UNIVERSITY OF CALIFORNIA,
IRVINE

Detection of naturally occurring RNA aptamers using a metagenomic DNA library

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Julio Alexander Polanco

Dissertation Committee:
Professor Andrej Lupták, Chair
Professor Melanie Cocco
Professor Donald Senear

2018

DEDICATION

To

my parents, Julio and Nora Polanco, my siblings Karen Armstrong and Joshua Polanco,

and my niece Bella Hogan

in recognition of their support and unconditional love,

and to my best friends Alysia Ahmed, Richard Daily and Eddie Ng

for their unceasing support and wisdom.

Unless we take one step forward, we cannot take the next, much less a thousand.

Daisaku Ikeda
Courage

TABLE OF CONTENTS

Page

LIST OF FIGURES iv

LIST OF TABLES vi

ACKNOWLEDGMENTS vii

CURRICULUM VITAE viii

ABSTRACT OF THE DISSERTATION xii

ABBREVIATIONS xiv

CHAPTER 1: Naturally occurring RNA aptamers 1

CHAPTER 2: Design of a metagenomic DNA library 9

CHAPTER 3: *In vitro* selection using a metagenomics DNA library 22

CHAPTER 4: Deep sequencing and structure probing of enriched DNA libraries 35

CHAPTER 5: *RAB3C* and *FGD3* aptamer expression 84

CHAPTER 6: Materials and methods 100

REFERENCES 125

LIST OF FIGURES

Page

Figure 1.1 Secondary and tertiary structures of the adenosine aptamer 3

Figure 1.2 Riboswitch-mediated gene expression in prokaryotes 6

Figure 2.1 Genome sizes 12

Figure 2.2 Gel analysis of high molecular weight genomic DNA 13

Figure 2.3 PCR amplification of the metagenomic DNA library 15

Figure 2.4 Size-selection of the metagenomic DNA library 15

Figure 3.1 *In vitro* selection scheme 23

Figure 3.2 *In vitro* selection elution profiles 26

Figure 3.3 Overview of ATP clones derived from recombinant DNA 28

Figure 3.4 Overview of cAMP clones derived from recombinant DNA 29

Figure 4.1 SHAPE-sequencing flowchart 37

Figure 4.2 Overview of computational pipelines 40

Figure 4.3 Features of ATP clone 336.3 57

Figure 4.4 Features of ATP clone 302.67 60

Figure 4.5 Features of ATP clone 230.5 62

Figure 4.6 SHAPE-profile for *Mus musculus* ERV17B 67

Figure 4.7 Features of cAMP clone 669.83 69

Figure 4.8 Features of cAMP clone 292.17 70

Figure 4.9 Features of cAMP clone 282.33 72

Figure 4.10 Features of cAMP clone 479.33 72

Figure 4.11 Deep sequencing read alignment along *D. discoideum* 5.8S and 26S rRNA reference sequence 75

Figure 4.12 Column binding assay of a rRNA segment amplified from the ATP-enriched RNA library 76

Figure 5.1 Flowchart for total RNA extraction and preparation 87

Figure 5.2 *RAB3C* and *FGD3* primer map and adenosine-dependent expression 88

Figure 5.3 RAB3C aptamer expression analyzed by qPCR 91

Figure 5.4 Effect of purine-synthesis inhibitors; azathioprine (AZA) and methotrexate (MTX) on *RAB3C* expression 93

Figure 5.5 Column binding assay using total RNA isolated from SH-SY5Y cell culture 95

Figure 5.6 Visual phenotype of SH-SY5Y cells by light microscopy 97

LIST OF TABLES

		Page
Table 2.1	ribosomal DNA sequence searches for unique expansion segments	19
Table 4.1	Percent survival of trimmed deep sequencing reads	42
Table 4.2	Enriched sequences from the ATP aptamer *in vitro* selection	46-48
Table 4.3	Enriched sequences from the cAMP aptamer in vitro selection	49-52

ACKNOWLEDGMENTS

I would like to acknowledge the committed support from my thesis advisor, Dr. Andrej Lupták. The objectives and results of this project have certainly challenged our outlook on genomic selections, and for that I appreciate his persistent and determined support throughout the years.

I would like to thank my committee members, Drs. Donald Senear, Melanie Cocco, Thomas Poulos, and Michael Mulligan for their advice and constructive feedback throughout the various stages of the research in the dissertation. Their time and participation were essential for the development and maturity of this research project.

In addition, I would like to extend my appreciation for the help and support I received from all the Lupták lab members especially Drs. Randi Jimenez, Dana Ruminski, and Cassandra Burke for their mentorship and friendship.

I thank Elsevier, Inc. and the American Chemical Society for permission to include copyrighted material in my dissertation.

I also thank the University of California, Irvine for the dissertation fellowship.

Finally, I would like to thank my family members for their unwavering support throughout the years. To my mother who has taught me boundless compassion and courage, and to my father, for his wisdom and encouragement in life.

CURRICULUM VITAE

Julio Alexander Polanco

Education:

2011-2018	University of California, Irvine	Ph.D.	Biological Sciences
2007-2011	University of California, Irvine	B.S	Biological Sciences

Field of study:

RNA biology, gene expression, structural biology, deep sequencing technology, computational biology and data science.

Research Experience:

July 2012 – Mar 2018

University of California, Irvine
Dr. Andrej Lupták, Faculty Advisor

Department of Pharmaceutical Sciences
Department of Molecular Biology and Biochemistry
Mapping and elucidating the role of eukaryotic RNA aptamers using *in vitro* selections and gene expression assays.

Summer 2011

Kings College London, London U.K.
Dr. Maria R. Conte

Randall Division of Cellular & Molecular Biophysics
This summer position focused on structural determination of the La-Related Protein 7 (LaRP 7) using NMR techniques. LaRP 7 is a RNA binding protein known to regulate gene expression in eukaryotes.

Jun 2008 – Jun 2011

University of California, Irvine.
Dr. Thomas Poulos

Department of Molecular Biology and Biochemistry.
Worked on developing experimental approaches to determine the structure and function of a *Leishmania major* peroxidase (LmP). The LmP protein is expressed by a eukaryoticparasite known to cause cutaneous leishmaniasis.

Summer 2009 University of California, San Francisco.
Dr. Paul Ortiz de Montellano

Department of Pharmaceutical Chemistry. Worked on structural characterization and purification of a *Mycobacterium tuberculosis* protein for drug design.

Jan 2008 - Jun 2008 University of California, Irvine.
Dr. Luis Mota-Bravo

Practiced aseptic techniques. Antibiotic resistance assays. Bio-hazardous materials training.

Awards and Honors:

University of California, Irvine:

2017	Robert Warner Award for Outstanding Achievement in Nucleic Acid Biochemistry
2014	Robert Warner Award for Outstanding Achievement in Nucleic Acid Biochemistry
2013	National Academies Ford Foundation Fellowships Program – Honorable Mention
2012	Graduate Minority Biomedical Research Support - Initiative for Maximizing Student Development Fellowship (NIH-MBRS-IMSD) NIH Grant # GM055246
2012	National Science Foundation Graduate Research Fellowship Program - Honorable Mention
2011	Minority Health and Health Disparities International Research Training (MHIRT) scholar. Hosted by Dr. Maria R. Conte, Randall Division of Cellular and Molecular Biophysics, King's College, London, U.K.
2008-2011	Minority Science Programs (MSP) Scholar: Minority Biomedical Research Support (MBRS) Minority Access to Research Careers (MARC) Minority Health and Health Disparities International Research Training (MHIRT)
2010	Excellence in Research: "Structural Studies of the *Leishmania* Ascorbate Peroxidase."
2009	University of California, Irvine: Dean's Honor List: Fall 2008, Winter 2009.

Certifications:

(2017) Introduction to Python Fundamentals. E-course offered by Microsoft.
(2017) Google Developers: Python Introduction. E-course offered by Google.

Research Talks

2017 RNA group: J.A. Polanco, A. Lupták. “Mapping eukaryotic aptamers by *in vitro* selection and next-generation sequencing” (Oral presentation)

2014 Cell Symposia - Regulatory RNAs Berkeley, California: J.A. Polanco, A. Lupták. “Mapping functional eukaryotic aptamers by *in vitro* selection of genomic aptamers.” (Poster presentation)

2012 American Association for the Advancement of Science. Boston, MA: J.A. Polanco, A. Lupták. “Mapping eukaryotic riboswitches by using *in vitro* selection (SELEX) of genomic aptamers.” (Poster presentation)

2010 American Association for the Advancement of Science. Washington DC: J.A. Polanco, V.S. Jasion, T.L. Poulos. “Structural Characterization of *Leishmania Major* Ascorbate Peroxidase.” (Poster presentation)

2010 Sigma Xi Scientific Conference. Raleigh, NC: J.A. Polanco, V.S. Jasion, T.L. Poulos. “Structural Characterization of *Leishmania Major* Ascorbate Peroxidase.” (Poster presentation)

2009 American Association for the Advancement of Science. San Diego, CA: J.A. Polanco, V.S. Jasion, T.L. Poulos. “Structural Studies of *Leishmania* Ascorbate Peroxidase.” (Poster presentation)

Publications:

Abdelsayed MM, Ho B, Vu MM, Polanco JA, Spitale RC, Lupták A. Multiplex Aptamer Discovery through Apta-Seq and its Application to ATP aptamers Derived from Human-Genomic SELEX. *ACS Chemical Biology*. 2017. 12:8 (2149-2156).

Jimenez RM, Polanco JA, Lupták A. Chemistry and Biology of Self-Cleaving Ribozymes. *Trends in Biochemical Sciences*. 2015. 40:11 (648-61).

Ho B, Polanco JA, Jimenez R, Lupták A. Discovering Human RNA Aptamers by Structure-Based Bioinformatics and Genome-Based *In Vitro* Selection. *Methods in Enzymology*. 2014. 549:2 (29-46).

Jasion VS, Polanco JA, Meharenna YT, Li H, Poulos, TL. Crystal Structure of *Leishmania Major* Peroxidase and Characterization of the Compound I Tryptophan Radical. *The Journal of Biological Chemistry*. 2011. 286:28 (24608-24615).

Teaching Experience:

University of California, Irvine:

Departments of Molecular Biology and Biochemistry, and Pharmaceutical Sciences

Teaching Assistant, 2013-2017
Molecular Biology (Bio99)
Experimental Microbiology Laboratory (M118L)
Molecular Biology Laboratory (M116L)
Human Physiology (Phrmsci120)
Medicinal Chemistry (Chem177L)

Broader Impacts:

2016 Southern California Regional Junior Science and Humanities Symposium (JSHS)
Volunteer judge for a panel of distinguished high school students presenting their own research projects for a JSHS scholarship.

2015 Irvine Unified School District (IUSD) science fair
Volunteer judge for science fair projects presented by students in grades 6-12.

2014 Disability Services Center – Notes Provider
Volunteer note provider for a UC Irvine human physiology course offered in the Fall. Approximately 12 hours of community service provided. This volunteer service is ongoing.

2013 Irvine Unified School District (IUSD) science fair
Volunteer judge for science fair projects presented by students in grades 6-12.

2012 Participation in Graduate and GK-12 education program (Outreach, Research Training and Minority Science Programs)
Lecture on the central dogma of biology presented to Santa Ana high school students in Santa Ana, CA.
Discussed career opportunities in the field of science to high school and middle school students in Santa Ana, CA.

ABSTRACT OF THE DISSERTATION

Detection of naturally occurring RNA aptamers using a metagenomic DNA library

By

Julio Alexander Polanco

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2018

Professor Andrej Lupták, Chair

Over the past few decades, the idea of RNA transcripts as passive genetic messengers within biological systems has changed significantly. RNAs, unlike DNA, possess an intrinsic capability to relay genetic information by actively participating in biochemical interactions that influence cellular activity. Riboswitches, or ligand-sensitive genetic elements, are functional RNAs within nascent and messenger RNA transcripts that consist of a structurally conserved aptamer domain and an expression platform capable of regulating gene expression at the level of transcription or translation.

Despite the numerous examples of prokaryotic riboswitches, the thiamine pyrophosphate (TPP) riboswitch is the only known functional riboswitch in eukaryotes. The existence of multiple human adenosine aptamers encoded within the human genome suggest that eukaryotic riboswitch activity may be more prevalent than previously thought. To investigate the diversity of eukaryotic aptamers and elucidate their biological roles, we designed an *in vitro* selection scheme to select for naturally occurring RNA aptamers from a metagenomic DNA library and monitored the expression of the Ras-related protein 3C

(*RAB3C*) adenosine aptamer by reverse transcriptase – quantitative polymerase chain reaction (RT-qPCR).

Libraries enriched for adenosine triphosphate (ATP) and cyclic adenosine monophosphate (cAMP) RNA aptamers were analyzed by deep sequencing and Apta-seq methods. Computational analysis of enriched libraries by deep sequencing showed that a significant fraction of sequences from both ATP and cAMP selections were repetitive elements. Sequence alignments and Apta-seq analysis found that these specific repetitive elements include endogenous retroviral elements and CG-rich repeats, and not simple repeats such as telomeric or ribosomal RNA elements.

Binding activity assays showed that individual sequences isolated from the enriched libraries lost specificity to either ATP or cAMP. Binding activity was recovered upon reintroducing a fraction of the enriched library suggesting that ligand binding may require a binding partner.

RT-qPCR assays revealed changes in 3′ untranslated region (UTR) expression of the *RAB3C* gene in the presence of adenosine. Column binding assays using total RNA isolated from these expression assays showed that these human adenosine aptamers bind to ATP in the context of their natural transcripts. This result highlights the biological significance of naturally occurring aptamers in eukaryotes.

ABBREVIATIONS

ATP – adenosine triphosphate

ADP – adenosine diphosphate

AMP – adenosine monophosphate

GTP – guanosine triphosphate

cAMP - 2′-3′ cyclic adenosine monophosphate

cGMP - 2′-3′ cyclic guanosine monophosphate

TPP – thiamine pyrophosphate

RAB3C – Ras-related protein - 3C

FGD3 – faciogenital dysplasia 3

SELEX – systematic evolution of ligands by exponential enrichment

bp – base pair(s)

RT-PCR – reverse transcription – polymerase chain reaction

NCBI BLAST – national center for biotechnology information basic local alignment search tool

rDNA – ribosomal DNA

rRNA – ribosomal RNA

PTC – peptidyl transferase center

mRNA – messenger RNA

tRNA – transfer RNA

UCSC – University of California, Santa Cruz

snRNA – small nuclear RNA

Apta-seq – aptamer deep sequencing

SHAPE-seq – selective hydroxyl acylation and analysis by primer extension coupled with deep sequencing

RT– reverse transcriptase

cDNA – complementary DNA

RPM – read per million

NA – no alignment

ERV – endogenous retrovirus

ERVK – endogenous retrovirus group k

RLTR17B – retroviral long terminal repeat group 17 b

SINE – short interspersed nuclear element

LINE – long interspersed nuclear element

ULP1 – ubiquitin-like protease 1

Slx4 – structure-specific endonuclease subunit

UTR – untranslated region

AZA – azathioprine

MTX – methotrexate

DMSO – dimethyl sulfoxide

NAI – 2-(azidomethyl)nicotinic acid acyl imidazole

BAM – binary alignment map

M-MLV – moloney murine leukemia virus

THE1B – ERV mammalian apparent long terminal repeat retrotransposon

CHAPTER 1

Naturally occurring RNA aptamers

I. Aptamers

Aptamers are oligonucleotide sequences that are capable of binding to small molecules with high affinity and specificity. To date, a significant number of aptamer sequences have been characterized, and have been found to bind targets including biological metabolites, hydrophobic molecules, and even extracellular proteins with affinities that rival antibodies [1], [2]. The large number of aptamer targets, as demonstrated by these studies, is without a doubt an interesting property of nucleic acids given the condition that any given sequence consists of only four possible residues. Despite this limitation, aptamers are capable of binding to a variety of ligands by structural variation in physical space, forming a substrate-specific binding pocket defined by structural motifs generated by intermolecular base-pairing, and nucleotide insertions. Collectively, these structural motifs form a tertiary structure that provides the necessary interactions of the nucleotide bases with the target substrate.

Initial aptamer studies considered the statistical probability and structural diversity that a given sequence will bind to a specific target by *in vitro* selection. Through a series of column binding and elution experiments aptamers that bound to the target ligand were enriched from RNA transcribed from a random synthetic DNA pool of 10^{15} diversity containing flanking regions of known composition [2], [3]. This selective enrichment revealed a significant number of aptamers present for a given target suggesting that there is more than one solution to a given target-binding problem. This observation holds true with a number of *in vitro* selections published to date where a number of biological molecules have been used as aptamer binding targets [3]–[6].

Among the best biologically significant aptamers that have been characterized is the adenosine aptamer, which has been observed in multiple independent *in vitro* selections [6]–[8]. The adenosine binding motif, also known as the Sassanfar motif, and its consensus secondary structure is illustrated in Figure 1.1a. Despite very little overall sequence conservation, all selected sequences were found to form an adenosine-specific binding pocket in a structurally conserved manner. The conserved 11-nucleotide loop and the opposing G are nucleotides necessary to interact with the adenosine triphosphate (ATP) ligand. A NMR solution structure of the adenosine aptamer bound to adenosine monophosphate (AMP) reveals that A10 and G11 of the 11-nucleotide loop base-stack with the adenosine base. Meanwhile the opposing G (G30) base pairs with G17 to establish an intricate hydrogen bond network that interacts with the 2´ and 3´ hydroxyl group of AMP (Fig. 1.1b) [7], [9], [10]. These interactions are facilitated by helices flanking the conserved 11-nucleotide loop and opposing G, which is required to form the binding pocket (Fig. 1.1b). In other words, target specificity requires specific structural motifs determined to form the binding pocket capable of initiating contacts with the target molecule. These observations demonstrate the value of *in vitro* selection, which can be further tailored to discover and study relevant naturally occurring functional RNAs.

Figure 1.1

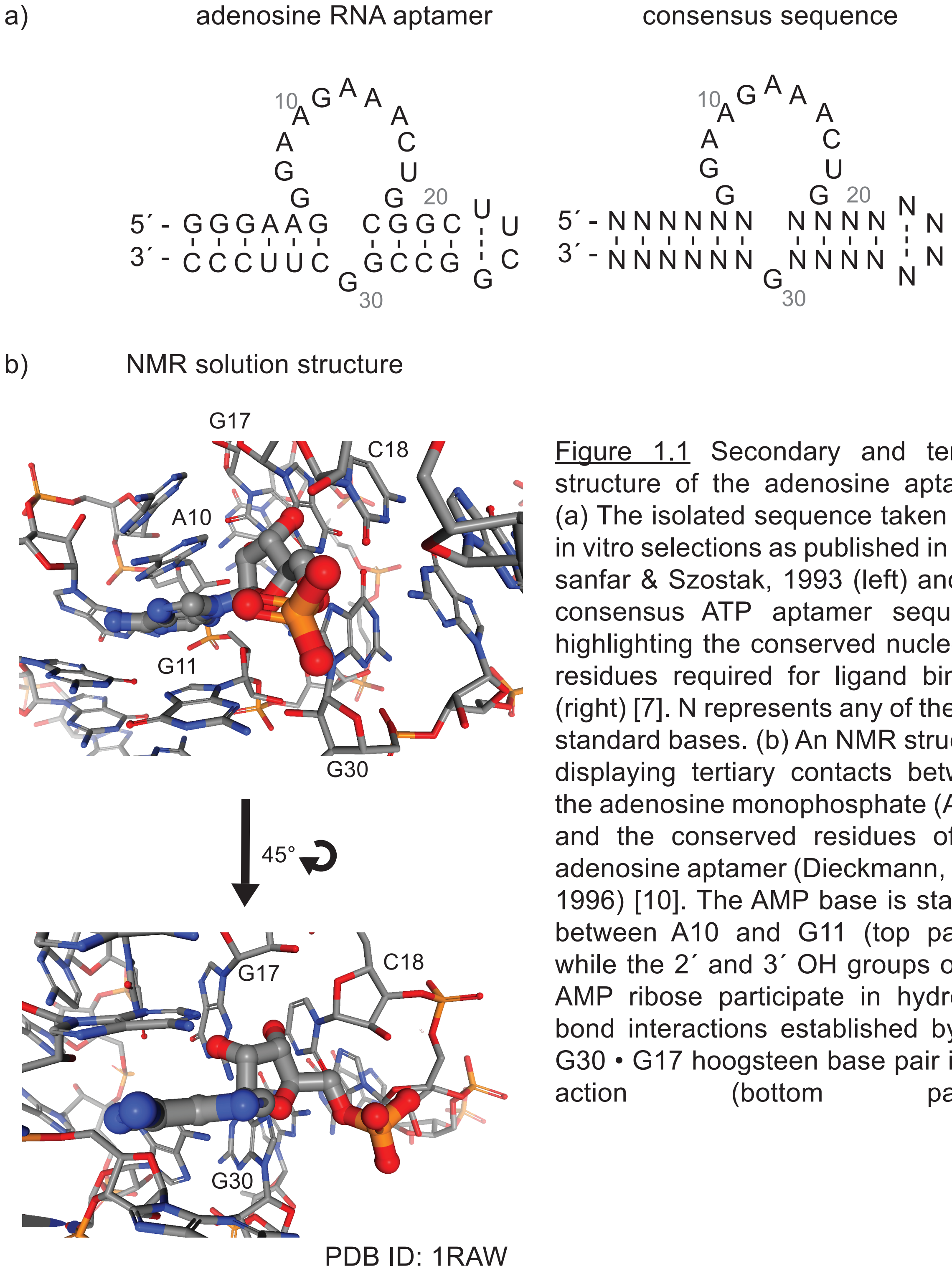


Figure 1.1 Secondary and tertiary structure of the adenosine aptamer. (a) The isolated sequence taken from in vitro selections as published in Sassanfar & Szostak, 1993 (left) and the consensus ATP aptamer sequence highlighting the conserved nucleotide residues required for ligand binding (right) [7]. N represents any of the four standard bases. (b) An NMR structure displaying tertiary contacts between the adenosine monophosphate (AMP) and the conserved residues of the adenosine aptamer (Dieckmann, et al. 1996) [10]. The AMP base is stacked between A10 and G11 (top panel), while the 2´ and 3´ OH groups of the AMP ribose participate in hydrogen bond interactions established by the G30 • G17 hoogsteen base pair interaction (bottom panel).

II. *In vitro* selection

Over the past two decades, *in vitro* selection, also known as Systematic Evolution of Ligands by EXponential Enrichment or SELEX, has served as a powerful tool for the discovery of novel DNA and RNA aptamers. Since then, extensive studies have highlighted the structural diversity within a given random pool of DNA or RNA sequences [11]. Initial *in vitro* selection studies utilized pools of synthetic random DNAs flanked by fixed, primer-binding sequences that also allowed for transcription of equally diverse RNAs, to determine the frequency of aptamers capable of binding a target molecule [3], [5].

Recently, *in vitro* selection studies have incorporated genomic DNA as a source for aptamer selections rather than synthetic DNA populations [12]–[14]. RNA aptamers selected using this approach not only serve as direct evidence of biological significance, but in some cases also revealed striking structural similarities to that of synthetically derived aptamers. For example, as in the case of the adenosine aptamer, both synthetic and genomic DNA selections revealed a number of structurally conserved aptamer sequences [6], [12], [13], [15]. These adenosine-binding motifs are also sequence-independent as has been seen with previous synthetic *in vitro* selections and represent a rare example of convergent molecular evolution spanning both genomic and synthetic sequence space. Discovery of the adenosine aptamer within the human genome motivated the investigation outlined in the following chapters to address questions involving the frequency and diversity of aptamers within genomes of eukaryotic organisms, characterization of structural motifs required for binding, and the genomic location and subsequent biological function, if any, for a given aptamer.

III. Naturally occurring RNA aptamers

RNA aptamers exist in nature as part of riboswitches. Riboswitches are ligand-dependent genetic modulators encoded within nascent RNA transcripts. Functional riboswitches consist of an aptamer domain, and an expression platform, to regulate gene expression (Fig. 1.2a). Upon ligand binding by the aptamer, changes in structural conformation due to formation of the ligand-binding pocket, result in transcriptional or in some cases translational regulation (Fig. 1.2b). In this functional example, the aptamer domain serves as a molecular sensor by responding to ligand availability. The *S*-adenosyl methionine (SAM) riboswitch, for example, undergoes a conformational change upon binding of SAM. In this example, a binding event induces a conformational change within the aptamer domain allowing the formation of a downstream terminator stem loop capable of terminating transcription [14]. To date, many riboswitch families, including the SAM riboswitch, and the thiamine pyrophosphate (TPP) riboswitch, have been characterized in bacteria [16]. While many of these prokaryotic riboswitches clearly show the active role of RNA-mediated gene regulation at the level of transcription or translation, analogous examples of riboswitch mechanisms within eukaryotic systems are scarce.

Functional eukaryotic riboswitches remain evasive. Structure-based searches have largely been unsuccessful due to partial conservation of the aptamer domain and little to no conservation of the expression platform. Despite this challenge, the TPP riboswitch, found to be functionally conserved within all three domains of life, serves as a single example of a functional riboswitch in eukaryotes [17]. Interestingly, studies of the TPP riboswitch in plants and fungi have shown thiamine-dependent 3´ processing of nascent RNA coding for proteins required for thiamine synthesis [17], [18]. Despite the limited

Figure 1.2

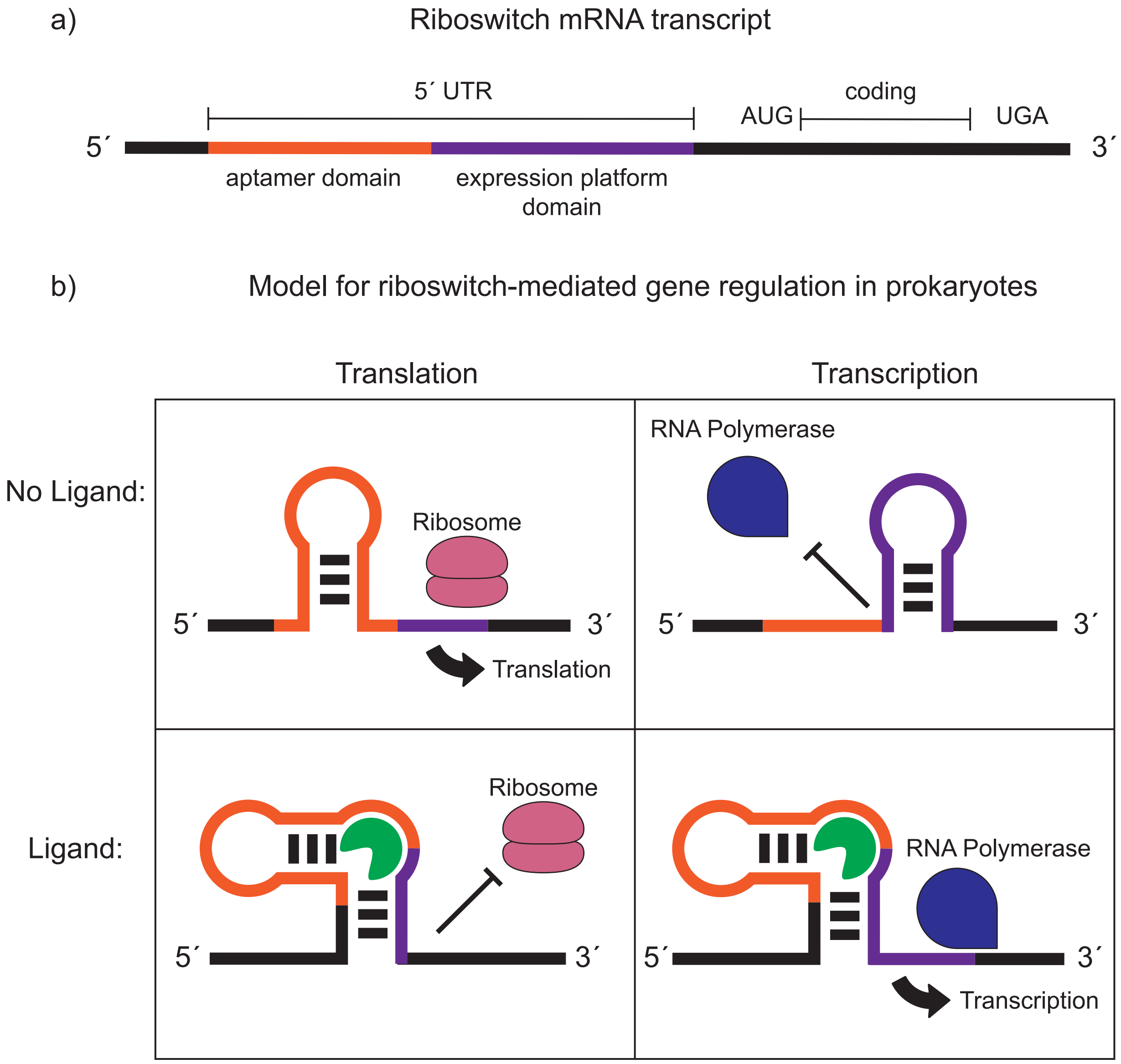


Figure 1.2 Aptamers are naturally expressed in prokaryotes within nascent RNA transcripts as part of riboswitches. Riboswitches contain both an aptamer domain and a highly variable expression platform (a). Upon transcription, structural reorganization of the RNA transcript due to the formation of a ligand binding pocket and can regulate gene expression at the level of transcription or translation. Two examples of riboswitch-mediated gene regulation are shown in (b). In the left panel, mRNA translation is inhibited by the presence of the ligand (green) due to sequestration of the ribosome binding site. In the right panel, the formation of the ligand binding pocket disrupts the downstream terminator hairpin loop encoded by the expression platform. Disruption of this terminator loop reveals a transcription promotor, allowing transcription to continue.

examples of eukaryotic riboswitches, the complexity of gene regulation achieved by such functional RNAs, such as alternative splicing as seen with the TPP riboswitch, bring to light the significance of their contribution to cellular processes. Characterization of these functional RNA transcripts introduce a novel mode of genetic regulation that can most certainly become a target for future applications.

IV. Objectives of the dissertation

Today, a significant effort is placed on characterizing RNA aptamers from synthetic libraries via *in vitro* selection for drug and clinical applications [19]. As a result, very little effort has been placed on describing the natural abundance and the biological purpose of eukaryotic RNA aptamers, despite the two known examples within the human and plant genomes. The *in vitro* selection study performed by Curtis and Liu revealed the presence of naturally occurring GTP aptamers within a relatively small metagenomic library. While their claim regarding biological function is speculative at best, identification of a GTP aptamer sequence demonstrates the propensity of using *in vitro* selection to discover biologically relevant RNA aptamers from genomic DNA. This fact is supported by the classification of the TPP eukaryotic riboswitch within filamentous fungi by the Breaker group, which strongly suggests that nascent RNA may play a significant role in gene regulation than previously thought.

To further investigate this mode of genetic regulation, the first objective of this study was to develop a diverse metagenomic DNA library consisting of various genetic model organisms to provide a platform for the discovery and characterization of naturally occurring RNA aptamers by *in vitro* selection. Acquisition of this metagenomic DNA library is then followed by a set of experiments designed to validate the diversity and unbiased representation of the organisms comprised within the DNA library. These efforts and their findings are presented in chapter 2. The second objective is to generate and investigate sequences arising from selective enrichment of RNA aptamers using the previously mentioned metagenomic DNA library with specificity and affinity for adenosine triphosphate (ATP) and cyclic adenosine monophosphate (cAMP). Experimental data from these *in vitro* selections are detailed in chapter 3.

The third and most significant objective was to structurally characterize and investigate a biological role, if any, for a given aptamer derived from genomic *in vitro* selections. This goal was divided into two mutually exclusive parts including extensive analysis of the resulting enriched libraries mentioned above, and surveying two known human adenosine aptamers within the *RAB3C* and *FGD3* genes by *in vitro* cell culture assays. Using deep sequencing coupled with structure probing techniques and computational analyses we studied the dominant RNA populations within the enriched libraries that survived *in vitro* selection rounds. Sequencing and structural data obtained from the metagenomic *in vitro* selections is discussed in chapter 4, while *RAB3C* and *FGD3* expression data is discussed in chapter 5. While these approaches simply begin to draw potential biological functions, their findings provide a baseline for future metagenomic library endeavors, which not only includes ongoing aptamer discovery, but also in refining future efforts towards discovery of additional eukaryotic riboswitches and naturally existing functional RNAs.

CHAPTER 2

Construction and design of a metagenomic DNA library

I. Introduction

Genomic SELEX was introduced by Singer and Gold in 1997 when a small metagenomic library consisting of human, yeast and *Escherichia coli* genomic DNA were pooled and used to generate a DNA library. [20]. The library was designed with a fixed forward DNA adapter and randomized 3´ tail to anneal and extend denatured genomic DNA for amplification. After amplification, gene-specific PCR analyses were used by the authors to determine sequence diversity and test genome representation of the metagenomic DNA library [20]. While GC sequence-specific biases were detected during amplification, all organisms were found to be represented proportionately with very little introduction of artificial mutations. Similarly, Salehi-Ashtiani et. al. designed a genomic pool by partial digestion of human genomic DNA using DNase I to study the presence of self-cleaving ribozymes encoded within the human genome. After enzymatic digestion, hairpin adapter sequences of known sequence were ligated onto the fragmented genomic DNA. Ligated DNA products were then subjected to a single stranded DNA-specific nuclease and then amplified by primer extension to generate a human genomic DNA library [21]. Both of these approaches yielded successful genomic DNA libraries and provided a platform for studying functional and biologically relevant RNA molecules. These techniques inspired the following approaches to generate a diverse eukaryotic metagenomic DNA library to further study functional RNAs including RNA aptamers that may exist within eukaryotic genomes.

II. Design of the metagenomic library DNA adapters

Sequence-specific synthetic adapters were designed for the metagenomic library with the following requirements: 1) The overall sequence composition must not be gene specific to avoid amplification artifacts that may arise during amplification of the DNA library or during *in vitro* selection amplification steps, 2) the primer sequences alone, and in any combination, must not generate artificial amplicons that result from self-priming or unwanted primer extension, 3) the forward adapter sequence must contain a T7 RNA polymerase promoter to generate RNA transcripts for downstream experiments, 4) the adapter sequences must be Illumina-compliant for easy library preparation to submit for deep sequencing on the Illumina HiSeq platform. Both forward and reverse adapters are capable of hybridizing onto the Illumina flowcell and are 62 and 33 bps in length, respectively. Each sequence set was checked and verified for minimal complementarity using the Integrated DNA technologies oligo analyzer tool. Using these online tools and PCR amplification, all permutations of each DNA adapter pair did not exhibit any unexpected and stable base-pairing or self-hybridizing interactions.

III. Assembly of the metagenomic DNA library

The first challenge during assembly of a large metagenomic DNA library was to consider sequence diversity and to ensure equal representation of each organism. Prior to any quantification, the organisms to be represented in the genomic library were chosen using the following prerequisites: 1) the organism must fall within the eukaryotic domain of life 2) the organism has served as a genetic model in various fields of biological research and 3) a published, and preferably highly annotated, DNA sequence assembly is readily accessible. A list of these organisms and their genome sizes is shown in Figure

2.1. Considering the total number of nucleotides present within this metagenomic sample, at approximately 1 x 10^{10}, the minimum sequence diversity required must approach 1 x 10^{10} sequences to ensure single-nucleotide resolution of every genetic model organism. In comparison, the sequence diversity of synthetic DNA libraries previously used for *in vitro* selections, at approximately 10^{15} molecules, means that a metagenomic library with a diversity of 10^{10} falls well within capacity of performing an aptamer selection. Furthermore, an equivalent number of genome copies of every organism must be introduced to prevent any genomic DNA biases towards any given organism. To do this, the C-value, or equivalent mass value per single copy (n) of genomic DNA derived from genome assembly records was used to determine the input amount of DNA needed during assembly of the combined genomic DNA sample.

Genomic DNA from *Anopheles gambiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe* was extracted by crushing frozen whole tissue samples above liquid nitrogen by mortar and pestle followed by phenol chloroform extraction. As for the remaining organisms, genomic DNA samples were kindly donated by various research labs whom routinely extract total genomic DNA. For quality control, all genomic DNA samples were visualized on a 1 % agarose gel stained with ethidium bromide (Fig. 2.2). Each genomic DNA sample varied in the amount of total RNA and total genomic DNA present, presumably due to the method of extraction. Using a spectrophotometer in tandem with the imaging software, approximate values for high molecular weight DNA bands were calculated to compensate for variances in DNA preparation. This method was chosen over size-selective gel extraction to maintain high molecular weight DNA, and sample diversity. Once pooled,

Figure 2.1

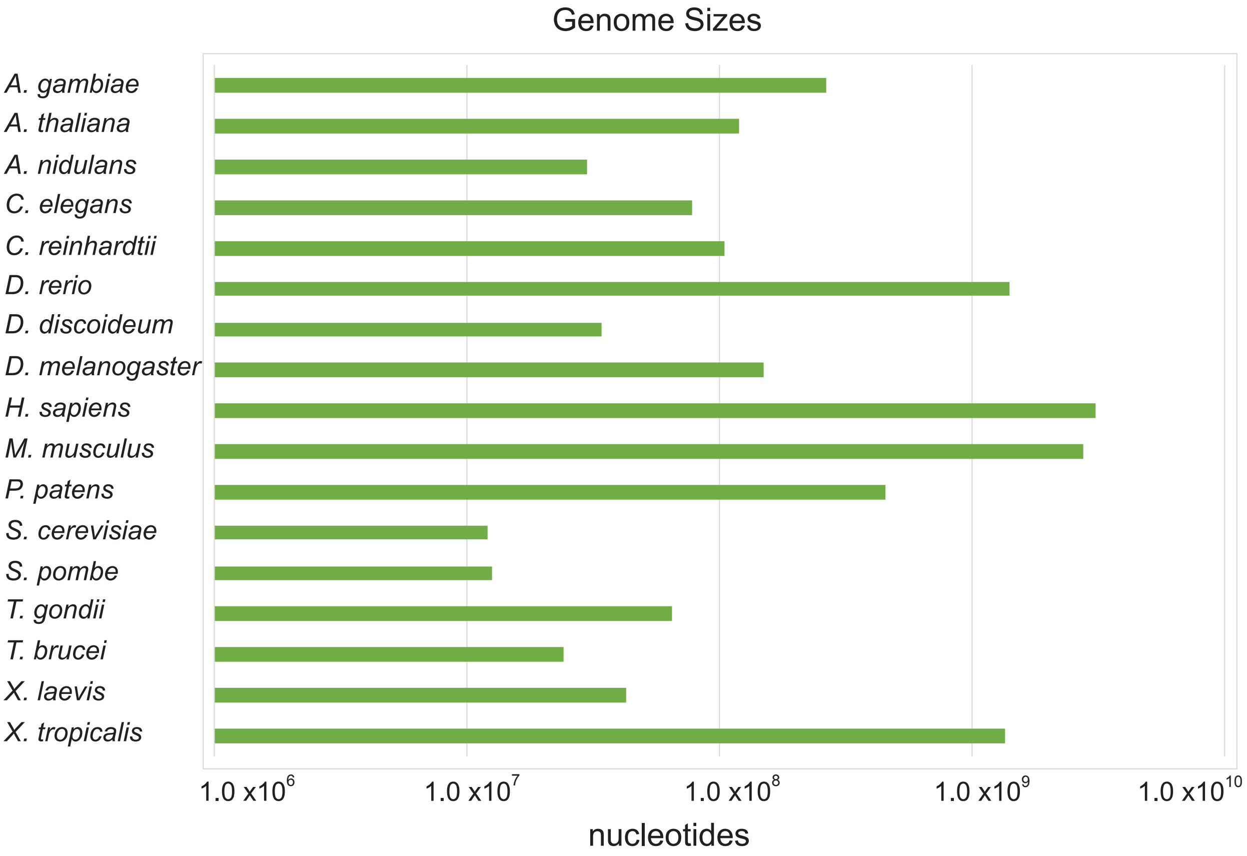


Figure 2.1 A list of the seventeen organisms used to generate the metagenomic DNA library and their respective genome lengths are represented above. The human, mouse and zebrafish genomes are among the largest in length at nearly two orders of magnitude larger than that of baker's yeast (*S. cerevisiae*). To avoid representation biases during metagenome assembly, equivalent copies of each organism were pooled by using known c-values.

Figure 2.2

Gel analysis of high molecular weight genomic DNA

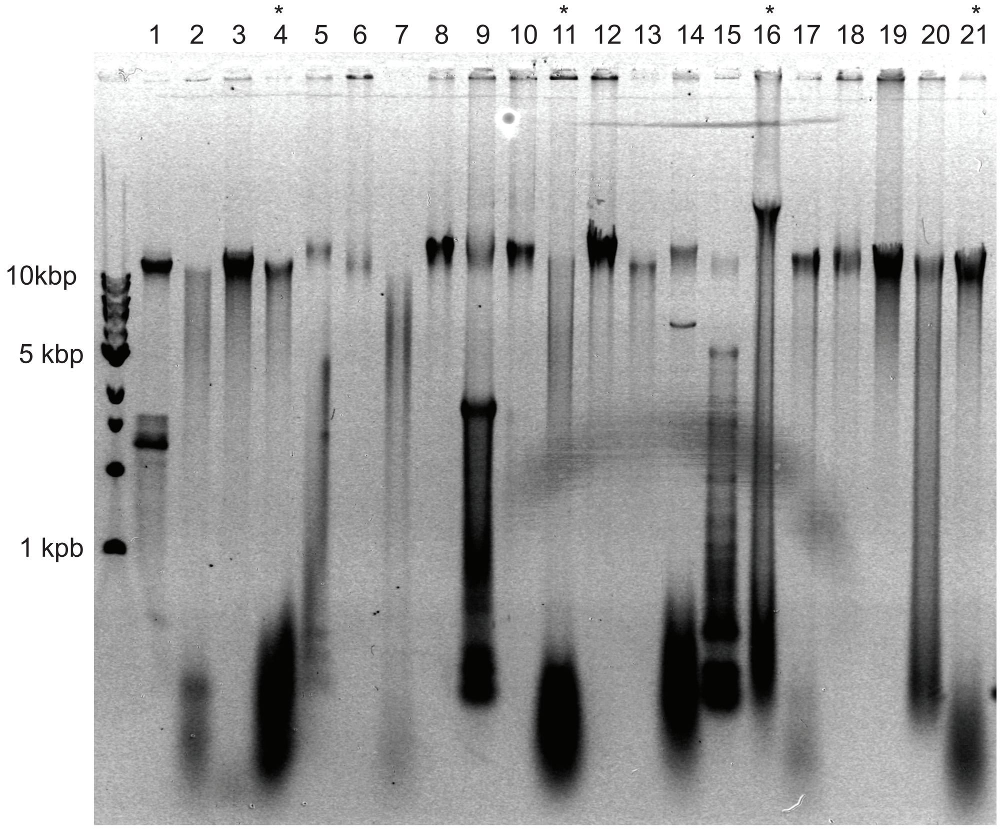


Figure 2.2 A 0.8 % agarose gel showing high molecular weight genomic DNA extracted from various eukaryotic organisms. Low molecular weight banding/smearing (<500 bp) correspond to total RNA co-extracted with genomic DNA (NaOH hydrolysis not shown). 1) *A. gambiae* 2) *A. thaliana* 3) *A. nidulans* 4) *B. distachyon* 5) *C. elegans* 6) *C. reinhardtii* 7) *D. rerio* 8) *D. discoideum* 9) *D. melanogaster* 10) *H. sapiens* 11) *M. truncatula* 12) *M. musculus* 13) *P. patens* 14) *S. cerevisiae* 15) *S. pombe* 16) *S. purpuratus* 17) *T. gondii* 18) *T. brucei* 19) *X. laevis* 20) *X. tropicalis* 21) *Z. mays*. Lanes with an asterisk indicate the organisms that were not included in the final metagenomic DNA library.

the high molecular weight genomic DNA was subjected to physical agitation to obtain a desired fragment size.

A number of procedures exist for fragmentation of high molecular weight DNA including nuclease-driven fragmentation and mechanical shearing [22]. Currently, enzymatic fragmentation of DNA involves employing commercially available mutant transposases to insert 19 base pair sequences along a target DNA molecule [23], [24]. However, this approach requires transposase-specific insertion sequences that are typically composed of partial inverted repeats. Not surprisingly, slight biases in insertion sites have also been identified with this methods and are significantly amplified after PCR library preparations [24], [25]. For this reason, traditional methods involving mechanical shearing and enzymatic ligation was chosen. Pooled genomic DNA was sheared nonspecifically by sonication using the Acoustic S2 sonicator. This process generated a distribution of fragmented genomic DNA ranging from 200 bps to 500 bps with a mean size of 300 bps as expected (Fig. 2.3a). Prior to size selection and initial amplification, the fragmented DNA pool was first prepared for enzymatic ligation of the synthetic adapters.

IV. Adapter ligation

Detailed experimental setup for adapter ligation of the metagenomic DNA library can be found in the Materials and Methods chapter (Chapter 6). Briefly, fragmented genomic DNA was extended and repaired using DNA polymerase to generate blunt ends. Cohesive ends were introduced to the repaired genomic DNA by the addition of a 3´-deoxyadenosine overhang and phosphorylation of 5´ ends by the Klenow fragment of DNA polymerase I and T4 phosphonucleotide kinase, respectively. Synthetic DNA oligos described previously, containing the appropriate 3´deoxythymidine overhangs and 5´

Figure 2.3

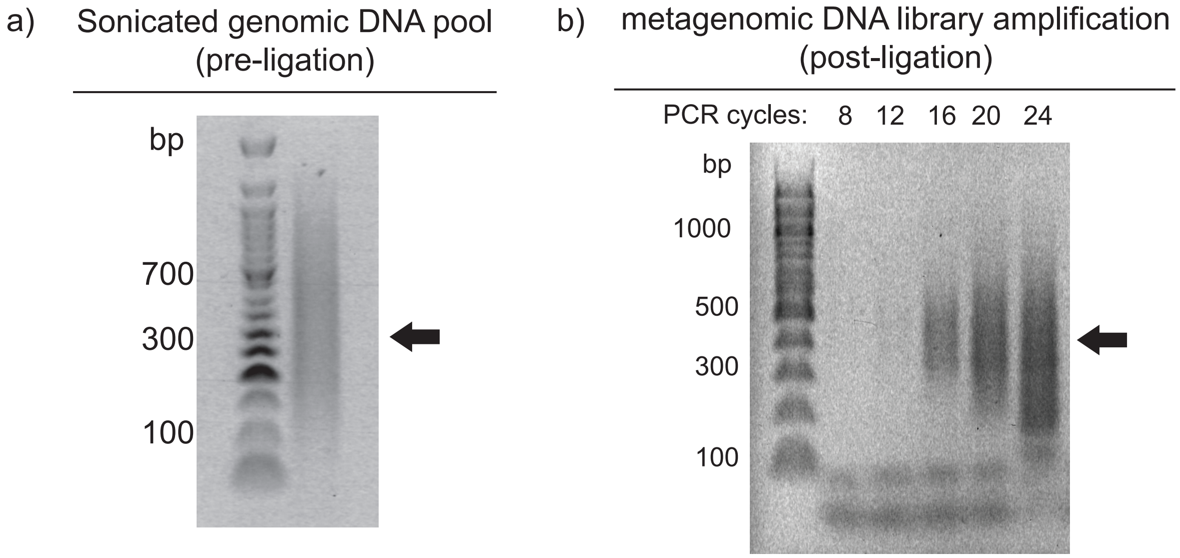

a)
Sonicated genomic DNA pool
(pre-ligation)
bp
700
300
100
b)
metagenomic DNA library amplification
(post-ligation)
PCR cycles: 8 12 16 20 24
bp
1000
500
300
100

Figure 2.3 A set of agarose gels confirming steps taken during assembly of the metagenomic DNA library. Pooled high molecular weight metagenomic DNA was sheared by acoustic agitation using a Corvaris S2 shearer using precalibrated settings to generate a mean of 300 bp fragments (a). After fragmentation, priming of the DNA ends, and in vitro ligation of DNA adapters, the ligated metagenomic pool was amplified by PCR and visualized on an agarose gel (b).

Figure 2.4

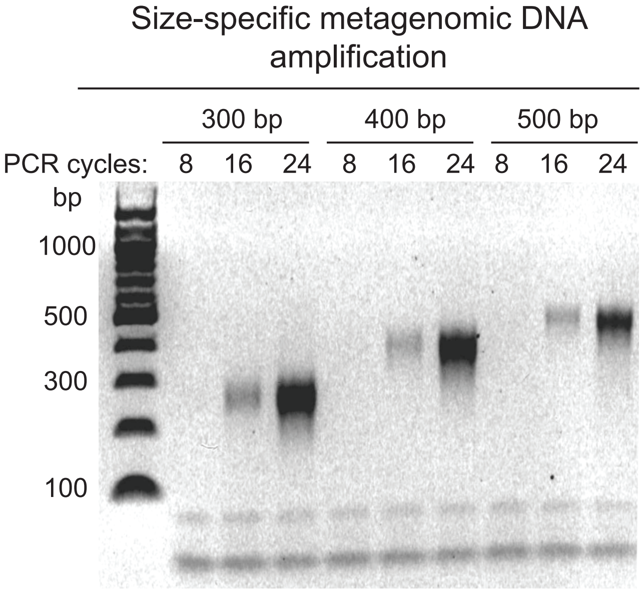

Size-specific metagenomic DNA
amplification
300 bp
400 bp
500 bp
PCR cycles: 8 16 24 8 16 24 8 16 24
bp
1000
500
300
100

Figure 2.4 After initial amplification of the metagenomic DNA library, bands corresponding to 300, 400 and 500 base pairs were excised from an agarose gel and reamplified by PCR. These products were visualized on an agarose gel (left) to confirm sequence length and to be used for in vitro selection.

phosphorylation modifications were ordered from IDT and purified by gel electrophoresis. Pre-annealed adapter oligos were introduced to genomic DNA with cohesive ends and the final ligation reaction was set up at 16 °C for 30 minutes using DNA ligase.

V. Metagenomic library diversity

To confirm high sequence diversity within the metagenomic library, semi-quantitative PCR amplification, or by fractionation of the PCR reaction at every 4th cycle, was performed to estimate the original number of molecules within the ligated starting material. The results of these collected fractions after gel electrophoresis are illustrated in Figure 2.3b. This initial amplification of the genomic library revealed a distribution of amplified genomic DNA from 200 to 500 bps at PCR cycle 16. To estimate library diversity and establish a size-specific DNA pool, the 300 bp, 400 bp and 500 bp products were excised and eluted from an agarose gel. Re-amplified products were then analyzed using an agarose gel to confirm sequence length (Fig. 2.4). Using mass values from the loading control and densitometry analysis, a mass was approximated for the amplified DNA library. To calculate library diversity, we corrected the mass value for the rate of logarithmic amplification and the appropriate dilutions. With this approximation, the final diversity of the metagenomic library was calculated to be $1x10^9$. This value is one order of magnitude smaller than the desired $1x10^{10}$ molecules needed for single nucleotide resolution for each organism represented in the metagenomic library (Fig. 2.1). Despite this result, however, since each genome was pooled by mass copy number (C-value normalization), each organism within the library remains equally represented.

To validate genome representation, the final metagenomic DNA library was submitted for deep sequencing analysis using the Illumina HiSeq 2500 sequencer. By obtaining deep sequencing reads, the origin of each sequence taken from the pool can be assigned using readily available genome databases such as the NCBI BLAST nucleotide reference database and the UCSC genome browser. First, sequences were prepared using an in-house computational pipeline designed to remove sequencing artifacts. Additional details on this pipeline can be found in the Methods chapter. The resulting trimmed-sequences were then studied for sequence composition *via* genomic alignments and text-based sequence searches.

Sequence searches for ribosomal DNA (rDNA) genes within the deep sequencing file were performed as a method to ensure that each organism is represented within the metagenomic DNA library and that no biases exist for a given organism. To do this, we focused on species-specific expansion segments (ESs) of the large ribosomal RNA. ESs are regions of ribosomal RNA that arise from the common core (the peptidyl transferase center (PTC)) and are characterized by having distinct structures and functions that are required for ribosome biogenesis and activity [26], [27]. Over the course of evolution, divergence of eukaryotic ribosomal RNAs has introduced variations in these regions resulting in species-specific ESs [27]. We used these distinct features of the large ribosomal RNA to perform our sequence searches.

DNA sequences of the reference large ribosomal RNA and their respective Clustal Omega alignments were downloaded from the University of Texas Comparative RNA Website and Project (CRW) and used as references to perform these searches [28]. Note that due to incomplete genomic assemblies for some of the 17 organisms, not all copies of all rDNA genes for the organisms are available. While the genomes of most of these organisms have been sequenced and partially assembled, the repetitive characteristic of ribosomal DNA and transfer RNA genes create challenges during the assembly pipeline, particularly for organisms such as *A. thaliana* and *Chlamydomonas reinhardtii*, which contain an unusually large number of ribosomal RNA genes [28]–[30]. Nevertheless, for those organisms that have a complete genomic assembly, we searched for species-specific ESs beginning near PTC to determine their depth within the metagenomic DNA deep sequencing file.

For every organism with a fully assembled reference genome, a unique ribosomal RNA expansion segment was identified. Based on each expansion segment, we then calculated a depth quotient with the respective organisms' ribosomal DNA copy number taken from values available in the literature [29], [31]–[37]. These organisms, rDNA copy numbers, ES abundances within the sequencing file, and depth quotients are reported in Table 2.1. Values within the sense and antisense columns correspond to the number of ES sequences found within the forward (Read1) and reverse (Read2) sequencing files. The totals represent the sum of both forward and reverse reads and were used to calculate a depth quotient using the rDNA copy number.

Abundance values for each of these unique expansion segments correlated with their respective ribosomal DNA copy number represented by the depth quotient values. With the exception of *A. thaliana*, *D. melanogaster* and *D. discoideum,* all rDNA segments are equally represented in the metagenomic library. It is important to note that ribosomal DNA copy numbers found in the literature are approximate. For organisms like S*. cerevisiae* and *A. thaliana*, various ribosomal DNA copy numbers have been reported over the past decade [29], [34]. Some genomes such as *A. thaliana* and *D. melanogaster*,

Table 2.1

rDNA sequence searches for unique expansion segments

Organism (Subject)	Coordinate	rDNA copies	Sense	Antisense	Total	Depth quotient
A. thaliana	4839 - 4861	570	138	152	290	0.5
C. elegans	5287 - 5310	55	2	3	5	0.1
D. discoideum	2181 - 2201	200	54	30	84	0.4
D. melanogaster	2055 - 2078	170	67	117	184	1.1
H.sapiens	2537 - 2560	212	8	8	16	0.1
M. musculus	2848 - 2869	165	11	7	18	0.1
S. cerevisiae	89 - 109	150	13	13	26	0.2
S. pombe	5517 - 5542	85	8	6	14	0.2
T. brucei †	5180 - 5201	56	5	7	12	0.2
T. gondii	4222 - 4243	110	2	4	6	0.1
X. tropicalis	2988 - 3010	630	39	33	72	0.1
P.T.C.	5216 - 5236		720	722	1442	
H75 - H78	4998 - 5018		445	527	972	

Table 2.1 A text-based sequence search for unique rRNA expansion segments (ES) present in each organism listed above was performed. The number of matches for each expansion segment found within read 1 (sense) and read 2 (antisense) deep sequencing files are reported above, along with the total number (sum of sense and antisense matches) observed. Published rDNA copies were used to calculate a copy number ratio (depth quotient) between the observed matches and known rDNA copies in order to determine if any organism-specific biases are present in the metagenomic DNA library [29-37]. P.T.C = peptidyl transferase center. H75-H78 = Helix 74 and Helix 78 of domain V of *S. cerevisiae* rRNA. † indicates polyfragmented rDNA is known to occur within the genome.

are also known to have some of the largest numbers of rDNA copies and may therefore show large amounts of variation in copy number [29], [31], [38], [39].

VI. Conclusion

To date, only a handful of *in vitro* selections have been performed using a metagenomic DNA library. The main objective of the experiments described in this chapter were to generate a diverse metagenomic DNA library consisting of genomes of 17 equally represented eukaryotic genetic model organisms to be used for *in vitro* selection. Incorporating these many organisms within a DNA library, however, presents a number of challenges including equal distribution of genomic DNA to avoid introducing any biases. Despite differences in genome lengths, the mass copy number (C-values) for each organism was used to assemble an unbiased metagenomic library. After adapter ligation and amplification by PCR, we found that the final diversity of the size-selected metagenomic DNA library was within 1 x 10^9, effectively making this assembly one order of magnitude under-sampled when compared to the ideal diversity of 1 x 10^{10} – the diversity required for single-nucleotide sampling of each organism. While this is true, ribosomal DNA abundances determined by organism-specific expansion segments show that each organism remains equally represented. Altogether, this means that there are minimal organism-specific biases present within the metagenomic library. Instead, additional deep sequencing analysis of the metagenomic DNA library revealed that a significant bias exists for simple sequence repeats. These sequence repeats are found naturally within the genomes of these eukaryotic organisms, such as those found in centromeres and telomeres, and are therefore expected to have a larger abundance.

Additional insight into these genomic repeats within the metagenomic library are further discussed in Chapter 4.

The purpose of generating this eukaryotic metagenomic library was to provide a platform for functional metagenomics, including discovery of naturally existing functional RNAs such as aptamers and riboswitches. Today, metagenomic DNA libraries are mediums for new discoveries as they provide a wealth of sequencing data that can be used in conjunction with high throughput DNA sequencing and sophisticated machine learning algorithms [40], [41]. These approaches are used today with metagenomics libraries generated from environmental samples including sea water collection, soil sampling and fecal matter [42], [43] With advances in high resolution DNA sequencing methods, sequence pattern recognition algorithms, and sequence assembly tools, these metagenomic libraries have already contributed to discovery of novel high processivity enzymes [44]. As the computational efforts evolve to screen through vast amount of sequencing data, it is exciting to see what a eukaryotic metagenomic library may contribute to the field of metagenomics.

CHAPTER 3

In vitro selection using a metagenomic DNA library

I. Introduction

In vitro selection is a powerful technique that can be used to enrich RNA aptamers from a diverse library of molecules. This goal can be achieved by applying selective pressure on a large population of sequences to subtract RNAs with nonspecific binding and collect RNAs capable of binding to a specific target. By using an agarose matrix column with a covalently attached ligand of interest, RNA aptamers can be separated by fractionation and enriched over multiple iterations of this technique. Over a few rounds, nonspecific RNAs are subtracted from the starting population through a series of short washes thereby enriching functional RNA sequences on the column matrix. These RNA aptamers are then collected as elution fractions via free-ligand exchange using the target ligand in excess to avoid biases due to binding competition. After collection of the RNA elution fractions, an RT-PCR step regenerates an enriched dsDNA library that can be used for subsequent rounds until a significant level of RNA binding is measured. In this chapter, a purified RNA and co-transcriptional *in vitro* selection were designed to enrich for steady-state and kinetically driven RNA aptamers. A summary of the *in vitro* selection cycle and the difference between these two strategies is illustrated in Figure 3.1. With the help of radioactive labeling, and semi-quantitative RT-PCR, RNA binding can be monitored over each round by scintillation counting and gel electrophoresis

To facilitate ligand binding during *in vitro* selection rounds, RNA refolding and buffer conditions were used as described in previously published *in vitro* selection work [13], [21]. For *in vitro* selections performed here, ionic concentrations were chosen to most

Figure 3.1

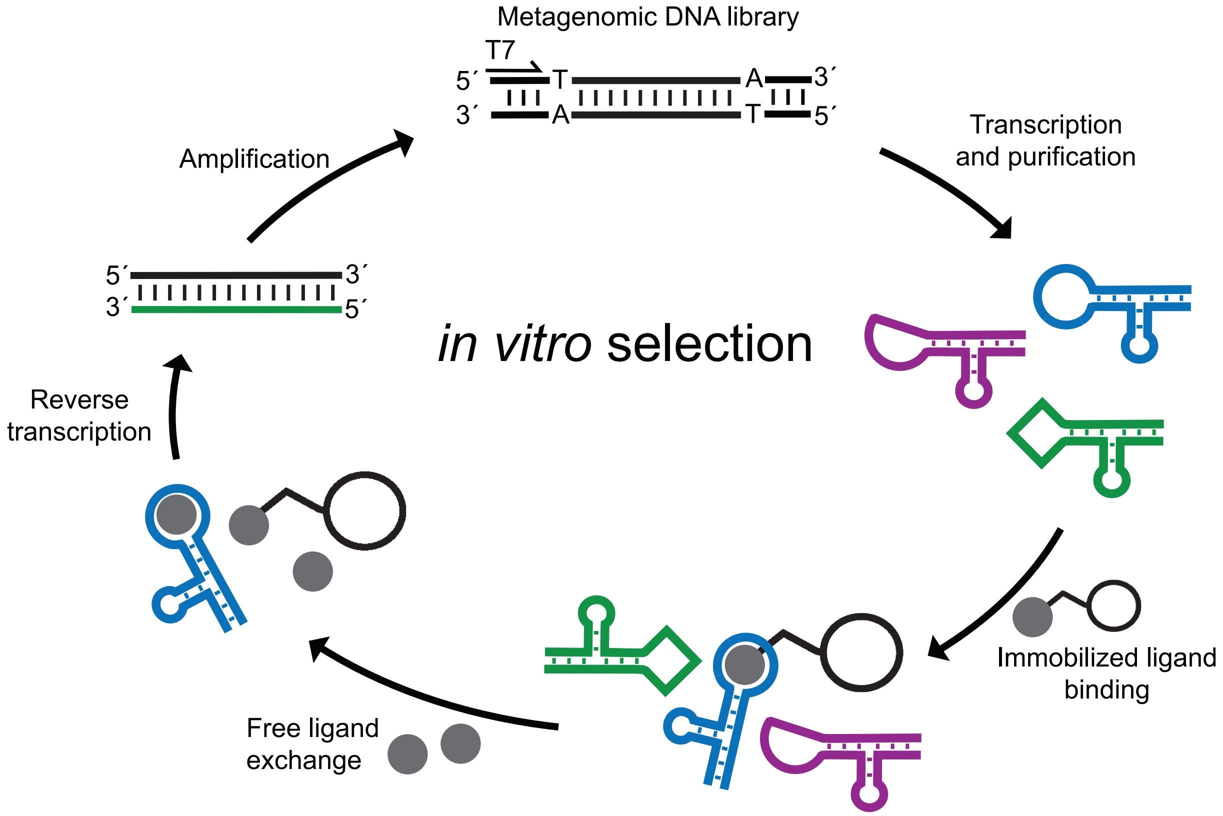


Figure 3.1 RNA aptamer *in vitro* selection scheme. One cycle of *in vitro* selection begins with transcription of the metagenomic DNA library from a T7 promoter located on the forward adapter. RNA transcripts are then purified by gel electrophoresis and resuspended in binding buffer containing divalent ions. RNAs are melted and refolded before subjecting them to an agarose matrix column containing the immobilized ligand of interest (ATP, cAMP or cGMP). Nonspecific RNAs are collected by washing the column with binding buffer and saved for downstream quantitative analysis. RNAs with affinity for the target ligand are eluted and collected using binding buffer containing the respective ligand by free ligand exchange. These elution fractions are then used as a template for RT-PCR to regenerate an enriched DNA library that can be used for the next round of selection. For co-transcriptional selections, transcription takes place in the presence of the immobilized ligand agarose matrix. Wash and elution fractions are immediately collected for quantitation by gel electrophoresis. Bands corresponding to the elution fractions are excised and prepared for RT-PCR. Each cycle is repeated again until a significant fraction of RNAs are found to elute from the column by free ligand exchange.

accurately represent the natural intracellular ionic environment. While the intracellular ion concentrations may vary between each organism represented in the metagenomic library, it is assumed that these variances are not highly divergent due to the establishment of membrane potentials needed for osmoregulation and cellular activity [45], [46]. Furthermore, because RNA folding is also mediated by magnesium concentrations, each selection was adjusted for an effective magnesium concentration based on the ability of the target metabolite to coordinate magnesium ions. These details along with a list of reagents used for *in vitro* selection can be found in the methods chapter.

II. Target Metabolites

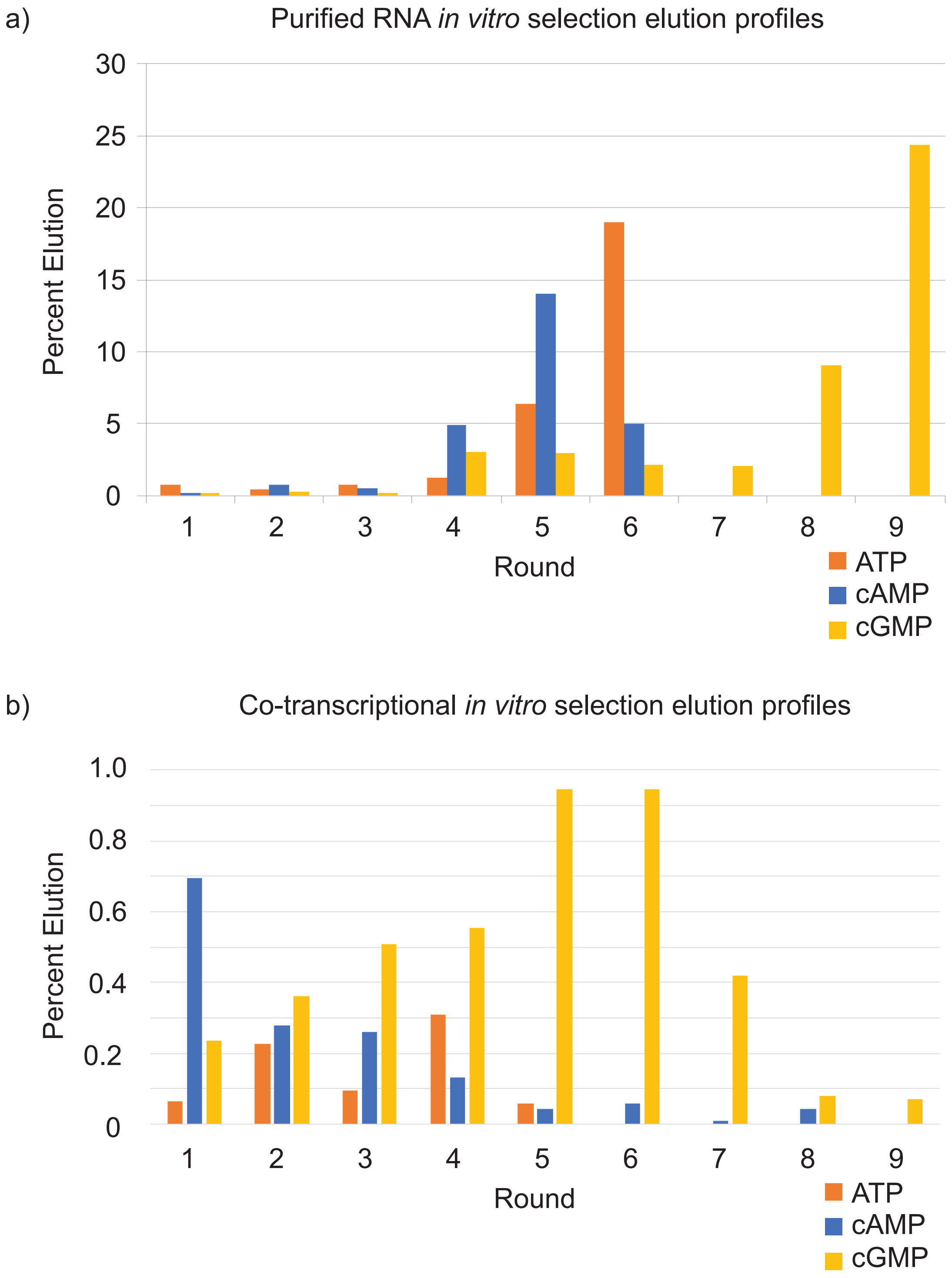
A series of *in-vitro* selections were performed using agarose-linked metabolite targets including 3´-5´ cyclic adenosine monophosphate (cAMP), 3´-5´ cyclic guanosine monophosphate (cGMP), and 5´ adenosine triphosphate (ATP). Each target was immobilized on an agarose matrix by covalent attachment to the C8 atom of the base and were purchased through a commercial supplier. The cAMP and cGMP ligands were chosen because of their known roles as secondary messengers within cellular signal transduction pathways, including cell cycle regulation, cytoskeletal rearrangement, and transcription regulators. ATP was also selected as a target ligand because of the endogenous availability within living systems and because it is required for essential biochemical reactions and processes. Furthermore, all of these targets mentioned are readily available within the nucleus in the presence of transcription factors and may very likely act as targets for a given aptamer to regulate genetic and cellular activity.

These small molecules are highly associated with important biological process and have been shown, at least for prokaryotic organisms, to interact with RNA riboswitches to influence gene regulation on the level of transcription and translation. These riboswitches and their associated aptamers high affinity and specificity for their respective target and serve as molecular sensors that play a direct role in gene expression. Such riboswitch mechanisms have yet to be discovered within eukaryotes, with the exception of the thymine pyrophosphate (TPP) riboswitch, which has been shown to regulate mRNA isoform variants for the thymine pyrophosphate mRNA [47]. In this chapter, data obtained from these series of metagenomic *in vitro* selections are reported. The 300 bp size-selected metagenomic DNA library, previously introduced in chapter 2, was used as the starting material for each selection. Below, binding data obtained from both purified RNA and co-transcriptional selections are described and investigated.

III. *In vitro* selection by UREA-PAGE

In order to obtain size-specific RNAs and to avoid aborted nonspecific RNA products during selection, 300 nt radiolabeled RNA products were excised from a denaturing polyacrylamide gel followed by elution and ethanol precipitation, as detailed in Chapter 6. After a few enrichment rounds, a significant increase in the percent elution of RNA was observed for cAMP round 5, cGMP round 9, and ATP round 6 (Fig. 3.2). These results suggest that for each target, a population of RNAs exhibit binding activity to their respective ligands. To immediately address whether this was true, cAMP round 5 and ATP round 6 libraries were inserted into the TOPO-TA vector for transformation and sanger sequencing. PCR amplicons derived from successful transformants were sequenced to determine the most dominant sequence within the population and to perform genomic alignments using BLAST and UCSC genome browser [48], [49]. In

Figure 3.2



a) Purified RNA *in vitro* selection elution profiles

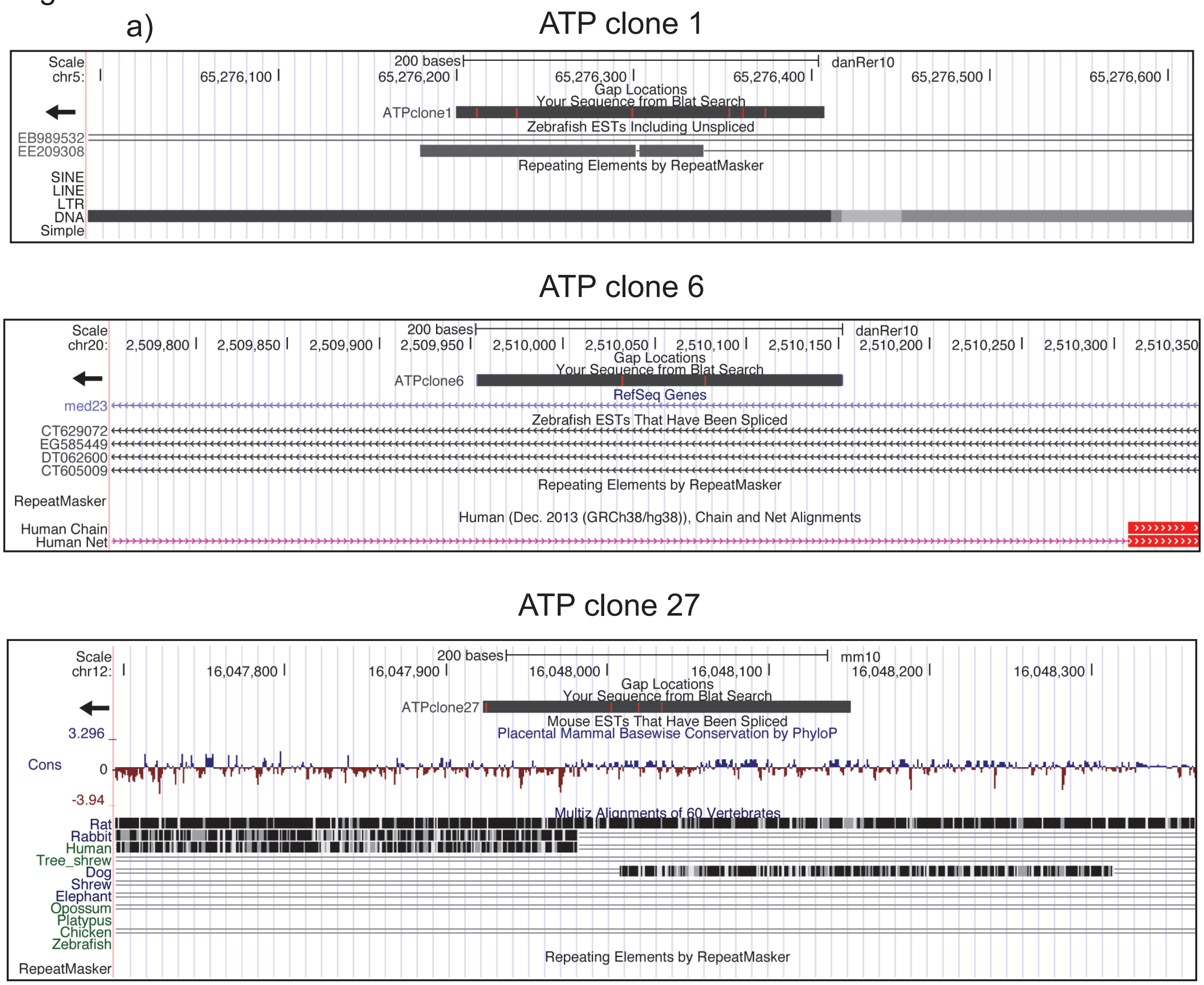
b) Co-transcriptional *in vitro* selection elution profiles addition, these sequences were then used as templates for clone-specific binding activity *in vitro*, discussed further below.

From the ATP *in vitro* selection, three unique sequences from round 6 were successfully isolated; ATP clone 1, 6, and 27. Using the BLAST database, ATP clones 1 and 6 were identified to have originated from the zebrafish (*D. rerio*) genome, while ATP clone 27 had high sequence similarities with the mouse (*M. musculus*) genome. Upon closer inspection of these clones and their genomic contexts with the UCSC genome browser, only ATP clone 6 was identified as residing within an intron of a highly conserved *med23* gene (Fig. 3.3). On the other hand, ATP clone 1 was found to be unique to the zebrafish genome as a relatively large ~500-600 nt tandem repeat annotated as DNA25TWA1_DR. In addition, ATP clone 27 distinctively mapped to a specific locus within chromosome 12 of the mouse genome rendering it in the middle of an annotation desert. For all three ATP clones, *in vitro* binding activity was extremely low with high degrees of variability (Fig. 3.3b). This irreproducibility was surprising, given that these sequences presumably represent a fraction of the most abundant sequences within the enriched library. In addition, no common features were immediately observed between all three sequences, suggesting that these isolated sequences are either co-enriched with an unidentified ATP-aptamer sequence, or simply represent insufficient selective pressure.

Due to difficulty in Sanger sequencing, only two cAMP clones were obtained from the cAMP enriched library. Primary sequences of these clones (cAMP clone 6 and 11) were also submitted to the nucleotide NCBI BLAST database (Fig. 3.4). Unfortunately, cAMP clone 6 bore no primary sequence similarities to any organism within the genomic

Figure 3.3

a)


ATP clone 1

ATP clone 6

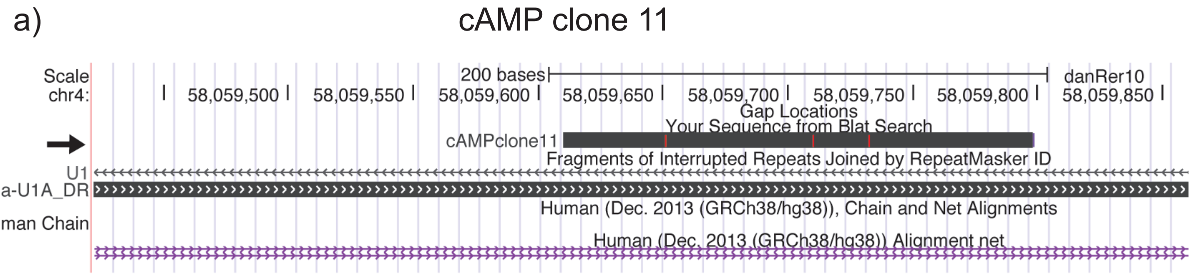
ATP clone 27

b) ATP-agarose column binding assay

Clone	Trial 1	Trial 2	Trial 3
ATP 1	17.40	0.94	14.00
ATP 6	7.50	10.40	0.96
ATP 27	7.47	ND	ND

Figure 3.3 Overview of ATP clones isolated from recombinant DNA and Sanger sequencing. (a) UCSC genome browser sequence alignments along the mouse (*M. musculus*) genome. (b) Each clone was subjected to a column binding assay to measure ATP affinity. Each trial corresponds to a binding experiment performed on different days. Values in the table represent the percent of total radiolabeled RNAs eluted from the ATP column.

Figure 3.4


a)
cAMP clone 11
Scale
chr4:
200 bases
danRer10
58,059,500
58,059,550
58,059,600
58,059,650
58,059,700
58,059,750
58,059,800
58,059,850
Gap Locations
Your Sequence from Blat Search
cAMPclone11
Fragments of Interrupted Repeats Joined by RepeatMasker ID
U1
Dada-U1A_DR
Human (Dec. 2013 (GRCh38/hg38)), Chain and Net Alignments
Human Chain
Human (Dec. 2013 (GRCh38/hg38)) Alignment net

b) Binding Assay (Percent Elution)

Clone	Trial 1	Trial 2	Trial 3
cAMP 6	9.40	0.96	1.88
cAMP 11	0.90	1.4	ND

Figure 3.4. A cAMP clone identified by vector cloning followed by Sanger sequencing. A UCSC Genome Browser BLAT alignment of cAMP clone 11 along the zebrafish (*D.rerio*) genome is shown in (a). cAMP clone 11 was found to overlap a DNA repeat element (Dada-U1A_DR) and is found along multiple locations along chromosome 4. The alignment map shown here illustrates one case in which the cAMP clone 11 overlaps antisense to a known U1 snRNA. The red dashes along the solid cAMP clone 11 sequence indicate alignment mismatches. The table (b) consists of percent elution values from a series of binding assays performed on cAMP clones isolated from the cAMP enriched library. No robust cAMP-agarose column binding for cAMP clone 11 was detected over experimental triplicates. ND = No signal detected.

database, including the noncomplex/repeat sequence database. Clone 11, however, had partial similarity to a repetitive element within the zebrafish genome. This repetitive element, Dada-U1A_DR, is a transposable element found repeatedly in chromosome 4, and is intermittently found to overlap functional U1 small nuclear RNA (snRNA). Similar to clones isolated from the ATP selection, these cAMP clones had no apparent primary sequence similarity and did not show robust binding to cAMP. For detailed information on these cAMP clones and more, refer to the metagenomic deep sequencing data in Chapter 4.

The lack of robust binding of isolated sequences from both parallel *in vitro* selections was surprising. After multiple experimental replicates, it was apparent that these sequences do not demonstrate affinity for either ATP or cAMP. Rather than to continue to isolate and study sequences by vector cloning, the ATP and cAMP enriched libraries were prepared for deep sequencing and Apta-seq [50]. By obtaining deep sequencing data, RNAs within each library can be clustered and rationally organized by computational algorithms to help determine RNA molecule enrichment. In complement with Apta-seq, a technique that combines deep sequencing with structure probing of RNA, structural information of these RNAs can also be extracted as a function of ATP concentration. Doing so would further characterize these RNA sequences into predefined families and can aide in resolving the difference between nonspecific ‘free-rider’ sequences and potential RNA aptamer candidates. The experimental details and subsequent data generated from these deep sequencing and structure probing approaches for the ATP and cAMP enriched libraries is covered in Chapter 4.

IV. *In vitro* selection of RNA aptamers by co-transcriptional binding

Vu et.al. previously observed that the human *FGD3* adenosine aptamer only bound ATP co-transcriptionally [13]. This observation supports the argument that ligand binding by the aptamer is dominated by a kinetically-driven mechanism. In an attempt to directly select these aptamers from the metagenomic library, a co-transcriptional assay was developed. To perform this assay, the agarose-linked metabolite was presented to the RNA library during transcription followed by immediate fractionation and gel purification. RNAs enriched with the co-transcriptional assay would be expected to adopt a transient metabolite-binding conformation that may otherwise be disrupted by denaturation and purification steps mentioned previously. The series of washes and elutions are relatively similar; RNAs in the ligand-binding conformation are subjected to a series of quick washes to remove nonspecific binders and collected after elution from the immobilized target by free-ligand exchange.

Similar to the purified RNA *in vitro* selection assay, radiolabeled RNA molecules were generated to monitor sequence enrichment throughout rounds of selection. To accurately quantify RNAs eluted by free-ligand exchange, each fraction was loaded onto a denaturing acrylamide gel for visualization by autoradiography. Each fraction, or lane, corresponding to eluted RNAs of interest were excised, eluted and prepared for RT-PCR to regenerate the DNA library for the next round of transcription and selection. After several sounds of selection, the fraction of eluted RNAs was quantified by autoradiography and are reported in Figure 3.2. For all co-transcriptional selection rounds, no sample showed a significant increase in RNA binders across all parallel selections. This observation is evident by the fact that for every completed round, we recovered only 1 % of the radiolabeled RNA population for RT-PCR. While this result can be expected for rounds 1 and 2, subsequent rounds were expected to demonstrate an increase in binding as a result of enrichment of target-specific RNA aptamers. The lack of RNA binding was perplexing because each selection demonstrated successful recovery of eluted RNAs, as verified by RT-PCR. However, upon transcription of these recovered libraries, no binding was detected even after multiple experimental attempts to observe an elution fraction of more than 1 %. These efforts were set aside to further investigate RNA aptamers recovered from the parallel (purified RNA) *in vitro* selections.

V. Conclusion

In vitro selection is a method that has previously been used to successfully identify RNA aptamers from synthetic and genomic libraries. This technique was adapted in this study to select for ATP, cAMP and cGMP specific RNA aptamers from a diverse eukaryotic metagenomic pool. To do this, two *in vitro* selection schemes were designed to isolate naturally occurring genomic aptamers. The purified-RNA approach was designed to isolate thermodynamically stable RNA aptamers by first purifying and refolding RNA molecules prior to binding, while the co-transcriptional approach was designed to immediately subject RNA aptamers in the presence of the ligand. In both cases, the fraction of eluted RNAs were measured by radiolabeling RNA molecules and observing an increase in binding affinity after each subsequent round of selection.

The co-transcriptional selection method was designed to select for naturally existing aptamers that undergo a kinetic mode of ligand binding as has been previously observed with the *FGD3* aptamer [13]. After multiple selection rounds, including rounds designed to optimize and the method of detection, no significant binding and elution was detected. This result however does not indicate the absence of naturally occurring aptamers that bind co-transcriptionally, but rather demonstrate the challenges faced during experimental setup. For example, one immediate issue was the presence of readily available free ATP and immobilized ATP present during *in vitro* transcription. During selection, it was periodically observed that the immobilized ATP-agarose beads remained highly radioactive even after fractionation and denaturing washes. This result suggests that a covalent association between the synthesized RNA molecules (assuming successful incorporation of the radionucleotide) and the ATP-agarose matrix occurred thereby removing ligand-specific aptamers. It is unclear whether immobilized ATP is incorporated into the RNA backbone, although this is likely not the case due to agarose-linkage groups which may inhibit this from occurring. Despite a number of attempts to optimize free ATP concentrations to prevent this from occurring, these attempts resulted in either decreased RNA yields due to low ATP levels or only intensified the radioactive signal on the ATP-agarose matrix. Since not any one optimization attempt seemed to improve the number of RNAs eluted from the column, the co-transcriptional binding assays and the respective libraries were frozen and stored for study at a later date.

In contrast to the co-transcriptional selections, *in vitro* selection rounds using purified RNAs displayed an increase in ligand-specific binding. By round 5, both ATP and cAMP selections showed significant elution of bound RNAs upon the addition of free ligand. Sequences isolated from ATP round 5 were found to align with the mouse or zebrafish genomes but had no apparent specificity for ATP as demonstrated by *in vitro* binding assays. Similarly, cAMP clones also mapped to the zebrafish genome and also showed no specificity for cAMP after *in vitro* binding assays. The lack of ATP and cAMP specificity from both sets of isolated clones was unexpected. Sequences isolated here were expected to bind to their respective ligands based on the assumption that dominant sequences within these enriched libraries represent a large fraction of ligand-binding RNAs. If this is true, then there is a high statistical chance that sampling a small subset of individual sequences would reveal aptamers capable of binding ATP or cAMP. It is important to note that this assumption implies that any aptamer recovered by this strategy exists as a mutually exclusive sequence and does not strongly interact with other potential RNAs within the library.

The recovery of a low-complexity repeats (ATP clone 1 and cAMP clone 11) suggest that these sequences are somehow associated with RNAs that may have affinity for ATP and is a selection ‘free-rider’. For additional evidence and discussion on this matter, refer to the rRNA sequence data in chapter 4. Altogether, data reported here in this chapter are insufficient to draw a convincing conclusion to validate successful enrichment of these cloned RNA sequences. It more apparent, however, that the naturally abundant repetitive elements in genome sequence space may obstruct selective pressures to a greater extent than previously thought. In Chapter 4, deep sequencing data highlights how the survival of specific repetitive elements observed here may have arisen with both the ATP and cAMP selections.

CHAPTER 4

Deep sequencing and structural analysis of enriched DNA libraries

I. Introduction

Previous chapters of this thesis focused on experimental design of an *in vitro* selection to identify and isolate ligand-specific RNA aptamers from a diverse metagenomic library. After a few rounds of selection, the most prevalent sequences derived from two enriched metagenomic libraries containing ATP and cAMP RNA aptamers were briefly studied by Sanger sequencing. The aim for the experiments in this chapter focus on systematically analyzing these populations by deep sequencing and study bona fide RNA aptamers by testing for binding affinity and identifying distinct structural features by molecular probing. Presented here are confounding data that challenge the idea of naturally occurring RNA aptamers generated by using state-of-the-art DNA sequencing technology, and biochemical probing.

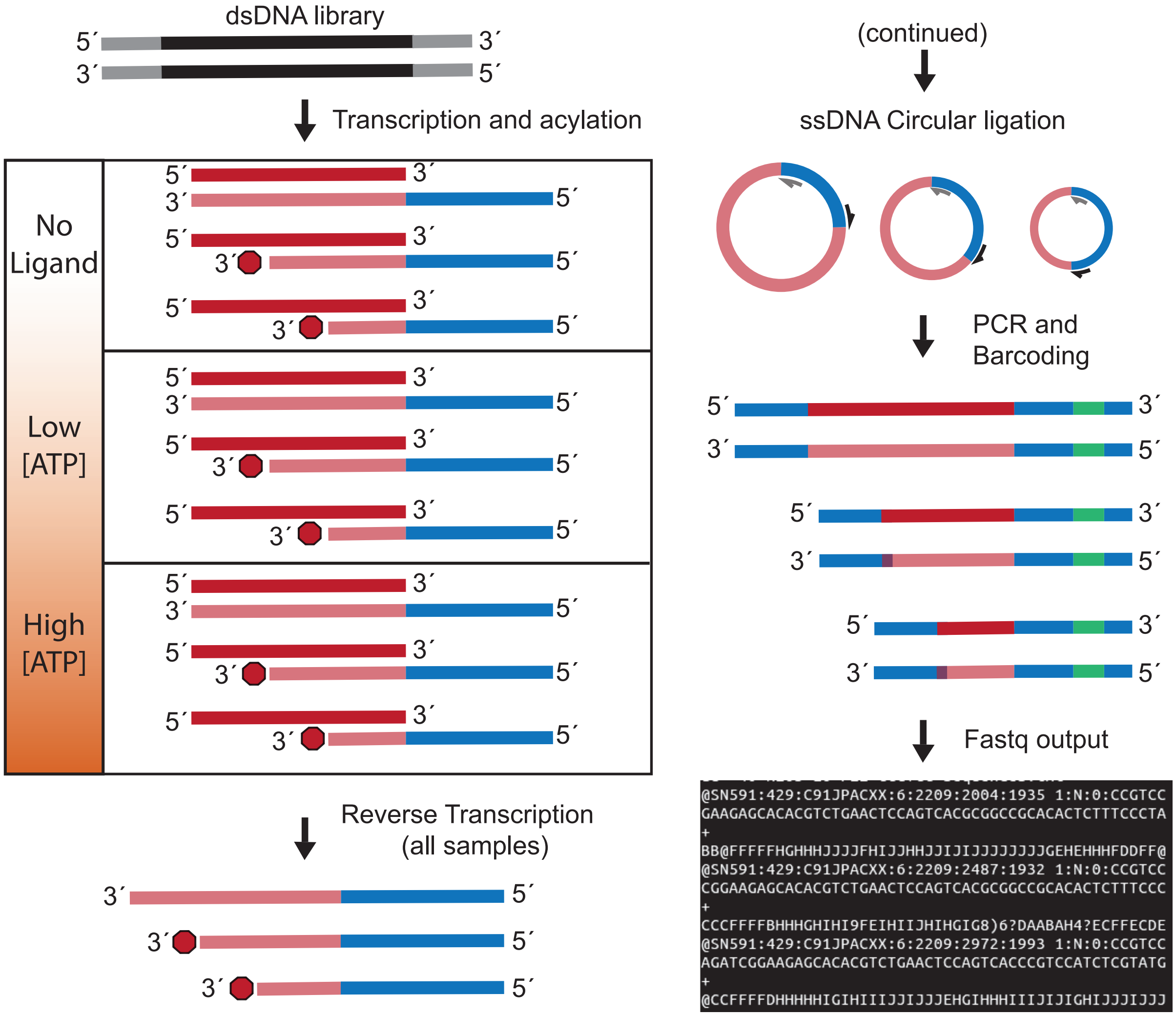
Traditionally, denaturing gel electrophoresis, vector cloning coupled with molecular probing have been used to study RNA aptamers [51]–[53]. While this traditional approach has worked in the past to characterize aptamers like the adenosine aptamer, it is limited by the number of RNAs that can be studied simultaneously. This is a particularly large limiting factor when dealing with large DNA and RNA libraries scaled at 1 x10^{10} sequences. Sampling of these unique sequences in a given library by traditional vector cloning methods is restricted by the high probability of cloning the most dominant sequence within the population. In other words, a statistically large sample size is required to confidently observe alternative lower-abundance populations that may also represent a valid solution to the aptamer selection. The ability to observe alternative lower-abundance sequence families provides a unique opportunity to study population genetics and therefore common features that may be required for survival.

To achieve this goal, the Illumina HiSeq deep sequencer was used to measure these sequence families that are virtually undetectable by traditional vector cloning and sanger sequencing. To further capitalize on the deep sequencing platform, a molecular probing method was also incorporated during preparation of the RNA library for sequencing. Since the metagenomic DNA library was designed with Illumina-compliant adapters, preparing any metagenomic sample for deep sequencing required an additional two-step PCR procedure to introduce barcodes. Chemically-probed RNA libraries were reverse transcribed, self-ligated and amplified, as illustrated in (Fig. 4.1), to generate an Illumina-compliant cDNA library for deep sequencing. Altogether, the purpose of this approach was to characterize dominant sequence families resulting from an *in vitro* selection of RNA aptamers present within the ATP and cAMP enriched libraries, and to extract structural data accompanied by these sequences using selective hydroxyl acylation and analysis by primer extension (SHAPE).

SHAPE probing followed by deep sequencing (SHAPE-seq) is a relatively new method that takes advantage of the high throughput sequencing capacity of the Illumina HiSeq instrument to report chemical reactivities of an RNA molecule that can then be used to shed light on structural features at single nucleotide resolution. A detailed illustration of the experimental steps taken for SHAPE-Seq is provided in Figure 4.1. Initially, flexible 2′ hydroxyl groups along an RNA molecule are acylated by an electrophilic reagent, or SHAPE reagent. Selective modification of an RNA molecule is highly correlated to 2′ hydroxyl availability to the SHAPE reagent as a result of local

Figure 4.1

SHAPE - sequencing flowchart



```
@SN591:429:C91JPACXX:6:2209:2004:1935 1:N:0:CCGTCC
GAAGAGCACACGTCTGAACTCCAGTCACGCGGCCGCACACTCTTTCCCTA
+
BB@FFFFFHGHHHJJJJFHIJJHHJJIJIJJJJJJJJGEHEHHHFDDFF@
@SN591:429:C91JPACXX:6:2209:2487:1932 1:N:0:CCGTCC
CGGAAGAGCACACGTCTGAACTCCAGTCACGCGGCCGCACACTCTTTCCC
+
CCCFFFFBHHHGHIHI9FEIHIIJHIHGIG8)6?DAABAH4?ECFFECDE
@SN591:429:C91JPACXX:6:2209:2972:1993 1:N:0:CCGTCC
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGTCCATCTCGTATG
+
@CCFFFFDHHHHHIGIHIIIJJIJJJEHGIHHHIIIJIJIGHIJJJIJJJ
```

Figure 4.1 A flowchart illustrating the experimental steps taken to generate a SHAPE-seq cDNA library for *in vitro* selected RNA libraries. First, full length RNAs are transcribed and purified by gel electrophoresis. A semi-logarithmic titration from low to high concentration of ligand is then prealiquoted into separate tubes (represented by the solid red lines; boxes indicate one sample along the titration). The RNA library is then equally divided into these tubes and chemically modified by a SHAPE reagent (red octagon). Reverse transcription then extends a SHAPE-RT primer (blue) up to the modification site, generating an RT stop profile for all RNA molecules in the library. For every sample along the titration, a single-stranded DNA Circular ligation and PCR step generates illumina compliant double-stranded DNA. A single end sequencing procedure generates a fastq file containing the sense (top) strand where the first nucleotide sequenced corresponds to the preceeding nucleotide that was chemically modified.

hydrogen bonding networks, electrostatic interactions and structural restraints determined by nearest neighbor interactions. These nucleotides prone to modification generally include single-stranded regions and highly flexible nucleotides [52]. In turn, the acylation event generates RNA nucleotide adducts that cannot be successfully read through by natural reverse transcriptases. This property is exploited by generating cDNAs that are extended up to the nucleotide immediately preceding the modified nucleotide. After cDNA synthesis, these reverse transcription stops (RT-stops) are measured by subsequent PCR amplification and gel electrophoresis [51]–[53]. It is important to note that these SHAPE cDNA libraries have previously been shown to be highly reproducible on deep sequencing platforms, so long as the rate of modification per RNA molecule is proportional to avoid null or saturated RNA probing that inhibits structural characterization [50], [54]–[57].

A given acylation event along an RNA molecule is dependent on the local structural arrangement that may make the ribose sugar accessible in three-dimensional space. Because of this, these sites may become more or less reactive upon undergoing a conformational change that is required for ligand binding. With this in mind, the ATP and cAMP enriched RNA libraries were independently probed along a semi-logarithmic titration for their respective ligands from 1 μM – 10 mM and 1 nM to 1 μM, respectively. Each sample, or point along the titration, is then prepared for deep sequencing by first being converted into Illumina-complaint dsDNA libraries using a unique RT primer that includes an Illumina adapter 5′ overhang, represented in blue in Figure 4.1. Each cDNA sample is then separately subjected to a single stranded DNA circular ligation step and subsequently amplified using Illumina PCR amplification primers. Each sample is then barcoded with a unique reverse Illumina flowcell primer for reference during file processing using bioinformatics software.

After obtaining sequencing data, computational methods were utilized to trim, count, align, and process sequences to bin them into related families based on primary sequence. Algorithms and procedures to initially process these sequences (including trimming and alignment against genome sequence databases) were taken from published approaches that have demonstrated successful transcriptomic analysis of RNA sequencing data [58]. While a number of these software packages exist for analyzing such RNA sequence data, these packages are ultimately tailored to specific datasets to process sequenced total RNA and mRNA sequences, some of which are not required here [58]–[61]. For this reason, only specific modules from these software packages were used to assemble a final software pipeline needed to process metagenomic sequencing data generated in this study. A flowchart consisting of the software programs used to process deep sequencing data and SHAPE-seq data is shown in Figure 4.2. Note that a significant portion of these modules and their manuals are readily available online and can be easily accessed for local execution within a Linux/UNIX shell environment. Furthermore, because some steps required subtle procedures such as sorting by nucleotide length, extraction of RT stop counts, and reformatting/transforming sequence coordinate data into data matrices, in-house programs were written and assembled to generate the SHAPE reactivity data presented below. Further details pertaining to differences between deep sequencing and SHAPE-seq software pipelines uses can be found in the computational methods section.

Figure 4.2

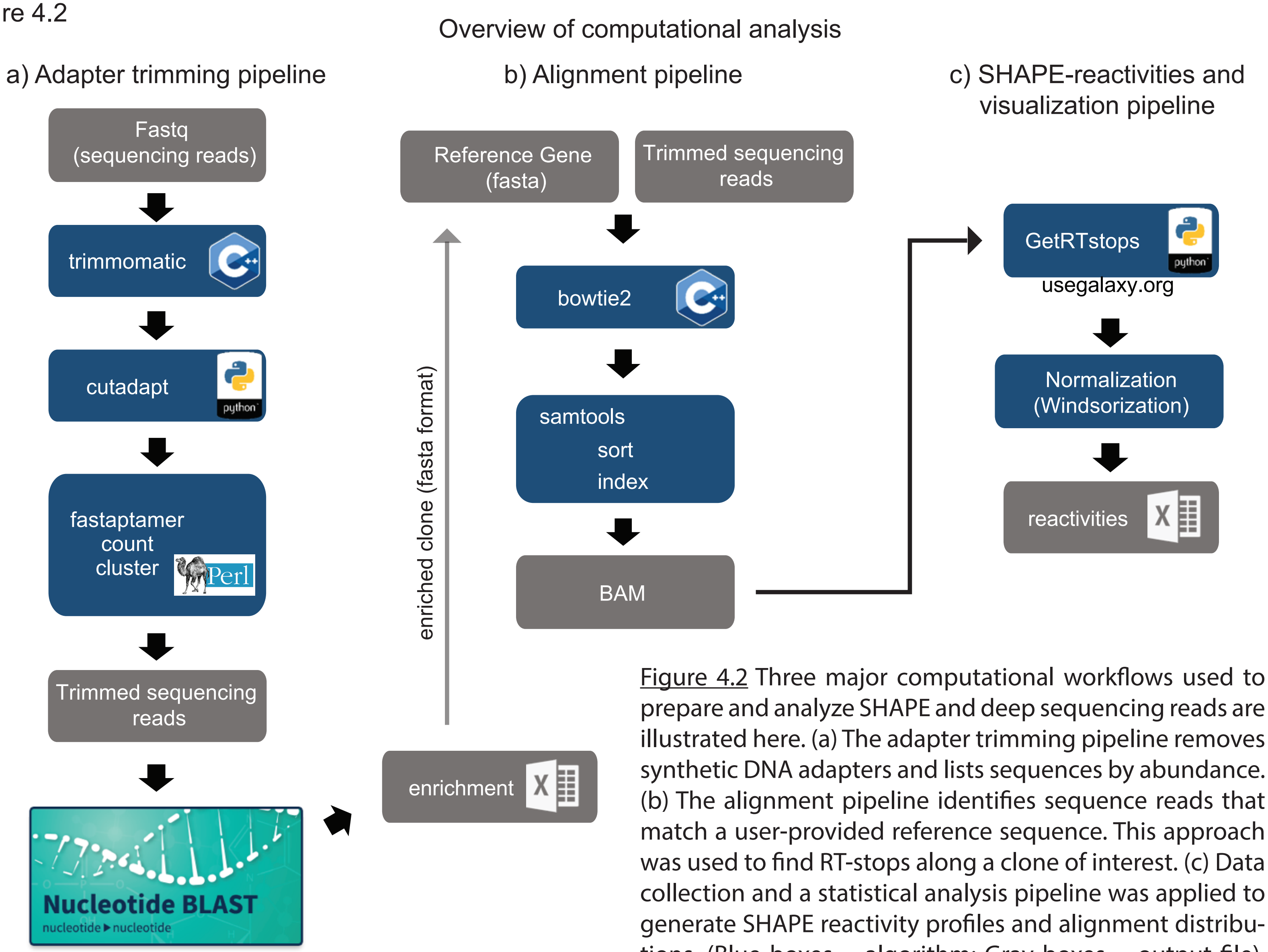


Figure 4.2 Three major computational workflows used to prepare and analyze SHAPE and deep sequencing reads are illustrated here. (a) The adapter trimming pipeline removes synthetic DNA adapters and lists sequences by abundance. (b) The alignment pipeline identifies sequence reads that match a user-provided reference sequence. This approach was used to find RT-stops along a clone of interest. (c) Data collection and a statistical analysis pipeline was applied to generate SHAPE reactivity profiles and alignment distributions. (Blue boxes = algorithm; Gray boxes = output file).

II. Deep sequencing of the metagenomic libraries

Files received from deep sequencing of the naïve and enriched DNA libraries were trimmed and processed by the trimmomatic and cutadapt programs to remove adapter sequences and reads less than 24 nucleotides in length. Each paired-end deep sequencing dataset arrived as two files corresponding to sense (read 1) and antisense (read 2) strands. Input reads, surviving reads, and percent survival after each processing step for each of these files are reported in Table 4.1. Note that the percent survival values among each paired end sample set are consistent. This implies successful barcode assignment for a given sequencing cluster identified on the flowcell by the illumina HiSeq 2500 sequencer. Reads that were unsuccessfully assigned a barcode are ignored and discarded to avoid assigning sequences that may have originated from other sequencing samples.

It is important to note, that although each metagenomic library submitted is 300 bps in length, all DNA deep sequencing reads obtained here are only 150 nucleotide reads. Furthermore, no sequencing assembly software was used to regenerate 300 bp reads. Efforts to assemble full 300 bp sequences using such software was inefficient and unreliable due to the sophisticated sequence-pattern recognition algorithm that ultimately generated sequences that did not match raw sequence reads provided by the Illumina sequencer. For this reason, for all deep sequencing data, read 1 and read 2 were processed individually and treated as separate sequencing jobs.

After trimming of adapter sequences, each sequence file for each dataset was processed by the fastapatmer-count algorithm to generate a ranked list of sequences in fasta format. Unlike the original fastq file, this counted fasta file consolidates sequence duplicates and reports them as a list of unique sequence reads in the order of abundance.

Table 4.1

Percent survival of trimmed input reads

Experiment	Input Reads	Timmomatic	Cutadapt	Surviving Reads	%
naïve DNA library - Read 1	18584326	140381	2511580	15937057	85.8%
naïve DNA library - Read 2	18584326	107937	2866875	15614327	84.0%
ATP-enriched library - Read 1	11702953	548413	1018338	8424378	72.0%
ATP-enriched library - Read 2	11702953	498912	1068563	8404077	71.8%
cAMP-enriched library - Read 1	19510453	270698	1475575	17767510	91.1%
cAMP-enriched library - Read 2	19510453	218412	1902158	17393337	89.1%
ATP - No SHAPE	3984503	1105983	183051	2697711	67.7%
ATP - No Ligand	1625491	1284109	27575	314173	19.3%
ATP - 1 µM	11497221	1603027	682074	3381152	29.4%
ATP - 3 µM	3205475	2743085	33453	429403	13.4%
ATP - 10 µM	7374528	6430584	285740	714309	9.7%
ATP - 30 µM	13032385	10011061	632724	2392623	18.4%
ATP - 100 µM	19637030	16800269	665764	2176088	11.1%
ATP - 300 µM	10222642	6931357	963421	2333549	22.8%
ATP - 1 mM	10917498	8483296	325428	2156292	19.8%
ATP - 3 mM	9094388	6232528	330192	2571733	28.3%
ATP - 10 mM	14827444	971917	1005953	12857891	86.7%
cAMP - No SHAPE	266	241	12	229	86.1%
cAMP - No Ligand	106	80	2	78	73.6%
cAMP - 1 nM	423	169	8	161	38.1%
cAMP - 10 nM	209	198	3	195	93.3%
cAMP - 100 nM	123	100	9	91	74.0%
cAMP - 1 µM	80	73	4	69	86.3%
cAMP - 10 µM	95	93	8	85	89.5%
cAMP - 100 µM	44252	2656	116	2540	5.7%
cAMP - 1 mM	149	133	9	124	83.2%

These reads are then renamed as a series of three numbers as the sequence header where each number represents an abundance rank number, the number of times the sequence appeared within the file, and a normalized read per million (RPM) value, respectively.

One immediate observation was the large discrepancy in RPM values assigned to top ranking sequences between the naïve and both enriched metagenomic libraries. Top ranking sequences within the naïve metagenomic library included highly repetitive tandem repeats such as telomeric, centromeric, (CA)n, (GT)n, and DNA satellite repeats but exhibited low RPM values suggesting that these simple DNA repeats are only a small fraction of total unique reads sequenced. Nonetheless, it should be taken into consideration that these simple sequence repeats do represent a significant population bias within the metagenomic library. Interestingly, these repeats were also found in both the ATP and cAMP enriched libraries at low abundance and low RPM values. This promising observation indicates successful selective pressure against low-complexity sequences present within the naïve metagenomic library.

To study deep sequencing files corresponding to the ATP and cAMP enriched libraries, each dataset was further processed by the fastaptamer_cluster module to distribute reads related by primary sequence. This allowed for counted sequence reads to be binned into sequence families that differ by more than 10 % or 15 nucleotides. These families are then arbitrarily ranked and individually written into a new fastaptamer_cluster fasta file. In addition, family rank number and edit distances from the family seed sequence is appended to the existing header. Since this is a computationally intensive process, this procedure was not performed with high sensitivity for the naïve metagenomic sequencing dataset due to the large diversity of sequences present in that file. Instead, the naïve library fastaptamer-count file was used as a reference to determine the level of enrichment for a given sequence taken from the fastaptamer-cluster file of the enriched libraries. Enrichment was calculated by taking the ratio of RPM values for a given sequence present within the respective enriched and naïve sequencing file.

$$\text{Sequence enrichment} = \text{RPM}_{\text{enriched}} / \text{RPM}_{\text{naïve}}$$

Finally, these cluster-ranked sequences were individually threaded into genomic alignment algorithms (BLAST and UCSC Genome Browser) to determine their genomic origin using publicly available genome databases. Any significant features, coordinates and annotated elements found near these mapped sequences were recorded and studied.

To organize these data, a list of these significantly enriched sequences from both the ATP and cAMP selection libraries were divided and defined into the following three groups; exclusive matches, repeat matches, and no alignment (NA). An exclusive match was defined as a sequence with a single genomic coordinate and high sequence match within a specific organism. A repeat match is defined as a sequence that is found to have multiple genomic coordinates usually due to association with a mobile genetic element but can typically be identified within a specific organism. Note that this category also includes simple tandem repeat sequences. Finally, NA matches are defined as complex or simple tandem repeat sequences that have an unknown genomic origin as determined by alignment tools. For reference, all sequences classified within these groups showed significant enrichment within their respective libraries and are potential candidates for binding affinity assays. Further details including sequence attributes, genomic coordinates, and sequence were tabulated and are explained below.

Enriched sequences from the ATP-selection were predominantly identified as repetitive elements and categorized as repeat matches (Table 4.2). Note, that although a number of these sequences were classified within the ‘exclusive-match’ group, the vast majority either flank or partially overlap known repetitive elements. High sequence similarity and specific genomic coordinates indicated for the ‘exclusive-match’ set of sequences are what distinguishes them from the Repeat and NA groups. Sequences within the Repeat group are simple tandem DNA repeats and were found to successfully align across multiple loci and even throughout multiple organisms in some cases. This result made it difficult to further characterize and study these sequences and were therefore unheeded. In contrast, sequences in the NA group exhibited a mixture of characteristics that made it difficult to categorize as either exclusive or repetitive. For example, most of the NA-group sequences show partial alignment (<50 %) to specific genetic elements including mouse *scf1a1* and mouse ERVK using UCSC genome browser but are completely unrecognized using NCBI BLAST. This may be largely due to either incomplete assembly of online genomic databases (unlikely in the case for mouse) or may be an example of sequence divergence within the enriched population. In either case, this detail remains to be solved.

Clustered sequences from the cAMP-selection were also categorized into groups as mentioned above (Table 4.3). BLAST and UCSC genome browser alignments showed that a large fraction of these enriched sequences map to fixed loci within specific organisms including mouse, zebrafish and green algae. Interestingly, these alignment

Table 4.2

Enriched sequences from the ATP aptamer *in vitro* selection

a) Exclusive matches

Sequence (Header)	Enrichment	Organism	Coordinates	Strand	Notes
167-325-38.58-2	59.35	*M. musculus*	chrUn_GL456239		Some conservation in chicken, platypus and zebrafish.
327-233-27.66-5	230.5	*M. musculus*	chr16:3,235,156		Simple repeat (CTGTG)n.
336-228-27.06-6	41.63	*M. musculus*	chr10:3,111,355		Low vertebrate conservation. LTR regions flank matching sequence.
558-167-19.82-13	82.58	*M. musculus*	chr7:64,087,469		Low vertebrate conservation. Multiple alignmentsfound within the mouse genome.
592-160-18.99-15	35.83	*M. musculus*	chr7:64,084,339		Simple repeat (CTGTG)n.
616-157-18.64-16	64.28	*M. musculus*	chr4:47,442,010	sense	First 60 nts overlaps RLTR20A4 ERVK (LTR)
714-142-16.86-19	140.5	*M. musculus*	chromosome 4		Multiple matches along chromosome 4
763-135-16.02-21	89	*M. musculus*	chr2:131,723,327		Some conservation found in rats and zebrafish. Overlaps a simple tandem repeat
867-123-14.60-28	81.11	*M. musculus*	chromosome 7		Multiple matches along chromosome. Simple repeat (CTGTG)n.
914-119-14.13-29	58.88	*M. musculus*	chr15:36,223,504	antisense	Mapped within intron 14 of spag1 with weak conservation in mammals.
1045-107-12.70-30	52.92	*M. musculus*	chr4:98,584,579	sense	Mapped within an intron of *inad1* . Match overlaps a SINE B4 element.
1210-97-11.51-37	191.83	*M. musculus*	chr5:113,113,638	sense	Maps to an intron of KIAA1671 mRNA. Large overlap with RLTR7B ERVK (LTR)
1353-90-10.68-44	89	*M. musculus*	chr17:5,866,712	antisense	Match with intron 1 of *snx9* . Conservation in vertebrates.

Table 4.2 (cont'd)

b) Repeat matches

Sequence (Header)	Enrichment	Repeat Type	Strand	Notes
327-233-27.66-5	230.5	Simple Repeat (CTGTG)n	antisense	Sequence is related to 632-153-18.16-17
558-167-19.82-13	82.58	Simple Repeat (TACTG)n	sense	
592-160-18.99-15	35.83	Simple Repeat (CTGTG)n	sense	
616-157-18.64-16	64.28	ERVK (LTR)	sense	Multiple mouse-specific ERVKs. Annotated LTRs are highly divergent.
746-138-16.38-20	140.5	Simple Repeat (CA)n	sense	
770-134-15.91-23	23.06	Simple Repeat (CACTG)n	antisense	
1045-107-12.70-30	52.92	SINE (B4)	sense	
1145-100-11.87-33	NA	Simple Repeat (ACATG)n	antisense	
1253-95-11.28-38	27.51	SINE (B4)	sense	
1316-92-10.92-43	60.67	ERVK (LTR)	sense	Mouse-specific ERVK. This sequence is highly related to 1210-97-11.51-37

Table 4.2 (cont'd)

c) No alignment

Sequence (Header)	Enrichment	Strand	Notes
117-407-48.31-1	166.59	antisense	Intron 4 of *scf1a1* in mouse. Unannotated repetitive element. Very little sequence conservation in vertebrates
187-308-36.56-3	126.07		
255-266-31.58-4	59.58	sense	Simple Tandem Repeat (ATCTG)n. 90 % identity to Intron 68 of *ryr* , cardiac mRNA in zebrafish.
347-224-26.59-7	37.45	sense	Simple Repeat (CAGAC)n. Overlaps an ERVK (LTR) element found in mice.
431-194-23.03-8	46.06		Very low sequence identity to *Plasmodium ovale* genome.
540-170-20.18-10	336.33		
550-168-19.94-12	36.56		Simple Repeat (TGAG)n.
632-153-18.16-17	302.67	sense	Simple Repeat (TG)n. Mapped redundantly within intron 1 of the *c1qtnf1* gene in mice. Flanked by ERVK (LTR)
714-142-16.86-19	140.5		
777-133-15.79-24	83.11	sense	Simple Repeat (TG)n.
846-125-14.84-27	247.33		Sequence has high similarity to 540-170-20.18-10
1167-99-11.75-34	61.84	sense	Simple Tandem Repeat. Maps within an ERVK (LTR) in mice.
1210-97-11.51-37	191.83	sense	Partial overlap with an ERVK (LTR) found in mice
1295-93-11.04-42	184		
1403-88-10.45-46	174.17		Very little identity with mouse genome. No nearby gene annotation.
1481-85-10.09-48	168.17	sense	Simple tandem repeats present. Overlaps ends of ERVK (LTR). Antisense to *tmem11* 6 gene in mice (intron 1).

Table 4.2 Individual sequences from the ATP-aptamer enriched DNA library and their respective enrichment values are listed above. Following an NCBI BLAST alignment, sequences were categorized as either (a) Exclusive sequences - having high identity to a reference genome, (b) Repetitive matches - repetitive in nature or belonging to an annotated repetitive element, or (c) No alignment - no significant and unique alignment attributes. Sequences within the 'no aligment'

Table 4.3

Enriched sequences from the cAMP aptamer *in vitro* selection

a) Exclusive matches

Sequence (Header)	Enrichment	Organism	Coordinates	Strand	Notes
5-1126-63.37-5	107.41	*M. musculus*	chr17:53,689,651	antisense	100% identity. Unique to mouse. 500 bps downstream of *sgol1.*
9-1055-59.38-8	169.66	*D. rerio*	chr11:7,167	sense	Spans last intron-exon junction of *myg1* gene. High conservation with humans, frogs.
11-936-52.68-9	219.50	*M. musculus*	chr1:85,598,517	sense	Spans first exon-intron junction of *sp110* . Fragmented vertebrate conservation.
14-891-50.15-12	172.93	*M. musculus*	chr12:100,633,405	sense	Overlapping with ERVK (LTR). Sits within of *Rps6ka5* . Weak vertebrate conservation.
27-714-40.19-21	669.83	*C. reinhardtii*	XM_001700120.1	sense	Predicted protein (partial mRNA sequence) in *Chlamydomonas reinhardtii* .
31-623-35.06-24	292.17	*M. musculus*	chr16:3,985,480	antisense	Maps to exon 3 of *slx4* gene. Nucleotide conservations in vertebrates.
34-591-33.26-27	138.58	*D. rerio*	chr8:7,298,409	sense	Maps to exon 1 of cenpa (unique to zebrafish).
36-565-31.80-28	265.00	*C. reinhardtii*	XM_001694233.1	antisense	Similarity to a predicted protein (partial mRNA sequence) in *Chlamydomonas reinhardtii* .
37-549-30.90-29	128.75	*M. musculus*	chr12:11,226,208		No nearby genes. Conservation in vertebrates.
41-527-29.66-33	123.58	*M. musculus*	chr1:91,521,177	sense	Spans 12th exon-intron junction of *traf3ip1.* Found near 3´ end of some ESTs.
63-444-24.99-48	208.25	*M. musculus*	chr13:58,937,464	sense	Spans intron 12 of *ntrk2* . mRNA found 500 bps downstream of 3´ UTRs. Vert. conservation.

Table 4.3 (cont'd)

b) Repeat matches

Sequence (Heade	**Enrichmen**	**Repeat Type**	**Strand**	**Notes**
2-2041-114.87-3	67.57	Simple Tandem Repeat	sense	Low score match with *D. rerio* . Overlaps a Kolobok T2-DR.
29-632-35.57-22	122.66	ubiquitin-protein ligase RBBP6-lik	antisense	Multiple sequence matches overlay RBBP6-like (*D.rerio* -specific) mRNAs
32-619-34.84-25	99.54	None annotated		Multiple matches. *D.rerio* -specific.
33-602-33.88-26	282.33	Simple Tandem Repeat	sense	Gapped repeat ~ 604 bps in *D. rerio* .
38-543-30.56-30	169.78	LTR11 DR	sense	
39-538-30.28-31	252.33	None annotated	antisense	Sits between LTR1 and Copia-6-I (Gypsy) repeat elements.
40-533-30.00-32	166.67	Simple Tandem Repeat		Heavy ungapped 90 nt repeat in *D. rerio* . Found on chr. 1, 6, and 10.
41-527-29.66-33	123.58	DIRS (LTR)	antisense	Heavy number of repeats throughout the zebrafish genome.
42-521-29.32-34	244.33	None annotated		Multiple matches. *D.rerio* -specific.

Table 4.3 (cont'd)

c) No alignment

Sequence (Header)	Enrichment	Strand	Notes
1-2747-154.61-1	69.33	sense	Simple tandem repeat (GTGTGCTGCT)n. in mice. No vertebrate conservation.
2-2041-114.87-2	67.57		No sequence similarity found using NCBI:BLAST
4-1544-86.90-4	147.29		Gapped repeat (triplet of purines)
6-1123-63.21-6	97.25		Simple tandem repeat (TG)n
7-1090-61.35-7	86.41		Simple tandem repeat (TG)n
12-905-50.94-10	124.24		
14-891-50.15-12	172.90		
15-889-50.04-13	50.04		
20-786-44.24-13	126.40		
44-511-28.76-36	479.33		No sequence similarity found using NCBI:BLAST
50-493-27.75-40	95.69		

Table 4.3 (cont'd)

d) Gene ontology analysis

Gene	GO class	Organism(s)	Evidence
sgol1	Regulation of heart contraction. Meiotic chromosome segregation.	Mm, Dr, Sp	Morpholino studies in D.rerio. Inferred from genomic sequence assembly.
myg1	Intracellular molecular function	Hs, Dr, An, Mm	RNA seq. Sequence orthology. *Related to sequence >9-1055-59.38-8-1-0.*
sp110	DNA binding. Metal ion binding.	Hs, Mm, Dr, Xt	Sequence orthology. Inferred from genomic sequence assemblies.
slx4	Slx1 - Slx4 complex, meiotic dsDNA break-processing *, DNA binding **	Hs, Mm, An, Dm, Xt*, Sp*, Sc*, Dr**	Sequence orthology. Inferred from genomic sequence assemblies. Phylogenetic analyses.
cenpa	mitotic spindle orientation/cytokinesis, nucleosomal DNA binding.	Hs, Mm, Xt	Histone pulldown assay. Inferred from genomic sequence assemblies.
traf3ip1	Centrosome, cytoskeleton*, cilium assembly **	Hs, Mm, Xt+** Dr*	Sequence orthology. Some phylogenetic analyses. *Related to sequence >63-444-24.99-48-1-0*
ntrk2	ATP binding, protein tyrosine kinase activity, vasculogenesis *	Hs+*, Mm+*, Dr	Receptor tyrosine kinase pulldown, Immunofluorescence. Sequence orthology.

An -	*Aspergillus nidulans*
Dm -	*Drosophila melanogaster*
Dr -	*Danio rerio*
Hs -	*Homo sapiens*
Mm -	*Mus musculus*
Sc -	*Saccharomyces cerevisiae*
Sp -	*Schizosaccharomyces pombe*
Xt -	*Xenopus tropicalis*

Table 4.2 Sequences from the cAMP enriched DNA library and their respective enrichment values are listed above. After an NCBI BLAST alignment, sequences were categorized as either (a) Exclusive sequences - high identity to a reference genome, (b) Repetitive matches - repetitive or belonging to an annotated repetitive element, or (c) No alignment - no significant or unique sequence patterns. Sequences within the 'no aligment' category generally have weak sequence identity. (d) Sequences along conserved mRNAs were submitted for ontological analysis to determine if expression of these genes include a common cellular pathway.

windows show varied levels of conservation across vertebrate genomes and are either actively processed as mRNAs or belong to known nascent RNAs. For those matches along highly conserved genes, we performed a gene ontological analysis (Table 4.3d) to determine if gene products belong to specific cellular signaling pathways. Gene names and their gene ontology classes are listed in Table 4.3d, along with experimental evidence taken from genetic model organisms. Interestingly, common classes include DNA-binding activity and DNA replication mechanisms including centrosomal and mitotic spindle orientation. These features strongly support evidence in the literature that show that cAMP regulates cell-cycle progression by upregulating DNA binding proteins and adenylyl cyclases [62]–[64].

Sequences classified as repeats were, not surprisingly, found to align redundantly along multiple locations along a given genome with varying similarity scores and were simply disregarded. NA-group sequences were primarily simple tandem repeats of low complexity that could not be mapped and were also disregarded for further study. Altogether, enrichment values for all sequences were significantly larger compared to those seen for the ATP selection. This observation, in addition to the overall lack in the number of simple sequence repeats present within the cAMP selection, greatly contrasts those observations made earlier for the parallel ATP selection. The top four most enriched sequences from the cAMP selection were used to generate *in vitro* RNA clones and used as a reference to produce SHAPE profiles from Apta-Seq data.

To generate clones from these sequencing data for Apta-Seq and *in vitro* binding analysis, the sequences from each selection were manually assembled using the read 2 deep sequencing files. Briefly, 3′ ends of read 1 sequences were used to find overlapping complementary 3′ ends of at least 30 base pairs from the read 2 (antisense) sequence file using a text-based search algorithm. Once identified, the 3′ end of read1 was in-silico extended using the matching read 2 sequence as a template to generate full sequences. For reference, each sequence that underwent this process is named starting with the enriched library from which it was derived, and the corresponding enrichment value (i.e. ATP 336.3). Complete sequences were then re-examined to confirm genomic origin and subsequently used as a reference to generate SHAPE-reactivity data and tested for binding activity for their respective target ligands by *in vitro* column binding.

III. SHAPE profiles and *in vitro* binding data

Prior to generating reactivity plots for highly enriched sequences, SHAPE-sequencing reads were trimmed and processed by the trimmomatic and cutadapt programs. In total, 1.05 x10^8 reads were acquired for the ATP Apta-Seq dataset, while 4.57 x10^4 reads were recovered from the cAMP Apta-Seq dataset. Once adapters and short reads were omitted, roughly 30 % and 8 % of these ATP and cAMP reads were suitable for study, respectively. This low yield of processed reads was somewhat expected because of the necessary steps that were taken to remove self-ligated single stranded DNAs during library preparation; however, the very low recovery of cAMP reads was unanticipated. Due to the design of this experiment, RT stops generated by this primer extension reaction do not generate fully extended cDNAs and therefore amplify as a distribution of sequences from 70 - 300 bps. Although steps were taken to carefully remove these self-ligated 70 bp artificial byproducts by gel excision, this step must be performed for every sample (including other SHAPE-sequencing pools that were sequenced in parallel) and therefore increases the likelihood of contamination. This fact may also explain the survival percentage discrepancies observed between each titration point along both ATP and cAMP SHAPE-sequencing samples. Unfortunately, due to the extremely low number of reads recovered for the cAMP dataset, no sufficient data was available to generate reliable SHAPE profiles for cAMP selection sequences. Instead, SHAPE profiles presented below were generated for highly enriched ATP selection sequences.

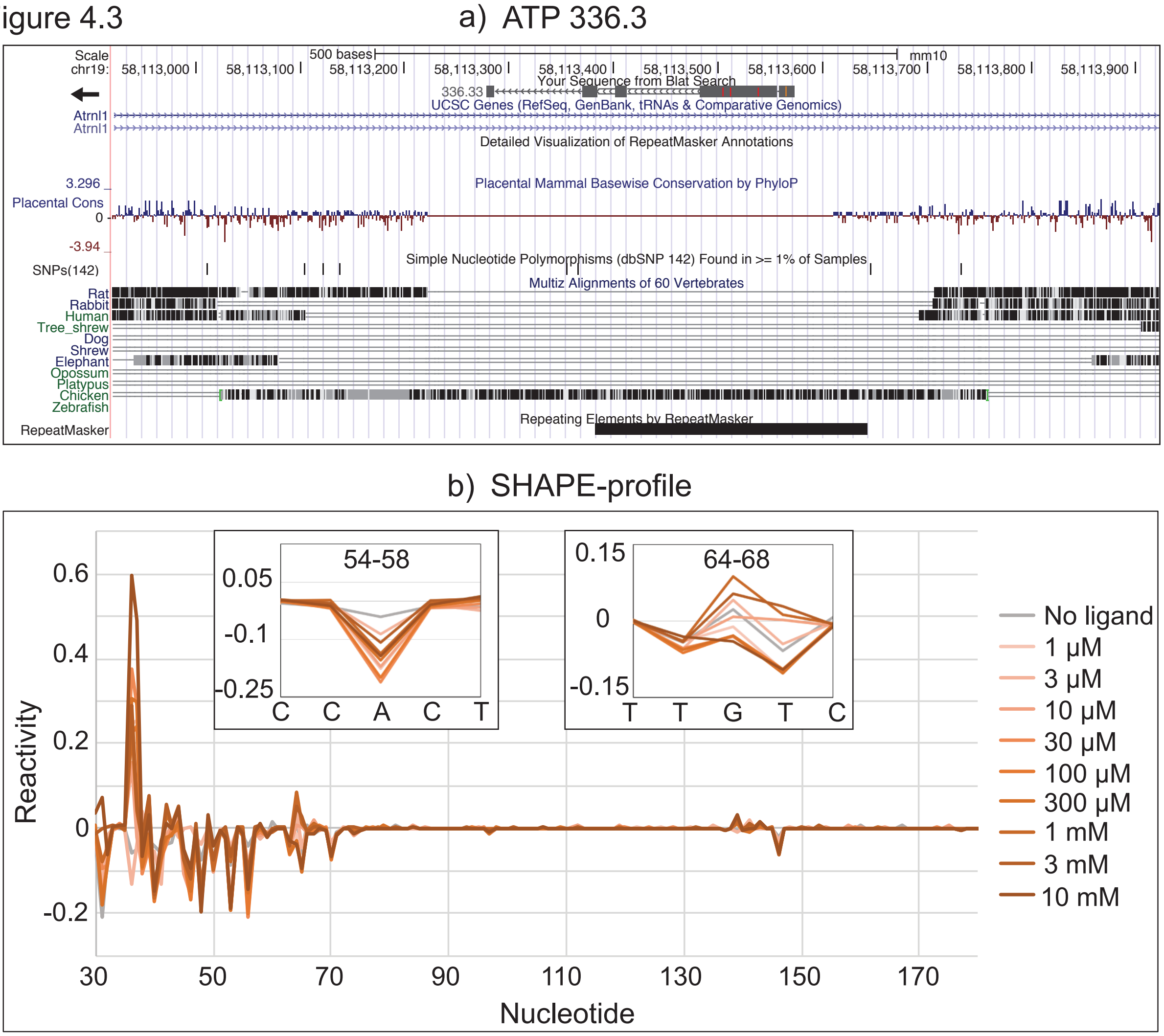
Apta-Seq reads were aligned to an ATP clone reference sequence on a local UNIX environment using the bowtie2 alignment algorithm to map RT stops (5′ ends of reads) along the sequence at nucleotide resolution. RT stop count, map quality scores, and nucleotide position are written in binary alignment map files generated by the bowtie2 and Get RT stops program. These files were further processed by additional python software modules needed to extract RT stop counts with respect to nucleotide positions and to assemble spreadsheets to generate SHAPE profiles. This process was repeated for every sequence file corresponding to each concentration of ATP tested. Once assembled in an excel spreadsheet, these data were normalized and rescaled to generate reactivity profiles for each candidate aptamer (refer to Fig. 4.2 and methods in Chapter 6 for a detailed normalization procedure). For simplicity, ATP and cAMP clones discussed below are named after their respective enrichment value instead of lengthy sequence headers (Table 4.2 - 4.3).

i. ATP clone – 336.3

Data taken from deep sequencing revealed that ATP clone 336.3 is the most abundantly enriched clone present within the ATP enriched library. Surprisingly, this clone was found to have no significant similarity to any known genomic reference sequence. In fact, sequence-cluster data show that two of the top five most enriched families, including ATP 336.3 and 247.33, do not have any similarity to known genomic reference sequences. This observation was further supported by the fact that the seed sequence corresponding to the ATP clone 336.3 cluster, was not immediately identified within the naïve metagenomic library. Instead, a related sequence containing a few point mutations was used as a proxy to estimate the enrichment for ATP clone 336.3. Interestingly, this related sequence showed a weak alignment match with the mouse genome suggesting that this sequence may have arisen from a mouse ERV repeat element (Fig. 4.3a). If this is the case, ATP 336.3 contains repetitive elements that complicates any effort attempting to determine the genomic origin. Furthermore, not only are ERV elements common in rodents and mammals, but their ability to transpose across a genomic landscape makes it difficult to narrow down and even assemble ERV genes due to their highly repetitive qualities.

Despite this difficulty in sequence alignment, a SHAPE reactivity profile was generated from Apta-Seq data using the most common (seed) sequence for ATP 336.3. Normalized stops at every concentration of ATP were rescaled and plotted as a function of reactivity scores based on accessibility by acylation versus nucleotide position (Fig. 4.3b). Using this approach, a large number of fraction of modified nucleotides mapped to the 5′ end of the clone. This means that a large number of RT stops were heavily detected near the 5′ end, and very little stops near the 3′ end, suggesting that this profile is heavily biased despite having already corrected for this (see SHAPE methods). This observation may be a result of partial similarities to the mouse ERV repeat element because 1) figure 4.3a shows partial alignment to an ERV masker 2) the SHAPE-seq alignment procedure

Figure 4.3



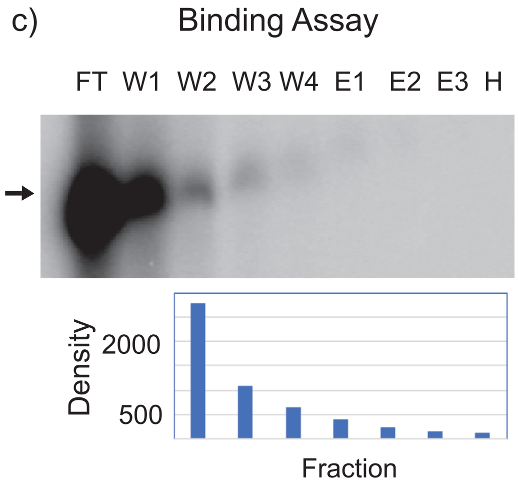


Figure 4.3 Detailed overview of ATP clone 336.3. NCBI BLAST alignment revealed no significant similarity to any genomic reference. A low-scoring alignment match found using the UCSC genome browser against the mouse genome is shown in (a). Portions of ATP 302.67 were found to match a mouse-specific (low conservation) ERV repeat masker. b) The SHAPE profile of ATP 336.3 reveals two nucleotides that exhibit ATP-dependent modification. c) ATP clone 336.3 did not exhibit binding affinity to ATP in an in vitro column binding assay. No significant band intensity was detected in elution fraction 1 (E1) by gel electrophoresis. (FT-flowthrough, W-wash, E-elution, H - denaturing wash).

simply reports matching reads to a reference sequence (ATP 336.3) from a pool of SHAPE reads representing the entire library. Other highly enriched sequences within this library have been found to also map with the ERV repeat element in mice with higher similarity scores (Table 4.2), thereby skewing this dataset. Nevertheless, these data imply a highly structured 5′ region. C36 and T37 become most accessible for modification at higher concentrations of ATP. A56 remains protected throughout all concentrations of ATP, suggesting this nucleotide becomes sequestered during a structural rearrangement in the presence of ATP. In addition, from 1 μM to 1 mM ATP, G66 remains relatively exposed to modification, while T67 remains protected. At 3 and 10 mM ATP, this trend at G66 and T67 is broken, leaving either both positions exposed or protected, respectively. Even with this observation, however, no robust trend could be extracted for changes in nucleotide accessibility as a function of ATP concentration.

To determine whether ATP 336.3 was able to bind to ATP *in vitro*, a simple ATP-agarose column binding assay was performed. Fractions containing radiolabeled ATP 336.3 RNA taken from this binding assay were ran on a denaturing gel. Autoradiography and line densitometry analysis in figure 4.3c showed no significant elution of ATP 336.3 RNA from an ATP-agarose column after free-ligand exchange with ATP. This procedure was repeated three times to reduce any experimental error; however, no significant binding was detected.

ii. ATP clone – 302.67

ATP clone 302.67 contains a number of irregular (TG)n repeats that are not immediately apparent. Consequently, this sequence scores fairly low after submission for genomic alignment. An NCBI BLAST alignment reported a number of poor alignments that predominantly map to repetitive elements within the mouse genome. BLAT alignments of ATP clone 302.67 mapped exclusively as fragmented alignments and were unevenly dispersed throughout the entire mouse genome. After carefully reviewing the best alignments, a strong correlation was found between co-localization of ATP 302.67 and ERV mobile elements. One curious example is shown in Figure 4.4a. In this example, multiple fragmented alignments were observed within the intron of a highly conserved mouse gene, *c1qtnf1*. Only partial alignment matches are seen throughout this intron and overall display very little conservation among vertebrate genomes. Of note, this example is distinctively different from ATP 336.3, as a simple sequence similarity alignment shows very little primary sequence conservation. However, this implies that these repeat motifs are a unique feature to mice and are correlated with ERV insertion sites.

The SHAPE-reactivity plot for ATP clone 302.67 is shown in Figure 4.4b. Nucleotide resolution for this clone was significantly greater and evenly distributed throughout the entire sequence, with four major peaks at T70, A124, A151, and A162. Nucleotide T70 exhibits the most dramatic change in accessibility from low to high ATP. Nucleotides A124, A151 and A162 remain largely protected at all concentrations of ATP with the exception of 1 mM and 10 mM ATP at A124. A162 exhibits higher modification at 3 mM ATP with respect to lower concentrations of ATP, but once again becomes protected at 10 mM ATP. The apparent unpredictable trend seen here is hard to justify. One assumption may include that a particular subset of the 302.67 population may respond to higher concentrations of ATP differently due to a few point mutations. Recall that the reference sequence represents the most common sequence in that family, and there is no guarantee that SHAPE-seq reads from other distantly related sequences may

Figure 4.4

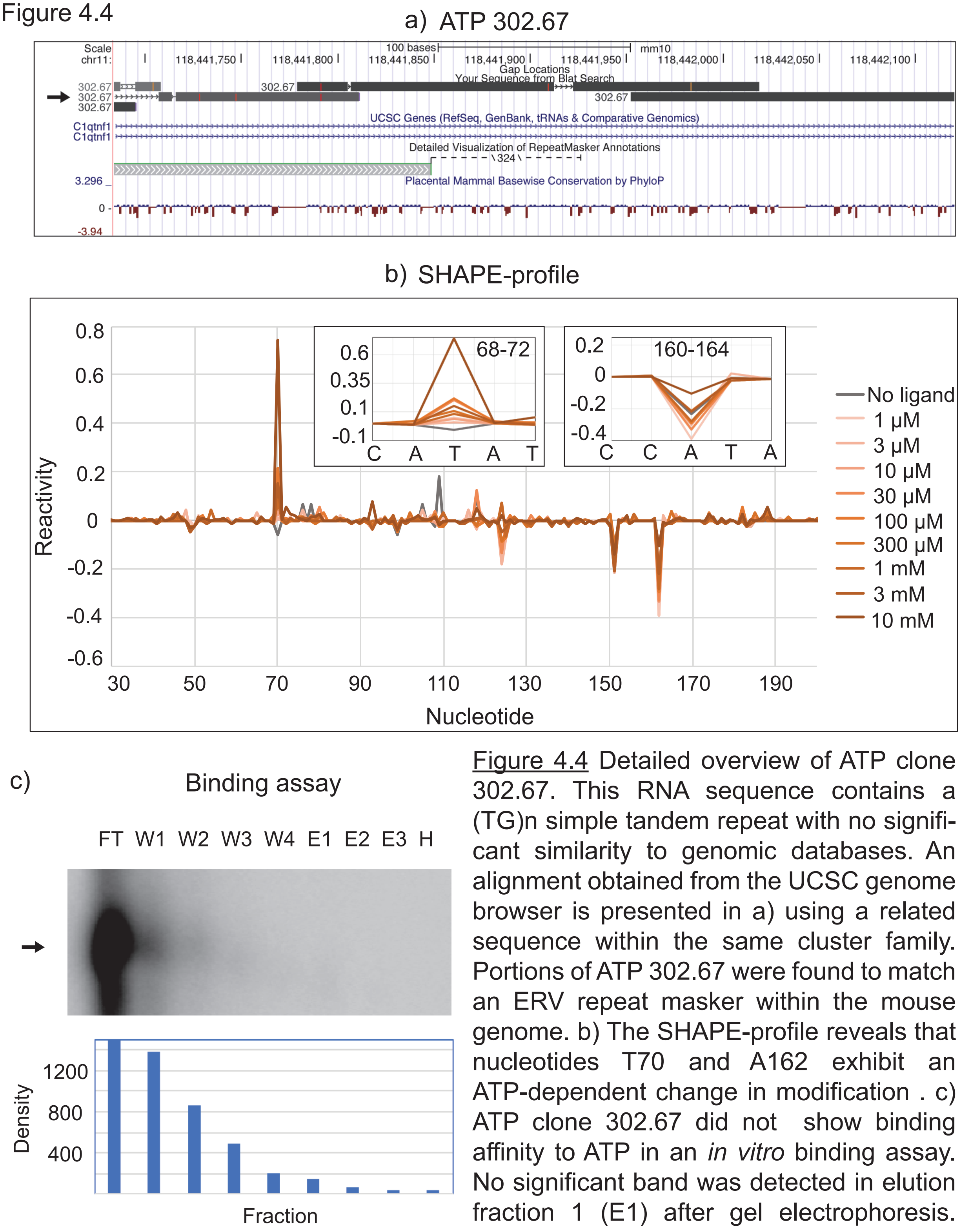


Figure 4.4 Detailed overview of ATP clone 302.67. This RNA sequence contains a (TG)n simple tandem repeat with no significant similarity to genomic databases. An alignment obtained from the UCSC genome browser is presented in a) using a related sequence within the same cluster family. Portions of ATP 302.67 were found to match an ERV repeat masker within the mouse genome. b) The SHAPE-profile reveals that nucleotides T70 and A162 exhibit an ATP-dependent change in modification . c) ATP clone 302.67 did not show binding affinity to ATP in an *in vitro* binding assay. No significant band was detected in elution fraction 1 (E1) after gel electrophoresis. (FT-flowthrough, W-wash, E-elution, H - denaturing wash, B-beads).

be picked up in this analysis. This is further exasperated by the ERV repeat element that has, thus far been present in a large fraction of sequences derived from the ATP enriched library.

The *in vitro* binding assay for ATP 302.67 is shown in Figure 4.4c. As before, radiolabeled ATP 302.67 RNA was incubated on an ATP-agarose matrix to detect binding by elution of free ATP. Fractions collected from this *in vitro* binding assay were run on a denaturing polyacrylamide gel and exposed by autoradiography. From this experiment, no significant band was observed for any of the elution fractions (E1, E2 or E3), demonstrating affinity for ATP. A densitometry analysis confirms this result, showing nearly all of the radiolabeled ATP 302.67 unbound before the completion of the second wash. This experiment was experimentally replicated three times with no significant changes in the detection of labeled RNA.

iii. ATP clone - 230.5

ATP 230.5 was identified as a simple tandem repeat (CTGTG)n sequence present along multiple locations within the mouse genome and represents the most highly enriched sequence from the ATP enriched library, having the best genomic alignment score. UCSC genome browser alignment against the mouse genome includes both complete and fragmented genomic alignments throughout chromosome 16. In general, fragmented alignment matches are a product of the UCSC genome browser tool attempting to identify query sequences as a partially processed sequence, such as nascent RNA, matching the reference genome. To get a better understanding of this repetitive feature, the best alignment match (Fig. 4.5a) was studied for genomic context. ATP clone 230.5 resides immediately between two (CTGTG)n simple tandem repeats

Figure 4.5

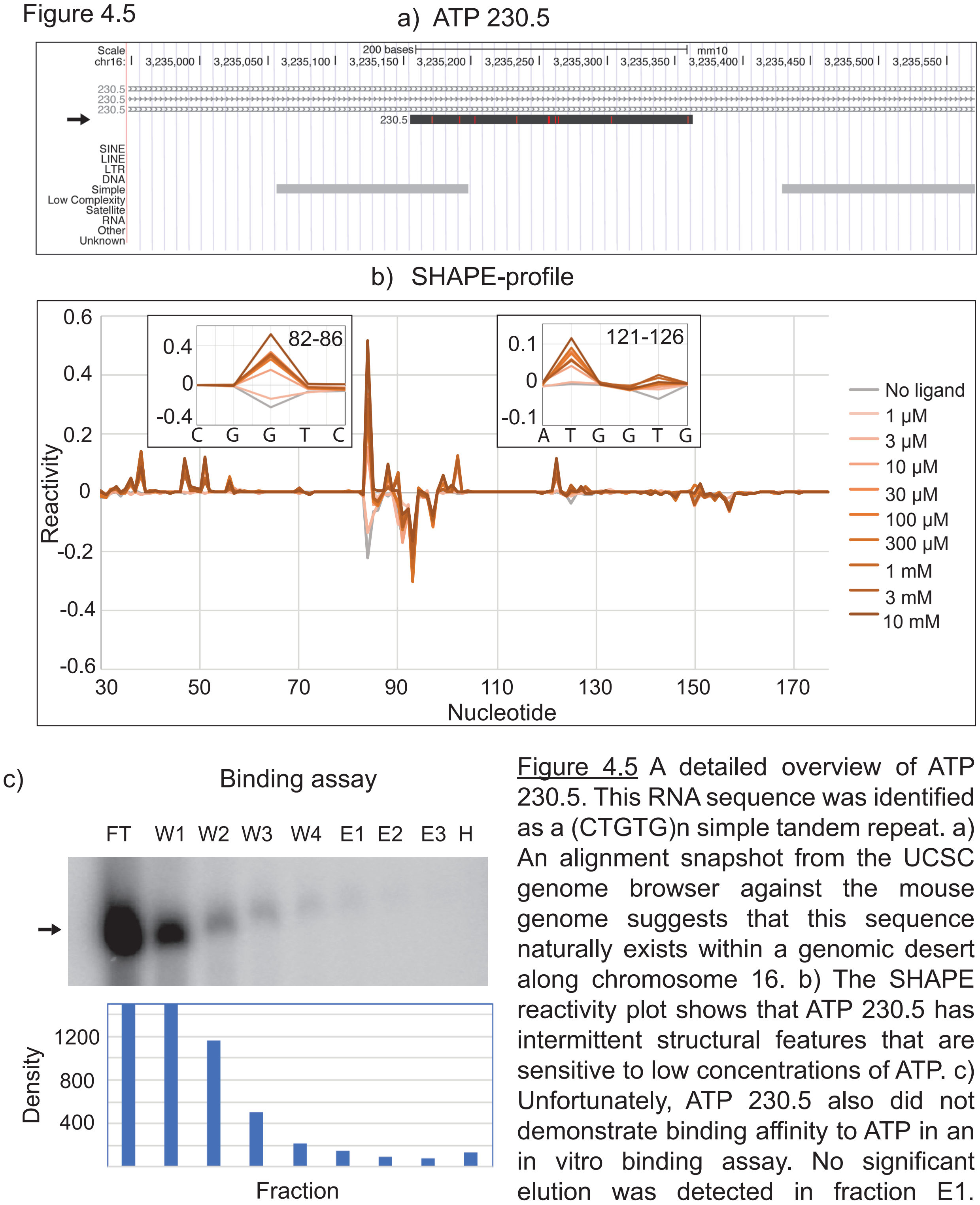


Figure 4.5 A detailed overview of ATP 230.5. This RNA sequence was identified as a (CTGTG)n simple tandem repeat. a) An alignment snapshot from the UCSC genome browser against the mouse genome suggests that this sequence naturally exists within a genomic desert along chromosome 16. b) The SHAPE reactivity plot shows that ATP 230.5 has intermittent structural features that are sensitive to low concentrations of ATP. c) Unfortunately, ATP 230.5 also did not demonstrate binding affinity to ATP in an in vitro binding assay. No significant elution was detected in fraction E1. (FT-flowthrough, W-wash, E-elution, H - denaturing wash, B-beads).

within a region of very low vertebrate conservation. To no surprise, additional (CTGTG)n repeats are present throughout this region and explain the high number of fragmented alignments. Furthermore, this region of chromosome 16 is a genomic desert, with the nearest known annotated RNA / mRNA observed at roughly 50 kbp in each direction from the coordinates shown in figure 4.5a. These characteristics strongly suggest that any RNA aptamer present in this region does not participate in *cis* regulation of gene expression.

A SHAPE reactivity profile for ATP 230.5 is provided in Figure 4.5b. Nucleotide reactivity is intermittently dispersed throughout the entire sequence, with the most significant signal falling within nucleotides 82 – 86 and 121-126 (Fig. 4.5b inset). G84 exhibits the largest positive change in 2′ hydroxyl accessibility from low to high concentration of ATP. This dominant peak suggests that a significant structural feature, likely a nucleotide bulge, is exposed at concentrations of ATP above 10 μM. Interestingly, C93 and A97 remain highly protected from acylation across all concentrations of ATP. T38, A47, C51, and T122 all become accessible to modification above 10 μM ATP. Altogether this profile suggests that ATP clone 230.5 exhibits changes in structural features in the presence of low concentrations of ATP. Unfortunately, no strong correlation between structure and ATP concentration was observed. This is evident by a seemingly sporadic shift in the magnitude of SHAPE reactivities throughout the RNA, particularly at position T122 (Fig. 4.5b inset).

To determine whether this structural response to ATP constitutes binding, an *in vitro* column binding experiment was performed. Fractions containing radiolabeled RNAs from this experiment were loaded onto a denaturing gel and developed using autoradiography (Fig. 4.5c). From this experiment, no significant binding was detected within any of the elution fractions. RNA bands corresponding to ATP 230.3 are present in the first two elution fraction lanes; however, the continued signal decay suggests that this RNA nonspecifically binds to the ATP-agarose column matrix. RNA with high specificity and affinity for ATP would have appeared as band with higher density compared to the preceding wash fractions. Densitometry analysis confirms that this observation is not the case.

iv. ATP – ERVK RLTR17B

To further investigate whether the high number of mouse ERV repeats found by deep sequencing alignments were significant, SHAPE-seq data analysis was applied to the most frequently mapped ERV type observed. On average, ATP selection sequences that were identified with ERV-like sequences were highly related to the mouse ERVK RLTR17B element on chromosome 5 in mouse. This ERVK element at this genomic location represents an ancient viral insertion site evident by flanking simple tandem repeats and short interspersed nuclear element (SINE). This region consists of a number of additional repetitive elements including other ERV types, long interspersed nuclear elements (LINEs), and SINEs and is relatively inactive with only a few uncharacterized RNA transcripts known to be expressed. Furthermore, the ERVK element is not conserved in mammals and shows little conservation with other ERV elements within the mouse genome including active ERV3.

The SHAPE profile against the ERVK RLTR17B element is presented in Figure 4.6. The purpose of this experiment was to determine if RNA encoded within this region of the mouse genome demonstrated sensitivity to free ATP. In total, an average of 4.5K stops were obtained from each ATP titration sample along the ERVK RLTR17B sequence with a larger distribution of stops towards the 3′ end. Although the majority of peaks remain relatively unchanged, four nucleotides A231, T345, A408, and T433 show ATP-dependent changes in modification. Nucleotides at position 400-450 are most accessible at all ATP concentrations indicating a relatively open RNA structure. A408 and T433 show the most dramatic changes in structural rearrangement from 100 μM to 1mM. At position A408, an increase in RT stops are observed together with neighboring nucleotides as ATP concentration increases. A peak shift is observed from A408 to T407 as well as sustained levels of acylation at A408 and T409 indicating a relatively open RNA structure at 10 mM ATP. Similarly, T433 shows the largest change in accessibility at various ATP concentrations. However, no robust correlation between ATP concentration and accessibility of A231, T345, A408, and T433 could be extrapolated. This may be due to an unforeseen bias in both deep sequencing and SHAPE-seq reads due to closely related ERV repetitive elements. This deviation is discussed further below.

v. cAMP clone – 669.83

With an enrichment value of 669-fold, cAMP 669.83 is the most enriched sequence observed from any of the metagenomic *in vitro* selection experiments performed here. The cAMP 669.83 sequence is GC-rich and does not contain simple tandem repeats. NCBI nucleotide BLAST found a 100 % match to the *Chlamydomonas reinhardtii* genome, a unicellular eukaryote belonging to the family of green algae. In fact, the entire assembled cAMP 669.83 sequence was found within the two largest exons of a known mRNA expressed in *C. reinhardtii* (Fig. 4.7, red boxes). Unfortunately, the protein product for this mRNA transcript has yet to be characterized, however peptide BLAST suggests that this protein encodes two peptidase domains (white directional arrow, Fig. 4.7) that co-align with the cAMP 669.83 sequence. The mRNA transcript is 4.4 k nts in length and contains two repeated sequence segments, each of which contain one predicted peptidase domain. Alas, this transcript resides in a sequence assembly gap, masked by small but significant number of simple sequence repeats in this region.

To determine if cAMP 669.83 alone retained affinity for cAMP, radiolabeled RNA was added to a cAMP-agarose column. As before, unbound RNA was removed through a series of washes, and cAMP – specific RNA was eluted by free ligand exchange. Each fraction collected from this binding assay was loaded onto a denaturing acrylamide gel and subsequently imaged by autoradiography. The gel showed 'sticky' RNAs adhering to the cAMP-agarose matrix, evidenced by a consistent band present throughout all the wash fractions. Densitometry analysis shows that that the first elution (E1) band is slightly larger than the preceding wash fraction, suggesting that a small number of RNAs were eluted from the column upon exchange with free cAMP. Despite the subtle peak in elution (Fig. 4.7b) this result was found to vary slightly during experimental replication and does not demonstrate robust binding of cAMP 669.83 to cAMP.

vi. cAMP clone – 292.17

cAMP 292.17 is the third most enriched sequence derived from the cAMP *in vitro* selection and was found to span antisense to exon 3 of the Slx4 gene on chromosome 16 in mouse using NCBI nucleotide BLAST. A detailed view of this region using the UCSC genome browser is illustrated in Figure 4.8a. While the gene itself is conserved amongst most vertebrates, this exon region is only sequence conversed in mice and rats. Slx4, or SLX4 structure-specific endonuclease subunit homolog was first identified in yeast

Figure 4.6
a)

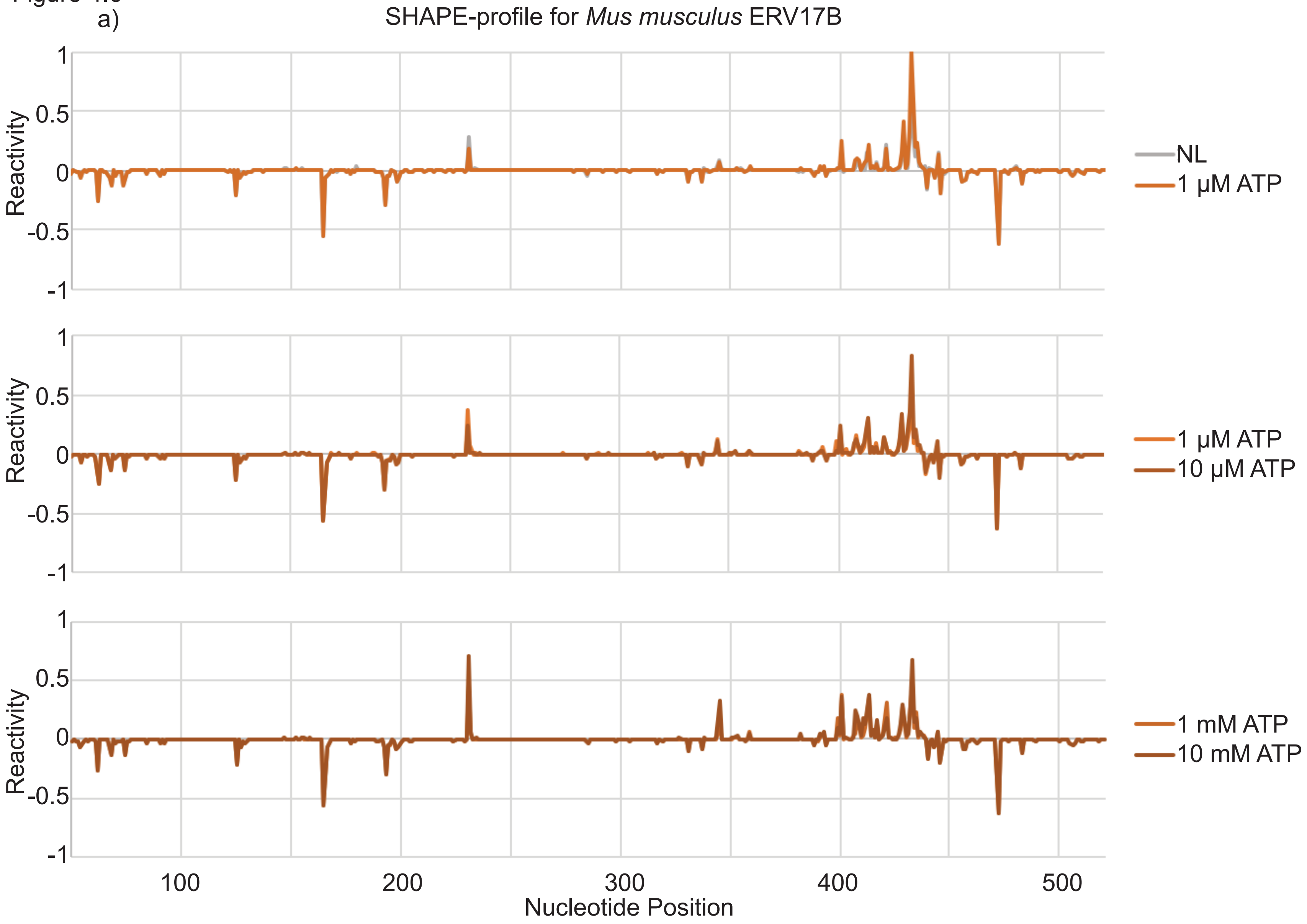

SHAPE-profile for *Mus musculus* ERV17B
Reactivity
1
0.5
0
-0.5
-1
NL
1 µM ATP
Reactivity
1
0.5
0
-0.5
-1
1 µM ATP
10 µM ATP
Reactivity
1
0.5
0
-0.5
-1
1 mM ATP
10 mM ATP
100
200
300
400
500
Nucleotide Position

Figure 4.6 (continued)

b)

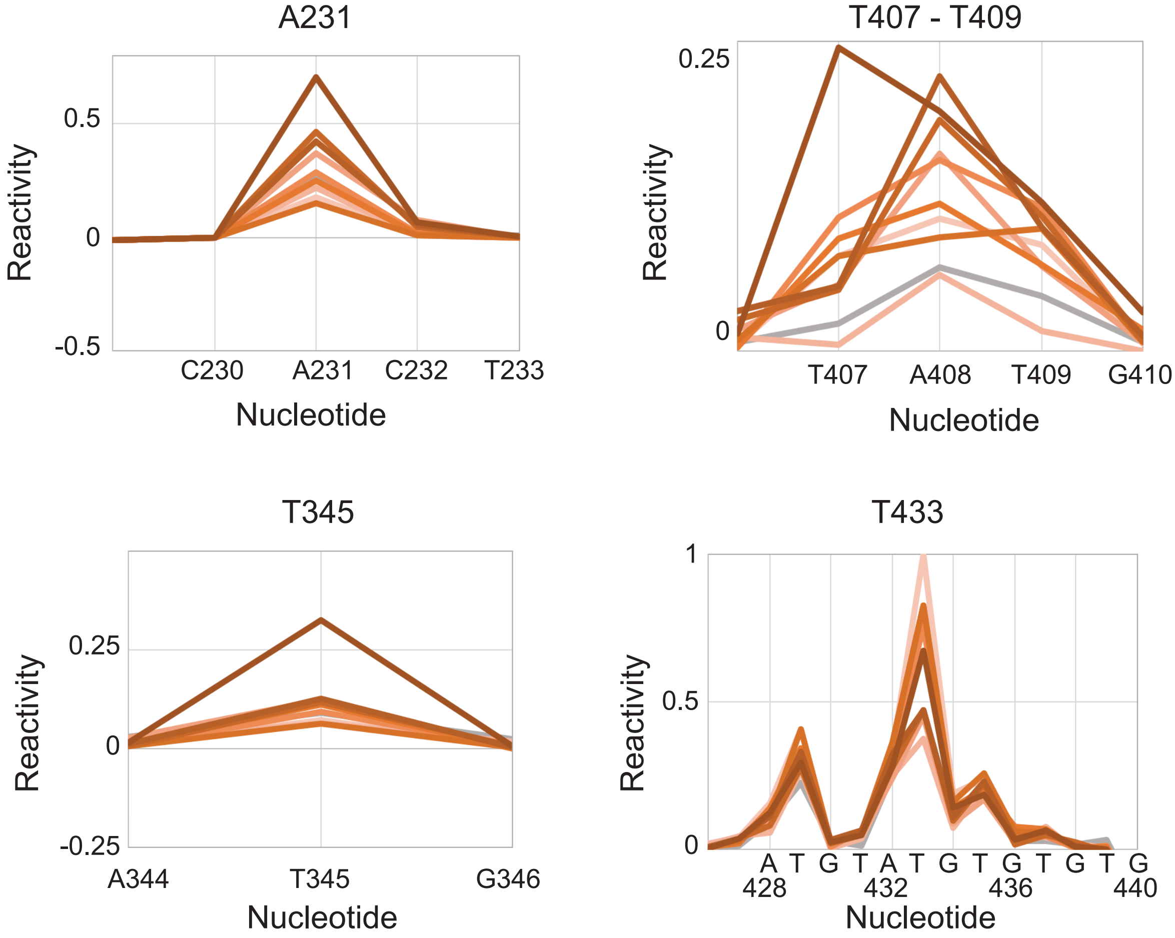
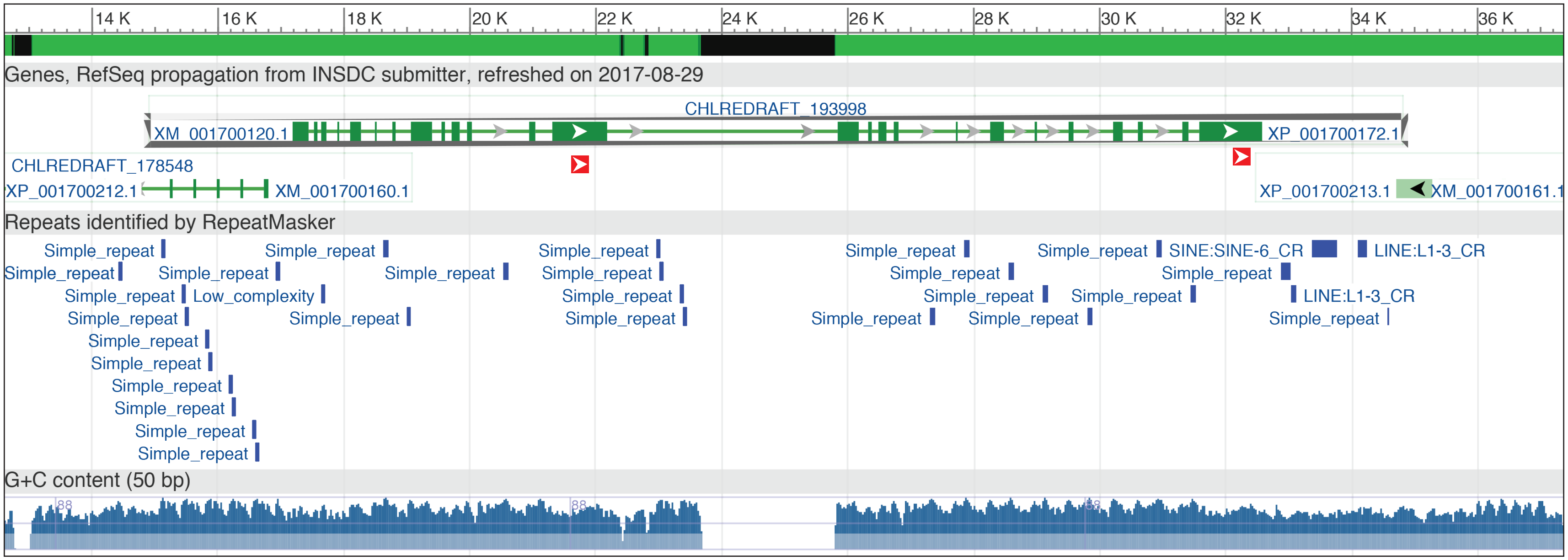


Figure 4.10 A detailed SHAPE-profile for a dominant ERV (ERVK RLTR17B) element present within the ATP enriched pool is illustrated above. To generate this profile, SHAPE-sequencing reads were aligned onto the a 1.5 kbp window containing the entire mouse ERVK RLTR17B element using the bowtie2 algorithm. A total of 4.5 K RT-stops were identified strictly within the 500 bp ERVK element as shown in (a). Four peaks corresponding to A231, A408, T345 and T433 demonstrated the largest changes in chemical modification as a result of increasing ATP concentration. Detailed views of these positions are shown in (b). A231, and A408 show larger reactivity values suggesting that these regions may be less structured and become more relaxed and flexible with increasing concentrations of ATP. In addition, T433 shows the largest change in accessibility at various ATP concentrations; However, no strong correlation between ATP concentration and reactivity could be extrapolated.

Figure 4.7

a) cAMP 669.83 alignment against the *Chlamydomonas reinhardtii* genome



b) cAMP-agarose column binding assay

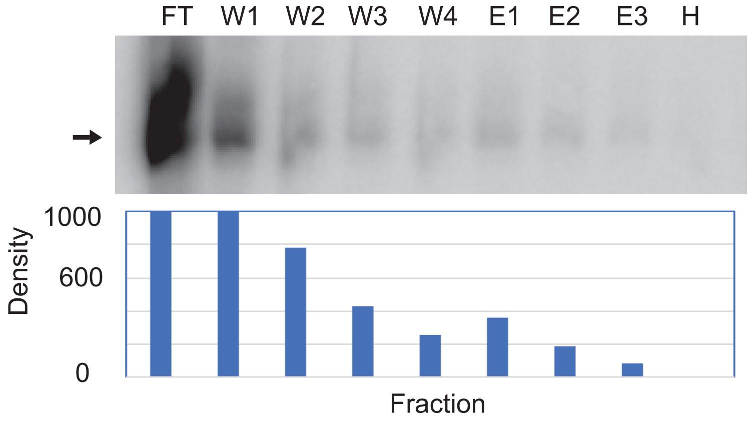


Figure 4.7 Sequence alignment and binding profile for cAMP clone 669.83. (a) A detailed NCBI BLAST alignment window is presented above. The segmented green boxes correspond to a predicted ubiquitin-like protease 1 gene. The red boxes denote the direction of the cAMP 669.83 sequence along the mRNA transcript. Blue boxes represent high GC simple repeat maskers. (b) In an *in vitro* column binding experiment, cAMP 669.83 demonstrates little to no affinity for the cAMP ligand. Oddly, radiolabeled RNAs are observed to slowly pass through the cAMP column over a series of washes (W) and elutions (E).

Figure 4.8

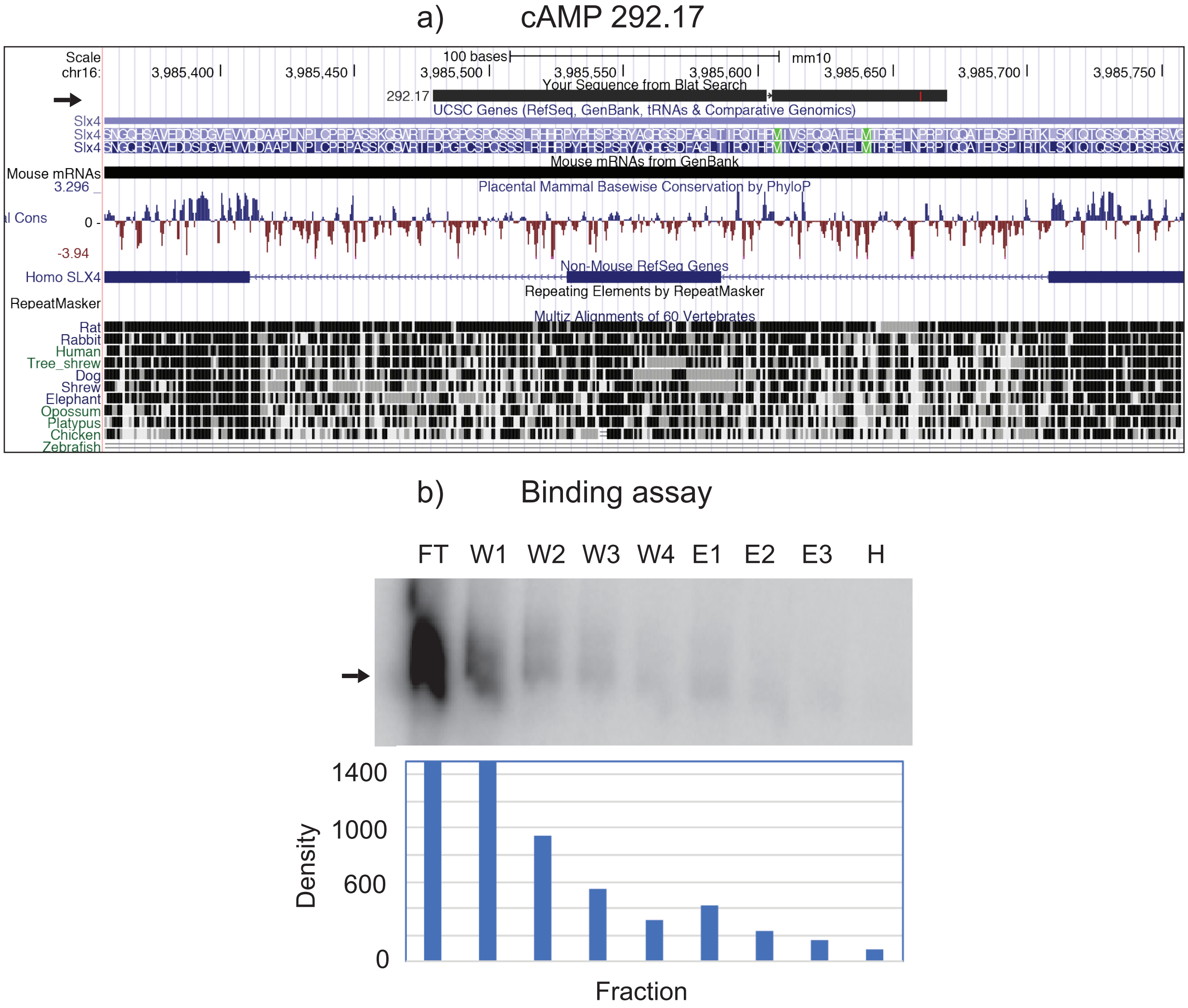


Figure 4.8 Detailed overview of cAMP 292.17. Using NCBI BLAST, cAMP 292.17 was found to have high sequence similarity to exon 3 of the mouse Slx4 gene (a). Slx4 is a highly conserved structure-specific DNA endonuclease that has been shown to be involved in DNA repair mechanisms. A portion of the cAMP 292.17 sequence overlaps the exon of an Slx4 homolog present in *H.sapiens*. Unfortunately, the cAMP 292.17 RNA sequence did not exhibit affinity for cAMP in an *in vitro* column binding assay (left).

(*S. cerevisiae*) and encodes a protein responsible for recognizing secondary structure motifs in DNA associated with DNA replication and repair mechanisms.

As with previous clones, cAMP 292.17 was tested for cAMP binding by *in vitro* column binding assay. The gel electrophoresis results and a subsequent densitometry plot for this experiment are presented in Figure 4.8b. Radiolabeled cAMP 292.17 RNAs were bound and eluted over a series of washed and elutions as previously descripted (refer to methods). Similar to cAMP 669.83, a small fraction of RNAs were retained throughout all wash fractions, with only an increase in band intensity in the first elution (E1) fraction. The densitometry plot suggests comparable densities between wash 3 and elution 1; however, no robust binding was detected. This experiment was repeated three times to minimize experimental error, and no significant binding was observed in any trial. Additional experiments were performed including ATP-agarose and ADP-agarose matrices, and elution with their respective free ligands. Despite these efforts, no significant binding events were detected for either target.

vii. cAMP 282.33 and 479.33

A UCSC BLAT alignment window for cAMP 282.33 is shown in Figure 4.9. cAMP 282.33 was identified as a simple tandem repeat unique to the zebrafish genome. Figure 4.9 illustrates one example of cAMP 282.33 flanked by an upstream copia-6-I DNA repeat element and a downstream LTR_DR element. Both of the flanking elements are antisense to cAMP 282.33 and are highly prevalent throughout the genome. In fact, these repeat elements flanking cAMP 282.33 are consistent for nearly all matches found in zebrafish. One exclusive alignment shows that these elements even span homologous loci that encode conserved genes in other organisms (Fig. 4.9, bottom panel).

Figure 4.9 cAMP 282.33 alignment against the *Danio rerio* genome

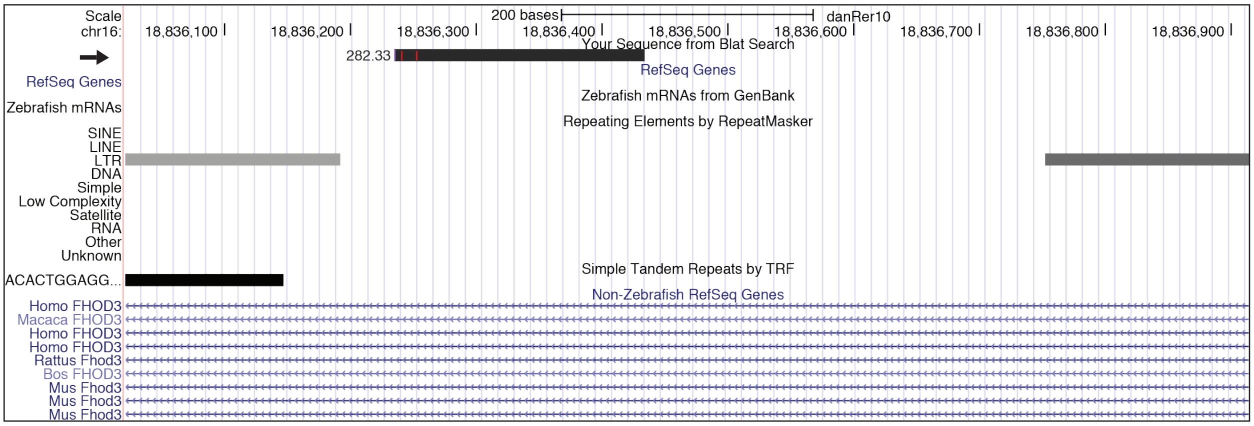


Figure 4.9 A snapshot of the cAMP 282.33 sequence alignment from the UCSC genome browser is shown above. The cAMP 282.33 sequence was identified as a simple tandem repeat unique to the zebrafish genome. This sequence is flanked by an upstream copia-6-I DNA repeat element (light gray) and a downstream LTR_DR element (darker gray), both of which are antisense to the 282.33 sequence. Oddly, these simple repeat elements are highly prevalent in this region of the zebrafish genome.

Figure 4.10 Column binding assay using cAMP 479.33

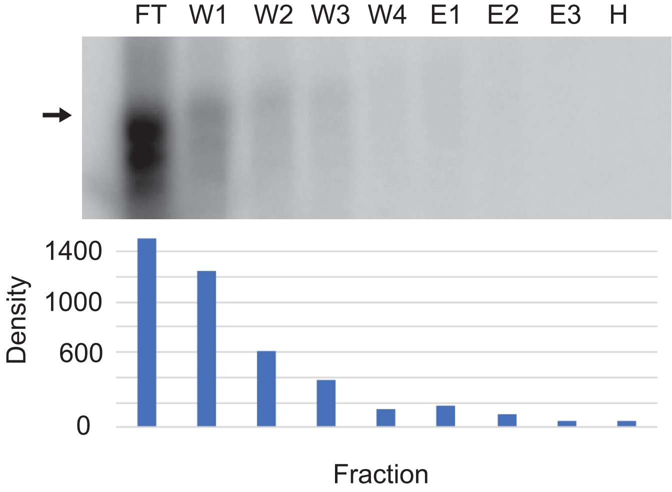


Figure 4.10 *In vitro* binding assay results for cAMP 479.33. Radiolabeled cAMP 479.33 show a sustained level of RNA being washed off the cAMP-agarose column. Quantification of band density reveals that a small but detectible signal is observed at the first elution fraction (E1). This small signal however was difficult to reproduce and was found to be insufficient to conclude that cAMP 479.33 demonstrates affinity for the cAMP target.

Despite being the second most enriched RNA sequence, cAMP 479.33 was found to have no significant genomic alignment match to any organism. Additional genomic alignments attempts were performed against low-complexity genomic databases with no success, which indicate that cAMP 479.33 may not be a repetitive element. While no genomic origin could be assigned, an *in vitro* column binding assay was performed to determine if high enrichment was due to cAMP binding (Fig. 4.10). Elution fractions loaded on a denaturing acrylamide gel showed no significant elution of cAMP 479.33 after a series of washes followed by free cAMP exchange. Bands quantified by pixel density (Fig. 4.10, lower panel) show a sustained level of RNAs collected from the fourth wash (W4) to the first elution (E1), but no large signal is observed for any of the elution fractions suggesting cAMP 479.33 has little to no affinity for cAMP. This experimental set up was repeated three times using cAMP-agarose and free cAMP to minimize experimental error; however, no significant variation to these data were observed.

viii. rRNA alignments

To get a better understanding why repetitive sequences were present within both the ATP and cAMP enriched libraries, we measured the distribution of sequence reads that distinctively map to a reference gene containing the large ribosomal subunit sequence. The large ribosomal subunit which encodes the peptidyl transferase center, is a highly conserved sequence and is found in every living organism [65]. Furthermore, rRNA genes are naturally repetitive within a given genome and their copy numbers vary with no apparent correlation to genome size. This alignment distinguishes whether the ATP *in vitro* selection enriched for nonspecific RNA by aggregation (i.e. simple repeat sequences) or if selective pressure was applied to specific repeat types, supporting enrichment of functional RNAs.

To proceed with this analysis, we performed a local sequence alignment of deep sequencing reads using the bowtie2 algorithm along the *D. discoideum* 26S ribosomal DNA reference sequence. This organism was chosen because it has one of the best annotated genomes and because a high resolution molecular structure of its large ribosomal subunit is available [28], [66]–[68]. The number of aligned sequences between the naïve and ATP-enriched library were compared using the Integrated Genome Browser (IGB) visualization software (Fig. 4.11). This visualization software plots the number of reads as peaks along the reference sequence in the 5′-3′ direction where values above the reference sequence represent reads matching the sense strand, and those underneath represent the number of matches along the antisense strand. For convenience, green boxes denote *D. discoideum* 5.8S and 26S rRNA genes and the red box represents the peptidyl transferase center. For both the naïve library and the ATP aptamer enriched library, seven 300-bp peaks are intermittently spread throughout the reference sequence (Fig. 4.11). Of these seven peaks, five major peaks near nucleotide positions 1000, 2000, 2500, 3500 and 4200 can be seen for both datasets. With the exception of a small population of sequences at the 3′ end of the 2500 peak, we observed an overall reduction of sequences in the ATP aptamer enriched library compared to the naïve metagenomic library. This reduction suggests that selective pressure was applied against rRNA sequences after a few rounds of *in vitro* selection.

To verify if any of these peaks demonstrated any potential adenosine-binding activity, we used gene-specific primers to PCR-amplify these specific segments from the

Figure 4.11

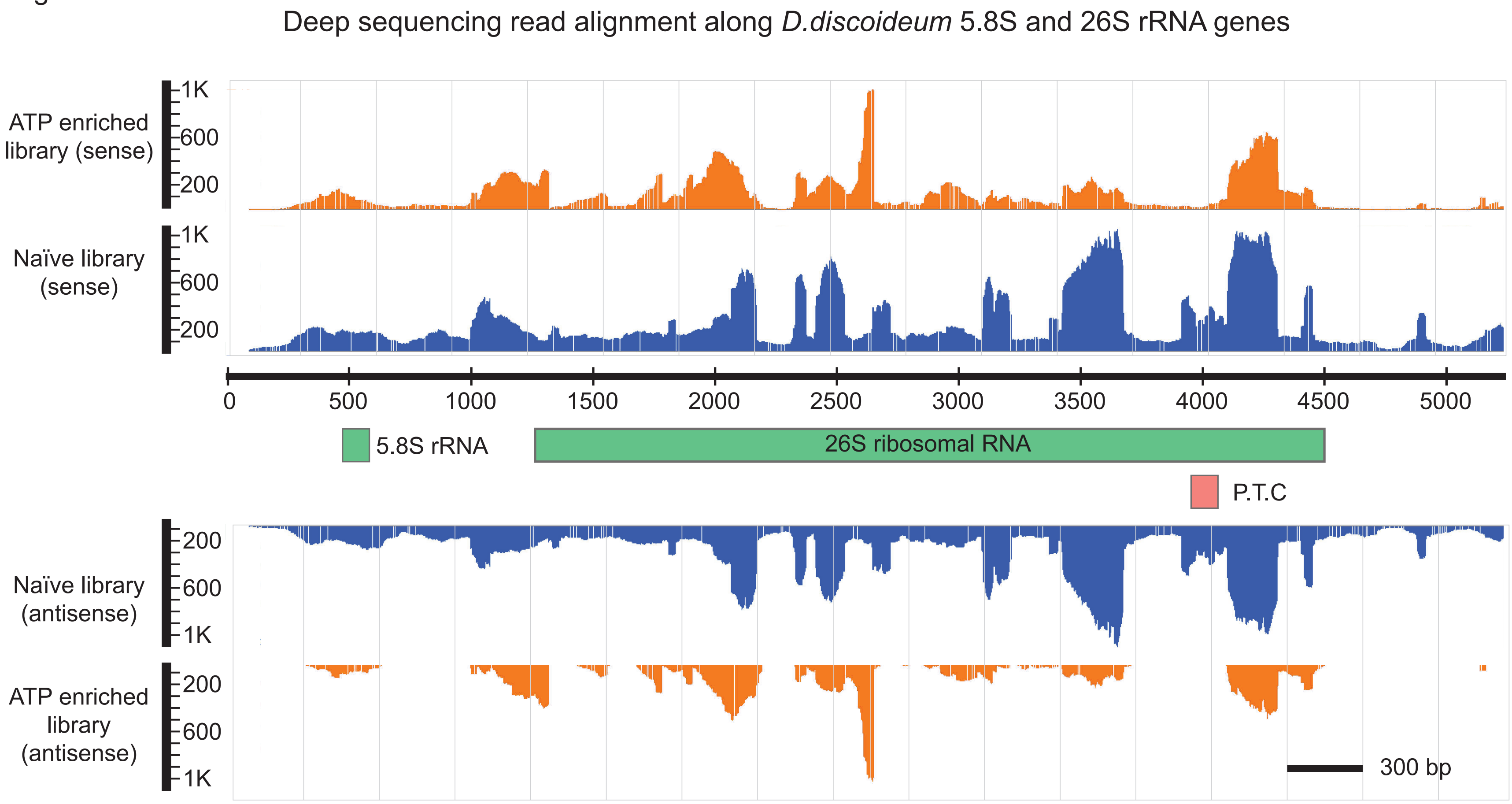
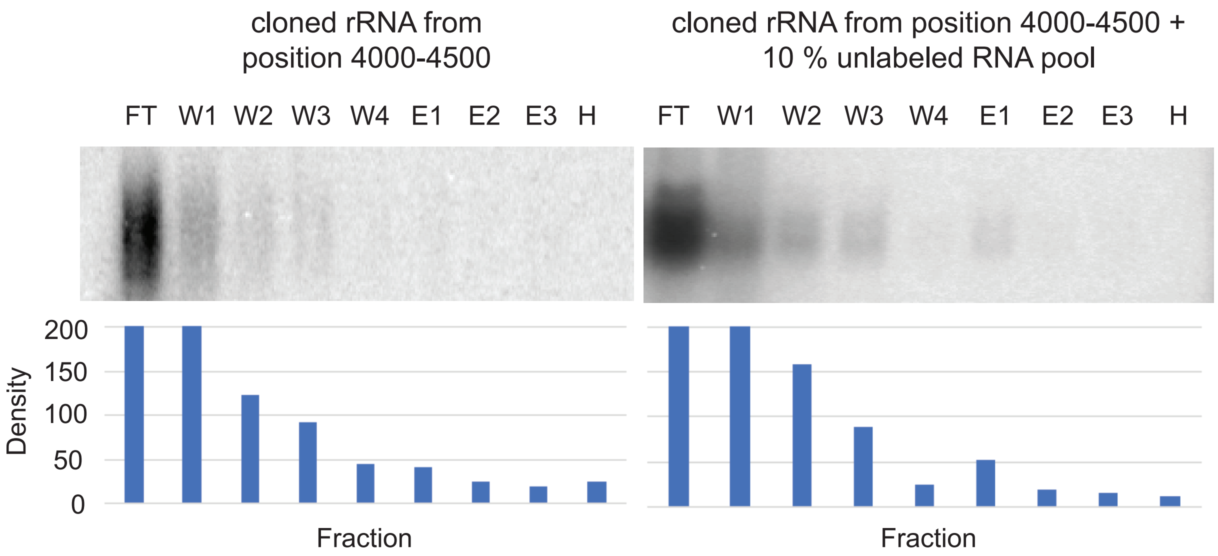


Figure 4.11 Alignment of the naïve metagenomic library (blue) and the ATP-aptamer enriched library (orange) deep sequencing reads along the *D. discoideum* 5.8S and 26S rDNA reference sequence. Y-axis values correspond to the number of sequence reads that match a specific coordinate (x-axis) along the rDNA reference sequence. Peaks above the coordinate axis correspond to sequences that align to the sense strand, while peaks below align to the antisense strand. For scaling, a 300 bp reference is annotated to represent the length of an intact RNA sequence derived from the naïve or enriched metagenomic library.

Figure 4.12

Column binding activity of a sequence derived from a segment of *D. discoideum* 26S rRNA



cloned rRNA from position 4000-4500

cloned rRNA from position 4000-4500 + 10 % unlabeled RNA pool

Figure 4.12 A rRNA segment of the *D.discoideum* 26S rRNA (taken from the distribution between position 4000 and 4500 in Fig. 4.11) was amplified from the ATP aptamer enriched library and tested for column binding activity. (a) The binding profile of this rRNA sequence alone using ADP-agarose is shown above (left). No significant ADP binding activity was detected. (b) Interestingly, in the presence of 10 % (v/v) ATP aptamer enriched RNA library, a significant elution (right; E1) was detected after a series of washes (W1-4) using an ADP-agarose column and ADP elution buffer. ATP binding activity for this specific rRNA segment supplemented with the enriched RNA library was also observed but at a reduced level (data not shown).

enriched library and performed a column binding assay (Fig. 4.12). Similar to observations made with ATP and cAMP clones above, all cloned rRNA segments did not demonstrate any specificity to ATP, ADP, or cAMP alone. However, in the presence of 10 % (v/v) unlabeled enriched RNA pool, the cloned rRNA segment between position 4000-4500 weakly bound and eluted off an ADP-agarose column (Fig. 4.12b). This result suggests that this cloned rRNA sequence alone does not have affinity for neither ATP, ADP, nor cAMP but demonstrates potential to bind ADP in the presence of the ATP aptamer enriched RNA library. In a similar assay, binding activity was not observed upon using a transfer RNA library (data not shown). This observation supports the idea that specific RNA-RNA interactions contribute towards ADP binding activity for sequences derived from the enriched library.

IV. Conclusions

In this chapter, deep sequencing and chemical molecular probing approaches were used to identify and detect structural features of RNAs derived from an *in vitro* selection. Firstly, the naïve and enriched RNA libraries were sequenced and examined to establish the most abundant and enriched sequences. Sorting and clustering methods made it easier to browse and study these sequences to determine their genomic origin and similarities, if any. As expected, the majority of the naïve metagenomic library sequence reads were highly diverse and are well distributed across any given genome after performing genomic alignments. However, the most abundant sequences in the naïve library are telomeric and centromeric sequences commonly found in each organism represented within the metagenomic library. These specific highly abundant genomic repeats were not observed in the enriched metagenomic libraries, suggesting successful removal of high-frequency starting sequences.

While the enriched libraries do not contain these large numbers of telomeric and centromeric sequences, other genomic repetitive elements were observed. This may be explained by the fact that these elements, like the ERV repeat element, are also commonly found in a number of different organisms and are also slightly enriched within the naïve metagenomic library. As observed in the ATP selection, this initial ERV bias within the naïve library considerably reduced the probability of obtaining a more diverse family of ATP aptamers by round 5. Nearly 70 % of the sequences recovered by deep sequencing suggest that they have originated from an ERV element. Altogether, this means that the ATP *in vitro* selection required additional selective pressure for ATP specificity or against ERV-specific aptamers should any non ERV-like sequence be found. Furthermore, this result contrasts our previous ATP *in vitro* selection using a human genomic library, which revealed a smaller fraction of ATP aptamers that partially aligned with human ERV elements. In this case, at least two adenosine aptamers were enriched for two different intronic sequences not associated with a known repeat element in humans. The addition of genomic DNA of various organisms therefore introduces an additional challenge that may require higher stringency during *in vitro* selection.

Four dominant ATP sequences were extracted from the ATP selection, all of which originated from the mouse genome. Of these sequences, ATP 336.33 was the most highly enriched sequence followed by ATP 302.67 and ATP 247.33. Both ATP 336.33 and ATP 302.67 mapped to segments of a specific ERV repeat masker element type (ERV17B) using the UCSC BLAT genome browser. While the origin of ATP 247.33 could not be identified by NCBI nucleotide BLAST, a sequence alignment using Clustal Omega revealed that ATP 247.33 is most related to ATP 336.33 in primary sequence with two 20-25 nt segments of high similarity. Additional sequence alignments did not show any primary sequence conservation between ATP 247.33 and a number of different ERV types, suggesting that ATP 247.33 may represent either an uncharacterized ERV repeat or an entirely new sequence from *in vitro* selection due to low similarity scores and having ERV-like sequence features. Finally, ATP 203 was characterized as a simple tandem repeat found in both mouse and zebrafish, with no significant expressed data found near matching loci.

Due to the high prevalence of mouse ERV matches amongst all these ATP selection sequences, a SHAPE profile was generated using a mouse genomic ERV sequence as a reference. The purpose of this was to determine if any valuable structural data could be consolidated onto a common reference sequence to both minimize ERV biases and to shed light on structural features of RNAs originating from the mouse ERV genome. A SHAPE profile against ERV17B revealed a relatively open and accessible 3′ end, likely attributed to a relaxed single-stranded or highly flexible region. Two nucleotides in this region, A408 and T433, seem to respond to an increase in ATP concentration; however, no strong correlation was observed due to seemingly sporadic changes in nucleotide modification. This subtle and inconsistent measurement may correspond to the flexibility of these nucleotides that are prone to modification due to experimental variations during RNA annealing. The fact that no flanking nucleotides demonstrate a compensatory in structural rearrangement by chemical modification indicates that these structural features are ATP independent.

Since ERV repetitive elements dominated the selection and were exacerbated by the large quantity of ERV elements in the mouse genome, isolation of a single RNA sequence was challenging. It should be noted that even though one of these clones may have ATP binding activity, it is probable that this gain of function may be a result of ERV site insertion and integration. This observation has been reported previously from *in vitro* selection studies using a human genomic library, and is one example that supports this argument [13]. In contrast, whether ATP binding is encoded within the ERV-genome remains to be seen. Deep sequencing data show high similarity to a mouse ERV17B element located on chromosome 5, which suggests that this may not be the case given the large number of ERV elements present along the mouse genome.

Repetitive biases were also observed within the cAMP selection. Despite the high enrichment of *C. reinhardtii*-related sequences, a noticeable amount of TG and CA repeats were also measured. The *C. reinhardtii* genome is GC rich with a large number of GC-rich transposable elements and simple repeat elements. It is apparent that these repeats tend to aggregate and were co-enriched as a result of association with the most abundantly enriched sequence. This factor is a significant drawback during SHAPE-seq analysis due to the fact that these repetitive elements will interfere with alignment SHAPE RT stops. Sequence repeats that may actually have arisen from an entirely unrelated locus, may align with the reference sequence and produce overestimated alignments by the bowtie2 algorithm. Even though the alignment algorithm takes mismatches and insertions into account, the low complexity and low information content of these sequences can significantly skew any SHAPE profile that has a reference sequence with any simple repeat features. Consequently, this is observed as a stable peak along a reactivity plot throughout an entire ATP titration as seen for the cases mentioned above. To address this issue, a customized stringent alignment score table should be configured in the bowtie2 alignment program. Doing so will allow the program to align sequences with reliable base calls and avoid aligning extremely low complexity sequences onto the reference sequence.

Repetitive sequences aside, several dominant cAMP sequences were identified from the cAMP *in vitro* selection library. Interestingly, many of these sequences were found to align to specific genomic loci with high similarity. cAMP 669.83 was the most highly enriched sequence taken from the cAMP selection and was matched to a predicted mRNA transcript expressed in *C. reinhardtii* with high similarity. *C. reinhardtii* is a unicellular eukaryote that resides in soil-dwelling communities and is a model organism studied for cell cycle and intracellular signaling regulation. Intriguingly, this species of green algae is known to contain the largest class of cytochrome P450 and guanylyl / adenylyl cylases [30]. *In vivo* experimental data provided in literature have shown that both cAMP and cGMP molecules serve critical functions for mating, nutrient acclamation, and flagellar function and regulation via intracellular signaling in *C. reinhardtii* [30].

Unfortunately, assembly and annotation of the *C. reinhardtii* genome is incomplete, making it necessary to rely on tools such as the conserved domains database (CCD) to identify protein domains through evolutionarily conserved protein sequences and architectures [69]. Using CCD, cAMP 669.83 was found to be related to an mRNA transcript that corresponds to a predicted peptidase expressed within *C. reinhardtii* that belongs to the Ubiquitin-like protease 1 (ULP1) superfamily of proteases. This family of proteins has been shown to be involved in peptide maturation and cell cycle regulation in yeast and have been associated with a variety of viral genomes [70]. Although no direct evidence has been found for the function of this mRNA in *C. reinhardtii*, ULP1 has been shown to be highly conserved in eukaryotes and is closely tied to intracellular signaling pathways [70]. Nucleotide sequence conservation, however, varies significantly between homologous ULP1 genes.

In addition, cAMP 292.17 was found to match a portion of exon 3 of slx4 structure-specific endonuclease subunit homolog in mouse. This sequence spans the exon of a homologous gene present in humans and is also present in a large number of vertebrate genomes. Currently there are 11 known isoforms of SLX4 that have been detected in mice, the majority of which encode a structure specific DNA endonuclease. The genomic context of cAMP 292.17 in addition to the diversity of isoforms for this ULP1-like gene is reminiscent of the TPP riboswitch. Furthermore, intracellular signaling *via* cAMP has been shown to stimulate ubiquitin activity in a number of neurological pathways [64], [71]–[73]. This result, however, is further complicated by the lack of *in vitro* binding data observed from column binding assays.

No significant ATP-column binding was observed for any sequence obtained from both ATP and cAMP selections. From these assays, only cAMP 669.83 and cAMP 292.17 exhibited minimal interaction with cAMP showing only 5-7 % elution between experimental replicates. Aside from these two sequences, no additional sequence exhibited detectable ATP, cAMP, or ADP binding activity. Interestingly, some binding activity for clones derived from a segment of the large rRNA subunit was observed upon reintroducing a fraction of the original enriched library. In some cases, 10 – 16 % binding and elution was observed in samples that had 3:1 ratio of radiolabeled clone RNA to unlabeled pool of RNA derived from the enriched library.

By providing these supplemental RNAs this may also facilitate additional RNA-RNA interactions that necessitate ligand binding. However, due to the number of potential interactions, perhaps these ‘nonspecific’ interactions also prevent successful binding which would explain the high variability (10 -16 %) of binding and elution. If this is the case, the ligand-binding RNA complex must be able to be isolated using alternative approaches such as using a gel-shift assay. This prediction is not far-fetched given that the aptamer domain of the thymine pyrophosphate (TPP) riboswitch requires base pairing interactions between two distant segments of RNA sequences along the RNA transcript.

CHAPTER 5

Expression studies of the human adenosine aptamers

I. Introduction

Earlier *in vitro* selections and structure-based searches revealed two adenosine aptamers within the introns of the human *RAB3C* and *FGD3* genes. This *in vitro* selection utilized a human genomic DNA library as the starting material which was used to enrich for ATP-binding aptamers [74]. Most recently, deep sequencing analysis of this enriched library revealed additional adenosine aptamers including the *THE1B* aptamer which maps antisense to a repeat element derived from a sub-family of ERV Mammalian apparent LTR-retrotransposons [50]. After multiple rounds of selection, re-folded RNA and co-transcriptional binding assays revealed that the *RAB3C* and *FGD3* aptamers had high specificity and affinity for ATP with K_ds on the order of ~400 µM [50], [74]. Furthermore, mapping these sequences against the human genome revealed that both the *RAB3C* and *FGD3* aptamers reside within sequence conserved regions amongst primates [74].

RAB3C belongs to a large family of GTPases that have been shown to be involved in vesicular trafficking, particularly in cells with high exocytosis activity such as neuronal cells [75], [76]. *RAB3C* is primarily expressed in the brain and within the adrenal glands at low levels, and mutations within in this family of proteins have been shown to lead to neurological and developmental disorders such as Warburg Micro Syndrome [77]. On the other hand, *FGD3* belongs to a family of FGD1 RhoGEF (guanine nucleotide exchange factors) that has been shown to regulate actin cytoskeleton rearrangement in the processes of cell growth and differentiation [78]. Mutations in the FGD1 family have been shown to negatively affect these cellular processes resulting in unusual cell shape and morphology and have been linked to Aarskog-Scott syndrome (faciogenital dysplasia), characterized by skeletal and genital abnormalities [78]. Interestingly, both *RAB3C* and *FGD3* are a part of cell signaling pathways that are ATP and GTP dependent. These characteristics suggest that their expression may be regulated by the level of ATP available. In this chapter, we sought to test and explore whether the ATP aptamer embedded within these genes may constitute a novel mode of genetic regulation by serving as a small-molecule sensor for mRNA expression.

II. *RAB3C* and *FGD3* expression in SH-SY5Y

To gain a better understanding of the biological role of these ATP aptamers, we sought to understand if expression of these genes, including the aptamer itself, is responsive to exogenous adenosine. Using SH-SY5Y cell line, we measured targeted genes and their respective ATP aptamers expression level by qRT-PCR. RNA levels for each gene and their respective ATP aptamers were measured using total RNA extracted from SH-SY5Y tissue cultures and analyzed by qRT-PCR. The SH-SY5Y cell line was derived from human neuroblastoma tissue that grows as a mixture of suspended and adherent cells. Studies have shown that treating SH-SY5Y cell cultures with all-trans retinoic acid can induce neuronal differentiation by inducing expression of a non-selective cation channel [79]. Both undifferentiated and differentiated cells have been used to study neuronal cytoskeleton regulation, Parkinson’s disease, and Alzheimer’s disease [79]–[81]. Adhered SH-SY5Y cells grow in clusters and typically form aggregation islets that grow both laterally and vertically. Upon adhering, SH-SY5Y cells also exhibit phenotypic features were observed and recorded over range of adenosine concentrations.

SH-SY5Y human neuroblastoma cells were cultured in DMEM/F12, supplemented with 10% fetal bovine serum (FBS) and 1 % penicillin (10 000 IU/mL)/streptomycin (10 000 μg/mL). The culture was maintained in a 5 % CO_2 incubator with a humidified atmosphere at 37 ˚C. Cultures were treated with different concentrations of adenosine (10 μM to 50 mM) or vehicle control for 24 hours. After incubation, cultures were then harvested and lysed for extraction of total RNA by Trizol extraction (Fig. 5.1). Normalized total RNA was reverse transcribed using random decamer primers and M-MLV reverse transcriptase. Gene-specific primers were used to amplify *RAB3C*, *FGD3*, and β-actin from cDNA samples for quantitative PCR (qPCR). Cycle threshold (Ct) values obtained from qPCR were plotted against a standard curve for each primer set to determine RNA levels for each condition. Expression of the housekeeping gene, β-actin, was used to monitor basal cellular activity. It should be noted that throughout these experiments, expression data were not normalized to β-actin levels unless otherwise noted. Instead, segments of the *RAB3C* or *FGD3* mRNA transcripts were used as a reference to determine intronic aptamer expression as compared to spliced mRNA. A map of the primer sets used to generate these gene-specific amplicons along *RAB3C* and *FGD3* is illustrated in Figure 5.2.

Using this approach, we found that the SH-SY5Y neuroblastoma cell line robustly expresses *RAB3C* across each condition while inconsistent expression was observed for *FGD3* (Fig. 5.2). This observation is consistent with findings in the literature where expression levels of *RAB3C* is 3.9 Transcript Per Million (TPM) and *FGD3* is less

Figure 5.1

Total RNA extraction and preparation

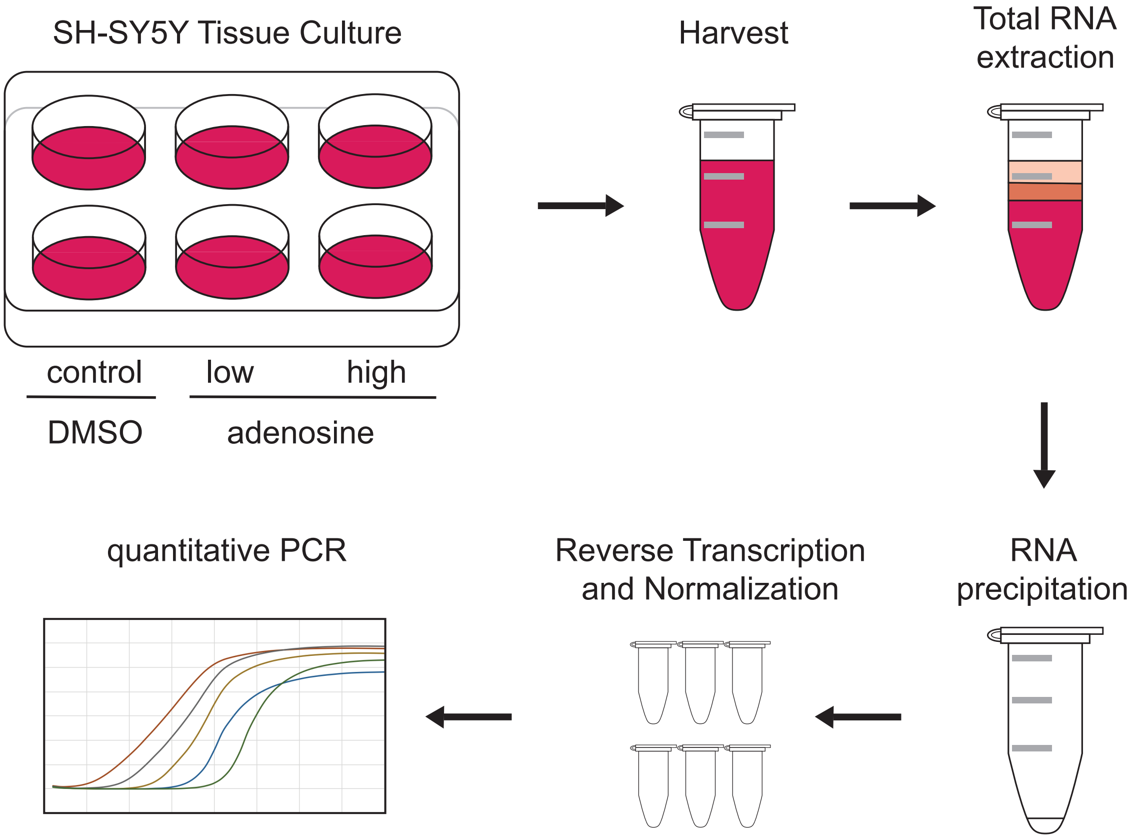
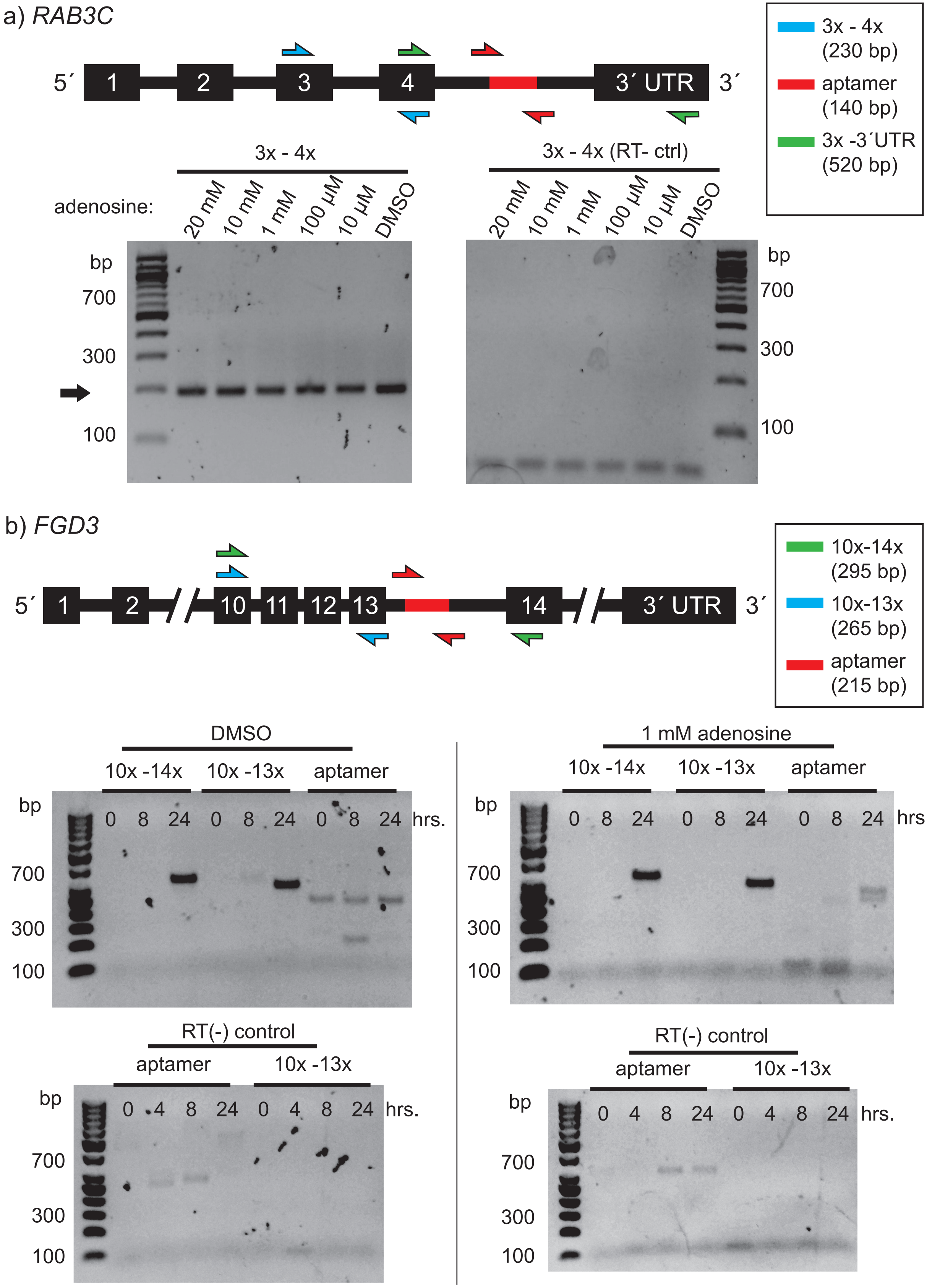


Figure 5.1 Illustrated above is a flowchart depicting the steps performed to extract, isolate and purify total RNA from SH-SY5Y after incubation with the desired small molecule compound. SH-SY5Y were grown to 80 % confluency prior to harvesting by trypsinization and centrifugation. Total RNA was isolated by phenol-chloroform extraction and precipitated using ethanol and sodium acetate. RNA pellets were washed, quantified and normalized to a final volume of 100 ng/µl prior to RT-qPCR. Using gene-specific primersets, Cq values for segments of *RAB3C* and *FGD3* RNA were measured and plotted against a standard curve.

Figure 5.2


a) RAB3C
5´
1
2
3
4
3´ UTR
3´
3x - 4x
(230 bp)
aptamer
(140 bp)
3x -3´UTR
(520 bp)
3x - 4x
3x - 4x (RT- ctrl)
adenosine:
20 mM
10 mM
1 mM
100 µM
10 µM
DMSO
20 mM
10 mM
1 mM
100 µM
10 µM
DMSO
bp
700
300
100
bp
700
300
100
b) FGD3
5´
1
2
10
11
12
13
14
3´ UTR
3´
10x-14x
(295 bp)
10x-13x
(265 bp)
aptamer
(215 bp)
DMSO
10x -14x
10x -13x
aptamer
bp
0 8 24 0 8 24 0 8 24 hrs.
700
300
100
1 mM adenosine
10x -14x
10x -13x
aptamer
bp
0 8 24 0 8 24 0 8 24 hrs.
700
300
100
RT(-) control
aptamer
10x -13x
bp
0 4 8 24 0 4 8 24 hrs.
700
300
100
RT(-) control
aptamer
10x -13x
bp
0 4 8 24 0 4 8 24 hrs.
700
300
100

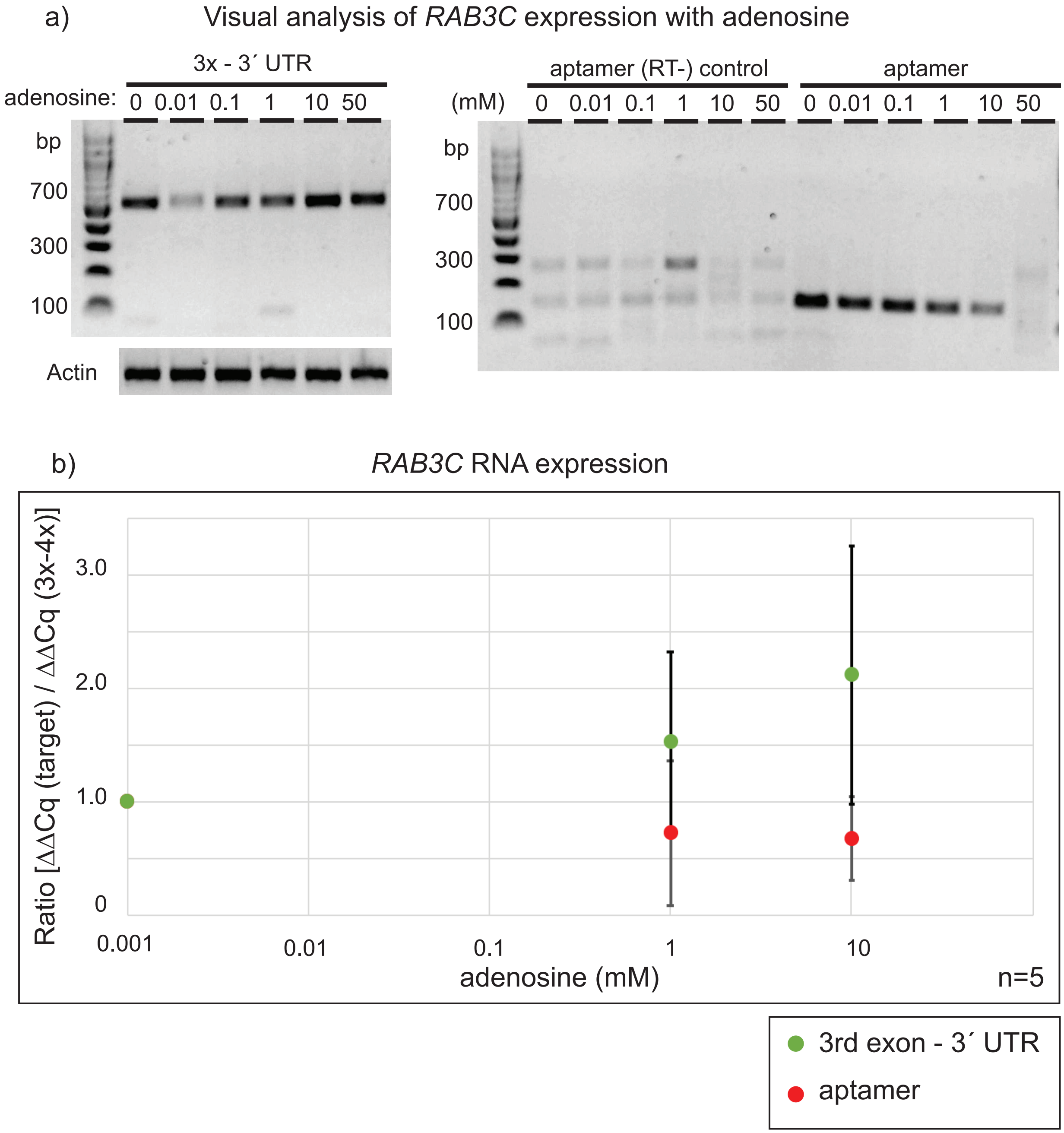
than 0.5 TPM in SH-SY5Y cell line [82], [83]. A standard RT-PCR amplification followed by gel electrophoresis shows that the *RAB3C* 3rd exon – 4th exon spliced mRNA product is stably expressed at all concentrations of adenosine. In a similar experiment performed for *FGD3*, *FGD3*-specific amplicons were not detected reliable despite multiple attempts to generate gene-specific primers. With difficulty in *FGD3* detection, a time course experiment was performed to determine if *FGD3* expression was dependent on cell-adherence. To do this, total RNA from SH-SY5Y that was incubated in the presence or absence of adenosine (1 mM) was extracted at 0, 8, and 24 hours after passaging cells. Figure 5.2b shows RT-PCR products obtained for *FGD3* in this time course experiment. Amplicons obtained in this experiment do not match the expected size of *FGD3*-specific products. At 8 hours and in the absence of adenosine, two amplicons were observed with the lower one corresponding to the *FGD3* aptamer. Interestingly, this band is not observed in the presence of adenosine. In either case, however, the nonspecific amplification requires that this experiment be interpreted cautiously. A number of *FGD3* mRNA variants are known to be expressed, and it could be that some of these products correspond to a larger unprocessed segment of nascent *FGD3* RNA. This argument is weakened by the absence of additional bands that indicate active processing with or without adenosine. With the lack of sequencing data, the results obtained here for *FGD3* RNA expression are inconclusive. For this reason, we focused additional experiments primarily on *RAB3C*.

III. Adenosine-dependent expression of the *RAB3C* aptamer

To determine whether changes in RNA levels for the *RAB3C* aptamer could be observed in an adenosine-dependent manner, two additional biological replicates were obtained. Cq values corresponding to the *RAB3C* aptamer, 3rd exon – 4th exon, and 3′ UTR amplicons were obtained by qPCR and observed by gel electrophoresis (Fig. 5.3). Gel analysis showed an inverse relationship in RNA levels between the *RAB3C* aptamer and *RAB3C* 3´ UTR amplicons (Fig. 5.3a). At 10 μM adenosine, the amplicon corresponding to the *RAB3C* 3´ UTR is present at low levels indicated by the faint 520 bp band (figure 5.3a bottom left panel). Conversely, the 140 bp amplicon corresponding to the *RAB3C* aptamer is saturated at 0 and 10 μM adenosine. The intensity of this band diminishes with increasing concentration of adenosine (Fig. 5.3a, bottom right panel). Using the *RAB3C* exon 3rd exon – 4th exon mRNA as a reference for baseline mRNA expression, the changes in *RAB3C* aptamer and *RAB3C* 3′ UTR RNA levels were plotted (Fig. 5.3, n=5). At 1 μM adenosine, both the *RAB3C* aptamer and the 3′ UTR amplicon begin to show differences in expression with respect to the upstream mRNA exons. At 10 mM adenosine, a 2-fold increase of the *RAB3C* 3′ UTR was detected by both qPCR and gel electrophoresis. In comparison, the *RAB3C* aptamer exhibited a small decrease at 1 mM adenosine and remains relatively unchanged at 10 mM adenosine.

Despite the biological replicates performed, this trend was not found to be statistically significant and is partly due to the low level of the *RAB3C* aptamer expression [84]. qPCR data collected (not shown) showed that *RAB3C* mRNA is expressed at least one order of magnitude less than that of actin control; however, the *RAB3C* aptamer is expressed ten-fold lower still. In fact, *RAB3C* aptamer is detected at sub-femtomolar concentration, creating challenges for total RNA extraction and recovery. In addition, this analysis does not take into account the expression of upstream RNA, including *RAB3C* 5´ UTR and exons 1 and 2. It is possible that by measuring *RAB3C* exons 3 and 4, such

Figure 5.3



a) Visual analysis of *RAB3C* expression with adenosine

b) *RAB3C* RNA expression

Figure 5.3 *RAB3C* adenosine aptamer expression analyzed by RT-qPCR. Total RNA isolated from SH-SY5Y cells incubated in various concentrations of adenosine were analyzed by gel electrophoresis after qPCR (a). Cq values corresponding to the *RAB3C* aptamer, and 3′ UTR amplicons were normalized against *RAB3C* 3rd exon - 4th exon mRNA levels. Error bars above represent the standard error of the mean of 5 biological replicates.

variant isoforms exons may also exhibit changes in expression with respect to the upstream exons as a function of adenosine concentration. This was evident while performing a similar analysis as above, but instead using the *RAB3C* 2nd exon – 3rd exon (data not shown) amplicon as a reference. Strangely, doing so revealed no change in *RAB3C* aptamer or 3′ UTR expression. While steady expression of the 2nd and 3rd exons across all adenosine concentrations was observed, the inverse relationship previously observed by the *RAB3C* aptamer and 3′ UTR were not observed suggesting that perhaps an isoform variant containing the adenosine aptamer is more abundant at low adenosine concentrations.

IV. Azathioprine and methotrexate screen on SH-SY5Y cultures

If *RAB3C* aptamer RNA expression is dependent on adenosine concentration as the previous observation suggests, then these changes should also be apparent with further reduction in adenosine levels. To test this hypothesis, SH-SY5Y cultures were incubated with two drugs known to inhibit de novo purine biosynthesis; azathioprine (AZA) and methotrexate (MTX). Azathioprine is a nucleoside analogue that is a direct inhibitor of the purine biosynthesis pathway [85]. Methotrexate is a folic acid analogue and a dihydrofolate reductase (DHFR) inhibitor, which is responsible for de novo synthesis of nucleosides [86], [87]. For this experiment, 23 nM and 230 nM of AZA, and 7.8 and 78 nM MTX were added to SHSY-5Y cultures to effectively reduce endogenous adenosine levels. The higher concentrations used for these compounds fall at approximately 10 % of IC_{50} and were chosen for their minimal effect on cell growth. After a 24-hour incubation, total RNA was extracted and *RAB3C* levels were measured using qRT-PCR as before.

Figure 5.4

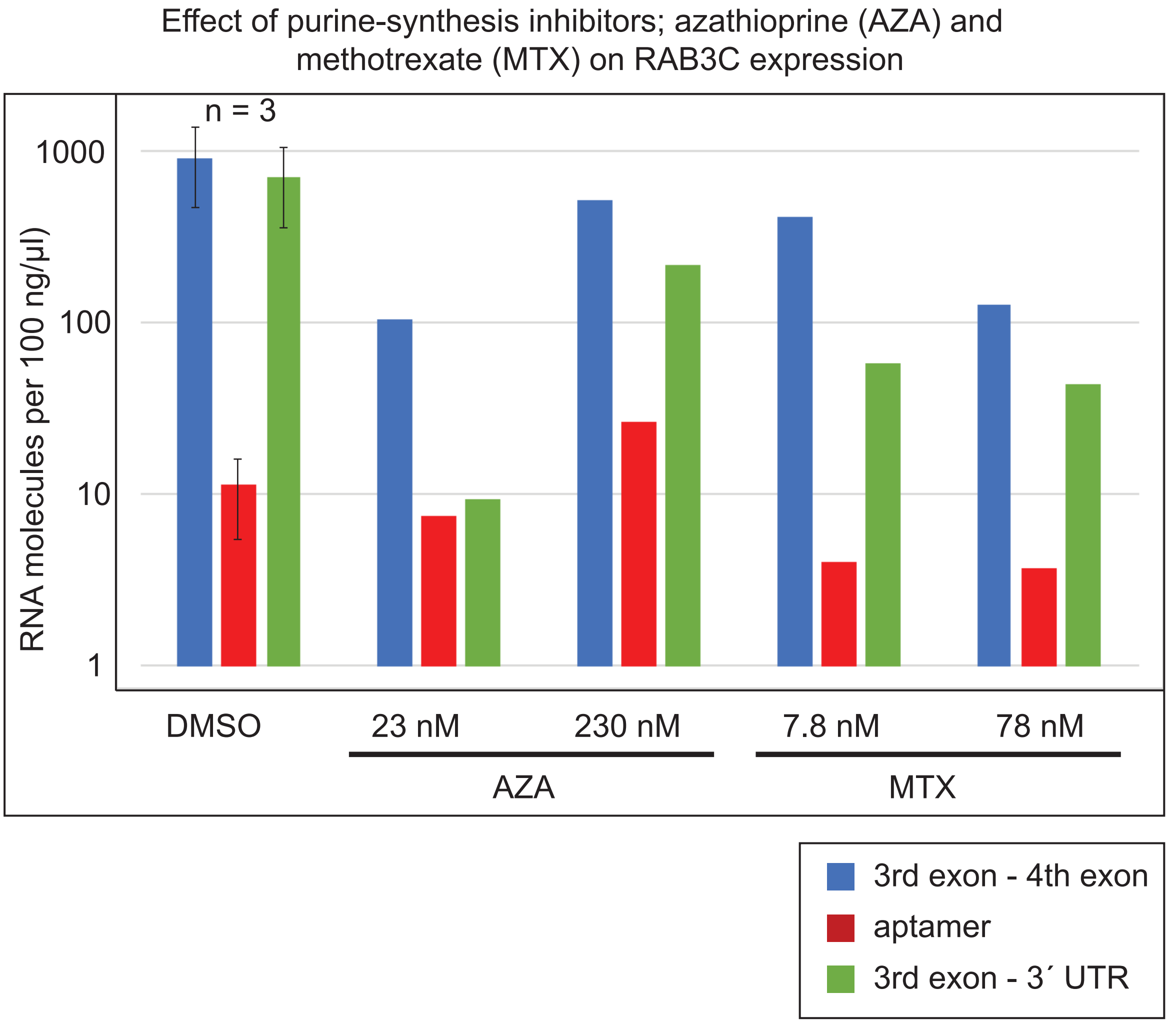


Figure 5.4. A SH-SY5Y *RAB3C* expression panel in the presence of low and high amounts of azathioprine (AZA) or methotrexate (MTX) is shown above. The DMSO vehicle control data is comprised of a biological triplicate, while the remaining samples represent the best of a biological duplicate dataset. In the presence of AZA, the ratio between the *RAB3C* 3´ UTR and *RAB3C* aptamer is reduced compared to the control. At 7.8 nM MTX, the *RAB3C* 3´ UTR appears to decrease significantly with respect to *RAB3C* 3rd - 4th exon, which remains at the same level as found in the control. These observations apparently agree with the trends observed previously in the presence of adenosine, however, additional replicates must be acquired in order to distinguish these observations from speculation.

Figure 5.4 shows *RAB3C* RNA levels measured in SH-SY5Y after a 24-hour period in the presence of AZA, or MTX. Note that the DMSO control represents an average of biological triplicates, while the remaining conditions represent the best data from a biological duplicate dataset. Furthermore, these data are presented as absolute RNA levels for each condition (Of note: Input total RNA is normalized prior to RT-PCR). At 23 nM AZA, no change in *RAB3C* mRNA or in *RAB3C* aptamer expression is detected compared to the DMSO control. At 230 nM AZA, however, the aptamer seems to be present at a higher level compared to the control. Interestingly, the 3rd exon – 3′ UTR product exhibits the lowest change in this condition and is consistent with the trend of expression made previously in the presence of adenosine. This observation is simply speculative as additional biological replicates are needed. Finally, a sustained decrease in *RAB3C* RNA levels were observed for both low and high MTX concentrations. This result may be due to MTX toxicity, which was repeatedly observed throughout the experimental trials and by light microscopy.

V. Column binding assay using SH-SY5Y total RNA

To analyze the *FGD3* and the THE1B adenosine aptamer binding activities in the context of the human transcriptome, we subjected total RNA isolated from SH-SY5Y cells incubated in either DMSO or 10 mM adenosine to an ATP column binding assay. Total RNA was equilibrated in binding buffer, introduced to ATP-agarose beads, and washed with 5 column volumes of binding buffer and eluted with free ATP. Elution fractions and the last wash were reverse-transcribed using primers specific for *FGD3* and THE1B and subsequently amplified using nested primers by qPCR.

Figure 5.5

Column binding assay using total RNA isolated
from SH-SY5Y

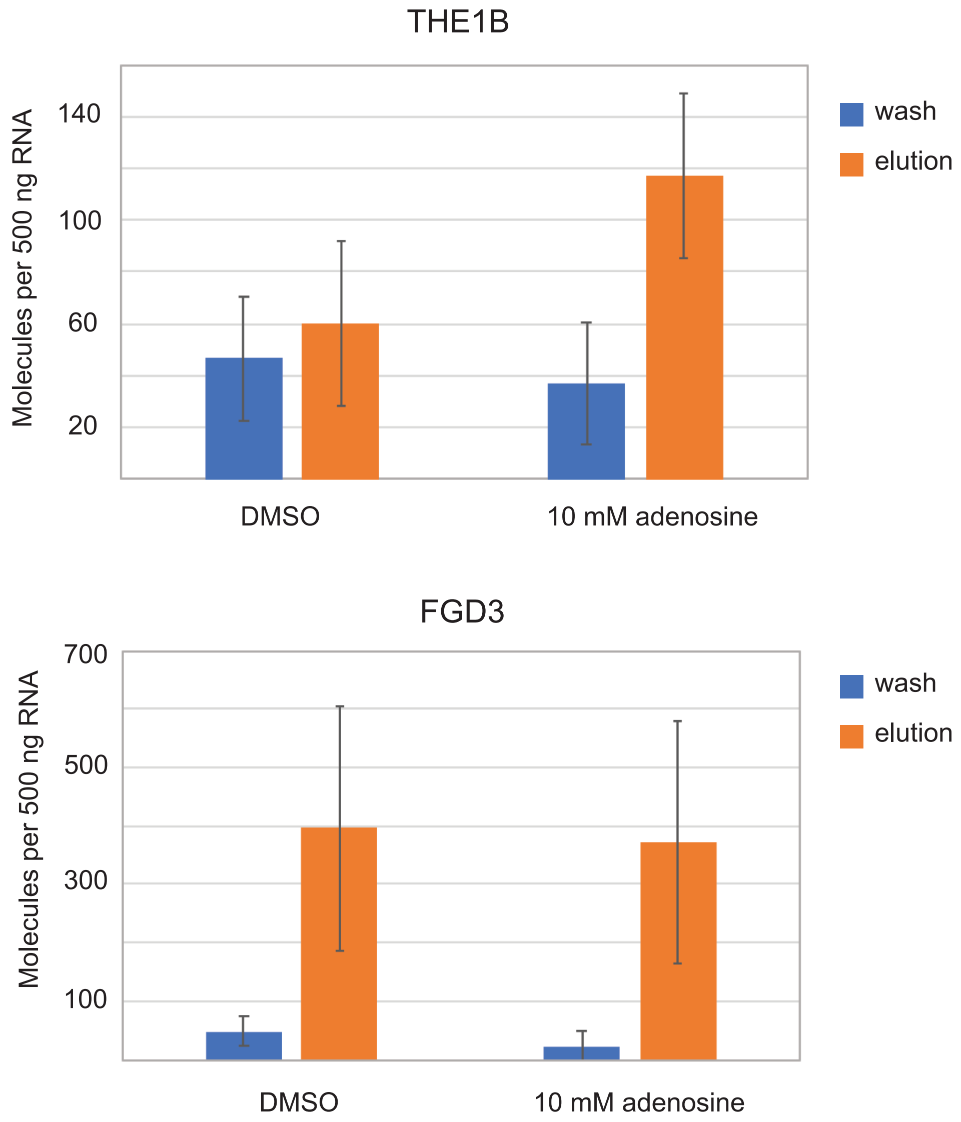


Figure 5.5 Binding activity of previously reported *THE1B* and *FGD3* adenosine aptamers [50]. Total RNA was isolated from the human SH-SY5Y cell line following incubation in the absence (DMSO) or presence of 10 mM exogenous adenosine. Starting with 500 ng of input total RNA, binding of the *FGD3* and *THE1B* aptamers were measured by performing RT-qPCR on the last wash and ATP elution fractions. The error bars shown represent the standard error of the mean of three biological replicates.

Both *FGD3* and THE1B aptamers bound and eluted from the ATP-agarose matrix. This was observed by detecting at higher levels of RNA within the elution fraction than preceding wash fraction (Fig. 5.5). Interestingly, the *FGD3* aptamer exhibited robust binding while the THE1B aptamer binding was only detected using RNA isolated from the SH-SY5Y cultures incubated in 10 mM adenosine. This result suggests that both the *FGD3* and THE1B aptamers derived from the human transcriptome have the capacity to bind ATP, and that binding activity of the THE1B aptamer is regulated by the presence of exogenous adenosine.

VI. Changes in SH-SY5Y cell structure

One interesting phenotype observed in these experiments was the accumulation of neuron-like processes when SH-SY5Y was subjected to adenosine and the production of spheroid clusters while in the presence of AZA. These changes in cell structure were observed by light microscopy and are shown in Figure 5.6. For this experiment, a population of cells were monitored over a period of 24 hours in the presence of 10 mM adenosine or 230 nM AZA. At the initial time point, SH-SY5Y cells are intermittently dispersed and exhibit an oblong cellular morphology with occasional neuron-like processes extending from the cell body. Over time and in the presence of adenosine, these processes become elongated and generate small networks of neuron-like extensions. In the presence of AZA, however, not only do these processes seem to be inhibited, but the cell body immediately begins to clump and round up over time then finally detach from the surface and remain as a purely suspended culture. In both cases, normal growth can be re-established upon reintroducing fresh medium without adenosine.

Figure 5.6

Light microscopy of SH-SY5Y cell culture

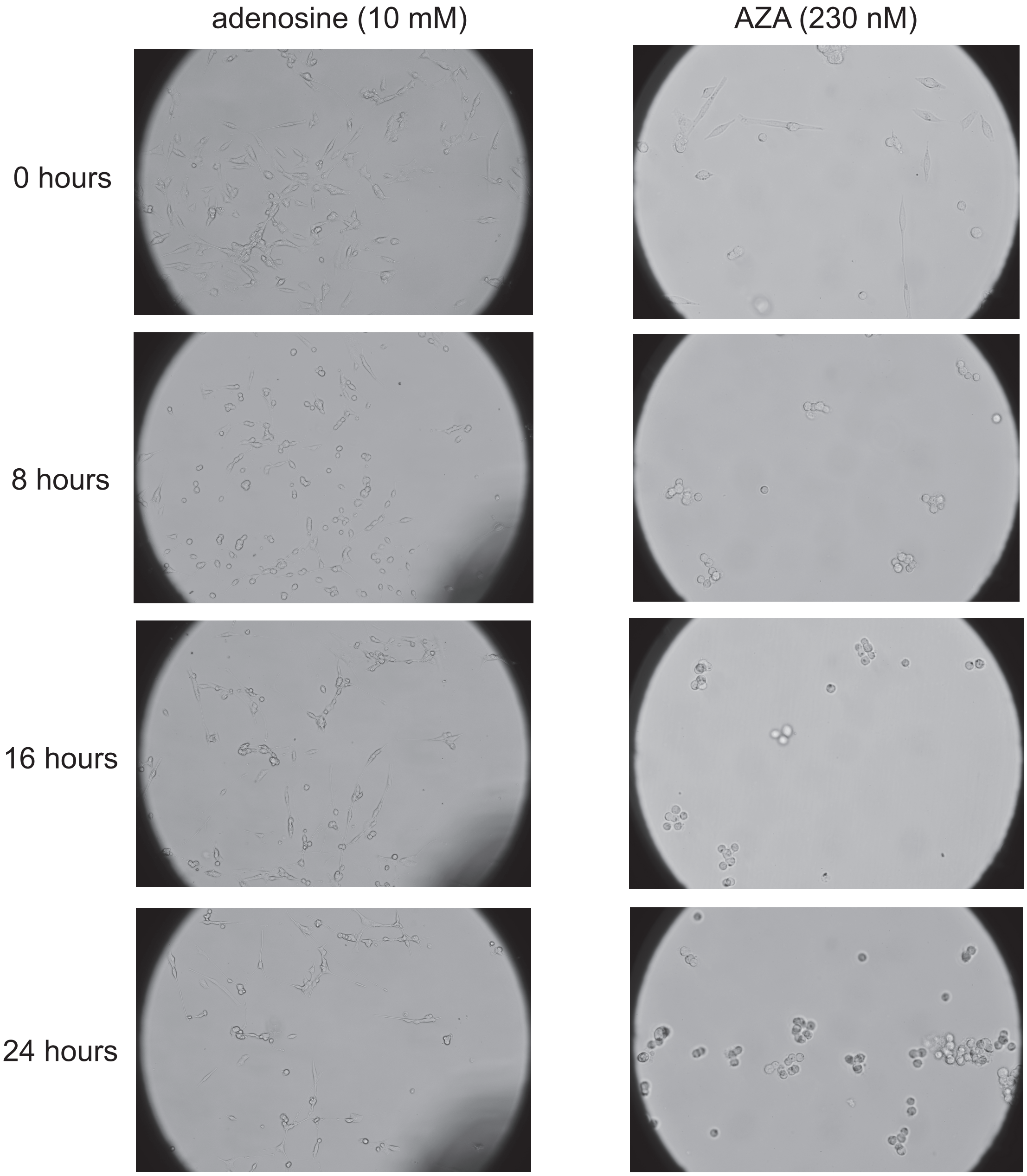


Figure 5.6 The SH-SY5Y cell culture visualized under a light microscope. Over a 24-hour incubation in the presence of adenosine or azathioprine (AZA), a series of images were taken using visible light. Striking differences in cellular morphology were observed. SH-SY5Y cells adopted a more neuron-like cell body in the presence of adenosine (right column), while those subjected to AZA generated clusters of aggregated spheriods (left column).

These results suggest that adenosine might play a role in regulating cellular morphology in SH-SY5Y cells, however, further studies are necessary to draw this conclusion.

VII. Conclusion

In this study, the expression of two naturally occurring adenosine aptamers within the human *RAB3C* and *FGD3* genes were studied closely. Both aptamers reside within an intron located near the 3´ ends of these genes and are predicted to play a role in adenosine-dependent regulation of their respective mRNAs. To determine if this is in fact true, RNA expression was measured by qRT-PCR using a neuroblastoma cell line incubated at various levels of adenosine and purine-synthesis inhibitors. Using gene-specific primers for mRNA and the adenosine aptamers, mRNA levels and their respective introns were monitored for adenosine-dependent changes.

Unfortunately, consistent *FGD3* expression was difficult to ascertain under normal conditions for the SH-SY5Y cell line. *FGD3* expression was highly variable over each biological replicate and consistent amplification of nonspecific amplicons made this study erratic. Recently, other types of cell lines including Karpas-707 or HMC-1, a lymphoid cancer cell line derived from the bone marrow and peripheral blood, respectively, have been shown to express *FGD3* and can be employed for studying the *FGD3* aptamer in the future [82], [83].

In the case of *RAB3C*, it was found that *RAB3C* aptamer expression decreased with the addition of exogenous adenosine. At the same time, transcripts containing the *RAB3C* 3´ UTR increased in the presence of adenosine, despite consistent mRNA expression up upstream exons measured at various adenosine conditions. Preliminary data using AZA, a purine-synthesis inhibitor, seem to maintain reduced levels of *RAB3C* 3´ UTR, and increased *RAB3C* aptamer levels. Most importantly, however, this adenosine-dependent result remains strictly correlational. Further experimentation such as antisense interference or genetic modification is required to establish if adenosine-aptamer binding is responsible for this effect. Collectively, these findings are a first step towards delineating the role of ATP aptamers in gene regulation *in vitro*.

CHAPTER 6

Materials and methods

I. Construction of a metagenomic DNA library

i. Total genomic DNA extraction and preparation

A total of 15 genetic model organisms were gathered to generate a metagenomic pool. Whole tissues from each organism were used to extract high molecular weight genomic DNA by crushing above liquid nitrogen using a mortar and pestle followed by phenol chloroform extraction and ethanol precipitation. Pelleted genomic DNA was washed with 75 % ethanol, allowed to dry then resuspended in 10 mM Tris-HCl pH 8.0. Each genomic DNA sample was transferred into a single test tube for adapter ligation preparation. During library preparation, it is critical that the starting genomic DNA stock consists of high molecular weight because the DNA will be subjected to sonication for fragmentation prior to adapter ligation. To maintain equal genomic diversity, C-values, or the mass number value representing one copy of a given genome, were used to pool the appropriate amount of genomic DNA. Traditional *in vitro* selections used a random synthetic DNA pool of $1x10^{15}$ sequences while a metagenomic DNA library comprising of these 17 organisms has a theoretical sequence diversity of $1x10^{10}$ (Fig. 2.1) [3], [5]. Diversity of the metagenomic DNA library can be approximated experimentally by quantifying initial amplification. This procedure is described further below.

ii. Fragmentation of genomic DNA

High molecular weight genomic DNA from each organism was aliquoted into a single 0.5 ml Eppendorf tube to achieve a final mass of 3 µg in a final volume of 130 µl in

10 mM Tris·HCl pH 8.0. This mixture was submitted to the University of California, Irvine Genomics High Throughput Facility (UCI-GHTF) for fragmentation *via* sonication. The UCI-GHTF uses a Covaris S2 acoustic shearer tuned to the appropriate parameters to shear high molecular weight genomic DNA to 300 bp ±100 bps. The efficiency of the sheared genomic DNA product was verified by running a small aliquot on a 2 % agarose gel stained with ethidium bromide (Fig. 2.3a). Sheared genomic DNA runs as a smeared band with the highest intensity near 300 bps.

iii. Repairing genomic DNA ends

Due to nonspecific shearing of dsDNA by sonication, single stranded overhangs that are generated after sonification are processed to generate double stranded blunt-ended products. To generate blunt ends, a nucleotide polymerization reaction was performed using 200 µM deoxynucleotide triphosphates, 50 mM NaCl, 10 mM Tris·HCl pH 8.0, 10 mM $MgCl_2$, 1 mM DTT, T4 DNA polymerase, and sheared genomic DNA. The reaction is then incubated at 12 °C for 15-30 minutes. The product was purified using a DNA binding column, eluted in double deionized water and quantified using a UV-visible light (UV-vis) spectrophotometer.

iv. Addition of 5´ phosphate group onto genomic DNA

DNA ligase requires a nucleotide monophosphate at the 5′ position of the ligation site. To prepare the genomic DNA for ligation, a 5´ phosphate group is added enzymatically using T4 phosphonucleotide kinase. The following reagents are combined in a PCR tube and incubated at 37 ˚C overnight: 70 mM Tris·HCl pH 7.6, 10 mM $MgCl_2$, 5 mM DTT, 5 % polyethylene glycol 8000, 2 mM spermidine, 200 nM dATP, end-repaired genomic DNA, 10 units of T4 polynucleotide kinase, and double deionized water. The phosphorylated products were then purified using a DNA binding column and eluted with double deionized water to be quantified by a spectrophotometer.

v. Addition of 3´ dA overhangs

To generate complementary cohesive ends needed for efficient for adapter ligation, deoxyadenosine overhangs are introduced to the 3´ ends of the genomic DNA. This reaction is set up using 15 units of the Klenow fragment of DNA polymerase I (3´-5´ exo-), phosphorylated genomic DNA product, 10 mM Tris·HCl pH 8.0, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, 100 µM dATP, and double deionized water for a final volume of 50 µL. The reaction is then incubated in a thermocycler at 37 °C for 15 minutes. The DNA product is then purified and quantified using a DNA binding column and a spectrophotometer.

vi. DNA adapter design

The DNA library adapter sequences consist of a total of four synthetic oligonucleotides. The left 5′ adapter (L5A) contains a T7 promoter for *in vitro* transcription, a 5´ blunt end, and a 3´ deoxythymidine (dT) overhang. For ligation purposes, a 5´ phosphate modification was added to the complementary left 3′ adapter (L3A) sequence. Similarly, the right 3′ adapter (R3A) contains a 5´ phosphate modification, and a 3´ blunt end. The complementary strand (R5A) was ordered with a 3´ dT overhang. In addition, the following requirements were considered during adapter design: adapter sequences should not be complementary so that they do not interfere with directional ligation of DNA and these sequences should not form primer dimers or amplification byproducts during PCR.

Left 5′ Adapter (L5A):

5′TAGATCTTAATACGACTCACTATAGGGAGACACTCTTTCCCTACACGACGCTCTT CCGATCT

Left 3′ Adapter (L3A):

5′GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCTCCCTATAGTGAGTCGTATTA AGATCTA

Right 3′ Adapter (R3A): GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

Right 5′ Adapter (R5A): CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

vii. DNA library adapter ligation

DNA oligonucleotides were purchased from a commercial supplier and synthesized with the appropriate 5´ phosphate group as needed to promote DNA ligation. To prepare for the adapter ligation onto end-repaired metagenomic DNA, each adapter oligonucleotide was diluted to a final 25 μM stock solution. The ligation reaction was set up in a clean PCR tube containing the entire volume of the 3´ dA genomic DNA product, 50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 1 mM ATP, 10 mM DTT, 30 units of T4 DNA ligase, 3 μL of 25 μM stock adapter solution and double deionized water to a final volume of 50 μl. After mixing, the reaction was incubated at 16 °C for 30 minutes and prepared for PCR amplification using a DNA purification kit.

viii. PCR amplification

After ligation of the library adapters, the DNA products were amplified using standard PCR. To maintain sequence diversity, the entire (purified) adapter ligated DNA product from the previous step was used to set up a PCR reaction containing: 10 mM Tris-HCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 200 µM dNTPs, and 2 µM of both forward and reverse primers (L5A and R5A). During PCR, an aliquot was collected at every four cycles for a total of 6 fractions over 24 PCR cycles to prevent excessive amplification of nonspecific byproducts. Each of these fractions was then visualized on a 2 % agarose gel for density analysis. For steps requiring gel excision for size-selection, amplicons corresponding to the desired length were excised using a nuclease free razor from a Sybr-gold stained agarose gel. The excised gel pieces were eluted in 300 mM KCl overnight, precipitated in 100 % ethanol, pelleted by centrifugation, and washed in 75 % ethanol. DNA pellet was resuspended in 10 µl 10 mM Tris-HCl, pH 8.0. To recover the excised fragments, re-amplification by PCR was performed to verify the size and molecular weight of the desired product.

viiii. Sequence diversity measurement

Sequence diversity was estimated using the band intensities generated by the PCR procedure mentioned above. The 6 fractions collected, which pertain to a 4-cycle fractionation series of a 24-cycle amplification of the metagenomic library, were loaded onto an ethidium bromide agarose gel along with a loading control and a standard (Fig. 2.3b). Using ImageJ software, pixel densities were extracted for the 300 bp amplicons and the loading control. Using the known mass value for the loading control, a molar quantity was estimated for the metagenomic library amplicon and corrected for sample dilution. Approximated mass values were then used to extrapolate the starting amount of DNA molecules by assuming efficient logarithmic amplification of 2^n, where n represents the number of cycles, and using Avogadro's constant.

II. Enrichment of RNA aptamers from a metagenomic library

i. Transcription

To prepare for round one of *in vitro* selection, a transcription reaction was set up containing 4 mM unlabeled rUTP/rGTP/rCTP, 250 µM unlabeled rATP, 3 nM α ^{32}P-labeled rATP, 16 mM $MgCl_2$, 10 % DMSO, 2 mM spermidine, 10 mM DTT, 0.01 % Triton-x, 40 mM Tris-HCl pH 8.0, 10 µl of purified metagenomic DNA, and one unit of T7 RNA polymerase. Transcription reactions were incubated at 37 ˚C for 3 hours and purified by polyacrylamide gel electrophoresis (PAGE) in the presence of 7 M urea. Bands corresponding to the full-length transcription products were excised and eluted over 3 hours in 300 mM KCl solution. An RNA pellet was collected after ethanol precipitation in the presence of glycoblue at -20 °C by centrifugation at 15000 g for 10 minutes. Pellets were washed three times with 75 % ethanol, dried, and resuspended in 50 µl of 1x binding buffer (BB) containing 140 mM KCl, 10 mM NaCl, 5 mM $MgCl_2$ and 20 mM Tris-HCl, pH 8.0. To ensure a large sequence diversity, round 1 of *in vitro* selection was performed using larger volumes (i.e. 500 µl transcription, 100 µl fractions). Subsequent rounds of selection were scaled down to 50 µl transcription volumes and 50 µl fractions.

ii. *In vitro* selection

C8-linked ATP-agarose beads, commercially available from Sigma Aldrich, were resuspended in 10 mM Tris-HCl pH 8.0 to a final concentration of 1 mg/ml. To prepare the beads, 30 µl of 1 mg/ml ATP-agarose beads were aliquoted onto a Spin-X spin-column and pre-equilibrated in 1x BB. All centrifugation steps for bead equilibration, and fraction collections were performed on an Eppendorf tabletop centrifuge at 2K x g for 1 minute. Fractions for selection were performed as follows; resuspended RNAs were aliquoted onto pre-equilibrated ATP-agarose beads and allowed to incubate at room temperature on a 3D rotator for 20 minutes to promote binding. This first flow-through fraction was collected by centrifugation and aliquoted into a clean PCR-tube. Next, ATP-agarose beads were resuspended in 50 µl of 1x BB and allowed to incubate on a 3D rotator for 5 minutes. These wash steps were repeated four times and individually collected into clean PCR-tubes. To collect elution fractions, ATP-agarose beads were resuspended in 50 µl of 1x elution buffer (EB) containing 5 mM ATP, 140 mM KCl, 10 mM NaCl, 10 mM $MgCl_2$ and 20 mM Tris-HCl, pH 8.0 and allowed to incubate for 30 minutes on a 3D rotator at room temperature. After collecting a total of 4 elution fractions, ATP-agarose beads were then presented with a harsh elution buffer containing 7 M urea and allowed to sit for 5 minutes before collecting in a clean PCR-tube. The remaining ATP-agarose beads were resuspended in 1x BB and collected into a clean PCR-tube. Radioactivity present within each fraction was then measured using a liquid scintillation counter and reported as CPM. To determine the fraction of eluted RNAs within a given selection experiment, the sum of the eluted fractions (CPM) were divided by the CPM sum of all fractions collected. After each round, CPM data is then reported as a fraction of RNAs eluted over each successive round.

iii. RT-PCR

Selected RNAs were subjected to reverse transcription (RT) by combining all four elution fractions in a 0.5 ml Eppendorf tube and precipitated with 100 % ethanol at -20 ˚C overnight. After centrifugation at 15000 g for 10 minutes, the RNA pellet was washed three times with 75% ethanol and allowed to dry for two hours at room temperature before resuspending in 20 µl of double deionized water. 10 µl of the RNA resuspension was added to a reverse transcription master mix containing 2 µM metagenomic DNA RT-primer (AL2093), 200 µM dNTPs, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 50 mM Tris-HCl, pH 8.3 to a final volume of 20 µl. To promote cDNA synthesis, the above mixture along with a no RT control were incubated at 90 ˚C for 30 seconds in a thermocycler to melt stable RNA structures prior to adding reverse transcriptase. While sitting at 4 ˚C, one unit of superscript III reverse transcriptase was added and mixed by pipetting to the appropriate sample. To help with extension of difficult sequences, a temperature gradient over time was programmed as follows: 24 ˚C for 10 minutes, 40 ˚C for 50 minutes, 50 ˚C for 15 minutes, 55 ˚C for 15 minutes and finally 70 ˚C for 15 minutes. The final cDNA product was then used as a DNA template for PCR amplification.

Using commercially available Dreamtaq Green by Thermo Scientific, 20 µl of the reverse transcription reaction were diluted into a PCR master mix containing 1x Dreamtaq Green, 1 µM forward and reverse primers (L5A and R5A). A total of 24 PCR cycles were performed on a BioRad thermocycler. To avoid excessive amplification of template cDNAs and to avoid nonspecific PCR-amplicons, the PCR reaction was fractionated by collecting an aliquot every 4 cycles into a clean PCR tube. After completion of 24 cycles of PCR, all fractions were loaded onto a 2 % Agarose gel visualized by ethidium bromide staining. To proceed with the next round of *in vitro* selection, the amplicon with the least number of amplification cycles as visualized by agarose electrophoresis was selected to serve as the starting DNA template for the next round of *in vitro* transcription. *In vitro* selection rounds were repeated until a significant number of radiolabeled RNAs were detected within the elution fractions of each respective target. At the completion of the *in* *vitro* selection, the DNA library generated by RT-PCR was used for vector cloning and for generating deep sequencing libraries.

Metagenomic DNA reverse transcription primer:

5′CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

iv. Co-transcriptional *in vitro* selection

For co-transcriptional binding assays, each selection step was modified to take into consideration the availability of free ATP in the reaction mixture. A 20 µl reaction was set up containing 40 % (w/v) C8 ATP-agarose beads, 10 µl of 100 nM purified metagenomic DNA library, 4 mM rUTP/rGTP/rCTP-mix, 250 µM unlabeled rATP, and 75 nM α^{32}P-labeled rATP, 16 mM $MgCl_2$, 10 % DMSO 2 mM spermidine, 10 mM DTT, 0.01 % Triton-x, 40 mM Tris-HCl pH 8.0. The reaction was mixed by pipetting in a 0.5 mL Eppendorf tube and allowed to incubate at 37 ˚C for 30 minutes. Before loading the reaction mixture on a filter column, 2 units of RNase-free DNase I was added. Fractions were collected by centrifugation at 2000 g using a clean PCR-tube. After the initial centrifugation step, four 20 µl bead washes were performed with 1x BB, followed by four 20 µl elution fractions containing 1x EB. A 20 µl harsh elution fraction containing 7M urea was also collected following the elution fractions. Beads remaining on the filter were resuspended in 20 µl of 7M urea.

v. Reverse transcription of co-transcriptionally enriched RNAs

To prepare the RNA for reverse transcription, each fraction was resuspended in gel-loading solution (7M urea, 0.01 % bromophenol blue, 0.01 % xylene cyanol) and loaded onto a 7 % denaturing (7M urea) polyacrylamide gel. Using a phosphorimager, gel bands corresponding to the elution fractions were excised and eluted at room temperature in 300 mM KCl for 3 hours. RNA was collected by ethanol precipitation and washed with 75 % ethanol prior to resuspension in RT buffer. The same reverse transcription steps were executed as previously mentioned in the standard *in vitro* selection. Briefly, each elution sample was reverse transcribed using reverse transcriptase Superscript III in the presence of 2 µM metagenomic DNA RT-primer. cDNA products were then PCR-amplified using Dreamtaq Green polymerase and monitored at every 4 cycles to prevent any over-amplification.

Metagenomic DNA reverse transcription primer:

5′CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

vi. Vector cloning of *in vitro* selection libraries

After *in vitro* selection, the enriched DNA libraries generated during RT-PCR were further amplified by scaling up PCR protocols in preparation for vector cloning. Similar to RT-PCR, 2 µL of cDNA was mixed into a fresh 20 µl PCR master mix containing 1x Dreamtaq Green, and 1 µM of metagenomic DNA primers (L5A and R5A). To facilitate vector ligation, the DNA product was allowed to incubate for an additional 10 minutes at 72 ˚C after PCR to ensure full extension of PCR amplicons as well as generate 3´ dA overhangs, which are required for cloning into the pCR 2.1-TOPO vector.

Vector cloning was performed as indicated by the TOPO-TA cloning kit manual provided by the manufacturer. In summary, 1 µl of fresh PCR product was mixed with 1.6 ng/µl pCR 2.1-TOPO vector in 200 mM NaCl, and 10 mM $MgCl_2$. The cloning reaction was incubated for 5 minutes at room temperature. Commercially available One Shot TOP10 competent cells were transformed by heat shock using a 42 ˚C heat block for 30 seconds and immediately replacing the cells on ice. After recovering the transformed

TOP10 cells by adding 250 µl of S.O.C. media and shaking at 200 rpm for 1 hour in a 37 ˚C incubator, 100 µl of the culture was plated onto pre-warmed (Luria-Bertani) LB agar culture plates containing 40 mg/ml X-gaL, and 50 µg/mL kanamycin. Agar plates were incubated in a 37 ˚C incubator overnight and screened for blue or white colonies. A total of 48 white colonies, which are indicative of a successful transformation of recombinant DNA, were picked for long term storage by resuspension in 50 % glycerol and 50 % S.O.C media. These glycerol stocks were stored at -80 ˚C and used for culture inoculation for plasmid preparation and colony PCR.

vii. Sanger sequencing

To prepare clones for sequencing, each colony was used to inoculate LB liquid media containing 1 mg/ml kanamycin. Each culture was then incubated on a 37 ˚C shaker for 6 hours before lysing and and extracting plasmid DNA using the Qiagen Miniprep kit. The concentration of purified plasmid DNA was determined using a UV-vis spectrophotometer and were used for subsequent dilution into a sequencing reaction. All Sanger sequencing reactions were submitted to the Genewiz sequencing facility. Each sequencing reaction was premixed prior to submission and contained: 50 ng/µl plasmid DNA, and 25 pmol of M10 (pCR™2.1-TOPO) primer in a total of 15 µl. After submission, sequence traces were downloaded from the Genewiz website to be analyzed and mapped onto the pCR™2.1-TOPO vector map using the Snapgene vector mapping software. This software allowed for extracting the recombinant DNA sequence of interest corresponding to a clone of interest obtained from *in vitro* selection. These sequences were saved as fasta format files for downstream analysis using NCBI-BLAST.

viii. Colony PCR and column binding assays

In a 20 µl PCR tube, the following reagents were prepared for colony PCR: 10 µl of 2x Dreamtaq Green, 1 µM final of metagenomic library primers (L5A and R5A), and approximately 0.5 µl of colony DNA (glycerol stock). As before, all PCR cycles were performed with a 63 ˚C annealing step and 1-minute extension periods. In addition, fractions were collected over every 4 cycles to prevent amplification of DNA byproducts. All fractions were visualized on a 2 % agarose gel stained with ethidium bromide. Amplified DNA from this PCR reaction were used as a template for *in vitro* transcription and subsequent column binding.

To prepare RNAs for column binding, radiolabeled RNAs were first gel purified on a 7 M urea, 7 % polyacrylamide gel ran at 20 watts for one hour to remove free ATP. Radiolabeled RNAs were imaged by autoradiography and excised from the gel. Gel bits were eluted in 300 mM KCl, shaking for three hours and precipitated with 100 % ethanol. Pelleted RNAs were carefully washed with 75 % ethanol and dried over two hours before re-suspending them in double deionized water. Alternatively, radiolabeled transcripts could also be passed through a 500 µl G50 Sephadex column equilibrated in 1x BB to remove free ATP. This Sephadex procedure was occasionally employed as an experimental alternative during optimization.

Column binding began by first equilibrating 30 µl of 10 mg/ml ATP-agarose beads, cAMP-agarose beads, or ADP-agarose beads by washing three times with 30 µl with 1x binding buffer (BB) containing 140 mM KCl, 10 mM NaCl, 5 mM $MgCl_2$, and 20 mM Tris-HCl, pH 8.0. Depending on the experiment, RNA was added to the beads in one of the following ways: 1) resuspended RNA was worked up in the presence of Mg^{2+} by adding

10x BB and then transferred onto the beads. 2) resuspended RNA was worked up in the presence of Mg^{2+} by adding 10x BB then heat annealed at 70 ˚C for 1 minute and allowed to cool to room temperature for 5 minutes before being transferred onto the beads; or 3) resuspended RNA was worked up in the presence of Mg^{2+} by adding 10x BB, heat annealed at 70 ˚C for 1 minute and immediately transferred onto the beads and allowed to cool to room temperature. In each case, the RNA-agarose bead mixture was allowed to incubate at room temperature for 10 minutes prior to collecting the first flow-through fraction in a clean PCR tube. The column was then washed four times with 30 µl of 1x BB and collected in clean PCR tubes. Bound RNAs were then eluted over a series of four 30 µl fractions with 1x EB. Any remaining RNAs that may be present after elution were then denatured and collected in a clean PCR tube using 7 M urea. All centrifugation steps for fraction collection were carried out on an Eppendorf tabletop centrifuge at 2k x g for 1 minute at room temperature. Fractions were then subjected to either gel electrophoresis and visualized by autoradiography or counted with a liquid scintillation counter.

III. Structure probing and deep sequencing of RNA libraries

i. Selective hydroxyl acylation and analysis by primer extension (SHAPE)

To prepare the RNA for acylation, an *in vitro* transcription reaction was set up as described previously to generate full length RNAs from a DNA library. To do this, the following components were mixed in a 0.5 mL Eppendorf tube and incubated at 37 ˚C for 1 hour: 4 mM rNTPs, 16 mM $MgCl_2$, 10 mM DTT, 10 % DMSO, 2 mM spermidine, 10 mM DTT, 0.01 % Triton-x, 40 mM Tris-HCl pH 8.0, and 5 µl of library DNA from a fresh PCR. RNA products were purified by denaturing acrylamide electrophoresis (7 % urea-PAGE), eluted with 300 mM KCl for 3 hours and precipitated using 100 % ice-cold ethanol. Pellets were washed three times with 75 % ethanol and allowed to dry over two hours. Once dried, the RNA pellet was resuspended in 30 µl of nuclease-free double deionized water and placed on a UV-vis spectrophotometer to determine the final concentration.

To prepare for acylation, a separate set of 0.5 ml Eppendorf tubes were prepared as a titration of each free ligand previously used in the selection. For example, for the ATP-enriched RNA library, a set of tubes containing a final concentration of either 100 nM, 1 µM, 10 µM, 100 µM, 1 mM, and 10 mM ATP were prepared prior to adding RNA. In a master mix, the RNA library was diluted to 1 µM final in the presence of 1x BB and heat-annealed at 70 ˚C for 1 minute. Before adding 50 mM of 2-(azidomethyl)nicotinic acid acyl imidazole (NAI), the RNA was equally aliquoted over 8 tubes including a no-acylation control and a no ligand control [57]. Upon adding NAI to each tube, the reaction was incubated at room temperature for 30 minutes before quenching with 10x reaction volume using double deionized water. This process was repeated for the cAMP-enriched RNA library using a titration of free cAMP at 100 fM, 1 nM, 10 nM, 100 nM, 1 µM and 10 µM final.

The acylated RNA products were then prepared for reverse transcription by first performing an overnight precipitation using 300 mM KCl and 100 % ethanol. After three successive washes with 75 % ethanol and drying by vacuum aspiration, the RNA pellets were resuspended in a reverse transcription master mix containing 50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 5 mM dNTPs, 500 nM SHAPE-sequencing metagenomic reverse transcription primer and 10 units of superscript reverse transcriptase.

Metagenomic library SHAPE-sequencing reverse transcription primer:

5′AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTGCGGCCGCGTGACTGGAGTT
CAGACGTGTGCTCTTCCG

Following reverse transcription, a single-stranded DNA circular ligation reaction was set up using the entire volume of the completed reverse transcription reaction. This circular ligation step is required to amplify the cDNA library corresponding to the acylation modification events in the previous step by PCR (Fig. 4.1). To set up this ligation reaction, 50 mM MOPS pH 7.5, 10 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 50 µM ATP, 2.5 mM $MnCl_2$, 10 µl of reverse transcription cDNA product and 5 units/µl Circligase ssDNA ligase were well mixed and incubated at 60 ˚C for 5 hours on a thermocycler. Before PCR amplification, all reactions were ethanol precipitated once more to collect and wash the cDNA pellet. Upon re-suspending the cDNA pellet in nuclease-free double deionized water, this final product was used as a template for PCR in preparation for high throughput sequencing.

ii. Generating Illumina complaint DNA libraries for deep sequencing

Outlined here are the steps taken to prepare metagenomic DNA libraries for deep sequencing on the Illumina HiSeq 2500 instrument. It is important to note that the specific DNA oligos used for this procedure may vary depending on the sequencing platform used. These differences are attributed to the compatible flowcell chip used for a given sequencing instrument. The adapter sequences presented here are intended to work for the metagenomic DNA library described above and the flowcell chip model used for deep sequencing using the Illumina HiSeq 2500 sequencer.

To generate a compatible library for the Illumina flowcell, a two-step PCR primer extension was performed prior to sample submission. Buffer conditions for each PCR step can be found in the RT-PCR methods mentioned above. For all PCR amplification steps, a 50 µl reaction was set up using Dreamtaq Green supermix with a universal forward primer. For the first PCR step, the universal forward primer and a library-specific reverse primer were used. These primers include the primer-binding site required for sequencing both sense and antisense strands. In addition, this step also standardizes all samples for the barcoding step (step 2). If required, the amplified products from step 1 PCR can be pre-purified at this step before proceeding to step 2 to exclude self-ligated DNA artifacts that may have been introduced from the single-stranded ligation step during SHAPE-reverse transcription and ligation. For step 2, each sample is then primer extended and amplified using a unique reverse primer that corresponds to a pre-assigned barcode. The barcoding reverse primer is provided below where N's represent the 6-nucleotide assigned barcode. With this barcode assignment, the sequencing reads can be binned and analyzed independently by experimental condition. In other words, a unique barcode can be assigned to a specific enriched DNA library or a specific ligand titration experiment for SHAPE-sequencing. Prior to submitting barcoded DNA libraries, the final PCR-extended amplicons are first visualized by gel electrophoresis, pooled, then further analyzed by qPCR to ensure proper amplification. After confirmation, the pooled sample is then submitted to the sequencing facility.

Universal forward primer:

5′AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Metagenomic library – specific reverse primer (step 1 reverse primer):

5′GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Barcoding primer (step 2 reverse primer):

5′CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCT CTTCCG

IV. Computational analysis of SHAPE and deep sequencing reads

i. Computer Hardware Requirements

Scripts, programs and data visualization software were run on a desktop computer equipped with an AMD 2.4 Ghz quad-core processor with 8 Gb RAM running the Ubuntu 16.04 LTS operating system. All steps, with the exception of tabulated spreadsheets, were executed using the Linux/UNIX bash terminal and the Python 3.6.1 programming language. When appropriate, computationally intensive algorithms were executed on the UC Irvine High Performance Cluster (HPC) using Sun-Grid Engine compatible scripts to call alignment and trimming programs. Data files produced by the HPC were then transferred to a local desktop server for long-term storage and downstream processing. In addition, structure probing modules from The Galaxy Project (www.usegalaxy.org) online server were used to access RT-stop data and generate SHAPE reactivity datasets [88]. Trimming, alignment of reads, and generating SHAPE reactivities are described in further detail below.

ii. Removing synthetic adapters from sequencing reads

To remove artificial sequences introduced by library and deep sequencing preparation, two pattern-searching algorithms, encoded by the trimmomatic and cutadapt programs, were used to systematically remove metagenomic DNA adapter sequences and self-ligated DNA artifacts. Both programs are readily available online and can either be loaded onto The Galaxy Project online server or downloaded for execution on a local server. The trimmomatic and cutadapt programs use flexible pattern-searching algorithms that can identify, modify, remove, or extract a user-provided sequence pattern from a sequencing read file. Both programs iterate over a user-provided deep sequencing file based on two different pattern-searching computational theories [89], [90]. The trimmomatic program searches for patterns that match a user-provided adapter sequence file and accounts for differences of up to 10 % mismatch. The program scans sequencing reads in the 5′ to 3′ direction and removes matching segments such that the overlapping sequence and any upstream sequence is removed from the sequence read. This generates a list of downstream truncated sequence reads of variable length that are further processed by cutadapt.

While this is a very efficient process, this sometimes results in truncated reads that are less than 20 nucleotides in length and are not desired for generating downstream alignments. The cutadapt program was used omit these short sequences and additional artificial sequences introduced by RT read-throughs due to circular ligation (Fig. 4.1). The command lines provided below were executed to remove sequences that match the user provided DNA sequence. Conditional argument descriptions can be within the respective user manuals [89], [90]. In summary, both algorithms require a minimum overlap of 25 nucleotides and recognize sequences with up to a 10 % mismatch. This process generates a list of downstream sequence reads of variable length referred in the text as trimmed sequence reads (Fig. 4.2). A statistical report containing a detailed summary of processed reads are reported Table 4.1.

Trimmomatic command line:

```
java -jar /data/apps/trimmomatic/0.35/trimmomatic-0.35.jar SE -phred33 -trimlog
READ1-trimlog.txt READ1-Sequences.fastq READ1-trimmed.fastq
ILLUMINACLIP:READ1-adapters.fa:2:30:8
```

Cutadapt command line:

```
cutadapt -b GGGCAGACGTGCCTCACTAC --discard-trimmed -m 25 -o
/READ1-trimmedandcut.fastq READ1-trimmed.fastq
```

iii. Counting and clustering sequence reads.

After adapter trimming, the fastaptamer count and clustering programs were used to consolidate redundant sequencing reads into a counted fasta file. These fastaptamer programs are a part of the fastaptamer toolkit pipeline designed examine sequencing data, and can be downloaded from GitHub or The Galaxy Project online server [88], [91]. The fastaptamer count and clustering algorithms were used to count, normalize, rank unique sequences by abundance, and cluster sequences into related families [91]. Seed sequences, or the most abundant sequence taken from each family, were used for genomic alignments using the NCBI nucleotide BLAST server to determine the organism from which they were derived. These results are reported in Tables 4.2 and 4.3. Because the clustering algorithm is a computationally intensive process, clustering was only performed for the deep sequencing reads corresponding to the enriched DNA libraries.

Due to high sequence diversity, clustering of the naïve metagenomic library deep sequencing reads was not efficient.

fastaptamer count command line:

```
fastaptamer_count -i READ1-trimmedandcut.fastq -o /READ1_trimcounted.fasta
```

fastaptamer cluster command line:

```
fastaptamer_cluster -i READ1-trimcounted.fasta -d 15 -o READ1-clustered.fasta
```

iv. Sequence searches and enrichment calculations

To determine sequence enrichment, seed sequences highlighted by the fastaptamer clustering algorithm were identified within the naïve metagenomic library using text-based searches. This was done to extract the reads per million (RPM) normalization value required calculate enrichment values from the sequence header (for additional information, refer to the fastaptamer cluster user manual). To retrieve the RPM values, the Global Regular Expression Print (grep) function was invoked in a Linux/UNIX terminal shell for every clone as follows:

```
grep -B 1 AGATCGG READ1-clustered.fa
```

where AGATCGG represents the subject query (i.e. clone sequence), '-B 1' denotes retrieval of the preceeding line containing the sequence header, and READ1-clustered.fa is the input text file for which the search algorithm is iterated over. Matches are immediately printed to standard output and can be copied into an excel spreadsheet. Once RPM values were obtained from both the naïve and the respective enriched library for each clone, enrichment was calculated:

$$\text{Sequence enrichment} = \text{RPM}_{\text{enriched}} / \text{RPM}_{\text{naïve}}$$

v. Alignment of SHAPE and deep sequencing reads.

Sequence alignments were generated using the bowtie2 version 2.3.2 software. Bowtie2 is a well-established and flexible alignment program that has been used in many applications including ChIP-seq and RNA-seq [92]. The program features modes of permissive, strict, and custom alignment scoring parameters depending on the quality of the input sequencing reads. Bowtie2 generates binary alignment map (BAM) files, which contain sequence read identifiers, map coordinates along a reference sequence, alignment quality scores, mismatching nucleotides, and even base-call quality scores from the Illumina sequencer (fastq input required).

For deep sequencing alignments, such as those performed on rDNA reference sequences, we threaded trimmed sequence reads into the bowtie2 program using the default strict (local) alignment parameters. Under these constraints, reads with differences greater than 10 % are ignored and all matching sequences must exhibit 90 % match identity with the reference sequence. These details and how they are implemented in the command line can be found in the bowtie2 user manual [92].

For SHAPE-sequencing reads, enriched sequences identified in Tables 4.2 and 4.3 were used as reference sequences. Trimmed SHAPE-sequence reads were aligned using the bowtie2 program using the default permissive (soft) alignment parameters. Here, the 5′ and 3′ ends of sequence reads are soft-trimmed to achieve the best possible match identity with the reference sequence. This alignment mode, however, still requires 90 % match identity, and a minimum of 25 matching nucleotides. Sequences that have been soft-trimmed also display a reduced alignment score which can be observed using the visual graphic interface described further below.

Bowtie2 local (strict) alignment command line:

```
bowtie2 --local -N 1 -x /reference-index.dna -U /READ1-trimcounted.fa -S /READ1-aligned.bam
```

Bowtie2 default (soft) alignment command line:

```
bowtie2 -N 1 -x /reference-index.dna -U /READ1-trimcounted.fa -S / READ1-aligned.bam
```

vi. Visualization of sequence alignments

BAM files were visualized using the Integrated Genome Browser (IGB), an open source genome visualization tool [93]. The IGB can automatically call full sequence assemblies from a genomic database available on an online server to be loaded for whole-genome alignment files. In addition, BAM files containing sequence alignment data can also be loaded separately onto the IGB. The user-friendly interface allows BAM files to easily be loaded directly onto the IGB. In the visualization window, BAM alignment data is presented on an abundance (y-axis) versus reference coordinate (x-axis) bar graph (Fig. 4.11). Data points from these illustrations can be extracted, if needed, as tabulated values in a text file.

vii. Normalization and calculation of SHAPE reactivities

To extract RT-stop values from SHAPE-sequencing reads, BAM files were uploaded onto the Galaxy server and passed through the GetRTstops module. Briefly, this module generates a list of tabulated values consisting of RT stop frequency for every nucleotide position along the reference sequence. Table values are downloaded from the server and transferred onto an Excel spreadsheet using the transform_stop_counts.py program for statistical analysis.

For each dataset, raw RT stop values were first normalized to the mean of the 95th percentile values. RT-stop values along the first and last 30 nucleotides of the reference sequence were excluded to omit RT biases. Reactivity is defined as the difference in RT stops for a given position in the presence of the SHAPE-reagent to that of background RT stops (No SHAPE-reagent) control.

$$\text{Reactivity} = \text{RTstops}_{\text{(NAI)}} - \text{RTstops}_{\text{(DMSO)}}$$

Reactivity scores are adjusted on a scale from [-1, 1] by setting the top 5 % of values to 1 and the bottom 5 % of values to -1. On this scale, negative values correspond to nucleotides protected from chemical modification compared to natural RT stops (background), while positive values denote accessible nucleotides.

V. Cell culture expression assays

i. SH-SY5Y cell culture preparation

The SH-SY5Y cell line was thawed from cryo-preservation and seeded onto separate T-75 culture plates with DMEM media containing 10 % Fetal Bovine Serum (FBS), 10 % Amphotericin B, and 10 % penicillin / streptomycin. The culture was maintained in a 5 % CO_2 incubator with a humidified atmosphere at 37 ˚C. Using a cytometer, cell cultures were passaged appropriately to achieve 80 % confluency on the day of the experiment. The addition of 2 % DMSO + 10 mM adenosine was presented to adhered confluent cells one day prior to total RNA extraction.

ii. Total RNA extraction

Total RNA was harvested from SH-SY5Y using commercially available Ambion Trizol Reagent. All total RNA isolation steps were performed as listed in the provided user manual. Briefly, adhered cells were collected by washing each culture dish with 750 µl of Trizol Reagent and collected into a fresh 5 ml Eppendorf tube. 200 µl of chloroform was added to each sample to allow for phase separation. Extraction of the aqueous later was followed by RNA precipitation using 100 % isopropanol. The total RNA was pelleted by centrifugation and washed using 75 % ethanol prior to resuspension in RNase-free ddH_2O. Once resuspended, total RNAs were treated with DNase I to remove genomic DNA. The TRIzol extraction procedure was then repeated for all DNase I – treated total RNA samples to avoid protein contaminants for RT-qPCR.

iii. RT-qPCR of total RNA extracts

Prior to qPCR, purified total RNA was quantified by a UV-vis nanodrop spectrophotometer and normalized to 100 ng/µl. A 10 µl reaction containing 500 µM dNTPs, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 50 mM Tris-HCl, pH = 8.3 and 1 µM random decamer oligos (N_{10} reverse transcription) were well mixed and transferred to a thermocycler. To promote cDNA synthesis, the above mixture along with a no RT control were incubated at 90 ˚C for 30 seconds to melt stable RNA structures prior to adding reverse transcriptase. While sitting at 4 ˚C, one unit of Murine-MLV reverse transcriptase was added and mixed by pipetting to the appropriate sample. To help with extension of difficult templates RNA was incubated as follows: 24 ˚C for 10 minutes, 40 ˚C for 50 minutes, 50 ˚C for 15 minutes, 55 ˚C for 15 minutes and finally at 70 ˚C for 15 minutes.

The final cDNA product was then diluted 10-fold with double deionized water and used as a DNA template for qPCR.

For qPCR, white low-profile 96-well plates were used on a BioRad CFX connect real-time PCR detection system. First, a qPCR master mix containing 1x BioRad iTaq qPCR supermix and 300 nM gene-specific primers was premixed in a 0.5 mL Eppendorf tube. 10 % (v/v) of diluted cDNA product was added to the qPCR master mix then carefully aliquoted along the 96-well plate on ice. Reactions were mixed by tapping and centrifuged on a table top centrifuge for 1 min at 1 k x g after being sealed with optically clear adhesive.

iv. Adenosine, azathioprine (AZA) and methotrexate (MTX) screens

After passage into a fresh T-75 culture plate, SH-SY5Y cell cultures were incubated until 80 % confluency was achieved. On the day of the experiment, cells were treated with either; 10 nM, 100 nM, 1 mM, 10 mM, or 50 mM adenosine; 23 nM AZA, 230 nM AZA, 7.8 nM MTX, or 78 nM MTX premixed with fresh DMEM culture media containing 10 % Fetal Bovine Serum (FBS), 10 % Amphotericin B, 10 % penicillin / streptomycin, and 2 % DMSO. After 24 hours, SH-SY5Y cultures were harvested for total RNA extraction and analyzed by RT-qPCR as described above.

v. Light Microscopy

Images were taken using a Nikon D5100 digital camera mounted on an inverted microscope. Using the 10X objective, the tissue cultures were for adhered or suspended SH-SY5Y cluster aggregation. Images were taken under the 20X objective under visible light and processed using the Adobe Photoshop imaging software.

REFERENCES

[1] R. Stoltenburg, C. Reinemann, and B. Strehlitz, “SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands,” *Biomol. Eng.*, vol. 24, no. 4, pp. 381–403, Oct. 2007.

[2] S. D. Jayasena, “Aptamers: an emerging class of molecules that rival antibodies in diagnostics,” *Clin. Chem.*, vol. 45, no. 9, pp. 1628–1650, 1999.

[3] A. Ellington and J. W. Szostak, “In vitro selection of RNA molecules that bind specific ligands.,” *Nature*, vol. 346, no. 6287, pp. 818–822, Aug. 1990.

[4] A. R. Ferre-D’Amare and W. G. Scott, “Small Self-cleaving Ribozymes,” *Cold Spring Harb. Perspect. Biol.*, vol. 2, no. 10, pp. a003574–a003574, Oct. 2010.

[5] C. Tuerk and L. Gold, “Systematic Evolution of Ligands by Exponential Enrichment: RNA ligands to Bacteriophage T4 DNA Polymerase,” *Science*, vol. 248, no. 4968, pp. 505–510, Aug. 1990.

[6] D. H. Burke and L. Gold, “RNA aptamers to the adenosine moiety of S-adenosyl methionine: structural inferences from variations on a theme and the reproducibility of SELEX,” *Nucleic Acids Res.*, vol. 25, no. 10, pp. 2020–2024, 1997.

[7] M. Sassanfar and J. W. Szostak, “An RNA motif that binds ATP.,” *Nature*, vol. 364, no. 6437, pp. 550–553, Aug. 1993.

[8] P. Burgstaller and M. Famulok, “Isolation of RNA aptamers for biological cofactors by in vitro selection,” *Angew. Chem. Int. Ed. Engl.*, vol. 33, no. 10, pp. 1084–1087, 1994.

[9] T. Dieckmann, E. Suzuki, G. K. Nakamura, and J. Feigon, “Solution structure of an ATP-binding RNA aptamer reveals a novel fold.,” *RNA*, vol. 2, pp. 628–640, 1996.

[10] T. Dieckmann, S. E. Butcher, M. Sassanfar, J. W. Szostak, and J. Feigon, “Mutant ATP-binding RNA Aptamers Reveal the Structural Basis for Ligand Binding,” *J. Mol. Biol.*, vol. 273, pp. 467–478, Oct. 1997.

[11] R. Stoltenburg, C. Reinemann, and B. Strehlitz, “SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands,” *Biomol. Eng.*, vol. 24, no. 4, pp. 381–403, Oct. 2007.

[12] E. A. Curtis and D. R. Liu, “Discovery of Widespread GTP-Binding Motifs in Genomic DNA and RNA,” *Chem. Biol.*, vol. 20, no. 4, pp. 521–532, Apr. 2013.

[13] M. M. K. Vu, N. E. Jameson, S. J. Masuda, D. Lin, R. Larralde-Ridaura, and A. Lupták, “Convergent Evolution of Adenosine Aptamers Spanning Bacterial, Human, and Random Sequences Revealed by Structure-Based Bioinformatics and Genomic SELEX,” *Chem. Biol.*, vol. 19, no. 10, pp. 1247–1254, Oct. 2012.

[14] W. C. Winkler, A. Nahvi, N. Sudarsan, J. E. Barrick, and R. R. Breaker, “An mRNA-structure that controls gene expression by binding S-adenosylmethionine,” *Nat. Struct. Biol.*, vol. 10, no. 9, pp. 701–707, Nov. 2003.

[15] M. Sassanfar and J. W. Szostak, “An RNA motif that binds ATP.,” *Nature*, vol. 364, no. 6437, pp. 550–553, Aug. 1993.

[16] J. Miranda-Ríos, M. Navarro, and M. Soberón, “A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 17, pp. 9736–9741, 2001.

[17] A. Wachter, M. Tunc-Ozdemir, B. C. Grove, P. J. Green, D. K. Shintani, and R. R. Breaker, “Riboswitch Control of Gene Expression in Plants by Splicing and Alternative 3’ End Processing of mRNAs,” *PLANT CELL ONLINE*, vol. 19, no. 11, pp. 3437–3450, Nov. 2007.

[18] M. T. Cheah, A. Wachter, N. Sudarsan, and R. R. Breaker, “Control of alternative RNA splicing and gene expression by eukaryotic riboswitches,” *Nature*, vol. 447, no. 7143, pp. 497–500, May 2007.

[19] Y. Li *et al.*, “Development of RNA Aptamer-Based Therapeutic Agents,” *Curr. Med. Chem.*, vol. 20, no. 29, pp. 3655–3663, 2013.

[20] B. S. Singer, T. Shtatland, D. Brown, and L. Gold, “Libraries for genomic SELEX,” *Nucleic Acids Res.*, vol. 25, no. 4, pp. 781–786, 1997.

[21] K. Salehi-Ashtiani, “A Genomewide Search for Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene,” *Science*, vol. 313, no. 5794, pp. 1788–1792, Sep. 2006.

[22] R. Kubota *et al.*, “Double-Strand Breaks in Genome-Sized DNA Caused by Ultrasound,” *ChemPhysChem*, vol. 18, no. 8, pp. 959–964, Apr. 2017.

[23] B. P. Hennig *et al.*, “Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol,” *G3amp58 GenesGenomesGenetics*, vol. 8, no. 1, pp. 79–89, Jan. 2018.

[24] S. Picelli, Å. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, and R. Sandberg, “Tn5 transposase and tagmentation procedures for massively scaled sequencing projects,” *Genome Res.*, vol. 24, no. 12, pp. 2033–2040, Dec. 2014.

[25] A. Adey *et al.*, “Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition,” *Genome Biol.*, vol. 11, no. 12, p. R119, 2010.

[26] L. M. Gómez Ramos *et al.*, “Eukaryotic Ribosomal Expansion Segments as Antimicrobial Targets,” *Biochemistry (Mosc.)*, vol. 56, no. 40, pp. 5288–5299, Oct. 2017.

[27] L. M. Gómez Ramos *et al.*, "Yeast rRNA Expansion Segments: Folding and Function," *J. Mol. Biol.*, vol. 428, no. 20, pp. 4048–4059, Oct. 2016.

[28] J. J. Cannone *et al.*, "The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs," *BMC Bioinformatics*, vol. 3, no. 1, p. 2, 2002.

[29] R. E. Pruitt and E. M. Meyerowitz, "Characterization of the genome of Arabidopsis thaliana," *J. Mol. Biol.*, vol. 187, no. 2, pp. 169–183, 1986.

[30] S. S. Merchant *et al.*, "The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions," *Science*, vol. 318, no. 5848, pp. 245–250, Oct. 2007.

[31] E. Long and I. Dawid, "Repeated Genes in Eukaryotes," *Annu. Rev. Biochem.*, vol. 49, pp. 727–764, Jul. 1980.

[32] J. G. Gibbons, A. T. Branco, S. Yu, and B. Lemos, "Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans," *Nat. Commun.*, vol. 5, p. 4850, Sep. 2014.

[33] J. G. Gibbons, A. T. Branco, S. A. Godinho, S. Yu, and B. Lemos, "Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes," *Proc. Natl. Acad. Sci.*, vol. 112, no. 8, pp. 2485–2490, Feb. 2015.

[34] A. R. D. Ganley and T. Kobayashi, "Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data," *Genome Res.*, vol. 17, no. 2, pp. 184–191, Jan. 2007.

[35] R. Maleszka and G. D. Clark-Walker, "Yeasts have a four-fold variation in ribosomal DNA copy number," *Yeast*, vol. 9, no. 1, pp. 53–58, 1993.

[36] M. Berriman *et al.*, "The genome of the African trypanosome Trypanosoma brucei," *science*, vol. 309, no. 5733, pp. 416–422, 2005.

[37] J.-M. Guay, A. Huot, S. Gagnon, A. Tremblay, and R. Levesque, "Physical and genetic mapping of cloned ribosomal DNA from Toxoplasma gondii: primary and secondary structure of the 5S gene," *Gene*, vol. 114, no. 2, pp. 165–171, Jan. 1992.

[38] K. Gentile L., W. D. Burke, and T. H. Eickbush, "Multiple Lineages of R1 Retrotransposable Elements Can Coexist in the rDNA Loci of Drosophila," *Mol. Biol. Evol.*, vol. 18, no. 2, pp. 235–245, 2001.

[39] J. L. Jakubczak, W. D. Burke, and T. H. Eickbush, "Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects.," *Proc. Natl. Acad. Sci.*, vol. 88, no. 8, pp. 3295–3299, 1991.

[40] F. Berini, C. Casciello, G. L. Marcone, and F. Marinelli, "Metagenomics: novel enzymes from non-culturable microbes," *FEMS Microbiol. Lett.*, vol. 364, no. 21, Nov. 2017.

[41] P. I. Costea *et al.*, “Towards standards for human fecal sample processing in metagenomic studies,” *Nat. Biotechnol.*, Oct. 2017.

[42] S. G. Tringe *et al.*, “Comparative Metagenomics of Microbial Communities,” *Science*, vol. 308, no. 5721, pp. 551–554, Apr. 2005.

[43] J. C. Venter *et al.*, “Environmental genome shotgun sequencing of the Sargasso Sea,” *science*, vol. 304, no. 5667, pp. 66–74, 2004.

[44] A. Kusnezowa and L. I. Leichert, “In silico approach to designing rational metagenomic libraries for functional studies,” *BMC Bioinformatics*, vol. 18, no. 1, Dec. 2017.

[45] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, “Intracellular ion environment and membrane electric potential,” 2000.

[46] F. Lang, “Mechanisms and Significance of Cell Volume Regulation,” *J. Am. Coll. Nutr.*, vol. 26, no. sup5, p. 613S–623S, Oct. 2007.

[47] A. Wachter, M. Tunc-Ozdemir, B. C. Grove, P. J. Green, D. K. Shintani, and R. R. Breaker, “Riboswitch Control of Gene Expression in Plants by Splicing and Alternative 3’ End Processing of mRNAs,” *PLANT CELL ONLINE*, vol. 19, no. 11, pp. 3437–3450, Nov. 2007.

[48] W. J. Kent, “BLAT—the BLAST-like alignment tool,” *Genome Res.*, vol. 12, no. 4, pp. 656–664, 2002.

[49] E. M. Gertz, “BLAST scoring parameters,” *Manuscript*, 2005.

[50] M. M. Abdelsayed, B. T. Ho, M. M. K. Vu, J. Polanco, R. C. Spitale, and A. Lupták, “Multiplex Aptamer Discovery through Apta-Seq and Its Application to ATP Aptamers Derived from Human-Genomic SELEX,” *ACS Chem. Biol.*, vol. 12, no. 8, pp. 2149–2156, Aug. 2017.

[51] K. A. Wilkinson, S. M. Vasa, K. E. Deigan, S. A. Mortimer, M. C. Giddings, and K. M. Weeks, “Influence of nucleotide identity on ribose 2’-hydroxyl reactivity in RNA,” *RNA*, vol. 15, no. 7, pp. 1314–1321, Jul. 2009.

[52] E. J. Merino, K. A. Wilkinson, J. L. Coughlan, and K. M. Weeks, “RNA Structure Analysis at Single Nucleotide Resolution by Selective 2‘-Hydroxyl Acylation and Primer Extension (SHAPE),” *J. Am. Chem. Soc.*, vol. 127, no. 12, pp. 4223–4231, Mar. 2005.

[53] B. Wang, K. A. Wilkinson, and K. M. Weeks, “Complex Ligand-Induced Conformational Changes in tRNA ^{Asp} Revealed by Single-Nucleotide Resolution SHAPE Chemistry †,” *Biochemistry (Mosc.)*, vol. 47, no. 11, pp. 3454–3461, Mar. 2008.

[54] J. B. Lucks *et al.*, “Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq),” *Proc. Natl. Acad. Sci.*, vol. 108, no. 27, pp. 11063–11068, 2011.

[55] D. Loughrey, K. E. Watters, A. H. Settle, and J. B. Lucks, “SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing,” *Nucleic Acids Res.*, vol. 42, no. 21, pp. e165–e165, Dec. 2014.

[56] Y. Ding, C. K. Kwok, Y. Tang, P. C. Bevilacqua, and S. M. Assmann, “Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq,” *Nat. Protoc.*, vol. 10, no. 7, pp. 1050–1066, Jun. 2015.

[57] R. C. Spitale *et al.*, “Structural imprints in vivo decode RNA regulatory mechanisms,” *Nature*, vol. 519, no. 7544, pp. 486–490, Mar. 2015.

[58] Y. Tang *et al.*, “StructureFold: genome-wide RNA secondary structure mapping and reconstruction *in vivo*,” *Bioinformatics*, vol. 31, no. 16, pp. 2668–2675, Aug. 2015.

[59] S. Aviran *et al.*, “Modeling and automation of sequencing-based characterization of RNA structure,” *Proc. Natl. Acad. Sci.*, vol. 108, no. 27, pp. 11069–11074, 2011.

[60] N. A. Siegfried, S. Busan, G. M. Rice, J. A. E. Nelson, and K. M. Weeks, “RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP),” *Nat. Methods*, vol. 11, no. 9, pp. 959–965, Jul. 2014.

[61] J. Talkish, G. May, Y. Lin, J. L. Woolford, and C. J. McManus, “Mod-seq: high-throughput sequencing for chemical probing of RNA structure,” *RNA*, vol. 20, no. 5, pp. 713–720, May 2014.

[62] M. J. Smit, D. Verzijl, and R. Iyengar, “Identity of adenylyl cyclase isoform determines the rate of cell cycle progression in NIH 3T3 cells,” *Proc. Natl. Acad. Sci.*, vol. 95, no. 25, pp. 15084–15089, 1998.

[63] J. Meitzen, J. I. Luoma, C. M. Stern, and P. G. Mermelstein, “β1-Adrenergic receptors activate two distinct signaling pathways in striatal neurons: β1-Adrenergic receptors activate two pathways,” *J. Neurochem.*, vol. 116, no. 6, pp. 984–995, Mar. 2011.

[64] H. Huang, H. Wang, and M. E. Figueiredo-Pereira, “Regulating the Ubiquitin/Proteasome Pathway Via cAMP-signaling: Neuroprotective Potential,” *Cell Biochem. Biophys.*, vol. 67, no. 1, pp. 55–66, Sep. 2013.

[65] S. Smit, J. Widmann, and R. Knight, “Evolutionary rates vary among rRNA structural elements,” *Nucleic Acids Res.*, vol. 35, no. 10, pp. 3339–3354, May 2007.

[66] R. L. Chisholm, “dictyBase, the model organism database for Dictyostelium discoideum,” *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D423–D427, Jan. 2006.

[67] F. Weis *et al.*, “Mechanism of eIF6 release from the nascent 60S ribosomal subunit,” *Nat. Struct. Mol. Biol.*, vol. 22, no. 11, pp. 914–919, Nov. 2015.

[68] R. Sucgang, “Sequence and structure of the extrachromosomal palindrome encoding the ribosomal RNA genes in Dictyostelium,” *Nucleic Acids Res.*, vol. 31, no. 9, pp. 2361–2368, May 2003.

[69] A. Marchler-Bauer *et al.*, “CDD/SPARCLE: functional classification of proteins via subfamily domain architectures,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D200–D203, Jan. 2017.

[70] S.-J. Li and M. Hochstrasser, “A new protease required for cell-cycle progression in yeast.,” *Nature*, vol. 398, pp. 246–251, Mar. 1999.

[71] A. Tripathi, V. Saini, A. Marchese, B. F. Volkman, W.-J. Tang, and M. Majetschak, “Modulation of the CXC Chemokine Receptor 4 Agonist Activity of Ubiquitin through C-Terminal Protein Modification,” *Biochemistry (Mosc.)*, vol. 52, no. 24, pp. 4184–4192, Jun. 2013.

[72] N. Myeku, H. Wang, and M. E. Figueiredo-Pereira, “cAMP stimulates the ubiquitin/proteasome pathway in rat spinal cord neurons,” *Neurosci. Lett.*, vol. 527, no. 2, pp. 126–131, Oct. 2012.

[73] E.-J. Kim and Y.-S. Juhnn, “Cyclic AMP Signaling Reduces Sirtuin 6 Expression in Non-small Cell Lung Cancer Cells by Promoting Ubiquitin-Proteasomal Degradation via Inhibition of the Raf-MEK-ERK (Raf/Mitogen-activated Extracellular Signal-regulated Kinase/Extracellular Signal-regulated Kinase) Pathway,” *J. Biol. Chem.*, vol. 290, no. 15, pp. 9604–9613, Apr. 2015.

[74] M. M. K. Vu, N. E. Jameson, S. J. Masuda, D. Lin, R. Larralde-Ridaura, and A. Lupták, “Convergent Evolution of Adenosine Aptamers Spanning Bacterial, Human, and Random Sequences Revealed by Structure-Based Bioinformatics and Genomic SELEX,” *Chem. Biol.*, vol. 19, no. 10, pp. 1247–1254, Oct. 2012.

[75] C. Lin, Y. Lin, H. Liu, and L. Kao, “Characterization of Rab3A, Rab3B and Rab3C: different biochemical properties and intracellular localization in bovine chromaffin cells,” *Biochem J*, vol. 324, pp. 85–90, 1997.

[76] L. Zou *et al.*, “The GTPase Rab3b/3c-positive recycling vesicles are involved in cross-presentation in dendritic cells,” *Proc. Natl. Acad. Sci.*, vol. 106, no. 37, pp. 15801–15806, 2009.

[77] A. Gerondopoulos *et al.*, “Rab18 and a Rab18 GEF complex are required for normal ER structure,” *J. Cell Biol.*, vol. 205, no. 5, pp. 707–720, Jun. 2014.

[78] N. Pasteris, K. Nagata, A. Hall, and J. L. Gorski, “Isolation, characterization, and mapping of the mouse FGD3 gene, a new Faciogenital Dysplasia (FGD1; Aarskog Syndrome) gene homologue,” *Gene*, vol. 242, no. 1, pp. 237–247, 2000.

[79] J. EL Andaloussi-Lilja, J. Lundqvist, and A. Forsby, “TRPV1 expression and activity during retinoic acid-induced neuronal differentiation,” *Neurochem. Int.*, vol. 55, no. 8, pp. 768–774, Dec. 2009.

[80] P. Arun, C. N. Madhavarao, J. R. Moffett, and A. M. A. Namboodiri, “Antipsychotic drugs increase *N* -acetylaspartate and *N* -acetylaspartylglutamate in SH-SY5Y human neuroblastoma cells,” *J. Neurochem.*, vol. 106, no. 4, pp. 1669–1680, Aug. 2008.

[81] H. Xie, L. Hu, and G. Li, “SH-SY5Y human neuroblastoma cell line: in vitro cell model of dopaminergic neurons in Parkinson’s disease.,” *Chin. Med. J. (Engl.)*, vol. 123, no. 8, pp. 1086–1092, 2010.

[82] M. Uhlén *et al.*, “A human protein atlas for normal and cancer tissues based on antibody proteomics,” *Mol. Cell. Proteomics*, vol. 4, no. 12, pp. 1920–1932, 2005.

[83] L. Berglund *et al.*, “A genecentric Human Protein Atlas for expression profiles based on antibodies,” *Mol. Cell. Proteomics*, vol. 7, no. 10, pp. 2019–2027, 2008.

[84] L. Fagerberg *et al.*, “Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics,” *Mol. Cell. Proteomics*, vol. 13, no. 2, pp. 397–406, Feb. 2014.

[85] K. Fujita and Y. Sasaki, “Pharmacogenomics in Drug-Metabolizing Enzymes Catalyzing Anticancer Drugs for Personalized Cancer Chemotherapy,” *Curr. Drug Metab.*, vol. 8, no. 6, pp. 554–562, 2007.

[86] D. S. Goodsell, “The molecular perspective: methotrexate,” *The Oncologist*, vol. 4, no. 4, pp. 340–341, 1999.

[87] P. R. Rajagopalan, Z. Zhang, L. McCourt, M. Dwyer, S. J. Benkovic, and G. G. Hammes, “Interaction of dihydrofolate reductase with methotrexate: ensemble and single-molecule kinetics,” *Proc. Natl. Acad. Sci.*, vol. 99, no. 21, pp. 13481–13486, 2002.

[88] E. Afgan *et al.*, “The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update,” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W3–W10, Jul. 2016.

[89] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014.

[90] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet J.*, vol. 17, no. 1, p. pp–10, 2011.

[91] K. K. Alam, J. L. Chang, and D. H. Burke, “FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections,” *Mol. Ther. - Nucleic Acids*, vol. 4, p. e230, 2015.

[92] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nat. Methods*, vol. 9, no. 4, pp. 357–359, Mar. 2012.

[93] J. W. Nicol, G. A. Helt, S. G. Blanchard, A. Raja, and A. E. Loraine, “The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets,” *Bioinformatics*, vol. 25, no. 20, pp. 2730–2731, Oct. 2009.