UNIVERSITY OF CALIFORNIA,
IRVINE

Discovery and characterization of novel functional human RNAs by *in vitro* selection

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry
with a concentration in Chemical Biology

by

Michael Manh Khang Vu

Dissertation Committee:
Professor Andrej Luptak, Chair
Professor Jennifer Prescher
Professor Robert Spitale

2019

DEDICATION

First and foremost, to my wonderful family
My lovely and inspiring fiancé, Genia, who has been my rock.
Can we get married yet?
My precious son, Lex. You bring me endless happiness.
Never lose your wonderful warm heart and can-do attitude.
And my darling baby, Katya, who will love me soon

I am but an empty vessel and all scientific insights are just an intellectual manifestation of the coffee

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my appreciation to my committee chair, Professor Andrej Luptak for his guidance throughout the many years and really introducing me to the realm of RNA biology

I would like to thank my committee members, Professor Robert Spitale and Professor Jennifer Prescher. Rob, working with you in Medicinal Chemistry was quite the pleasure and I was very excited when you decided to come to UCI, your NAI was the first molecule I made! Jenn, you have a wonderful lab and I learned so much during my time there. I also have to mention your Chemical Biology class was my favorite class in my academic career.

I would also like to thank all the other professors in the department of Chemistry I've gotten to interact with over the years. I'm proud of the education I received and grateful for the support I've been given.

I also thank for Elsevier Ltd. permission to include Chapter Two of my dissertation, which was originally published in Chemistry & Biology and American Chemical Society for permission to include Chapter Four of my dissertation, which was originally published ACS Chemical Biology

Michael Vu

University of California, Irvine

EDUCATION

PhD Chemistry, *University of California, Irvine* 2012-2013, 2014-2019

BS Chemistry, BS Biology, *University of California, Irvine* 2007-2011
Honors in Chemistry
Chancellor's Scholar
Campus-wide Honors
Dean's Honor List

RESEARCH EXPERIENCE

Graduate Researcher, RNA Biochemistry, *Luptak Lab* 2012-Current
•Isolated and identified potential sequences for a human ribozyme
•As a senior member, mentored new students in techniques and lab procedure
•Mentored an undergraduate researcher

Staff Researcher, RNA Biochemistry, *Luptak Lab* 2011-2012
•Characterized aptamer structure and function for publication
•Trained new student in laboratory techniques
•Managed lab supplies, ordering, and vendor communications

Student Researcher, RNA Biochemistry, *Luptak Lab* 2009-2011
•Isolated the first human RNA aptamers
•Learned *in vitro* selection

TEACHING EXPERIENCE

Instructor, GRE & SAT, *Sherwood Test Prep* 2014-2016
•Taught students test taking strategies specific for their exam
•Provided review on relevant subjects

Lab Teaching Assistant, General Chemistry, *UC Irvine* Summer 2013, 2015, 2017, 2018
•Led lab sections by walking students through lab procedures and relevant chemical concepts Fall 2014
•Provided explanation of concepts and advice for lab reports at weekly office hours Winter 2012
•Graded and provided detailed feedback on reports.

Lab Teaching Assistant, Organic Chemistry, *UC Irvine* Summer 2016, 2017, 2018
•Led lab sections by walking students through lab procedures and relevant chemical concepts Winter 2013
•Provided explanation of concepts and advice for lab reports at weekly office hours
•Graded and provided detailed feedback on reports.

Head TA, Medicinal Chemistry, *UC Irvine* Winter 2015
•Led discussion sections summarizing and clarifying the lessons taught in lecture
•Led lab sections by walking students through lab procedures and relevant chemical concepts
•Prepared reagents and cells for weekly labs
•Organized and held review sessions for exam preparation
•Designed and graded quizzes
•Collaborated in the writing and grading of the exams

Teaching Assistant, Medicinal Chemistry, *UC Irvine* Winter 2016, 2017
•Led discussion sections summarizing and clarifying the lessons taught in lecture
•Led lab sections by walking students through lab procedures and relevant chemical concepts
•Prepared reagents and cells for weekly labs
•Organized and held review sessions for exam preparation
• Collaborated in the writing and grading of the exams

Lab TA, Medicinal Chemistry, *UC Irvine* Spring 2016, 2017 2019
•Led discussion sections summarizing and clarifying the lessons taught in lecture Winter 2016, 2017
•Led lab sections by walking students through lab procedures and relevant chemical concepts
•Prepared reagents and cells for weekly labs
•Organized and held review sessions for exam preparation
• Collaborated in the writing and grading of the exams

Teaching Assistant, General (Honors) Chemistry, *UC Irvine* Fall 2012, 2018
•Led discussion sections summarizing and clarifying the lessons taught in lecture Spring 2013
•Held weekly office hours and answered question daily via email and social media forums
•Collaborated in the grading finals and midterms

Teaching Assistant, Chemical Biology, *UC Irvine* Winter 2019
•Designed and presented lecture on modern RNA chemical biology
•Led discussion sections summarizing and clarifying the lessons taught in lecture
•Held weekly office hours and answered question daily via email and social media forums
•Graded quizzes and homework
•Collaborated in the grading finals and midterms

OUTREACH

Dancer, Dance Chemistry 2013-2014
•Took part in a series of videos aimed at teaching chemical concepts through dance
•These videos are now utilized as part of the general chemistry lab curriculum
•Videos found on this channel: www.youtube.com/dancechemistry

OTHER AWARDS

Pfizer Undergraduate Organic Research Award 2011
Summer Undergraduate Research Fellowship 2010
Contributions to the Chemistry Department Teaching Program- Continuing TA 2019

PUBLICATION

Vu, M.; et al; Convergent Evolution of Adenosine Aptamers Spanning Bacterial, Human, and Random Sequences Revealed by Structure-Based Bioinformatics and Genomic SELEX *Chem Biol* 2012, 9, 1247-54

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

ABSTRACT OF THE DISSERTATION

Discovery and characterization of novel functional human RNAs by *in vitro* selection

By

Michael Vu

Doctor of Philosophy in Chemistry

University of California, Irvine, 2019

Professor Andrej Luptak, Chair

The more we learn to read the instructions encoding the functions of all life, the more we realize how much we still have to learn. Our understanding of DNA encoded functions thus far has revolutionized medicine and biological study, but we growing to appreciate how much we still don't know about the roles of the vast majority of the RNAs that our DNA encodes. My work centers around revealing these undiscovered RNA functions, particularly in humans, and the characterization of these RNAs achieved by harnessing the characteristics of RNA as an evolvable, multi-functional, self-encoded polymer. I used *in vitro* evolution to isolate and identify RNAs that perform a specified function from the vast landscape of RNA encoded by our genome (genomic SELEX) and again to help define components and properties that grant each of the RNAs that function. I also developed multiplexed methods to chemically probe these RNAs to address the challenge of characterizing the diverse pools of candidates produced by genomic SELEX with current sequencing technology. I discovered the first small molecule aptamers in humans, with now eight known examples of ATP binding RNAs encoded in the human genome. Over the course of their study, I've developed CIM-seq and Mini-SELEX and aided in the development of Apta-seq as methods that can be used broadly in the characterization of in *vitro*

selection experiments. Finally, I have found evidence for the existence of a very intriguing ribozyme encoded in our DNA. These contributions inch us closer to understanding the many functions of RNA *in vivo*, provides topics for further study, and means with which to study them.

Chapter 1: RNA aptamer discovery by Genomic SELEX

Introduction

It is hypothesized that RNA or a similar nucleic acid polymer was the primary macromolecule of early protocellular life (Joyce 1989). With the myriad functions RNAs have demonstrated, it is easy to see why. The complex three-dimensional folds available to RNA allow it to achieve diverse functions in catalysis and gene regulation. This endows RNA the ability to fulfill the essential roles needed for primitive life. Since their discovery, ribozymes and riboswitches, catalytic RNA and gene regulating RNA elements, respectively, have sparked broad efforts towards discovering as much about these functional RNAs as possible. New discoveries in the field have become increasingly common and have found applications in biotechnology and therapeutics (Nimjee, Rusconi and Sullenger 2005) but we are still growing to appreciate the importance of RNA in present day biology.

Unlocking the genetic code has been one of the most fundamentally important advances in our understanding of biology, empowering us not only with the ability to predict characteristics of living things but even exert control over biological systems through their DNA. This power stems from the simple knowledge of what DNA segments encode for which proteins. That is just the tip of the genomic iceberg. Only ~1% of our DNA encodes for the proteins products we derive our genetic insights from, yet most of our DNA is still transcribed into RNA. This would indicate that RNA itself has functional importance beyond protein synthesis. The bulk of these RNAs have yet uncharacterized functions leaving an immense and exciting task of elucidating the functions of the remaining, non-coding, RNAs (ncRNAs). One RNA function we will focus on in particular is ligand binding performed by RNAs called aptamers.

Current approaches for discovering RNA aptamers and other functional RNA include bioinformatics and transcriptomics strategies. These approaches have had their successes; however, they also have key limitations. Transcriptomic discovery of functional RNA requires that the RNAs are transcribed in sufficient quantities in a given cell line/condition, missing potentially important RNA with low, transient, or cell specific expression. Bioinformatic approaches to functional RNA discovery search for additional examples of already characterized functional RNAs based on their ability to adopt similar secondary structure. The precise nucleotide sequence of an RNA is less important than the overall structure achieved; indeed, many classes of functional RNA display conserved folds and achieve similar functions with minimal sequence conservation. If the secondary structure of a functional RNA is known, a descriptor can be written and aligned to the genome of interest (Riccitelli and Lupták 2010). This reveals RNAs that can potentially adopt a similar structure, and presumably, function, to be confirmed *in vitro*. Another approach is to use comparative sequence analysis to identify highly conserved structures in RNAs. For example, riboswitches, gene regulating RNAs that employ aptamers, are found by searching for conserved secondary structures and sequences specifically in the 5′ UTR of bacterial mRNAs. The ligands they respond to are usually inferred by the genetic location and are tested *in vitro* and *in vivo*, but there are also several orphan riboswitches whose ligands have yet to be identified (Matylla-Kulinska, et al. 2012). A complementary method to discovery which overcomes many of these drawbacks applies *in vitro* selection to genomically encoded RNA

In vitro selection or SELEX (Systematic Evolution of Ligands by EXponential enrichment) was developed to evolve oligonucleotides capable of binding to a desired target starting from random sequence pools (Ellington and Szostak 1990) (Tuerk and Gold 1990)

Oligonucleotides with the chosen function are enriched from a diverse pool by an iterative process of separating the pool into functional and non-functional oligonucleotides and regenerating the pool by amplification of only those functional species. Genomic SELEX applies this method to genome derived libraries, in place of synthetic libraries, and biological targets to enrich functional RNAs or DNAs that are likely to have biological significance (Singer, et al. 1997). It does not require any prior knowledge of the desired oligonucleotides sequences or structures, overcoming one of the limitations of bioinformatics based functional RNA discovery, and enabling the search for novel functions and motifs. Using genomic SELEX has the added benefit of screening RNAs irrespective of expression, allowing for the discovery of functional RNA with low, transient, or condition-specific expression missed by transcriptomic approaches. Genomic SELEX is not able to screen RNA that require post-transcriptional modifications and the resulting sequences are not guaranteed to be expressed. However, the amount of genomic DNA with confirmed RNA expression increases constantly with deeper sequencing experiments being performed in different tissues and conditions.

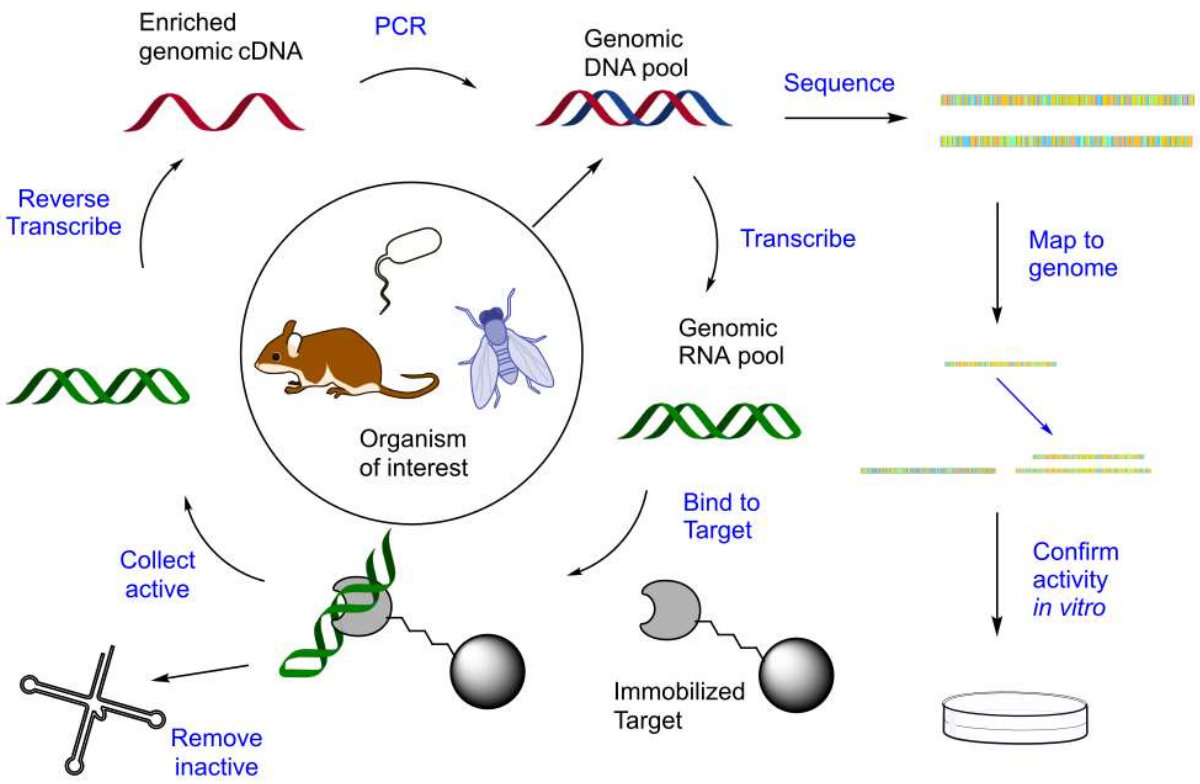


Figure 1-1 Genomic SELEX of RNA aptamers Genomic SELEX starts with isolation of DNA from the organism(s) of interest and addition of constant regions for amplifications. The genomic pool is transcribed into RNA and bound to the target of interest. Nonbinders are removed and those that bind are reverse transcribed into cDNA. PCR amplifies this enriched fraction to allow for the cycle to be repeated as necessary. Functional pools are sequenced and aligned the genome(s) of interest. Ligand binding of promising candidates are then confirmed *in vitro*.

The genomic library is created by extracting DNA from the organism(s) of interest, followed by random priming to introduce a handle for amplification while maintaining maximum diversity to sample all potential structures (Singer, et al. 1997). The library can then be size selected, if so desired, with longer libraries being more likely to yield complex structures (Sabeti, Unrau and Bartel 1997), but requiring more minimization downstream. The function of RNAs isolated, will be determined by the design of the separation strategy. When selecting for aptamers, ligand binding oligonucleotides, RNAs encoded by the genomic library are required to bind to the immobilized ligand of interest or be washed away. Potential aptamers are then collected and amplified via reverse transcription and PCR to comprise a pool lower in diversity

but enriched for ligand binding. This is repeated until the pool is deemed sufficiently active, with later rounds often increasing in stringency with more washes, lower incubation time, and counter selection with off targets. The resulting pool is sequenced and categorized by family or motif based on comparative sequence analysis. The origin of each sequence is found by comparing them to reference genomes from available databases and the functions of the selected sequences is then verified *in vitro*.

Genomic SELEX of RNAs has identified aptamers to proteins, RNAs and even small molecules. Protein-RNA interactions exist in several biological context and these interaction networks can be highly complex with multiple RNAs capable of binding a single protein. Using a protein target in genomic selections allows for identification of numerous RNA binding partners at once. RNAs that bind small molecules, such as the aptamer domain of riboswitches, could also be found using genomic SELEX. This chapter aims to review the genomic selection performed to date and highlight lessons for genomic SELEX experiments to come.

Lessons from early genomic selections

Shtatland and colleagues performed the first genomic SELEX to investigate whether MS2 bacteriophage could regulate gene expression as some other viruses are known to do. (Shtatland, et al. 2000). One of the phage's proteins, MS2 coat protein (MS2 CP), was known to bind MS2 mRNA via an RNUA tetraloop consensus site for at least two distinct functions: to repress translation of phage replicase and to form a component of the viral coat (Uhlenbeck, et al. 1983). Hence, it was a good candidate for genomic SELEX to elucidate interactions with its host's RNA, potentially as a means of regulating expression. Initial selection rounds produced

mostly sequences that fit the viral binding site consensus sequence. However, many of these initial isolates were granted this consensus sequence via participation of the flanking constant regions and mutations picked up along the selection process. As these selection artifacts would have no biological relevance, they were removed from the selection by switching constant regions on both ends and performing additional rounds of selection. This produced 183 isolates identified by cloning and sequencing. These isolates corresponded to 7 unique genetic loci containing an MS2 CP consensus binding sequence. The genomic-SELEX derived consensus sequence differed slightly from the viral consensus, utilizing a higher affinity RNCA tetraloop, discovered previously from random sequence SELEX, as opposed to the RNUA tetraloop present in MS2 mRNA. The RNCA tetraloop was used to inform an RNAMOT search across the *E. coli* genome to reveal 18 additional genomic locations not detected in the selection while missing 4 of the selected sequences. This was the earliest genomic selection, so some of these sequences may not have been detected in SELEX due to the low depth of the sequencing methods used at the time. Most of the sequences were represented by only one isolate, and the most abundant sequence, which corresponded to a region 210 nt from the start of the *rffG* gene, comprised 98 of the 183 isolates. Two of these candidates were confirmed to bind MS2 CP; the sequence mapping to the *rffG* gene and one to the *ebgR* gene. The *ebgR* sequence, which was also 210 nt from its the start codon, bound despite only having one isolate. One of the 24 isolates that did not contain the consensus sequence was shown not to bind MS2 CP.

This first genomic selection demonstrates a few themes which we will see repeated throughout selections to follow: first, genomic SELEX can find motifs that can inform bioinformatic discovery of additional genomic aptamers. Next, the binding of *ebgR* shows that even the least abundant selected RNAs can be functional. Finally, only a few sequences are

usually characterized post selection. The choice of which candidates to analyze further often depends on known motifs, genomic location, abundance, or even random selection. In the aforementioned MS2 genomic SELEX experiment, the remaining MS2 CP aptamers reported from both SELEX and RNAMOT, most of which were antisense sequences, were not tested for target binding. The sequences and genetic loci for selected sequences that were missing a consensus tetraloop were not reported.

Genomic SELEX was next used to find RNA aptamers that bind the *Drosophila melanogaster* Transformer 2 (Tra2) protein (Brunel, et al. 2001). This protein is necessary for female sexual differentiation and male fertility, and was known to directly bind both the doublesex (*dsx*) and fruitless (*fru*) mRNAs (Lynch and Maniatis 1995). It had also been implicated in the regulation of three other genes in the testes, suggesting it may have other RNA binding partners (Madigan, et al. 1996). For these reasons, it was chosen as a target for selection with known examples to evaluate the method. Most of the RNA sequences isolated for Tra2 binding were adenosine-rich, mirroring the results of a synthetic SELEX experiment targeting human Tra2 (Tacke, et al. 1998), and contain a CAAAA motif, similar to the CAACA motif of *dsx* RNA. While motifs similar to the *dsx* motif were found, the *dsx* RNA sequence itself was not amongst the sequenced clones. A DNA blot indicated that the *dsx* sequence was indeed present in the pool despite being missed by cloning and it is possible this sequence would have been detected in their selection with today's sequencing technology. Brunel et al. postulated that the low abundance of the *dsx* sequence known to bind Tra2 may result from absence of other biological participants in the selection condition, a potential limitation to genomic selections in identifying cooperative binding. For example, Tras2 *in vivo* works in concert with the Tra protein (Lynch and Maniatis 1996), and both proteins may be required to bind the *dsx* RNA. Another

possibility is that the *dsx* sequence is somehow less amplification competent, for example, by having a strong reverse-transcriptase pause site or another poorly amplifiable structural feature. The selection also enriched a family of sequences from satellite DNA compose mostly of either (AAUAACAUAG)n or (UG)n repeats, representing discovery of novel motifs for Tra2 binding. This selection shows that functional sequences are not guaranteed to be abundant in SELEX while discovering new and modified motifs, though these were not verified *in vitro*.

The low abundance of known aptamers continued to be a problem in the validation of early genomic selections. This was observed in the next genomic SELEX experiment aimed at finding splicing sites for the *Drosophila* pre-mRNA splicing factor B52 (SRp55) (Kim, et al. 2003). Identification of sites that require a specific splicing factor had been attempted but proved elusive. Previous efforts could not detect splicing defects in any specific genes of B52-null flies despite B52 deletions being lethal (Ring and Lis 1994). While B52 is essential for fly development, it shares functional redundancies with other splicing factors, complicating attempts to identify B52-specific RNA targets. To resolve this challenge, genomic SELEX for *Drosophila* RNAs that bind to B52 splicing factor was performed. The resulting pool was highly diverse, with most sequences represented by only one isolate and none were reported to appear more than four times. Out of 96 selected and sequenced clones, 87 bound B52 representing 55 unique genomic sequences spanning introns, exons, and intergenic regions. While it is unclear what role, if any, the many intergenic B52-binding sequences could play, those mapping to multi-exon genes were screened for B52-dependent splicing defects. Their assay used B52 null flies, which arrest in the 2nd larval stage, further limiting the study to genes expressed at that stage. Of the 20 selected sequences that mapped to intron containing genes, 15 were detectable at that larval stage and four of those displayed splicing defects. Each of the four sequence with B52 dependent

splicing defects had only one isolate in selection, strongly demonstrating that abundance in genomic SELEX is not determinant of biological importance. In fact, the *ftz* gene sequence, which is known to bind B52, was not found in their selection at all. The authors proposed that another protein partner could be required for B52 to bind *ftz* RNA, but an alternative explanation is that the *ftz* sequence was simply not detected with the low depth cloning method. Perhaps current HTS technology may have yet revealed the *ftz* sequences in their pool.

Genomic SELEX saw its first application towards RNA-RNA interactions in an experiment by Watrin et al. hoping to elucidate mechanisms of HIV retroviral transcription (Watrin, et al. 2009). The transactivation responsive (TAR) element of HIV-1 is a portion of the 5′ UTR of mRNA essential for viral replication. It binds to viral protein Tat and host proteins to allow for transcription of the retroviral genome (Dayton, et al. 1986). It forms an imperfect 59-nt hairpin and is able to form kissing-loop interactions within the retroviral genome to a sequence dubbed TAR*, the function of which is unknown (Jaeger and Tinoco 1993). It can also form kissing-loop interactions with *in vitro* selected Anti-Tar aptamer, R06, which forms a stem with an eight loop, six of which complement the apical TAR loop (Kolb, et al. 2006). Hence, selection from a pool of human genomic RNA was performed to identify sites were viral TAR could interact with human RNA. These authors identified seven genomic sequences from 41 candidates, but only two of the lower abundance sequences used the eight nucleotide loop of the artificial R06 aptamer. Most aptamers instead opt for the six nucleotide complementary loop akin to that of TAR* and all aptamers had at least five complementary nucleotides. Comparing the binding data to their reported abundance demonstrates that abundance in selection is also not an indicator of binding strength as there was no correlation. Unfortunately, the authors were unable to find their highest affinity binding sequence in transcript databases or via RT-PCR

attempts, and thus concluded that it did not correspond to a natural interaction. On the surface, this result highlights one of the drawbacks of genomic SELEX as a discovery tool. However, if one inputs their reported sequence into database searches today, one would find expression of many of the reported sequences, including their top binder, in human mRNA and ESTs. Instead this demonstrates of SELEX to discover RNAs that were, at the time, literally undetectable.

Genomic SELEX empowered by high-throughput sequencing

With many of the strongest and best known aptamers coming from low abundance isolates in genomic SELEX, it appears that the deeper the sequencing, the better the chances of finding biologically important aptamers. And so, we will see that as sequencing gets more powerful, so does genomic SELEX. This is clear in the next genomics selection for RNAs that bind *E. coli* host factor for replication (Hfq). Hfq is involved in regulating many genes by facilitating RNA-RNA interactions and genomic SELEX was able to shed light on the mechanism by showing that Hfq primary targets antisense transcripts (Lorenz, et al. 2010). After performing a genomic SELEX for *E. coli* RNA that could bind to Hfq, the selected clones were first verified to bind in a cellular environment before sequencing. Sequenced clones mapped primarily antisense to translation initiation sites and between ORFs but none of the known Hfq binding RNAs were among them. Since no Hfq binding RNAs were detected, Lorenz et. al turned to deep sequencing using the 454 method. The obtained 8865 sequences were clustered and mapped to the *E. coli* genome in 1552 genomic locations. They found several known Hfq binding RNAs but even then, only the one was among the 24 sequences reported with at least 50 reads each. When comparing their finding to RNAs found to be differentially expressed in Hfq-depleted cells, 52 of their sequences were upregulated and 42 were downregulated. From these

results Lorenz et. al concluded that their selection was successful but not exhaustive, as many of the best known and abundant RNAs associated with Hfq were still not found. They postulate that these known Hfq aptamers may have been weaker binders or difficult-to-reverse-transcribe sequences, hence, not picked up in selection. To characterize the pool further, the authors used MEME, a motif prediction software, to identify the binding motif, AAYAAYAA, which was next verified by DMS footprinting of three of the motifs containing RNAs. The motif was much more prevalent on antisense sequences; 365 times within protein coding genes and 712 times on strands opposite to protein coding genes. This data also suggests that Hfq binding in general is more common among antisense transcripts, and the expression of many of these antisense transcripts were verified in this study. In this genomic selection, deeper sequencing allowed for several known aptamers to be rediscovered along with over 1000 novel ones. Efforts were focused on antisense aptamers with an identifiable motif, however there are hundreds of sequences that went uncharacterized and even unreported among both motif containing sequences and those without one. The amount of completely uncharacterized sequences in Genomic SELEX continues to grow as sequencing gets more powerful.

Genomic SELEX experiments for RNA polymerase aptamers generated 15,000 bacterial and 1,300 yeast aptamer candidates and revealed a wide-spread mechanism to reduce gene expression by binding RNA polymerase, especially in antisense transcripts in *E. coli* (Sedlyarova, et al. 2017) and telomeric transcripts in *S. cerevisiae* (Klopf, et al. 2018). These RNA polymerase aptamers (RAP) bind to RNA polymerase and can either increase or decrease the expression of their downstream transcripts. Before this work, only two examples of RAPs allosterically regulating transcriptions were known. Genomic SELEX for *E. coli* genomic RNA that bind RNA polymerase revealed a startling 15,000 distinct RAPs, including known RAP, 6S

noncoding RNA. The vast majority, 95%, of these sequences were found within genes, mostly opposite the coding region, and covered 60% of all *E. coli* genes. These RAPs were highly diverse, and MEME-enabled motif searches resulted only in a single (CAN)n motif present in 12% of the sequences. From 15,000 RAPS, 85 were tested for inhibitory effects on transcriptions and 24 of promoted sequence and polymerase specific transcription termination. Transcription termination by two of these inhibitory RAPs were examined further and shown to be Rho dependent. Based on the inhibitory effect of the specific RAPs tested and the prevalence of RAPs on so many antisense transcripts, it was proposed that these RAPs are partially responsible for the low expression levels of antisense RNA in *E. coli*. This work was quite extensive, and yet generates several new scientific questions. The actual mode of binding is unclear since motif predictions describe only 12% of the sequences and the (CAN)n repeat does not appear to be among those reported to have either inhibitory effects or no effects. Common themes resurface as there seems to be correlation between abundance of the aptamer in the pool and the strength of the inhibitory effect. In fact, the two strongest inhibitors of those tested included the least abundant aptamer reported as well as one of the most abundant. This leaves thousands of aptamers to be explored as they cannot be ruled out based on abundance alone. With RAPs playing such a prominent role in *E. coli*, genomic SELEX was next used to investigate their role in yeast (Klopf, et al. 2018). Nearly 1,300 RAPs were found in the genomic SELEX of yeast RNA against RNA polymerase. While *E. coli* RAPs were primarily found on the antisense strand of DNA, the yeast RAPs were equally distributed among sense and antisense strands. Instead, yeast RAPs saw significant enrichment in telomeric and subtelomeric regions. The HOMER motif identification software was used to find three low complexity sequence repeats motifs, rich in either CCA, CAA, or AU, which represented about 18% of the sequences. Representatives of

both the CCA and the CAA motifs were shown to decrease transcription levels and RT-PCR confirmed lower expression downstream of RAPs vs upstream displaying the inhibitory effect of these RAPs. These selections in particular highlight the significant wealth of data generated from combining genomic SELEX with high throughput screening. Despite the extensive nature of these works, the activity of most of the sequences remain uncharacterized and potential motifs undiscovered.

Genomic SELEX for small molecule aptamers

The role of non-coding RNA is not limited to interactions with proteins or other RNA. RNA is also known to bind and recognize small molecules. Riboswitches are the predominant examples, which regulate gene expression in response to binding small molecule targets via their aptamer domains. SELEX for small molecule RNA aptamers is commonly used for medicinal (Nimjee, Rusconi and Sullenger 2005) or biotechnological purposes but has been underutilized for the discovery of functional RNA. The first genomic SELEX aimed at discovering functional RNAs that bind to small molecules enriched for ATP binding from human genomic RNA (Vu, et al. 2012) and will be discussed in Chapter 2.

Human ATP aptamers were quickly followed by another ubiquitous small molecule with diverse biological roles, GTP. Guanosine is utilized by the group I introns during self-splicing (Kruger, et al. 1982) and several riboswitches bind to guanine (Breaker 2011) and this genomic SELEX was able to reveal a novel mode of guanosine binding predominant in organisms outside the bacterial kingdom (Curtis and Liu 2013). As GTP and many other small molecules are identical between all organisms, SELEX can be used to probe for small molecule aptamers in several genomes at once. Genomic DNA from several phylogenetically diverse eubacteria (*E.*

coli, B. subtilis, and *B. fragilis*), archaebacteria (*H. marismortui, A. pernix,* and *M. jannaschii*), and eukaryotes *(H. sapiens* and *G. gallus*) were all used to simultaneously select for GTP binding. Curtis *et al*. found 73 aptamers; surprisingly, virtually all from either human or chicken genomes and mapped mostly to intergenic regions. Almost all analyzed sequences were G rich and predicted to form G quadraplexes, dubbed the G-motif. They selected one candidate for further study based on its length, expression, and conservation and it seems to bind to guanine containing molecules via incorporation of the ligand into the G quadraplex structure with no preference for the sugar or phosphate. This informed bioinformatic searches for other G quadraplex forming structures in the human genome and 20 of these sequences, selected at random, all bound GTP more efficiently than a random sequence pool. Many G quadraplexes with known regulatory functions were also confirmed to bind GTP more efficiently than a random pool. These authors concluded that guanosine binding is a common feature in human RNA, estimating that 75,000 sequences in the human genome have GTP binding constants relevant to physiological concentrations. Further inspection into their pool revealed a second motif, called the CA motif, which contains multiple repeats of CA rich sequences, for example ACAACA (Curtis and Liu 2014). These sequences seem to bind GTP by adopting a triplex structure, with no requirement for canonical base pairing, as AC repeats are sufficient for binding with no other nucleotide. After determining the nucleotide and repeat requirements for this motif, they used a bioinformatic search which revealed additional examples in several eukaryotes and none in either bacterial or archaeal genomes. Both characterized GTP binding motifs differ from those isolated from previous *in vitro* selections from random sequence pools and neither motif seems to rely on canonical base pairing structures making them elusive to bioinformatic discovery.

Conclusions

These genomic SELEX experiments have revealed the ligand binding functions of many classes of RNA across several organisms and these RNAs have been shown to modulate splicing and control expression in conjunction with their binding partners. The selections also demonstrated to us the strengths and weaknesses of genomic SELEX as a discovery tool. It does not produce an exhaustive list of genomic aptamers. Many known biological aptamers bind their targets in conjunctions with other macromolecules and often do not survive the selection despite their *bona fide* biological activity and it can be difficult to incorporate all potential chaperones and cofactors in the selection conditions. What genomic SELEX can do and has done is complement bioinformatic and transcriptomic approaches, often identifying candidates unavailable to the other discovery strategies. It enables the discovery of novel motifs and RNAs that escape detection of transcriptomic methods.

One lesson from these past genomic SELEX experiments is that abundance in selection is not a reliable indicator of importance *in vivo* or even strength of binding *in vitro*. Even RNAs with only one isolate have been shown several times to be examples of novel verified functions or rediscovered aptamers. This observation highlights the importance of high throughput sequencing in identifying these low abundant sequences. This result also leaves little to guide in the selection of candidate to test *in vitro* and *in vivo*. If even the least abundant aptamers can be important, how does one approach thousands of candidate sequences generated with current sequencing methods. Most of these are left uncharacterized if its functional motif or biological relevance is less obvious. This challenge is being addressed as high throughput methods to

characterize pools are developed and improved upon which can narrow down the pool of candidates or gather biochemical information on the entire pool at once.

The genomic SELEX experiments also show us that the motifs from genomic and random sequence SELEX can often differ, making bioinformatic searches based on motifs arrived at by traditional SELEX less informative. The aptamers used in nature do not necessarily need to be optimized enough to compete away the 10^{16} randomly synthesized RNAs used in traditional SELEX, and therefore may utilize motifs with more attenuated binding. The discovery of the biologically utilized motifs can inform bioinformatic searches to discovery additional aptamers as was demonstrated in the MS2 and GTP binding selections.

One of the earliest identified weaknesses of genomic SELEX is that identified RNA aptamers may not be expressed, and therefore not be of any real biological importance. In fact, there have been similar SELEX based experiments using a pool derived from transcripts specifically to ensure expression. These transcriptomic SELEX experiments have been used to identify transcripts that bind RNA binding protein, HEXM1 (Fujimoto, Nakamura and Ohuchi 2012), and a human microRNA precursor that binds folic acid (Terasaka, et al. 2016). While some of the earliest genomic selections encountered aptamers that were not known to be expressed at the time, submitting their reported sequences to current databases shows that the more recent RNA-seq data have since found them. This instead highlights the importance of genomic SELEX's ability to identify functional RNAs independent of their biological abundance.

Genomic SELEX has led us toward unexpected roles of RNA but there are still many RNA functions that could be expanded by SELEX. There is an especially small number of selections that utilize small molecules, which have the advantage of being amenable selections

using metagenomes. Genomic selections also do not have to be limited to ligand binding and have successfully identified catalytic RNAs such as self-cleaving ribozymes in humans (Salehi-Ashtiani, et al. 2006) and self-alkylating ribozymes in *A. pernix* (McDonald, et al. 2014) but is also under-utilized. Much more can also be done in the analysis of these selections. Empowered by high throughput sequencing, genomic SELEX generates data which requires high throughput methods of characterization in order to fully utilize. Many methods developed for *in vitro* selection in general have yet to be applied to genomic selection.

This dissertation will describe my use of genomic SELEX to discover novel classes of functional RNAs in humans and methods I developed for multiplexed characterization of *in vitro* selected candidates. Chapter 2 covers the genomic selection for and discovery of the first small molecule sensing RNAs in humans in the form of ATP binding aptamers (Vu, et al. 2012). The resulting pool was analyzed further in Chapter 3, which describes additional aptamers, outside the canonical ATP binding motif, found by high throughput sequencing of the ATP binding pool. Chapter 3 also includes characterization of individual aptamers candidates, and multiplexed conservation and covariation analysis by mutagenesis followed by further *in vitro* selection. Chapter 4 describes Apta-seq, a method for multiplexed aptamers discovery, structural characterization, and K_D analysis which was developed alongside colleagues, Michael Abdelsayed and Bao Ho (Abdelsayed, et al. 2017). In Chapter 5, I describe CIM-seq, a method I developed for multiplexed mapping of cleavage interference/enhancement of RNA function and its application on new ATP binding candidates. Chapter 6 introduces Mini-SELEX, a multiplexed method I designed to isolate minimized functional motifs from *in vitro* selected pools by truncation of the pool and further selection. Finally, Chapter 7 describes an *in vitro*

selection designed to discover RNAs that form covalent bonds to ATP encoded in the human genome resulting in what appears to be the first splicing ribozyme(s) in humans.

Chapter 2: Discovery of human ATP aptamers by Genomic SELEX: Convergent evolution of adenosine aptamers spanning bacterial, human, and random sequences revealed by structure-based bioinformatics and genomic SELEX

Introduction

Functional nucleic acids fold into specific structures that form the physicochemical foundations for their activity, which includes catalysis and ligand binding. Among the best characterized functional RNAs are aptamers, oligoribonucleotides *in vitro* selected to bind target molecules with high affinity and/or specificity (Ellington and Szostak 1990, Stoltenburg, Nikolaus and Strehlitz 2012, Tuerk and Gold 1990). In general, the aptamer structures consist of the ligand-binding pockets formed by folding of single-stranded regions of specific sequence and flanking helical regions that are typically sequence independent (Riccitelli and Lupták 2010). When the target is a small molecule (e.g., a metabolite), the aptamer often forms a buried binding site with extensive hydrogen bonding and van der Waals interactions that give rise to the ligand-binding characteristics. The flanking helical regions contribute to the strength of the interaction by stabilizing the ligand-binding structure but can also be used as transduction domains for signaling the presence of the ligand in the binding site by switching between two alternative pairing interactions (Breaker 2011).

One of the most extensively characterized aptamers is a simple motif that binds adenosine. The aptamer was first identified in an *in vitro* selection for ATP-binding RNAs (Sassanfar and Szostak 1993), and later in several independent selection experiments targeting adenosine-containing cofactors, including nicotinamide adenine dinucleotide (Burgstaller and Famulok 1994), S-adenosyl methionine (Burke and Gold 1997), and S-adenosylhomocysteine (Gebhardt, et al. 2000). The aptamer recognizes adenosine and its 5′ phosphorylated analogs by

an 11 nucleotide (nt) loop and a bulged guanosine located on the opposite strand. Solution structures of AMP-bound complexes revealed that the conserved recognition loop folds into a compact binding site that contacts the Watson-Crick face of the nucleobase and the 2′ position of the ribose (Dieckmann, et al. 1996, Dieckmann, et al. 1997, Jiang, et al. 1996), explaining the binding specificity for adenine ribonucleoside and tolerance to substitutions at the 8′ and 5′ positions (Sassanfar and Szostak 1993). The reproducible isolation of this motif from random sequences suggests that it is the simplest adenosine-binding structure, providing a striking example of convergent molecular evolution. Despite this structural convergence *in vitro*, the aptamer has not been identified in genomic sequences; therefore, it has not been known whether the convergence extends to biological systems and whether biological ligand-binding RNAs, particularly riboswitches, utilize this motif for regulatory or other functions.

Riboswitches are functional RNAs composed of aptamer domains and expression platforms, which bind molecular targets and regulate gene expression, respectively (Breaker 2011). Most riboswitches have been found in bacteria, with a single example described in algae, fungi, and plants but not in other eukaryotes (Cheah, et al. 2007, Kubodera, et al. 2003, Wachter, et al. 2007, Croft, et al. 2007). The aptamer domains of the riboswitches constitute a diverse set of RNA structures, ranging from three-way junctions to intricate pseudoknots (Montange and Batey 2008); however, these motifs are different from *in vitro*-selected aptamers targeting the same ligands. This observation suggests that *in vitro*-selected and biological aptamer domains follow separate evolutionary pathways, perhaps because aptamers evolved *in vitro* are optimized for target binding and efficiency of amplification during the selection process, whereas the riboswitch motifs have evolved to couple target binding to gene expression regulation. On the other hand, the methods used to identify the aptamer domains in synthetic and biological

sequences are distinct, relying on *in vitro* selection and phylogenetic conservation of the aptamer structure, respectively, and may thus bias the identified motifs. Moreover, *in vitro*-selected aptamers are typically smaller than riboswitches, in part because the length of the starting selection pool tends to be shorter than the typical riboswitch. In principle, however, the motifs recognizing the same ligand can be isolated from random sequences and identified in genomic sequences, especially if the methods used are amenable to the discovery of simple RNA motifs that may be abundant in both sequence sets.

To test whether *in vitro*-identified aptamer motifs exist in genomic sequences, we used structure-based searches (Riccitelli and Lupták 2010) and identified adenosine aptamers in a bacterium and several vertebrates, including humans. Furthermore, to identify all adenosine-binding RNAs encoded in the human genome, independent of structure, we performed an *in vitro* selection for ATP-binding transcripts of a human genomic pool (Salehi-Ashtiani et al., 2006) and discovered two more aptamers. It is surprising that both aptamers fold into the same structure as was identified *in vitro*. Two of the three human aptamers map to introns of expressed genes and may sense ATP concentration through a kinetic mechanism.

Structure-Based Search for Genomic Adenosine Aptamers

We performed the motif searches by designing a descriptor for the adenosine aptamer structure, based on the known sequence variants and solution structures (Burke and Gold 1997, Dieckmann, et al. 1997, Gebhardt, et al. 2000, Dieckmann, et al. 1996, Jiang, et al. 1996, Sassanfar and Szostak 1993, Burgstaller and Famulok 1994). The aptamers recognize any adenosine-containing molecule with exposed ribose and adenine moieties using a conserved 11 nt loop opposite to a single guanosine and flanked by two helical segments (Figures 2-1A and

21B). Because of its asymmetry, the adenosine-binding loop can be incorporated into the flanking sequences through either of two topologies, with the loop in either the 5′ or the 3′ strand of the aptamer, requiring two separate descriptors (Riccitelli and Lupták, 2010). Both topologies have been identified among *in vitro*-selected sequences, with a slight preference for the loop in the 5′ strand (Burke and Gold, 1997; Sassanfar and Szostak, 1993); thus, we designed descriptors for both orientations. Whereas the aptamer fold is simple, its information content (≥35 bits; Carothers et al., 2006) suggests that it should appear by chance no more than about once per 10 mammalian genomes, and it should be readily isolated from random libraries of diversities >5e10, as confirmed by several independent experiments (Burgstaller and Famulok, 1994; Burke and Gold, 1997; Gebhardt et al., 2000; Sassanfar and Szostak, 1993). We threaded the publicly available reference genomes through the structure descriptors and identified sequences capable of assuming the same fold. Our search revealed a number of potential aptamers but very few that perfectly matched the consensus structure identified *in vitro*. Most candidate aptamers had weaker paired regions or mutations in the binding loop, suggesting that they have lower affinity for adenosine than the optimal *in vitro*-selected aptamers (KD ∼1 μM).

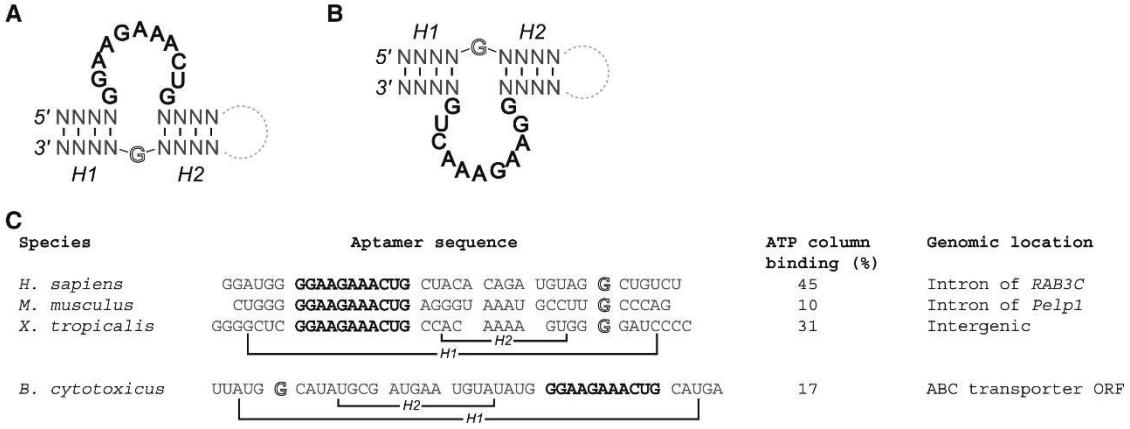


Figure 2-1 Genomic Adenosine Aptamers Uncovered Using Structure-Based Bioinformatics (A and B) Secondary structure descriptors for the *in vitro*-selected adenosine aptamer, with the recognition loop (shown in bold) in (A) the 5′ strand or (B) the 3′ strand. Both topologies have been identified *in vitro* (Burgstaller and Famulok, 1994; Sassanfar and Szostak, 1993). The bulged guanosine required for ligand binding is outlined on a strand opposite of the recognition loop. (C) Sequences of aptamers that showed robust ATP binding *in vitro*. Fraction of each RNA eluted from ATP-agarose beads in the presence of 5 mM ATP-Mg and genomic locations of the aptamers are listed on the right. Outer and inner helices of the structure are marked as H1 and H2, respectively. The *B. cytotoxicus* sequence has the recognition loop in the 3′ segment of the aptamer.

In vitro Activity of Aptamer Candidates

To test their *in vitro* activity, we transcribed the putative aptamers from synthetic templates corresponding to the genomic sequences, purified, and incubated with ATP-agarose beads at conditions similar to the ones used for *in vitro* selections (Sassanfar and Szostak, 1993). After extensive washing, the aptamers were eluted with 5 mM AMP or ATP-Mg and the fraction of specifically bound RNA was determined using scintillation counting. Several RNAs showed significant ATP binding and competitive elution, including sequences mapping to genomes of a bacterium and several vertebrates: Bacillus cytotoxicus, Western clawed frog, mouse, marmoset, chimp, and human (Figure 2-1C; the primate aptamers are almost identical, and their alignment is shown in Figure 2-2B).

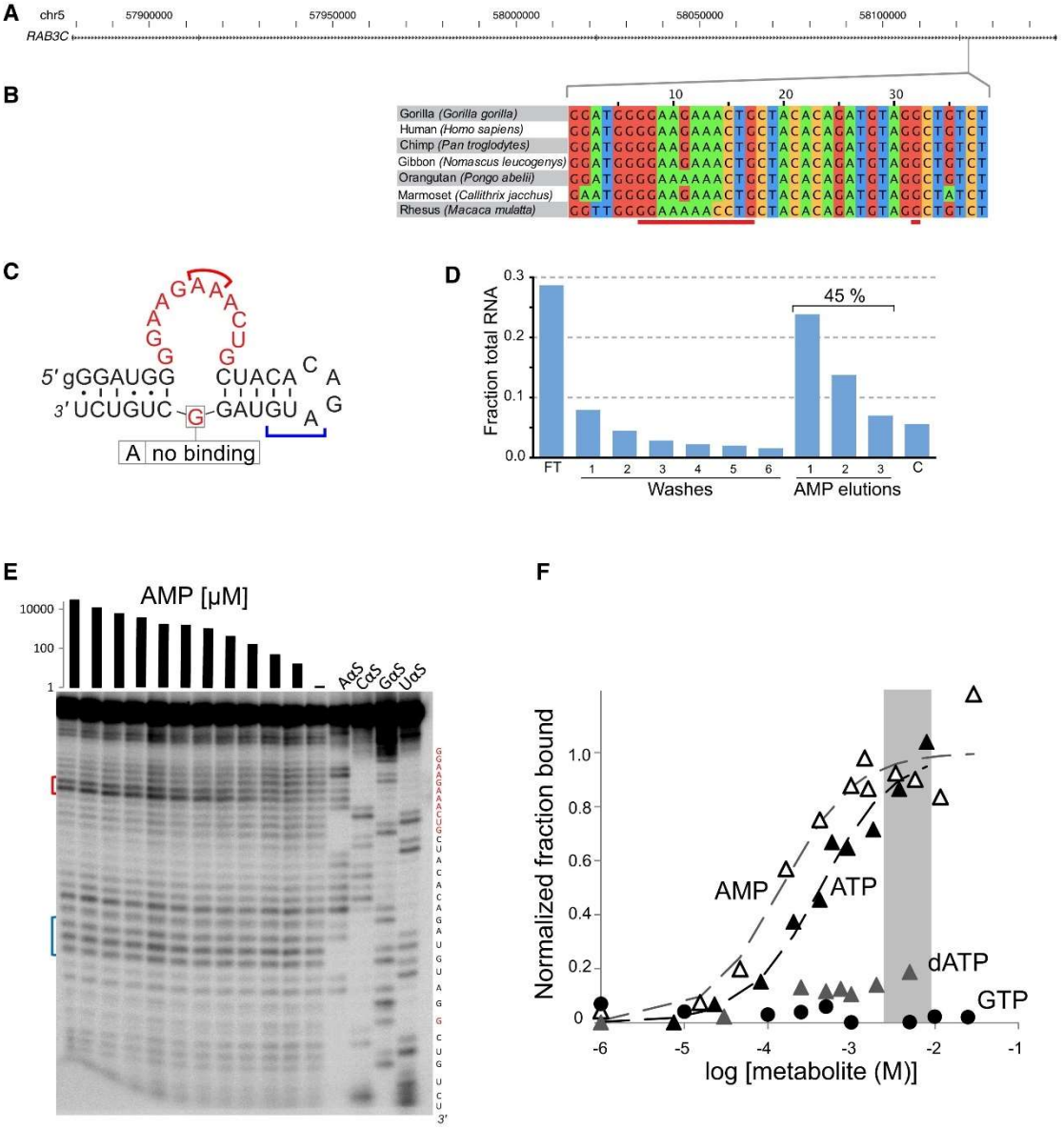


Figure 2-2 Human *RAB3C* Adenosine Aptamer Discovered Using Structure-Based Bioinformatics

(A) Mapping of the aptamer sequence on to the primate *RAB3C* gene.
(B) Sequence alignment of the primate *RAB3C* aptamers, with the adenosine-binding loop underlined in red. The marmoset and chimp sequences have activity similar to the human sequence (data not shown).
(C) Predicted secondary structure of the *RAB3C* aptamer, with the ligand-binding loop shown in red. Red and blue brackets indicate positions used to detect ligand binding and in-line probing reference bands, respectively, shown in (E). 5′ guanosine was introduced to facilitate *in vitro* transcription and is not present in the genomic sequences.
(D) Binding of ATP column and elution with AMP (FT, flowthrough; C, fraction of RNA bound to the column after three AMP elution steps).

(E) In-line probing (Soukup and Breaker, 1999) of the 3′-labeled aptamer at physiological-like conditions in the presence of AMP. The band intensities that change with AMP concentration correspond to the predicted binding loop positions, confirming the predicted secondary structure. The sequence of the RNA was determined using iodoethanol cleavage of aptamers with phosphorothioate-modified backbone at positions indicated above each lane. The aptamer shows a single-nucleotide heterogeneity at its 3′ end, resulting in doublets for all positions.
(F) Binding of the aptamer to ATP (closed triangles), AMP (open triangles), dATP (gray triangles), and GTP (closed circles) based on in-line probing experiments and fit using a single-binding-site model (see Figure S2-1 for probing in the presence of ATP, dATP, and GTP). The dissociation constants for ATP and AMP are ~400 μM and ~200 μM, respectively. Gray area corresponds to the concentration range of ATP in human tissues (Buchli et al., 1994; Kemp et al., 2007).

The bacterial aptamer resides in a gene coding for an ABC transporter, an ATP-binding permease protein. The aptamer bound to an ATP column and eluted specifically in the presence of free ATP or AMP (data not shown). The sequence was unique to *B. cytotoxicus* (NVH391-98); other Bacillus species carry mutations in both the adenosine recognition loop and the flanking helices. While these mutations do not individually abrogate ligand binding (Dieckmann et al., 1997), their combination resulted in complete loss of affinity for ATP-agarose, suggesting that the aptamer is not a conserved regulatory element among Bacillus species, but rather represents a possible gain-of-function variant of a diverged *Bacillus cereus* species (Lapidus, et al. 2008).

The aptamers found in the frog (*Xenopus tropicalis*) and mouse (*Mus musculus*) genomes also showed robust binding to ATP-agarose column and elution in presence of free ATP (data not shown). The frog aptamer maps to an intergenic region, and the mouse aptamer maps antisense to a MIRc SINE element in an intron of the *Pelp1* gene, which codes for an estrogen receptor coregulator. The distance from the end of the aptamer to the next exon of the *Pelp1* gene is only 88 nt, suggesting that the aptamer may influence the splicing of the exon, but the sequence is not conserved among rodents; therefore, it is unlikely it has a broad biological role.

Characterization of the Human *RAB3C* Aptamer

The primate aptamer maps to the last intron of the *RAB3C* gene (Figures 2-2A and 2-2B), which codes for a GTPase involved in regulation of exocytosis (Schonn, et al. 2010)) The human sequence bound an ATP column and eluted with AMP (Figure 2-2D) but not with GMP. Mutation of the "opposing" guanosine of the recognition loop to adenosine abolished ATP binding, as was shown previously for the *in vitro*-selected aptamers (Dieckmann et al., 1996). Partial hydrolysis of the RNA under physiological-like conditions (37°C, 140 mM KCl, 10 mM NaCl, 10 mM Tris-HCl or 10 mM potassium phosphate, pH 7.9, 1 mM spermidine, 1 mM $MgCl_2$), also known as in-line probing, showed significant adenosine-dependent changes in the canonical adenosine recognition loop (Soukup and Breaker, 1999) (Figure 2-2E). The apparent dissociation constants (K_{DS}) calculated from the characteristic increase of degradation in the middle of the binding loop were ~200 μM and ~400 μM for AMP and ATP, respectively (Figure 2-2F). No change in the degradation pattern was observed in the presence of 2′-dATP or GTP, confirming that the aptamer is specific for adenosine (Figure S2-1).

In vitro Selection of Human Adenosine Aptamers

The discovery of a robust human adenosine aptamer led us to consider two questions: (1) How many human aptamers exist? and (2) Do human adenosine aptamers form multiple structural families? To measure the distribution of all ATP aptamers in the human genome, independent of structure or cellular expression, we performed an *in vitro* selection experiment from an ~150 nt human genomic library (Salehi-Ashtiani et al., 2006). We carried out the selection essentially as described elsewhere (Sassanfar and Szostak, 1993), binding the pool of transcripts to ATP-agarose beads and eluting the bound fraction with ATP-Mg at physiological-

like salt conditions described earlier. This unbiased approach yielded another two aptamers: one that maps antisense to a junction between an ERV1 LTR repeat and its 3′ insertion site, and another that resides in an intron of the *FGD3* gene (Figure 2-3), which codes for a guanyl nt exchange factor regulating cell morphology and motility (Hayakawa, et al. 2008). It is surprising that both of these sequences were predicted to fold into the canonical structure found in the *RAB3C* and *in vitro*-selected adenosine aptamers. ATP column binding and "in-line" probing experiments of individual selection clones showed robust binding to ATP by the predicted recognition loop (KD ~130 μM) (Figure S2-2). The aptamer found straddling the ERV1 insertion site was not identified at other ERV1 insertion sites, suggesting that the ability to bind adenosine is not generally associated with antisense transcripts of the ERV1 and that this instance potentially represents a unique gain of function facilitated by a mobile genetic element.

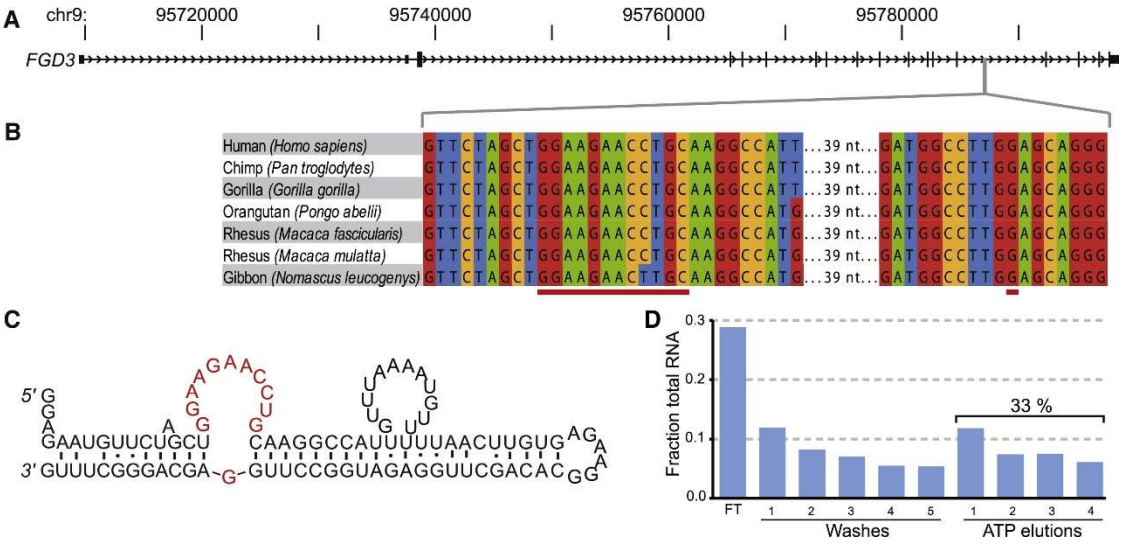


***Figure 2-3 In vitro*-Selected Human *FGD3* Adenosine Aptamer**

(A) Mapping of the aptamer to the *FGD3* intron.
(B) Aptamer sequence conservation among primates. Red bars indicate positions that form the adenosine-binding loop.
(C) Secondary structure of the aptamer based on in-line probing and thermodynamic structure predictions.
(D) Cotranscriptional binding of ATP-agarose beads and competitive elution with ATP-Mg under physiological-like conditions. See Figure S2-2 for the secondary structure of the clone sequence, full alignment of primate sequences, and in-line data with ATP.

Cotranscriptional Binding of ATP by the Human *FGD3* Aptamer

Because the *FGD3* aptamer clones included the primer binding sites introduced during the construction of the genomic library, we also tested genomic sequences lacking these artificial sequences. Even though the predicted secondary structure suggested that the adenosine aptamer fold forms in absence of the primer sequences, none of the genomic constructs bound ATP under equilibrium conditions when a purified RNA was incubated with either ATP beads or free ATP. The predicted secondary structure included potential alternative folds, suggesting that the lowest energy of the adenosine-free RNA may prevent formation of the adenosine-binding loop. We reasoned that, while the RNA had the ability to form the binding loop, as evidenced by the *in vitro*-selected clone, once purified using denaturing gel electrophoresis, the aptamer likely folded into a nonbinding conformation. However, given that some riboswitches control gene expression kinetically, sensing ligand concentrations during transcription in order to promote termination shortly after the synthesis of the aptamer domains (Wickiser, et al. 2005), we hypothesized that the adenosine-binding conformation of the *FGD3* aptamer could form transiently during transcription. To test this hypothesis, we incubated the *in vitro* transcription reaction in the presence of ATP beads and measured the fraction of the newly transcribed RNA that bound the beads and dissociated in the presence of free ATP. Under these conditions, a significant fraction of the aptamers bound ATP-agarose specifically (in the presence of 500 μM free ATP used for transcription) (Figure 2-3D). This result suggested that the *FGD3* aptamer may sense cellular ATP using a kinetic mechanism, although any biological role for this RNA remains to be elucidated.

Discussion

To date, most riboswitches have been identified in bacteria (Breaker, 2011), typically regulating transcription termination or translation initiation, and a single example has been described in algae, fungi, and plants where the thiamine pyrophosphate riboswitch regulates alternative splicing or mRNA processing (Cheah et al., 2007; Croft et al., 2007; Kubodera et al., 2003; Wachter et al., 2007). Given the broad distribution of bacterial riboswitches, the paucity of eukaryotic examples is surprising and perhaps reflects the need for novel computational and experimental tools for their discovery. Structure-based searches for known aptamer motifs and direct *in vitro* selection of ligand-binding RNAs used in this study show that both approaches yield new eukaryotic ligand-binding RNAs, although their regulatory function has yet to be demonstrated. The association of two of the vertebrate aptamers with retrotransposons also indicates that the evolution of functional RNAs in eukaryotic genomes may be strongly coupled to the activity of mobile genetic elements, potentially extending their gene regulatory functions to include eukaryotic riboswitches.

Our results show that a bacterium and several vertebrates harbor adenosine aptamers of the same structural family as have been discovered using *in vitro* selections, demonstrating that the convergent molecular evolution of adenosine binding encompasses genomic sequences. *In vitro* selection from a human genomic library revealed the same structural family, indicating that the convergent evolution of adenosine-binding RNAs extends to all human genomic aptamers. Previous studies identified several synthetic aptamers and biological riboswitches that bind the same ligand (e.g., adenine, flavin mononucleotide, S-adenosyl homocysteine) (Burgstaller and Famulok 1994, Gebhardt, et al. 2000, Gebhardt, et al. 2000, Mandal and Breaker 2004, Meli, et al. 2002, Mironov, et al. 2002, Weinberg, et al. 2010, Winkler, Cohen-Chalamish and Breaker

2002)), but adenosine is the only ligand for which the same aptamer motif has been identified in both random pools and genomes. The convergent molecular evolution spanning synthetic and genomic sequences thus mirrors that of another simple functional RNA, the hammerhead self-cleaving ribozyme (Hammann, et al. 2012, Salehi-Ashtiani and Szostak 2001)

The biological functions of the adenosine aptamers remain unknown. Our data suggest that the primate *RAB3C* and *FGD3* aptamers may sense the ATP concentration using a kinetic mechanism. In the case of the *FGD3* aptamer, we detected ATP binding only during transcription and not in equilibrium experiments, supporting the hypothesis that the genomic sequence assumes ATP-binding conformation only transiently and any biological function associated with the aptamer would have to be coupled to the cotranscriptional binding. The *RAB3C* aptamer KD for ATP is about an order of magnitude below the concentration of ATP, which, at 2 to 9 mM, is the most concentrated adenosine-containing small molecule in human tissues (Buchli and Boesiger 1994, Kemp, Meyerspeer and Moser 2007). Thus, under equilibrium conditions and normal ATP levels, the aptamer binding sites would be nearly saturated; however, if the cellular ATP concentration decreased about 10 times, a significant fraction of the aptamers would remain unoccupied. If the aptamer domain is connected to a regulatory process, it may be used to sense a large change in ATP concentration. Alternatively, if the aptamer is coupled to a cotranscriptional regulatory process, it may sense changes in ATP concentrations through a kinetic mechanism, whereby the occupancy of the aptamer domain would be proportional to the rate at which the ligand binds (Garst, Edwards and Batey 2011, Zhang, Lau and Ferré-D'Amaré 2010). A change of ATP concentration within the normal physiological range of 2–9 mM would lead to almost a 5-fold range of binding rates, and the range would be even greater if the ATP concentration deviates significantly from the normal

levels. Thus, any process kinetically coupled to the rate of binding by ATP could be sensed by this aptamer. Further experiments will be necessary to establish what, if any, biological processes these aptamers affect.

Significance

Aptamers are structured macromolecules *in vitro* evolved to bind a variety of molecular targets. In biological systems, aptamers form the ligand-binding domains of riboswitches—cofactor-dependent cis regulatory RNAs. The ligand-binding motifs identified *in vitro* and in riboswitches are diverse, even for the same ligand, suggesting that synthetic and biological aptamers follow separate evolutionary paths to achieve high-binding specificity and affinity. A paradigm of convergent molecular evolution among aptamers is represented by the adenosine-binding motif, which was identified in several independent *in vitro* experiments, but had not been previously found in genomic sequences. We used structure-based bioinformatics to measure the distribution of these aptamers in genomes and discovered that they map to organisms spanning from bacteria to humans. The human aptamer is conserved among primates, maps to an intron of the *RAB3C* gene, and binds ATP with dissociation constant about 10 times lower than physiological ATP concentration. Furthermore, to experimentally measure the distribution of all human adenosine-binding aptamers, independent of structure, we performed an *in vitro* selection from a genomic library and isolated two more aptamers. Both sequences fold into the same structure as was identified in random sequences and the *RAB3C* gene, demonstrating that its structural convergence extends to all human adenosine aptamers. One of the aptamers maps to an intron of the *FGD3* gene and exhibits full sequence conservation among primates, but the genomic construct binds ATP exclusively cotranscriptionally and not in equilibrium

experiments. The human adenosine aptamers may be kinetically controlled ATP sensors.

Acknowledgements

Nora Jameson performed the biochemical analysis of *RAB3C* aptamer. Stuart J Masuda contributed to structure-based searches and biochemical analysis of candidates. Dana Lin contributed to biochemical analysis of candidates. Rosa Larralde-Ridaura performed structure-based searches.

Experimental Procedures

Structure-Based Searches

We used the RNABOB program (courtesy of S. Eddy, Howard Hughes Medical Institute; ftp://selab.janelia.org/pub/software/rnabob/) to perform motif searches (Riccitelli and Lupták, 2010). The secondary structure of the adenosine aptamer was based on *in vitro*-selected RNAs that bind adenosine-containing molecules: ATP (Sassanfar and Szostak, 1993), nicotinamide adenine dinucleotide (Burgstaller and Famulok, 1994), S-adenosyl methionine (Burke and Gold, 1997), and S-adenosylhomocysteine (Gebhardt et al., 2000). In addition, we incorporated constraints based on the solution structures of the AMP-bound aptamers (Dieckmann et al., 1996, 1997; Jiang et al., 1996). The secondary structure was divided into the recognition loops (s1 and s3) flanked by single strict base pairs (r2 and r3 elements). The rest of the helices were allowed to contain G·U wobble pairs (h1 and h4 elements). Box 1 shows the descriptors for aptamers with the recognition loop in the 5′ (#1) and 3′ (#2) strands corresponding to the circularly permuted motif:

Box. 1 Descriptors for Aptamers with the Recognition Loop in the 5′ and 3′

#1
h1 r2 s1 r3 h4 s2 h4′ r3′ s3 r2′ h1′
h1 0:0 NNNN:NNNN
r2 0:0 N:N TGCA
s1 0 GGAAGAAMCUG
r3 0:0 N:N TGCA
h4 0:0 NNNN:NNNN
s2 0 NNN[10]
s3 0 G

#2
h1 r2 s1 r3 h4 s2 h4′ r3′ s3 r2′ h1′
h1 0:0 NNNN:NNNN
r2 0:0 N:N TGCA
s1 0 G
r3 0:0 N:N TGCA
h4 0:0 NNNN:NNNN
s2 0 NNN[10]
s3 0 GGAAGAAMCUG

RNA Transcription

RNA was transcribed at 37 °C for 1 to 3 h in a volume of 20 μL containing 40 mM Tris chloride, 10% dimethyl sulfoxide (DMSO), 10 mM dithiothreitol (DTT), 2 mM spermidine, 2.5 mM each CTP, GTP, and UTP, 250 μM ATP, 2.25 μCi [α-^{32}P]-ATP (Perkin Elmer, Waltham, MA, USA), 25 mM $MgCl_2$, one unit of T7 RNA polymerase, and 0.2 μM of DNA template. DMSO was used to increase transcript yields, as documented in previous studies (Chen and Zhang, 2005). The transcripts were purified using denaturing PAGE.

3′-Terminus Labeling

RNA ligation reactions were performed essentially as previously described (England et al., 1980). Briefly, RNA transcribed in the absence of [α-^{32}P]-ATP was PAGE purified and ligated at

37 °C for 3 h in a volume of 10 μL, containing RNA ligase buffer (New England Biolabs [NEB] Ipswich, MA, USA), 2 μCi [5′-^{32}P] cytidine 3′, 5′-bisphosphate (Perkin Elmer) and one unit of T4 RNA ligase (NEB) and PAGE purified again.

In-Line Probing

In-line probing reactions were performed as previously described (Regulski and Breaker, 2008; Soukup and Breaker, 1999). The 3′-end labeled RNA was incubated with varying amounts of ligand for 1 or 2 days at 37 °C in a buffer containing 140 mM KCl, 10 mM NaCl, 10 mM potassium phosphate, pH 7.9, or Tris chloride, pH 7.9, 1 mM $MgCl_2$, and 1 mM spermidine. Triphosphorylated ligands (ATP, dATP, GTP) were prepared as 1:1 complexes with Mg2+. AMP was used directly, without additional divalent metal ions. The partially hydrolyzed RNAs were resolved using denaturing PAGE, exposed to phosphorimage screens (Molecular Dynamics/GE Healthcare, Pittsburgh, PA, USA), and scanned by GE Typhoon phosphorimager. The band intensities were analyzed by creating line profiles of each lane using ImageJ, exporting the graphs into IgorPro, and fitting all peaks by Gaussian curves. The areas of the fitted curves were used to measure intensity changes of the binding loop and divided by intensities of a control band. The resulting ratios were plotted in Excel as a function of ligand concentration and modeled with a dissociation constant equation for a single ligand:

The model was fit to the data using a linear least-squares analysis and the Solver module of Microsoft Excel.

Sequence of the adenosine-binding loop were confirmed by running iodoethanol-cleaved RNAs with α-phosphorothioate nt incorporated at indicated positions (Gish and Eckstein, 1988; Ryder and Strobel, 1999). The dominant product of phosphorothioate sequencing, a 2′-3′ cyclic phosphate, is the same as in in-line probing (Soukup and Breaker, 1999), allowing a direct readout of the positions affected by the bound ligand.

In vitro Selection

The DNA pool used for the *in vitro* selection was derived from the human genome and described previously (Salehi-Ashtiani et al., 2006). Purified RNA transcripts were precipitated, dried, and resuspended in 200 µL binding buffer containing 140 mM KCl, 10 mM NaCl, 10 mM Tris chloride, pH 7.5, and 5 mM $MgCl_2$ and heated to 70 °C before loading on to C8-linked ATP-agarose beads (Sigma-Aldrich, St. Louis, MO, USA) equilibrated in the binding buffer. Flowthrough was collected after the columns were capped and shaken for 20 min at room temperature. The beads were washed with 200 µL of binding buffer, and potential aptamers eluted with the same buffer supplemented by 5 mM ATP·Mg with 30 min of shaking at room temperature. Each fraction was analyzed for radioactivity using a liquid scintillation counter. Elutions were pooled, desalted using YM-10 spin filters (Millipore, Billerica, MA, USA), precipitated, dried, and resuspended in H2O.

In vitro Selection Primers

Forward: 5′ GATCTGTAATACGACTCACTATAGGGCAGACGTGCCTCACTAC 3′

Reverse: 5′ CTGAGCTTGACGCATTG 3′

Reverse Transcription

RNA was reverse transcribed in 20 μL using the Promega reverse transcription buffer, 2 μM reverse primer, and RNA recovered from the previous selection round. The RNA and primer were annealed by heating at 65 °C and cooling to room temperature before 1 unit of Thermoscript (Invitrogen, Grand Island, NY, USA) and Improm II (Promega, Madison, WI, USA) reverse transcriptases each were added. The reaction was initiated for 5 min at 25°C, and then the temperature was ramped to 42°C, 50°C, 55°C, and 65 °C for 15 min each before the enzymes were inactivated at 85 °C for 5 min.

Amplification

DNA was amplified by using DreamTaq Master Mix (Fermentas, Glen Burnie, MD, USA), 2 μM forward primer, 2 μM reverse primer, and DNA from reverse transcription. DNA was initially denatured at 95 °C for 1.5 min (30 s for subsequent denaturing steps), annealed at 55 °C for 30 s, and extended at 72 °C for each cycle. Optimum number of PCR cycles was determined for each selection round by comparing 8-, 12-, 16-, and 20-cycle aliquots on agarose gel.

Cloning

After five rounds of selection, the pool was cloned using the Topo TA cloning kit (Invitrogen) and individual colonies were directly PCR amplified and sequenced. Only two unique sequences, the *FGD3* and ERV1, were found:

FDG3 (chr9:95,787,061-95,787,193)
GGTCCTGG GGAGTCTAAC CCCCGGGTTG GTGCCCTGAC AGGAGgATGT
TCTtGCTGGA AGAACCTGCA AGGCCATTGT TTAAgATGTT TTAACTTGTG
GGAAGACACA GCTTGGAGAT GGCCTTGGAG.

ERV1 (chr3:162359023-162359154)
AGAAGC AAGAAGGAAG AAAcTGCCAc CGTTGTCATC TTGTGTCTGa AATTGGTGGn
TTCTTGGTCT CACTGgCTTC AAGAATGAAG CCGTGGACCC TCGCGGTGAG
TGTTACAGTT CTTAAAGGCG GTGTCT.

Lowercase letters are mutations acquired during the *in vitro* selection or library construction, as compared with the reference genome. The putative adenosine-binding loop is highlighted in bold. The ERV1 aptamer contained an A-to-C mutation in the recognition loop (underlined) in all sequenced clones. The genomic sequence is likely to bind adenosine, because this variant has been observed among *in vitro*-selected clones (Sassanfar and Szostak, 1993). As with the other two aptamers, in-line probing showed an ATP-dependent change in the hydrolysis pattern for the predicted loop characteristic of this class of aptamers (Soukup and Breaker, 1999), confirming that the cloned sequence forms the same binding loop.

Co-transcriptional Binding Assay

RNA was transcribed in a Spin-X column (Corning, Corning, NY, USA) for 30 min at 37 °C in 50 µL solution containing 40 mM Tris chloride, 10% DMSO, 10 mM DTT, 5 mM each GTP, UTP, and CTP, 500 µM ATP, 3.75 µCi [α-^{32}P]-ATP (Perkin Elmer), 25 mM $MgCl_2$, 1 unit of T7 RNA polymerase, 50 pmol of DNA template, and C8-linked ATP-agarose beads, washed, and equilibrated in transcription buffer. The columns were centrifuged at 4,000 g for 1 min, and flowthrough was collected. The columns were washed in 50 µL of the binding buffer and centrifuged. Elutions were collected in the ATP·Mg elution buffer following 30 min of shaking. Each fraction was resolved on a denaturing PAGE gel to separate the full-length transcripts from

unincorporated [α-^{32}P]-ATP and shorter transcripts. The gels were exposed to phosphorimage plate and scanned, and the amount of full-length RNAs measured with ImageJ software.

Supplementary Figures

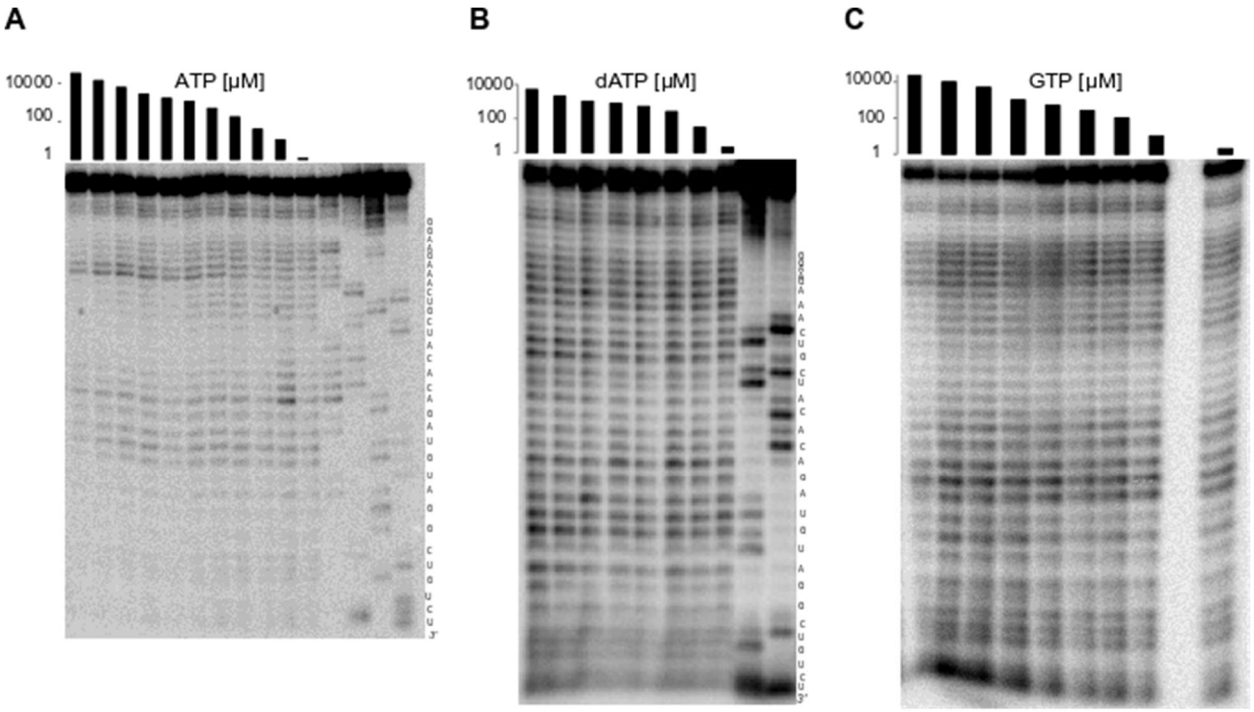


Figure S2-1, related to Fig. 2-2. In-line probing of the *RAB3C* aptamer with (A) ATP, (B) 2′-deoxy-ATP, and (C) GTP. The sequence lanes correspond to the aptamer with α-phosphorothioate nucleotides (indicated above each lane) incorporated into the sequence and cleaved with iodoethanol. The sequence of the aptamer is located on the right side of the gel with red nucleotides indicating the binding loop. The red brackets denote the binding loop positions sensitive to adenosine and the blue brackets denote the nucleotides that were used for normalization to calculate fraction of the aptamer bound by the ligand (Fig. 2-2).

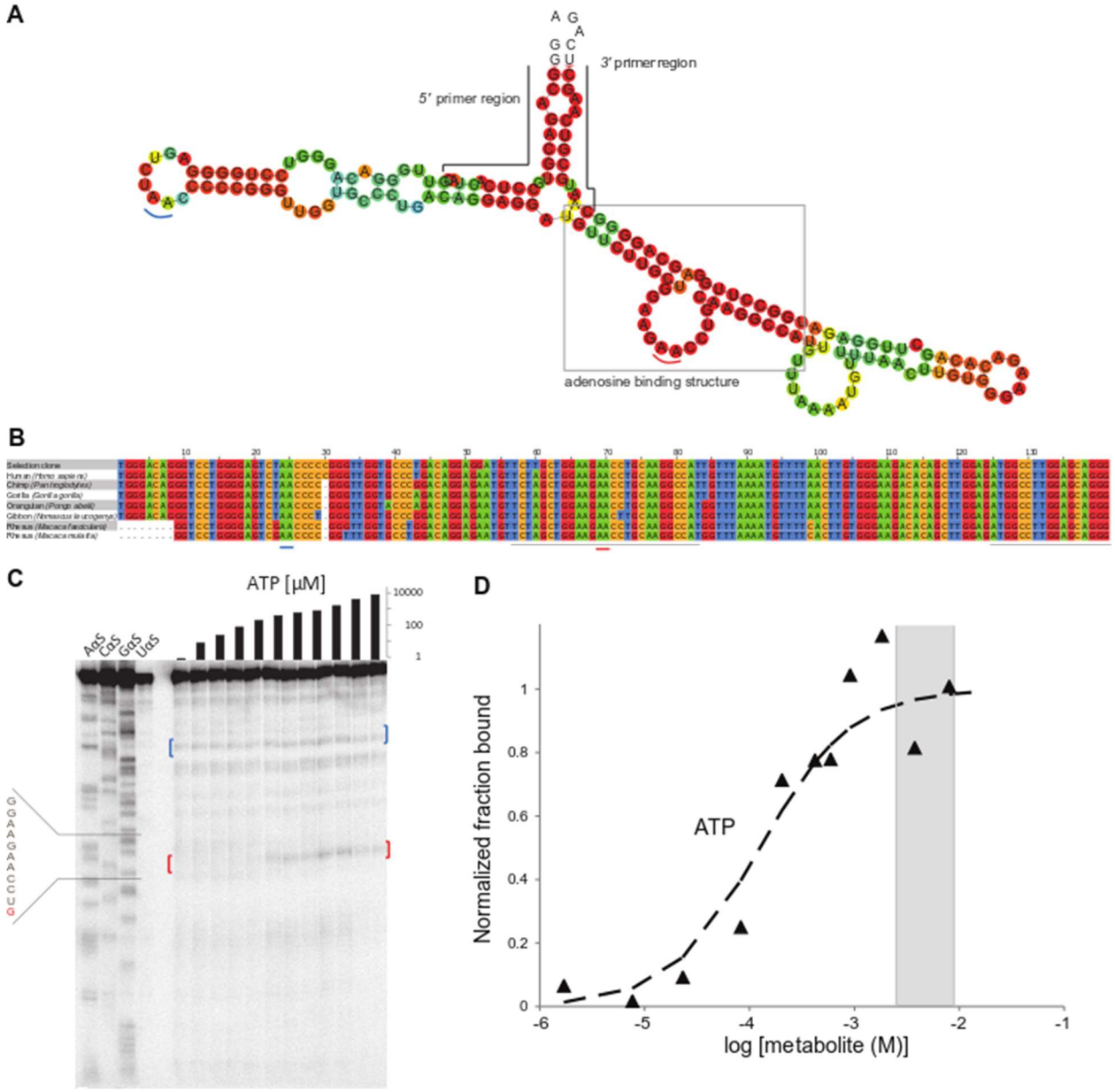


Figure S2-2, related to Fig. 2-3. Characterization of the human *FGD3* aptamer**.** (A) Structure of the *in vitro* selected aptamer clone, including non-genomic, primer-binding sequences, as predicted by RNAfold and colored by the probability of the structure (blue to red for the lowest to highest probability). Canonical adenosine aptamer structure presented in Fig. 201 is boxed. (B) Sequence alignment of the aptamer region in primates. Grey lines correspond to the boxed region in (A). Red and blue brackets indicate the nucleotides that correspond to the bracketed nucleotides in the in-line probing gel (C). Other areas of RNA hydrolysis correlate well with loops of the predicted secondary structure in (A). Red brackets denote the binding loop positions sensitive to the ATP concentration and the blue brackets denote the nucleotides that were used for normalization to calculate fraction of the aptamer bound by ATP. Sequence of the adenosine binding loop is indicated on the left side of the image and is based on phosphorothioate sequencing using A-,C-, and G- α- phosphorothioate (bands corresponding to the uracil positions are weak, therefore the U is labeled in grey in the sequence). (D) Solution binding of the aptamer to ATP, with an apparent KD of ~ 130 μM. Gray area in this graph and in Fig. 2-2 corresponds to the concentration range of ATP in human tissues.

Chapter 3 Characterizing novel ATP aptamers in humans

Introduction

Our appreciation for the roles RNA plays in biology has increased dramatically since the times of the central dogma of RNA (Cech and Steitz 2014), but as most of the function of non-coding RNAs were primarily discovered in less complex organism, it would be fair to assume they are just remnants of an RNA world. However, the known roles of RNA in our own bodies have exploded over the course of our lifetimes and we have come to appreciate the ~99% of our DNA that is non-protein coding. RNAs that thwart the central dogma are pervasive within our genomes. One third of our genes are thought to be regulated by miRNA (Hammond 2015), at least 40% of transcript occur in both directions (Rosikiewicz and Makałowska 2016), and ~77% of our multi-exonic genes are thought to have retained introns (Braunschweig, et al. 2014). Clearly the role of RNA in human biology is much grander than originally thought and many efforts have been made towards understanding the roles of all noncoding RNA in humans both out of fascination and due to their implications in disease (Esteller 2011).

The discovery of human aptamer for ATP was the earliest use of genomic SELEX for small molecule targets. Since then, additional examples of small molecule recognition by human RNA have been found in the form of a folic acid aptamers (Terasaka, et al. 2016) and evidence of widespread GTP aptamers (Curtis and Liu 2013, Curtis and Liu 2014) that are present in humans but absent in the bacterial genome associated with a heavy reliance on functional RNA. At the time of the original ATP genomic selection, cloning and sequencing was used to identify selected aptamers. However, this method can miss low abundant sequences in the pool, which has been observed to include important aptamers in previous genomic selection (See Chapter 1).

It's estimated that ~75,000 RNAs the human genome binds GTP, with nucleotide binding being so pervasive, it's likely that the ATP selected pool still contains undiscovered aptamers.

The three known ATP binding aptamers in humans all conformed to a previously discovered ATP binding motif, discovered through random-sequence SELEX (Sassanfar and Szostak 1993). When the ATP binding aptamer motif was first discovered, 16 clones contained a consensus loop GGWAGADNHTG help opposite a single G bulge by base paired stems on either side. This motif has returned in several selection and is responsible for binding of the adenosine moiety of nicotinamide adenine dinucleotide (Burgstaller and Famulok 1994), S-adenosyl methionine (Burke and Gold 1997), and S-adenosylhomocysteine (Gebhardt, et al. 2000) selected aptamers. It returned again in the genomic selection for ATP aptamers, but no new modes of ATP binding were discovered, despite novel motif discovery being one of the primary advantages of genomic SELEX. Identification of less abundant aptamers would not only increase our immediate knowledge of human RNA functions but could also elucidate novel motifs for ATP binding that can then be used to inform later bioinformatic experiments. This chapter will describe the characterization of ATP binding candidates discovered by high throughput sequencing

Additional sequences in the ATP binding pool

In order to identify low-abundance aptamers, the round 6 ATP binding pool from genomic selection (See Chapter 2) was submitted for sequencing on the Hiseq of the UCI Genomics High Throughput Facility. This sequencing generated reads that mapped to ~4,300 location in the human genome but only about 40 locations with at least four reads (reported in Table 3-1). The sequences with at least four reads were inputted into MEME motif prediction

software which did not identify a common motif among them. Closer inspection of the sequences did reveal some interesting candidates, however. While none of the additional sequences had the exact GGWAGADNHTG motif, three of them had similar sequences that could form across a bulged G. These sequences map to the *PRR5* gene on chromosome 22, the L1PA16 LINE element on chromosome 14, and the THE1B repeat element on chromosome 18. They each contained mutations or insertions not seen in previous selection random sequence SELEX where those modifications may have been disadvantageous enough to prevent enrichment.

The *PRR5* sequence is predicted to present the loop GGAAG**G**ACUG, substituting a G in place of what was thought to be a conserved A. Moreover, instead of a single bulged G, the loop is predicted to be held opposite a UGC bulge. The sequence mapping to the L1PA16 LINE element contains a loop that contains an insertion forming a 12-nucleotide loop GGAAGA**G**AACUG. The loop on the THE1B sequence, GGA**G**G**U**AACUG, has two of its conserved adenosines substituted. Despite these differences, all three were shown to bind immobilized ATP and elute with free ATP when selectively amplified from the pool. Site specific mutations to these putative recognition loops abolished ATP binding as seen in (Chapter 4), confirming the proposed binding site and expanding the functional variants of the ATP binding motif.

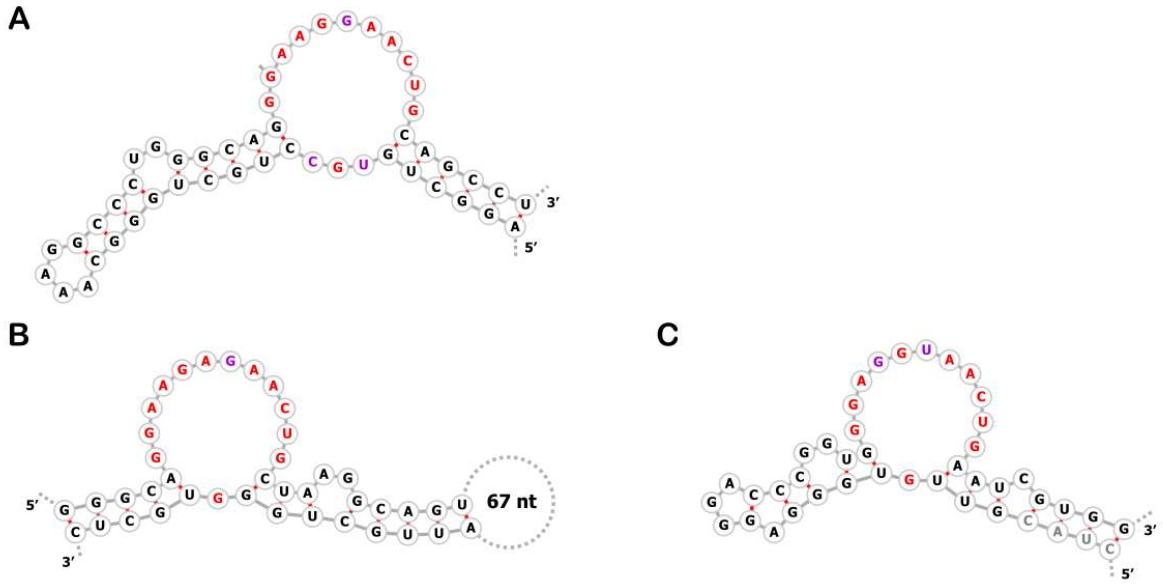


Figure 3-1 Proposed ATP binding motifs ATP binding domains for the (A) *PRR5*, (B) L1PA16, and (B) THE1B aptamers. Canonical ATP binding motif shown in red with modifications in purple. Greyed out letter correspond to the artificial constant region with is shown to form part of the stem of the THE1B structure

A candidate novel ATP binding motif

Most of the sequences found did not contain anything resembling the known ATP binding motif, presenting the opportunity to discover entirely novel modes of ATP binding. One of the most abundant of these sequences was investigated further. This sequence maps to chromosome 15, not to any known genes, but a human EST (BU853031) and is conserved among mammals. When amplified from the pool with sequence specific primers, this sequence seemed to bind and elute with ATP but not when the primer sequences were removed. While this had been the case for the *FGD3* aptamer, secondary structure prediction revealed an alternative fold for the genomic *FGD3* construct to explain this observation. The genomic *FGD3* aptamer was later shown to bind ATP cotranscriptionally. For the chromosome 15 sequence, free energy

secondary structure prediction performed with Vienna RNAfold showed no differences in predicted base pairing between the selected sequence and primer-less sequence, though the primer-less construct had lower confidence, possible suggesting a stabilizing effect from the primers. The primer-less construct was also tested for cotranscriptional binding of ATP but showed no detectable binding. In line probing was used to inform the secondary structure and identify regions responsive to ATP concentrations. In line data and genomic conservation data from BLAT were used to design 10 additional genomic constructs of various length, but none bound and eluted from ATP-agarose efficiently either purified or cotranscriptionally. With multiple experiments getting us no closer a functional genomic construct of chromosome 15 and several potential aptamers still in the pool, efficient analysis begs multiplexed experiments and data analysis.

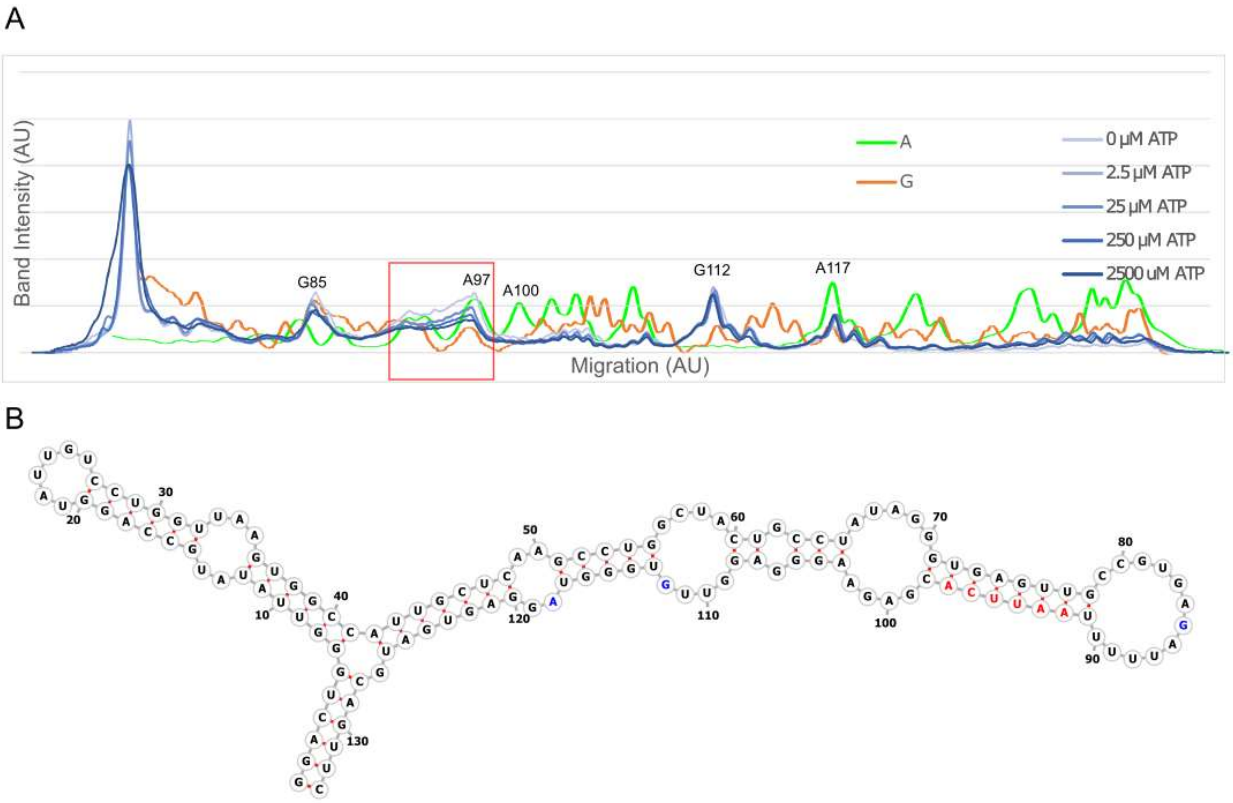


Figure 3-2 Potential ATP aptamer from Chr 15 (A) In line probing profiles for chromosome 15. Sequencing lanes show in green (A) and orange (G). Area most responsive to ATP concentration is boxed in red. (B) Thermodynamic structure prediction of the genomic construct informed by in line probing. Blue nucleotides correspond to areas of increase hydrolysis used to constrain the structure model. The regions protected from hydrolysis by ATP is indicated with red nucleotides.

Multiplexed mutagenesis and reselection

Mutagenesis is often used in random sequence *in vitro* selections to increase diversity and allow for the isolation of more optimal aptamers. Genomic SELEX is interested in biological aptamers, not necessarily optimal ones, so mutagenesis isn't usually used. Mutants would have little biological relevance except in the case of known SNPs. However, mutants can also provide information regarding which portions of a sequence are important for activity and recognition based on sequence conservation. They also provide information about secondary structure from covariation modelling. Thus, the ATP binding pool was mutagenized and subjected to additional rounds of selection in order to garner the benefits of these mutants for comparative sequence analysis. This is often performed on selection winners for both enhanced activity and information on sequence and structure conservation. Applying this strategy to a pool of active molecules all reveals conserved elements on the entire pool at once.

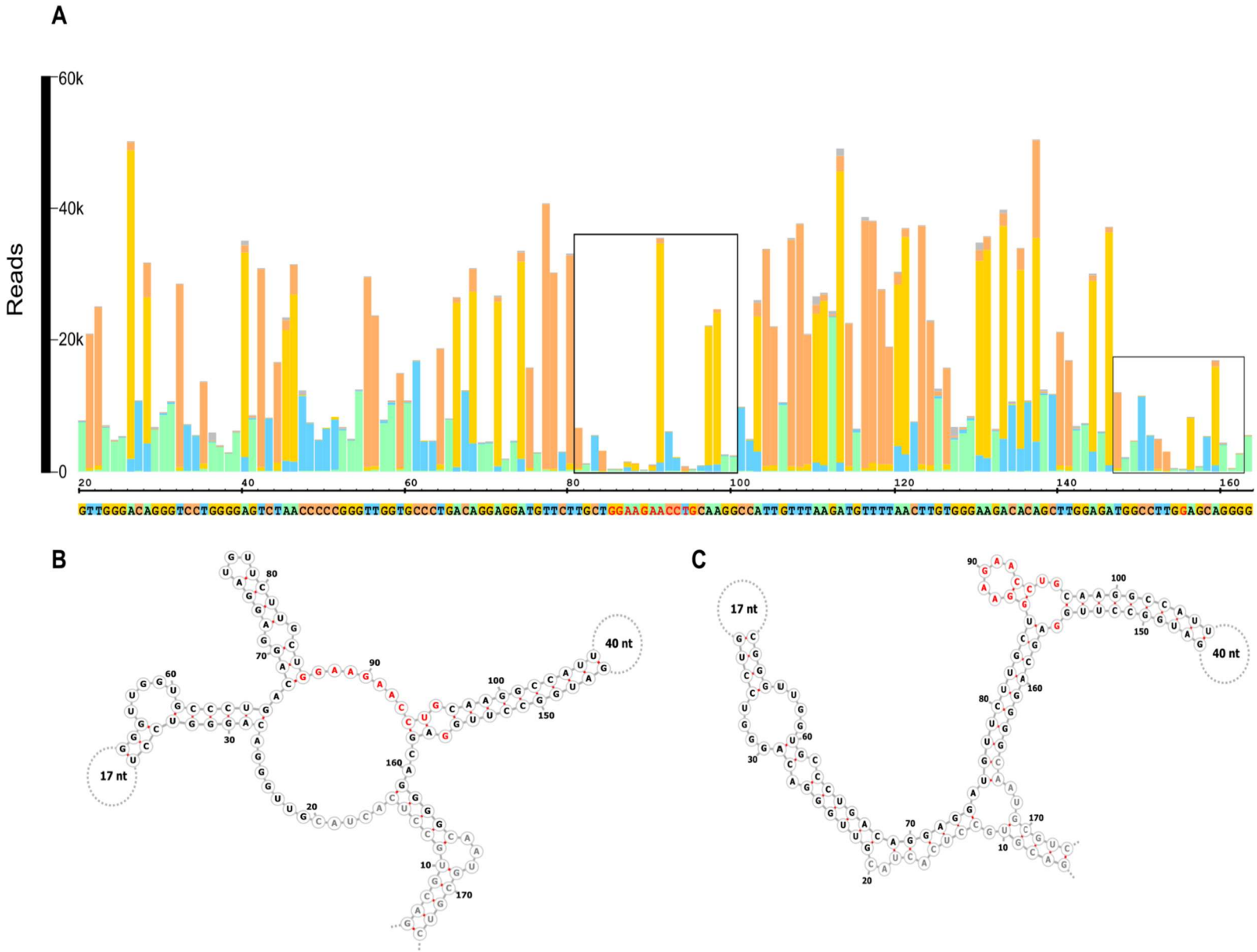
The round 6 pool was mutagenized by including mutagenic nucleotide analogs, 8-oxo-GTP and dPTP in PCR (Zaccolo, et al. 1996) and three additional rounds of enrichment for ATP binding were performed before sequencing. Sequencing reads were mapped to their parent sequence (from the previous sequencing) using bowtie2 and mismatches were compiled on IGB to reveal mutable regions. Mismatch pile ups were generated for 11 of the previous sequences (See supporting info). Biases in the mutation were observed with roughly 4-fold higher mutation rates for AT vs GC indicating the need for more 8-oxo-GTP in the initial mutagenesis. Conservation of nucleotides were evaluated with this bias in mind. To assist in secondary structure predictions sequences were clustered via fastaptamer (Alam, Chang and Burke 2015) and entire clusters were inputted into covariation based secondary structure predictors. The

known *FGD3* and ERV1 aptamers were used to evaluate the technique

Mutagenesis reveals conservation in ATP binding loop

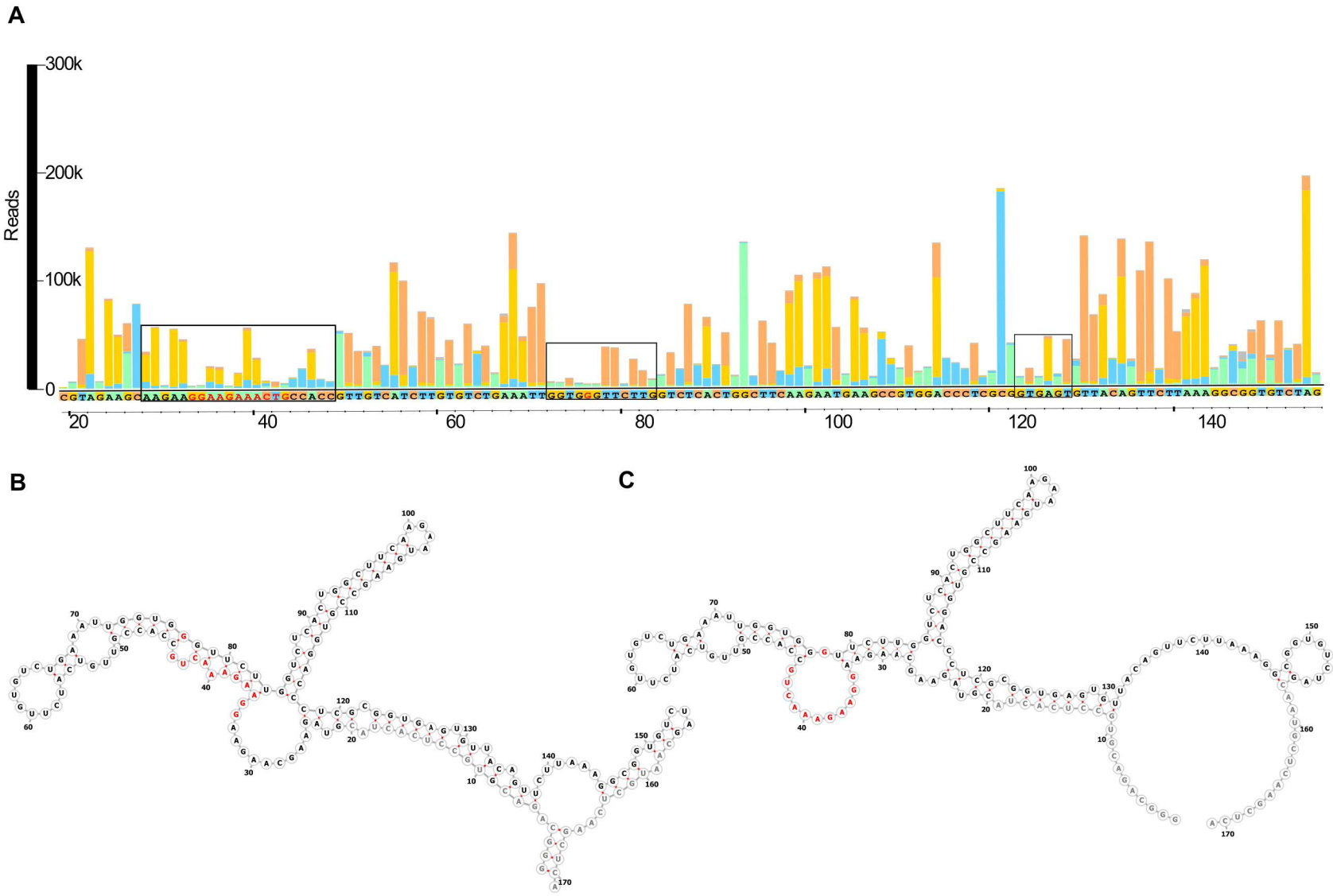
Comparing the mutants for the *FGD3* sequence reveals high level of conservation for all but one of the nucleotides in the ATP recognition loop, consistent with its proposed role in ATP binding. We also see low relative mutation rates in the bases directly flanking the consensus loop and the opposing G. The A to G mutation in the 7th nucleotide of the loops still conforms to the *in vitro* selection derived consensus from synthetic pools. Mutational data was also able to provide support for these conserved nucleotides being placed in the correct structure. Our proposed structure for the genomic *FGD3* ATP aptamer is not favored according to thermodynamic predictions (Vienna RNAfold) (Hofacker 2003) which predicts an alternate fold. However, with the mutant *FGD3* sequences submitted to the collection of covariation-based structure modelling software (Torarinsson and Lindgreen 2008), most of the programs included were able to predict the formation of stems that allow the recognition loop to be placed across from the requisite bulged G. RNAalifold (Bernhart, et al. 2008) was one such program and was chosen for use in subsequent analysis due to its accessible webserver and graphical output.

Figure 3-3 *FGD3* aptamer mutants. **(A)** Mismatch pile up of mutants of the *FGD3* aptamer generated by IGB. The bar color indicates the identity of the mutation green (A) blue (T) orange (C) gold (G) and height shows number of mutants. Areas with high levels of conservation (boxed) are observed around the ATP binding motif shown in red except for A to G mutations which maintain binding and base pairing. **(B)** Thermodynamic structure prediction of the *FGD3* aptamer. ATP binding nucleotides are shown in red. Constant region is shown in gray. **(C)** Thermodynamic structure prediction informed by mutants. ATP motif forming stems are correctly predicted.


A
Reads
60k
40k
20k
0
20
40
60
80
100
120
140
160
GTTGGGACAGGGTCCTGGGGAGTCTAACCCCCGGGTTGGTGCCCTGACAGGAGGATGTTCTTGCTGGAAGAACCTGCAAGGCCATTGTTTAAGATGTTTTAACTTGTGGGAAGACACAGCTTGGAGATGGCCTTGGAGCAGGGG
B
17 nt
40 nt
C
17 nt
40 nt

In the case of the ERV1 aptamer sequences, the ATP binding loop again saw very few mutations. Moreover, the shorter of the adjacent stems was also especially sensitive to mutation. There are also regions of surprisingly low mutation rates in the downstream structure, thought to fold independently from the ATP binding motif. This portion of the ERV1 sequence is not thought to contribute to ATP binding, but its low mutation rate may indicate otherwise. While the alifold structure prediction of ERV1 sequence cluster did not predict an active fold for the ATP binding motif, several sequence clusters were noted that originated from the ERV1 sequence but were not included in the cluster because it exceeded the edit distance used. This highly mutagenized group of ERV1 sequences with edit distance greater than 30 was used to predict a secondary structure in alifold and generated a fold congruent with our model for ATP binding. The higher rate of mutation in the subset of sequences allowed for more examples of covariation

Figure 3-4 ERV1 aptamer mutants. **(A)** Mismatch pile up of mutants of the ERV1 aptamer generated by IGB. The bar color indicates the identity of the mutation green (A) blue (T) orange (C) gold (G) and height shows number of mutants. Areas with high levels of conservation (boxed) are observed in the ATP binding portions shown in red as well as an additional region downstream of the binding motif. **(B)** Thermodynamic structure prediction of the ERV1 aptamer. ATP binding nucleotides are shown in red. Constant region is shown in gray. **(C)** Thermodynamic structure prediction informed by mutants. ATP motif forming stems are correctly predicted.


A
300k
200k
100k
0
Reads
20
40
60
80
100
120
140
B
C

As an abundant sequence with no motif recognized the chromosome 15 sequence was also examined further. However, there seemed to be no obvious regions of conservation. If anything, there is a slightly higher mutation rate in the mammalian conserved regions which includes the ATP responsive region indicated by in line. The region with most conservation seems be the area around ~130 nucleotides into the sequence where mammalian conservation drops off. The clustered sequences were submitted to RNAalifold to obtain a more informed prediction of the active structure which did differ from the previous model. This structure was used to design six additional genomic constructs for testing. However, none of these bind ATP convincingly.

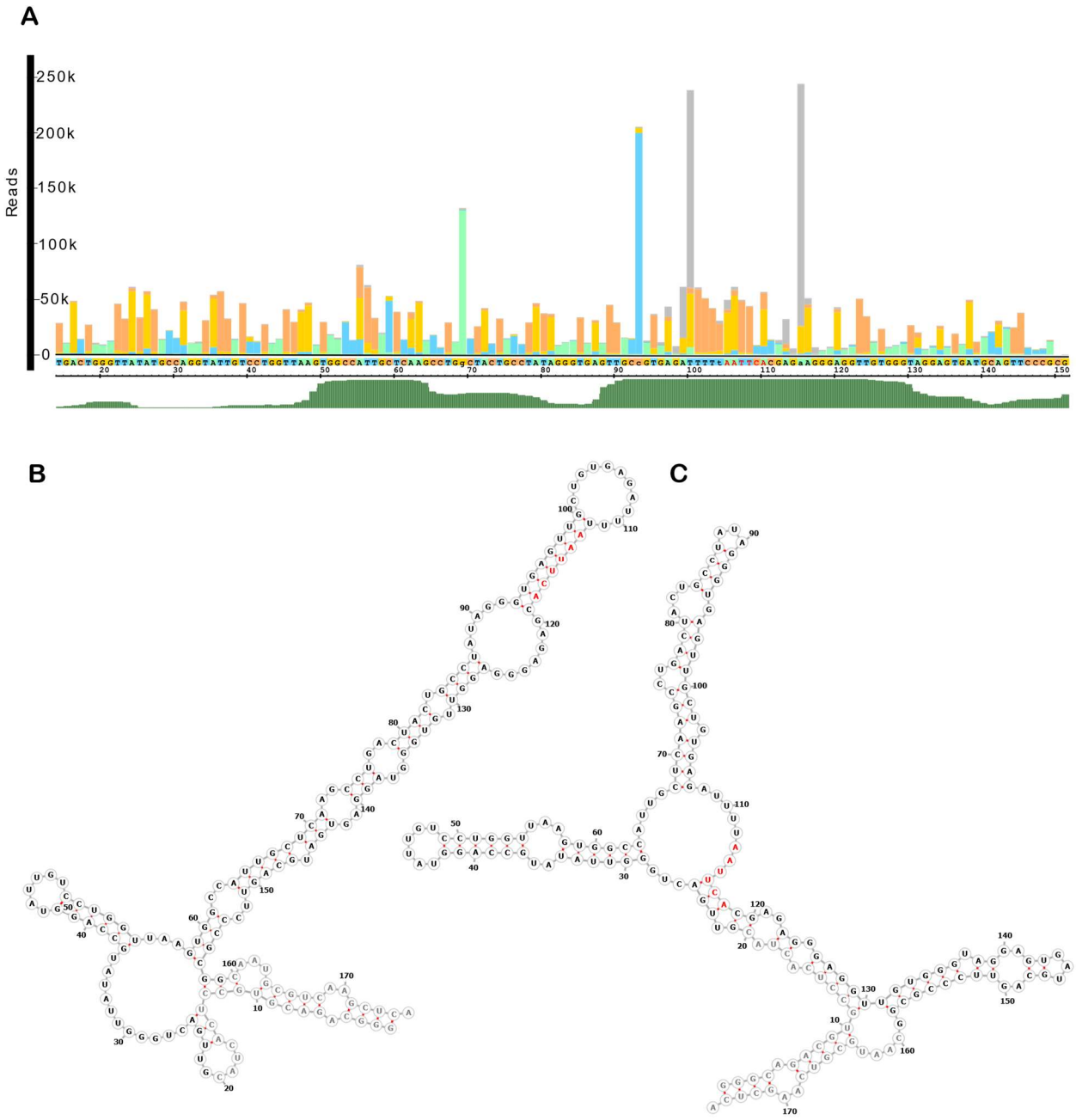


Figure 3-5 Mutants of the Chr15 aptamer. (A) Mismatch pile up of mutants of Chr15 sequence generated by IGB. The bar color indicates the identity of the mutated nucleotide green (A) blue (T) orange (C) gold (G) and height shows number of mutants. Mammalian conservation shown below in green. Nucleotides protected from hydrolysis by ATP are shown in red (**B**) Thermodynamic structure prediction of Chr15 sequence. Constant region is shown in gray. Nucleotides protected from hydrolysis by ATP are shown in red (**C**) Thermodynamic structure prediction informed by mutants. Constant region is shown in gray. Nucleotides protected from hydrolysis by ATP are shown in red

The mutagenesis, selection, and sequencing process is intended to generate data for several sequences at once. It is also meant to reveal potential candidate to tested based on this data. One such candidate is a sequence that map to chromosome 20 of from previous high throughput sequence. Mutants mapped to this sequence revealed a highly conserved area in one region of the sequences. Further investigation reveals that the region conserved in selection corresponds to a conserved element among mammals. Though the conservation in both SELEX and nature is intriguing, this sequence has not been shown to bind ATP.

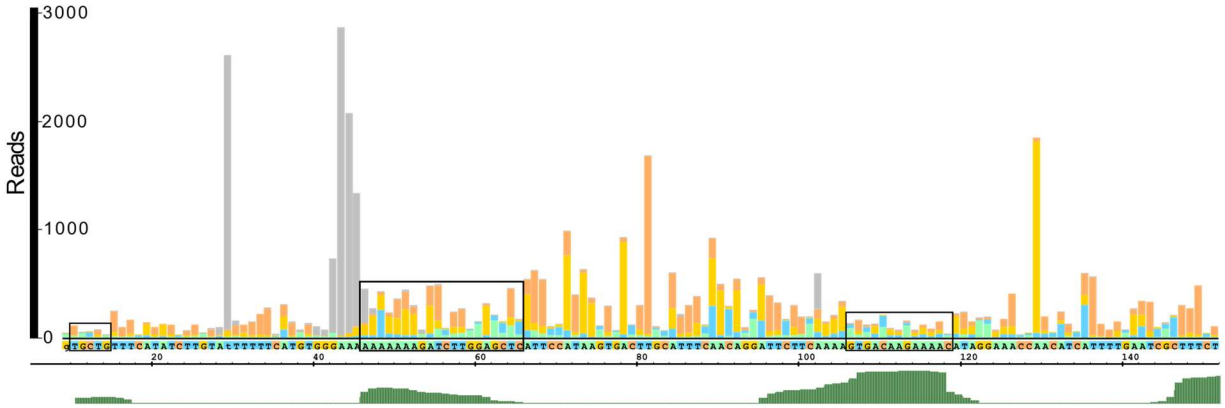


Figure 3-6 Mismatch pileup of chromosome 20 conserved sequence. Mismatch pile up of mutants of Chr15 sequence generated by IGB. The bar color indicates the identity of the mutated nucleotide green (A) blue (T) orange (C) gold (G) and height shows number of mutants. Mammalian conservation shown below in green. Areas of low mutation rate in SELEX (boxed) were seen to correspond with the mammalian conserved element

Conclusion

High throughput sequencing enabled genomic SELEX to discover three more examples of ATP binding RNAs encoded in the human genome. Their variation from the SELEX derived consensus sequence kept these sequences from being discovered from previous bioinformatic approaches but their appearance here could inform later attempts at bioinformatic discovery. Once again, this demonstrates the ability of genomic SELEX to identify functional RNAs

inaccessible by other methods. The amount of data generated from these experiments has presented a new challenge, as several functional RNAs are likely to exist in the large pool of candidates obtained. Because abundance in genomic SELEX does not indicate biological relevance, it's preferable to analyze as many of the sequences that come out of selection as possible. However, characterizing each of the 40 sequences with at least 4 reads is quite a task, let alone the 4,300 potential sequences with lower read counts. This requires multiplexed methods of biochemical analysis of *in vitro* selected pools to define their characteristics and identify the most promising candidates for further study

Mutagenesis and *in vitro* selection was shown to identify conserved elements of ATP binding and aid in secondary structure predictions for the entire pool at once. The experiment demonstrated sequence conservation in loops known to bind ATP, and the detected mutants enabled covariation-based structure prediction of the active fold. The variants of the ATP binding motif were unfortunately not detected in this experiment which may indicate they were selected away. Though, surprisingly, the chromosome 15 sequence with no known ATP binding motif survived the selection despite low ATP binding.

The chromosome 15 sequence in particular saw a great amount of effort towards its characterization. It had several constructs screened and structure prediction software, in line probing, and data from mutants seemed to get us no closer to isolating a genomic construct, reinforcing the need for multiplexed methods of characterization of selected pool. These sequences will be carried along throughout this dissertation as it describes the development of methods aimed at doing just that.

Materials & Methods

Sequencing & Read processing

The round 6 ATP binding pool (Chapter 2) was sequence on the HiSeq-MFG paired end 150 read length. Reads were merged with PEAR using default settings. Adapters were clipped with cutadapt:

```
-a CAATGCGTCAAGCTCAG
-g GATCTGTAATACGACTCACTATAGGGCAGACGTGCCTCACTAC
```

Read processing

Sequences were initially identified using fastaptamer count and cluster options. Clustering was run with an edit distance of 10 with no filter flag.

Mapping of sequences for identification

Reads where mapped to the hg38 human genome using bowtie2 options `-N 1 --local.-N 1` allows mismatches in the seed sequence. The mapped SAM file was viewed on Integrated Genome Browser (IGB) to identify sequences with multiple reads. Additional information for the sequences were obtained by submitting to human BLAT. Reference genomes for each sequence were constructed by retrieving the original unclipped read

Amplification

DNA was amplified by using DreamTaq Master Mix (Fermentas, Glen Burnie, MD, USA), 2 μM forward primer, 2 μM reverse primer, and DNA from reverse transcription. DNA was initially

denatured at 95 °C for 1.5 min (30 s for subsequent denaturing steps), annealed at 55 °C for 30 s, and extended at 72 °C for each cycle. Optimum number of PCR cycles was determined for each selection round by comparing 8-, 12-, 16-, and 20-cycle aliquots on agarose gel.

Selective amplification

Specific sequences were amplified from the round 6 ATP binding pool as above using primers from selection with 5-10 sequence specific nucleotides with annealing temperature adjusted accordingly.

RNA Transcription

RNA was transcribed at 37 °C for 1 to 3 h in a volume of 20 µL containing 40 mM Tris chloride, 10% dimethyl sulfoxide (DMSO), 10 mM dithiothreitol (DTT), 2 mM spermidine, 2.5 mM each CTP, GTP, and UTP, 250 µM ATP, 2.25 µCi [α-^{32}P]-ATP (Perkin Elmer, Waltham, MA, USA), 25 mM $MgCl_2$, one unit of T7 RNA polymerase, and 0.2 µM of DNA template. DMSO was used to increase transcript yields, as documented in previous studies (Chen and Zhang, 2005). The transcripts were purified using denaturing PAGE.

ATP-Agarose Binding Assay

Purified RNA transcripts were precipitated, dried, and resuspended in 200 µL binding buffer containing 140 mM KCl, 10 mM NaCl, 10 mM Tris chloride, pH 7.5, and 5 mM $MgCl_2$ and heated to 70 °C before loading on to C8-linked ATP-agarose beads (Sigma-Aldrich, St. Louis, MO, USA) equilibrated in the binding buffer. Flowthrough was collected after the columns were capped and shaken for 20 min at room temperature. The beads were washed with 200 µL of

binding buffer, and potential aptamers eluted with the same buffer supplemented by 5 mM ATP·Mg with 30 min of shaking at room temperature. Each fraction was analyzed for radioactivity using a liquid scintillation counter.

Co-transcriptional Binding Assay

RNA was transcribed in a Spin-X column (Corning, Corning, NY, USA) for 30 min at 37 °C in 50 µL solution containing 40 mM Tris chloride, 10% DMSO, 10 mM DTT, 5 mM each GTP, UTP, and CTP, 500 µM ATP, 3.75 µCi [α-^{32}P]-ATP (Perkin Elmer), 25 mM $MgCl_2$, 1 unit of T7 RNA polymerase, 50 pmol of DNA template, and C8-linked ATP-agarose beads, washed, and equilibrated in transcription buffer. The columns were centrifuged at 4,000 g for 1 min, and flowthrough was collected. The columns were washed in 50 µL of the binding buffer and centrifuged. Elutions were collected in the ATP·Mg elution buffer following 30 min of shaking. Each fraction was resolved on a denaturing PAGE gel to separate the full-length transcripts from unincorporated [α-^{32}P]-ATP and shorter transcripts. The gels were exposed to phosphorimage plate and scanned, and the amount of full-length RNAs measured with ImageJ software.

3′-Terminus Labeling

RNA ligation reactions were performed essentially as previously described (England et al., 1980). Briefly, RNA transcribed in the absence of [α-^{32}P]-ATP was PAGE purified and ligated at 37 °C for 3 h in a volume of 10 µL, containing RNA ligase buffer (New England Biolabs [NEB] Ipswich, MA, USA), 2 µCi [5′-^{32}P] cytidine 3′, 5′-bisphosphate (Perkin Elmer) and one unit of T4 RNA ligase (NEB) and PAGE purified again.

In-Line Probing

In-line probing reactions were performed as previously described (Regulski and Breaker, 2008; Soukup and Breaker, 1999). The 3′-end labeled RNA was incubated with varying amounts of ligand for 1 or 2 days at 37 °C in a buffer containing 140 mM KCl, 10 mM NaCl, 10 mM potassium phosphate, pH 7.9, or Tris chloride, pH 7.9, 1 mM $MgCl_2$, and 1 mM spermidine. Triphosphorylated ligands (ATP, dATP, GTP) were prepared as 1:1 complexes with Mg2+. AMP was used directly, without additional divalent metal ions. The partially hydrolyzed RNAs were resolved using denaturing PAGE, exposed to phosphorimage screens (Molecular Dynamics/GE Healthcare, Pittsburgh, PA, USA), and scanned by GE Typhoon phosphorimager. The band intensities were analyzed by creating line profiles of each lane using ImageJ.

Mutagenic PCR

The mutagenized pool was made by amplification of a 100-fold dilute sample of round 6 selected pool by using DreamTaq Master Mix (Fermentas, Glen Burnie, MD, USA), 2 μM forward primer, 2 μM reverse primer, and 200 μM 8-oxo-GTP and 200 μM dPTP (Trilink Biotechnologies). DNA was initially denatured at 95 °C for 1.5 min (30 s for subsequent denaturing steps), annealed at 55 °C for 30 s, and extended at 72 °C for each of 10 cycles. These PCR reactions were diluted another 100-fold and reamplified with no nucleotide analogs in 8 cycles.

***In vitro* Selection**

The DNA pool used for the *in vitro* selection was derived from the human genome and described previously (Vu, et al. 2012) (Salehi-Ashtiani, et al. 2006) Purified RNA transcripts were precipitated, dried, and resuspended in 200 μL binding buffer containing 140 mM KCl, 10 mM

NaCl, 10 mM Tris chloride, pH 7.5, and 5 mM $MgCl_2$ and heated to 70 °C before loading on to C8-linked ATP-agarose beads (Sigma-Aldrich, St. Louis, MO, USA) equilibrated in the binding buffer. Flowthrough was collected after the columns were capped and shaken for 20 min at room temperature. The beads were washed with 200 µL of binding buffer, and potential aptamers eluted with the same buffer supplemented by 5 mM ATP·Mg with 30 min of shaking at room temperature. Each fraction was analyzed for radioactivity using a liquid scintillation counter. Elutions were pooled, desalted using YM-3 spin filters (Millipore, Billerica, MA, USA), precipitated, dried, and resuspended in H2O.

Reverse Transcription

RNA was reverse transcribed in 20 µL using the Promega reverse transcription buffer, 2 µM reverse primer, and RNA recovered from the previous selection round. The RNA and primer were annealed by heating at 65 °C and cooling to room temperature before 1 unit of Thermoscript (Invitrogen, Grand Island, NY, USA) and Improm II (Promega, Madison, WI, USA) reverse transcriptases each were added. The reaction was initiated for 5 min at 25 °C, and then the temperature was ramped to 42 °C, 50 °C, 55 °C, and 65 °C for 15 min each before the enzymes were inactivated at 85 °C for 5 min.

Mapping of mutant sequences

More permissive mapping was performed mapping mutants to the reference genome. The parameter `-N 1 --local --score-min G,20, 4` were used. `--score-min G,20, 4` defines the minimum score for mapping by the function:

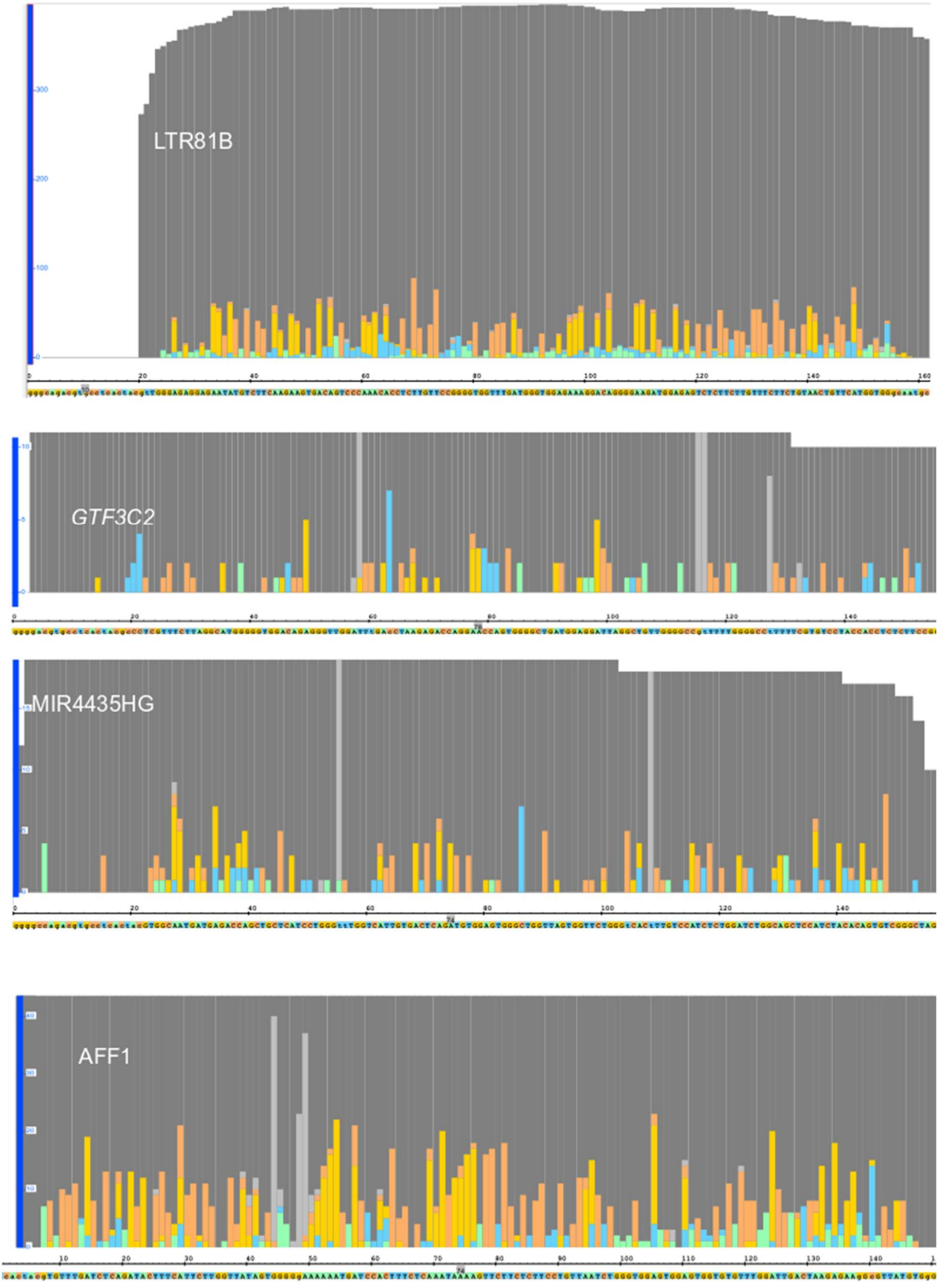
$$min\_score = 20 + 4*\ln(x)$$

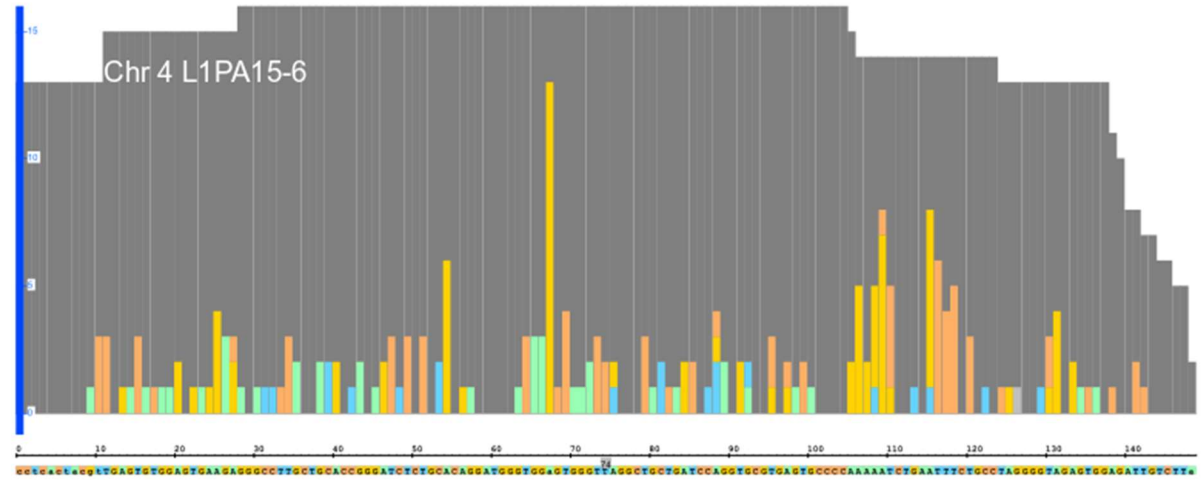
where x = read length. (Default: min_score = 20 + 8*ln(x))

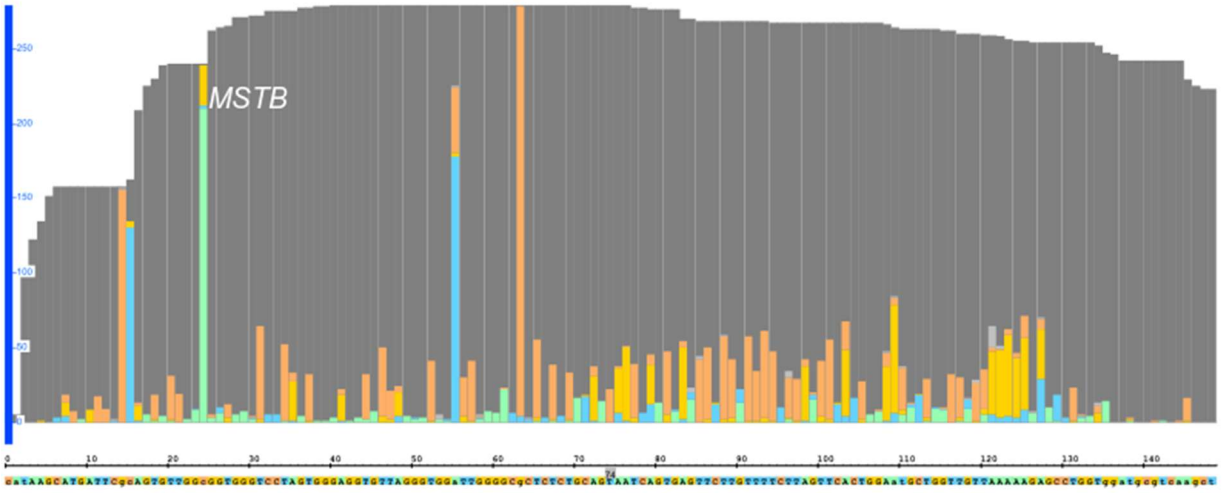
Mismatch pileups

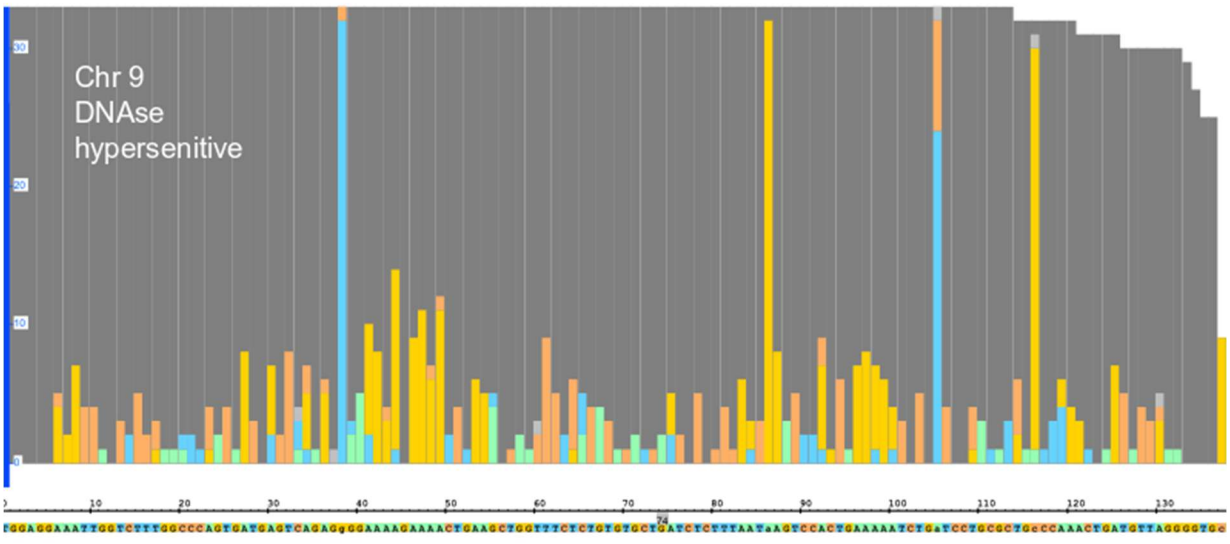
Mapped mutants were viewed on IGB. Mutations were observed using the track operation "mismatch pile up" and graphs were exported as a .svg file.

Supplementary Data




Chr 4 L1PA15-6


MSTB


Chr 9
DNAse
hypersenitive

Chr	strand	reads	Genetic loci	Sequence
2	-	40	LTR81B	acatTGGGAGcGGAGAATATGTCTTCAAGAAGTGACAGTCCCAAACACCTCTTGTTCCGGGGTGGTTTGATGGGTGGAGAAAGGACAGGGGAAGATGGAGAGTCTCTTCTTGTTTCTTCTGTAACTGTTCATGGTGGCAatgcgtagct
2	+	701	*GTF3C2*	ggggacgtgcctcactacgcCCTCGTTTCTTAGGCATGGGGGTGGACAGAGGGTTGGATTtGAcCTAAGAGACCAGGAACCAGTGGGGCTGATGGAGGATTAGGCTGTTGGGGCCgtTTTTGGGGCCtTTTTCGTGTCCTACCACCTCTCTTCCGG
2	+/-	530	MIR4435-2HG	ggggccagacgtgcctcactacGTGGCAATGATGAGACCAGCTGCTCATCCTGGGttTGGTCATTGTGACTCAGATGTGGAGTGGGCTGGTTAGTGGTTCTGGGtCACtTTGTCCATCTCTGGATCTGGCAGCTCCATCTACACAGTGTCGGGCTAG
3	+	4	CSNKA2IP	ntGGACTTTAGAGCCCCTTCATCTATGCCATGGACACTAGAAGCTTGAAAAGCTTCTTGGGGAGGAGGGTGGTTTGTGTCTGTGGAGGATGGTGTGATAAATTTGGATAAGGATAAGCTCAAAATcGTTcCAATAGACCTGATgcgtgca
4	+	127	*AFF1*	ctcactacgTGTTTGATCTCAGATACTTTCATTCTTGGTTATAGTGGGGgAAAAAATGATCCACTTTCTCAAATAAAAGTTCTTCTCTTCCTGTTAATCTGGGTGGAGTGGAGTGGTGTGTTTGGATTGACTAAGAGAAgGcGTTATGTGgG
4	-	28	L1PA15-16	cctcactacgtTGAGTGTGGAGTGAAGAGGGCCTTGCTGCACCGGGATCTCTGCACAGGATGGGTGGaGTGGGTTAGGCTGCTGATCCAGGTGCGTGAGTGCCCCAAAAATCTGAATTTCTGCCTAGGGGTAGAGTGGAGATTGTCTTa
6	-	20	L2A/ MER70B	TTGGGaAAGTAGTAAGAGTGCAGCCCAGGACGCCGTGGGagGAgGTGGAGGCCgGTGCCTGGCAAGCCTGCTTCCAACCTCCTGCCacgaccaCTTCaaGTGCCCCACCCTCCAGGtCcCTTaCTGCCTCCCCAgcaataccccaaacc
7	-	8	L1M2c	TTGACTGCACATGCAGGTCCTTTGGGCAGGTCAGCAGCATGTCCATGGTAGGGGTGTGGTCCCATGCTGCTACTGTGTCTTAGGCCCTGGGCATGGGCACACAGCGCTCTGTGGGCCCAGTGGCATGTCAGCTGTTAGGcggagcaatg
7	+	10	*PTPRN2*	tCGAAGTGAATCAGACGGTGGTGGCTTCTCCCCACCATCCACTCGGCAGGGGACAGGGTCTCCACGGGaAGCTGGTCACCCCACGTGacCGCCGTcCCCTGCCTCCTCTGGCGAGCTTTGTGTTCACCGGGCGGATTCTGGCAGCAACC
8	-	7	*CSMD1*	tTGGATGCTGTGGGATTGCAGGTGCATTCCTGATGCTGTGGATTGCAGGTGCATTCCAGATGCTGTGGGATCaCAGGTGCTTTCCGGATGCTGTGGGATTGTGCATTCTGGATGCTGTGGGATcGCAGGTGCATTCCaGACGCTGTGGG
8	-	4	*SGCZ*	atCACAGCATCAACTGGGATAGCTGGCTTGGAGGATTCACTATCAAGGcGACTCTCTCCAGcGGCTGaTAAGTTGGTGCTAcCGcTTAGCTaAGAGCTCACTTGAGGATGTGGGCTGGGAACCTTATTTCTTCTCCAaTGAGCaattcg
8	-	10	*MSTB*	catAAGCATGATTCgcAGTGTTGGcGGTGGGTCCTAGTGGGAGGTGTTAGGGTGGaTTGGGGCgCTCTCTGCAGTAATCAGTGAGTTCTTGTTTTCTTAGTTCACTGGAatGCTGGTTGTTAAAAAGAGCCTGGTggatgcgtcaagct
9	-	6	*KDM4C*	tCCTGAGCCTTGGGAGGTCGAGACTCCAGTGACCTCTGATCTTGCCATTGCCCTCCAGCTTGcGGGACAGcGTGAGaACCTGTCGCCCtAaCTGGAGaACAGTGGCATGccTTCaGCTAACTtCAACCTTCGCCTCCCAGcatcagcaa
9	-	20	chr9_839.1	TGGAGGAAATTGGTCTTTGGCCCAGTGATGAGTCAGAGgGGAAAAGAAAACTGAAGCTGGTTTCTCTGTGTGCTGATCTCTTTAATaAGTCCACTG

				AAAAATCTGaTCCTGCGCTGcCCAAACTGATGTTAGGGGTGcatgtggagagc
9	+	5	LTR38/ L1PA4	TGATTGGGcGAAAGCAGACTTGATGGGATGACACGCTTCTGAGTTTAGGAGGGGCTTCTGAGCAGAGCCCCAAATGGTCCGGAGCTGTGGCTCAAGTCTGGCTGCTTGCTGCTCAGAGACCAAAGACATGAAGTGAGGTGTGGTagcaa
9	-	9	*BRD3*	agctaggGGGCTGAGGTGGGGTGAACATGGTGTGAAAGCACCACACTCTGCAgCAGGGTGGCTGTCCTGCTTGTGCACCCTCTGGGGCACTCGCAGGCTGGGGTAGGGTGACCCCACCACCGTCCTTCTCTcGTTGaatgcgtcaagcc
10	-	16		gtgcctcactacgTGGTACATCCTGGTGGTATGTTCCGGGCTCTGGGCCCAGCTGGCTCGTAGTTAGGAGTGGCGGTGAgGGTGGAATGcGCTTTTGAGCCGTCCTGAGCCCTGAGAAGAGGTAGAGAGAGGCTGTTGTTGTACCTCCAG
11	+	14		ntTGGAGCCAGGAGCAGCTTGGTGGGGTGGGAGCTGATGTTAGGGTCAGCTCACAAGCCCTACAGTGTGTTGGGACCCTTGCTTTTGGTGGAAGCATCAGtATGTGCCAGGCACCCACTGCTCCTCCAGCCGTACAatgcgtcaagctcg
11	-	26	*KIRREL3*	gcctcactacgtGGTGTGTGCATGTGTCAATGTGCATGTGTGCACGTGAGTGCATGTGTCCATGTGCATGTGTACATGTGTGAGCATGTGCATGTGTGCATGCATGTGCATGTGTGTGCGTGTGTGCATATGTGTCCATGTGCATGTGa
12	-	4	AluJr	taatCCAGGTATTATAGCAAGcGCCTGTGGTCCCAGCTACGTGGGAGGCTGAGATGGGAGGATAGCTTGaGCCCAGAAGGTAAAGGCTACgGTGAGCCATGATAATTTCACTGCcCTCCAACCTGGGCAACAGAGCAAGcgtcgctcac
12	+	6	(GT)n	actacGTGTGTATGTtTGTGTGTGCGTGCGTATGCATGCATGCATATGTGTGTGGGATGTGTGTGCATGTAAGCATGTGCATATATGTGCACGTGTGGGGTGTGTGTTCATATGTGAATTCATATGTGAACATGTGTGAATGCATATGaG
14	+	27	L1P3	gcGTGCTAGCAGCTAGAATCTCATGCCAGTGGATCTTAGCTTGCTGGGCTCCGTGGGGGTGGGACCCACTGAGCCAGACCACTTGGCTCCCTGGCTTCAGCCCCCTTTCCAGGGGAGTGAATGGTTCTGTCTTGCTGGCATACCAGGTAg
14	+	12	SATR1	agacgtgcctcactacgtTGGGAGAGGAGAATATGTCTTCAAGAAGTGACAGTcCCCAAACACCTCTTGTTCCGGGGTGGtTTTGATGGGTGGAGAAAGGACAGGGGAAGATGGAGAGTCTCTTCTTGTTTCTTCTGTAACTGTTCATGGTGag
14	+	34	L1PA16	ctcactacgtGGCTTAGGGCA**GGAAGAGAACTG**CTAAGGCAGTTTCTCCTAGAAGATGAGACCTGCAGCCAGGTCCAGCTTGGTGACCTAGAACTGGTCTGCATGTGTCATTGCTGG**G**TGCTCCACCCTGCTCCCCTGAGATCATGTTGag
15	+	9	L1PA7	ATCTTGTGGGcGCCCCTCTGGGACAAAAATTCCAGAGGAAGGAACAAGCAGCAATCTTTGCCGTTCTGCAGTCTCCGCTGGTGATACCCAGGAAAACAGGGTCTGGAGTGGACCTCCAGCAAACTCCAACAGACCTGCAGCAGAGGGGGag
15	+	212	EST: BU853031	gggcctcactacgtTGACTGGGTTATATGCCAGGTATTGTCCTGGTTAAGTGGCCATTGCTCAAGCCTGgCTACTGCCTATAGGGTGAGTTGCcGTGAGATTTTtAATTCACGAGaAGGGAGGTTGTGGGTAGGAGTGATGCAGTTCCCGCGg
17	+	58	AC018521.4	gcctcactacgtCAAAAACaAAAAAAAACAGTGCCTGGCACAGTAAGGCTCAGACcAGCTGTTGTGAGAGGGCGGTGGGGTGGAAATACCCTCAGCTGA

				GGTTAGAAAGTTGTGTTCTAGGCCGGGCGTGGTGGGTCACGCCTGTAATCCCAG
17	+	148	RPTOR	gtgcctcactacgtTGCGTACTGTGTGCGTGTGCGTACTGTGTGTGTGCATACTGTGTGTGCGTACTGTGCaTGTGcGTACTGTGTGCATGTGCATACTGTGTGTGCATACTGTGTGTGTGCATACTGTGTGcGTGTGCgTACCGTGTGTGg
18	+	337	PTPRM	ggcagacgtgcctcactacgcAGCTCCATTTGTCTGCTCTGAGCCCAAGGAGGCTGCCACCCAGTTGGAGGCCATGGGGTGAGGGCAAGGGGAAGGCTTCCAGGAGGTGGGAGCACCCAAGAGGGGCAAGAGAAGGTGAGTAGTGATGCTGGAg
18	-	461	AC090125.1 THE1B	ggggcagacgtgcctcactacGTT**G**TGGGAGGGACCCGGTG**GGAGGTAACTG**AATCgTGGGATGAGTCTTTCCCGTACTGTcGTCATGATAGTGAATAAGTCTCATGAGgTCTGATGGtTTTTATATGGGGGAGcTTTCCCTTCgCAAACTCTag
18	+	4	*ATP9B*	nTGGTTAGCGTGCTCTCcgTGGTTgGCGTGCTCTCCGTGGTTAGCGTGCTCTCCGTGGTTAGCGTGCTCTCCGTGGTaAGCGTGCTCTCTCTGGTTAGCGTGCTCTCCGTGGTTAGCGTGCTCTCTGTGGTTAGCGaGcaatgcgtcaac
18	+	4	*CTPD1*	GtGGAGGACGGTGCAGGGACCTCGTGGTCTCTACTCCCaGTTTCTCTGCGGTGCcCCTTACaGGGTGTGAGGTCCTtGTGAGGAGGACGGTGCAGGGACCTCGTGGTCTCTAgTCCCgGTTTCTCcGCGGTGCTCCTTACGGGGTGTGAG
18	-	4	*PARD6G-AS1*	gCAGGAGGGTGGAGACCCCCACCCAGAGTGACTTCCCAAAAGGCAGCACTACCCGAtGAGGGACGCCACAGGAGGAGGACTaGGAGGGAGGGCAGTGGGCACGGTAGAGTCTTCTTTCCTGAGGCGACCCCAGCTTGTTGGACGAGGCgg
19	-	8	*LRG1*	atGGGAGCCGGGGGAGGCTTCTTGTTAGgGGCCTGGGGCCACAGTGACCAAGGCTGTCGAGATCATGTCTGATGGCTGCCGGGCCAGGGAAGCCCTGGAGGAGGAGGGGAAGCCCCAGGCTTCCAAGGGTTGTGGTGCaatgcgtcaag
19	+	111	*MRI1*	gGGAATCCGCgTCCTGGGAACCCGTGGAATCCGCgTCCTGGGAACCCGTGGAATCCGCgTCCTGGGAACCCGTGGAATCCGCgTCCTGGGAACCCGTGGAATCGGGTTGGATGCGCATGTGCGTGag
20	-	237	EST: EB385442	cctcactacgTGCTGTTTCATATCTTGTAtTTTTTCATGTGGGAAAAAAAAAAGATCTTGGAGCTCATTCCATAAGTGACTTGCATTTCAACAGGATTCTTCAAAAGTGACAAGAAAACATAGGAAACCAACATCATTTTGAATCGCTTTCTCAg
20	-	11	MLT1G3 / L2A repeat	TCACGTTCTCTTGGAGGGGGGTGTGGGTAGACAATAGAGAAGCAAGTAAATACACATTATGTTAGACTCTATAGAGAAAAATTAAATGAGGaGTTCTGGTTACTATGACCGTGCAACTAGATGGGCCCAAACTTATGGATTCTGTGGca
20	+	11	*VSTM2L*	tacgtAGGGGCGTGGTGCGACTTGCCTGAGGACACACAGATGCTGAACAGCAAAGCCGGGCTTGAGCCCAGCTTGTGGGACGACAGAGCCAGAGTTCGTAGCCTGCGCATCTGTCTGCTCCTGGCTGCCAAGCTGTGTGTGTGGGCAGG
21	+	9	AP000477.2, HERVL18-int	acgtgcctcactacgTGGTTGTCAGGGATGAATGTACAACACTGTATGCCTACAAGGGAACATAGTAACATCTTGGGCAGCTGTTACTATGTCTAAGGCCATCCAGTTTTGCAGCACCACCTTCCTGATCTGATCAACCTCATCGGTTAn
22	+	1003	*PRR5*	gggtGGAGGAGCCTGGATGCTGCCTGCAGGACCTCAGGCTGT**G**CCTGCTGGGCAAAGgCCCTGGGCAG**GGAAGGAACTG**CAGCCTCCACAGAGG

				GTGGATaTGGTGGAGAGGTGGGAGGCCAGCTCCTGTCATCCGAGGTC CAGGCAAGCCAG

Table S3-1. New Sequences from the ATP aptamer selection. Lower case letters denote nucleotides not matching the reference genome. Potential ATP binding motifs are shown in bold and flanking stems underlined. Genes, EST, retrotransposons, and predicted genes found in that genomic loci are reported where applicable

Chapter 4 ATP Apta-seq: Multiplex Aptamer Discovery through Apta-Seq and Its Application to ATP Aptamers Derived from Human-Genomic SELEX.

Introduction

Functional RNAs play central roles in regulating gene expression and catalyzing essential cellular reactions (Sharp 2009) RNAs evolved in the laboratory serve equally diverse functions, including as diagnostic and therapeutic aptamers that bind a wide variety of targets (Jijakli, et al. 2016). The vast majority of aptamers have been identified using *in vitro* selection (or SELEX), a molecular evolution technique based on selecting target-binding RNAs from highly diverse pools through serial rounds of enrichment and amplification. (Ellington and Szostak 1990) (Tuerk and Gold 1990) The RNA pools are transcribed from either synthetic (typically random) or genomic DNAs (Pobanz and Lupták 2016), and selections often yield multiple distinct motifs of highly variable abundance, target-binding affinities, and specificities.

The discovery of new aptamers is often hampered by the difficulty of identifying and characterizing the structural motifs that result from the selection process, because testing of individual sequences identified in selected pools tends to be a tedious process. Moreover, low copy number sequences, which may play key functional roles, can go undetected when only the dominant sequences are identified and tested individually. High-throughput sequencing can be applied to measure sequence diversity of selected pools and identify potential aptamers, (Alam, Chang and Burke 2015, Hoinka and Przytycka 2016), but their structural and binding characteristics have to be established individually for each sequence, often making this the limiting step in the discovery of novel aptamers. Indeed, many functional aptamers may go uncharacterized due to the challenge associated with testing the structure and affinities of the majority of sequences within a selected pool. This handicaps the description of selected RNAs

and is a key hurdle to overcome for the efficient discovery of functional aptamers that may have diverse and important functions.

One application of *in vitro* selections seeking to characterize all enriched RNAs, including low copy number sequences, is genomic (Singer, et al. 1997, Lorenz, von Pelchrzim and Schroeder 2006) and transcriptomic SELEX. (Fujimoto, Nakamura and Ohuchi 2012, Terasaka, et al. 2016, Chen, et al. 2003, Dobbelstein and Shenk 1995) These experiments reveal RNAs encoded by the genomes of predetermined species and are particularly important in the discovery of new instances of known functional RNAs, such as aptamers and ribozymes (Vu, et al. 2012) (Curtis and Liu 2013)(Salehi-Ashtiani, et al. 2006). In contrast to selections based on synthetic sequences, which are typically designed to identify a single (or a few) fittest functional RNA, genomic and transcriptomic selections are designed to map out all instances of a given function, such as binding of a target metabolite or protein. However, the relative abundance of *in vitro* selected RNAs is biased toward not only sequences that fulfill the selection criteria but also those that are highly amplifiable (sequences that transcribe, reverse-transcribe, and PCR amplify more efficiently than others (Pobanz and Lupták 2016, Takahashi, et al. 2016), thus the resulting distribution of the functional RNAs in the selected pools may be strongly skewed by these two properties.

The ability to more quickly characterize selection pools for the binding function would be enabled if the structures of many aptamers and their physical interactions with their targets could be determined in a single experiment. RNA probing with chemical reagents is a robust method for analyzing the structures of RNAs and characterizing important conformational changes that can occur due to interaction with small molecules and proteins; however, such experiments are usually accomplished on single RNA molecules (Soukup and Breaker 1999, Merino, et al. 2005,

Ding, et al. 2015). Information about structure and ligand-binding sites in RNAs can be extracted from experiments based on partial hydrolysis (in-line probing) and chemical modification of RNAs, using selective 2′-hydroxyl acylation (SHAPE) or base modification (e.g., by dimethyl sulfate)((Soukup and Breaker 1999, Merino, et al. 2005, Ding, et al. 2015). The SHAPE method detects 2′-OH accessibility and reactivity to acylation, thereby reporting on which aptamer segments are more flexible and reactive in response to a ligand (Wang, Wilkinson and Weeks 2008). Changes in structure can be probed under different environmental conditions or in the presence of varying concentrations of ligands, such as small molecules (Stoddard, et al. 2010) and proteins, (Fu, et al. 2014) and can be used to extract the dissociation constants for the RNA–ligand interactions (Wakeman and Winkler 2009). Several methods combine SHAPE with high-throughput sequencing (SHAPE-seq) to achieve single-nucleotide resolution of acylation reactivity on a diverse set of sequences simultaneously and couple the output to computational modules developed to yield genomic locations, intrinsic reverse transcriptase (RT) stops, SHAPE reactivities, and secondary structure models for each transcript (Lucks, et al. 2011, Loughrey, et al. 2014, Tang, et al. 2015). Although useful, most of these efforts have been largely descriptive with few examples of their use for novel biological discovery.

To overcome the limitations of single-sequence characterization of *in vitro* selected pools, we developed a multiplexed approach to couple RNA selection with structural and binding characterization of individual sequences within the selected pools. We marry selection with chemical probing of RNA structure to reveal the sequence, structural features, and ligand affinities of both dominant and minor species from the same pool. We use this technique, Apta-Seq, to discover and characterize both known and novel adenosine aptamers in the human genome. Our methodology not only increases the rate of novel aptamers discovery for our

studied ligand (ATP) but also has the potential to be applied to any ligand-pool pair, thereby greatly enhancing the speed of aptamer discovery and structural characterization.

Apta-seq

In order to establish the identity, secondary structure, and binding properties of aptamers in a single experiment, we combined *in vitro* selection, SHAPE-seq, and StructureFold into a pipeline (Figure 4-1), which provides all the information required for aptamer discovery. We applied this method to an *in vitro* selected pool derived from the human genome and enriched for ATP-binding aptamers, as described previously (Vu, et al. 2012). We chose adenosine/ATP as the target of the selection, because it has been used extensively over the past two decades, and as such, it is an excellent model system for the development of new *in vitro* selection and analysis methods. Adenosine aptamers with a conserved motif consisting of an 11-nucleotide binding loop and an opposing bulging guanosine, flanked by two helices (the Sassanfar–Szostak motif), were initially isolated from synthetic, random pools by *in vitro* selections targeting ATP (Sassanfar and Szostak 1993), nicotinamide adenine dinucleotide (Burgstaller and Famulok 1994), S-adenosyl methionine (Burke and Gold 1997), and S-adenosyl homocysteine (Gebhardt, et al. 2000). More recently, a genomic SELEX experiment revealed the same motif in two distinct loci in the human genome (Vu, et al. 2012): the *FGD3* aptamer resides in an intron of the *FGD3* gene, and the ERV1 aptamer maps antisense to a junction between an ERV1 LTR repeat and its 3′ insertion site. These were the only two aptamers revealed by traditional cloning; however, this approach tests only a small number of sequences and can miss low copy number aptamers, we therefore reanalyzed the pool using high-throughput sequencing and SHAPE to uncover other genomic aptamers.

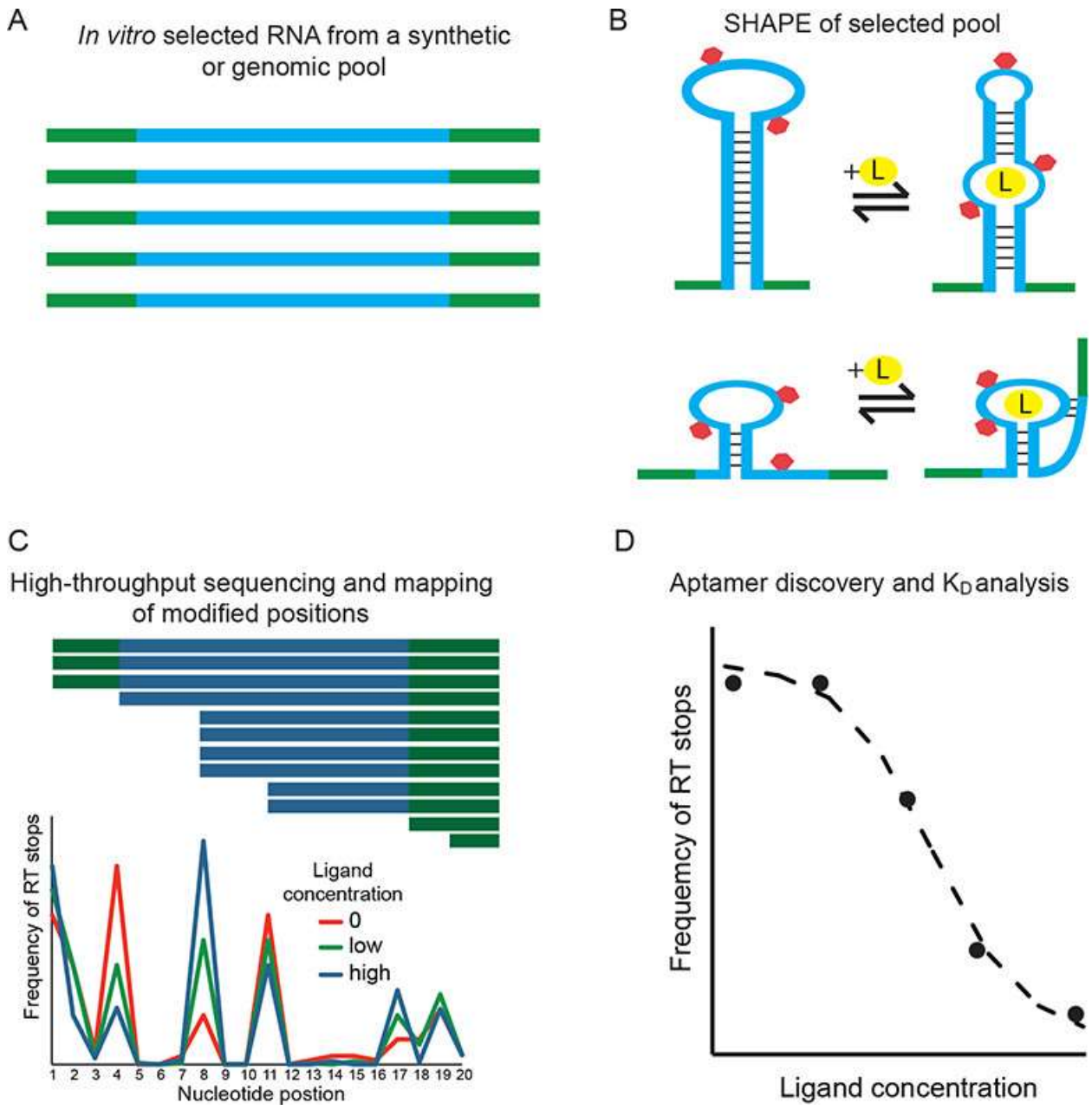


Figure 4-1 Apta-Seq scheme. The pipeline consists of (A) an *in vitro* selection, (B) SHAPE analysis of selected pool, and (C) high-throughput sequencing for determination of sequence identity, secondary structure, and (D) binding isotherms of individual aptamers.

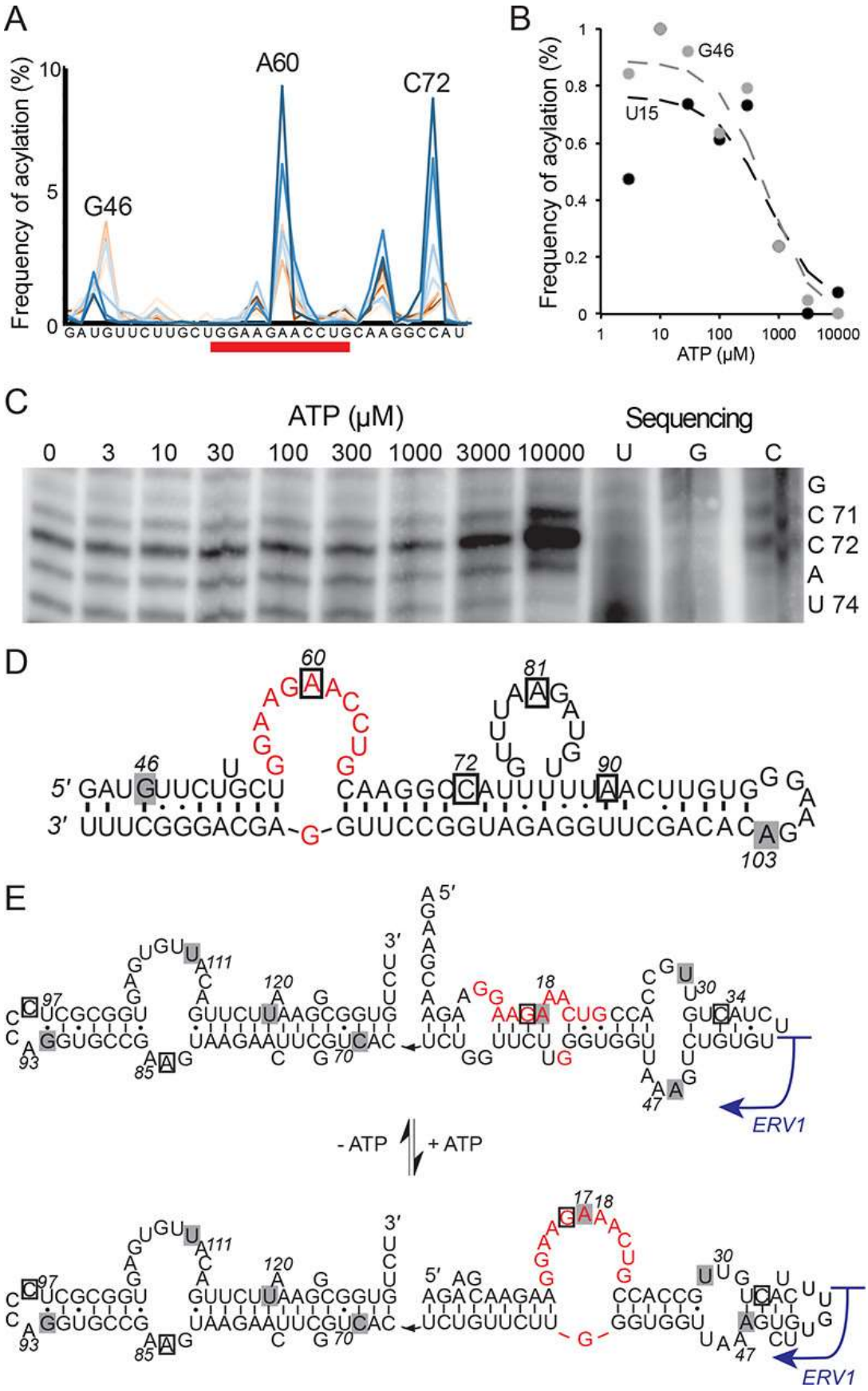
To test these potential aptamers, we adapted the SHAPE-Seq analysis coupled to StructureFold to the selected human genomic RNA pool at varying ATP concentrations. The pool was transcribed and purified as during the SELEX experiment, divided into 10 fractions,

and incubated with ATP at concentrations ranging from 3 μM to 10 mM, with two fractions set aside for control reactions (no-SHAPE and no-ligand SHAPE controls). The samples were acylated by 2-(azidomethyl)nicotinic acid acyl imidazole (NAI-N3) (Spitale, et al. 2015), reverse transcribed, and ligated to yield circular single-stranded DNAs, in which the ligation positions correspond to the reverse transcriptase termination points. These occur either due to acylation of the RNA template on the nucleotide immediately upstream of the termination point or at boundaries of RNA structural elements, such as paired segments. The no-SHAPE control reaction reveals the natural RT stops for each sequence in the pool, whereas the SHAPE experiments reveal additional RT stops corresponding to the acylated positions in the RNA, some of which change upon ligand binding (Figure 4-1B). Amplification of the ssDNA by PCR and barcoding each experimental condition in the ligand titration allow one-pot sequencing of the entire population. Subsequent barcode-based demultiplexing of the sequences reveals all sequences and SHAPE-dependent RT stops for each experimental condition. Nucleotide positions of all detected sequences were analyzed for RT termination and expressed as RT-stop frequency, which is defined as a fraction of total reads (per experiment) for a given sequence (Figure 4-1C).

The frequency of RT stops was used to obtain SHAPE profiles in the presence of increasing concentration of ATP-Mg (Figure 4-1C). The RT stops were resolved with single-nucleotide resolution, and the ATP dependence of SHAPE profiles allowed us to determine the apparent K_D for each aptamer at several positions. For the *FGD3* aptamer (Vu, et al. 2012), two positions (G46 and U15, numbered by position in the pool sequence and shown in Figures 4-2D and S1A) yielded a $K_D \sim 700$ μM (Figure 4-2B, Table 4-1). The Apta-Seq data show that the adenosine-binding loop of the Sassanfar–Szostak motif becomes more reactive toward the

SHAPE reagent at high concentrations of ATP at the third adenosine (A60, Figure 4-2D) of the binding loop (seen in the ATP-dependent increase of the A60 peak in Figure 4-2A). Previous in-line probing data for the *FGD3* and other ATP-binding aptamers have also demonstrated that the adenosine-binding loop becomes more susceptible to in-line attack at the third adenosine (equivalent to A60) (Vu, et al. 2012, Soukup and Breaker 1999). In the solution structures of the *in vitro* selected aptamers bound to AMP, the nucleotide equivalent to A60 makes direct contacts with the ligand through stacking (Dieckmann, et al. 1996, Jiang, et al. 1996), and the 2′-OH of the A60 equivalent appears partially solvent-exposed and hydrogen-bonded to the adjacent phosphate, which likely activates it for acylation (McGinnis, et al. 2012). The sugar–phosphate backbone of this nucleotide in the adenosine-bound conformation is thus highly sensitive to modification, and in the case of the *FGD3* aptamer, the binding loop must undergo a significant conformational change upon ligand binding, because the same position is weakly acylated in the absence of the ligand (Figure 4-2A, low-ATP traces, shown in shades of red)

Figure 4-2 Apta-Seq analysis of the human *FGD3* and ERV1 adenosine aptamers. (A) Graph of acylation positions at the Sassanfar–Szostak motif of the *FGD3* aptamer and varying concentrations of ATP (dark red through dark blue, for ATP concentrations ranging from 3 μM to 10 mM). The adenosine-binding loop is underlined in red. (B) Binding isotherms of the aptamer extracted from positions U15 (black circles) and G46 (gray circles). Both positions reveal a $K_D \sim 700$ μM. (C) PAGE analysis of a purified *FGD3* aptamer clone showing that RT stops for position C72 strongly increase at high concentrations of ATP, validating the SHAPE-Seq data shown in A. (E) Secondary structure of the *FGD3* aptamer with ATP binding loop in red. Boxed nucleotides indicate aptamer positions that show an increase (black outlines) or decrease (gray) in acylation with increasing ATP. (E) Predicted secondary structure change accompanying ATP binding by the ERV1 aptamer. Positions with ATP-dependent acylation changes are indicated as in D. The boundary of the ERV1 retrotransposon is indicated with a blue arrow


A
Frequency of acylation (%)
10
5
0
G46
A60
C72
GAUGUUCUUGCUGGAAGAACCUGCAAGGCCAU
B
Frequency of acylation (%)
1
0.8
0.6
0.4
0.2
0
G46
U15
1
10
100
1000
10000
ATP (μM)
C
ATP (μM)
0 3 10 30 100 300 1000 3000 10000
Sequencing
U G C
G
C 71
C 72
A
U 74
D
60
81
46
72
90
5′ GAUGUUCUGCU CAAGGCCAUUUUUAACUUGUG G G A
3′ UUUCGGGACGA G GUUCCGGUAGAGGUUCGACAC A G A
103
E
5′
3′
97
111
120
18
30
34
93
85
70
47
ERV1
- ATP
+ ATP
17
18
5′ AG
ERV1

To validate the results obtained from Apta-Seq of the pool with the reactivity of the purified *FGD3* aptamer, we performed a SHAPE analysis on an isolated clone of the aptamer. One of the positions with the most prominent ATP-dependent changes in SHAPE reactivity is C72 (Figures 4-2A and S4-1A), which maps to a domain adjacent to the adenosine binding loop (Figure 4-2D). We confirmed this result by conventional (PAGE-based) SHAPE analysis of the aptamer clone, which also revealed a strong, ATP-dependent increase in reverse transcription termination at position C72 (Figure 4-2C), reaffirming that the method reveals comparable data for the same aptamer within a highly heterogeneous pool of sequences. These results demonstrate that our high-throughput sequencing approach parallels more traditional structural analysis normally reserved to single clone analyses.

For the ERV1 aptamer, we derived the SHAPE reactivity using the Reactivity Calculation module of StructureFold, with RT stops from the no-ligand SHAPE data set as the negative control for the reactivity profile of the 10 mM ATP experiment (Figure S4-1B), and carried out structure predictions using the RNAprobing server of the Vienna RNA Package (Washietl, Will, et al. 2012). In the absence of ATP, the ERV1 structure is not predicted to form the ATP-binding motif, because the binding loop is sequestered within a stem flanked by two bulges. However, in the presence of ATP, the structure rearranges to form the classical ATP binding loop (Figure 4-2E). The binding loop showed similar acylation trends to those of the *FGD3* aptamer, but shifted by one nucleotide: the third guanosine (G16; Figure 4-2E) of the ATP binding loop becomes more accessible with increasing concentration of ATP, analogous to the change of A60 in the *FGD3* aptamer (Figure 4-2D). Altogether, the data are consistent with previous findings and structure predictions for the *FGD3* and ERV1 aptamers (Vu, et al. 2012), and the K_Ds we extracted from the Apta-Seq data (Table 4-1) are within 2-fold of the K_Ds

determined previously by in-line probing of individual aptamers. The *FGD3* and ERV1 aptamers are the dominant aptamers in the selected pool (Table 4-1), with a relative abundance of 0.38% and 0.49%, respectively, to all sequences. The sequences represent only a small fraction of all reads obtained from the Apta-Seq experiment, because the high-throughput sequencing output is dominated by sequences corresponding to short primer extensions of the reverse primers and are thus difficult to map to the genome uniquely.

The Apta-Seq pipeline also yielded novel, less abundant sequences not found in previous analyses. Our previously reported aptamers were discovered in an ATP column binding analysis of individual clones to find potential candidates to undergo structural analysis; Apta-Seq condenses the process to gain a large-scale analysis of the pool in one experiment in solution. Surprisingly, three new sequences also contained the Sassanfar–Szostak motif. The *PRR5* aptamer maps to the second intron of the *PRR5* gene or the first intron of the *PRR5*-ARHGAP8 fusion protein (Figure 4-3A). *PRR5* codes for a protein that is part of the mTORC2 complex (Woo, et al. 2007) and, like the *FGD3*, plays important roles in pathways that regulate cell growth, but the significance of the ATP/adenosine-binding aptamers in their introns remains unknown. The *PRR5* aptamer bound to ATP-agarose beads and eluted in the presence of free ATP (Figure 4-3C) and Apta-Seq data revealed a K_D of ~1.7 mM (Figure 4-3B and D, Table 4-1) at a relative abundance of 0.17% among the mapped sequences. Interestingly, mutations (including a 14-nt insertion) of the aptamer sequence throughout the genomes of primates preserve the aptamer structure, suggesting that it may be a functional RNA in primates (Figure 4-3E and 4-F).

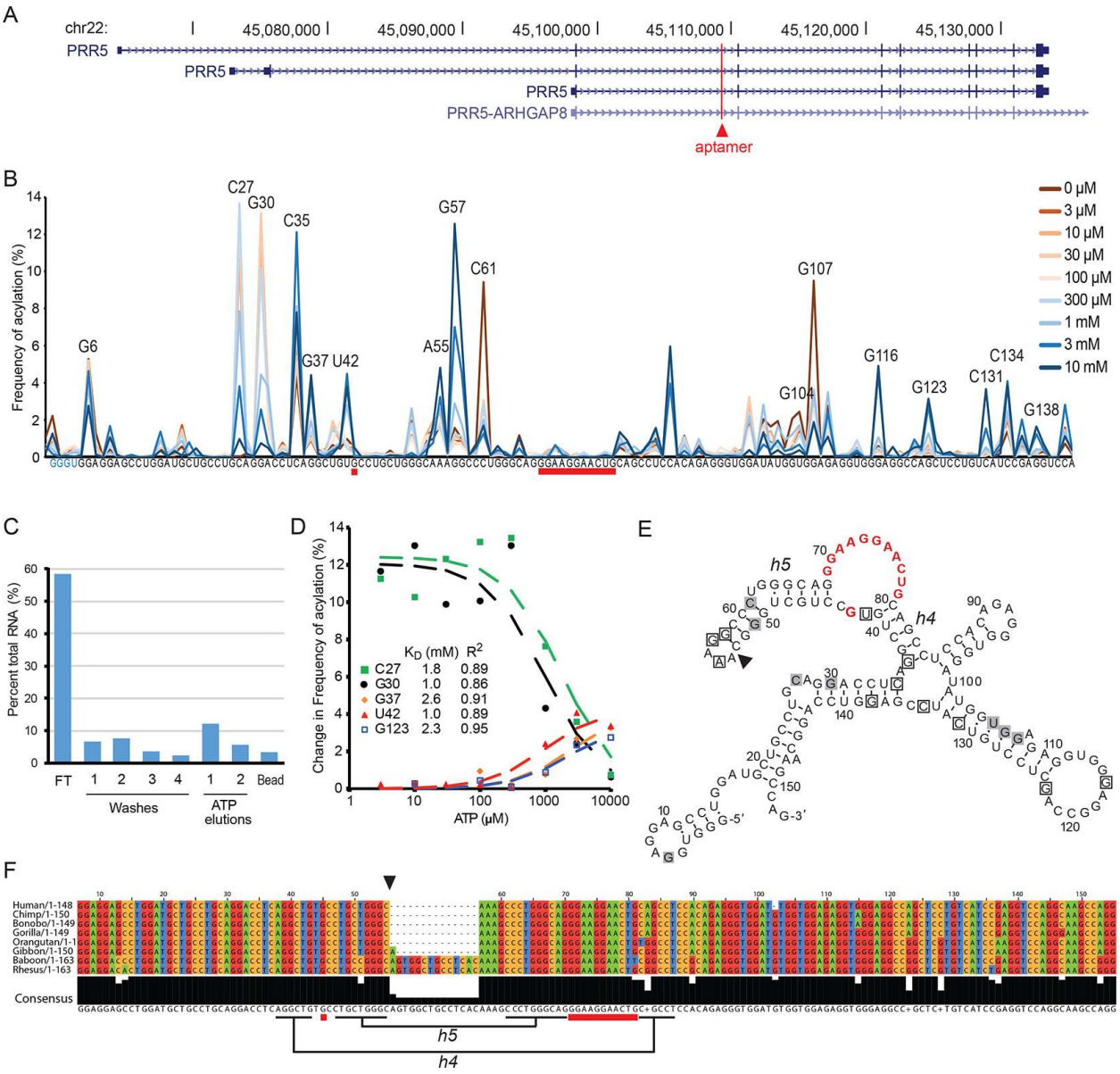


Figure 4-3 A novel adenosine aptamer, mapping to the *PRR5* gene in primates. (**A**) Location of the aptamer (red line) in the *PRR5* intron. The aptamer also appears in the first intron of a *PRR5–ARHGAP8* fusion. (**B**) Trace of acylation for the *PRR5* aptamer with varying ATP concentrations, ranging from 0 (dark red) to 10 mM (dark blue). The conserved adenosine-binding motif is underlined in red, and positions with ATP-dependent changes in RT stops are indicated. (C) Graph of column binding fractions of the *PRR5* sequence amplified out of the selected pool. The graph displays 18% of RNA eluting with free 5 mM ATP. (**D**) Binding isotherms of the aptamer modeled from the change in RT stops at positions indicated in the legend, yielding an average KD of 1.7 ± 0.6 mM (average deviation). (**E**) Predicted secondary structure of the *PRR5* aptamer with ATP binding loop shown in red. Boxed nucleotides indicate aptamer positions that show an increase (black outlines) or decrease (gray) in acylation with increasing ATP. Black triangle indicates sequence-insertion site for some primates. (**F**) Aptamer sequence conservation among primates. All mutations within the aptamer motif are conservative with respect to base-pairing interactions of the proposed structure: U49C mutation leads to a G•U wobble pair to be replaced with a canonical G–C pair in helix 5 (h5). Insertion at A54 (black triangle) extends the sequence into the loop of h5 and potentially extends the helical domain by four base-pairs, and the A81G mutation in h4 creates a G•U wobble pair from an A–U canonical base-pair.

The second aptamer (Figures 4-4A and 4-S2) was found in a long interspersed element (LINE), L1PA16, and the third (Figures 4-4B and 4-S3) maps antisense to a repeat element THE1B, which is derived from a subfamily of ERV Mammalian apparent LTR-retrotransposons (ERV-MaLR), and may be stabilized by fortuitous base-pairing with a part of the sequence derived from the forward primer of the pool (Figure 4-4B). The KDs derived from Apta-Seq (Figure 4-4C and D) were 1.3 mM and 1.1 mM (Table 4-1); however, the SHAPE profile of the THE1B aptamer afforded only a single peak (G32; Figures 4-4D and S4-3A) that yielded a KD model with a good fit (R2 ~ 0.95), whereas the L1PA16 aptamer exhibited several peaks from which the dissociation constants could be derived (Figures 4C and S2A). When transcribed from individual DNA templates and purified using PAGE, both aptamers bound to ATP-agarose beads and eluted in the presence of free ATP (Figures 4-4C and 4-3D). Mutations of key residues (Figure4- 4A and B) in the binding loops of these aptamers abolish binding to ATP columns (Figures 4-S2C and 4-S3C). These aptamers, together with the previously discovered adenosine (Vu, et al. 2012) and GTP (Curtis and Liu 2013) aptamers, indicate that ligand-binding RNAs are likely common in higher eukaryotes.

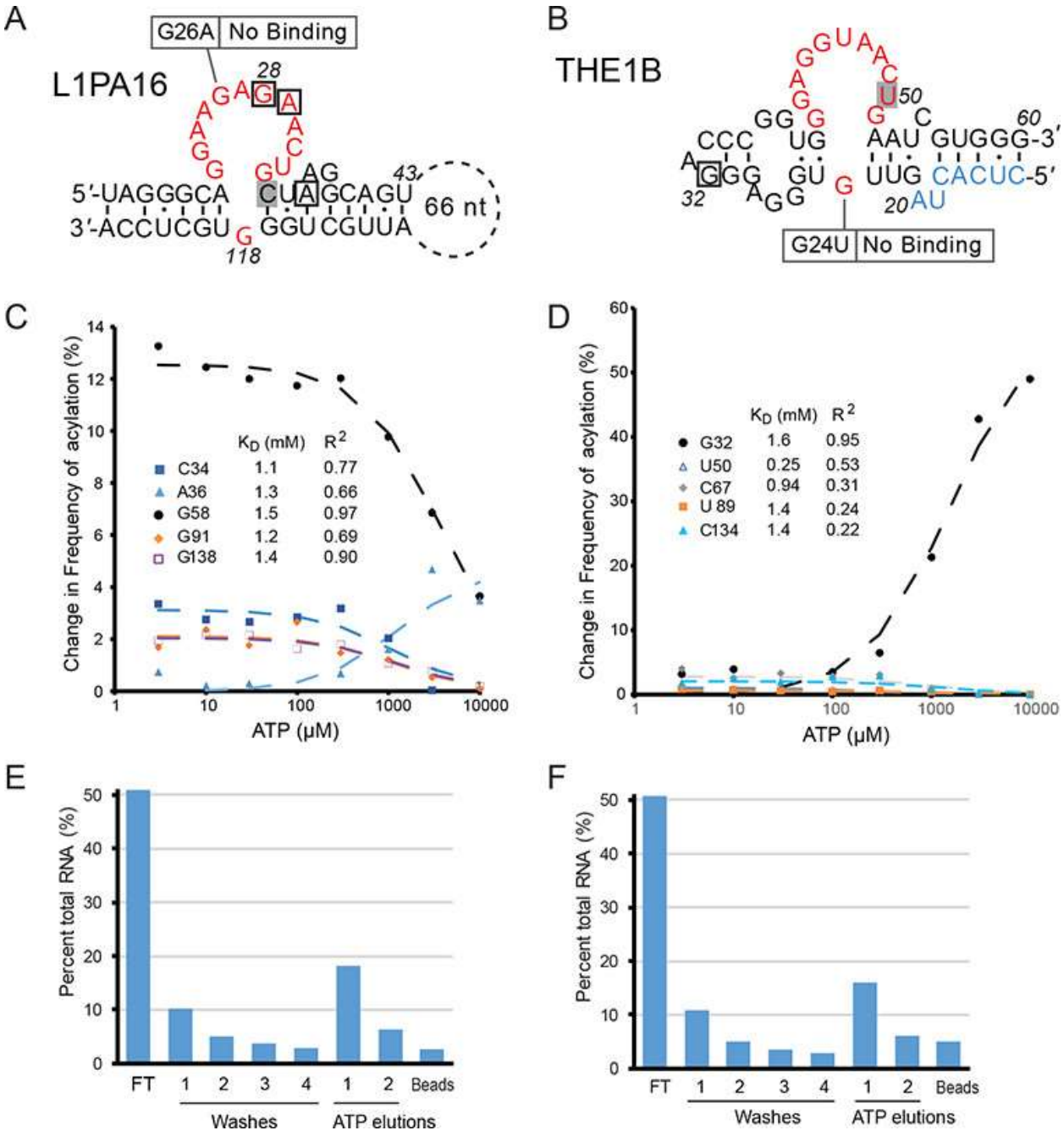


Figure 4-4 Novel adenosine aptamers revealed by Apta-Seq in human retrotransposons. (A) The Sassanfar–Szostak motif mapping to the L1PA16 LINE1 element, with ATP-dependent acylation positions indicated by squares and annotated as in Figures 4-2 and 4-3. Mutation of a key binding site residue (G26A) results in no ATP binding. (B) An aptamer mapping to a THE1B retrotransposon. Part of the 5′ sequence that originated from the forward primer-binding region of the DNA pool is shown in light blue. A mutation of an essential binding site residue (G24U) abolished ATP binding. Binding isotherms of the L1PA16 (C) and THE1B (D) aptamers modeled from the change in acylation at positions indicated in the legend, yielding a KD of 1.3 ± 0.1 mM and 1.1 ± 0.4 mM, respectively. Graph of ATP column binding of the L1PA16 (E) and THE1B (F) sequences amplified out of the selected pool, showing robust activity of the purified sequences and validating the Apta-Seq data. Full SHAPE-Seq profiles and secondary structure models of these aptamers are presented in Supporting Information Figures S4-2 and S4-3.

To analyze these Sassanfar–Szostak motifs for potential binding in the context of the human transcriptome, we isolated total RNA from four cell culture experiments (SHSY-5Y cells incubated with DMSO or DMSO + 10 mM adenosine and OV90 and MCF7 cells) and treated it in the same manner as during the SELEX experiment. The total RNA samples were annealed in the binding buffer, introduced to the ATP beads, washed with five column volumes of the binding buffer, and eluted with 5 mM ATP-Mg. RNAs isolated from the last wash fraction and ATP elution were reverse-transcribed using primers specific for the five aptamers described above and analyzed using nested primers by qPCR. Two of the aptamers, *FGD3* and THE1B, were detected at higher levels in the elution fractions than in the last washes in the SHSY-5Y cells. *FGD3* aptamer exhibited more robust binding to ATP beads and the levels of the aptamer appear insensitive to the presence of adenosine in the tissue culture medium (Figure S4-1B). In contrast, the THE1B aptamer showed significantly higher levels in the ATP elution fraction only in RNA extracted from SHSY-5Y cells incubated with 10 mM adenosine (Figure S4-3D and E). These experiments suggest that at least two of the aptamers described herein have the capacity to bind ATP within the context of their endogenous transcripts and the human transcriptome, and that their expression or activity can be modulated by exogenous adenosine.

In summary, we describe an efficient process of multiplex analysis of *in vitro* selected RNA pools. An *in vitro* selection experiment is combined with SHAPE-Seq and StructureFold analysis to efficiently and quantitatively analyze the selected pools at a single-nucleotide resolution. A number of high-throughput methods, many of which can be adapted for application in aptamer discovery, have recently been developed to analyze RNA–protein interactions (Buenrostro, et al. 2014, Tome, et al. 2014, Lambert, et al. 2014); however, all of these

techniques require immobilization of the nucleic acid on the surface of a sequencing chip and a labeled target molecule. Another recent application of high-throughput analysis of *in vitro* selected pools (of mRNA-displayed peptides) provides binding parameters for the target interactions, but the method requires the target molecule to be immobilized on solid support and relatively slow binding and dissociation kinetics (Jalali-Yazdi, et al. 2016). None of these techniques reveal information about conformational changes associated with target binding but can provide secondary structure constraints thorough covariation analysis of the active RNAs. Moreover, these techniques derive equilibrium binding constants from kinetics and presume an absence of rate-limiting conformational changes associated with target binding, whereas Apta-Seq detects target binding under equilibrium conditions. Apta-Seq makes it significantly easier to discover, sort, and characterize aptamers by measuring structural changes in RNAs in solution, providing a straightforward way to measure affinity for the target molecules. Importantly, Apta-Seq is a powerful enabling technological pipeline that is sure to expedite the transition from aptamer selection to the unraveling of their structure and binding affinity and thus their biological or biotechnological relevance. Furthermore, because Apta-Seq is performed in solution, the method can be used with a large number of targets, providing a label-free approach to studying specificity of the selected sequences. In contrast, high-throughput methods based on immobilization of the RNAs or their binding partners either require modification (immobilization or fluorescent labeling) of each target to measure the binding kinetics or only measure the rates of association, but not dissociation, when unlabeled off-target molecules are introduced in competition with labeled targets. Our results show that ATP binding by the Sassanfar–Szostak motif is in some cases coupled to significant remodeling of the RNA structure in adjacent domains. The method thus reveals not just ligand binding but also concomitant large-

scale conformational changes, facilitating multiplexed experimental discovery of potential riboswitches.

Acknowledgements

Michael Abdelsayed PhD. designed the primers, performed the acylation reactions, and performed sequencing gel analysis. Synthesis of the SHAPE reagent and abundance estimations were performed by Bao Ho PhD. Column binding of RNA extracts and qPCR were performed by Julio Polanco PhD.

Material and Methods

Transcription

RNAs were transcribed for 2 h at 37 °C in 400 μL of a solution containing 40 mM tris chloride; 10% dimethyl sulfoxide (DMSO); 10 mM dithiothreitol (DTT); 2 mM spermidine; 5 mM each rCTP, rGTP, rUTP, and rATP; 20 mM $MgCl_2$; one unit of T7 RNA polymerase; and ∼0.5 μM DNA template. Transcripts were purified by 7% polyacrylamide gel electrophoresis (PAGE) under denaturing conditions (7 M urea). RNA was eluted from the gel into 400 μL of 400 mM KCl and precipitated by adding 800 μL of 100% ethanol at −20 °C.

Primer Phosphorylation

Primer labeling was prepared in a total volume of 20 μL. Then, 20 μM of primer, 1× T4 Polynucletoide Kinase (PNK) ligase buffer (NEB), 1 unit of T4 PNK (NEB), and 0.5 μCi [γ-^{32}P] ATP was incubated at 37 °C for 1 h then purified from denaturing PAGE.

Synthesis of the SHAPE Reagent

The SHAPE reagent, 2-(azidomethyl)nicotinic acid acyl imidazole, was synthesized following a previously described protocol.(32)

Selective 2′-Hydroxyl Acylation and Primer Extension (SHAPE)

SHAPE reactions were prepared in a total volume of 10 μL. RNAs (pools or individual aptamer sequences) were resuspended in water and heated to 70 °C for 3 min. Then, 1 μM of purified RNA was added to a buffer containing 140 mM KCl, 10 mM NaCl, 10 mM tris chloride at pH 7.5, and 5 mM $MgCl_2$. A dilution series of 5′-adenosine triphosphate (ATP) was prepared using a 1:1 stock of ATP:Mg^{2+}. Then, 1 μM to 10 mM ATP was aliquoted to the RNA in a buffer and incubated at RT (~23 °C) for 1 min. Then, 50 mM 2-(azidomethyl)nicotinic acid acyl imidazole was added to the mixture, and the reaction was incubated for 45 min at RT. Reactions containing no SHAPE reagent were substituted with 10% DMSO. Reactions were precipitated with 10 μL 3 M KCl, 1 μL glycoblue, 89 μL H2O, and 300 μL 98% ethanol.

Primer Extension

The RNA pellet was reconstituted in a 10 μL reaction volume containing 0.1 μM 5′-[^{32}P]-radiolabeled reverse transcription DNA primer, 2 μL of 5 × M-MuLV Reverse Transcriptase Reaction Buffer (New England Biolabs), 1 unit of M-MuLV enzyme, and 500 μM each of deoxyribonucleotide triphosphate. Extensions were performed at 42 °C for 15 min. 400 mM NaOH was added, and the reaction was incubated at 95 °C for 5 min to hydrolyze the RNA. Reactions were precipitated with 10 μL of 3 M KCl, 1 μL of glycoblue, 89 μL of H2O, and 300 μL of 98% ethanol. Complementary DNA (cDNA) was resolved using 12% denaturing PAGE or amplified for high-throughput analysis.

KD Analysis

RT stops were normalized by dividing peak intensities in SHAPE profiles by the total number of reads per aptamer and experimental conditions. These normalized profiles (presented as overlays in Figures 4-2 and 4-3 as well as S1–S3) were analyzed by plotting peak intensities for each position in the aptamers as a function of ATP concentration (such as in Figures 4-2B, 4-3D, and 4-4C and D) and modeled with a dissociation constant equation for the ligand

where "Range" corresponds to the range of values above baseline that a peak in the SHAPE profile can assume. The model was fit to the data using a linear least-squares analysis and the Solver module of Microsoft Excel to extract ATP KD for each peak. Because some of the apparent binding constants are near the maximum of the titration and in many cases the rising SHAPE signal did not level off at the highest ATP concentration, we predominantly used the positions where the SHAPE signal decreased with ATP concentration and approached zero for KD modeling.

SHAPE-Seq Library and Primer Design

Round 6 of an *in vitro* selection for an ATP aptamer from a human genomic library (Vu, et al. 2012) was used as the library for Apta-Seq. Libraries were given individual barcodes based on concentration of the ligand used during SHAPE. Ten libraries in total were made using a ligand titration of 1–10 000 μM, no-ligand control, and no-SHAPE (only DMSO, as described above) control.

Apta-Seq primers contain a reverse primer sequence for the pool of interest and flanking Illumina primers in order to barcode and sequence primer extensions by Illumina Sequencing. Apta-Seq primers for reverse transcription primer extension were designed 5′ to 3′ with the following components:

Illumina forward primer reverse-complement, NotI digestion site, Illumina reverse primer, reverse primer for RNA of interest

5′-
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTGCGGCCGCGTGACTGGAGTTCAG
ACGTGTGCTCTTCCGATCCTGAGCTTGACGCA-3′

Primers for amplification were designed 5′ to 3′ with the following components:

Forward primer containing Illumina forward adapter and primer.

5′-

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC

T-3′

Reverse primer containing Illumina reverse adapter, barcode, and Illumina reverse primer.

5′-CAAGCAGAAGACGGCATACGAGAT [barcode]

GTGACTGGAGTTCAGACGTGTGCTCTTCCG-3′

Primer Extension for Apta-Seq

Primer extension was carried out with the Apta-Seq primer as described above. cDNA was self-ligated in a 20 μL reaction using CircLigaseII Reaction Buffer (Epicenter), 2.5 mM MnCl2, 50 μM ATP, and five units of CircLigase ssDNA Ligase. Reactions were precipitated with 10 μL of 3 M KCl, 1 μL of glycoblue, 89 μL of H2O, and 300 μL of 98% ethanol, and cDNA was reconstituted in 20 μL of H2O. A polymerase chain reaction (PCR) was performed using 1 μM each of forward and reverse primers, cDNA template, and DreamTaq Master Mix (Thermo Fisher) and amplified for 16 cycles (denaturing 94 °C, 30 s, annealing 55 °C, 30 s, and elongation 72 °C, 30 s). Amplicons were sequenced on Illumina HiSeq 2500 at the UCI Genomics Facility.

SHAPE-Seq Reactivity Mapping

Galaxy (https://usegalaxy.org/) and the StructureFold module (Tang, et al. 2015) were used to map and determine the SHAPE reactivity of aptamers. Forward reads of libraries were used to analyze RT stops. Adapters and primers (CAATGCGTCAAG) were clipped using Clip adapter sequences on Galaxy. Default settings were used except for a minimum sequence length of 10 and an output of both clipped and nonclipped sequences. Clipped libraries were then processed using StructureFold, a series of Web-based programs to characterize RNA that have undergone SHAPE modifications. Aptamers were mapped to selected libraries using Iterative Mapping on Galaxy. Default settings for mapping were used except for a minimum read length of 12 nucleotides and three mismatches allowed (-v 3). RT stop counts were calculated using Get RT Stop Counts on Galaxy. RT stop counts were derived using mapped files from Iterative Mapping and aptamers as the reference sequences. Output from the Get RT Stop Counts module was tabulated and normalized to the total stop counts for each aptamer. The percentage of counts for each position was derived by dividing each position by the total number of RT counts for each aptamer.

The new aptamers were mapped to the following sequences obtained from Illumina sequencing:

>L1PA16
ctcactacgtGGCTTAGGGCAGGAAGAGAACTGCTAAGGCAGTTTCTCCTAGAAGATGAG
ACCTGCAGCCAGGTCCAGCTTGGTGACCTAGAACTGGTCTGCATGTGTCATTGCTGG
GTGCTCCACCCTGCTCCCCTGAGATCATGTTGag
>THE1B
ggggcagacgtgcctcactaGTTGTGGGAGGGACCCGGTGGGAGGTAACTGAATCgTGGGATG
AGTCTTTCCCGTACTGTcGTCATGATAGTGAATAAGTCTCATGAGgTCTGATGGtTTTT
ATATGGGGGAGcTTTCCCTTCgCAAACTCTag
>*PRR5*
gggtGGAGGAGCCTGGATGCTGCCTGCAGGACCTCAGGCTGTGCCTGCTGGGCAAAGg
CCCTGGGCAGGGAAGGAACTGCAGCCTCCACAGAGGGTGGATaTGGTGGAGAGGTG
GGAGGCCAGCTCCTGTCATCCGAGGTCCAGGCAAGCCAG

Lowercase letters are mutations acquired during the *in vitro* selection or library construction, as compared with the reference genome. The Sassanfar–Szostak adenosine-binding loop is highlighted in bold.

Structure Prediction

The Reactivity Calculation module on Galaxy was used to obtain SHAPE reactivities by using the output from the Get RT Stop Counts module. The RT stop counts for 10 mM ligand concentration was used as the (+) library and no ligand for the (−) library, both in the presence of the SHAPE reagent. Default settings were used except for nucleotide specificity, which was changed to AUCG. To predict structure, the RNAprobing Web server was used. All default settings were used according to the Washietel et al. SHAPE method (Washietl, Hofacker, et al. 2012).

Total RNA Extraction

Human cell lines OV90 and MCF7, ovary- and breast-derived adenocarcinoma, respectively, were 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. Similarly, SHSY-5Y neuroblastoma cells were thawed and seeded in two separate T-75 flasks with DMEM/F12 culture media containing 10% FBS, 10% amphotericin B, and 10% penicillin/streptomycin. All cell cultures were passaged appropriately to achieve 80% confluency on the day of the experiment. In one SHSY-5Y experiment, 2% DMSO with 10 mM adenosine was added to adhered confluent SHSY-5Y cells 1 day prior to total RNA extraction.

Total RNA was harvested from all cell types using TRIzol Reagent (Ambion). All total RNA isolation steps were performed according to the user manual. Briefly, adhered cells were

collected by washing each culture dish with 750 µL of TRIzol Reagent and collected into a fresh 1.5 mL microcentrofuge tube. A total of 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 with 75% ethanol prior to resuspension in RNase-free H2O. 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 remove protein and residual DNA contaminants before column binding and RT-qPCR.

Column Binding Assay Using Total RNA Extracts

Prior to RNA binding, 10 µL of C8-linked ATP-agarose (1 mg mL–1; Sigma-Aldrich) was pre-equilibrated on a spin-filter with binding buffer (BB) containing 140 mM KCl, 10 mM NaCl, 5 mM $MgCl_2$, and 20 mM tris-HCl, at pH 8.0. For each tissue culture source (OV90, MCF7, SHSY-5Y, and SHSY-5Y incubated with 10 mM adenosine), 500 ng of total RNA was annealed in the presence of BB at 72 °C for 1 min and allowed to cool to RT over 5 min before addition onto ATP-agarose beads. A total of 11 fractions were collected starting with the flow-through, which was collected after a 30 min incubation. After flow-through collection, a total of four washes were collected over a span of 10 min using 10 µL of BB. After washing, four 30 min elution fractions were collected using an elution buffer (EB) containing 5 mM ATP, 140 mM KCl, 10 mM NaCl, 10 mM $MgCl_2$, and 20 mM tris-HCl, at pH 8.0. To remove any remaining RNAs attached to the column, one 10 µL denaturing wash was collected using 7 M urea following the elution fractions. Each fraction was precipitated using 300 mM KCl and 100% ethanol, pelleted by centrifugation, washed with 75% ethanol, and resuspended in ddH2O for reverse transcription.

RT-qPCR

Pellets from the last (fourth) wash and the elution fractions were resuspended in a 1× reverse transcription master mix containing 500 nM of aptamer-specific RT primers, 5 mM dNTPs, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, and 50 mM tris-HCl at pH 8.3. Resuspended RNAs were heat-treated at 90 °C for 30 s, followed by a 50 min incubation at 50 °C after the addition of superscript III reverse transcriptase. cDNA products were diluted 10-fold with nuclease-free H2O, and quantitative PCR was performed on a BioRad CFX Connect system using 1 μL of cDNA product, 500 nM of gene-specific primers, and BioRad iTaq supermix. The reverse primers were 3′-extended versions of the RT primers to ensure gene-specific amplification.

Supporting Info

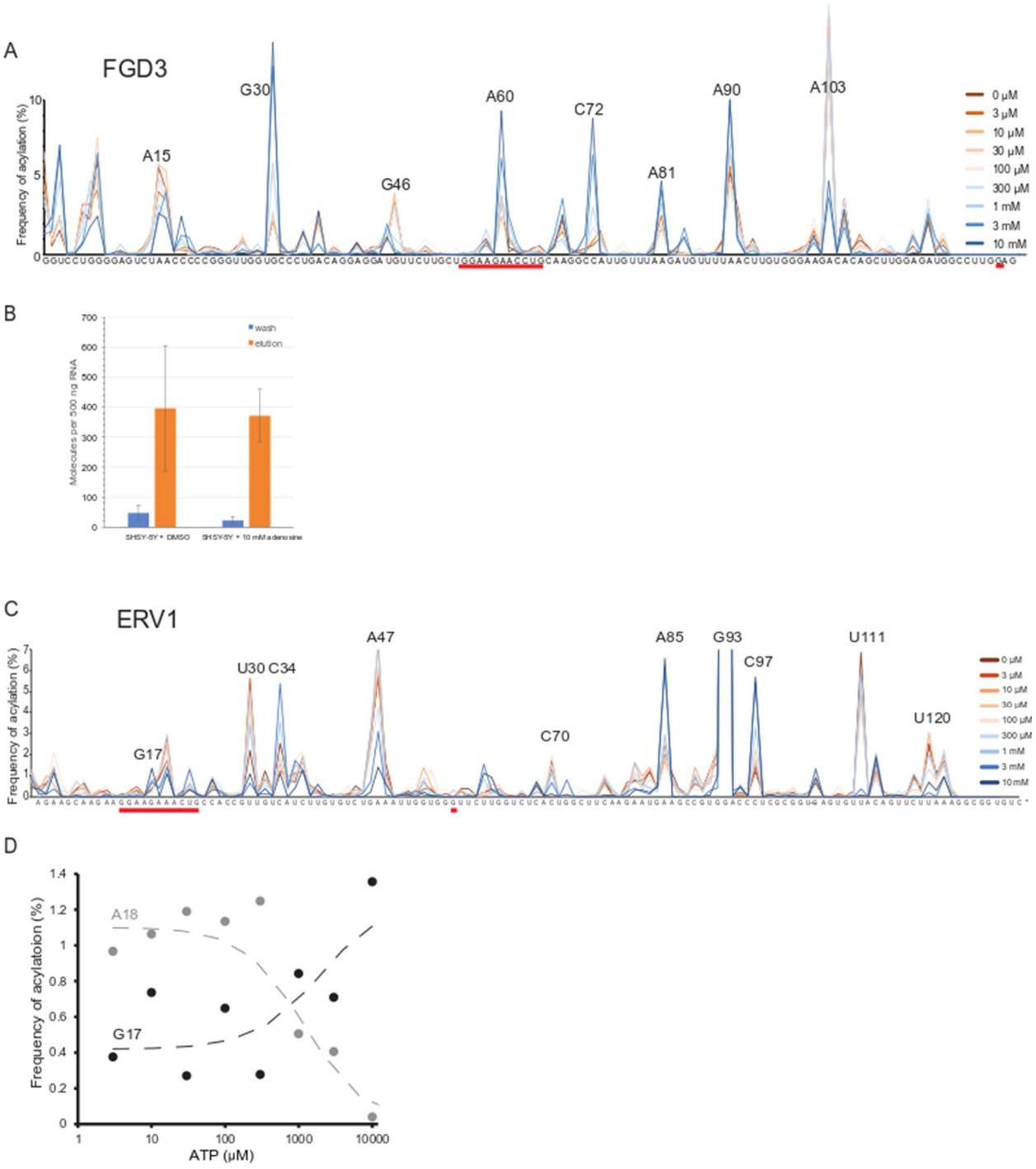


Figure S4-1. Related to Fig.4-2. Apta-Seq data of previously reported human *FGD3* and ERV1 aptamers (A) Trace of acylation frequency for the *FGD3* aptamer with varying ATP concentrations from no ATP to 10 mM ATP. The adenosine binding motif is underlined in red. (B) Binding of ATP beads by the *FGD3* sequence measured using RT-qPCR of 500 ng of total RNA isolated from the human SHSY-5Y cell line incubated in absence (DMSO) or presence of 10 mM exogenous adenosine. The aptamer quantities are expressed as the number of molecules in the

last wash and ATP elution fractions out of the starting 500 ng of total cellular RNA. Error bars are s.e.m. of three qPCR replicates. (C) Trace of acylation frequency for the ERV1 aptamer with varying ATP concentrations from no ATP to 10 mM ATP. The adenosine binding motif is underlined in red. (D) Binding isotherms of the ERV1aptamer modeled from the acylation

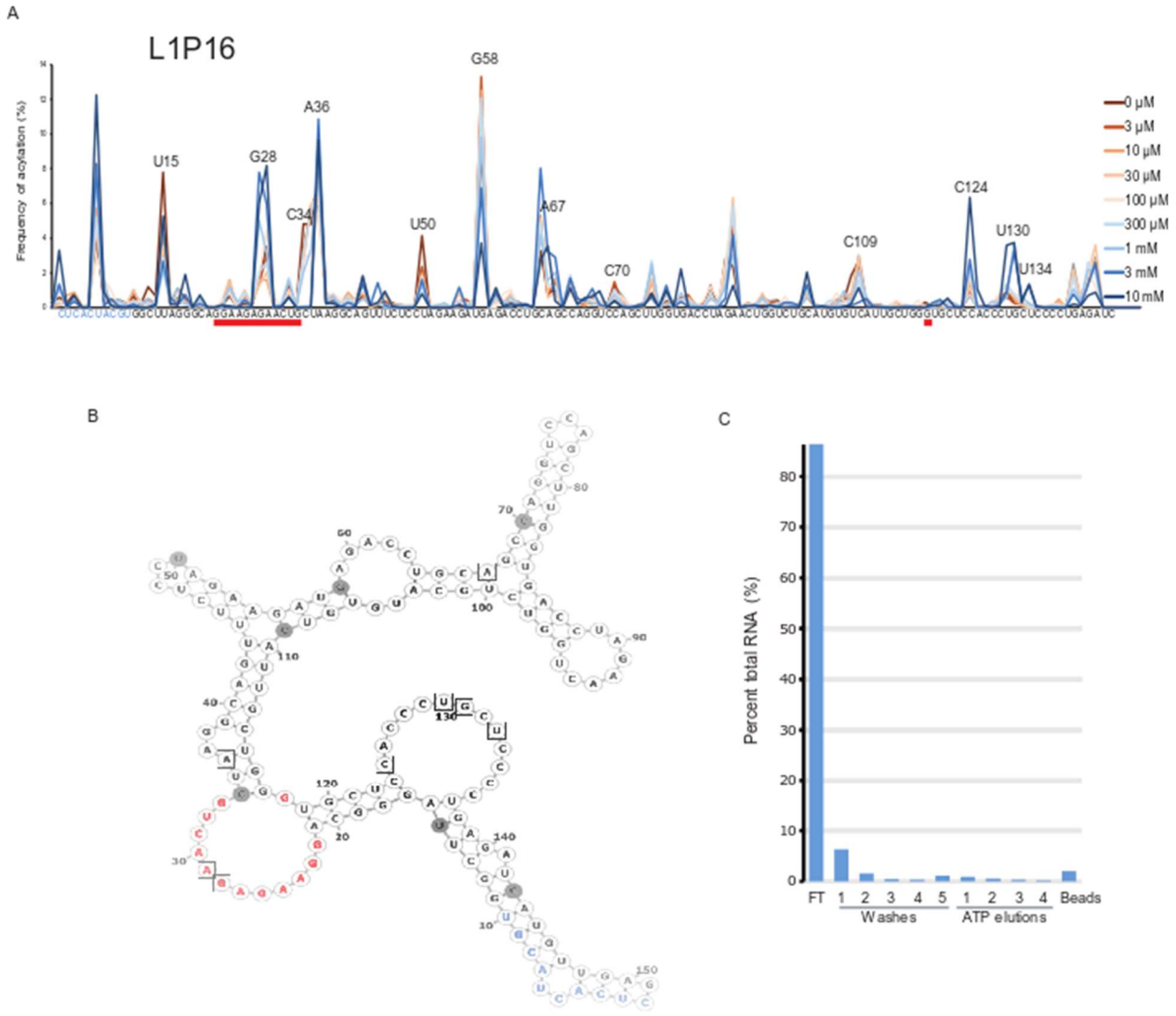


Figure S4-2. Related to Fig.4-4. Apta-Seq data of the human L1PA16 adenosine aptamer. (a) Trace of acylation frequency for the L1PA16 aptamer with varying ATP concentrations from no ATP to 10 mM ATP. The adenosine binding motifs underlined in red. (b) Secondary structure of the L1PA16 aptamer with ATP binding loop in red and nongenomic regions in blue. Nucleotides boxed in black indicate positions that show an increase in acylation with increasing ATP concentration whereas those that show a decrease are filled in grey. (c) Elution profile of the G26A mutant of the L1PA16aptamer, showing lack of binding to ATP beads (compared to the binding profile of the wild-type sequence shown in Fig. 4-4E). Wash 5 was incubated longer than wash 4, and equally long as the elutions, resulting in slight increase of RNA in the fraction.

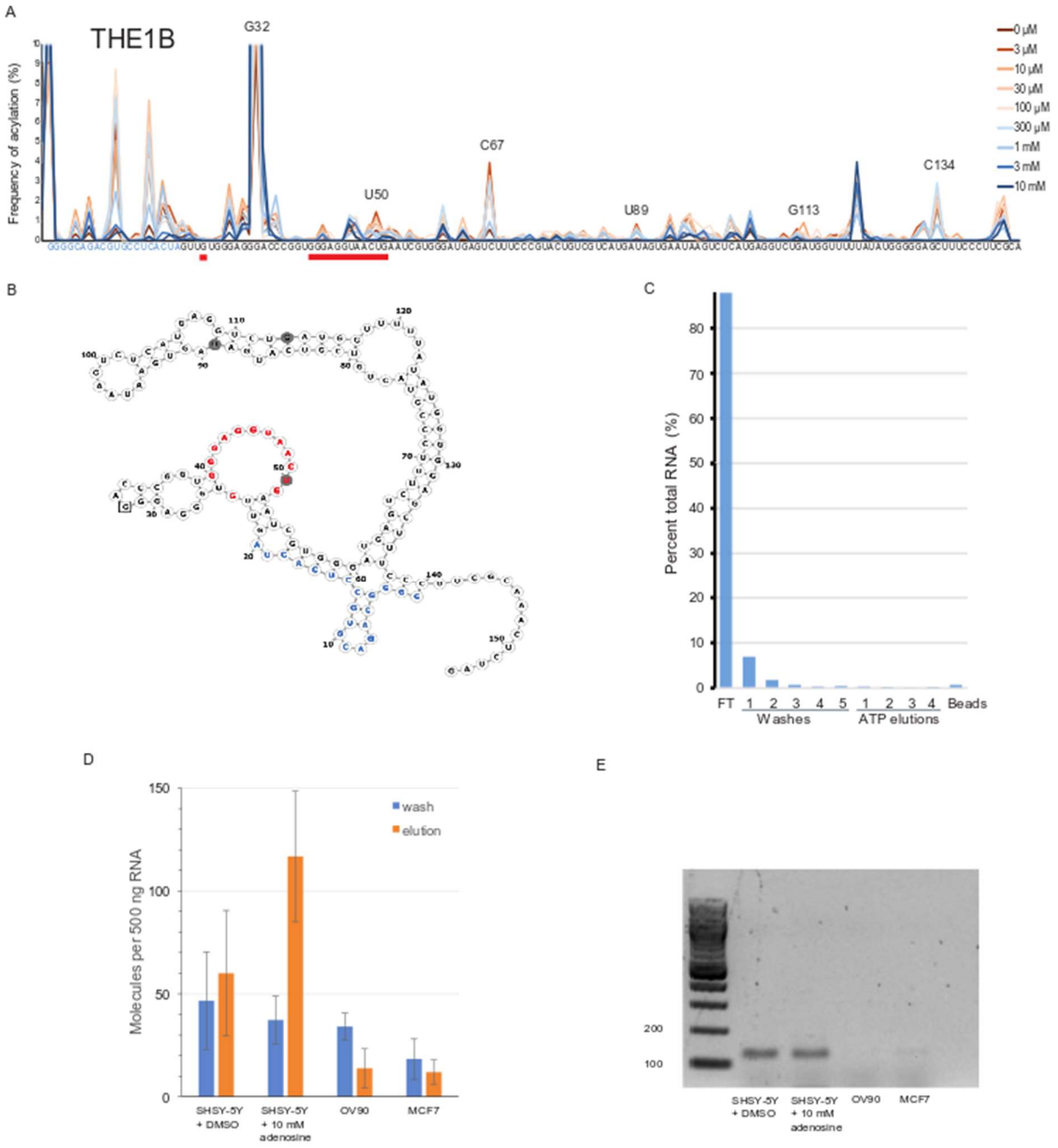


Figure S4-3. Related to Fig.4-4. Human THE1B adenosine aptamer. (A) Trace of acylation frequency for the THE1B aptamer with varying ATP concentrations from no ATP to 10 mM ATP. The adenosine binding motif is underlined in red. (B) Secondary structure of the THE1B aptamer with ATP binding loop in red and nongenomic regions in blue. Nucleotides boxed in black indicate positions that show an increase in acylation with increasing ATP concentration while those that show a decrease are filled in grey. (C) Elution profile of the G24U mutant showing lack of binding to ATP (compare with Fig. 4-4F for the binding profile of the wild-type sequence). (D) ATP beads binding by the THE1B sequence measured using RT-qPCR of 500 ng of total RNA isolated from the human cell lines indicated below the graph. The aptamer quantities are expressed as the number of molecules in the last wash and ATP elution fractions out of the starting 500 ng of total cellular RNA. Error bars are s.e.m. of three qPCR replicates. (E) Agarose gel electrophoresis of the THE1B aptamer amplified from the elution fractions in the indicated cell lines.

Chapter 5 CIM-seq enabled discover of ATP Aptamers

Introduction

In vitro selection is now a common tool for discovery of both natural RNAs and synthetic ones. When combined with high throughput sequencing, these selections can generate tens to thousands of candidate sequences (Sedlyarova, et al. 2017, Klopf, et al. 2018). Biochemical characterization of these sequences is often the limiting factor as each sequence must by synthesized and screened individually and thus further study is often limited to only the most abundant sequences or those with more readily discernable structures and motifs. This can leave thousands of potentially interesting aptamers uncharacterized. From previous genomic selections, we saw that the abundance of a sequence in selection does not correlate with its ability to function *in vivo* (See Chapter 1). Indeed, many aptamers discovered by other methods have been found in very low abundance in selections (Brunel, et al. 2001, Lorenz, et al. 2010).

This challenge applies to ATP binding selection as well. High throughput sequencing reveals thousands of sequences detected only a handful of times and tens of sequences with at least 4 reads (See Chapter 3). Only those with potential ATP binding motifs or with high abundance were initially tested. Multiplexed mutagenesis with reselection (See Chapter 3), and Apta-seq (See Chapter 4) were used to gather data on several sequences of the pool at once, but additional methods to characterize the pools are required for more comprehensive data on each sequence. This can be achieved by applying biochemical methods to the selected pools, followed by high throughput sequencing analysis, and demultiplexing the data by mapping to a reference genome of individual candidates as seen in Apta-seq.

One method of characterizing RNAs identifies functional regions by mapping the effect of strand cleavage to the function of the RNA. Cleaved fragments of an RNA are generated by

random incorporation of phosphorothioate nucleotide analogs by *in vitro* transcription and cleavage of the RNA backbone at incorporation sites by iodine treatment. These cleaved RNAs are then sorted by their ability to function (ligand binding for aptamers) and analyzed by gel electrophoresis. To map cleavage sites detrimental to the RNA, the RNA is transcribed in four parallel reactions corresponding to each of the four nucleotides and modified with a radioactive phosphate by ligation or phosphorylation. Once the labeled RNAs are separated by function gel electrophoresis of all four reactions, RNA cleavage sites that interferes with function are identified by their enrichment in the inactive fraction (unbound population for aptamers) and the nucleotide cleaved is indicated by it its position in the gel. This is similar in process and principle to nucleotide analog interference mapping (NAIM) (Cochrane and Strobel 2004) which instead probes for interference from the incorporation of the analog itself in the full-length transcripts and is subsequently cleaved prior to gel electrophoresis to obtain sequence information.

The cleavage interference mapping (CIM) method was multiplexed by combining it with RNA-seq strategies to allow interrogation of a pool of sequences at once. Analysis of CIM by sequencing eliminates the need for parallel transcription with each nucleotide and radiolabeling of the RNA, reducing the amount of individual reactions performed per RNA analyzed even further. The workflow follows the following steps: RNA fragments are generated by partial alkaline hydrolysis, active and inactive fragments are isolated, and both fractions are analyzed by RNA-Seq. While alkaline hydrolysis is biased by secondary structure, the hydrolysis profile can inform on each RNAs secondary structure and the depth of high throughput sequencing will still allow for interrogation of each nucleotide. This method was applied to the ATP binding pool described previously (Chapter 2) to characterize multiple aptamer candidates as well as verified sequences in the same pool.

CIM-Seq

Fragments of the ATP binding pool were generated via alkaline hydrolysis of RNA from the round 6 of the ATP selection (See Chapter 2). The hydrolyzed RNAs were precipitated and desalted before binding to ATP-agarose beads. Nonbinders were collected as flowthrough and binders eluted with urea and EDTA. Flowthrough and elution fractions were collected and reverse transcribed using the same primer design as in Apta-seq (See Chapter 4). The cDNAs were then circularized with Circ Ligase and amplified via PCR to introduce barcoding primers for sequencing. The resulting sequences were then mapped using bowtie2 to a reference genome of sequences already discovered in the pool from previous experiments. The start site of the sequences, which corresponds to sites of cleavage, were mapped and plotted for each sequence to produce cleavage profiles to be compared between flowthrough and elutions sequencing reactions

The sequenced flowthrough fraction produced about five times any many reads as the elution fraction. It was also much more diverse containing reads that mapped to about half of the sequences reported thus far. CIM analysis, however, was limited to sequences detected in the both fractions. As a whole, the pool predominantly mapped to the ERV1 aptamer, the *FGD3* aptamer, and the chromosome 15 conserved EST.

The ERV1 binding motif is thought to be very close to the 5′ end with the first nucleotide of the stem just three nucleotides past the pool primer. Thus, 5′ end mapping is not expected to be particularly informative for this aptamer. What is discernable is that the ERV1 aptamer has a strong propensity for hydrolysis in a specific site. Two very prominent peaks were present; one corresponding to full length and one to the cleavage in the downstream loop of ERV1. These

two peaks comprised 77% of the reads from ERV1 across the two fractions. This cleavage occurs in the predicted loop of a downstream independent structure. This could indicate that the ERV1 structure is heavily protected in from hydrolysis in other regions.

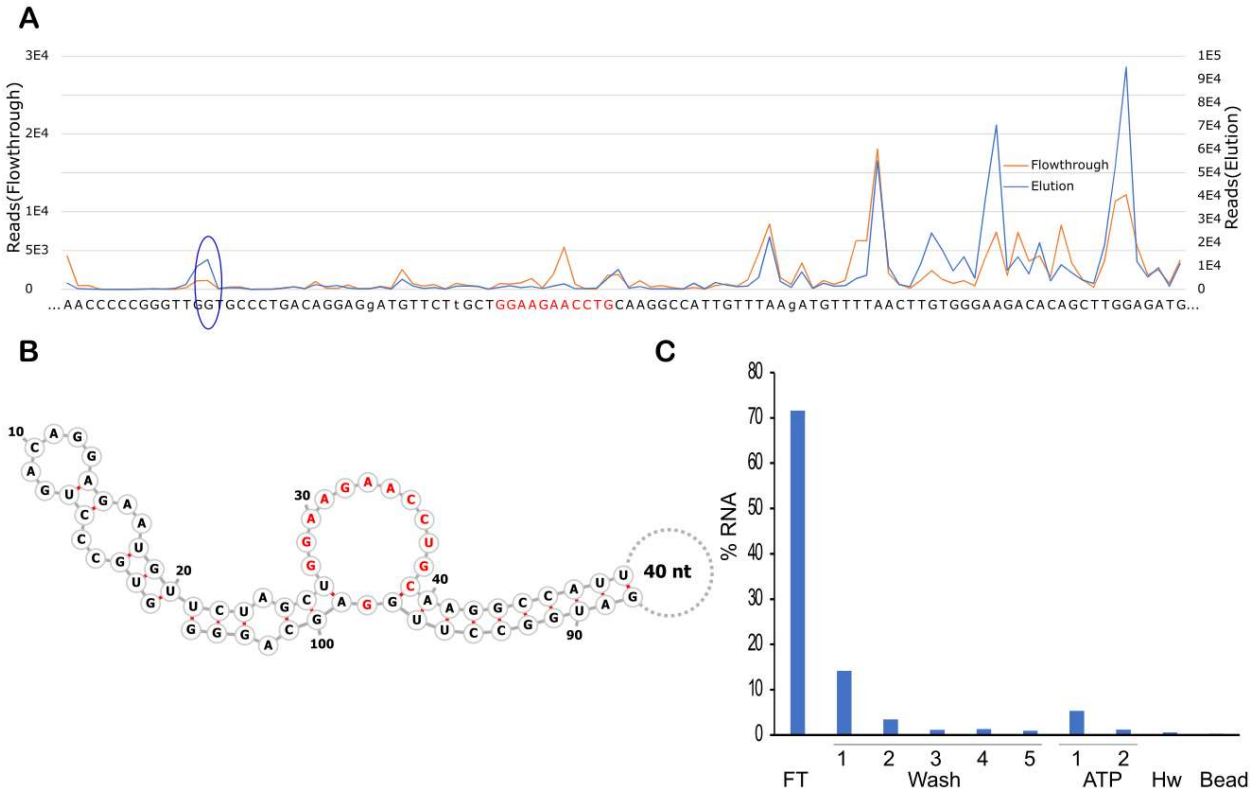


Figure 5-1 *FGD3* Aptamer CIM-seq (**A**) Trace of cleavage mapping of the *FGD3* aptamer for elution (Blue) and flowthrough (Orange) fractions. ATP binding loop shown in red and displays flowthrough enriched peak. (**B**) Predicted structure of the *FGD3* genomic construct derived from CIM. The nucleotides of the ATP binding motif are shown in red. (**C**) Graph of column binding fractions for the *FGD3* genomic construct. FT designates flowthrough fraction and HW designates denaturing wash. The graph displays 6% of RNA eluting with free 5 mM ATP

The *FGD3* aptamer was used to validate the method. In the cleavage interference mapping of the *FGD3* aptamer, we can see an inhibitory effect when the consensus loop is cleaved as shown by the peak in the flowthrough corresponding to the 4th adenosine in the loop. There is also enrichment in the elution for aptamers that start 27 nt upstream of the binding loop. This prompted the construction of the corresponding genomic construct with the 5′ start site informed by CIM. This construct showed some binding and elution (~6%) to ATP in equilibrium

conditions, while all the previous genomic constructs showed no binding (~0.6 %) unless bound cotranscriptionally. There was also a surprising amount of enrichment for sequences cleaved after the ATP binding motif. These segments are expected to lack the ATP binding motif and would not be expected to be enriched in the elution fraction. Two of these occurs in loops and the other in a stem near a bulged G, if the cleaved fragments found each other again during refolding of the fragmented pool, it's possible this could have relaxed strain on the folded *FGD3* aptamer structure

The most common sequences that mapped to the conserved Chr 15 EST were full length transcripts that were found primarily in the flowthrough fraction. The overall hydrolysis profile has peaks in both fractions corresponding the ATP protected region from in line probing. This would indicate that in the absence of ATP (such as the conditions of hydrolysis) this area is in a loop and is easily hydrolyzed. The absence of peaks in enriched the flowthrough fraction seemed to indicate cleavage interference but there were a few 5′ start sites that were enriched in the elution. These were used to make additional constructs to bind on an ATP column. The genomic construct (Fig 5-2 C) had 9% of its RNA bound to ATP beads, however only 4% competitively eluted with free ATP. Modest binding with weak elution from the beads represented a common trend for many of these constructs which require denaturing conditions to remove them from the beads. Perhaps the RNA has a higher affinity for ADP or AMP which may be present as a degradation product of the immobilized ATP. This hypothesis has not been fully experimentally confirmed.

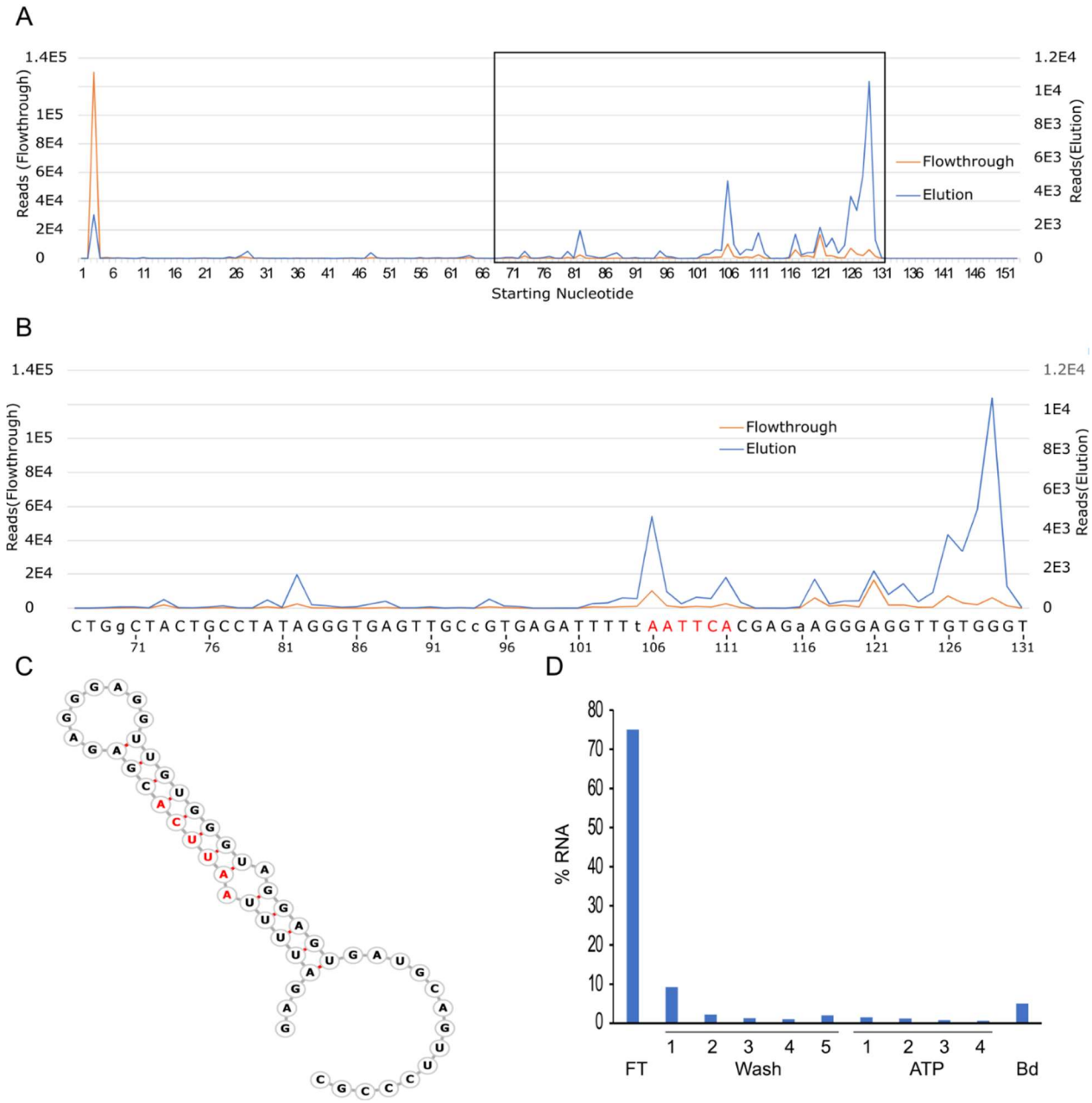


Figure 5-2 Chromosome 15 sequence CIM-seq (**A**) Full trace of cleavage mapping of the chromosome 15 sequence for elution (blue) and flowthrough (orange) fractions numbered from the start of the full sequence including constant regions (**B**) Boxed region of (A) showing peaks used to make genomic constructs. ATP sensitive domain from in line probing shown in red (**C**) Predicted structure of the chromosome 15 genomic construct derived from CIM the nucleotides of the ATP sensitive region are shown in red. (**D**) Graph of column binding fractions for the genomic construct. The graph displays some RNA bound to beads but not eluting with free 5 mM ATP.

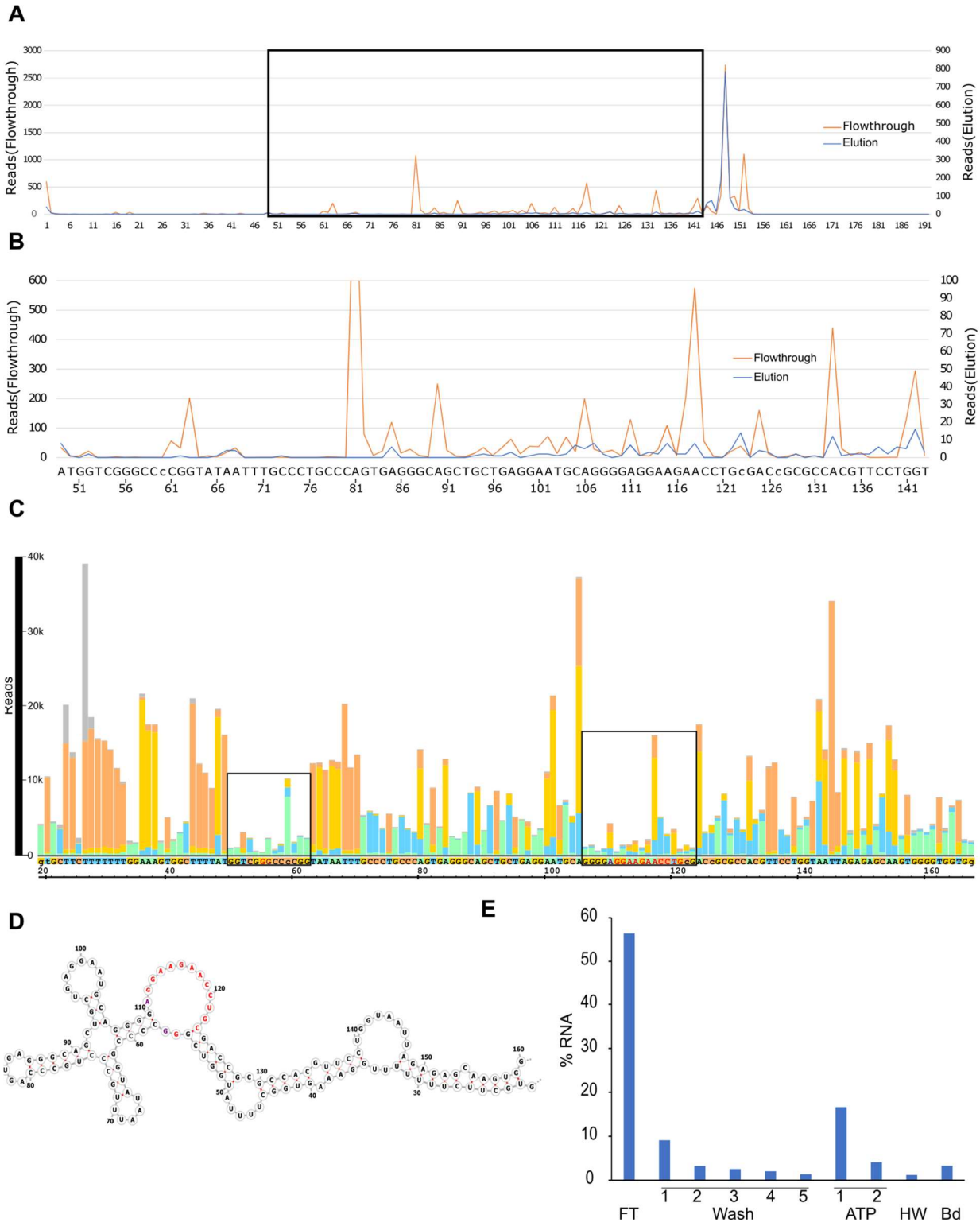
Additional ATP binding sequences

Reads that did not align to our reference of known aptamers were clustered via the fastaptamer software package. The top sequence for each cluster was retrieved with a grep command utilizing fastaptamer's header, and these top sequences were manually aligned with Jalview. Full length sequences were constructed from the aligned fragments and were compared to the human reference genome. This procedure revealed several sequences not identified in earlier sequencing experiments, two of which had the potential to fold into the ATP binding structures. A genomic sequence mapping antisense to the first exon of *CSDC2* on chromosome 22 and one mapping to the (-) strand of an ALR/Alpha repeat region on chromosome 21. A third sequence, mapping to an intron of *MUC20*-OT13 on chromosome 3, contained two potential instances that were similar to the 11-nucleotide recognition loop but seems to lack the ability to fold into the canonical ATP binding motif. All these new sequences were used as a reference genome to map the CIM fragments as well as the mutagenized pool to generate CIM traces and mismatch pileups for each.

While the *CSDC2* sequence contains a GGWAGADNHTG sequence, attempts to fold it into the canonical adenosine-binding structure leaves an extra A preceding this sequence and an extra G in the opposing bulge, leaving its activity uncertain. Cleavage interference mapping indicated that this motif is indeed active, as there is a significant peak on the 4th adenosine of the proposed binding loop which appears to interfere with ATP binding. Analysis of *CSDC2* mutants also reveals significant conservations in the proposed binding loop, the opposing G, and the upstream nucleotides forming the shorter stem. This conservation is weakest for the extra A and G of the binding motif, where mutation can actually lead to formation of the canonical ATP binding motif. This sequence was selectively amplified from the pool with sequence specific primers and confirmed to bind and elute efficiently with ATP in agreement with CIM and

conservation data. One of the Cs forming the ATP binding stem arose from an unintentional mutation picked up during selection. The genomic sequence contains an A instead which would weaken the fold the ATP binding motif, indicating that this sequence may be an artifact of selection. However, the mismatch pileup also shows that sequences with the natural A did indeed survive the selection after mutagenesis.

Figure 5-3 *CSDC2* aptamer candidate CIM-seq (**A**) Full trace of cleavage mapping of the *CSDC2* aptamer for elution (blue) and flowthrough (orange) fractions numbered from the start of the full sequence including constant regions (**B**) Boxed region of (A) showing putative ATP binding motif in red with a flowthrough peak greatly enriched. (**C**) Mismatch pile up of mutants mapped to *CSDC2* aptamer. Color indicates the identity of the mutated nucleotide: green (A) blue (T) orange (C) gold (G). Bar height shows number of mutants Putative ATP binding sequence is displayed in red. High levels of conservation are displayed in the 11 nt loop, the opposing G, and surrounding nucleotides (Boxed). Large A118G mutants in the recognition loop maintains consensus sequence. C60A mutants correspond to the genomic sequence (**D**) Predicted structure of the *CSDC2* ATP binding motif are shown in red. Sequence in numbered from the start of the full sequence (**D**) Graph of column binding fractions for *CSDC2* aptamer. FT represents flowthrough, $_{DW}$ represents denaturing wash. The graph displays ~20% of total RNA elute off the beads with ATP confirming its activity.


A
Reads(Flowthrough)
Reads(Elution)
Flowthrough
Elution
B
Reads(Flowthrough)
Reads(Elution)
Flowthrough
Elution
ATGGTCGGGCCcCGGTATAATTTGCCCTGCCCAGTGAGGGCAGCTGCTGAGGAATGCAGGGGAGGAAGAACCTGcGACcGCGCCACGTTCCTGGT
C
Reads
D
E
% RNA
FT
Wash
ATP
HW
Bd

The ALR/Alpha sequence contains a perfect instance of the known ATP binding motif. However, this RNA only contains the ATP binding sequence due to an A to G mutation picked up in selection that provides the first G of the consensus loop. Moreover, the forward primer provides two of the base pairs for the 5′ stem of the motif, making this aptamer very likely an artifact of selection. Mapping the cleavage products of the ALR/Alpha sequence shows several peaks in the ATP binding motif that seem to be enriched for in the flowthrough fraction in agreement with its role in ATP binding. One major peak in the enriched in the elution fraction correlates to the loop internal to the binding motif. This again may point to the ability of these aptamers to reform as a dimer in solution with a smaller propensity for alternative folding. Comparative analysis of the mutants demonstrates high levels of conservation for the recognition loop, opposing G, and flanking stems of the ATP binding motif as expected. It's interesting to note that the bases which are thought to participate in base pairing with the primer sequence is held especially constant in the mutagenesis as the primer does not allow for compensatory mutations.

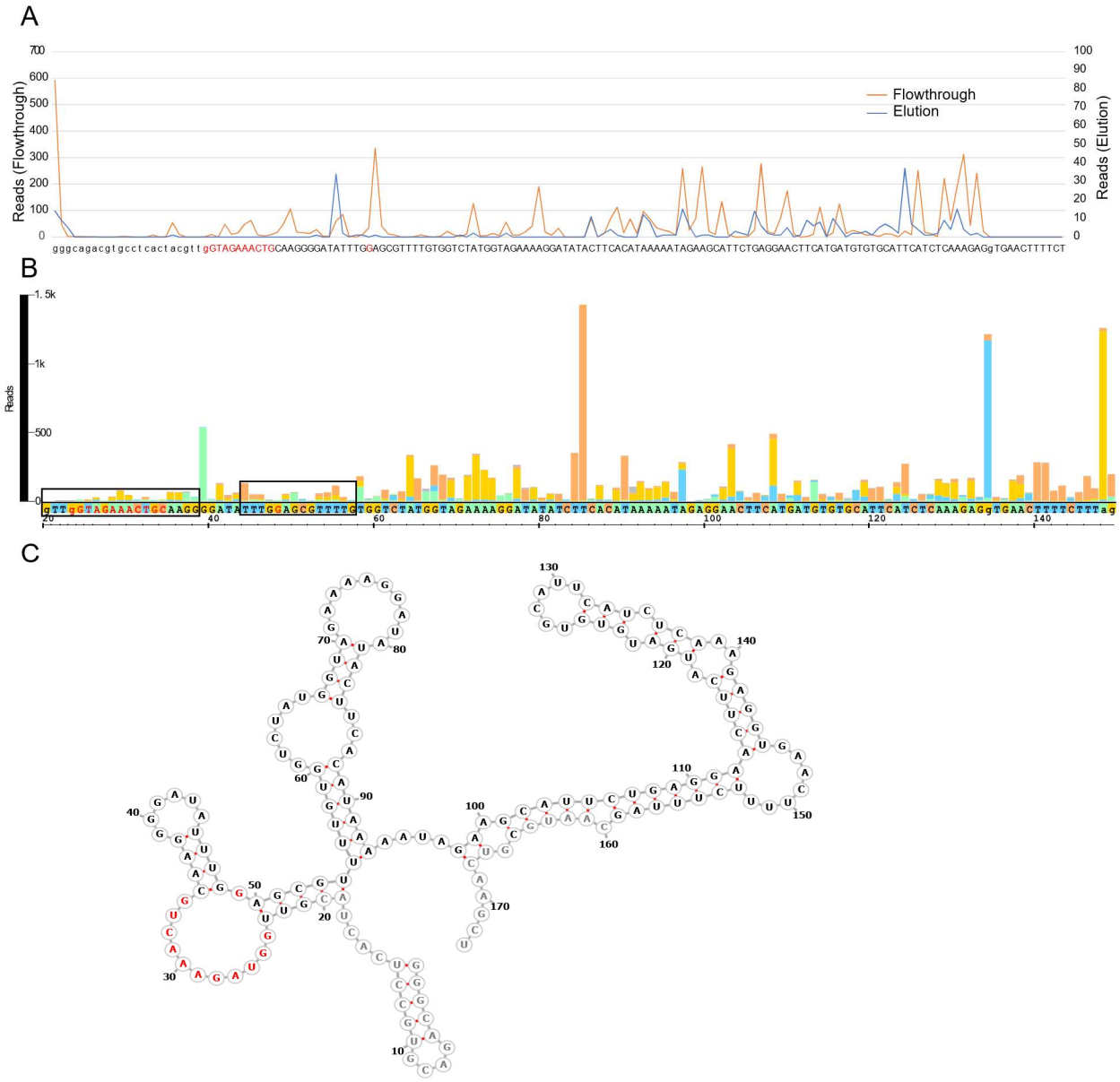


Figure 5-4 ALR/Alpha aptamer CIM-seq (**A**) Trace of cleavage mapping of the ALR/Alpha aptamer for elution (blue) and flowthrough (orange) fractions. (**B**) Mismatch pile up of mutants mapped to ALR/Alpha aptamer. Color indicates the identity of the mutated nucleotide: green (A) blue (T) orange (C) gold (G). Bar height shows number of mutants. ATP binding sequence is displayed in red. Areas of high conservation (boxed) correspond to the 11 nt loop, the opposing G, and surrounding nucleotides with allowed mutations conserving base pairing. (**C**) Predicted structure of the ALR aptamer. the nucleotides of the ATP binding motif are shown in red. Constant regions are displayed in gray

The sequence mapping to *MUC20* overlapping transcript1 has two instances of potential ATP binding loop that do not conform exactly to the 11-nucleotide recognition loop from previous study. However, there does not seem to be a region within the sequence that can base pair with the surrounding sequence in order to form the prerequisite ATP binding structure. Cleavage interference mapping suffers from low coverage depth for the sequence but would seems to indicate the downstream copy is more sensitive to cleavage than the upstream copy. The pile up of mutants show no significant conservation in either of the copies but, seemingly contrary to CIM data, suggest the upstream copy is the more conserved. When selectively amplified from the pool, the sequence does in fact bind ATP with 20% eluting with free ATP. The mode of binding is not yet known; however, this modified recognition sequence is found 28 times within a ~1,200 nucleotide stretch within *MUC20*-OT1 intron making this motif potentially very important for its function.

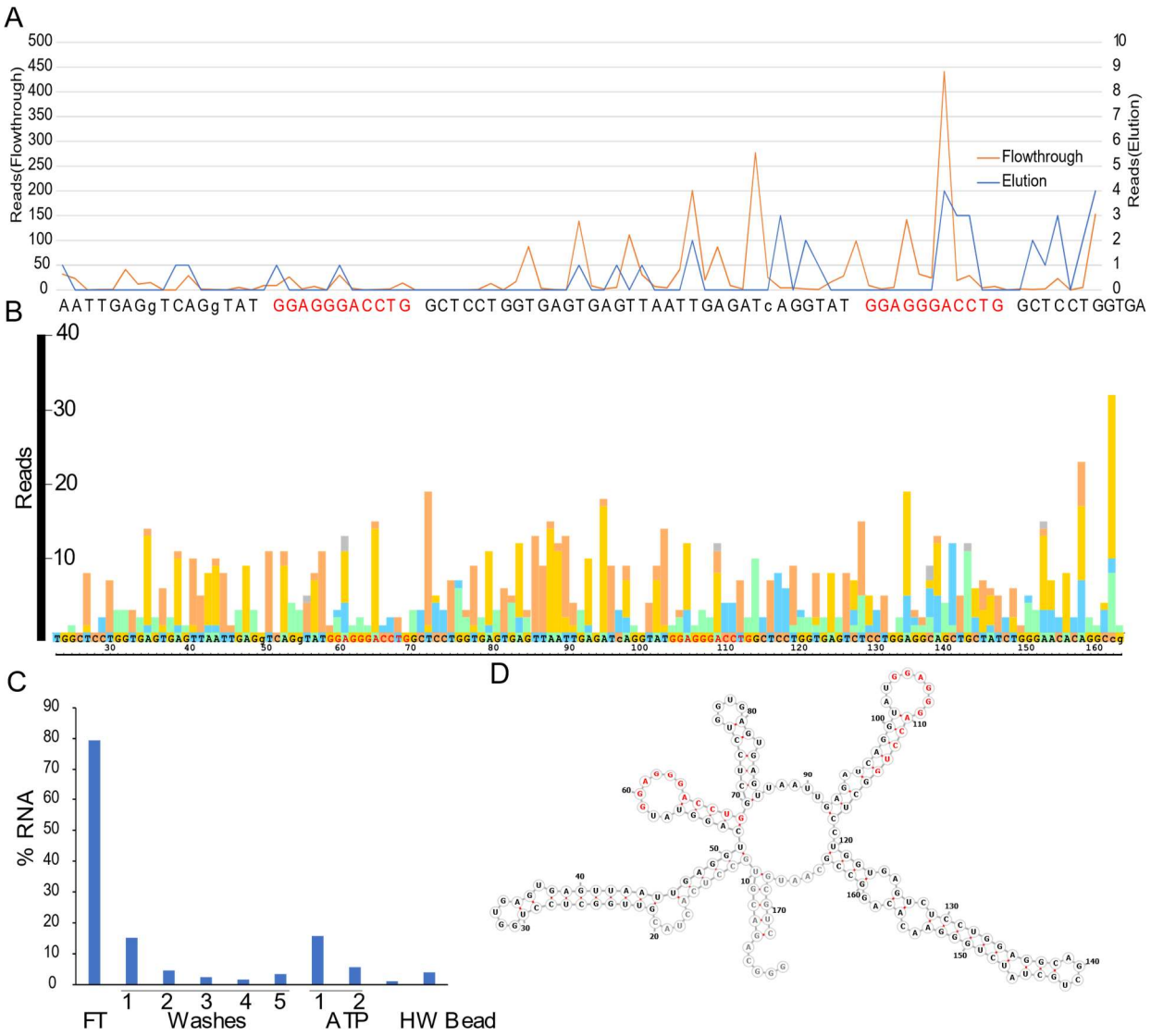


Figure 5-5 *MUC20*-OT1 aptamer candidate CIM-seq (**A**) Partial trace of *MUC20*-OT1 RNA for elution (blue) and flowthrough (orange) fractions showing putative ATP binding loops in red. Note the low reads on the elution axis. (**B**) Mismatch pile up of mutants mapped to *MUC20*-OT1 sequence. Color indicates the identity of the mutated nucleotide: green (A) blue (T) orange (C) gold (G). Bar height shows number of mutants. Potential ATP binding loops are displayed in red. (**C**) Graph of column binding fractions for *MUC20*-OT1 aptamer. The graph displays ~20% of total RNA elute off the beads with 5 mM ATP. (**D**) Predicted thermodynamic structure of the *MUC20*-OT1 ATP binding motif are shown in red. Constant regions displayed in gray

Conclusions

ATP binding by RNAs is much more common than initially thought with at least eight examples found via *in vitro* selection and one by structure-based searches. Some of these are have been artificially enhanced via mutations or the constant regions from selection, but many of them are likely to bind ATP *in vivo*. Combined with the widespread occurrence of GTP aptamers (Curtis and Liu 2013, Curtis and Liu 2014)and the presence of a folic acid aptamer (Terasaka, et al. 2016), these results could indicate ligand binding is a prominent function of human RNA. The reasons for ATP binding are potentially diverse, as these sequences have now been found in introns, exons and retrotransposons.

Cleavage interference mapping was able to generate biochemical data on several sequence in the ATP binding pool in parallel. These data generated from the known aptamers were congruent with what we expect while also revealing additional information leading to an improved genomic construct for the *FGD3* ATP aptamer and the potential ATP aptamer on the EST of chromosome 15. The method provided biochemical evidence of binding for the new candidate, *CDC2* which was then confirmed *in vitro* demonstrating its potential in guiding the selection of candidates.

Multiplexing this method with high throughput sequencing generated data for several sequences, however the amount of sequences analyzed is limited by the abundance of the sequence in the pool. Reliable CIM data requires adequate sampling of each nucleotide which often is not detected for low abundance sequences. The inability of CIM to provide evidence for the binding of the *MUC20* OT1 RNA is likely evidence of this. Future efforts may benefit from depleting the pool of the dominant sequences to allow more even representation and from higher depth sequencing. Some information on secondary structure could be inferred from the partial

hydrolysis profile, however they correspond to the structure of the RNA in the absence of ligand, which may change upon ligand binding. Perhaps future applications of this method could perform hydrolysis of the RNA pool in the presence of ligand, to simultaneously refine secondary structure predictions.

Material and Methods

Transcription

RNAs were transcribed for 2 h at 37 °C in 400 μL of a solution containing 40 mM tris chloride; 10% dimethyl sulfoxide (DMSO); 10 mM dithiothreitol (DTT); 2 mM spermidine; 5 mM each rCTP, rGTP, rUTP, and rATP; 20 mM $MgCl_2$; one unit of T7 RNA polymerase; and ∼0.5 μM DNA template. Transcripts were purified by 7% polyacrylamide gel electrophoresis (PAGE) under denaturing conditions (7 M urea). RNA was eluted from the gel into 400 μL of 400 mM KCl and precipitated by adding 800 μL of 100% ethanol at −20 °C.

Alkaline Hydrolysis

Round 6 of an *in vitro* selection for an ATP aptamer from a human genomic library was hydrolyzed for 2, 5, and 10 minutes in 50 mM $NaHCO_3$ and 1 mM EDTA at 72 °*C* pH 10. Aliquots of each fraction were run on a gel and the 5 minute hydrolysis was chosen as the RNA with the greatest distribution of RNAs.

CIM-Seq Library and Primer Design

Round 6 of an *in vitro* selection for an ATP aptamer from a human genomic library (Vu, et al. 2012) was used as the library for CIM-Seq. Primers were used as described in Apta-seq (Abdelsayed, et al. 2017).

Apta-Seq primers contain a reverse primer sequence for the pool of interest and flanking Illumina primers in order to barcode and sequence primer extensions by Illumina Sequencing. Apta-Seq primers for reverse transcription primer extension were designed 5′ to 3′ with the following components:

Illumina forward primer reverse-complement, NotI digestion site, Illumina reverse primer, reverse primer for RNA of interest

```
5′-
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTGCGGCCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCTG
AGCTTGACGCA-3′
```

Primers for amplification were designed 5′ to 3′ with the following components:

Forward primer containing Illumina forward adapter and primer.

```
5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′
```

Reverse primer containing Illumina reverse adapter, barcode, and Illumina reverse primer.

```
5′-CAAGCAGAAGACGGCATACGAGAT [barcode] GTGACTGGAGTTCAGACGTGTGCTCTTCCG-3′
```

Primer Extension for CIM-Seq

Primer extension was carried out with the Apta-Seq primer as described above. cDNA was self-ligated in a 20 µL reaction using CircLigaseII Reaction Buffer (Epicenter), 2.5 mM MnCl2, 50 µM ATP, and five units of CircLigase ssDNA Ligase. Reactions were precipitated with 10 µL of 3 M KCl, 1 µL of glycoblue, 89 µL of H2O, and 300 µL of 98% ethanol, and cDNA was reconstituted in 20 µL of H2O. A polymerase chain reaction (PCR) was performed using 1 µM each of forward and reverse primers, cDNA template, and DreamTaq Master Mix (Thermo Fisher) and amplified for 16 cycles (denaturing 94 °C, 30 s, annealing 55 °C, 30 s, and

elongation 72 °C, 30 s). Amplicons were sequenced on Illumina HiSeq 2500 at the UCI Genomics Facility.

CIM-seq Mapping

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

```
-a CAATGCGTCAAGCTCAG
```

```
-g GATCTGTAATACGACTCACTATAGGGCAGACGTGCCTCACTAC
```

Bowtie2 was used to map the sequences to the reference genome. The parameter `-N 1 --local` were used.

Identification of unmapped reads

Sequences were identified using fastaptamer count and cluster options. Clustering was run with an edit distance of 10 with no filter flag. Top read per cluster was retrieve via grep

```
grep -A 1 '[0-9]- [0-9]\{1,\}-1-0'
```

and aligned manually on Jalview

Supporting information

Other instances of the potential ATP binding motif in the *MUC20*-OT1 intron

>hg38_dna range=chr3:195710200-195711689 strand=+

GTATAGTGAGAATTAAGCCAGGTGGAGAGGCCTGGCTACTGGTGAGTGAGGTAACTGAGATCAGGTAT**GGAGAGGAC**
TGGCTCCTGGTGAGTGAGTTAATTGAGATGAGGTAT**GGAGGGACCTG**TCTCCTGGTGACTGAGTTAATTGAGATGAG
GTGT**GGAGGGACCTG**GCTCCTGGTGAGTGAGTTAATTGAGATCCATTA**GGGAGGGACCTG**GCTCCTGGCGAGTGAGT
TAATTGACACCCGGTAT**GGAGGGACCTG**GCTCCTGGCGAGTGATTTAATTGAGACCCGGTGT**GGAGGGACTTG**GCTC
CTGGTGAGTGAGTGAATTGAGATGAGGTAT**GGAGGGACCTG**TCTCCTGGTGAGTGAGTTAATTGAGATGAGGTGT**GG**
AGGGACCTGTCTCCTGGTGAGTGAGTTAATTGAGATGAGGTGTGGAGGGACCCGGCTCCTGGTGAGTGAGTTAATTG
AGACTCGGTGT**GGAGGGACTTG**GCTCCTGGTGAGTGAGTGAATTGAGATGAGGTAT**GGAGGGACCTG**GCTCCTGGTG
AGTGAGTGAATTGAGATGAGGTGT**GGAGGGACCTG**TCTCCTGGTGAGTGAGTTAATTGAGATGAGGTAT**GGAGGGAC**
CTGTCTCCTGGTGAGTGAGTTAATTGAGATCAGGTAT**GGAGGGACTTG**GCTCCTGGTGAGTGAGTTAATTGAGACTC
TGTGT**GGAGGGACCTG**GCTCCTGGTGAGTGAGTTAATTGAGATGAGGTAT**GGAGGGACCTG**GCTCCTGGTGAGTGAG
TTAATTGAGATCAGGTAT**GGAGGGACCTG**GCTCCTGGTGAGTGAGTTAATTGAGATGAGGTGT**GGAGGGACCTG**GCT
CCTGGTGAGTGAGTTAATTGAGATGAGGTAT**GGAGGGACCTG**GCTCCTGGTGAGTGAGTTAATTGAGATCAGGTAT**G**
GAGGGACCTGGCTCCTGGTGAGTGAGTTAATTGAGATGAGGTAT**GGAGGGACCTG**GCTCCTGGTGAGTGAGTGAATT
GAGATCAGGTAT**GGAGGGACTTG**GCTCCTGGTGAGTGAGTTAATTGAGACTCTGTGT**GGAGGGACCTG**GCTCCTGGT
GAGTGAGTTAATTGAGATGAGGTAT**GGAGGGACCTG**GCTGCTGGTGAGTGAGGTAT**GGAGGGACCTG**GCTCCTGGTG
AGTGAGTTAATTGAGGTCAGGTAT**GGAGGGACCTG**GCTCCTGGTGAGTGAGTTAATTGAGATCAGTTAT**GGAGGGAC**
CTGGCTCCTGGTGAGTGAGTTAATTGAGATGAGGTAT**GGAGGGACCTG**GCTCCTGGTGAGTCTCCT**GGAGGCAGCTG**
CTATCTGGGAACACAGGCACAGGTGGGAACAGACCTTCACTTCCTGCTCACTTAGGTTCAGTGAGTTCTCCAAACCA
GCCTCCCAGGAATGCCATTCAACATGGCTGTGAGGAGAATAAAGAAGAGAGCCTGACTCCTCTCCTGAGGCCCCTTC
CCCACCCTGAGCCAGCAGGATCCAC

Chr	Strand	Genetic loci	Sequence
3	+	*MUC20*-OT1	gtaatacgactcactatagggcagacgtgcctcactacgtTGGCTC CTGGTGAGTGAGTTAATTGAGgTCAGgTAT**GGAGGGACCTG**GCTCC TGGTGAGTGAGTTAATTGAGATcAGGTATGGAGGGACCTGGCTCCT GGTGAGTCTCCT**GGAGGCAGCTG**CTATCTGGGAACACAGGCcgcaa tgcgtc
4	+	*RNF212*	CCTAGAGTGAACCCTAATGTAAACCATGGACGTTGGGTGATAATGA TGTATTCAAGTAGGATCATCAGTTTTAACACATGGACTGCTCTGGT GGAGGATGAGGATGATGGGGGAGCT
14	+	*CINH1*	tGAGGTCAGTGGTCTTGGGAGGTGTGGAGGGTAACTGACTTAAGAT GGGGCATAAGTGAAcGCTCTGGAGGCTACAACTGTTGTGTAcCTTG ATCTGGGT
15	+	*ANAX2*	gtaatacgactcactatagggcagacgtgcctcactacgtTGTGGC CTGATCTGAATCAATATTTGTTTGCTCCCTACTGACTTGTTGCAGC TTCATGCTCTCACAGCCTCTCTCTGCTGGTATCCTGATCTGGGGTT GGAGTGGGTGGAAAGAGACCGGCAGGAAGAATGATGTCACTTGGCA gcaatgc
21	-	ALR/Alpha	gtaatacgactcactatagggcagacgtgcctcactacgtt**gGTAG AAACTG**CAAGGGGATATTTGGAGCGTTTTGTGGTCTATGGTAGAAA AGGATATACTTCACATAAAAATAGAAGCATTCTGAGGAACTTCATG ATGTGTGCATTCATCTCAAAGAGgTGAACTTTTCTTTagcaatgcg tcaagct
22	-	*CSDC2*	gggcagacgtgcctcactacgtGCTTCTTTTTTTGGAAAGTGGCTT TTATGGTCGGGCCcCGGTATAATTTGCCCTGCCCAGTGAGGGCAGC TGCTGAGGAATGCAGGGGA**GGAAGAACCTG**cGACcGCGCCACGTTC CTGGTAATTAGAGAGCAAGTGGGGTGGTGggcaatgcgtcaa

Table S5-1 Additional sequences found in the round 6 pool detected by CIM-Seq. Lower case letter designate nucleotides not matched to the reference human genome. Bold letters indicate potential binding motifs.

Chapter 6 Mini-SELEX

Introduction

It is rare for the entire sequence of an RNA isolated from *in vitro* selection participates in the desired function. This is especially true of selections utilizing longer libraries, which are more likely to yield RNAs with complex folds (Sabeti, Unrau and Bartel 1997). Further biochemical analysis of selected sequences often starts with identification of the minimal motif necessary for function. This is achieved by manual removal of nucleotides and individual *in vitro* confirmation of function. Rational truncations of an RNA, however, require reliable knowledge of the secondary structure which can be informed by phylogeny or chemical probing. Identification of the minimal motif by this process can be tedious and especially elusive when secondary structure data is unavailable and virtually impossible to apply to an entire selected pool. Bioinformatic tools, such as MEME have been developed to help predict motifs by searching for common features in the sequences, but do not perform well with highly diverse pools with multiple rare motifs, which has been the case in genomic SELEX.

Minimization of selected sequences has also been achieved by utilizing *in vitro* selection. Minimal motifs for both the pyrimidine nucleotide synthase ribozyme (Wang and Unrau 2005) and streptavidin DNA aptamer (Bittker, Le and Liu 2002) were inferred by derivatizing each sequence into individual pools by non-homologous recombination, *in vitro* selection of these pools, and comparative analysis of the resulting sequences. The recombinant winners of these selections were compared to each other to find regions common to all populations. This strategy has been limited to pools derived from single sequences, but the SELEX platform can accommodate much higher diversity and is able to screen up to 10^{16} molecules at a time. An already enriched pool, like the ATP binding pool, can be further diversified and selected to

characterize multiple sequences at once without going near the capacity of the SELEX method. Using recombination to identify minimal motifs is not a diversification strategy amenable to genomic selections as recombinant sequences would lose their biological relevance. Instead Mini-SELEX creates differently sized versions of each sequence from truncations on either end. This also isolates the minimized fragments directly instead of producing data used to infer the motif.

Mini-SELEX isolates functional motifs from a selected pool by generating random truncations of each sequence in the pool and subjecting them to additional rounds of selection. This was developed method to minimize selected pools in a multiplexed fashion. While manual truncations usually stop when a sequence reaches activity close to that of the original, the selective pressure in Mini-SELEX favors isolation of optimized motifs. Two truncations strategies are described to encompass both 5′ truncations and 3′ truncations. These can be applied separately in parallel or in series to achieve truncation of both ends. Both of these strategies allow for the constant regions to be switched, which has the added benefit of removing selection artifacts generated by the participation of the constant regions (Shtatland, et al. 2000).

Truncations on the 5′ end of the RNAs are generated by alkaline hydrolysis of the RNA followed by reverse transcription with an oligo composed of the existing 3′ primer with an additional constant region that will serve as a new 5′ primer binding site (See chapter 4). While hydrolysis of RNA generates random fragments, only those with an intact 3′ primer binding region will anneal to the reverse transcription oligo to produce cDNA with a constant 3′ end and variable start sites. Circularization of the cDNA produces a template with two constant regions flanking a variable region now with variable sizes. The 3′ truncations are produced by reverse

transcription using an oligo composed of a randomized (N10) region for random annealing on to RNA and a new orthogonal 3′ primer binding sequence.

Mini-SELEX of the ATP binding pool

The previously selected pool of ATP aptamers contains numerous examples of aptamers that make use of the well-studied ATP binding motif, as well as sequences where no motif was identified. This makes it an ideal pool for method validation while simultaneously prospecting for new potential motifs. The pool was truncated from the 5′ end via alkaline hydrolysis and reverse transcription with an oligo containing the selection reverse primer and designed to introduce a 5′ constant region amenable for Illumina sequencing. The resulting cDNA was circularized and amplified to generate the starting pool for Mini-SELEX. Gel purification was used in the first round of selection to remove most of the full-length sequences from the starting pool to allows for more informative lengths to be selected. 3′ truncation was performed between the 4^{th} and 5^{th} rounds. The selected pool was sequenced and mapped using bowtie2 to a reference genome containing all previously reported sequences as well as a sequence corresponding to the ligation of both primers. Unmapped reads were then clustered using fastaptamer for identification.

Minimization of the ATP binding motif

Reads at least 40 nucleotides long, the length of the minimal motif for artificial ATP binding aptamers (Sassanfar and Szostak 1993), were mapped to previously discovered aptamers, resulting in 1006 reads for *FGD3* aptamer, 120 to our recurring chromosome-15 sequence, and 14 to the ERV1 aptamer. The majority of the *FGD3* mapped reads, 79%, shared a

common 5′ start site 46 nucleotides downstream of the original genomic construct and retains the first three nucleotides from the 3′ constant region. RNAfold is able to predicts the ATP binding structure for this sequence from Mini-SELEX with high confidence, but removal of the 3′ unnatural nucleotides greatly disrupts this prediction. Nonetheless, an *FGD3* aptamer without these three nucleotides was still generated for *in vitro* testing in the hopes of generating a functional genomic construct. Despite the loss of the 3′ nucleotides, 13% of this RNA bound ATP beads and eluted with free ATP. Selection-introduced mutations were removed and 6% of the resulting genomic construct bound ATP linked agarose and eluted with free ATP. This is 10-fold higher than the previously published genomic sequence which binds to ATP cotranscriptionally but showed very weak binding when gel purified and refolded.

As additional rounds of selection had previously favored the ERV1 aptamer sequences post mutagenesis (See Chapter 3), this drop in abundance might not be expected. However, the ATP binding motif in the ERV1 aptamer is very close to its 5′ end, making ATP binding highly sensitive to 5′ cleavage. Moreover, previous data from the *in vitro* selected ERV1 mutants indicate that sections outside of the canonical adenosine-binding motif are quite conserved as well (See Chapter 3). This observation could indicate that these downstream sequences participate in ATP binding and put the ERV1 sequence at a disadvantage when truncated. Those ERV1 sequences that did survive Mini-SELEX were primarily uncleaved sequences.

Most of the reads that mapped to the chromosome 15 sequence of previous chapters correspond a full length read or one starting from a 5′ start site found previously by cleavage interference mapping (Chapter 5). These sequences have been shown to bind the ATP column but do not elute with free ATP. Nonetheless it appears Mini-SELEX has isolated the most active motif.

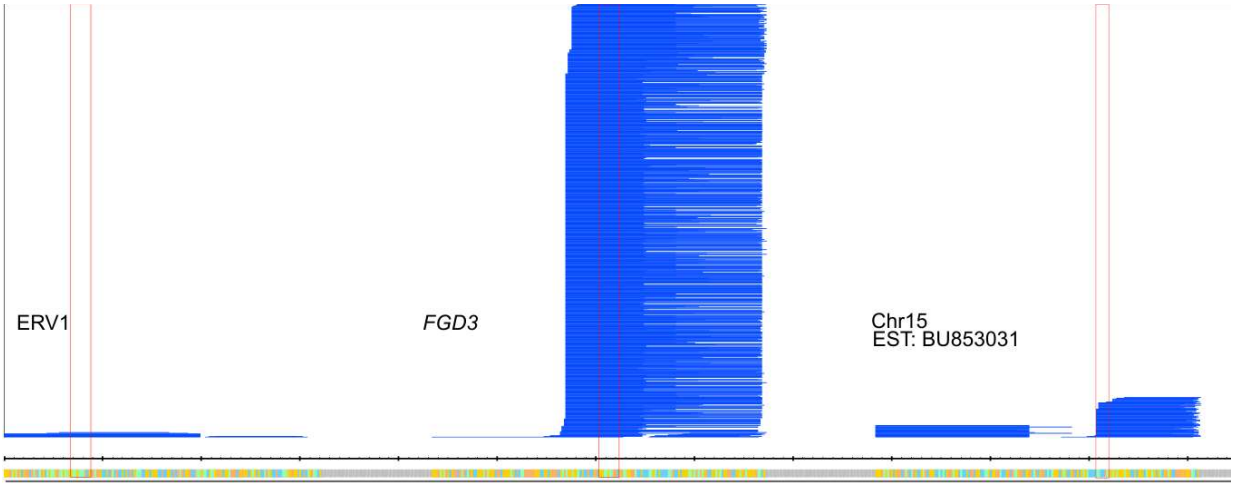


Figure 6-1 Mapping of Mini-SELEX Reads (blue) mapping to the ERV1 *FGD3* and Chr15 reference sequences. Red boxes correspond to the ATP binding loop of ER1 and *FGD3* and the ATP sensitive region of chromosome 15. Height indicates abundance with a maximum of 1000 reads.

These data are consistent with previous biochemical characterization of these RNAs. The remaining sequences previously discussed in this dissertation where not initially detected in this experiment. Referring back to the CIM-seq (See Chapter 5), it is apparent that many of these sequences were detected in the flowthrough but very few were detected in the elution fraction (which had lower reads in general) which makes them easily selected away. More permissive mapping revealed that at least a few of these sequences did survive the minimization.

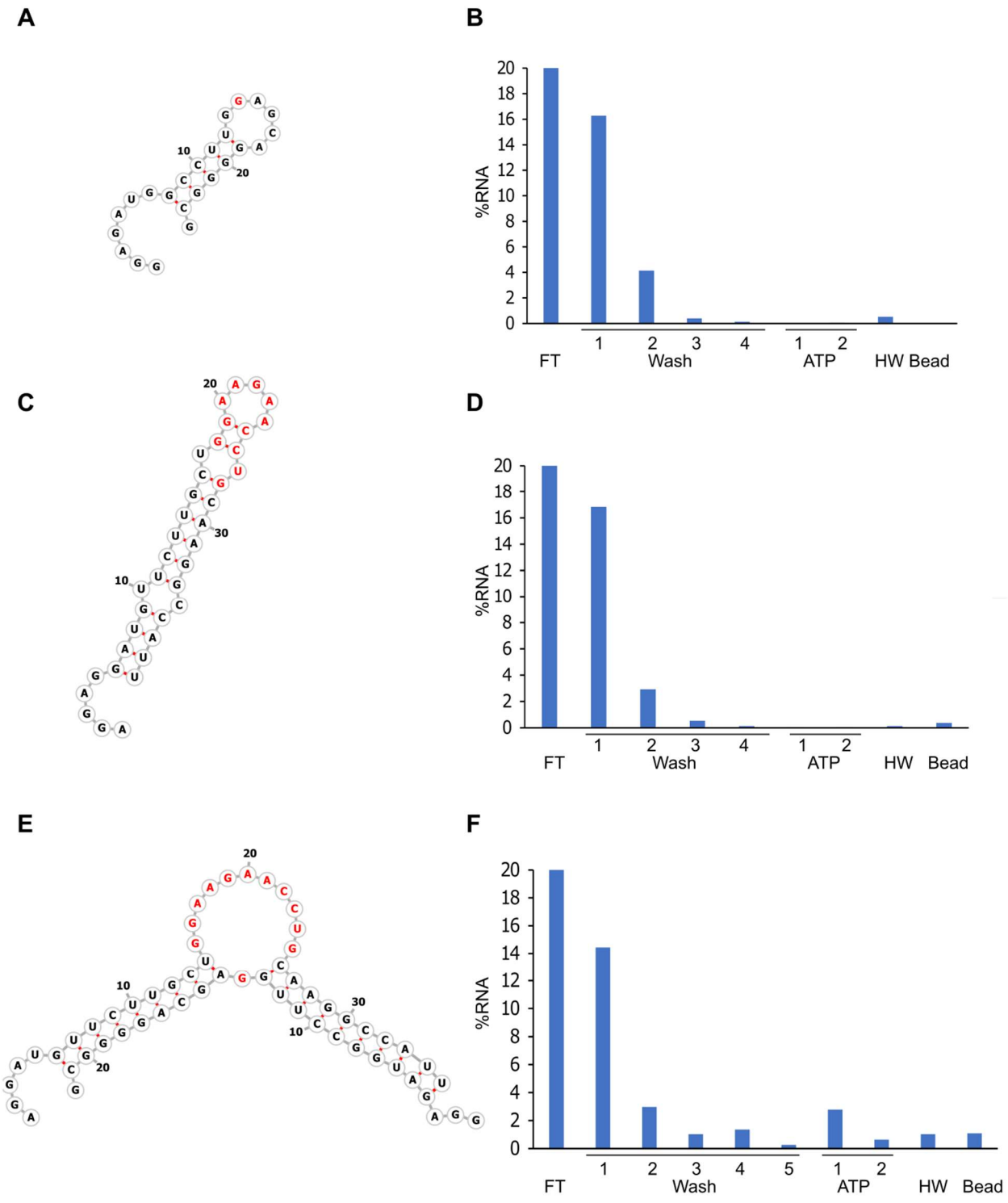
Delving deeper into the sequencing

Sequences shorter than 40 nts were thought less likely to yield function motifs, but because the majority of the sequences fell below the 40 nt size threshold for mapping, a second more permissive mapping was performed, revealing sequences that primarily mapped to *FGD3* (~7000 reads), ERV1 (~300 reads), the chromosome 15 sequence (~1800 reads), and *CSDC2* (~1000) selected sequences. Table S6.1 reports regions selected and depth of those regions in this mapping. This mapping is likely too permissive of short reads to highlight functional motifs

and is not part of motif discovery by Mini-SELEX. It is performed to evaluate whether the initial mapping was too stringent and that unmapped sequences don't correspond to useful human genomic reads. In this particular instance, a deeper look into these often discarded sequences yielded interesting results for further examination.

Two of the top three represented regions mapped to two separate portions of *FGD3*. One population contained the 11 nt consensus sequences with flanking nucleotides and the other population mapped to the predicted opposing strand containing the opposing G and flanking nucleotides. These fragments do not bind ATP on their own but do when they are both present, indicating they likely survived selection as an obligate heterodimer. While this is an interesting phenomenon, this result could complicate analysis of selections in general as RNAs do not have to be independently capable of the selected function. However, when the monomers of a dimeric aptamer are evaluated as single isolates, they will then appear inactive.

Figure 6-2 Split *FGD3* aptamers (**A**) Strand of the split aptamer from selection with opposing G in red (**B**) Graph of column binding fractions for this RNA showing no binding. (**C**) Strand of the split aptamer from selection containing the 11 nt recognition loop with consensus sequence displayed in red. (**D**) Graph of column binding fractions for this RNA showing no binding. (**E**) Proposed structure of the heterodimeric ATP aptamer (**F**) Graph of column binding fractions for this RNA showing ~3% binding.


A
B
%RNA
FT
Wash
ATP
HW
Bead
C
D
%RNA
FT
Wash
ATP
HW
Bead
E
F
%RNA
FT
Wash
ATP
HW
Bead

The ERV1 aptamer also saw enrichment of multiple small fragments of its sequences. While one fragment contained the 11-nt consensus loop, the other was not the segment thought to bind across from this loop, but part of the downstream region which is predicted to form an independent structure. These fragments can, however, form a nine base pair stem together which would actually sequester the ATP binding loop.

The remaining short sequences are primarily capable of forming short hairpins (See supplementary info) and are reported in Table S6-1. Many of them can also bind to each-other as most have pyrimidine rich regions complementary to the purine rich regions common to many of the fragments. The base pairing network available to these short fragments may account for their survival in selection as they can form hairpins, dsRNA, or RNA-DNA duplexes to facilitate transcription. Reverse transcription and PCR amplification also both favor short RNAs.

Identification of unmapped reads

Identification of unmapped reads is usually done to evaluate the mapping and potential sequences not included in the reference genome, which was how the *CSDC2*, *MUC20*-OT1 and ALR/Alpha ATP aptamers were discovered (See Chapter 5). These unmapped reads are first counted and clustered with fastaptamer for ease of identification and manually aligned on Jalview. Among the unmapped sequences was a cluster of sequences that correspond to two fragments of the *FGD3* aptamer ligated together. The ligation products are thought to have arose from nucleophilic attack of the 5′ hydroxyl of one fragment to the electrophilic 2′, 3′ cyclic phosphate of another upon EtOH precipitation which occurs nonenzymatically in dehydrating conditions (Costanzo, et al. 2016). This represents a fortuitous example of internal loop minimization not intended in the minimization design. More examples of these ligation products

were observed in the unselected truncated pool, including ligations between sequences originating from separate genetic loci. These ligations are rare events, demonstrating the ability of SELEX to amplify very rare species. A construct was made to match the most abundant of these internally-minimized *FGD3* aptamers and tested for binding on an ATP-agarose column. This *FGD3-derived* aptamer, which is not likely to exist biologically, actually displayed the strongest binding of any *FGD3* sequence tested. This extraordinary binding could have proved vital in the survival of what most likely were very rare sequences in the pool and also supports the hypothesis that the genomic sequence that flanks the natural *FGD3* aptamer weakens its affinity to ATP. A more modulated, weaker binding is likely to be necessary for the aptamer to be at all sensitive to ATP concentrations in human cells and this aptamer's affinity may be tuned to those concentrations.

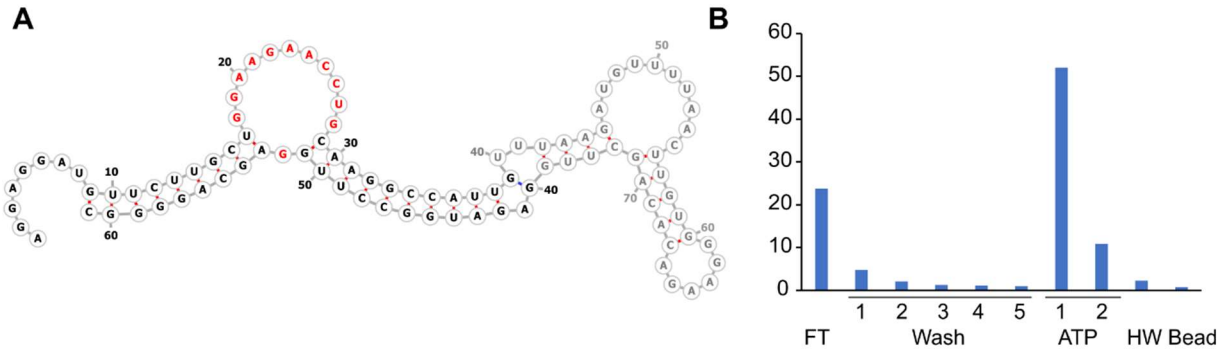


Figure 6-3 *FGD3* based aptamer (**A**) Proposed secondary structure of the aptamer in agreement with thermodynamic prediction. ATP binding motif is displayed in red. Removed nucleotides from the original *FGD3* sequence is shown in gray. New phosphodiester bond shown in blue. (**B**) Graph of column binding fractions displaying ~60% of RNA binding and eluting in 5 mM ATP.

Conclusion

The Mini SELEX method was able to define the minimal motifs for the *FGD3* sequence while providing evidence that ATP binding by the ERV1 aptamer is more complex than originally thought. The method allows for multiplexed motif characterization, though some of these sequences may die off as more fit sequences take their place in each generation. This particular instance was flooded with short sequences that are likely to be favored in the amplification process. Like full-length sequences, the short sequences can be suppressed in future applications with gel purification. They don't directly prohibit the identification of motifs, as they will often fall below minimum mapping thresholds, but they can take up valuable read depth.

Internal minimization was not designed into this method, as internally minimized sequence lack biological relevance in genomic SELEX, but it is an intriguing development that points to ATP binding inhibition by the internal stem loop flanked by the *FGD3* ATP binding motif. The observed recombination events are random as evidenced by chimeric sequences detected in the CIM-seq, and the likelihood that recombination will occur between two sequences mapping to the same genomic location are based entirely on the abundance of each in the pool.

The isolation of a split *FGD3* aptamers was another interesting find. Split aptamers are often used for molecular sensors, are rationally designed from existing aptamers, not isolated by *in vitro* selection. This particular split ATP aptamer is likely too weak to be useful for diagnostic applications. This discovery does however have greater implication on evaluating selection results in general. Perhaps sequences tested in selection that do not seem to be functional are

carried through as dimers or even trimers. This hypothesis has led to testing seemingly inactive sequences in the presence of the parent pool in several selection in our lab but has not had a concrete example where a previously inactive sequence was found to bind in a pool dependent manner.

These minimization strategies described have since been used to minimize selected pools for cGMP, luciferase, Oregon Green, and nickel binding by myself and other members of our lab, however, the results of these are not yet ready for presentation.

Materials and Methods

RNA Transcription

RNA was transcribed at 37 °C for 1 to 3 h in a volume of 20 µL containing 40 mM Tris chloride, 10% dimethyl sulfoxide (DMSO), 10 mM dithiothreitol (DTT), 2 mM spermidine, 2.5 mM each CTP, GTP, and UTP, 250 µM ATP, 2.25 µCi [α-^{32}P]-ATP (Perkin Elmer, Waltham, MA, USA), 25 mM $MgCl_2$, one unit of T7 RNA polymerase, and 0.2 µM of DNA template. DMSO was used to increase transcript yields, as documented in previous studies (Chen and Zhang, 2005). The transcripts were purified using denaturing PAGE.

In vitro Selection

The DNA pool used for the *in vitro* selection was derived from the human genome and described previously (Salehi-Ashtiani et al., 2006). Purified RNA transcripts were precipitated, dried, and resuspended in 200 µL binding buffer containing 140 mM KCl, 10 mM NaCl, 10 mM Tris chloride, pH 7.5, and 5 mM $MgCl_2$ and heated to 70 °C before loading on to C8-linked ATP-

agarose beads (Sigma-Aldrich, St. Louis, MO, USA) equilibrated in the binding buffer. Flowthrough was collected after the columns were capped and shaken for 20 min at room temperature. The beads were washed with 200 µL of binding buffer, and potential aptamers eluted with the same buffer supplemented by 5 mM ATP·Mg with 30 min of shaking at room temperature. Each fraction was analyzed for radioactivity using a liquid scintillation counter. Elutions were pooled, desalted using YM-10 spin filters (Millipore, Billerica, MA, USA), precipitated, dried, and resuspended in H2O.

In vitro Selection Primers

Forward: `5′ TAGATCTTAATACGACTCACTATAGGGAGACACTCTTTCCCTACACGACGCTCTTCCGATCT 3′`

Reverse: `5′ CTGAGCTTGACGCATTG 3′`

Reverse Transcription

RNA was reverse transcribed in 20 µL using the Promega reverse transcription buffer, 2 µM reverse primer, and RNA recovered from the previous selection round. The RNA and primer were annealed by heating at 65 °C and cooling to room temperature before 1 unit of Thermoscript (Invitrogen, Grand Island, NY, USA) and Improm II (Promega, Madison, WI, USA) reverse transcriptases each were added. The reaction was initiated for 5 min at 25°C, and then the temperature was ramped to 42°C, 50°C, 55°C, and 65 °C for 15 min each before the enzymes were inactivated at 85 °C for 5 min.

Amplification

DNA was amplified by using DreamTaq Master Mix (Fermentas, Glen Burnie, MD, USA), 2 μM forward primer, 2 μM reverse primer, and DNA from reverse transcription. DNA was initially denatured at 95 °C for 1.5 min (30 s for subsequent denaturing steps), annealed at 55 °C for 30 s, and extended at 72 °C for each cycle. Optimum number of PCR cycles was determined for each selection round by comparing 8-, 12-, 16-, and 20-cycle aliquots on agarose gel.

Alkaline Hydrolysis

Round 6 of an *in vitro* selection for an ATP aptamer from a human genomic library was hydrolyzed for 2, 5, and 10 minutes in 50 mM $NaHCO_3$ and 1 mM EDTA at 72 °*C* pH 10. Aliquots of each fraction were run on a gel and the 5 minute hydrolysis was chosen as the RNA with the greatest distribution of RNAs.

Primer Design

Round 6 of an *in vitro* selection for an ATP aptamer from a human genomic library (Vu, et al. 2012) was used as the library for CIM-Seq. Primers were used as described in Apta-seq (Abdelsayed, et al. 2017).

Apta-Seq primers contain a reverse primer sequence for the pool of interest and flanking Illumina primers in order to barcode and sequence primer extensions by Illumina Sequencing. Apta-Seq primers for reverse transcription primer extension were designed 5′ to 3′ with the following components:

Illumina forward primer reverse-complement, NotI digestion site, Illumina reverse primer, reverse primer for RNA of interest

```
5′-
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTGCGGCCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCTG
AGCTTGACGCA-3′
```

Primers for amplification were designed 5′ to 3′ with the following components:

Forward primer containing Illumina forward adapter and primer.

`5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′`

Reverse primer containing Illumina reverse adapter, barcode, and Illumina reverse primer.

`5′-CAAGCAGAAGACGGCATACGAGAT [barcode] GTGACTGGAGTTCAGACGTGTGCTCTTCCG-3′`

Primer Extension

Primer extension was carried out with the Apta-Seq primer as described above. cDNA was self-ligated in a 20 μL reaction using CircLigaseII Reaction Buffer (Epicenter), 2.5 mM MnCl2, 50 μM ATP, and five units of CircLigase ssDNA Ligase. Reactions were precipitated with 10 μL of 3 M KCl, 1 μL of glycoblue, 89 μL of H2O, and 300 μL of 98% ethanol, and cDNA was reconstituted in 20 μL of H2O. A polymerase chain reaction (PCR) was performed using 1 μM each of forward and reverse primers, cDNA template, and DreamTaq Master Mix (Thermo Fisher) and amplified for 16 cycles (denaturing 94 °C, 30 s, annealing 55 °C, 30 s, and elongation 72 °C, 30 s).

Mapping of reads

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

`-a CAATGCGTCAAGCTCAG -g GATCTGTAATACGACTCACTATAGGGCAGACGTGCCTCACTAC -m 40`

Bowtie2 was used to map the sequences to the reference genome. The parameter `-N 1 --local` were used. For more permissive mapping no minimum length was used during clipping and bowtie1 scoring options were adjusted to `-min_score G,40, 8`

Identification of unmapped reads

Sequences were identified using fastaptamer count and cluster options. Clustering was run with an edit distance of 10 with no filter flag. Top read per cluster was retrieve via grep

```
grep -A 1 '[0-9]-[0-9]\{1,\}-1-0'
```

and aligned manually on Jalview

Chr	Strand	Genetic loci	Depth	Sequence
2	+/-	MIR4435-2HG	5	**ggggccagacgtgcctcactacGTGGCAATGATG**AGACCAGCTG CTCATCCTGGGttTGGTCATTGTGACTCAGATGTGGAGTGGGCT GGTTAGTGGTTCTGGGtCACtTTGTCCATCTCTGGATCTGGCAG CTCCATCTACACAGTGTCGGGCTAG
2	-	LTR81B	9	gtaatacgactcactatagggcagacgtgcctcactacgtTGGG AGAGGAGAATATGTCTTCAAGAAGTGACAGTCCCAAACACCTCT TGTTCCGGGGTGGTTTGATGGGTGGAGAAAGGACAGGGGAAGAT GGAGAGTCTCTTCTTGTTTCTTCTGT**AACTGTTCATGGTGGgca atgc**
3	+	*MUC20*-OT1	1	gtaatacgactcactatagggcagacgtgcctcactacgtTGGC TCCTGGTGAGTGAGTTAATTGAGgTCAGgTATGGAGGGACCTGG CTCCTGGTGAGTGAGTTAATTGAGATcAGGTATGGAGGGACCTG GCTCCTGGTGAGTCTCCTGGAGGC**AGCTGCTATCTGGGAACACA GGCc**gcaatgcgtc
15	+	ANAX2	6	gtaatacgactcactatagggcagacgtgcctcactacgtTGTG GCCTGATCTGAATCAATATTTGTTTGCTCCCTACTGACTTGTTG CAGCTTCATGCTCTCACAGCCTCTCTCTGCTGGTATCCTGATCT G**GGGTTGGAGTGGGTGGAAAGAGACCGGCAGGAA**GAATGATGTC ACTTGGCAgcaatgc
22	-	*CSDC2*	851	gggcagacgtgcctcactacgtGCTTCTTTTTTTGGAAAGTGGC TTTTATGGTCGGGCCcCGGTATAATTTGCCCTGCCCAGTGAGGG CAGCTGCTGAGGAATGCAGGGGAGGAAGAACCTGcGACcGCGCC ACGTTCCTGGT**AATTAGAGAGCAAGTGGGGTGGTGgg**caatgcg tcaa
15	+	EST BU853031	1326	gggcctcactacgtTGACTGGGTTATATGCCAGGTATTGTCCTG GTTAAGTGGCCATTGCTCAAGCCTGgCTACTGCCTATAGGGTGA GTTGCcGTGAGATTTTtAATTCACGAGaAGGGAGGTTGTG**GGTA GGAGTGATGCAGTTCCCGCGg**
			144	gggcctcactacgtTGACTGGGTTATATGCCAGGTATTGTCCTG GTTAAGTGGCCATTGCTCAAGCCTGgCTACTGCCTATAGGGTGA GTTGCcGTGAGATTTTt**AATTCACGAGaAGGGAGGTTGTGGGTA GGAGTGATGCAGTTCCCGCGg**
3	-	ERV1	188	gggcagacgtgcctcactacgtAGAAGCAAGAAGGAAGAA AcTGCCAcCGTTGTCATCTTGTGTCTGaAATTGGTGGGTT CTTGGTCTCACTGgCTTCAAGAATGAAGCCGTGGACCCTC GCGGTGAGTG**TTACAGTTCTTAAAGGCGGTGTCTag**caat gc
			70	gggcagacgtgcctcactacgt**AGAAGCAAGAAGGAAGAA AcTGCCAc**CGTTGTCATCTTGTGTCTGaAATTGGTGGGTT CTTGGTCTCACTGgCTTCAAGAATGAAGCCGTGGACCCTC GCGGTGAGTGTTACAGTTCTTAAAGGCGGTGTCTagcaat gc
9	+	*FGD3*	3837	gggcagacgtgcctcactacgtTGGGACAGGGTCCTGGGGAGTC TAACCCCCGGGTTGGTGCCCTGAC**AGGAGgATGTTCTtGCTGGA AGAACCTGCAAGGCC**ATTGTTTAAgATGTTTTAACTTGTGGGAA GACACAGCTTGGAGATGGCCTTGGAGCAGGGgcaatgc
			1667	gggcagacgtgcctcactacgtTGGGACAGGGTCCTGGGGAGTC TAACCCCCGGGTTGGTGCCCTGACAGGAGgATGTTCTtGCTGGA AGAACCTGCAAGGCCATTGTTTAAgATGTTTTAACTTGTGGGAA GACACAGCTTG**GAGATGGCCTTGGAGCAGGGgc**aatgc
			800	gggcagacgtgcctcactacgtTGGGACAGGGTCCTGGGGAGTC TAACCCCCGGGTTGGTGCCCTGAC**AGGAGgATGTTCTtGCTGGA AGAACCTGCAAGGCCATTGTTTAAgATGTTTTAACTTGTGGGAA GACACAGCTTGGAGATGGCCTTGGAGCAGGGgca**atgc

Table S6-1 Sequences detected in the round 5 Mini-SELEX pool. Full parent sequences are provided in normal font. Bold regions were detected in the Mini-SELEX pool

Hairpin formation in small fragments

Chapter 7 -An ATP utilizing human ribozyme

Introduction

Ribozymes are functional RNA molecules capable of catalyzing enzymatic reactions, much like protein enzymes, and were a surprising development to the world of biochemistry that started with the discovery of self-sufficient splicing of what is now called a group I intron in *Tetrahymena* (Kruger, et al. 1982). It would later be revealed that the enzyme responsible for the synthesis of all the other enzymes, the ribosome itself, was also a ribozyme. Since then, ribozymes, both naturally occurring and synthetic, have been discovered at an accelerated rate. We have even learned that human ribozymes are not limited to the ribosome with the discovery of self-cleaving RNA in humans and these RNA appear to be products of recent evolution (Salehi-Ashtiani, et al. 2006).

Over the course of the genomic SELEX for ATP aptamers (See Chapter 2), RNA species were observed that remained bound to an ATP-agarose matrix even after denaturation in 7 M urea and 20 mM EDTA. Because EDTA chelates the Mg^{2+} ions that are often necessary for RNA folding and urea destabilizes helix formation, this behavior is unlikely to arise from non-covalent interactions. However, if the human genome codes for an RNA capable of reacting with ATP as its substrate, an ATP-utilizing ribozyme, it would likely withstand harsh denaturing conditions. Thus, it was hypothesized, based on the preliminary results, that the human genome codes for a ribozyme that uses ATP as its substrate accounting for the covalent linkage. To identify and characterize these RNAs, an *in vitro* selection was designed to enrich potential ribozymes by capturing them mid-catalytic cycle.

Naturally occurring ribozymes primarily catalyze transfer of amino acyl or phosphoryl groups (Doudna and Cech 2002) so a reaction involving ATP is not surprising. Though not yet observed in natural ribozymes, one such potential function for the ATP utilizing ribozyme includes 5′ capping. A 5′ capping ribozyme was first found while selecting for an amino acid-activating ribozyme. Adapting to unexpected results from early rounds of selection, RNAs were instead enriched for denaturing resistant bond formation to UTP-agarose (Huang and Yarus 1997) resulting in and RNA that could catalyzed the formation of a 5′-5′ cap using a variety of triphosphates. A similar ribozymes appeared in a subsequent selection designed to enrich for an RNA polymerase ribozyme (Zaher, Watkins and Unrau 2006). The genomic SELEX discovery strategy has already been applied to natural ribozymes discovery for two types of ribozymes; four self-cleaving ribozymes (Salehi-Ashtiani, et al. 2006) found in the same human genomic pool as the hypothesized ATP utilizing ribozyme, and a self-alkylating ribozyme in *A. pernix* (McDonald, et al. 2014). Enrichment of a specific ribozyme requires some design in the method of collection. Human self-cleaving ribozymes were required to cleave from concatemeric RNA and retrieved by size (Salehi-Ashtiani, et al. 2006). *A.pernix* self-alkylating RNAs were selected by allowing RNAs to react with electrophilic small molecules attached to a biotin handle (McDonald, et al. 2014). Limited to the knowledge that the novel potential ribozyme in humans is covalently linked to beads at some point and survived the ATP binding selection, this selection required that the RNAs react to ATP immobilized on beads and remain bound after several denaturing washes. Two methods of collection were used; bound RNA was allowed to react again with ATP in solution, potentially freeing itself for collection or reverse transcribed directly on the beads themselves.

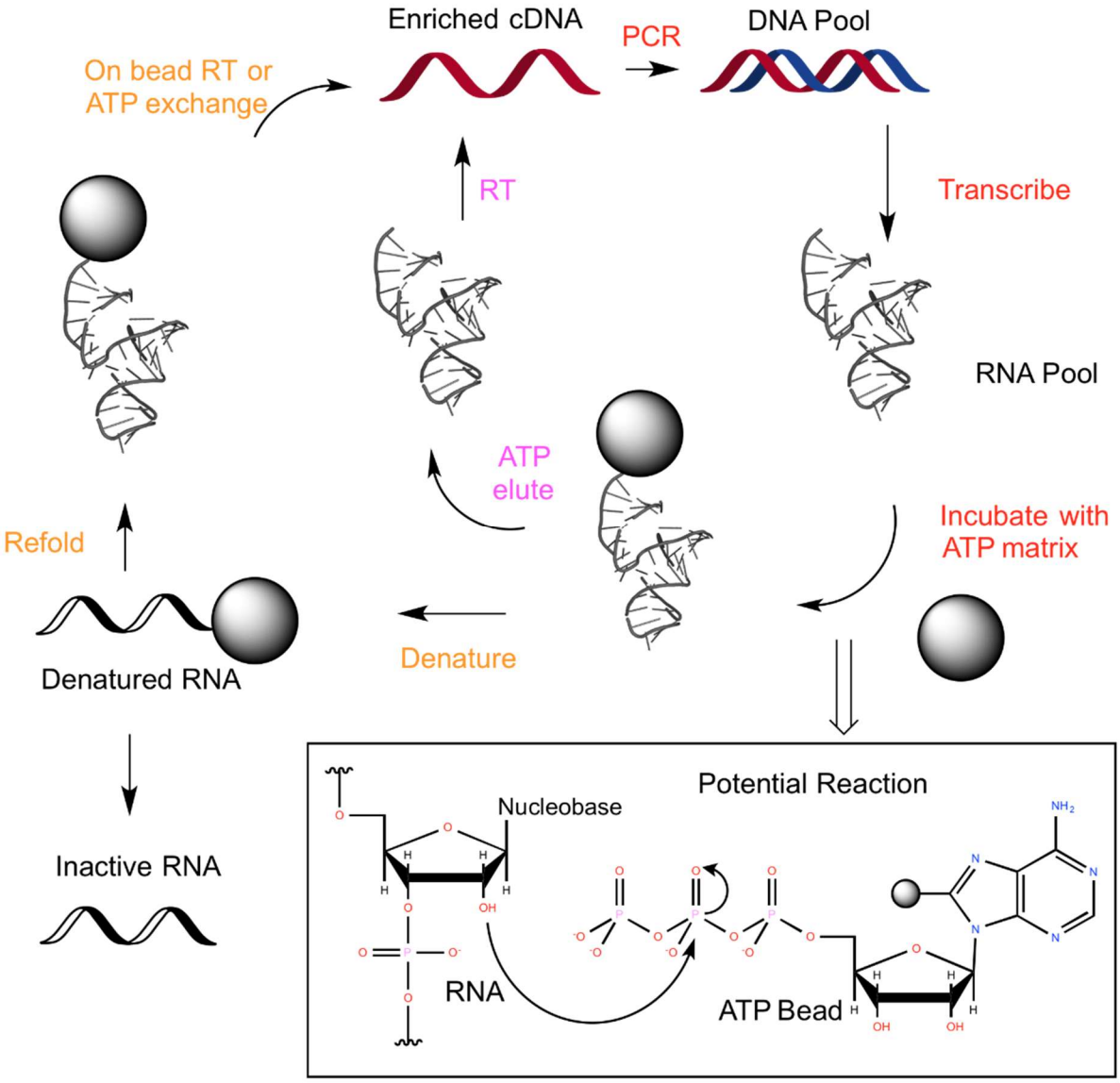


Figure 7-1 Selection Schemes. (A) RNA transcribed from the DNA pool will be incubated with ATP linked to a solid support for binding. From there, the selections diverge in technique. The aptamer selection (SA) competitively elutes potential aptamers with free ATP and reverse transcribed to yield cDNA following the magenta path. The ribozyme selections follow the orange path. The RNA-ATP reaction mixture will be denatured to remove RNA not covalently attached. The RNA will then be refolded to allow elution by substrate exchange (selection SX) or on bead reverse transcription (selection SB). The cDNA obtained from all selections returns to the red path and will be amplified to produce an enriched pool of DNA that can be used for subsequent rounds of selection repeating the cycle. **(B)** Potential reaction. A covalent linkage could result from nucleophilic attack of a ribose hydroxyl with any one of the electrophilic phosphates of ATP

Genomic selection for ATP reactive ribozymes

The RNAs transcribed from a pool of human genomic DNA were incubated with agarose-linked ATP, attaching the reactive RNA to the solid matrix. Washes with urea and EDTA abolish tertiary structure and remove non-covalently attached RNA from the ATP matrix. Presumably only RNAs with a covalent linkage to ATP remain and are refolded in the original binding buffer. Two strategies were initially used to collect and amplify the active sequences (See Fig 7-1): 1) If the RNAs linked to beads are capable of multiple turnovers, they exchange the bead-bound ATP for free ATP and are eluted from the beads to be collected for RT-PCR and subsequent rounds of selection or sequencing (selection SX); 2) in parallel, RT-PCR is performed directly on the agarose matrix to amplify sequences remained bound to ATP after denaturation and refolding (selection SB).

Rounds selected	% On Bead	% Exchanged	% Total	Estimated Diversity
0	0.06%	0.02%	0.08%	~10^9
1	0.51%	1.66%	2.17%	~10^6
2	0.68%	0.33%	1.01%	~10^5
3	0.63%	0.77%	1.40%	~10^3
4	0.35%	11.65%	12%	~10^1

Table 7-1 – *SB Selection summary* % RNA in populations of interest are shown with estimated diversity. Enrichment is observed in the exchanging population by round 4.

Four rounds from selection SB produced low on-bead retention but enriched the population of RNA that exchanges with free ATP (Table 1). Attempts to push the selection onto further rounds failed to yield amplifiable full length cDNA from either fraction. PCR of collected fractions instead produced smaller than expected products. The multiple attends, however, successfully demonstrated both reproducible resilience to denaturing and substrate exchange upon refolding. This selection was halted with an average of 12% RNA remaining after denaturing over the 4 trials (See Fig 7-2) as the DNA itself for this pool also

contained these shorter amplifiable sequences, which would thwart attempts to PCR even after attempted purifications.

The selection collecting RNAs that exchanged (SX) failed to amplify after two trials. This collection strategy was reutilized in response to the result of selection SB. This new selection (SBX) started from the round 4 pool of selection SB but actually displayed a decline in activity over the next two rounds and was aborted (Table S1). The SB round 4 pool provided much more resistance to cloning attempts than previous selection pools. The two clones eventually isolated were successfully sequenced mapped to the same location as the ATP aptamers first two ATP aptamers found previously (Chapter 2). When the *FGD3* and ERV1 sequences were incubated on the ATP matrix column, they had only 2.7% and 1.5% of total RNA remaining on beads after denaturing (Table S2) which would not account for the activity observed by the pool.

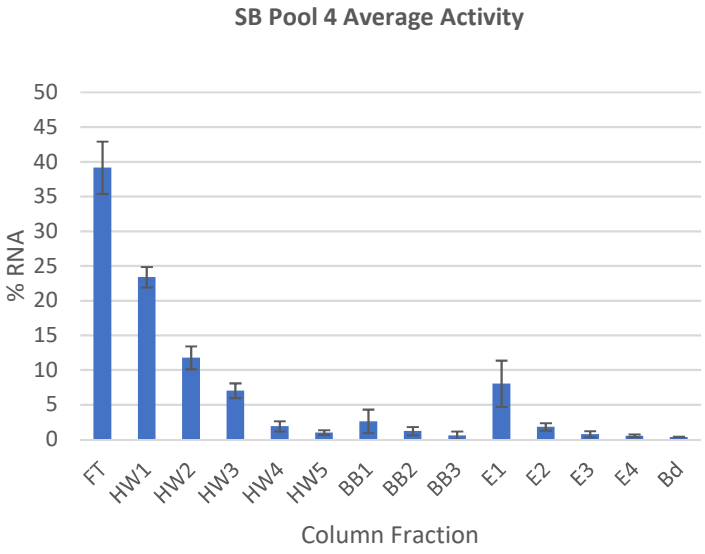


Figure 7-2 - SB selection round 4 column analysis. While the matrix retained 60% of the RNA after flowthrough (FT). Roughly 12 % of the RNA remained after denaturing (HW) and refolding (BB). The majority of the surviving RNA was observed exchanging (E) while few remained on bead (Bd)

Continued ribozyme selection

Repeated biochemical experimentation and cloning attempts require PCR to regenerate the DNA pool, however, amplification of the round 4 pool of SB generates amplicons shorter than the expected pool size. These amplicons PCR much more efficiently than the full-length pool and return when the full length pool is gel purified and amplified. This eventually resulted in the loss of the pool. These short amplicons are not thought to be the active species, as the pool is gel purified before reacting with ATP. Several attempts to reproduce the active pool included selecting from previous rounds, regenerating cDNA from frozen transcripts, and reamplification of the pool from samples prepared for sequencing.

An additional collection strategy was added in the form of elution with pyrophosphate. Since ATP exchange points to a reversible reaction, the ATP-RNA reaction likely creates a pyrophosphate product. Incubating with pyrophosphate could drive the reaction backwards, releasing the RNA from the ATP. DNA recovered from any of these three strategies where carried on to the next round and these additional efforts generated a pool that underwent four more rounds of selection. The latest rounds from each of the previously described selection (SB, SX, and SBX) were also combined and *in vitro* selected for two additional rounds using these conditions (SBC).

The SB round 8 pool was sequenced at very higher depth and when mapped to the human genome contained thousands of peaks. To narrow down the focus of candidates, the sequences were counted and clustered via fastaptamer. A representative sequence from the top 100 clusters were retrieved using a grep command taking advantage of fastaptamer's generated header. These sequences ranged in depth from 102-307,106 reads and contain several GT rich repeats. The

most abundant sequences mapped to the (-) strand of chr 12 in the intron of a predicted gene, to the (-) strand of the 4th intron of the *TNS3* gene overlapping with a simple (AC)n repeat, to the (-) strand of 19th intron of the *GAK* gene and (CACA)n repeat, and a portion of the 4th intron of a long intergenic noncoding RNA on chromosome 13 which overlaps with both a HUERV-p1-int LTR and a simple (TG)n repeat on the (+) strand. Many of the clustered sequences were shorter than the starting pool and closer examination revealed many were missing internal regions of the genomic sequence. Each of the abundant sequences had several clusters among the top 100 representing them, most with a uniform 5′ within the cluster but variable lengths and 3′ ends. Mapping of these sequences via BLAT reveal that the 5′ and 3′ usually map to the same gene, but with a 10-100 nucleotide gap. Alignment of these sequences to each other confirms several sequences seem to be missing internal regions and some sequences even had 3′ ends originating from other sequences in the pool. The clustered top 100 sequences also contained the *FGD3* and ERV1 aptamers and the mammalian conserved chromosome 15 sequence both as full length sequences and as seemingly spliced products but in lower abundance than the other reported sequences. These three sequences in addition to the other previously discovered aptamers have all been shown to bind to immobilized ATP, but none saw retention of RNA on beads under denaturing conditions.

Ribozyme or splicing substrate?

The chromosome 12 sequence was the most abundant full-length sequence. It maps to a the 3rd intron of Geneid predicted gene, chr12_1170, on the (-) strand ~1,800 nucleotide downstream of the 3rd exon. When washed in native conditions ~5% remain linked to ATP beads and under denaturing conditions 2% remain. This is not significantly higher than observed in ATP aptamers tested thus far which have ranged between ~ 0.5-3% RNA remaining

on beads after denaturing. Several examples of splicing were detected for this sequence. Most of these sequences seemed to be spliced at the base or bulges of predicted stems, though the repetitive nature of sequence leaves the precise segment removed ambiguous for some of these variants. There is also an example of 3′ end of the *FGD3* aptamer being spliced to the 5′ end of the chr12_1170 aptamer. This *FGD3* fragment includes the half of the ATP binding motif which hold the opposing G and starts from the internal loop.

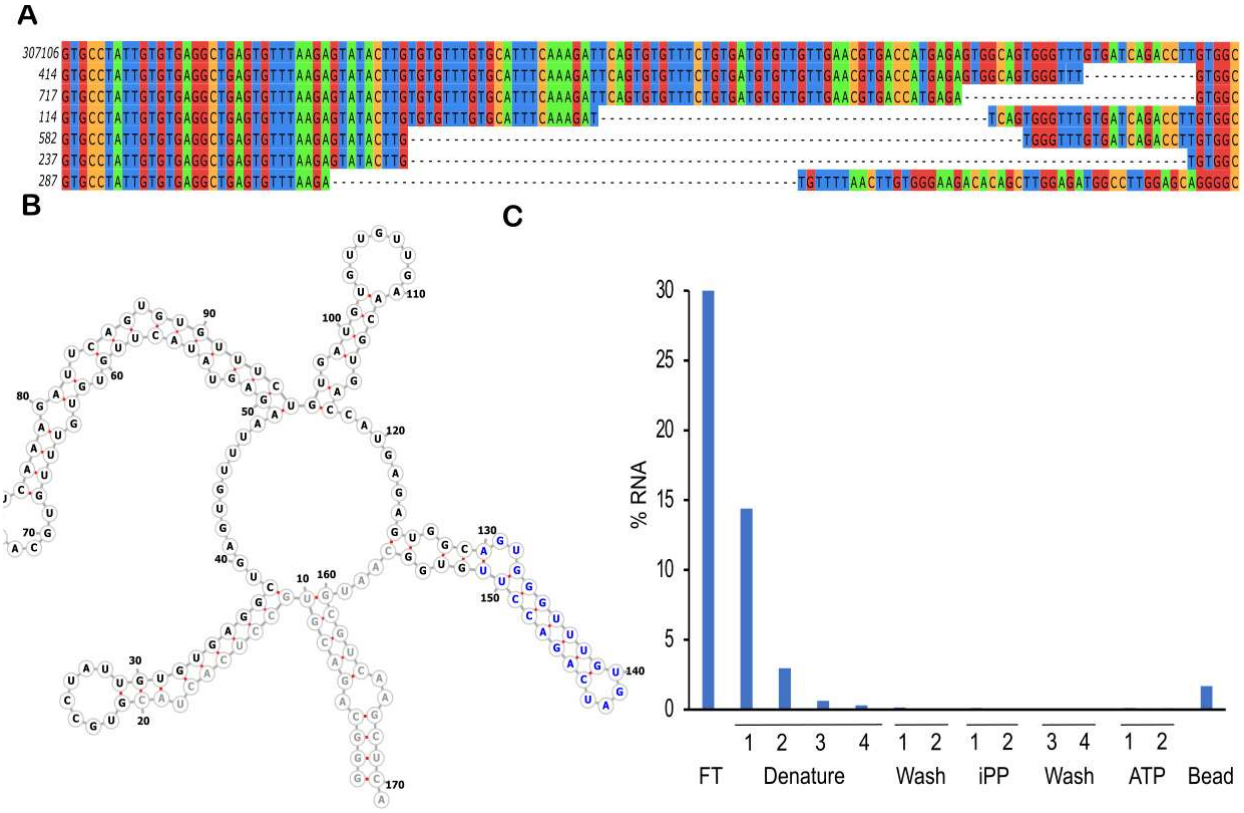
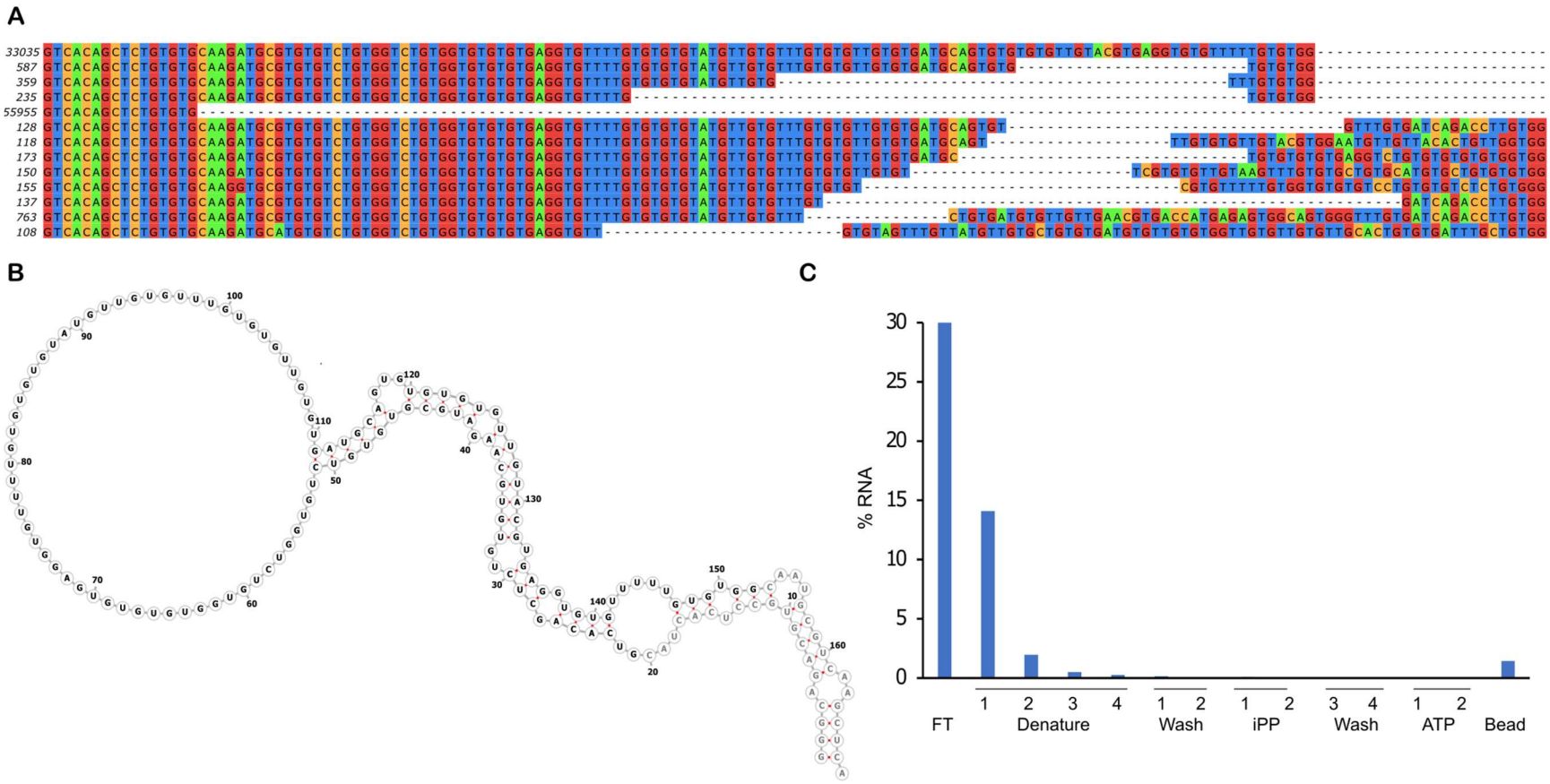


Figure 7-3 chr12_1170 intronic RNA (**A**) Alignment of sequences representing cluster from the chromosome 12 sequence. Number on the left indicate abundance of that sequence. The bottom sequence contains a 3′ end that instead correspond to the *FGD3* sequence (**B**) Thermodynamic prediction of the secondary structure of the sequence mapped to chr12_1170. Most commonly removed nucleotides are shown in blue. Artificial primer sequence is shown in gray (C) Graph of column binding fraction displaying ~2% binding.

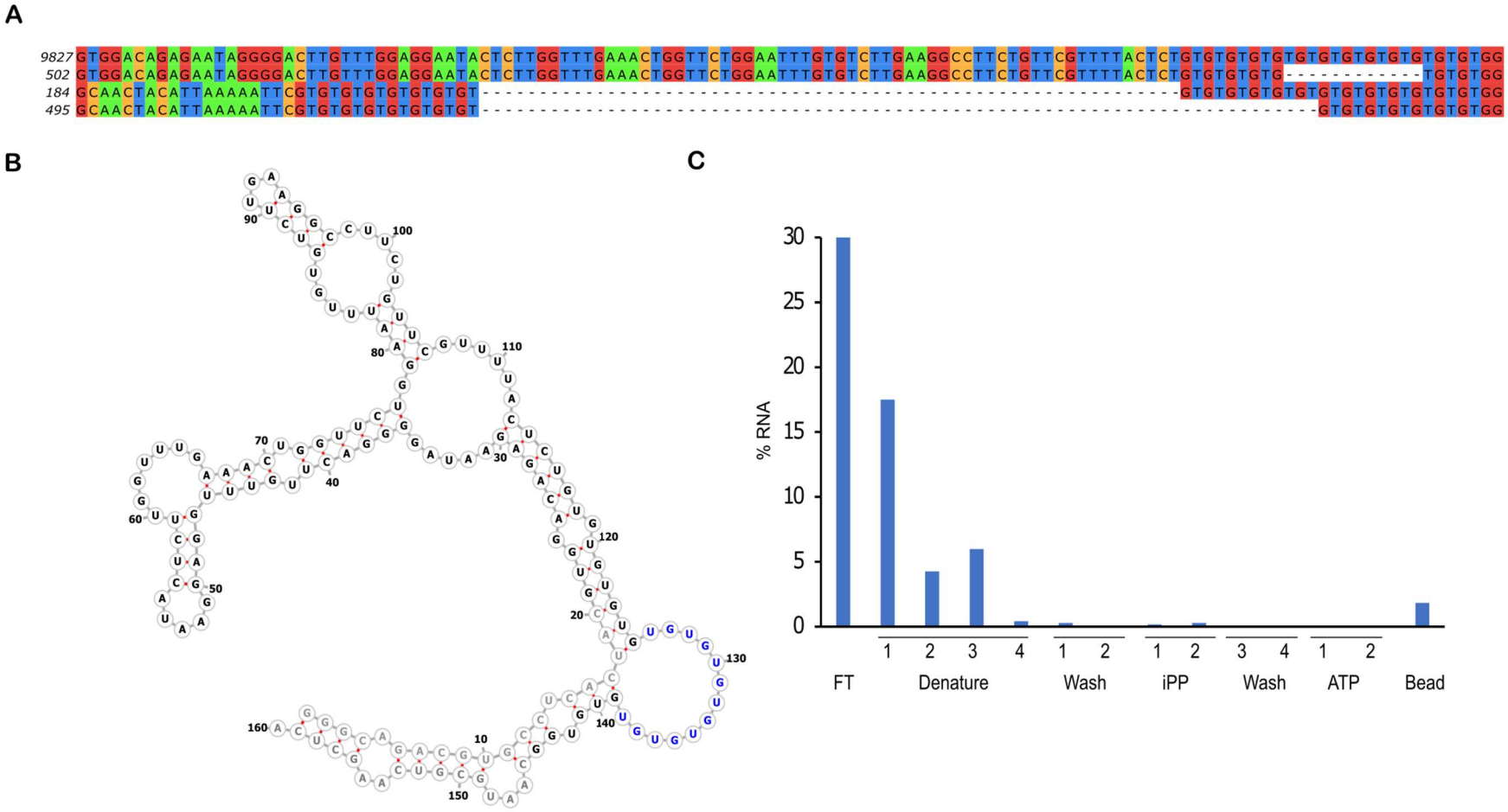
The *TNS3* sequence was actually the most abundant sequence, though in a very short spliced form. Its full-length sequence was still amongst the top sequences. It maps to 4th intron of the *TNS3* (tensin 3) gene on the (-) strand, the same direction as the gene. It lies ~6000 nt from the 4th exon where the gene overlaps with a simple (AC)n repeat. About 1.6 % of this RNA remained attached to beads under denaturing conditions. Not only was it detected as several spliced variants, but also attached to several other genomic sequences from the pool. It is predicted to have an unstructured 60 nt loop which seems to be retained in all but the shortest, but most dominant, sequence.

Figure 7-4 *TNS3* intronic RNA. **(A)** Alignment of sequences representing cluster from *TNS3* sequence. Number on the left indicate abundance of that sequence bottom eight alignments are correspond to the 3′ end of sequences found elsewhere in the pool. (**B**) Thermodynamically predicted secondary structure of the sequence mapped to *TNS3* intron showing a large loop with no predicted structure. Artificial primer sequence is shown in gray. (C) Graph of column binding fractions for the *TNS3* sequence on an ATP pool including elutions with saturate pyrophosphate and 5 mM ATP


A
33035
587
359
235
55955
128
118
173
150
155
137
763
108
B
C
% RNA
30
25
20
15
10
5
0
1 2 3 4
1 2
1 2
3 4
1 2
FT
Denature
Wash
iPP
Wash
ATP
Bead

The chromosome 13 sequence maps to the 4th intron of a long intergenic non-coding (LINC) RNA (17,000 nts from exon 5) and the 30th intron of a Genescan predicted gene (~3,000 nts from exon 31). It's 5′ end also maps to both the HUERS-P1-int ERV family LTR while the 3′ segment maps to a simple (TG)n repeat. Roughly 2.5% of this RNA remained on beads in denaturing conditions. Only one spliced variant was detected in the top 100 clusters removing six units from its (TG) repeat which could corresponds to six UGs predicted to form a 12 nt loop. Its its 3′ end also seems to have been spliced to the 5 end of a sequence mapping to the *PIGN* genes 13th/14th intron and L1ME4c LINE element on chromosome 18 often replacing or extending the it's own TG repeats. Because GT rich sequences are common in this pool, there is a possibility some other sequence also shares a similar 3′ end composed of $(TG)_{13}G$ followed by reverse primer.

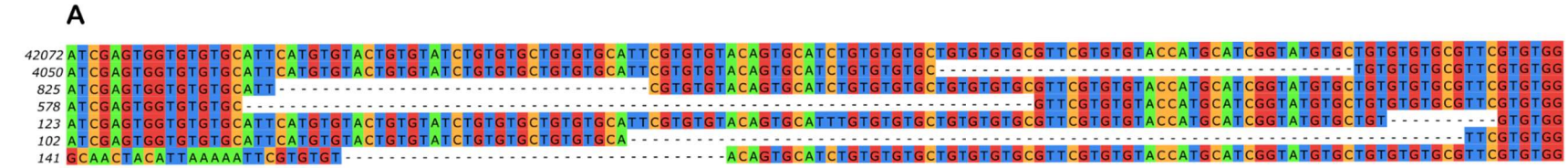
Figure 7-6 Chr 13 LINC intronic RNA. **(A)** Alignment of sequences representing cluster from LINC sequence. Number on the left indicate abundance of that sequence. Bottom alignment contains the 5′ end corresponding to another sequence (**B**) Thermodynamically predicted secondary structure of the LINC RNA. Constant regions are shown in grey. Most commonly removed nucleotides are shown in blue (C) Graph of column binding fractions for the LINC sequence on an ATP pool including elution with 5 mM ATP

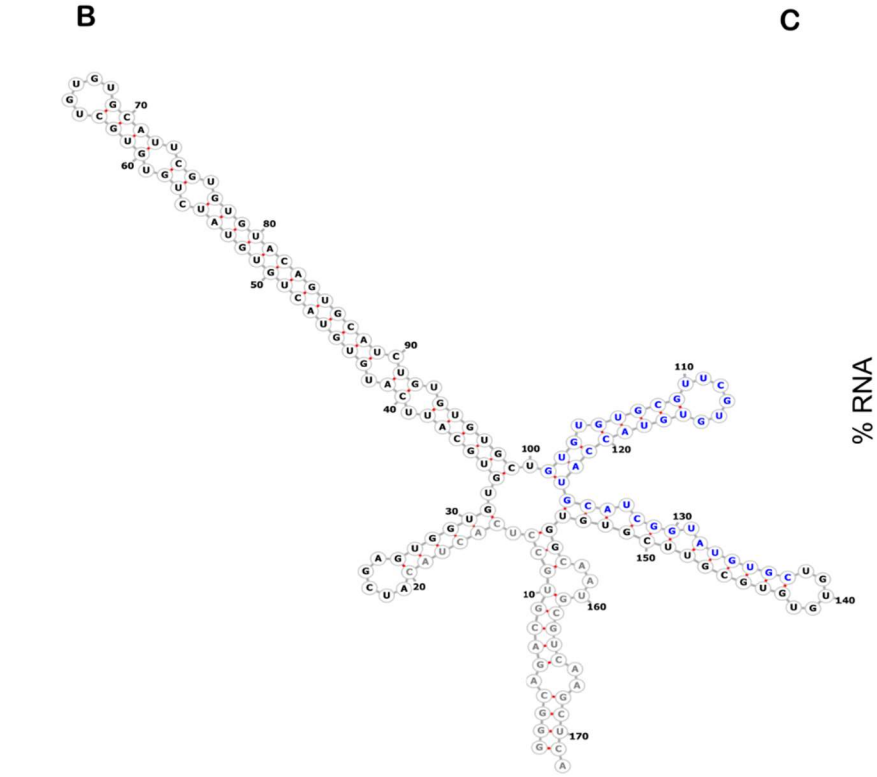

A
9827 GTGGACAGAGAATAGGGGACTTGTTTGGAGGAATACTCTTGGTTTGAAACTGGTTCTGGAATTTGTGTCTTGAAGGCCTTCTGTTCGTTTTACTCTGTGTGTGTGTGTGTGTGTGTGTGTGTGG
502 GTGGACAGAGAATAGGGGACTTGTTTGGAGGAATACTCTTGGTTTGAAACTGGTTCTGGAATTTGTGTCTTGAAGGCCTTCTGTTCGTTTTACTCTGTGTGTGTG----------TGTGTGG
184 GCAACTACATTAAAAATTCGTGTGTGTGTGTGTGT----------GTGTGTGTGTGTGTGTGTGTGTGTGTGG
495 GCAACTACATTAAAAATTCGTGTGTGTGTGTGTGT----------GTGTGTGTGTGTGTGG
B
C
% RNA
30
25
20
15
10
5
0
1 2 3 4 1 2 1 2 3 4 1 2
FT
Denature
Wash
iPP
Wash
ATP
Bead

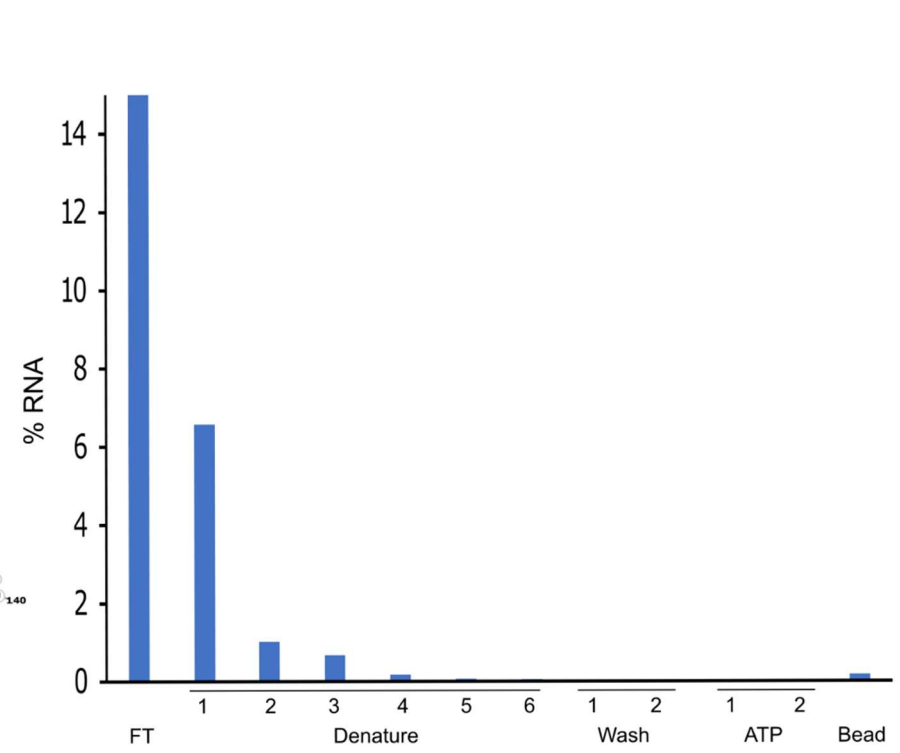
142

The sequence mapping to the *GAK* (cyclin G assosiaked kinase) gene is found in the 19th intron about 650 nts from the 19th exon on the (-) strand in the same orientation as the gene. It is also part of a (CACA)n repeat. Less than half a percent of this RNA remained on ATP beads in denaturing conditions. It's predominent spliced variant removes a stem loop from it's structure. As was the case with the chr 13 LINC sequence, it's 3′ end was also found attached to the 5′ end of the *PIGN* sequence.

Figure 7-5 *GAK* intronic RNA. **(A)** Alignment of sequences representing clusters mapping to the *GAK* intron. Number to the left indicate abundance of that sequence. (**B**) Thermodynamically predicted secondary structure of the *GAK* RNA. Constant regions are shown in grey. Most commonly removed nucleotides are shown in blue (C) Graph of column binding fractions for the *GAK* sequence on an ATP pool including elutions with saturated pyrophosphate and 5 mM ATP

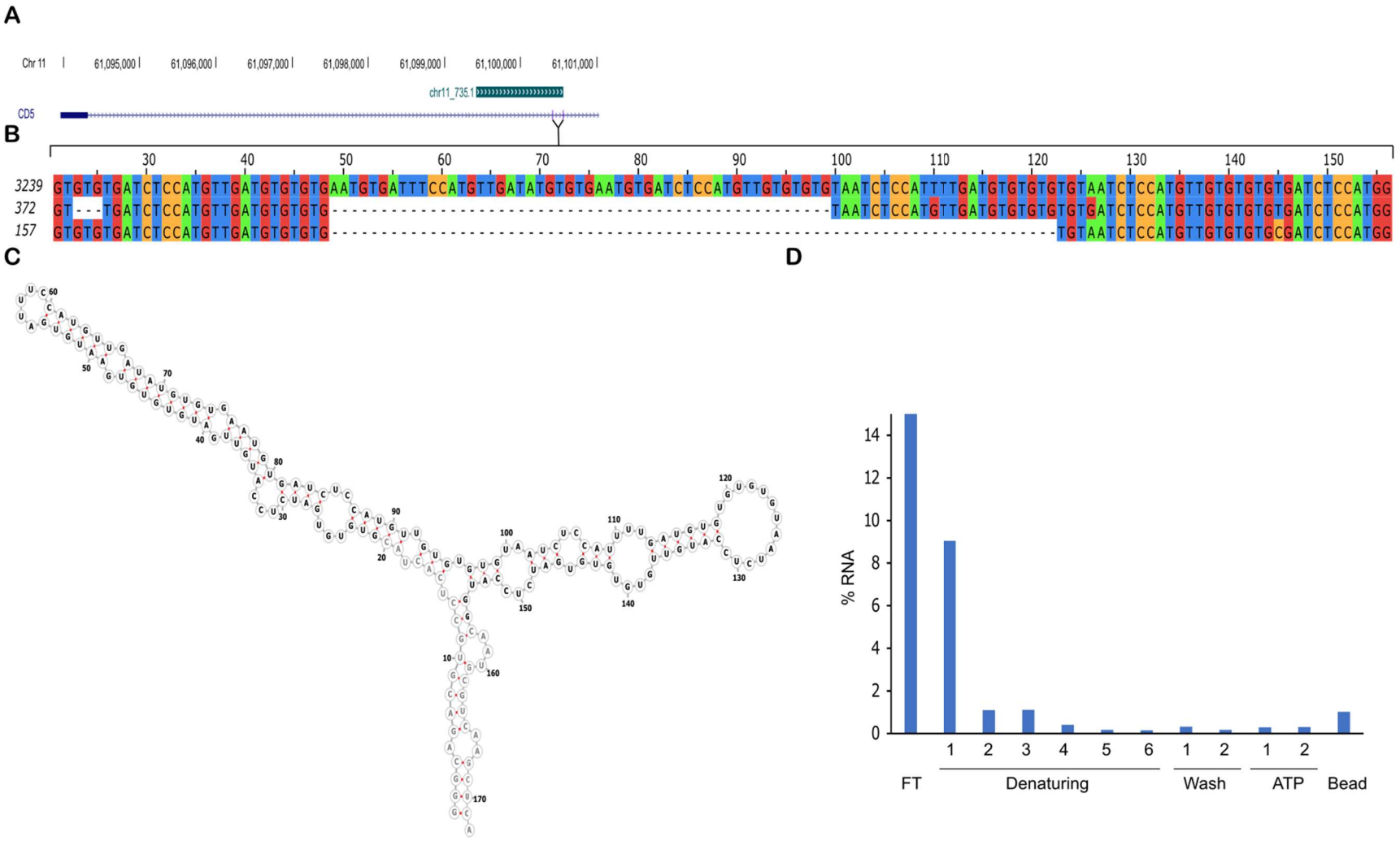

A
42072
4050
825
578
123
102
141


B


C
% RNA
0
2
4
6
8
10
12
14
1
2
3
4
5
6
FT
Denature
1
2
Wash
1
2
ATP
Bead

The last sequence examined mapped to the (-) strand of chromosome 11, antisense to the first intron of the *CD5* gene (~6000 nt from exon 1) and the 3′ end of a Geneid predicted gene, chr11_735, which is composed entirely of that one exon. About 1.6% of this RNA remained bound to ATP-agarose under denaturing conditions. Both spliced variants of this sequence share the same 5′ splice site which included 6 bp stem loop. At first glance, removal of these sequences involve major rearrangement of the RNA nearly the entire length of each spliced out regions is involved with basepairing with the other half of the RNA. However, inputting the spliced out region on it's own shows that it is also capable forming and independent stem, which may be the arrangement used during splicing. It is actually the participation of the forward primer that seems to bias the RNA away from structure in the prediction.

Figure 7-7 *CD5* intronic RNA. **(A)** Genomic location with respect to the *CD5* gene (blue) and the predicted gene (green) (**B)** Alignment of sequences representing cluster from *CD5* sequence. Number on the left indicate abundance of that sequence (**C**) Secondary structure of the sequence mapped to *CD5*. Nucleotides from the constant region (gray) are seen to base pair with removed portions of the sequence. (**D**) Graph of column binding fractions for the *CD5* sequence on an ATP pool including elution 5 mM ATP


A
Chr 11
61,095,000
61,096,000
61,097,000
61,098,000
61,099,000
61,100,000
61,101,000
chr11_735.1
CD5
B
3239
372
157
C
D
% RNA
FT
Denaturing
Wash
ATP
Bead

Apparent splicing events were observed in all abundant sequences in the pool and also observed in the parallel SBC selections. In fact, this pool was dominated by these spliced products of these RNAs with very few reads corresponding to full length version. The bead retention after denaturing was relatively low for each sequence and proved inconclusive. As the selection enriched for RNA covalently attached to ATP, the splicing may be facilitated by nucleophilic attack by adenosine akin to use of guanosine in group I self-splicing introns (Kruger, et al. 1982). However, the nucleophilic guanosine is covalently attached to the splice site by the catalytic portion of the RNA. If the human ribozyme is working in *trans*, RNAs detected to be covalently attached to the ATP beads are substrates. We can see evidence of activity, then, in the presence of both ribozyme and substrate, but we will not see activity for ribozyme alone if the ribozyme is a poor substrate for itself. Activity could potentially be seen by incubating each sequence with labeled RNA deemed to be a good substrate based on the sequencing collected, however a more direct readout of splicing is preferable to detection of a covalent intermediate.

Alignment of the splice site indicates the most conserved nucleotides are the UGU directly 5′ of the splice site and the UGU starting from the 3rd nucleotide of the spliced-out intron giving us some idea of the substrate specificity. Preliminary structure predictions do not seem to indicate base pairing between exon and intron to facilitate cleavage as is the case with group I introns, but instead topologies where both exons are in close proximity, for example via hairpin

formation of the intron, however the predicted structure may not represent the active structure.

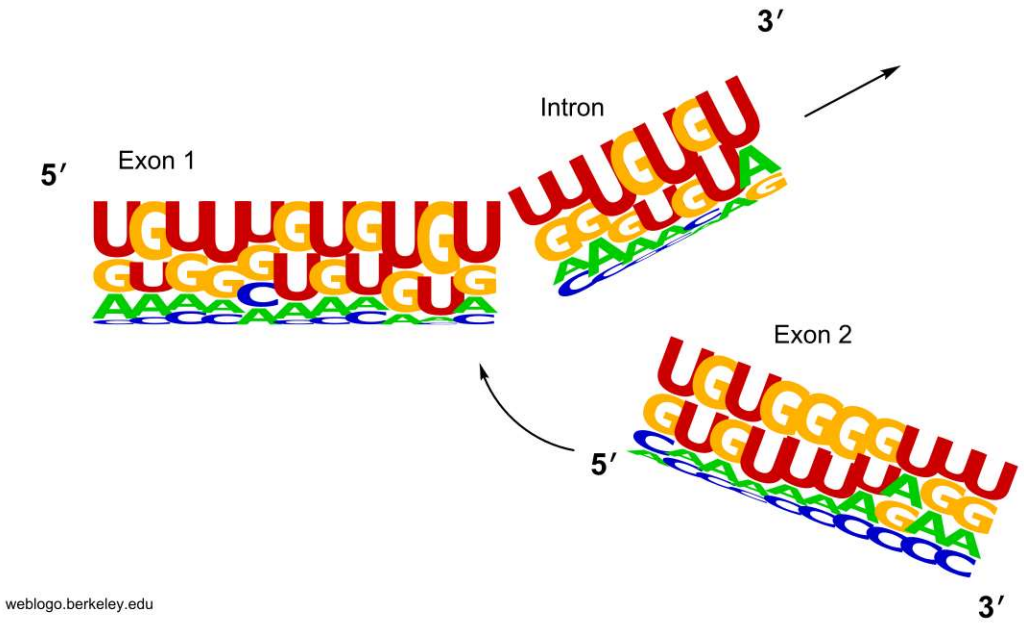


Figure 7-8 Sequence composition of splice site. Logo showing composition of introns and exons near the splice site. Sequences are more UG rich closer to the splice site and become more evenly distributed as you move away. The most conserved regions are the 3′ UGU of the first exon and the UGU of the intron starting three nucleotides from the splice site which are preceded by two nucleotides with low nucleotide preference.

Conclusions

To borrow the words of NSF GFRP reviewer 1, "An outwardly spurious observation of unusually strong ATP binding by aptamers [has been developed] into an impactful, testable hypothesis." The *in vitro* selection for RNAs covalently attached to ATP produced an intriguing pool of RNA that could represent further examples of functional RNAs playing a role in human biology. Sequencing data suggests that this ribozyme is catalyzing the splicing of other RNAs. Borrowing words once more, this "has the potential to transform our understanding of human and eukaryotic gene regulation" -NSF reviewer 3. The putative introns are predicted to form structures independent of the retained exons or are capable of forming independent structures as

an alternative structure. These map primarily in introns or predicted introns, however, none of them occur at the predicted splice site.

The presence of spliced products for each sequence, including ATP aptamers, and examples of RNAs spliced together from two separate genomic RNAs, together indicate that the ribozyme acts in *trans*, complicating the identification of the active sequence. Enriched sequences may be the best substrates of the ribozyme and not the ribozyme itself. For this reason, it is not clear if any of the sequences reported are the ribozyme itself or the preferred substrates for a splicing ribozyme that targets GU rich regions. This could explain the continued presence of ATP aptamers as they would be the most likely sequences to be found near the immobilized ATP. Alternatively, the ribozyme could utilize the known ATP binding domain to position ATP for nucleophilic attack. The splicing also explains some difficulties encountered in selection as the ribozyme can only be carried forward when it acts as a substrate for itself and continued splicing could remove the active motif. We also learn the identity of the persistent short sequences that come out of RT-PCR after collecting fractions from what was a gel purified sample.

The sequences and structure predictions of splice sites provides some insights about the splicing substrate but there is still much biochemical analysis to be done. Given what we know about this potential splicing ribozyme, it is uncertain whether its covalent attachment to beads is an effective way to screen candidates *in vitro*. Identification of the reactive site had previously been planned by mapping ATP dependent reverse transcription termination at low temperature and analyzed by sequencing. However, with more knowledge of the function of this human ribozyme, new methods can be applied to screen for activity or select further from these pools (Olson and Müller 2012).

Materials and Methods

Pool construction

The construction of the DNA pool derived from the human genome was described previously (Salehi-Ashtiani et al. 2006) using genomic DNA from whole human blood and ligating selection primers below

In vitro Selection Primers

Forward: 5´ GATCTGTAATACGACTCACTATAGGGCAGACGTGCCTCACTAC-3´

Reverse: 5´ CTGAGCTTGACGCATTG 3´

RNA Transcription

RNA was transcribed at 37 °C for 1 to 3 h in a volume of 20 ml containing 40 mM Tris chloride, 10% (v/v) dimethyl sulfoxide (DMSO), 10 mM dithiothreitol (DTT), 2 mM spermidine, 2.5 mM each CTP, GTP, and UTP, .250 mM ATP, 2.25 mCi [a-^{32}P]-ATP (Perkin Elmer, Waltham, MA, USA), 25 mM $MgCl_2$, one unit of T7 RNA polymerase, and 0.2 mM of DNA template. DMSO was used to increase transcript yields, as documented in previous studies.17 The transcripts were purified using denaturing PAGE.

Substrate exchange selection procedure

Purified RNA transcripts were precipitated, dried, and resuspended in 100 ml binding buffer containing 140 mM KCl, 10 mM NaCl, 10 mM Tris chloride, pH 7.5, and 5 mM $MgCl_2$ and heated to 70 °C before loading on to C8-linked ATP-agarose beads (Sigma-Aldrich, St. Louis, MO, USA) equilibrated in the binding buffer. Flow-through was collected after the columns were capped and shaken for 20 min at room temperature. The beads were denatured in 100 µl of a denaturing solution with 7 M urea and 0.5 M EDTA for 2 minutes at 72 ℃ then washed in the same buffer. The pool was refolded with washes of 100 µl of the binding buffer and potential ribozymes were eluted with the same buffer supplemented by 5 mM ATP-Mg with 30 min of shaking at room temperature. Each fraction was analyzed for radioactivity using a liquid scintillation counter. Elutions were pooled, desalted using YM-10 spin filters (Millipore, Billerica, MA, USA), precipitated, dried, and resuspended in H2O.

Pyrophospahate collection selection procedure

Purified RNA transcripts were precipitated, dried, and resuspended in 100 ml binding buffer containing 140 mM KCl, 10 mM NaCl, 10 mM Tris chloride, pH 7.5, and 5 mM $MgCl_2$ and heated to 70 °C before loading on to C8-linked ATP-agarose beads (Sigma-Aldrich, St. Louis, MO, USA) equilibrated in the binding buffer. Flow- through was collected after the columns were capped and shaken for 20 min at room temperature. The beads were denatured in 100 µl of a denaturing solution with 7 M urea and 0.5 M EDTA for 2 minutes at 72 ℃ then washed in the same buffer. The pool was refolded with washes of 100 µl of the binding buffer and subsequently washed with the same buffer supplemented by with saturated pyrophosphate with 30 min of shaking at room temperature. Each fraction was analyzed for radioactivity using a

liquid scintillation counter Each fraction was analyzed for radioactivity using a liquid scintillation counter. Elutions were pooled, desalted using YM-10 spin filters (Millipore, Billerica, MA, USA), precipitated, dried, and resuspended in H2O.

Solid matrix collection selection procedure

Purified RNA transcripts were precipitated, dried, and resuspended in 100 ml binding buffer containing 140 mM KCl, 10 mM NaCl, 10 mM Tris chloride, pH 7.5, and 5 mM $MgCl_2$ and heated to 70 °C before loading on to C8-linked ATP-agarose beads (Sigma-Aldrich, St. Louis, MO, USA) equilibrated in the binding buffer. Flow- through was collected after the columns were capped and shaken for 20 min at room temperature. The beads were denatured in 100 µl of a denaturing solution with 7 M urea and 0.5 M EDTA for 2 minutes at 72 ℃ then washed in the same buffer. The pool was refolded with washes of 100 µl of the binding buffer and subsequently washed with the same buffer supplemented by 5 mM ATP.Mg with 30 min of shaking at room temperature. Each fraction was analyzed for radioactivity using a liquid scintillation counter. The solid matrix was collected and resuspended in 1X reverse transcriptions buffer (Promega, Madison, WI, USA)

Reverse Transcription

RNA was reverse transcribed in 20 µL using the Promega reverse transcription buffer, 2 mM reverse primer, and RNA recovered from the previous selection round. The RNA and primer were annealed by heating at 65 °C and cooling to room temperature before 1 unit of Thermoscript (Invitrogen, Grand Island, NY, USA) and Improm II (Promega, Madison, WI, USA) reverse transcriptases each were added. The reaction was initiated for 5 min at 25 °C, and

then the temperature was ramped to 42 °C, 50 °C, 55 °C, and 65 °C for 15 min each before the enzymes were inactivated at 85 °C for 5 min.

Amplification

DNA was amplified by using DreamTaq Master Mix (Fermentas, Glen Burnie, MD, USA), 2 μM forward primer, 2 μM reverse primer, and DNA from reverse transcription. DNA was initially denatured at 95 °C for 1.5 min (30 s for subsequent denaturing steps), annealed at 55 °C for 30 s, and extended at 72 °C for each cycle. Optimum number of PCR cycles was determined for each selection round by comparing 8-, 12-, 16-, and 20-cycle aliquots on agarose gel.

Sequence diversity estimate

The maximum estimated diversity was taken as the product of the fraction of RNA collected and the starting RNA diversity of that round. The loss of diversity from each round is likely greater due to sequence bias in reverse transcription-PCR and disproportionate isolation of sequence repeats during selection.

ATP matrix reactivity assay

Individual clones were tested for reactivity by the same procedure described for the selection procedures up to analysis by liquid scintillation counter.

Cloning

The pool was cloned using the Topo TA cloning kit (Invitrogen) and individual colonies were directly PCR amplified and sequenced.

M13 Primers

M13 forward - 5´-GTAAAACGACGGCCAG-3´

M13 reverse - 5´-CAGGAAACAGCTATGAC-3´

Chr	Strand	Reads	Genetic loci	Sequence
1	+	6	*ABCD3*	gggcaGACTCCTGGCTTTCTAGCTGGAGCAACTGTTTGATAGTTTCATTTGCTGAAGTGTTCAAGATTAGGGAAGGGAATGGATTTTgtgcctcactacg
1	-	17	*LAMTOR5-AS1* MER89 LTR	tctgggcagacggcctcactaGCTATAGCAAGCTTGGGCCAAGTAGCTTTTGCTTGTCTCATTTGGTCTTCGTTTATTTCCACAATCCTTCCAAATGCTA
2	-	9	L1MDa LINE	acacgcacggggCTGCCTTaCTGTGAACATTTTTAAGAATTCCACTTTGCACTGTTTTTGAAAAGATCTCTTTGCATAGCTTTGTGGTTACTCTAAATGT
2	-	34		gaaatgggcacgtgcctcaCTCTGGTTGCTCCGAAGTCCATGCTTGCTGGGCTCTTGGTTTGCCTTTGCTTTTTCTCTGGTTTGTTTGGGACATGGTTGA
2	-	6	(ACATGCAC)n	gggcagacgtgcctcactacGTGTGCATGTGTGCATGTGTGCATGTGTGAGTATGTGTGCATGTGTGCATGTGTGCATGTGTGTATGTGTGCCtTTGTGT
4	+	32	*RNF212* L1MC1 LINE	gggcagacgtcactacgtGTGAACCCTAAcGTAAACCATGGACGTTGGGTGATAATGATGTATTCAAGTAGGATCATCAGTTTTAACACATGGACTGCct
4	-	43	*GAK* (CACA)n	ccacgggcagacgtgtcTACTGTGTATCTGTGTGCTGTGTGCATTCGTGTGTACAGTGCATCTGTGTGTGCTGTGTGTGCGTTCGTGTGTACCATGCATC
7	+	4	*MAD1L* (TGTG)n	gggcagacgtgtcaCTGCCTGTGCATGTGAGCATGCATGTGGCTGCCTGTACATGTGTGTGCACGCATGTATGTGGCTGCCTGTTCATGTGTGCccacgc
7	-	64	*TNS3* (AC)n	gatgggcagactgctcctacGCACAGCTCTGTGTGCAAGATGCGTGTGTCTGTGGTCTGTGGTGTGTGTGAGGTGTTTTGTGTGTGTATGTTGTGTTTGT
8	+	5	*NRG1*	gggcagacGTTAAGTAAAGTGGGAGGAAAGTGGTTGTGGCTATAAAAGTGCCACATGAGGAATGCTTGTAATGATGGAATTGTCTGTTTCgcctcaccgt
11	-	10	*CD5*	actgggcagacgtgcctcacTGTGATCTCCATGTTGATGTGTGTGTAATCTCCATGTTGATGTGTGTGTGTGATCTCCATGTTGTGTGTGTGATCTCCAT
11	+	58	*KIRREL3*	gggcagacgtgcctctacGTGTCTGCATGTGAAAGTGTGGGCATATGTGTGTATGCGTGTATAGATGTGTGAGCATGTGCACGTGTGTGTGCATGTGTca
12	+	82	*RAD5*	gggcaacGGCCTActACTAGATGTATaATTTTGTATGATGCAATATTGGAGTAAAGTGTCATGTTTTCTGCAACTTACTCTCAATTCATTCAACTAgtcg
12	-	6		gacgtgcctcacgggcATACcTGTGTGTTTGTGCATtTCAAAGATTCAGTGTGTTTCTGTGATGTGTTGTTGAACGTGACCATGAGAGTGGCAGTGGGTT
12	+	4	*NCOR2* (TG)n (TGTGAG)n	GgGCATGTGCGCATGTATCTGTGTGGGTGTCTGCATGTGAGTATGGGTGTGTGTGCACGACTGTGAGTGTGCATATGAGTGTGAGgacgcctcactacgt
12	-	3	*TMEM132D*	gggcagacgtgcctcacTACGTACGATTGTTTTCTCATTCGTCATCTGCCTTCTGAGACAGTGGGTGGGATGTGcTGTGGCAGGCAGAGACCATGTGAAC
13	+	1328	LINC00423 HUERS-P1-int	ggtcacacgtgcCTTAgTGGACcGAGAATAGGGGACTTGTTTGGAGGAATACTCTTGGTTTGAAACTGGTTCTGGAATTTGTGTCTTGAAGGCCTTCacc
15	+	4	EST: BU853031 AC087633.2	gggcagacgtgccTCTGACTGGGTTATATGCaAGGTATTGTCCTGGTTAAGTGGCCATTGCTCAAGCCTGACTACTGCCTATAGGGTGAGTTGCTCACTA

15	-	3	AC022523.1	cgtgcctcaccgtgggcAGATAGTTTTCTCTGTTTTAGAGACAA GGAAACTGAGTAAAGTACCTCAACAAGTTATTGTGATGGTGCAA TGAAAAGAAACG
16	+	21	*PHKB*	gggcgtcctcactacgtCGAGATTGCCTAATGATGCATTTCTCA GAATGTATCTTGGTCATTAAGCAATGCATGACAGTGTGTGTGTG TGTGTGTagacg
16	-	4		cgtccacgggCAGAGcTCATGTTAGCACTGTTATTGCTCATAAT GATGATGGTCTTGATGATGATGATGGTCaTGGTGATGATGACGG TGATGATGGTGG
18	+	13	*CEP291*	cgtccacgggCAGAGcTCATGTTAGCACTGTTATTGCTCATAAT GATGATGGTCTTGATGATGATGATGGTCaTGGTGATGATGACGG TGATGATGGTGG
19	-	10	*MIER2*	ccactacgtgggcagacGTGTCTGATGAGGATGTACTTTGGTGT ATTGAGGTGTCTGATGAGGATGTGCTTTGGTGTATTGAGGAGTC TGATGAGGATGT
19	+	5		GGGCATTTGACAAACGCCGCTGAGAATTTCAGGATGACACGAAG GAACCTTTGATTGGATTTGTGGATGTGCCTGTGAAAGCGTGgac gtgcctcacacg
20	-	130		cgtgcctcactacgtgggcagAAAAAAAGATCTTGGAGCTCATT CCATAAGTGACTTGCATTTCAACAGGATTCTTCAAAAGTGACAA GAAAACATAGGA
X	-	3	*FAM9B*	ctgcctcgggcagagactacgtGATTGGTAATGTCTGAGCATTG TGAAAAACAGATGTCCTGACATAGAACCAATTGGATTTTTTACC AGTTGAGGGAGG
X	-	4	*PLAC1*	tgcctcgggcagacgactacgTAATGACATGAGTAAATGCTCAT GGTCTAACTTAAGTGGGGGAAAGGAGGATACACAAACATACAGT GAAATCCCAAGT

Table S7-1. Sequences identified in the round 6 SB pool. Lowercase letters correspond to nucleotides that do not match the human reference genome.

References

Abdelsayed, M.M., 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 Chemical Biology* 12, no. 8 (2017).

Alam, Khalid K., Jonathan L. Chang, and Donald H. Burke. "FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections." *Molecular Therapy - Nucleic Acids*, 2015.

Bernhart, Stephan H., Ivo L. Hofacker, Sebastian Will, Andreas R. Gruber, and Peter F. Stadler. "RNAalifold: Improved consensus structure prediction for RNA alignments." *BMC Bioinformatics*, 2008.

Bittker, Joshua A., Brian V. Le, and David R. Liu. "Nucleic acid evolution and minimization by nonhomologous random recombination." *Nature Biotechnology*, 2002.

Braunschweig, Ulrich, et al. "Widespread intron retention in mammals functionally tunes transcriptomes." *Genome Research*, 2014.

Breaker, Ronald R. "Prospects for Riboswitch Discovery and Analysis." *Molecular Cell.* Vol. 43. no. 6. 9 16, 2011. 867-879.

Brunel, Christine, Bernard Ehresmann, Chantal Ehresmann, and Michael McKeown. "Selection of genomic target RNAs by iterative screening." *Bioorganic and Medicinal Chemistry*, 2001.

Buchli, Reto, and Peter Boesiger. "Comparison of methods for the determination of absolute metabolite concentrations in human muscles by 31P MRS." *Magnetic Resonance in Medicine*, 1994.

Buenrostro, Jason D., et al. "Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes." *Nature Biotechnology*, 2014.

Burgstaller, Petra, and Michael Famulok. "Isolation of RNA Aptamers for Biological Cofactors by In Vitro Selection." *Angewandte Chemie International Edition in English*, 1994.

Burke, Donald H., and Larry 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 Research*, 1997.

Cech, Thomas R., and Joan A. Steitz. "The noncoding RNA revolution - Trashing old rules to forge new ones." *Cell.* 2014.

Cheah, Ming T., Andreas Wachter, Narasimhan Sudarsan, and Ronald R. Breaker. "Control of alternative RNA splicing and gene expression by eukaryotic riboswitches." *Nature* (Nature Publishing Group) 447, no. 7143 (5 2007): 497-500.

Chen, L., S. W. Yun, J. Seto, W. Liu, and M. Toth. "The fragile X mental retardation protein binds and regulates a novel class of mRNAs containing u rich target sequences." *Neuroscience*, 2003.

Cochrane, Jesse C., and Scott A. Strobel. "Probing RNA Structure and Function by Nucleotide Analog Interference Mapping." In *Current Protocols in Nucleic Acid Chemistry*, by Jesse C. Cochrane, & Scott A. Strobel. 2004.

Costanzo, Giovanna, et al. "Non-enzymatic oligomerization of 3′,5′ cyclic AMP." *PLoS ONE* (Public Library of Science) 11, no. 11 (11 2016).

Croft, M. T., M. Moulin, M. E. Webb, and A. G. Smith. "Thiamine biosynthesis in algae is regulated by riboswitches." *Proceedings of the National Academy of Sciences*, 2007.

Curtis, Edward A, and David R Liu. "A naturally occurring, noncanonical GTP aptamer made of simple tandem repeats." *RNA biology*, 2014.

Curtis, Edward A., and David R. Liu. "Discovery of widespread GTP-binding motifs in genomic DNA and RNA." *Chemistry and Biology*, 2013.

Dayton, Andrew I., Joseph G. Sodroski, Craig A. Rosen, Wei Chun Goh, and William A. Haseltine. "The trans-activator gene of the human T cell lymphotropic virus type III is required for replication." *Cell*, 1986.

Dieckmann, T, E Suzuki, G K Nakamura, and J Feigon. "Solution structure of an ATP-binding RNA aptamer reveals a novel fold." *RNA (New York, N.Y.)*, 1996.

Dieckmann, Thorsten, Samuel E. Butcher, Mandana Sassanfar, Jack W. Szostak, and Juli Feigon. "Mutant ATP-binding RNA aptamers reveal the structural basis for ligand binding." *Journal of Molecular Biology* (Academic Press) 273, no. 2 (10 1997): 467-478.

Ding, Yiliang, Chun Kit Kwok, Yin Tang, Philip C. Bevilacqua, and Sarah M. Assmann. "Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq." *Nature Protocols*, 2015.

Dobbelstein, M., and T. Shenk. "In vitro selection of RNA ligands for the ribosomal L22 protein associated with Epstein-Barr virus-expressed RNA by using randomized and cDNA-derived RNA libraries." *Journal of Virology*, 1995.

Doudna, Jennifer A., and Thomas R. Cech. "The chemical repertoire of natural ribozymes." *Nature.* 2002.

Ellington, Andrew D., and Jack W. Szostak. "In vitro selection of RNA molecules that bind specific ligands." *Nature* 346, no. 6287 (1990): 818-822.
Esteller, Manel. "Non-coding RNAs in human disease." *Nature Reviews Genetics.* 2011.
Fu, Yang, Kaila Deiorio-Haggar, Mark W. Soo, and Michelle M. Meyer. "Bacterial RNA motif in the 5′ UTR of rpsF interacts with an S6:S18 complex." *RNA*, 2014.
Fujimoto, Yuki, Yoshikazu Nakamura, and Shoji Ohuchi. "HEXIM1-binding elements on mRNAs identified through transcriptomic SELEX and computational screening." *Biochimie*, 2012.
Garst, Andrew D., Andrea L. Edwards, and Robert T. Batey. "Riboswitches: Structures and mechanisms." *Cold Spring Harbor Perspectives in Biology.* Vol. 3. no. 6. Cold Spring Harbor Laboratory Press, 2011. 1-13.
Gebhardt, Kirsti, Afshin Shokraei, Eshrat Babaie, and Bjørn H. Lindqvist. "RNA aptamers to S-adenosylhomocysteine: Kinetic properties, divalent cation dependency, and comparison with anti-S-adenosylhomocysteine antibody." *Biochemistry* 39, no. 24 (6 2000): 7255-7265.
Hammann, Christian, Andrej Luptak, Jonathan Perreault, and Marcos De La Peña. "The ubiquitous hammerhead ribozyme." *RNA.* Vol. 18. no. 5. 5 2012. 871-885.
Hammond, Scott M. "An overview of microRNAs." *Advanced Drug Delivery Reviews.* Vol. 87. Elsevier B.V., 6 29, 2015. 3-14.
Hayakawa, Makio, et al. "Novel insights into FGD3, a putative GEF for Cdc42, that undergoes SCF FWD1/β-TrCP -mediated proteasomal degradation analogous to that of its homologue FGD1 but regulates cell morphology and motility differently from FGD1." *Genes to Cells* 13, no. 4 (4 2008): 329-342.
Hofacker, Ivo L. "Vienna RNA secondary structure server." *Nucleic Acids Research*, 2003.
Hoinka, Jan, and Teresa Przytycka. "AptaPLEX – A dedicated, multithreaded demultiplexer for HT-SELEX data." *Methods*, 2016.
Huang, Faqing, and Michael Yarus. "5‘-RNA Self-Capping from Guanosine Diphosphate †." *Biochemistry*, 1997.
Jaeger, John A., and Ignacio Tinoco. "An NMR Study of the HIV-1 TAR Element Hairpin." *Biochemistry*, 1993.
Jalali-Yazdi, Farzad, Lan Huong lai, Terry T. Takahashi, and Richard W. Roberts. "High-Throughput Measurement of Binding Kinetics by mRNA Display and Next-Generation Sequencing." *Angewandte Chemie - International Edition*, 2016.
Jiang, F., R. A. Kumar, R. A. Jones, and D. J. Patel. "Structural basis of RNA folding and recognition in an AMP-RNA aptamer complex." *Nature*, 1996.
Jijakli, Kenan, et al. "The in vitro selection world." *Methods.* Vol. 106. Academic Press Inc., 8 15, 2016. 3-13.
Joyce, Gerald F. "RNA evolution and the origins of life." *Nature.* 1989.
Kemp, Graham J., Martin Meyerspeer, and Ewald Moser. "Absolute quantification of phosphorus metabolite concentrations in human muscle in vivo by 31P MRS: A quantitative review." *NMR in Biomedicine.* Vol. 20. no. 6. 10 2007. 555-565.
Kim, Soyoun, Hua Shi, Dong Ki Lee, and John T. Lis. "Specific SR protein-dependent splicing substrates identified through genomic SELEX." *Nucleic Acids Research.* 2003.
Klopf, Eva, et al. "Nascent RNA signaling to yeast RNA Pol II during transcription elongation." *PLoS ONE* (Public Library of Science) 13, no. 3 (3 2018).
Kolb, Gaëlle, et al. "Endogenous expression of an anti-TAR aptamer reduces HIV-1 replication." *RNA Biology*, 2006.
Kruger, Kelly, Paula J. Grabowski, Arthur J. Zaug, Julie Sands, Daniel E. Gottschling, and Thomas R. Cech. "Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena." *Cell*, 1982.
Kubodera, Takafumi, et al. "Thiamine-regulated gene expression of Aspergillus oryzae thiA requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR." *FEBS letters* 555, no. 3 (12 2003): 516-20.
Lambert, Nicole, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A. Sharp, and Christopher B. Burge. "RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins." *Molecular Cell*, 2014.
Lapidus, Alla, et al. "Extending the Bacillus cereus group genomics to putative food-borne pathogens of different toxicity." *Chemico-Biological Interactions* (Elsevier Ireland Ltd) 171, no. 2 (1 2008): 236-249.
Lorenz, C., et al. "Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts." *Nucleic Acids Research*, 2010.
Lorenz, Christina, Frederike von Pelchrzim, and Renée Schroeder. "Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels." *Nature Protocols* 1, no. 5 (12 2006): 2204-2212.

Loughrey, David, Kyle E. Watters, Alexander H. Settle, and Julius 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 research*, 2014.
Lucks, J. B., et al. "Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)." *Proceedings of the National Academy of Sciences*, 2011.
Lynch, Kristen W., and Tom Maniatis. "Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila* doublesex splicing enhancer." *Genes and Development*, 1996.
Lynch, Kristen W., and Tom Maniatis. "Synergistic interactions between two distinct elements of a regulated splicing enhancer." *Genes and Development*, 1995.
Madigan, S J, P Edeen, J Esnayra, and M McKeown. "att, a target for regulation by tra2 in the testes of *Drosophila* melanogaster, encodes alternative RNAs and alternative proteins." *Molecular and Cellular Biology*, 1996.
Mandal, Maumita, and Ronald R. Breaker. "Adenine riboswitches and gene activation by disruption of a transcription terminator." *Nature Structural and Molecular Biology* 11, no. 1 (1 2004): 29-35.
Matylla-Kulinska, Katarzyna, Jennifer L. Boots, Bob Zimmermann, and Renée Schroeder. "Finding aptamers and small ribozymes in unexpected places." *Wiley Interdisciplinary Reviews: RNA.* 2012.
McDonald, Richard I., John P. Guilinger, Shankar Mukherji, Edward A. Curtis, Won I. Lee, and David R. Liu. "Electrophilic activity-based RNA probes reveal a self-alkylating RNA for RNA labeling." *Nature chemical biology*, 2014.
McGinnis, Jennifer L., Jack A. Dunkle, Jamie H.D. Cate, and Kevin M. Weeks. "The mechanisms of RNA SHAPE chemistry." *Journal of the American Chemical Society*, 2012.
Meli, Marc, Jacques Vergne, Jean-Luc Décout, and Marie-Christine Maurel. "Adenine-Aptamer Complexes." *Journal of Biological Chemistry* (American Society for Biochemistry & Molecular Biology (ASBMB)) 277, no. 3 (1 2002): 2104-2111.
Merino, Edward J., Kevin A. Wilkinson, Jennifer L. Coughlan, and Kevin M. Weeks. "RNA structure analysis at single nucleotide resolution by Selective 2′-Hydroxyl Acylation and Primer Extension (SHAPE)." *Journal of the American Chemical Society*, 2005.
Mironov, Alexander S., et al. "Sensing small molecules by nascent RNA: A mechanism to control transcription in bacteria." *Cell* (Cell Press) 111, no. 5 (11 2002): 747-756.
Montange, Rebecca K., and Robert T. Batey. "Riboswitches: Emerging Themes in RNA Structure and Function." *Annual Review of Biophysics* (Annual Reviews) 37, no. 1 (5 2008): 117-133.
Nimjee, Shahid M., Christopher P. Rusconi, and Bruce A. Sullenger. "Aptamers: An Emerging Class of Therapeutics." *Annual Review of Medicine*, 2005.
Olson, Karen E., and Ulrich F. Müller. "An in vivo selection method to optimize trans-splicing ribozymes." *RNA*, 2012.
Pobanz, Kelsey, and Andrej Lupták. "Improving the odds: Influence of starting pools on in vitro selection outcomes." *Methods.* 2016.
Riccitelli, Nathan J., and Andrej Lupták. "Computational discovery of folded RNA domains in genomes and in vitro selected libraries." *Methods.* Vol. 52. no. 2. 10 2010. 133-140.
Ring, H Z, and J T Lis. "The SR protein B52/SRp55 is essential for *Drosophila* development." *Molecular and Cellular Biology*, 1994.
Rosikiewicz, Wojciech, and Izabela Makałowska. "Biological functions of natural antisense transcripts." *Acta Biochimica Polonica.* 2016.
Sabeti, Pardis C., Peter J. Unrau, and David P. Bartel. "Accessing rare activities from random RNA sequences: The importance of the length of molecules in the starting pool." *Chemistry and Biology*, 1997.
Salehi-Ashtiani, K., and J. W. Szostak. "In vitro evolution suggests multiple origins for the hammerhead ribozyme." *Nature*, 2001.
Salehi-Ashtiani, Kourosh, Andrej Lupták, Alexander Litovchick, and Jack W Szostak. "A Genomewide Search for Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene." *Science*, 2006.
Sassanfar, Mandana, and Jack W. Szostak. "An RNA motif that binds ATP." *Nature*, 1993.
Schonn, Jean Sébastien, et al. "Rab3 proteins involved in vesicle biogenesis and priming in embryonic mouse chromaffin cells." *Traffic*, 2010.
Sedlyarova, Nadezda, et al. "Natural RNA Polymerase Aptamers Regulate Transcription in E. coli." *Molecular Cell*, 2017.
Sharp, Phillip A. "The Centrality of RNA." *Cell.* Vol. 136. no. 4. 2 20, 2009. 577-580.
Shtatland, Timur, et al. "Interactions of Escherichia coli RNA with bacteriophage MS2 coat protein: genomic SELEX." *Nucleic Acids Research*, 2000.

Singer, Britta S., Timur Shtatland, David Brown, and Larry Gold. "Libraries for genomic SELEX." *Nucleic Acids Research*, 1997.
Soukup, Garrett A., and Ronald R. Breaker. "Relationship between internucleotide linkage geometry and the stability of RNA." *RNA* 5, no. 10 (10 1999): 1308-1325.
Spitale, Robert C., et al. "Structural imprints in vivo decode RNA regulatory mechanisms." *Nature*, 2015.
Stoddard, Colby D., Rebecca K. Montange, Scott P. Hennelly, Robert P. Rambo, Karissa Y. Sanbonmatsu, and Robert T. Batey. "Free State Conformational Sampling of the SAM-I Riboswitch Aptamer Domain." *Structure*, 2010.
Stoltenburg, Regina, Nadia Nikolaus, and Beate Strehlitz. "Capture-SELEX: Selection of DNA aptamers for aminoglycoside antibiotics." *Journal of Analytical Methods in Chemistry* 1, no. 1 (2012).
Tacke, Roland, Masaya Tohyama, Satoshi Ogawa, and James L. Manley. "Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing." *Cell*, 1998.
Takahashi, Mayumi, et al. "High throughput sequencing analysis of RNA libraries reveals the influences of initial library and PCR methods on SELEX efficiency." *Scientific Reports*, 2016.
Tang, Yin, et al. "StructureFold: Genome-wide RNA secondary structure mapping and reconstruction in vivo." *Bioinformatics*, 2015.
Terasaka, Naohiro, Kazuki Futai, Takayuki Katoh, and Hiroaki Suga. "A human microRNA precursor binding to folic acid discovered by small RNA transcriptomic SELEX." *RNA*, 2016.
Tome, Jacob M., Abdullah Ozer, John M. Pagano, Dan Gheba, Gary P. Schroth, and John T. Lis. "Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling." *Nature Methods*, 2014.
Torarinsson, Elfar, and Stinus Lindgreen. "WAR: Webserver for aligning structural RNAs." *Nucleic acids research*, 2008.
Tuerk, Craig, and Larry Gold. "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase." *Science* 249, no. 4968 (1990): 505-510.
Uhlenbeck, Olke C., Jannette Carey, Paul J. Romaniuk, Peggy T. Lowary, and Dorothy Beckett. "Interaction of r17 coat protein with its rna binding site for translational repression." *Journal of Biomolecular Structure and Dynamics*, 1983.
Vu, M.M.K., 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." *Chemistry and Biology* 19, no. 10 (2012).
Wachter, A., 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." *THE PLANT CELL ONLINE* (American Society of Plant Biologists (ASPB)) 19, no. 11 (11 2007): 3437-3450.
Wakeman, Catherine A., and Wade C. Winkler. "Analysis of the RNA backbone: structural analysis of riboswitches by in-line probing and selective 2'-hydroxyl acylation and primer extension." *Methods in molecular biology (Clifton, N.J.)*, 2009.
Wang, Bin, Kevin A. Wilkinson, and Kevin M. Weeks. "Complex ligand-induced conformational changes in tRNAAsp revealed by single-nucleotide resolution SHAPE chemistry." *Biochemistry*, 2008.
Wang, Qing S., and Peter J. Unrau. "Ribozyme motif structure mapped using random recombination and selection." *RNA*, 2005.
Washietl, Stefan, et al. "Computational analysis of noncoding RNAs." *Wiley Interdisciplinary Reviews: RNA*. 2012.
Washietl, Stefan, Ivo L. Hofacker, Peter F. Stadler, and Manolis Kellis. "RNA folding with soft constraints: Reconciliation of probing data and thermodynamic secondary structure prediction." *Nucleic Acids Research*, 2012.
Watrin, Marguerite, Frederike Von Pelchrzim, Eric Dausse, Renée Schroeder, and Jean Jacques Toulmé. "In vitro selection of RNA aptamers derived from a genomic human library against the TAR RNA element of HIV-1." *Biochemistry*, 2009.
Weinberg, Zasha, et al. "Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes." *Genome Biology*, 2010.
Wickiser, J. Kenneth, Wade C. Winkler, Ronald R. Breaker, and Donald M. Crothers. "The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch." *Molecular Cell*, 2005.
Winkler, W. C., S. Cohen-Chalamish, and R. R. Breaker. "An mRNA structure that controls gene expression by binding FMN." *Proceedings of the National Academy of Sciences*, 2002.
Woo, So Yon, et al. "PRR5, a novel component of mTOR complex 2, regulates platelet-derived growth factor receptor β expression and signaling." *Journal of Biological Chemistry*, 2007.

Zaher, Hani S., R. Ammon Watkins, and Peter J. Unrau. "Two independently selected capping ribozymes share similar substrate requirements." *RNA*, 2006.
Zhang, Jinwei, Matthew W. Lau, and Adrian R. Ferré-D'Amaré. "Ribozymes and riboswitches: Modulation of RNA function by small molecules." *Biochemistry.* 2010.