes of strain-promoted alkyne-azide cycloaddition catalysis

Introduction

The discovery of the first catalytic RNA has compelled the RNA world hypothesis, which postulates RNA to be the macromolecule responsible for biological processes in the early evolution of life before peptide-based polymer rose to dominance[30,138]. The chemical repertoire of naturally occurring ribozymes encompass phosphotransferases and ribosomal peptidyl transferase, and can be found in regulation of gene expression and protein synthesis[139,140]. *In vitro* evolution has extended ribozymes' chemical capabilities, giving rise to, albeit artificial, transformations such as template-directed RNA elongation, Diels-Alder cycloaddition, and amide bond synthesis[141–144]. The advents of these selected ribozymes have in turn supported the hypothesis that RNAs were the primary catalytic unit of primordial life on Earth.

As the expansion of artificial ribozymes' reaction catalogs continue, it became evident that these chemical transformations not only should take advantage of highly orthogonal and robust chemistries, but also present a versatile application to biological investigation, so as to reinforce the notion that ribozymes are a chemically adaptable and specific catalytic entity. Aforementioned efforts in the ribozyme evolution community have yielded RNA molecules capable of a wide range of reactions, yet these chemical conversions might not necessarily be useful in direct exploration of cellular processes due to their lack of bioorthogonality and potential toxicity. Among the extensively utilized approaches of chemical biology is the click

reaction, where an azido precursor reacts with another containing an alkyne functionality, leading to the formation of a 1, 2, 3-triazole adduct[145]. The reaction typically requires a copper (I) catalyst, and has been used widely in chemical synthesis due to its high yield, reactive kinetics, and reliable stability[146]. However, a copper-based system might not be physiological inert,

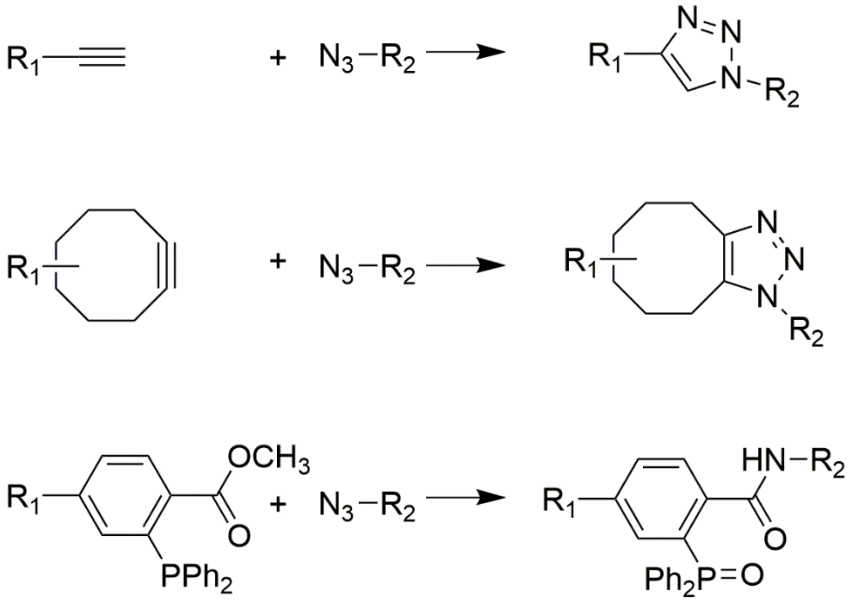


Figure 5-1: Bioorthogonal chemistries utilizing azide-precursor for bioconjugation. (*top)* Copper-catalyze click reaction, (*middle*) Strain-promoted azide-alkyne cycloaddition, (*bottom)* modified Staudinger ligation.

because copper has been found to generate hydroxyl radicals, leading to nucleotide degradation[147]. Due to such limitation, investigations aimed at monitoring and visualizing cellular processes in real time have adopted other bioorthogonal azide-coupling reactions such as modified Staudinger ligation[148] and copper-free, strain-promoted azide-alkyne cycloaddition (SPAAC)[149] (Figure 5-1). The copper-free cycloaddition takes advantage of the lack of Cu-mediated cytotoxicity, as well as the metabolic stability of its reagents, and is consequently more suitable for bioconjugations[150]. This reaction incorporates cyclooctynes derivatized with ring strained and electron-withdrawing elements in order to promote the formation of the triazole product. The copper-free click chemistry has enjoyed many application in bioorthogonal

studies, where appropriate reagents transfected into a biological system remain inert to chemical entities within, yet become reactive exclusively towards intended synthetic agents. Such approaches have proven powerful in labelling of azido sugars on the surface of the cell, purification of azido-peptide using cyclooctynes pulldown, and identification of azido amino acids installing enzymes within *E. coli*[151–153].

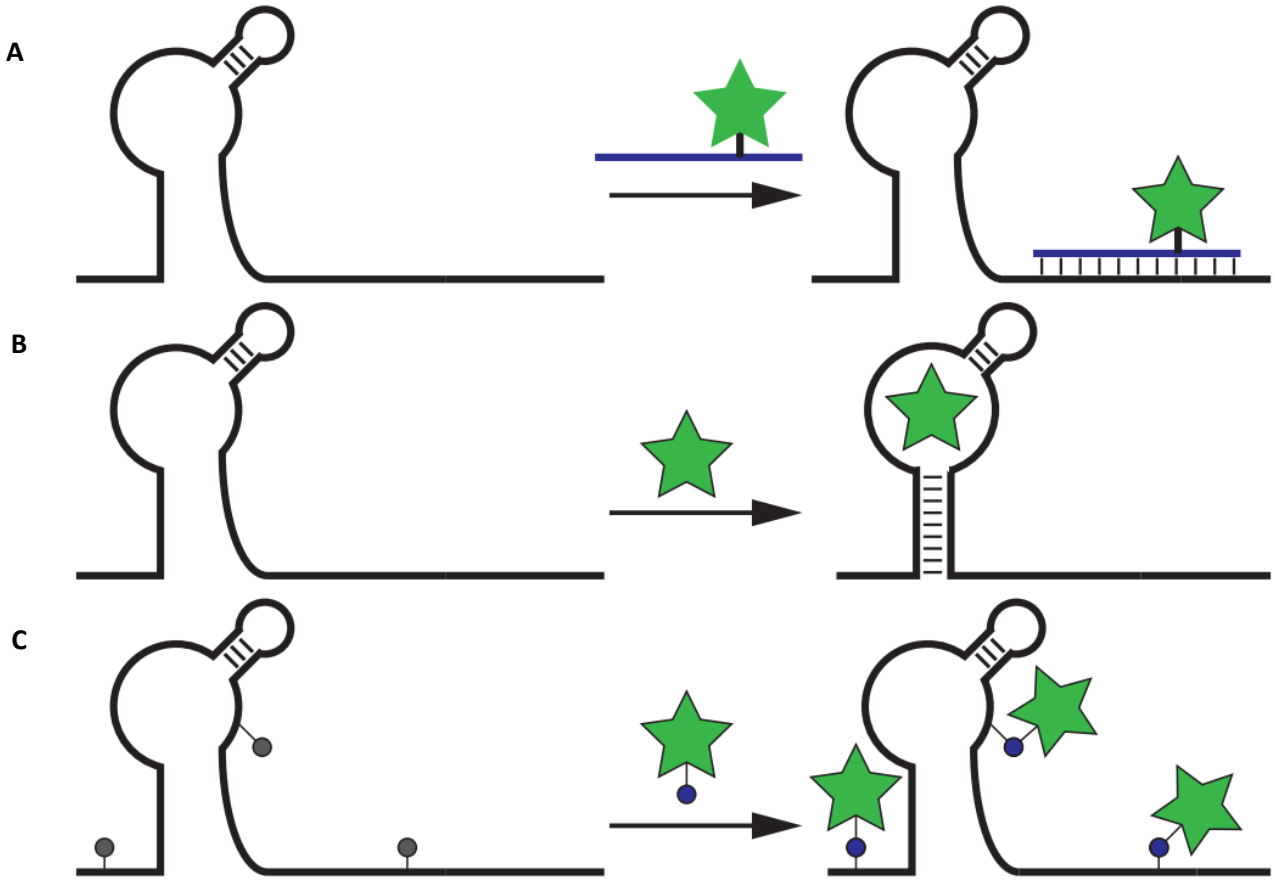


Figure 5-2: Strategies to visualize RNAs *in vivo.* Imaging of RNAs in living system typically uses (A) *in situ* hybridization, where an anti-sense strand containing a fluorescent tag hybridizes to the target sequence. (B) An aptamer domain fused to the target sequence that binds to a fluorophore. (C) Chemically modified RNAs containing bioorthogonal functional group for fluorescent conjugation.

A typical application of bioorthogonal click chemistry in live cells is fluorescent visualization of cellular processes. Current efforts in fluorescent detection of RNAs in living systems have started with *in situ* hybridization (ISH)[154]. ISH is particularly useful when visualizing

RNA containing a known sequence. The general approach involves hybridization of a DNA/RNA strand containing a fluorophore via canonical base-paring to a target sequence, followed by detection, known widely as FISH. This technique has shown to be highly specific, as Larson C. *et al* have shown to be capable of single nucleotide selectivity[155]. However the approach, as well as its derivatives (such as molecular beacon and forced intercalation), relies on exogenous chemicals that might not be sufficiently membrane permeable, not synthesized by cellular machineries, and requires the target sequence to be known in advance[156]. Researchers have also turned their attention to chimeric or transgenic constructs containing aptamer domains specific to fluorophores such as Spinach, Broccoli, and Mango[16,17,157]. These aptamer domains were evolved with *in vitro* selection to be highly selective towards small molecule fluorophores, and thus can be affixed onto a sequence of interest to track its location within the cell in real time. Other methods take advantage of covalent modifications to the RNA sequences by introducing unnatural ribonucleotides such as 5-ethylnyluridine, which has been shown to be phosphorylated into its triphosphate equivalence readily within the cell[158], or azido-UTP, which has be to transfected into the cell[159]. These approaches, however adopt the versatile click chemistry, necessitate extensive modification of the RNAs, and might not suffice in visualizing cellular processes with timescale incomparable to the transfection efficiency of nucleotide analogs, or concentration in milieu.

A better approach to real time, *in vivo* imaging of RNAs using strain-promoted azide-alkyne cycloaddition would be a multi-turnover ribozyme catalyzing the reaction appended onto the RNA of interest. Previously described approaches have limited applicability towards low abundance RNA, where the fluorescent signal depends on the stoichiometry of hybridized,

bound, or conjugated fluorophore. A multi-turnover ribozyme approach promotes signal amplification, reducing the amount of fluorescent probes used *in* vivo, as a single copy is capable of generating many fluorescent molecules. The copper-free cycloaddition has been optimized for very high yield, although it still has a slow reaction rate at low reactant concentrations due to its bimolecular nature. Furthermore, the concentration of click reagents needs to be sufficiently low so that they do not interfere with cellular processes, yet still enough to emit quantifiable signal for tracking. These challenges encourage an enzymatic approach, where a multi-turnover ribozyme can accelerate the cycloaddition at sub-micromolar range of azide-alkyne. Given the catalytic repertoire of *in vitro* selected ribozymes, a ribozyme proficient of SPAAC would make an attractive target, and the rate enhancement capability would make *in vivo* imaging of RNA molecules widely applicable to many biological systems.

Designing an *in vitro* selection for a multi-turnover ribozyme of SPAAC

Since the target of the *in vitro* selection is a ribozyme enhancing a robust and well characterized chemistry, the non-catalyzed rate of the strain-promoted azide-alkyne cycloaddition needs to be established. As the application of the ribozyme is within the *in vivo* imaging toolkit, the click reaction should provide a fluorescent output well beyond the range of cellular autofluorescence[160]. The reagents should also include a chemical handle for covalent modification, in the event of separation of active RNA sequences, or immobilization of such reagents onto solid support. As earlier reports of multi-turnover ribozyme selections of Michael and Diels-Alder addition also depended on the reproducibility of these transformations, a SPAAC ribozyme should rely on the robustness of the reaction to identify the active RNA sequences. However, the aforementioned selections were designed to isolate active species by covalently

attaching the RNA sequences to the newly formed products. This approach might yield unintended single-turnover ribozymes, and do not reliably provide a solution to trace rate enhancement across multiple rounds of selections.

Thus the reagents for the SPAAC ribozyme selection should be fluorogenic in nature. In addition, they should be resistant, and ideally immune, to physiological-like conditions, while exhibit satisfactory solubility. Dibenzocyclooctyne-amine provides the strained alkyne moiety, along with an amino handle for coupling, while 3-azido-7-hydroxycoumarine serves as the fluorogenic azide (Figure 5-3). The resulting 1,2,3 triazole product emits an intense blue signal

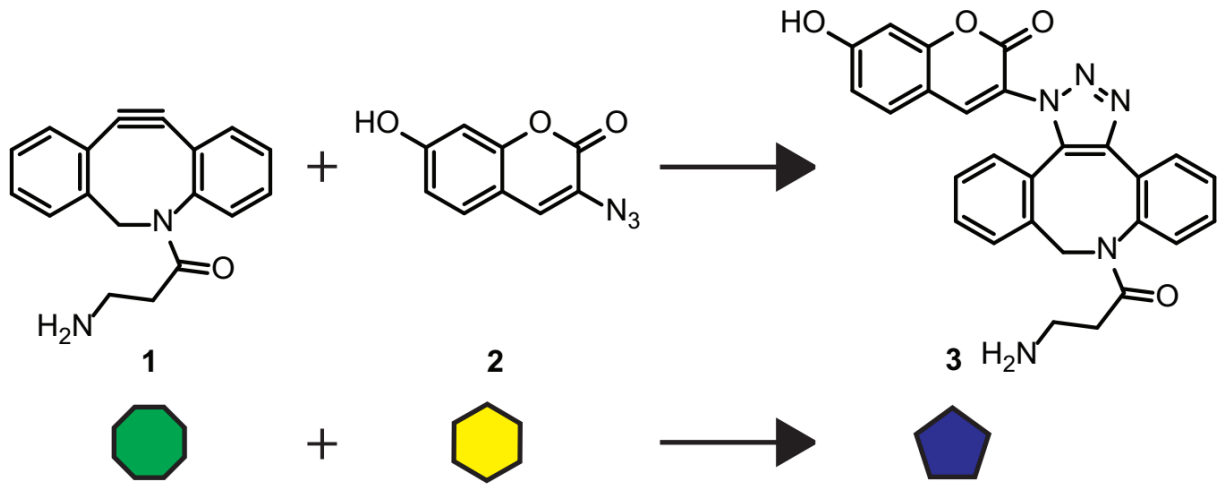


Figure 5-3: Strain-promoted azide-alkyne cycloaddition between dibenzocyclooctyne (1) and 3-azido-7-hydroxycoumarin generates 1, 2, 3-triazole adduct (3). The substrates and product are represented by the polygonal shapes for ease of clarification in subsequent text.

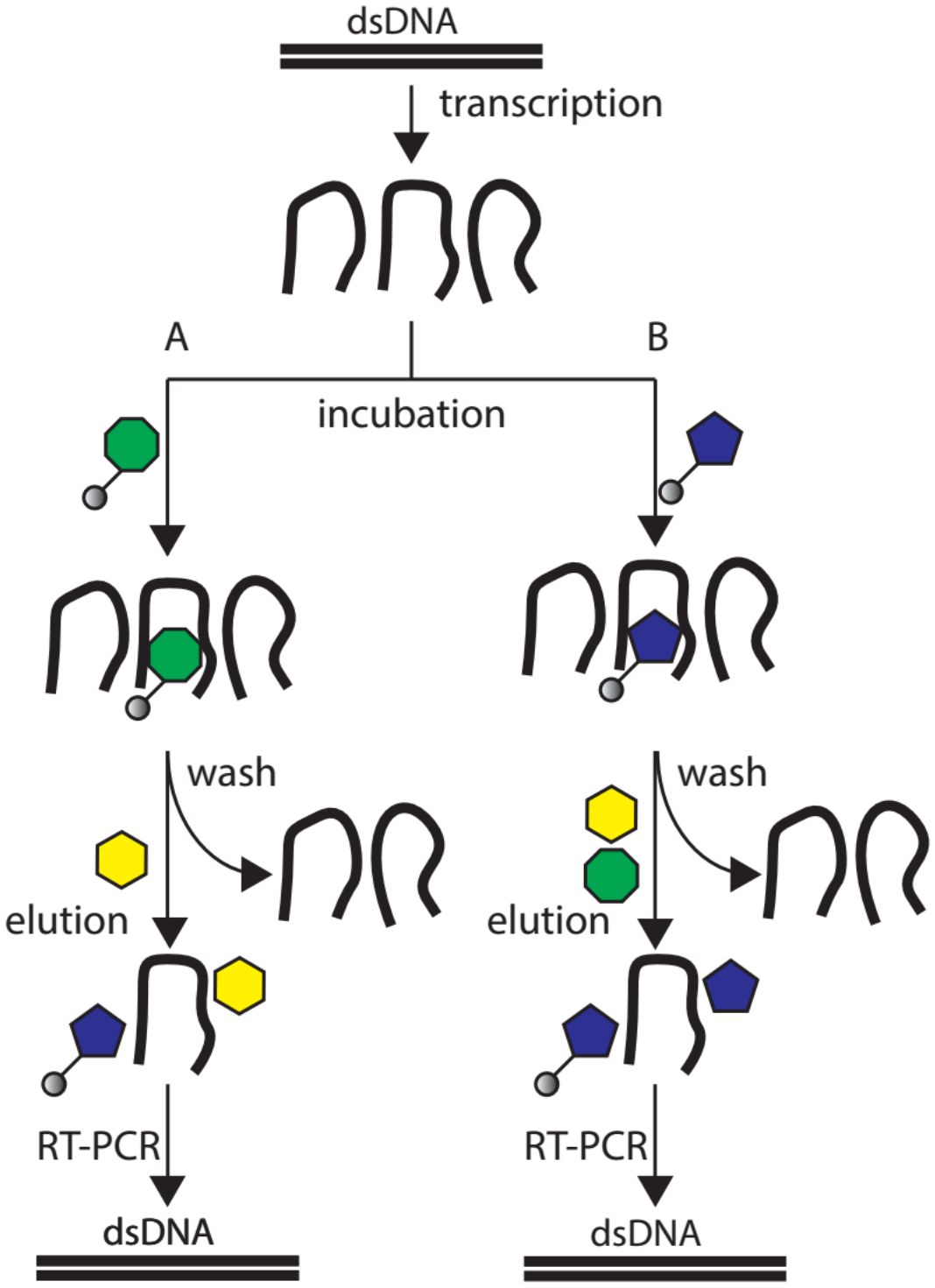


Figure 5-4: Selection strategies of a SPAAC ribozyme. Strategy A involves incubation of RNA sequences on DBCO-beads, followed by elution with **2**. Approach B requires annealing of sequences onto clicked-adduct agarose, succeeded by elution of both click reactants.

and has been used to detect nucleotides in cells[161,162]. This emission is useful when determining rate enhancement of selected RNAs across multiple rounds, as well as characterizing the non-enzymatic rate of addition.

Due to the bimolecular nature of the reaction, the rate of the non-catalyzed SPAAC can be determined using pseudo-first order approximation, where a reagent is supplied in excess, and the other reagent is added over a range of concentrations, upon which the product formation is determined via fluorescence. Given that the alkyne molecule has an amine group, the dibenzocyclooctyne-amine can be used as the excess reagent by immobilizing a high concentration onto *N*-hydroxysuccinimide matrix, while the coumarin derivative serves as the limiting reagent. The output signal of the clicked product can be scanned to spectroscopically determine the emission maximum, which will be used for future detection *in vivo*. The instrument of choice would be a fluorimeter that has a chamber that can accommodate a small volume of reaction. This is a practical consideration as the reference reaction is extremely efficient, therefore the concentrations of the reactants have to be sufficiently low. As the mixing of the reagents due to diffusion might cause a delay in emission of fluorescence, the endpoint kinetic measurement also needs to be adjusted accordingly.

As mentioned above, selection strategies that aim to isolate catalytic RNAs through tagging newly formed products with the selected sequences often yielded single-turnover ribozymes[163]. To navigate around this pitfall, the SELEX of a SPAAC ribozyme should be considered as an affinity selection, as evidenced that an enzyme should have higher affinity for the transition state, than for the reactants and products. This approach has inspired the designs of many transition state analogues in contemporary therapeutics targeting enzymes[164,165].

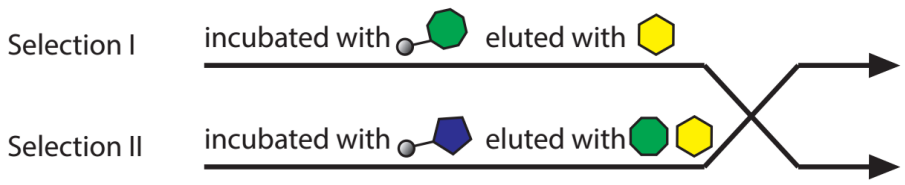


Figure 5-5: Two selections were performed for SPAAC multi-turnover ribozymes. Initially each selection followed their own elution format. The format was exchanged to discourage selection of azide-alkyne aptamers.

However, as the transition state of the strain-promoted cycloaddition, by definition, cannot be isolated, the selection can be directed towards obtaining RNA sequences with high affinity for both the azide and alkyne modules. This can be accomplished by annealing the RNA sequences onto either SPAAC substrate, then eluting with the other click reactant, or incubating the transcripts with the clicked triazole, followed by elution with the azide and alkyne molecules. The first approach assumes the RNAs of interest bind onto immobilized **1** with adequate affinity, and bring **2** to sufficient proximity and stabilize the transition state for the cycloaddition to occur. The second strategy exploits the tendency of the active sites of ribozymes to release the product, while obtaining the next turnover's reagents for the succeeding transformation. Both of these designs share the pitfall in that the selected sequences might not be inherently catalytic, *i.e.* RNAs with binding sites for the alkyne and azide separately would come out of the selection. Two parallel selections were performed using these formats, both started with a synthetic DNA library of 10^{14} diversity. For the first four rounds, each selection followed their own distinguished format. Once both selections have reached their fourth iterations, the formats were interchanged. This approach encouraged the selected sequences to have affinity for both the alkyne and azide molecules, and capability to bring these chemical species to proximity, thus accelerating the rate of the copper-free click reaction.

Fluorescence measurement played an important role in the design of the selection, from rate determination of background click reaction, to validation of rate enhancement across rounds of selection. Initially a single chamber fluorimeter was used to measure the output signal of triazole **3**, however as more reaction conditions were tested, the measurements were recorded using a 96-wells plate fluorimeter. The single-chamber fluorimeter was still useful in establishing control measurements, such as emission spectrum of click reactants and products. The 96-wells plate fluorimeter has enabled parallel measurements of reactions with different stoichiometry of alkyne and azide, as well as direct comparison of selection results between the two formats. The instrument was used to create a standard curve for clicked triazole fluorescence, which was used to reverse calculate unknown concentration of **3** on agarose

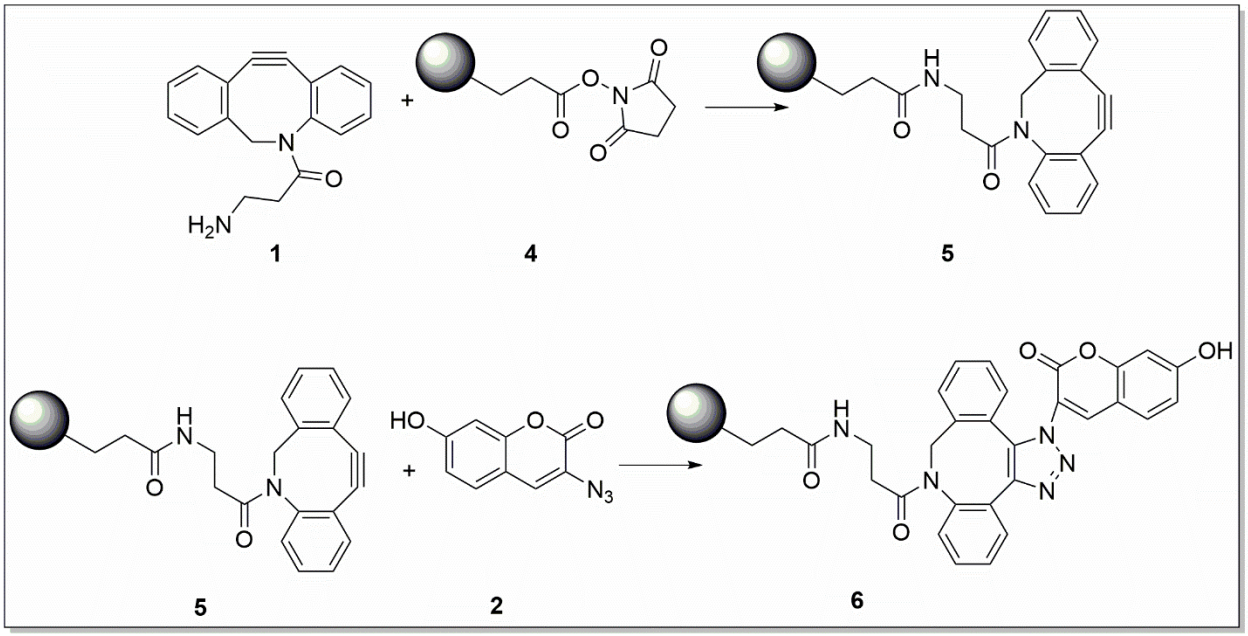


Figure 5-6: DBCO and 1, 2, 3-triazole agarose coupling

matrix, and in solution.

Beads coupling for the selection also required some considerations. First and foremost, due to the two formats, **1** and **3** needed to be immobilized onto agarose matrix. As DBCO-amine

does not exhibit any inherent fluorescence, the amount of compound covalently attached to NHS sepharose can be measured indirectly by saturating the coupled beads with **1**, followed by emission measurement of the **3** to determine the initial concentration of DBCO-amine present. Similarly, the 1, 2, 3-triazole beads can be analyzed by fluorescent spectroscopy. In addition, the storage condition for the reactants should ensure long term stability. As both compounds are commercially available in solid form, concentration stocks were prepared in anhydrous dimethyl sulfoxide, and diluted samples for coupling and selecting purposes were prepared in phosphate buffer. This was due to Tris-based buffer having been suggested by the vendor to facilitate unwanted side reactions with the click reagents. In addition, beads capacity was pivotal to the selection's success. As the yield of a purified transcription hovered around micro molar range, the amount of DBCO or triazole immobilized should be greater than the amount of RNAs introduced onto the beads to drive the equilibrium of the binding event to the bound state, thus guaranteeing that the sequences are likely to be annealed onto the immobilized ligands. Consequently, the coupling reaction should yield a high amount of ligands attached on beads. The coupling efficiency could be assessed by fluorescent reading for both selection formats.

Furthermore, selection condition was intentionally set to physiological-like. For the eventual application of the SPAAC ribozyme was *in vivo* imaging, it was integral that the selected RNAs adopted the appropriate conformation suitable for catalysis within cellular environment. In addition, since both selection formats included an elution step which maintains the addition of both click reactants, their concentrations needed to be sufficiently low so as to allow the RNA sequences to bind to the newly formed **3**, whereas high concentrations of **1** and **2** would lead

to near instantaneous formation of the adduct, and consequently cause the sequences to be selected for **3** affinity, not SPAAC catalysis.

Enriched DNA pool after each round of selection was transcribed, purified, and suspended in 1 μM of DBCO-amine in physiological-like buffer. Equal volume of 1 μM 3-azido-7-hydroxycoumarin was added to each reaction, making the final concentration of each click reactant at 0.5 μM. The rate of the cycloaddition in the presence of RNAs were monitored on 96-wells plate fluorimeter. A reaction without RNAs served as the no-enzyme control. The emission was monitored on a 0.5 hour basis, while the 24-hour time point was considered completion.

Experimental setup

The fluorescence measurements of rate determination for the non-catalyzed click reaction between **1** and **2**, as well as validation of beads coupling synthesis were completed on the JASCO FP-6300 located in the Corn Laboratory. The emission spectrum protocols included an excitation wavelength at 395.0 nm, measurement range from 415 to 615 nm, at a scanning speed of 100 nm/min and photomultiplier voltage of 750 V. The instrument could accommodate a single cuvette of 100 μL capacity. Solution of **1** and **2** were prepared at twice the reaction concentration and equal volume, thus once combined into the cuvette, resulted in the previously described final concentrations. Parallel measurements of fluorescence signal of the clicked reaction in the presence of RNAs were accomplished on BioTek Gen5 Microplate Reader set to endpoint kinetics mode. Samples were excited at 385 nm and emission spectrum started from 425 to 525 nm, with a resolution of 5 nm per step. The measurements included a 100 msec delay for mixing of reagents. In order to ensure all parallel reactions started at the same time,

multichannel pipettes were used to withdraw and combine the alkyne and RNA solution with the azide.

Determination of reference reaction rate and beads capacity via fluorescence measurements

The rate constant of the non-catalyzed click reaction was determined using pseudo first-order kinetics measurement between **1** and varying concentrations of **2** which resulted in the following integrated rate law (Appendix Figure 5-1):

$$Rate = 13.6\ (\pm 1.22)\ M^{-1}s^{-1}\ [DBCO\text{-}amine][3\text{-}azido\text{-}7\text{-}hydroxycoumarin]$$

For the quantification of DBCO immobilized on NHS beads, a standard curve of fluorescence signal from known concentrations of clicked triazole product at 0.01, 0.1, 1.0, and 10 µM was generated (Appendix Figure 5-1). Sepharose beads with clicked triazole covalently attached were measured for fluorescence, which approximated to 60 µM of DBCO-amide on the matrix per selection volume. Another version of dibenzocyclooctyne derivative were also available with the biotin chemical handle. The reagent can be coupled to streptavidin agarose to form DBCO-agarose, and was carried out with much success. However, RNA sequences might evolve to adopt affinity towards the biotin-streptavidin complex.

Fluorescence measurement of SPAAC in the presence of selected RNAs

Two selections were performed, following and exchanging formats as previously described. RNAs were labelled with [^{32}P] α-ATP to evaluate the amount present in each fraction. As the selections were not designed to isolate RNAs that could bind to the click reactants, the eluted fractions did not maintain a large amount of radioactivity, indicating high level of transcribed sequences.

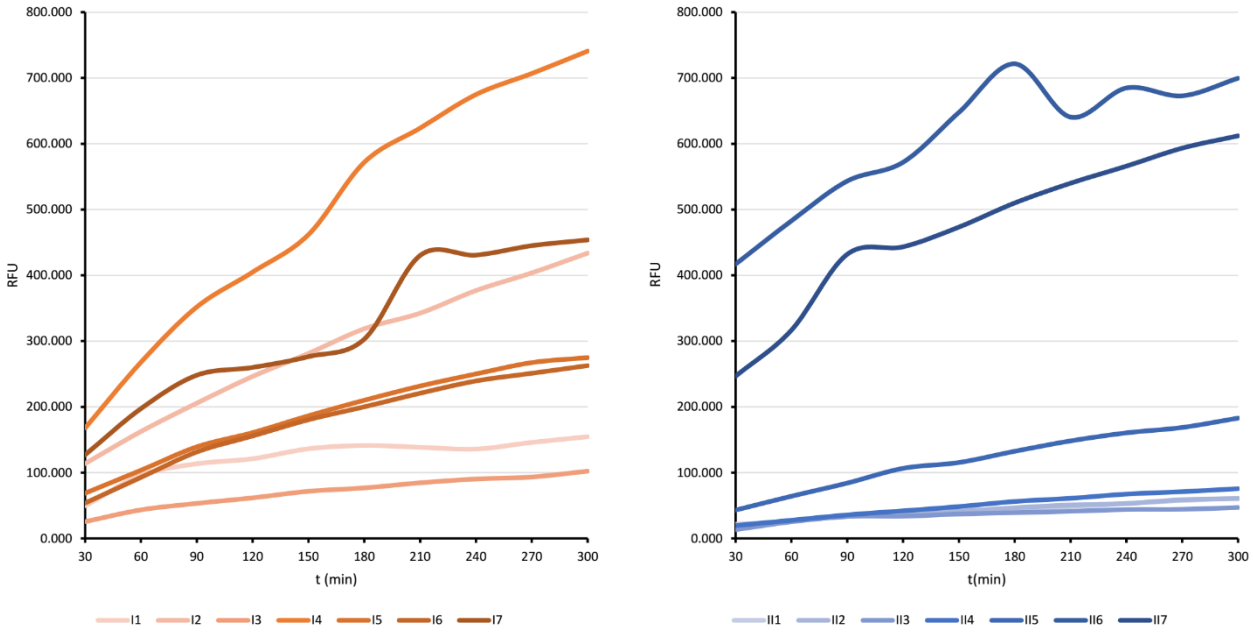


Figure 5-7: Fluorescent measurements of SPAAC selections following format I and II. Fluorescent detection of product formation over time of selection I (left) and II (right) showed promising differences in rate enhancement of cycloaddition in the presence of selected RNAs.

The results for background rate for the click reaction, as well as beads coupling efficiency, were considered while conditions for the SPAAC ribozyme selection were optimized. For instance, in format I, the elution buffer contained 10 µM 3-azido-7-hydroxycoumarin, and in format II, both elution buffers had 10 µM of each reactant. As defined by the experimentally determined rate equation, these concentrations were deliberated as the reference click reaction should be sufficiently slow to accommodate the release of the triazole **3** from any potential ribozyme's active site, followed by the binding of said ribozyme to the next turnover's reactants. In other words, should the click reagents' initial concentration stay too high, their binding event to the RNAs sequences might not proceed before the cycloaddition, rendering the selections ineffective. Seven parallel selection rounds were performed. Bulk fluorescence measurements suggested that the final round of selection II displayed promising results, while the effect of selection I was less pronounced. Subsequently, the last two rounds of each

selection became the point of focus for potential SPAAC ribozymes. We predict that as the RNA sequences selected in later round would give enhancement of fluorescent output, however the

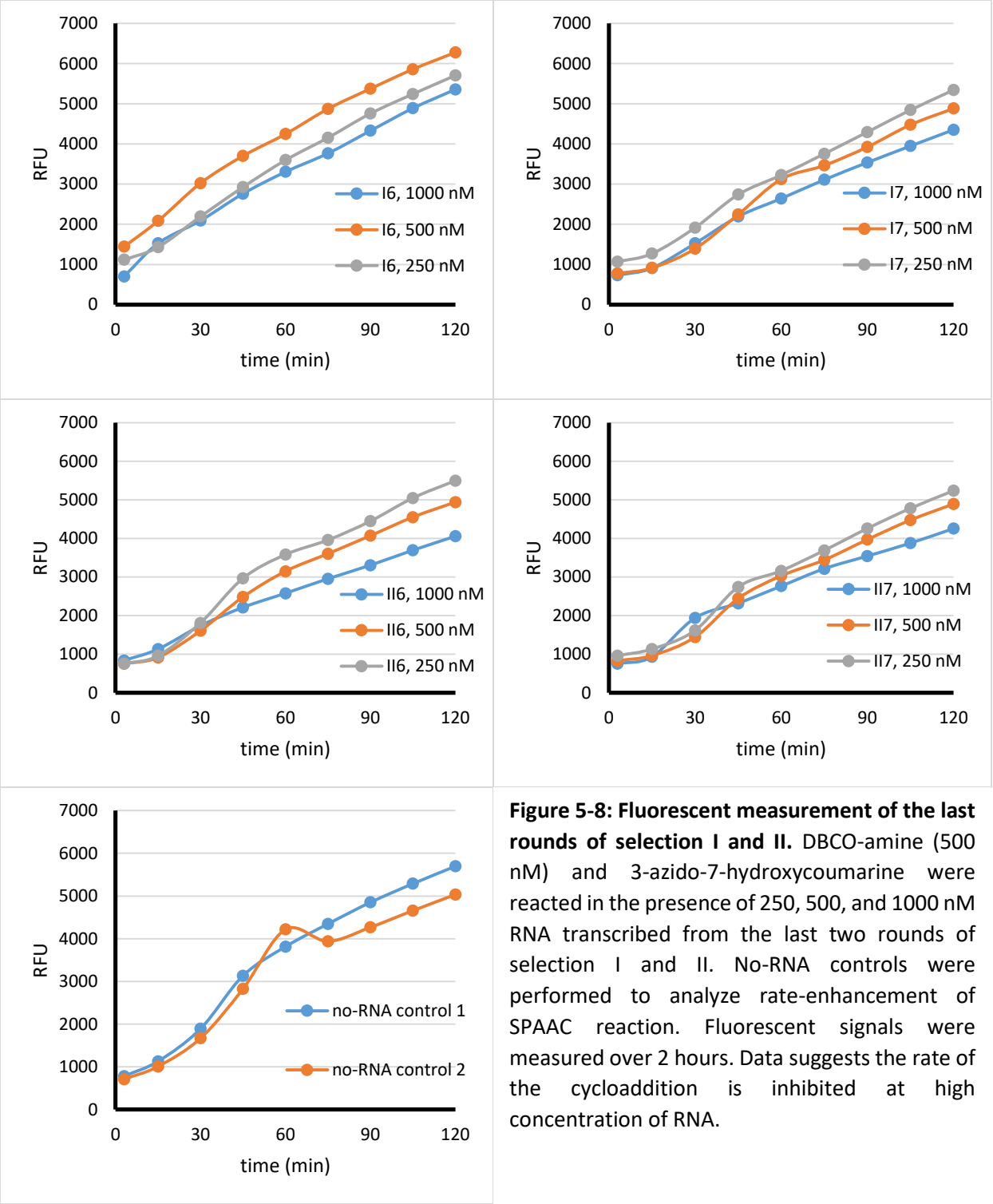


Figure 5-8: Fluorescent measurement of the last rounds of selection I and II. DBCO-amine (500 nM) and 3-azido-7-hydroxycoumarine were reacted in the presence of 250, 500, and 1000 nM RNA transcribed from the last two rounds of selection I and II. No-RNA controls were performed to analyze rate-enhancement of SPAAC reaction. Fluorescent signals were measured over 2 hours. Data suggests the rate of the cycloaddition is inhibited at high concentration of RNA.

data has been inconclusive at best (Figure 5-8).

The final rounds of both selections were transcribed, purified, and prepared by serial dilution to achieve concentration of 250, 500, and 1000 nM in physiological-like condition, while the concentrations of DBCO-amine and 3-azido-7-hydroxycoumarin remained constant at 500 nM. If these selected RNAs indeed exhibited multi-turnover catalysis, overall fluorescence signal should increase as RNAs become more concentrated. Surprisingly, the readouts of these measurements were higher in the earlier rounds of selection A and B, while rate enhancement over the non-catalyzed click reaction was observed. These suggestions, however, became confounding over the next set of data, where the concentration of RNAs was increased to 1 μM. The fluorescence signal between the samples and no-RNA control did not indicate any significant enhancement. In addition, the 1 μM RNA reaction in the presence of 0.5 μM azide and alkyne showed a lower fluorescence across all rounds, suggesting an inhibitory effect by the RNAs, where the affinity of the sequences might have play a role in keeping the reagents from completing the strain-promoted cycloaddition. This could be a side effect from the protocols of the selection, as they were designed to isolate RNA binders to the alkyne and azide species, rather than active sequences with catalytic activities.

The sequences obtained from both selections were submitted for high-throughput sequencing. Raw reads were assembly with ABySS to generate consensus sequences. However, these sequences did not show any significant rate enhancement to the strain-promoted azide-alkyne cycloaddition.

Conclusion

Our data suggested that the *in vitro* selection of multi-turnover ribozyme of strained-promoted azide-alkyne cycloaddition has isolated RNA sequences with affinity for

dibenzocyclooctyne and 3-azido-7-hydroxycoumarin. These sequences were examined for rate enhancement capability via fluorescent measurement of their 1, 2, 3-triazole adduct. Unfortunately the enriched pools have not demonstrated accelerated formation of clicked product. The selected sequences have shown potential affinity for the click reagents, as product formation was inhibited at high concentration of RNAs. The selection design should be re-contemplated, should the outcome of proximity-based ribozyme remains elusive.

Materials and Methods

Transcription, purification, reverse transcription, and amplification performed in this selection followed previously described protocols in Chapter 2.

Primers used for reverse transcriptions and amplification of selected RNAs

Forward primer: 5′-GCTAATACGACTCACTATAGGGAGATCACTTACGGCACC-3′

Reverse primer: 5′-CGCTGTCCCGAGCCTTGG-3′

Dibenzocyclooctyne sepharose synthesis

DBCO-agarose was generated by coupling DBCO-amine (Sigma-Aldrich) with *N*-hydroxysuccinimide activated sepharose (NHS) obtained from GE Healthcare Life Sciences. DBCO-amine was purchased as a solid and prepared as a solution in dimethylsulfoxide at 1 mM. The stock solution of DBCO-amine was aliquoted and diluted to 100 mM in a total volume of 500 µL phosphate buffer. NHS-sepharose (250 µL) was collected from the container via wide loading tip, and washed with copious amount of phosphate buffer. The DBCO-amine aliquot and NHS-activated sepharose were combined in a SpinX filter and let shake gently at room temperature overnight. Completed reaction was washed and store in phosphate buffer at 4°C.

1, 2, 3-triazole sepharose synthesis

DBCO-agarose from the previous description was subsequently reacted with 3-azido-7-hydroxycoumarin to generate their respective 1, 2, 3 adduct sepharose. DBCO-agarose (200 µL) were obtained and drained via centrifugation, followed by addition of 3-azido-7-hydroxycoumarin (100 mM in phosphate buffer). The azide reagent was purchased from A&Q Nano under commercial name iON-Blue as a solid, and was prepared as a solution of 1 mM in DMSO. The click reaction was run for 2 hours at room temperature, washed with copious amount of phosphate buffer, and stored in buffer at 4^{o}C.

Quantification of 1, 2, 3-triazole sepharose coupling efficiency

Clicked triazole-sepharose were measured for fluorescence. A standard curve of fluorescent signals from known concentrations of 1, 2, 3-triazole adduct (0-10 µM of **3** in phosphate buffer). A linear equation representing the regression of signal was generated and used to determine concentration of triazole immobilized on sepharose.

Pseudo-first order measurement of SPAAC reaction

The rate constant of the non-catalyzed click reaction was determined using pseudo first-order kinetics measurement between **1** at 100 µM and varying concentrations of **2** of 0.25, 0.5, and 1 µM. Reagents were prepared in phosphate buffer and combined to a final volume of 100 µL in cuvette. The emission signal was measured on Masco FP-6300 over time. A final measurement at t=72 hour was taken as completion time point. Fluorescent values were plotted and fitted onto a first-order regression, upon which the rate constant was determined.

Fluorescent measurement of the SPAAC reaction in the presence of selected RNA sequences

RNAs selected from each round was purified and diluted to equal concentrations, followed by incubation with 1 µM DBCO-amine on 96-wells plate. Subsequently, 1 µM of the

azide reagent at equal volume was added, yielding a final concentration of each reactant at 500 nM, after which the fluorescence measurement began.

Library preparation for high-throughput sequencing of selected pool

Procedures to transform and subMect the RNA sequences selected for SPAAC multi-turnover ribozyme followed previously described methods from Chapter 4. As the synthetic pool has different constant regions, the primers used for Apta-Seq of these sequences must thereby contain appropriate sequences for reverse transcription, ligation, and amplification.

Forward primer: 5′-CACGACGCTCTTCCGATCT GGG AGA TCA CTT ACG GCA C-3′

SHAPE reverse primer: 5′-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT GCGGCCGC GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC CGACGCGCTGTCCC-3′

Circ-ligation reverse primer: 5′-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC CGA CGCGCTGTCCCCG-3′

REFERENCES

1. Westhof, E. & Auffinger, P. RNA Tertiary Structure. *Encycl. Anal. Chem.* 5222–5232 (2000). doi:10.1002/9780470027318.a1428
2. Ren, A., Rajashankar, K. R. & Patel, D. J. Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch. *Nature* **486,** 85–89 (2012).
3. Chen, L. *et al.* The isolation of an RNA aptamer targeting to p53 protein with single amino acid mutation. *Proc. Natl. Acad. Sci.* **112,** 10002–10007 (2015).
4. Ray, P. & White, R. R. Cell-SELEX Identifies a 'sticky' RNA Aptamer Sequence. *J. Nucleic Acids* **2017,** (2017).
5. Nahvi, A. *et al.* Genetic control by a metabolite binding mRNA. *Chem. Biol.* **9,** 1043 (2002).
6. Breaker, R. R. Riboswitches and the RNA world. *Cold Spring Harb. Perspect. Biol.* **4,** 1–15 (2012).
7. Grundy, F. J. & Henkin, T. M. The S box regulon: A new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. *Mol. Microbiol.* **30,** 737–749 (1998).
8. Gelfand, M. S., Mironov, A. A., Jomantas, J., Kozlov, Y. I. & Perumov, D. A. A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Trends Genet.* **15,** 439–442 (1999).
9. Montange, R. K. & Batey, R. T. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature* **441,** 1172–1175 (2006).
10. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA

ligands to bacteriophage T4 DNA polymerase. *Science (80-.).* **249,** 505–510 (1990).

11. Ellington, A. & Szostak, J. In vitro selection of RNA molecules that bind specific ligands. *Nature* (1990).

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

13. Aminova, O. & Disney, M. D. in **669,** 209–224 (2010).

14. Paegel, B. M. & Joyce, G. F. Microfluidic Compartmentalized Directed Evolution. *Chem. Biol.* **17,** 717–724 (2010).

15. Huang, H. *et al.* A G-quadruplex–containing RNA activates fluorescence in a GFP-like fluorophore. *Nat. Chem. Biol.* **10,** 686–691 (2014).

16. Filonov, G. S., Moon, J. D., Svensen, N. & Jaffrey, S. R. Broccoli: Rapid selection of an RNA mimic of green fluorescent protein by fluorescence-based selection and directed evolution. *J. Am. Chem. Soc.* **136,** 16299–16308 (2014).

17. Dolgosheina, E. V *et al.* RNA Mango Aptamer-Fluorophore: A Bright, High-Affinity Complex for RNA Labeling and Tracking. *ACS Chem. Biol.* **9,** 2412–2420 (2014).

18. Ishihama, A., Shimada, T. & Yamazaki, Y. Transcription profile of Escherichia coli: Genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Res.* **44,** 2058–2074 (2016).

19. Watrin, M., Von Pelchrzim, F., Dausse, E., Schroeder, R. & Toulmé, J. J. In vitro selection of RNA aptamers derived from a genomic human library against the TAR RNA element of HIV-1. *Biochemistry* **48,** 6278–6284 (2009).

20. 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. *Chem. Biol.* **19,** 1247–54 (2012).

21. Zimmermann, B., Bilusic, I., Lorenz, C. & Schroeder, R. Genomic SELEX: A discovery tool for genomic aptamers. *Methods* **52,** 125–132 (2010).

22. Famulok, M. Oligonucleotide aptamers that recognize small molecules. *Curr. Opin. Struct. Biol.* **9,** 324–329 (1999).

23. Jayasena, S. D. Aptamers: An emerging class of molecules that rival antibodies in diagnostics. *Clin. Chem.* **45,** 1628–1650 (1999).

24. Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C. & Breaker, R. R. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell* **113,** 577–586 (2003).

25. Welz, R. & Breaker, R. R. Ligand binding and gene control characteristics of tandem riboswitches in Bacillus anthracis. *Rna* **13,** 573–82 (2007).

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

27. Martini, L., Ellington, A. D. & Mansy, S. S. An in vitro selection for small molecule induced switching RNA molecules. *Methods* **106,** 51–57 (2016).

28. Curtis, E. a & Liu, D. R. Discovery of widespread GTP-binding motifs in genomic DNA and RNA. *Chem. Biol.* **20,** 521–32 (2013).

29. Sassanfar, M. & Szostak, J. An RNA motif that binds ATP. *Nature* (1993).

30. Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31,** 147–57 (1982).

31. 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).

32. Forster, A. C. & Symons, R. H. Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell* **49,** 211–220 (1987).

33. Winkler, W. C., Nahvi, A., Roth, A., Collins, J. a & Breaker, R. R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428,** 281–6 (2004).

34. Webb, C., Riccitelli, N., Ruminski, D. & Lupták, A. Widespread occurrence of self-cleaving ribozymes. *Science (80-.).* 5 (2009).

35. Unrau, P. J. & Bartel, D. P. RNA-catalysed nucleotide synthesis. *Nature* **395,** 260–263 (1998).

36. Lohse, P. A. & Szostak, J. W. Ribozyme-catalysed amino-acid transfer reactions. *Nature* **381,** 442–4 (1996).

37. Jadhav, V. R. & Yarus, M. Acyl-CoAs from coenzyme ribozymes. *Biochemistry* **41,** 723–729 (2002).

38. Sengle, G., Eisenführ, A., Arora, P. S., Nowick, J. S. & Famulok, M. Novel RNA catalysts for the Michael reaction. *Chem. Biol.* **8,** 459–473 (2001).

39. Agresti, J. J., Kelly, B. T., Jäschke, A. & Griffiths, A. D. Selection of ribozymes that catalyse multiple-turnover Diels-Alder cycloadditions by using in vitro compartmentalization. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 16170–5 (2005).

40. Peattie, D. A. & Gilbert, W. Chemical probes for higher-order structure in RNA. *Proc. Natl. Acad. Sci.* **77,** 4679–4682 (1980).

41. Ehresmann, C. *et al.* Probing the structure of RNAs in solution. *Nucleic Acids Res.* **15,** 9109–9128 (1987).

42. Tullius, T. D. & Greenbaum, J. A. Mapping nucleic acid structure by hydroxyl radical cleavage. *Curr. Opin. Chem. Biol.* **9,** 127–134 (2005).

43. Ding, F., Lavender, C. A., Weeks, K. M. & Dokholyan, N. V. Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat Methods* **9,** 603–608 (2012).

44. Costa, M. & Monachello, D. *RNA Folding*. **1086,** (Humana Press, 2014).

45. Regulski, E. E. & Breaker, R. R. In-line probing analysis of riboswitches. *Methods Mol. Biol.* **419,** 53–67 (2008).

46. 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). *J. Am. Chem. Soc.* **127,** 4223–31 (2005).

47. Soukup, G. & Breaker, R. Relationship between internucleotide linkage geometry and the stability of RNA. *Rna* 1308–1325 (1999).

48. Weeks, K. M. Advances in RNA secondary and tertiary structure analysis by chemical probing. *Curr. Opin. Struct. Biol.* **20,** 295–304 (2010).

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

50. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci.* **106,** 97–102 (2009).

51. Turner, R., Shefer, K. & Ares, M. Safer one-pot synthesis of the 'SHAPE' reagent 1-methyl-7-nitroisatoic anhydride (1m7). *RNA* **19,** 1857–1863 (2013).

52. Dolan, G. F., Akoopie, A. & Müller, U. F. A faster triphosphorylation ribozyme. *PLoS One* **10,** 1–12 (2015).

53. Abdelsayed, M. M. *et al.* Multiplex Aptamer Discovery through Apta-Seq and Its Application to ATP Aptamers Derived from Human-Genomic SELEX. *ACS Chem. Biol.* acschembio.7b00001 (2017). doi:10.1021/acschembio.7b00001

54. Siegfried, N. a, Busan, S., Rice, G. M., Nelson, J. a E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11,** 1–14 (2014).

55. Loughrey, D., Watters, K. E., Settle, A. H. & Lucks, J. B. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res.* **42,** e165–e165 (2014).

56. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11,** 129 (2010).

57. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6,** 26 (2011).

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

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

60. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

61. Bentley, D. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

62. Tome, J. M. *et al.* Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nat. Methods* **11,** 683–8 (2014).

63. Jalali-yazdi, F., Lai, L. H., Takahashi, T. T. & Roberts, R. W. Affinity Measurements Hot

Paper High-Throughput Measurement of Binding Kinetics by mRNA Display and Next-Generation Sequencing Communications Angewandte. **90089,** 4007–4010 (2016).

64. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–1123 (2009).

65. Liu, F., Somarowthu, S. & Pyle, A. M. Visualizing the secondary and tertiary architectural domains of lncRNA RepA. *Nat. Chem. Biol.* **13,** 282–289 (2017).

66. Ruscito, A. & DeRosa, M. C. Small-Molecule Binding Aptamers: Selection Strategies, Characterization, and Applications. *Front. Chem.* **4,** 1–14 (2016).

67. Wittmann, A. & Suess, B. Engineered riboswitches: Expanding researchers' toolbox with synthetic RNA regulators. *FEBS Lett.* **586,** 2076–2083 (2012).

68. Francis, S. H., Busch, J. L., Corbin, J. D. & Sibley, D. cGMP-dependent protein kinases and cGMP phosphodiesterases in nitric oxide and cGMP action. *Pharmacol. Rev.* **62,** 525–63 (2010).

69. Hanafy, K. A., Krumenacker, J. S. & Murad, F. NO, nitrotyrosine, and cyclic GMP in signal transduction. *Med. Sci. Monit.* **7,** 801–819 (2001).

70. Murad, F. Cyclic guanosine monophosphate as a mediator of vasodilation. *J. Clin. Invest.* **78,** 1–5 (1986).

71. Lucas, K. A. *et al.* Guanylyl cyclases and signaling by cyclic GMP. *Pharmacol. Rev.* **52,** 375–414 (2000).

72. Sprenger, J. U. & Nikolaev, V. O. Biophysical Techniques for Detection of cAMP and cGMP in Living Cells. *Int. J. Mol. Sci.* **14,** 8025–8046 (2013).

73. Lolley, R. N. & Lee, R. H. Cyclic GMP and photoreceptor function. *FASEB J.* **4,** 3001–8

(1990).

74. White, R. E., Kryman, J. P., El-Mowafy, a M., Han, G. & Carrier, G. O. cAMP-Dependent Vasodilators Cross-Activate the cGMP-Dependent Protein Kinase to Stimulate BKCa Channel Activity in Coronary Artery Smooth Muscle Cells. *Circ. Res.* **86,** 897–905 (2000).

75. Tamayo, R., Pratt, J. T. & Camilli, A. Roles of Cyclic Diguanylate in the Regulation of Bacterial Pathogenesis. *Annu. Rev. Microbiol.* **61,** 131–148 (2007).

76. Gomelsky, M. cAMP, c-di-GMP, c-di-AMP and now cGMP: bacteria use them all! *Mol. Microbiol.* **79,** 562–565 (2011).

77. Sudarsan, N. *et al.* Riboswitches in Eubacteria Sense the Second Messenger Cyclic Di-GMP. *Science (80-.).* **321,** 411–413 (2008).

78. Lee, E. R., Baker, J. L., Weinberg, Z., Sudarsan, N. & Breaker, R. R. An Allosteric Self-Splicing Ribozyme Triggered by a Bacterial Second Messenger. *Science (80-.).* **329,** 845–848 (2010).

79. Batey, R. T., Gilbert, S. D. & Montange, R. K. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature* **432,** 411–415 (2004).

80. Lorenz, C. *et al.* Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts. *Nucleic Acids Res.* **38,** 3794–3808 (2010).

81. Lambert, N. *et al.* RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Mol. Cell* **54,** 887–900 (2014).

82. Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29,** 659–664 (2011).

83. Pobanz, K. & Lupták, A. Improving the odds: Influence of starting pools on in vitro

selection outcomes. *Methods* **106,** 14–20 (2016).

84. 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,** 1788–92 (2006).

85. Margarucci, L. *et al.* Collagen stimulation of platelets induces a rapid spatial response of cAMP and cGMP signaling scaffolds. *Mol. Biosyst.* **7,** 2311 (2011).

86. Burhenne, H. & Kaever, V. *Cyclic Nucleotide Signaling in Plants*. **1016,** (Humana Press, 2013).

87. Campbell, J. C. *et al.* Structural Basis of Cyclic Nucleotide Selectivity in cGMP-dependent Protein Kinase II. *J. Biol. Chem.* **291,** 5623–5633 (2016).

88. Gotrik, M. R., Feagin, T. A., Csordas, A. T., Nakamoto, M. A. & Soh, H. T. Advancements in Aptamer Discovery Technologies. *Acc. Chem. Res.* **49,** 1903–1910 (2016).

89. Blind, M. & Blank, M. Aptamer Selection Technology and Recent Advances. *Mol. Ther. - Nucleic Acids* **4,** e223 (2015).

90. Cox, J. C. *et al.* Automated selection of aptamers against protein targets translated in vitro: from gene to aptamer. *Nucleic Acids Res.* **30,** e108 (2002).

91. Schütze, T. *et al.* Probing the SELEX process with next-generation sequencing. *PLoS One* **6,** 1–10 (2011).

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

93. Underwood, J. G. *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods* **7,** 995–1001 (2010).

94. Talkish, J., May, G., Lin, Y., Woolford, J. L. & McManus, C. J. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20,** 713–720 (2014).

95. Opoku-Temeng, C., Zhou, J., Zheng, Y., Su, J. & Sintim, H. O. Cyclic dinucleotide (c-di-GMP, c-di-AMP, and cGAMP) signalings have come of age to be inhibited by small molecules. *Chem. Commun.* **52,** 9327–9342 (2016).

96. Lori, C. *et al.* Cyclic di-GMP acts as a cell cycle oscillator to drive chromosome replication. *Nature* **523,** 236–239 (2015).

97. Sintchak, M. D. *et al.* Structure and mechanism of inosine monophosphate dehydrogenase in complex with the immunosuppressant mycophenolic acid. *Cell* **85,** 921–930 (1996).

98. Roembke, B. T. *et al.* A cyclic dinucleotide containing 2-aminopurine is a general fluorescent sensor for c-di-GMP and 3′,3′-cGAMP. *Mol. Biosyst.* **10,** 1568 (2014).

99. Ryder, S. P. & Strobel, S. a. Nucleotide analog interference mapping. *Methods* **18,** 38–50 (1999).

100. Cote, R. H. & Brunnock, M. A. Intracellular cGMP concentration in rod photoreceptors is regulated by binding to high and moderate affinity cGMP binding sites. *J. Biol. Chem.* **268,** 17190–8 (1993).

101. Traut, T. W. Physiological concentrations of purines and pyrimidines. *Mol. Cell. Biochem.* **140,** 1–22 (1994).

102. 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. *J. Mol. Biol.* **255,** 589–603 (1996).

103. Lindgreen, S., Gardner, P. P. & Krogh, A. Measuring covariation in RNA alignments: Physical realism improves information measures. *Bioinformatics* **22,** 2988–2995 (2006).

104. Doudna, J. A. & Cech, T. R. The chemical repertoire of natural ribozymes. *Nature* **418,** 222–228 (2002).

105. Shu, D. & Guo, P. A viral RNA that binds ATP and contains a motif similar to an ATP-binding aptamer from SELEX. *J. Biol. Chem.* **278,** 7119–7125 (2003).

106. Saran, D., Frank, J. & Burke, D. H. The tyranny of adenosine recognition among RNA aptamers to coenzyme A. *BMC Evol. Biol.* **3,** 26 (2003).

107. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* **2,** 3 (2012).

108. Driscoll, H. E., Vincent, J. J., English, E. L. & Dolci, E. D. Metagenomic investigation of the microbial diversity in a chrysotile asbestos mine pit pond, Lowell, Vermont, USA. *Genomics Data* **10,** 158–164 (2016).

109. Miller, R. R. *et al.* Metagenomic Investigation of Plasma in Individuals with ME/CFS Highlights the Importance of Technical Controls to Elucidate Contamination and Batch Effects. *PLoS One* **11,** e0165691 (2016).

110. Beavo, J. a & Brunton, L. L. Cyclic nucleotide research — still expanding after half a century. *Nat. Rev. Mol. Cell Biol.* **3,** 710–718 (2002).

111. Montminy, M. R., Sevarino, K. A., Wagner, J. A., Mandel, G. & Goodman, R. H. Identification of a cyclic-AMP-responsive element within the rat somatostatin gene. *Proc. Natl. Acad. Sci.* **83,** 6682–6686 (1986).

112. Sureka, K. *et al.* The Cyclic Dinucleotide c-di-AMP Is an Allosteric Regulator of Metabolic

Enzyme Function. *Cell* **158,** 1389–1401 (2014).

113. Daniel, P. B., Walker, W. H. & Habener, J. F. Cyclic AMP signaling and gene regulation. *Annu. Rev. Nutr.* **18,** 353–83 (1998).

114. Siess, W. Cross-talk of cGMP- and cAMP-signaling pathways in human platelets. *Blood* **101,** 4230–4230 (2003).

115. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science (80-.).* **316,** 1497–1502 (2007).

116. Darnell, R. B. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip. Rev. - RNA* **1,** 266–286 (2010).

117. Konig, J. *et al.* iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J. Vis. Exp.* 1–7 (2011). doi:10.3791/2638

118. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. in *Current Protocols in Molecular Biology* 21.29.1-21.29.9 (John Wiley & Sons, Inc., 2015). doi:10.1002/0471142727.mb2129s109

119. Auerbach, R. K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci.* **106,** 14926–14931 (2009).

120. Smola, M. J., Calabrese, J. M. & Weeks, K. M. Detection of RNA–Protein Interactions in Living Cells with SHAPE. *Biochemistry* **54,** 6867–6875 (2015).

121. Flynn, R. A. *et al.* Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. *Nat. Protoc.* **11,** 273–290 (2016).

122. Spitale, R., Crisalli, P. & Flynn, R. RNA SHAPE analysis in living cells. *Nat. Chem. …* **9,** 18–20 (2013).

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

124. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38,** 1767–1771 (2009).

125. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

126. Smeds, L. & Künstner, A. ConDeTri - A Content Dependent Read Trimmer for Illumina Data. *PLoS One* **6,** e26314 (2011).

127. Chang, J. T. in *Transcriptomics and Gene Regulation* 99–113 (2016). doi:10.1007/978-94-017-7450-5_4

128. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. 2C- Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

129. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: Short oligonucleotide alignment program. *Bioinformatics* **24,** 713–714 (2008).

130. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

131. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

132. Freese, N. H., Norris, D. C. & Loraine, A. E. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* **32,** 2089–2095 (2016).

133. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–6 (2011).

134. Giardine, B. Galaxy: a platform for interactive large-scale genome analysis. *Genom Res* **15,** 1451–1455 (2005).

135. Tang, Y. *et al.* StructureFold: Genome-wide RNA secondary structure mapping and reconstruction in vivo. *Bioinformatics* **31,** 2668–2675 (2015).

136. Bellaousov, S., Reuter, J. S., Seetin, M. G. & Mathews, D. H. RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res.* **41,** 471–474 (2013).

137. Dupont, D. M., Larsen, N., Jensen, J. K., Andreasen, P. A. & Kjems, J. Characterisation of aptamer–target interactions by branched selection and high-throughput sequencing of SELEX pools. *Nucleic Acids Res.* **43,** gkv700 (2015).

138. Robertson, M. P. & Joyce, G. F. The Origins of the RNA World. *Cold Spring Harb. Perspect. Biol.* **4,** a003608–a003608 (2012).

139. Scott, W. G., Martick, M. & Chi, Y. I. Structure and function of regulatory RNA elements: Ribozymes that regulate gene expression. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1789,** 634–641 (2009).

140. Fedor, M. J. & Williamson, J. R. The catalytic diversity of RNAs. *Nat. Rev. Mol. Cell Biol.* **6,** 399–412 (2005).

141. Ekland, E. & Bartel, D. RNA-catalysed RNA polymerization using nucleoside triphosphates. *Nature* (1996).

142. Technikova-Dobrova, Z. *et al.* Selection of ribozymes that catalyse multiple-turnover Diels-Alder cycloadditions by using in vitro compartmentalization. *Chem. Biol.* **15,** 4610–

38 (2013).

143. Wiegand, T. W., Janssen, R. C. & Eaton, B. E. Selection of RNA amide synthases. *Chem. Biol.* **4,** 675–683 (1997).

144. Robertson, M. P. & Joyce, G. F. Highly efficient self-replicating RNA enzymes. *Chem. Biol.* **21,** 238–45 (2014).

145. Rostovtsev, V. V, Green, L. G., Fokin, V. V & Sharpless, K. B. A stepwise huisgen cycloaddition process: copper(I)-catalyzed regioselective 'ligation' of azides and terminal alkynes. *Angew. Chem. Int. Ed. Engl.* **41,** 2596–9 (2002).

146. Kolb, H. C., Finn, M. G. & Sharpless, K. B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem. Int. Ed. Engl.* **40,** 2004–2021 (2001).

147. Gaetke, L. Copper toxicity, oxidative stress, and antioxidant nutrients. *Toxicology* **189,** 147–163 (2003).

148. Saxon, E. Cell Surface Engineering by a Modified Staudinger Reaction. *Science (80-.).* **287,** 2007–2010 (2000).

149. Agard, N. J., Prescher, J. a & Bertozzi, C. R. A strain-promoted [3 + 2] azide-alkyne cycloaddition for covalent modification of biomolecules in living systems. *J. Am. Chem. Soc.* **126,** 15046–7 (2004).

150. Baskin, J. M. *et al.* Copper-free click chemistry for dynamic in vivo imaging. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 16793–7 (2007).

151. Agard, N., Baskin, J. & Prescher, J. A comparative study of bioorthogonal reactions with azides. *ACS Chem. …* **1,** (2006).

152. Nessen, M. A. M. *et al.* Selective enrichment of azide-containing peptides from complex

mixtures. *J. proteome …* 3702–3711 (2009).

153. Link, a J. *et al.* Discovery of aminoacyl-tRNA synthetase activity through cell-surface display of noncanonical amino acids. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 10180–5 (2006).

154. Weil, T. T., Parton, R. M. & Davis, I. Making the message clear: Visualizing mRNA localization. *Trends Cell Biol.* **20,** 380–390 (2010).

155. Larsson, C., Grundberg, I., Söderberg, O. & Nilsson, M. In situ detection and genotyping of individual mRNA molecules. *Nat. Methods* **7,** 395–397 (2010).

156. Mannack, L. V. J. C. J. C., Eising, S. & Rentmeister, A. Current techniques for visualizing RNA in cells. *F1000Research* **5,** 775 (2016).

157. Kellenberger, C. A., Wilson, S. C., Sales-Lee, J. & Hammond, M. C. RNA-based fluorescent biosensors for live cell imaging of second messengers cyclic di-GMP and cyclic AMP-GMP. *J. Am. Chem. Soc.* **135,** 4906–4909 (2013).

158. Jao, C. Y. & Salic, A. Exploring RNA transcription and turnover in vivo by using click chemistry. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 15779–84 (2008).

159. Sawant, A. A. *et al.* A versatile toolbox for posttranscriptional chemical labeling and imaging of RNA. *Nucleic Acids Res.* **44,** e16 (2016).

160. Monici, M. in 227–256 (2005). doi:10.1016/S1387-2656(05)11007-2

161. Sivakumar, K. *et al.* A Fluorogenic 1 , 3-Dipolar Cycloaddition Reaction of 3-Azidocoumarins and. *Org. Lett.* **6,** 4603–4606 (2010).

162. Li, K., Lee, L., Lu, X. & Wang, Q. Fluorogenic 'click' reaction for labeling and detection of DNA in proliferating cells. *Biotechniques* **49,** 525–527 (2010).

163. Levy, M., Griswold, K. & Ellington, A. Direct selection of trans-acting ligase ribozymes by

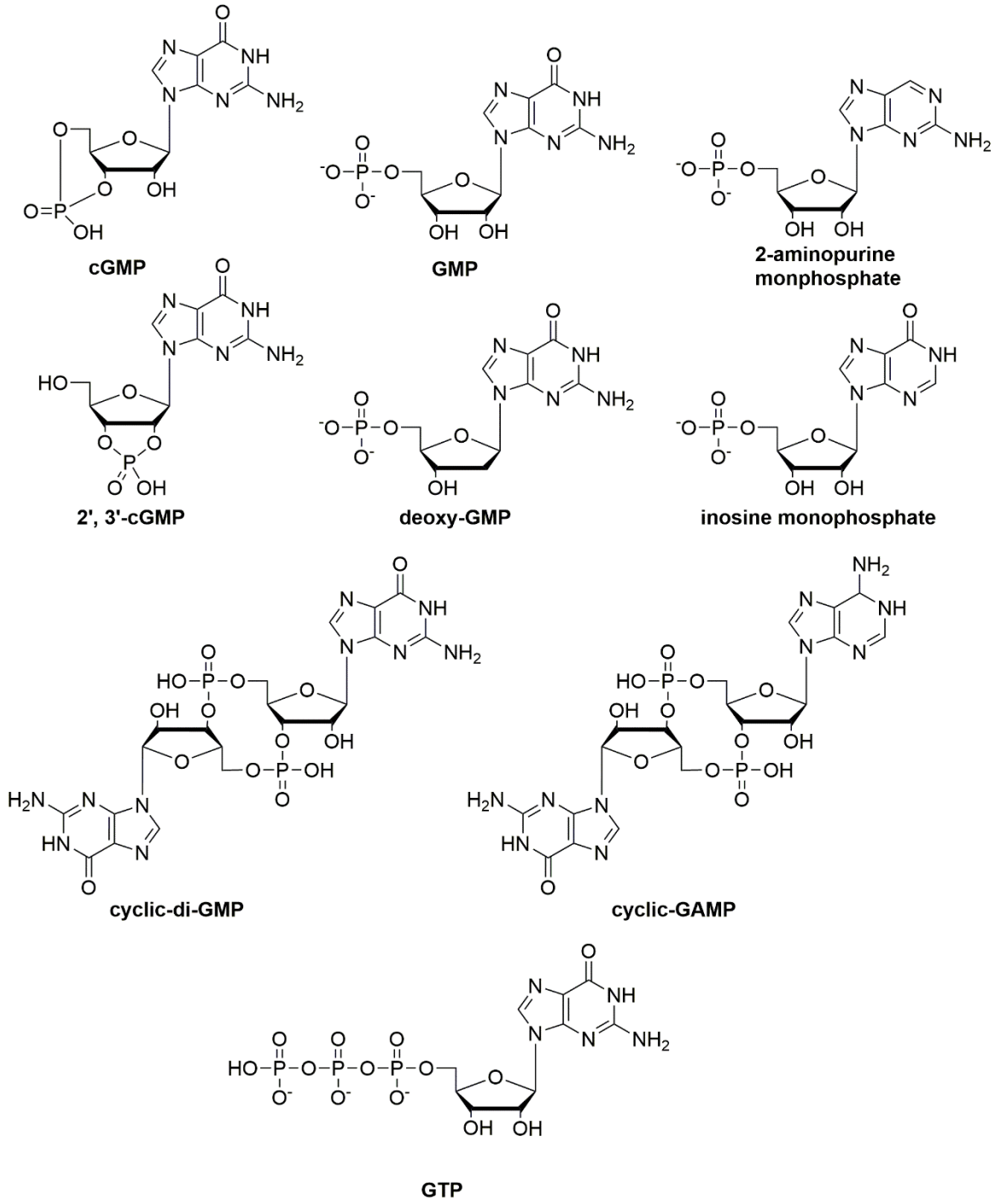
in vitro compartmentalization. *Rna* 1555–1562 (2005). doi:10.1261/rna.2121705.attempt

164. Schramm, V. L. Transition states, analogues, and drug development. *ACS Chemical Biology* **8,** 71–81 (2013).

165. Pagano, P. J. & Chong, K. T. In-Vitro Inhibition of Human-Immunodeficiency-Virus Type-1 by a Combination of Delavirdine (U-90152) with Protease Inhibitor U-75875 or Interferon-Alpha. *J. Infect. Dis.* **171,** 61–67 (1995).

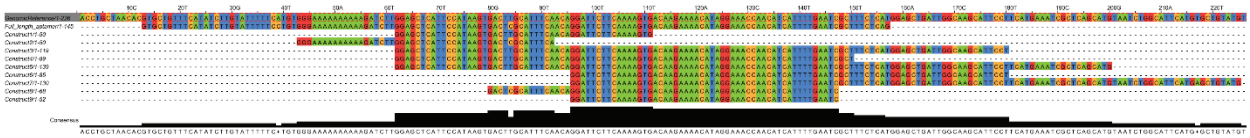
APPENDIX

Chapter 2

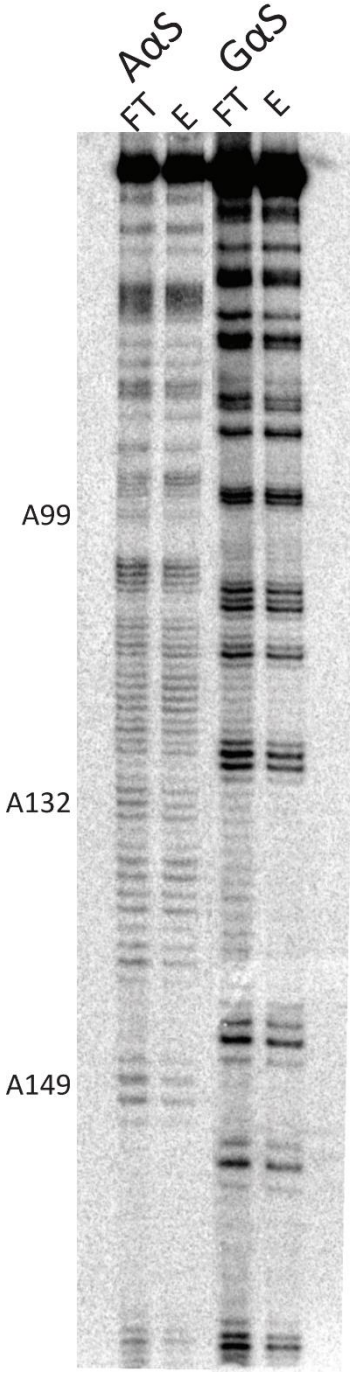


Appendix Figure 2-1: Guanosine derivatives used for structure probing of cGMP aptamers

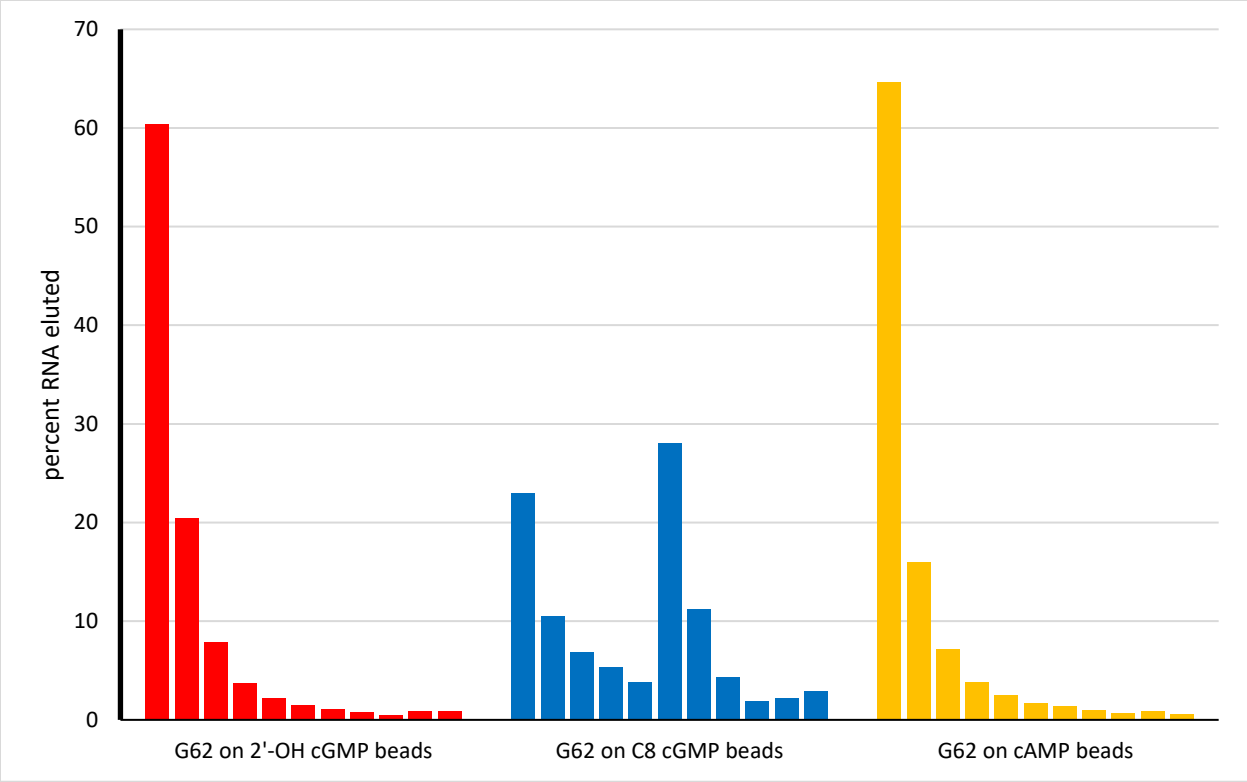
Appendix Table 2-1: Notable sequences from *in vitro* selection of cGMP aptamers		
Aptamer	**Location**	**Abundance (%)**
G2	Chr1:114474090-114474160	0.002
G3	Chr1:121484860-121484970	0.009
G12	Chr2:201982450-201982490	0.014
G13	Chr2:242287530-242287700	0.002
G16	Chr3:143261380-143261470	0.002
G17	Chr4:37279990-37280040	0.002
G18	Chr4:38910640-38910700	0.003
G20	Chr4:77529810-77529890	0.005
G21	Chr5:13780910-13780960	0.005
G22	Chr5:66308640-66308695	0.004
G23	Chr5:92821380-92821430	0.009
G24	Chr5:145442640-145442690	0.002
G27	Chr6:80199790-80199875	0.013
G28	Chr7:151946040-151946240	0.002
G31	Chr8:69075650-69075780	0.005
G32	Chr8:126033340-126033395	0.002
G34	Chr8:133250170-133250230	0.002
G35	Chr9:11283800-11283960	0.002
G36	Chr9:43314700-43314805	0.005
G37	Chr9:80768010-80768150	0.005
G38	Chr9:113703430-113703480	0.002
G39	Chr9:127939965-127940035	0.002
G40	Chr9:140567040-140567200	0.484
G50	chr14:82,619,315-82,619,444	0.069
G51	Chr15:55005050-55005180	0.223
G53	Chr15:60034970-60035120	0.006
G54	Chr15:72945430-72945510	0.002
G55	Chr15:74376410-74376485	0.013
G56	Chr15:75549290-75549370	0.014
G58	Chr16:48078428-48078548	0.002
G59	Chr16:87945550-87945610	0.058
G60	Chr17:71458065-71458115	0.002
G61	CHr19:27738630-27738710	0.003
G62	Chr20:5659560-5659770	92.38
G65	Chr21:11039790-11039990	0.027
G66	ChrX:9251420-9251585	3.493



Appendix Figure 2-2: Alignment of genomic constructs of G62 aptamer.



Appendix Figure 2-3: NAIM of G62 full-length aptamer



Appendix Figure 2-4: G62 aptamer's specificity on different cyclic-nucleotide agaroses.

Chapter 3

Appendex Table 3-1: Genomes included in the metagenomic library	
Organism	**Genome size**
A. gambiae	2.64E+08
A. thaliana	1.20E+08
A. nidulans	3.00E+07
B. distachyon	2.70E+08
C. elegans	7.80E+07
C. reinhardtii	1.05E+08
D. rerio	1.41E+09
D. discoideum	3.41E+07
D. melanogaster	1.49E+08
H. sapiens	3.09E+09
M. truncatula	3.10E+08
M. musculus	2.77E+09
P. patens	4.54E+08
S. cerevisiae	1.21E+07
S. pombe	1.25E+07
S. purpuratus	9.30E+08
T. gondii	6.48E+07
T. brucei	2.41E+07
X. laevis	4.25E+07
X. tropicalis	1.36E+09
Z. mays	2.00E+09

Appendix Table 3-2: human cGMP aptamers discovered in the metagenomics selection		
Aptamer	**Location**	**Abundance (%)**
MG1	chr1:29,456,751-29,456,860	0.00305
MG2	chr1:80,518,287-80,518,395	0.00946
MG3	chr1:153,495,357-153,495,482	0.009937
MG4	chr1:192,467,265-192,467,381	0.00312
MG5	chr1:231,424,353-231,424,478	0.002305
MG6	chr2:6,358,347-6,358,495	0.012341
MG7	chr2:26,139,649-26,139,749	0.008237
MG8	chr2:57,562,821-57,562,928	0.002502
MG9	chr2:83,976,538-83,976,627	0.00603
MG10	chr2:134,866,036-134,866,132	0.00745
MG11	chr2:149,435,147-149,435,250	0.00343
MG12	chr2:149,435,147-149,435,250	0.003514
MG13	chr3:41,884,830-41,884,933	0.002319
MG14	chr3:140,371,541-140,371,637	0.010486
MG15	chr3:162,359,014-162,359,161	0.135906
MG16	chr3:162,561,381-162,561,490	0.006705
MG17	chr4:25,240,046-25,240,150	0.003612
MG18	chr4:107,445,639-107,445,734	0.001954
MG19	chr5:29,295,612-29,295,704	0.010851
MG20	chr6:116,177,684-116,177,792	0.008026
MG21	chr7:52,369,379-52,369,503	0.128554
MG22	chr7:105,595,496-105,595,588	0.00357
MG23	chr8:47,244,617-47,244,719	0.006845
MG24	chr8:53,881,622-53,881,719	0.046454
MG25	chr8:83,794,018-83,794,127	0.007548
MG26	chr8:138,766,498-138,766,600	0.024584
MG27	chr9:95,787,045-95,787,200	0.142484
MG28	chr10:124,771,952-124,772,044	0.014098
MG29	chr11:113,519,044-113,519,138	0.041971
MG30	chr12:43,003,868-43,003,963	0.007042
MG31	chr13:98,198,912-98,199,007	0.003669
MG32	chr15:41,324,092-41,324,496	0.025019
MG33	chr15:95,463,392-95,463,540	0.38236
MG34	chr19:9,651,413-9,651,499	0.004498
MG35	chr20:5,659,560-5,659,770	52.99039

Appendix Table 3-3: human cAMP aptamers discovered in the metagenomics selection		
Aptamer	**Location**	**Abudance (%)**
MA1	chr1:121,483,910-121,485,578	0.140997074
MA2	chr1:94,970,551-94,970,638	0.011455474
MA3	chr1:56,396,001-56,396,089	0.012919707
MA4	chr2:6,358,378-6,358,520	0.027734305
MA5	chr2:76,084,621-76,084,709	0.011369343
MA6	chr2:223,680,353-223,680,451	0.05477956
MA7	chr2:234,633,820-234,633,982	0.028940145
MA8	chr2:236,174,516-236,174,658	0.095175178
MA9	chr2:237,615,602-237,615,780	0.124201454
MA10	chr3:18,833,991-18,834,078	0.005254014
MA11	chr3:196,625,333-196,626,099	0.004737226
MA12	chr4:1,066,137-1,066,260	0.074503646
MA13	chr4:88,047,782-88,047,958	0.016106569
MA14	chr5:121,627,025-121,627,113	0.003875912
MA15	chr6:58,776,039-58,780,313	0.045563502
MA16	chr6:136,487,760-136,487,978	0.004306569
MA17	chr8:27,214,342-27,214,463	0.062359121
MA18	chr8:34,049,523-34,049,611	0.01369489
MA19	chr8:69,075,627-69,075,715	0.021016057
MA20	chr9:95,787,037-95,787,143	0.069594158
MA21	chr11:126,320,666-126,320,796	0.173899262
MA22	chr13:112,318,896-112,319,159	0.103616054
MA23	chr14:82,619,371-82,619,543	0.02988759
MA24	chr15:55,004,996-55,005,117	0.061067151
MA25	chr15:95,463,378-95,463,498	0.339788306
MA26	chr17:78,607,790-78,608,802	0.010249635
MA35	chr20:5659609-5659813	78.51185642
MA36	chr21:30,406,180-30,406,268	0.003875912
MA37	chr21:46,322,592-46,322,696	0.001981022
MA38	chr22:38,521,107-38,521,195	0.002670073

Chapter 4

Cutadapt installation

sudo pip install --upgrade cutadapt

Cutadapt was coded in python. Package 'build-essential' and 'python-dev' might be needed on the operating system. The --upgrade option was only necessary if a previous version of cutadapt was already installed.

Bowtie2 installation

Bowtie2 can be downloaded from its website on Sourceforge.

```
make
sudo make install
```

Samtools and bcftools were installed for formatting of .sam files, as the extension resulted in very large file size. These programs were provided together as a package.

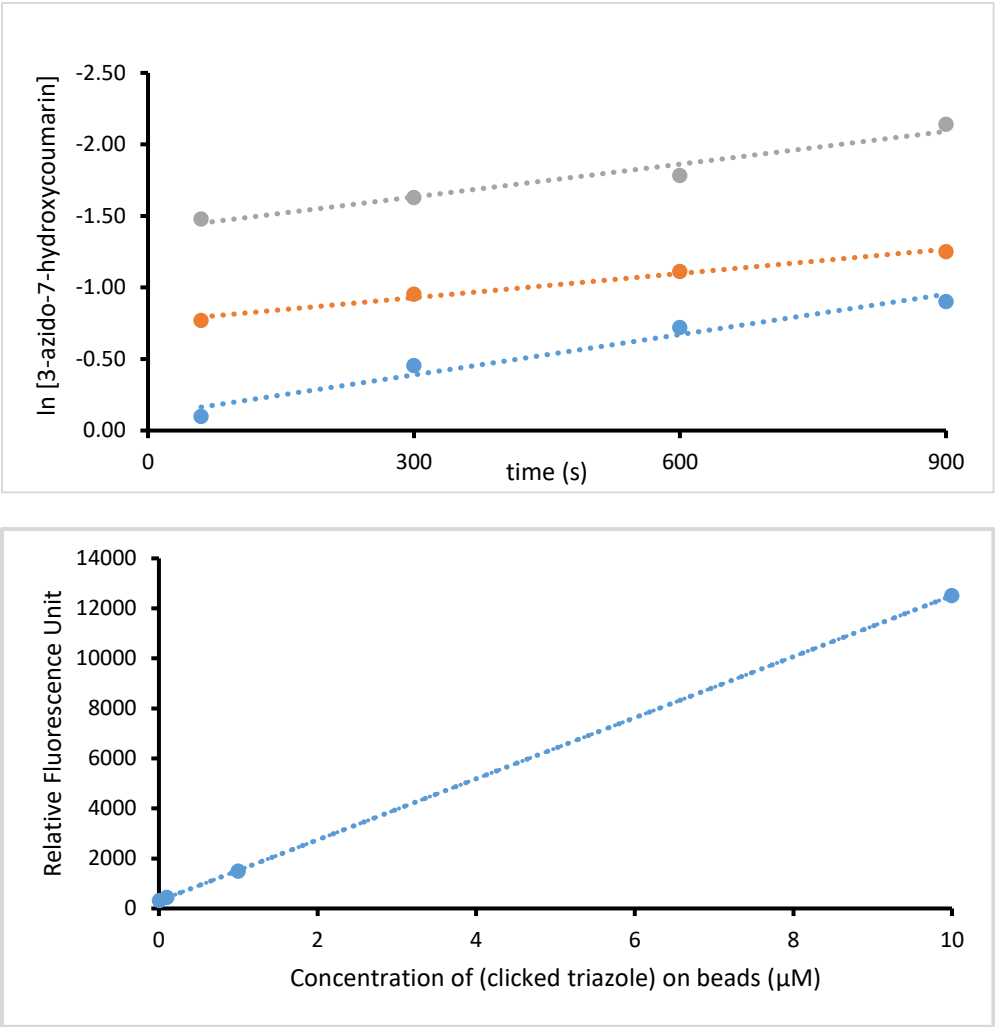
```
git clone git://github.com/samtools/htslib.git
git clone git://github.com/samtools/bcftools.git
cd bcftools; make
```

FASTX Installation

Short sequences in a fastq file were collapsed using FASTX.

```
wget http://cancan.cshl.edu/labmembers/gordon/files/fastx_toolkit-0.0.12.tar.bz2
tar -xMf fastx_toolkit-0.0.12.tar.bz2
cd fastx_toolkit-0.0.12
./configure
make
sudo make install
```

Chapter 5



Appendix Figure 5-1: *(top)* Pseudo-first order measurement of non-catalyzed SPAAC reaction. *(bottom)* Standard curve generated from fluorescent signal of known 1, 2, 3-clicked triazole concentrations.