EXPANDING THE RNA WORLD:
IDENTIFYING, SELECTING, AND DESIGNING UNIQUELY STRUCTURED RNAs

by
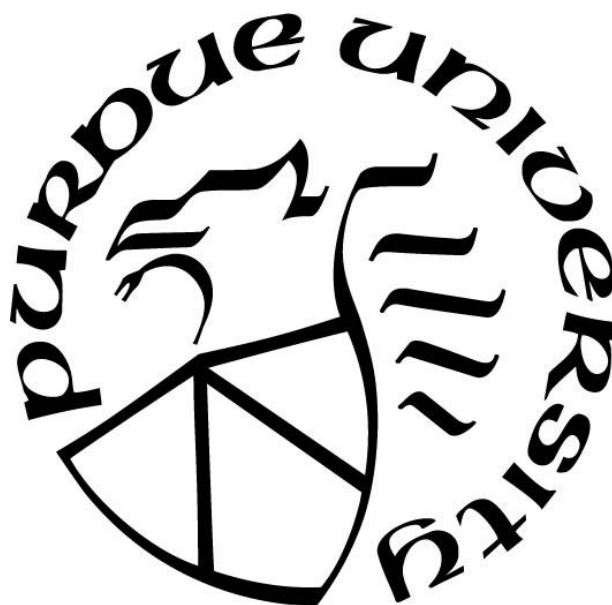
Samantha Waylin Lee

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Biochemistry
West Lafayette, Indiana
December 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Barbara Golden, Chair

    Department of Biochemistry

Dr. Frederick Gimble

    Department of Biochemistry

Dr. Jeremy Lohman

    Department of Biochemistry

Dr. Cynthia Stauffacher

    Department of Biological Sciences

Approved by:

    Dr. Andrew Mesecar

        Head of the Departmental Graduate Program

This work is dedicated to all the women in science that have laid the groundwork for this new era of female scientists and to the next-generation of science-loving girls out there.

Never give up!

ACKNOWLEDGEMENTS

I would like to acknowledge and thank the individuals that contributed to my unforgettable graduate school experience. First and foremost, I would like to thank my advisor, Dr. Barbara Golden for her endless support, both professionally and personally. Without whom, I would not be the scientist that I am today or the scientist that I aspire to become. I'm so thankful that she took a chance on this whacky purple-haired junior scientist so many years ago.

I would also like to express my gratitude to my thesis advisory committee: Dr. Frederick Gimble, Dr. Jeremy Lohman, and Dr. Cynthia Stauffacher. They have devoted a considerable amount of time towards my graduate school career and I will forever appreciate their willingness to chat and lend suggestions in the hallway.

The Golden Lab has always been a small, but tight-knit group. I am extremely grateful for the scientific discussions and more importantly, the friendships that we've built. Thank you to Dr. Ji Chen, Dr. Aamir Mir, Gary Gan, and Emma Lendy. On that note, I would like to acknowledge my friends in the Tran Lab that have generously shared their knowledge on next-generation sequencing library preparation: Dr. Siwen Wang and Dr. Yu-Hsuan (Karen) Lai.

Last but never least, I must express my sincerest appreciation for all the love, encouragement, and laughs from my family, friends, and fur babies throughout these years. A huge thank you to my parents, John and Bihyu, for always being my biggest cheerleaders. To my aunt Betty and uncle George for their continuous love, support, and encouragement. To my sister Cheryl for always being my adventure buddy. To all my puppies that have brought smiles to my face when I most needed them: Bella, Flik, Lady, Wally, and Teddy. Finally, to all the friends I've made during my time at Purdue, the fun times we've had together will never be forgotten.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

5' UTR	5 prime untranslated region
aaRS	aminoacyl tRNA synthetase
ATP	adenosine triphosphate
cDNA	complementary DNA
Chr	chromosome
CoTC	co-transcriptional cleavage ribozyme
CPEB3	cytoplasmic polyadenylation element binding protein ribozyme
dNTPs	deoxyribonucleotide triphosphate mix
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
EMSA	electrophoretic mobility shift
FPLC	fast protein liquid chromatography
HDV ribozyme	hepatitis delta virus ribozyme
HHRz	hammerhead ribozyme
mRNA	messenger RNA
MWCO	molecular weight cutoff
NAD+	oxidized form of nicotinamide adenine dinucleotide
NADH	reduced form of nicotinamide adenine dinucleotide
NGS	next-generation sequencing
nt	nucleotide
NTPs	nucleotide triphosphate mix
PCR	polymerase chain reaction
PEG	polyethelyne glycol
PES	polyethersulfone
RCF	relative centrifugal force
rDNA	ribosomal DNA
RNA	ribonucleic acid
RNase	ribonuclease
RNAseq	RNA sequencing

RPM	revolutions per minute
RT	reverse transcription
SDS	sodium dodecyl sulfate
STARzyme	specific tRNA aminoacyl ribozyme
T-box	tRNA-binding element
Tris-HCl	tris-base buffered with sodium hydroxide
UNKN	unknown genomic region
VS ribozyme	Varkud satellite

ABSTRACT

The cosmos of noncoding RNAs (ncRNAs) has been thriving in recent years; so much so that researchers are discovering them much faster than they can uncover their functions. The subset of these RNAs that have been characterized have been noted to perform and regulate a plethora of remarkably diverse and essential biological functions. This diversity in function is accompanied by a large array of dynamic and elegantly folded 3-dimensional structures. In this collection of work, we will journey through the discovery of the first catalytic noncoding RNAs (ribozymes), explore a new method for identifying uniquely structured ribozymes, and detail the design of a technique to select for highly structured RNAs with a high affinity for an RNA binding partner. Although these topics vary widely within the field of RNA, this work strives to showcase the integral relationship between intricate macromolecular structures with their chemical and cellular functions.

1. RIBOZYMES: PROTEIN OPTIONAL

1.1. Perspective

The perplexing and revolutionary discovery of the first catalytic RNAs shook biological sciences at its core, the central dogma. The long-standing idea that RNA is nothing but a middleman between DNA information storage and the proteins that they encode, was abandoned in the early 1980s with the curious discovery of a self-splicing RNA and the realization that the catalysis carried out by RNase P is actually performed by its RNA subunit. Although originally proposed decades earlier, The RNA World Hypothesis gained little traction prior to the discovery of catalytic RNAs. These RNA enzymes, ribozymes, have since been identified in all domains of life and carry out essential cellular functions such as protein synthesis and gene regulation. In this chapter, we will walk through the unexpected beginnings of ribozyme discovery to the modern biotechnological advances that have lead to *in silico* discovery of new ribozymes.

1.2. Catalytic RNA takes center stage

The story began in 1977 when splicing was first reported by Philip Sharp and Richard Roberts' labs (*1, 2*). At around the same time, Thomas Cech, an assistant professor at the University of Colorado Boulder, was studying extrachromosomal DNA and chromatin structures. These studies led him to the extrachromosomal ribosomal DNA (rDNA) in the ciliated protozoan, *Tetrahymena thermophila*. He soon discovered that the 26S ribosomal RNA contains a sizable intron (*3*). To elucidate the splicing mechanism, his lab created an *in vitro* splicing assay by isolating the nuclei from *T. thermophila* (*4*). When separating the spliced intron from the mature rRNA, Grabowski *et al*. observed that the intron was converted to a circular RNA (*5*). Even more interesting, the cyclic intron contained an additional guanosine that was not encoded by the rDNA, but by an exogenous guanine cofactor necessary for the splicing reaction to occur (*6, 7*). As a result of this work, the hypothesis arose that excision and subsequent circularization of the intron, and ligation of the 26S rRNA exons likely occurs in a single phosphoester-transfer mechanism (*7*). By this time, they had a detailed understanding of the splicing mechanism, but had many questions about the enzyme involved in the reaction. They had assumed the enzyme responsible was highly stable and tightly associated with the rRNA because it resisted comprehensive SDS and protease treatments (*7*).

Unable to detect a protein component, it was proposed that the pre-rRNA itself was responsible for splicing; and by using an *in vitro* transcription system, they showed that the RNA sequence itself is able to self-splice (*8*). Only 3 short years after the Cech lab's first published work on *T. thermophila* rRNA, they reported the first ribozyme (*8*).

About a decade before Thomas Cech's first work on *Tetrahymena*, Sidney Altman was working on tRNA structure and function at the Medical Research Council Laboratory in Cambridge, England. Altman set out to explore the structure-function relationship in tRNAs with addition and deletion mutations, which are inducible by acridine (*9, 10*). It had been shown previously in a pulse-chase experiment that precursor tRNA transcripts are processed into shorter, mature tRNAs containing modifications (*11-13*). When creating the single-nucleotide mutations, he noted that precursor tRNAs were not processed into their shorter, mature form *in vivo*. Knowing that precursor tRNAs are longer than mature ones, he devised a protocol to purify and sequence *E. coli* pre-tRNA^{Tyr} (*14, 15*). By adding wild-type and mutant precursor tRNAs to *E. coli* extract borrowed from Hugh Robertson, Altman showed that tRNA maturation proceeded with wild-type precursor tRNAs but not with the mutant precursor tRNAs (*16*). Robertson, having purified and characterized *E. coli* RNase III, was an expert in ribonucleases (*17*). Together, they identified the novel RNAse that specifically clips the 5' leader sequence off of *E. coli* tRNAs, which they named RNase P (*16*). Altman continued his work on RNase P as a faculty member at Yale University, where his lab discovered that the RNase P co-purifies with an RNA component and that the RNA is essential for catalysis (*18*). The year after the Cech lab published their discovery of self-splicing RNA, the Altman lab reported that RNase P is a ribozyme (*19*)!

1.3. Ribozymes are ubiquitous

The decade following Thomas Cech and Sydney Altman's revelations, the field of ribozymes was continuing full steam ahead. In 1986, a new self-splicing RNA was described in a yeast mitochondrial intron. The observation of the accumulation of circular introns *in vivo* pointed towards a self-splicing intron. It was discovered that these new group II introns do not require GTP for self-splicing (*20*). Additionally, the ligation junction of the circular intron is joined by a distinct 2'-5' phosphodiester bond, creating a branched lariat rather than a circle (*20*). This type of lariat was previously observed in pre-mRNA splicing (*21, 22*). This comparison has lead to the

hypothesis that the spliceosome may be a ribozyme or may have evolved from group II introns. In the realm of self-splicing introns, a more complex variation of the group I intron includes the lariat capping ribozyme that was identified while investigating the extrachromosomal rDNA of the protist, *Didymium iridis*. This complex intron encodes two distinct ribozymes, a canonical group I intron along with a group I-like intron that participates in mRNA maturation (*23*).

The same year that the group II intron was reported, two independent labs reported self-cleavage and ligation of plant virus satellite RNAs and viroids. Both of these examples in the avocado sunblotch viroid and the satellite RNA of the tobacco ringspot virus use the hammerhead ribozyme to separate monomeric copies of their RNA genomes from the multimeric transcript created through rolling circle replication (*24-26*). Since then, thousands of hammerhead ribozymes have been identified in all domains of life (*27-29*). Similar in function to the hammerhead ribozyme, is the hairpin ribozyme. They both function in rolling circle replication of plant virus satellite RNAs, but differ in structure and sequence (*24, 30-32*). The human hepatitis delta virus (HDV) is the human pathogen-associated counterpart to plant virus satellite RNAs. It is a subviral satellite of the human hepatitis B virus and the only ribozyme known to be necessary for the propagation of a human pathogen (*33-35*). A personal favorite in terms of structure, the HDV ribozyme sequence folds into a sophisticated double-nested pseudoknot structure (*36, 37*). The Varkud satellite (VS) ribozyme is another group of self-cleaving ribozyme found in the mitochondria of *Neurospora*. Structurally distinct from the aforementioned nucleolytic RNAs, it has been hypothesized that VS DNA is a type of retrotransposon because it encodes a reverse transcriptase (*38, 39*). All of these self-cleaving ribozymes catalyze a reversible cleavage and ligation reactions, where the termini of the cleavage products carry a 2', 3'-cyclic phosphate and a 5' hydroxyl.

1.4. The ribosome is a ribozyme

The ribosome, while first described in the 1943 by Albert Claude, as a "ribonucleic acid in the cytoplasm on particulate or granular structures" and later observed by George Palade and Phillip Siekevitz by electron microscopy in 1956, has eluded scientists for decades in regards to its catalytic mechanism for peptide synthesis (*40-46*). Many electron micrographs, cryoelectron microscopy images, and x-ray crystallographic structures have shed light on the overall

architecture of the massive multi-subunit ribonucleoprotein complex. But it wasn't until 2000 when Thomas Steitz's lab released a 2.4 Å crystal structure of the large ribosomal subunit from *Haloarcula marismortui* that the mystery of peptidyl transfer reaction, that is the essence of protein synthesis, was solved (*47, 48*). The observation of an all-RNA active site, with no amino acid side-chains within 18 Å of the site of the peptidyl transfer reaction led to the conclusion that the 23S rRNA is responsible for catalysis (*47*). Previous hypotheses suggested that the role of rRNA was to act as a scaffold for the multitude of ribosomal proteins to dock. We now know that it is quite the opposite. The ribosomal proteins provide structure and stability to the catalytic rRNA.

1.5. Riboswitches as ribozymes

Prior to the discovery of the first natural riboswitch, nucleic acid aptamers were being selected and evolved *in vitro* to bind a plethora of molecules. In 2002, the nature's first aptamer was revealed; an RNA regulatory element in bacteria that could detect metabolites and toggle gene expression in response to the metabolic state of the cell. Direct metabolite binding of the vitamin-derivates thiamine pyrophosphate, flavin mononucleotide, and adenosylcobalamin to their respective mRNAs was shown to induce conformational changes in RNA structure that in turn, would alter gene expression through transcriptional termination or translational initiation (*49-51*). These riboswitches were soon found to bind amino acids, purines, cofactors, metals, anions, and even other RNAs! tRNA riboswitches will be discussed further in Chapters 3 and 4. Two years following the first reports of riboswitches, a riboswitch-ribozyme hybrid was identified. The peculiar *glmS* riboswitch in *Bacillus subtilus* combines the gene regulation ability of a riboswitch with the self-cleavage activity of a nucleolytic ribozyme by using the metabolite that it binds as a cofactor (*52*). Specifically, the *glmS* gene encodes glutamine-fructose-6-phosphate amidotransferase, an enzyme that creates glucosamine-6-phosphate (GlcN6P) (*53*). The riboswitch domain embedded in the 5' untranslated region of *glmS* has the ability to sense GlcN6P by direct binding, which induces cofactor-mediated self-cleavage of the transcript and subsequent degradation (*52, 54*). Further complexity was added to this negative feedback mechanism when it was discovered that the GlcN6P binding site of the riboswitch also binds a variety of hexose compounds to inhibit GlcN6P binding and increase *glmS* mRNA abundance (*55*). The ability of this riboswitch to respond to the overall metabolic state of the cell is the first of its kind to be described. To date, the repertoire of *glmS* riboswitch-ribozymes with highly conserved structure

and consensus sequences has been expanded to include at least 18 gram-positive bacteria species (*56, 57*).

1.6. Data mining for ribozymes

Modern advances in whole genome and transcriptome sequencing, coupled with the power of computational analysis have led to the discovery of 4 new structural classes of self-cleaving ribozymes. The first class of ribozymes to be discovered by bioinformatics searches was the twister ribozyme (*58*). By using the simple fact that multiple hammerhead ribozyme sequences will often occur adjacent to one another, a comparative sequence and secondary structure search was employed on the genes and neighboring sequences that are known to encode hammerhead ribozymes (*28, 58*). Although twister ribozymes are functionally synonymous to hammerhead ribozymes, they have their own unique structure and sequence conservations (*58, 59*). By exploiting the same technique used to identify twister ribozymes, 15 other conserved RNA motifs were pinpointed (*60*). When tested for *in vitro* self-cleavage activity, only three showed catalytic activity and were named twister sister ribozyme, hatchet ribozyme, and pistol ribozyme (*60*).

1.7. Objectives

Ribozyme discovery began with the functional characterization of splicing and the structure-function relationship in tRNAs, where researchers stumbled upon the revolutionary discovery that RNA can act as a catalyst. Following this revelation, ribozymes started popping up everywhere. After the initial flood of discoveries, identification of new ribozyme classes simmered. The rise of high-throughput computational analysis revealed several new ribozymes with unique structures. Most important, these studies demonstrated that there are more ribozymes buried within the genomes of living organisms. The work described in Chapter 2 details the development of an assay for selective sequencing of 2', 3'-cyclic phosphate-terminal RNAs. This method combines a traditional biochemical approach with modern next-generation sequencing and bioinformatics analysis. The objective of developing such an assay is to uncover new classes of ribozymes, which is detailed in Chapter 3. Finally, looking to further expand the repertoire of structured RNAs, we dabble into *in vitro* selection of RNA-binding aptamers.

2. A METHOD FOR SELECTIVE SEQUENCING OF CYCLIC PHOSPHATE-TERMINAL RNAs

2.1. Declaration of collaborative work

The work described in this chapter is the result of a collaborative effort in the Golden lab. Samantha Lee performed the work described in this chapter and is the author of the chapter and its figures. However, Dr. Aamir Mir was instrumental to the early stages of this project. He established the protein purification and 2', 3'- cyclic phosphate ligation protocols. He also played a major role in the development of the library construction protocol and performed many of the earlier versions of library construction and subsequent sequencing analysis. I would also like to acknowledge Dr. Elizabeth Tran, Dr. Siwen Wang, and Dr. Yu-Hsuan Lai for their helpful feedback on NGS library construction and sequencing analysis. I would like to thank Dr. Catherine Hill and her lab for providing mosquito specimens, Dr. Vikki Weake and her lab for providing *D. melanogaster* starter cultures and materials, and Dr. James Clemens for his guidance with *Drosophila* cultivation.

2.2. Introduction

The universe of noncoding RNAs (ncRNAs) has been thriving in recent years; so much so that researchers are discovering them faster than they can uncover their functions. The subset that has been characterized have been noted to perform and regulate a plethora of remarkably diverse biological functions. Familiar ncRNAs such as microRNA (miRNA), piwi-interacting RNA (piRNA), small interfering RNA (siRNA), and long noncoding RNA (lncRNA) are involved in various levels of gene regulation. While guide RNA (gRNA) has been at the forefront recently with the popularity of CRISPR gene editing. Small nucleolar RNAs (snoRNA) orchestrate the modification of other RNAs; such as small nuclear RNA (snRNA), ribosomal RNA (rRNA), transfer RNA (tRNA). These modifications add stability and activity to many ncRNAs. For example, the U6 snRNA contains a stably capped 2', 3'-cyclic phosphate. This post-translational modification is necessary for stability and activity of the spliceosome. This unique 3'-terminus can occur in the absence of translational modification, such as the cleavage products of small self-cleaving ribozymes and tRNA splicing intermediates.

The modern collection of ncRNAs includes small RNAs derived from introns like the stable intronic sequence RNA (sisRNA) found in herpesviruses and the circular RNAs (circRNA) that structurally resemble pathogenic viroids. Among other curious ncRNAs are the tRNA-derived small RNA (tsRNA), double-stranded break-induced small RNA (diRNA), centromeric RNA (cenRNA), toxic small RNA (tsRNA), small conjugation specific RNA (scnRNA), transcriptional enhancer element (eRNA), and a Y RNA derivative (YsRNA). This is to just name a handful and the list continues to grow. The compelling question is how are these ncRNAs able to evade degradation by ribonucleases? Is it possible that they contain modifications? Perhaps by using a 2', 3'-cyclic phosphate cap? Take the tRNA-derived tsRNA for example. These small ncRNAs originate from tRNAs that are halved at the anticodon loop to form epigenetically inheritable regulatory molecules. tRNA splicing intermediates are also clipped at the anticodon loop and the 5' half will temporarily possess a 2', 3'-cyclic phosphate. These may be mere coincidences, but it certainly poses interesting questions.

Here, we have designed a method for selective sequencing of 2', 3'-cyclic phosphate terminal RNAs by combining a biochemical selection with next-generation sequencing techniques. We capitalized on the relaxed specificity of *Arabidopsis thaliana* tRNA ligase to select for terminal cyclic phosphates then employed a customized, rigorous cDNA library preparation protocol.

2.3. Methods

2.3.1. Expression and purification of A. thaliana tRNA ligase and S. cerevisiae Tpt1

The *A. thaliana* tRNA ligase gene in pET-28a and *S. cerevisiae TPT1* in pET-53 were obtained from Dr. Stanley Fields lab at the University of Washington (*61*). Plasmids were transformed into BL21-CodonPlus(DE3)-RIPL competent cells (Stratagene). Single colonies were inoculated into 500 mL of Luria Broth containing 100 μg/mL ampicillin and 35 μg/mL chloramphenicol. Cultures were grown in a shaking incubator at 37°C until $OD_{600} = 0.6$. The cultures were placed on ice for 30 minutes, then ethanol was added to 2% and isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to 0.4 μM. The cultures were shaken in an 18°C incubator for 20 hours before harvesting, followed by freezing at -80°C. Cell pellets were thawed on ice and resuspended in Lysis Buffer (50 mM Tris-HCl at pH 8.0, 100 mM NaCl, 1 mM imidazole, and 5 mM β-mercaptoethanol) and

rotated for 30 minutes at 4°C. Phenylmethanesulfonyl fluoride (PMSF) was added to 1 mM prior to sonication, then Triton X-100 was added to 0.1%, and the lysate was centrifuged for 30 minutes at 20,000 RCF. The supernatant was incubated with 2 mL of equilibrated Nickel-NTA Agarose (Qiagen) and rotated for 2 hours at 4°C. Beads were centrifuged for 2 minutes at 100 RCF and the collected beads were washed twice with 25 mL Wash Buffer (50 mM Tris-HCl at pH 8.0, 100 mM NaCl, 10 mM imidazole, and 5 mM β-mercaptoethanol). Washed beads were poured into an Empty Disposable PD-10 column (GE Healthcare Life Sciences). Elution Buffer (50 mM Tris-HCl at pH 8.0, 100 mM NaCl, 300 mM imidazole, and 5 mM β-mercaptoethanol) was added and twelve 1 mL fractions were collected. The peak fractions were pooled and dialyzed overnight using Spectra/Por membrane MWCO: 12-14,000 (Spectrum Laboratories) with 1 L of Dialysis Buffer (50 mM Tris-HCl at pH 7.5, 100 mM NaCl, 0.1 mM EDTA, and 0.1 mM DTT).

2.3.2. Total RNA isolation

Using TRIzol Reagent (Invitrogen) and Direct-zol RNA MiniPrep Plus (Zymo Research), total RNA was extracted from third instar *D. melanogaster* Oregon-R, fourth instar *A. aegypti* Liverpool, and fourth instar *A. gambiae* PEST larvae according to manufacturer's protocol. An additional phenol:chloroform extraction was performed to remove residual contaminants.

2.3.3. Ligation of 2', 3'-cyclic phosphate terminal RNAs

200 μg of total RNA was incubated with Ligation Buffer (2 mM Tris-HCl at pH 7.5, 20 mM potassium acetate, 60 μM spermine, 1.2 mM magnesium acetate, 100 μM DTT, and 0.1% Triton X-100), 12.5% PEG4000, 1 μM 5' preadenylated and 3' biotinylated TruSeq Universal adapter, and 32 μM of *A. thaliana* tRNA ligase for 3 hours at 30°C. To remove excess adapter, the ligation reaction was electroporated on a 6% polyacrylamide gel containing 7 M urea. The gel area containing RNA with higher molecular weights than the adapter was excised with a clean razor blade crushed though a syringe, and frozen in 10 mM Tris-HCl at pH 7.5, 1 mM EDTA, and 250 mM NaCl. The gel mixture was thawed and rocked overnight at 4°C. The gel pieces were removed by filtration through a PES membrane, followed by ethanol precipitation of the eluted RNA.

2.3.4. Biotin pull-down for enrichment of adapter-ligated RNAs

Dynabeads MyOne Strepavidin C1 (Invitrogen) was used according to manufacturer's protocol to enrich for adapter-ligated RNAs from the gel purified RNA pool.

2.3.5. Tpt1 reaction

To remove the 2' phosphate at the ligation-junction, the RNA recovered from the streptavidin-biotin pull-down was incubated with 20 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 2.5 mM spermidine, 100 μM DTT, 0.4% Triton X-100, 1 mM NAD^+, and 20 μM *S. cerevisiae* Tpt1 for 2 hours at 30°C (*61*). The reaction was stopped by phenol:chloroform extraction followed by ethanol precipitation.

2.3.6. RNA fragmentation

RNA was fragmented using Ambion RNA Fragmentation Reagents (Invitrogen) for 30 seconds at 70°C. The fragmented RNA was ethanol precipitated and resuspended in 21 μL of H_2O.

2.3.7. Reverse transcription

Fragmented RNA was mixed with 1 μM RT primer and heated at 95°C for 2 minutes then cooled at room temperature for 10 minutes. TIGRT buffer (20 mM Tris-HCl pH 7.5, 200 μM NaCl, 10 mM $MgCl_2$, and 5 mM DTT) and 2 μL TGIRT-III enzyme (InGex) was mixed and incubated at room temperature for 30 minutes (*62-64*). At which point, 1.25μM dNTPs was added and the reaction mixture was incubated at 60°C for 1 hour. Reaction components were removed by phenol:chloroform extraction followed by ethanol precipitation.

2.3.8. Removal of terminal phosphates and RNA

cDNA was incubated with 8 units of rSAP and CutSmart Buffer (New England BioLabs) for 15 minutes at 37°C. rSAP reaction was incubated with 5 units of RNase H (ThermoFIsher Scientific) and 20 μg of RNase A (Invitrogen) for 30 minutes at 37°C. The reaction was stopped by phenol:chloroform extraction followed by ethanol precipitation.

2.3.9. Index adapter ligation

Single-stranded cDNA was incubated with 1 uM TruSeq Index Adapter, 10 units of T4 RNA Ligase (New England BioLabs), reaction buffer, 25% PEG8000, and 25 μM ATP for 16 hours at 25°C. To remove excess adapter, the ligation reaction was run on a 6% polyacrylamide gel containing 7 M urea. The gel area containing RNA with higher molecular weights than the adapter was viewed by UV shadowing, excised using a clean razor blade, crushed though a syringe, then frozen in 10 mM Tris-HCl at pH 7.5, 1 mM EDTA, and 250 mM NaCl. The gel mixture was

thawed and rocked overnight at 4°C. The gel pieces were removed by filtration through a PES membrane, followed by ethanol precipitation of the eluted RNA.

2.3.10. Library enrichment

Single-stranded cDNA libraries were PCR enriched with Phusion High-Fidelity PCR Kit (New England BioLabs). 20 cycles were completed according to manufacturer's thermocycling conditions with an annealing temperature at 70.7°C. To remove excess primer and primer dimers, the PCR reaction was run on a 6% polyacrylamide gel containing 7 M urea. The gel area containing DNA with higher molecular weights than the primer dimers was excised, crushed though a syringe, and rocked overnight at 4°C in 10 mM Tris-HCl at pH 7.5, 1 mM EDTA, and 250 mM NaCl. The gel pieces were filtered for removal, followed by ethanol precipitation of the eluted RNA. The gel extraction was repeated to remove residual low molecular weight species that may interfere with sequencing.

2.3.11. Next-generation sequencing

Illumina MiSeq 2x150, 300 cycles, paired-end sequencing was performed by the Purdue Genomics Core Facility using sequencing primers complementary to the TruSeq Universal adapter and the 3' end of the TruSeq index adapters.

2.3.12. Computational analysis of RNAseq data

All computational analysis was performed on the Snyder Purdue Community Cluster. Using FASTX-Toolkit and Trimmomatic, the raw reads from the MiSeq 2x150 runs were quality clipped (> Phred Quality Score 20), adapter trimmed, and reads less than 30 nt were removed (*65, 66*). The filtered reads were mapped to the April 2014/June 2017 version of the *Aedes aegypti* genome (AaegL3/AaegL5.0), the May 2014 version of *Drosophila melanogaster* genome (BDGP6), or the April 2014 version of *Anopheles gambiae* genome (AgamP4) using Bowtie 2 version 2.3.2 (*67*). The genomic assemblies and annotation files (gff3) were downloaded from Ensembl and VectorBase databases (*68-70*). An analysis pipeline to create a list of peaks was carried out using a combination of SAMtools, BamTools, BEDTools, and Bioconductor (*71-74*). For our purposes, a peak is defined as any defined region that contains one or more mapped read(s). The output of this analysis pipeline was an ordered list of genomic locations starting with the locations containing the highest coverage per base. Data was visualized using Integrative Genomics Viewer

version 2.3.92 (*75, 76*). Peak calling and cleavage site prediction was performed manually to search for the distinct drop-off cliff-shaped peaks observed at 2', 3'-cyclic phosphate-to-adapter ligation junctions. The genomic coordinate of the 2', 3'-cyclic phosphate was mapped based on the location of the sharp drop-offs. To cross-reference the information contained in the annotation files, the sequences of the top peaks were fed into a BLAST (Basic Local Alignment Search Tool) (*77*).

2.3.13. Oligonucleotides

Oligonucleotides were synthesized by Integrated DNA Technologies.

Table 2.1 TruSeq adapter sequences. 6-nucleotide barcode sequence shown in blue. 5rApp denotes 5' adenylation, 3ddC denotes 3' dideoxycytidine, and 3Bio denotes 3' biotin.

Oligo name	Sequence 5'-3'
TruSeq 1 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 2 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 3 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 4 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTT3ddC

Table 2.1 continued

TruSeq 5 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 6 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 7 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 8 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 9 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 10 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 11 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq 12 Adapter	5rApp/NNNNNGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTT3ddC
TruSeq Universal Biotin	5rApp/NNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT/3Bio/

Table 2.2 Oligonucleotides for reverse transcription and library enrichment

Oligo name	Sequence 5'-3'
RT primer	AATGATACGGCGACCACCGAGATCTACACTCTTTC
Forward primer	AATGATACGGCGACCACCGAGATCTAC
Reverse primer	CAAGCAGAAGACGGCATACGA

2.4. Results

2.4.1. Analysis of RNA sequencing results

Sequencing libraries were created using the total RNA extracted from three species of Diptera: *A. gambiae*, *D. melanogaster*, and *A. aegypti*. Full chromosomally assembled and well-annotated genomes of *A. gambiae and D. melanogaster* resulted in relatively simple paired-end alignment of the reads (Table 2.3).

The *A. aegypti* genome is five times larger than that of *A. gambiae* (*70*). This, in combination with the large number of highly repetitive genomic elements in *A. aegytpi*, poses issues when creating a high quality chromosomal assembly. At the time of this work, the *A. aegypti* genomic assembly, AaegL3, consisted of 4,757 scaffolds. Large numbers of scaffolds create problems with read mapping, which is evident in the low overall alignment rates across all *A. aegypti* samples (Table 2.4). Both pared-end and single-end alignments were performed on both raw reads and filtered reads to provide a comparison of overall alignment rates. The datasets highlighted in red were chosen for further analysis. Overall alignment rates of paired-end versus single-end are similar (within 2-5%). Paired-end alignment was chosen because the method provides superior alignment across regions containing repetitive sequences. With the *A. aegypti* genome consisting of approximately 50% repetitive transposable elements, alignment accuracy is paramount (*78, 79*).

Table 2.3 Alignment summary for *A. gambiae* and *D. melanogaster.* Paired-end alignment using Bowtie 2 (v2.3.2) to BDGP6 (Ensembl release 35) or to AgamP4 (Ensembl release 35).

Organism	Total reads	Adapter trimmed & Quality clipped reads	% Reads lost	Aligned reads	Overall alignment rate
A. gambiae	2,364,042	289,800	87%	119,363	41.19%
D. melanogaster	2,952,997	1,102,350	62%	580,670	52.68%

Table 2.4 Comparative analysis of *A. aegypti* datasets when using various alignment methods.
All alignments performed using Bowtie 2 (v2.3.2) to AaegL3 (Ensembl release 35).

Sample ID	Total reads	Adapter trimmed & Quality clipped reads	% Reads lost	Alignment type	Aligned reads	Overall alignment rate
Aae01	2,628,906	675,278	74%	*Paired-end	190,594	28.22%
				*Single-end	203,776	30.18%
				^Paired-end	567,695	21.59%
				^Single-end	709,986	27.01%
Aae02	1,194,770	645,082	46%	*Paired-end	66,603	10.32%
				*Single-end	71,672	11.11%
				^Paired-end	293,174	24.54%
				^Single-end	349,282	29.23%

*Alignment performed using Adapter trimmed & Quality clipped reads (filtered reads)
^Alignment performed using Total reads (raw reads)

The current assembly, AaegL5, contains roughly half the number of scaffolds: 2,310 (*70*). The other 2,447 scaffolds have been assembled into three distinct chromosomes. We would expect the overall alignment rate to AaegL5 to increase; yet we observe a decrease (Table 2.5). The likeliest explanation for this is the removal of duplicate repetitive sequences that flank the scaffolds. These flanking sequences are necessary for genome assembly and act much like unique edges to puzzle pieces. These flanking sequences overlap when assembled so are essentially duplicated before scaffolds are assembled into chromosomes. When the AaegL5 released, a complete annotation file did not accompany it. The lack of a complete annotation file hindered further analysis. The remainder of the analysis was completed using alignments to the AaegL3 assembly.

Table 2.5 Comparative analysis of *A. aegypti* overall alignment rates to two different genomic assemblies. All alignments performed using Bowtie 2 (v2.3.2) to AaegL3 or AaegL5.0.

Sample ID	Total reads	Adapter trimmed & Quality clipped reads	% Reads lost	Reads used in alignment	Overall alignment rate to AaegL3	Overall alignment rate to AaegL5
Aae01	2,628,906	675,278	74%	Filtered	28.22%	3.60%
				Raw	21.59%	16.61%
Aae02	1,194,770	645,082	46%	Filtered	10.32%	2.94%
				Raw	24.54%	21.55%

2.4.2. RNAs with a 2', 3'-cyclic phosphate terminus were enriched compared to other RNAs

Direct evidence that our method is specifically selecting 2', 3'-cyclic phosphate capped RNAs can be seen when examining spliceosomal snRNAs. The U6 snRNA contains a stably capped 2', 3'-cyclic phosphate (*61, 80*). By comparison, other snRNAs such as U2 and U5 snRNAs do not posses this modification, but exist in an equimolar ratio to U6 snRNA. Figure 2.1 shows that reads mapping to U6 snRNAs across all samples are greatly enriched over other spliceosomal snRNAs. Notably, no U5 snRNA background was detected in *D. melanogaster*. In *A. aegypti* 02, U6 snRNA levels were 461 times higher than U2 snRNA background, demonstrating that our RNAseq method is specifically enriching 2', 3'-cyclic phosphate terminal RNAs.

Figure 2.2 displays genomic views of mapped reads. Genomic coordinates and gene annotation is displayed on the X-axis and number of reads mapped is shown on the Y-axis. Thus, the amplitude of the peak corresponds to the number of reads mapping to a specific genomic coordinate. A particularly unique feature is observed in the shape of these peaks. An abrupt, cliff-like drop-off is consistently observed at the 3' end of the U6 snRNA gene, the location of the 2', 3'-cyclic phosphate. The sudden drop in base coverage can be attributed to the ligation junction of the TruSeq Universal sequencing adapter to the 2', 3'-cyclic phosphate. All datasets displayed a similar trend.

Comparison of snRNA read counts

	D. melanogaster	A. gambiae	A. aegypti 01	A. aegypti 02
U6 snRNA	1621	714	1170	1845
U5 snRNA	0	49	1	4
U2 snRNA	8	2	12	3
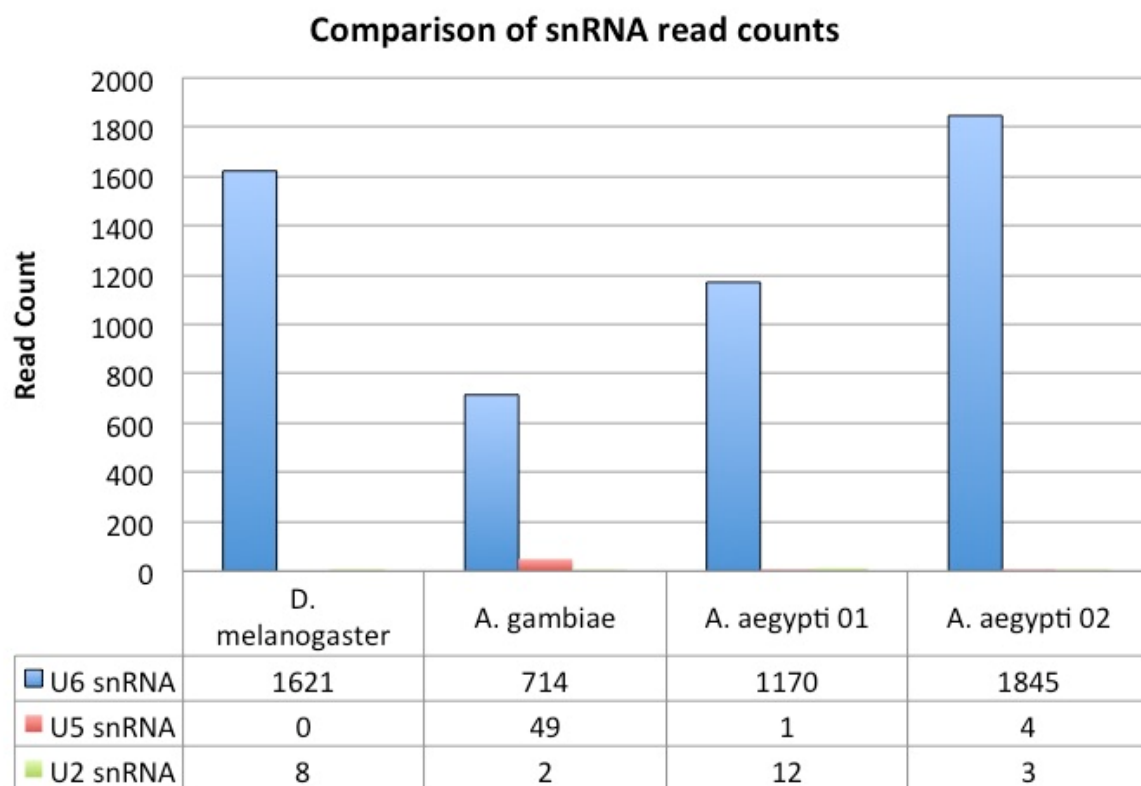
Figure 2.1 U6 snRNA is highly enriched relative to other spliceosomal RNAs
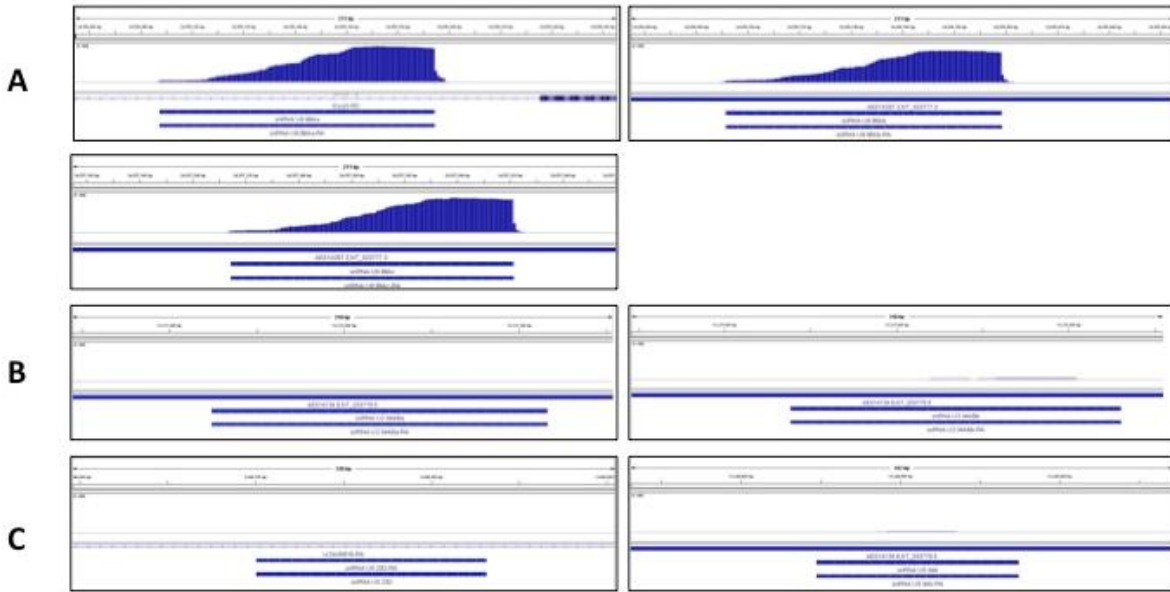
Figure 2.2 Genomic view and coordinates snRNAs in *D. melanogaster*. (A) Genomics view of all three copies of U6 snRNA (B) Select U5 snRNA genome locations (C) Select U2 snRNA genome locations. Data viewed and images taken on Integrative Genome Viewer 2.6.3. This is a selection of snRNA genomic views for *D. melanogaster*.

2.4.3. Analysis of the top peaks

Peak calling is traditionally performed computationally to identify genomic areas that are enriched with a higher number of aligned reads compared to the rest of the genome. Because we performed a selective RNAseq experiment and did not observe full genomic coverage after read alignment, manual peak calling was performed. As described in Section 2.3.4, an ordered list of genomic locations containing the highest coverage per base was generated. For *A. gambiae* and *D. melanogaster*, the annotated locations of the top 100 peaks largely skew towards unassembled regions of the genomic assembly, i.e. UNKN and contigs/unmapped scaffolds (Figure 2.3 and 2.4). As mentioned previously, difficulties in complete genome assembly arise from highly repetitive sequences, which coincidentally, are where self-cleaving ribozymes are likely to reside (*28, 79, 81*). This innate quality presents challenges in extracting full-length sequences to test for *in vitro* self-cleavage activity. Analysis could not be performed for *A. aegypti*, due to the lack of a chromosome-level genomic assembly with accompanying annotations. Additionally, genomic copies of U6 snRNA were in high abundance among the top 30 peaks in all 3 datasets. In *D.*

melanogaster, all 3 genomic copies were present in the top 30 peaks. For *A. gambiae*, 2 of the 5 copies were detected, while 11 out of 14 were detected in *A. aegypti*.
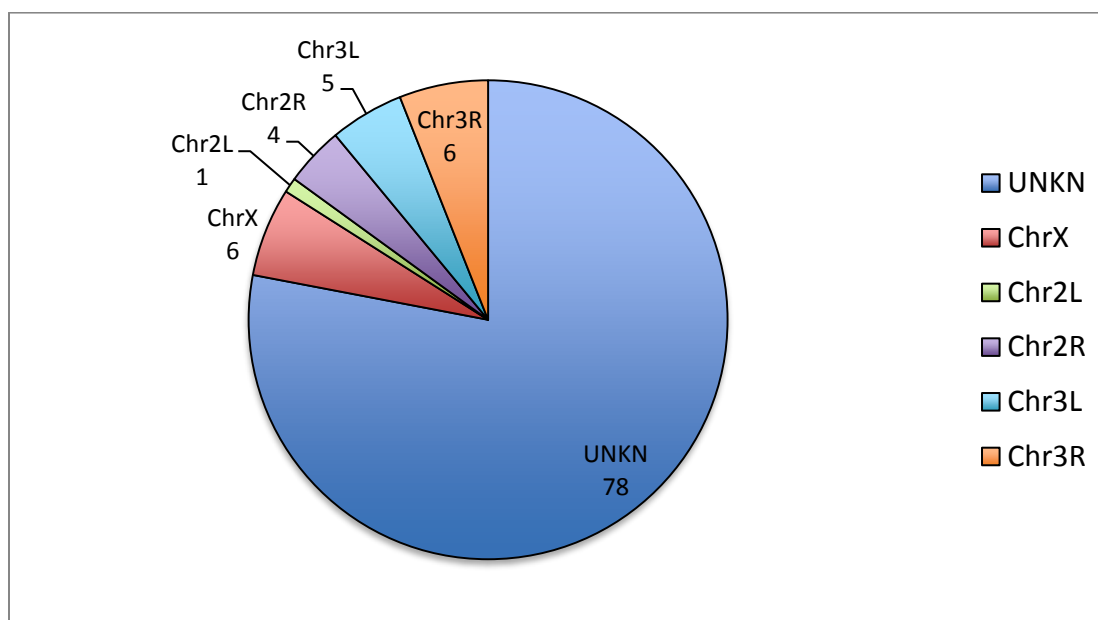
Figure 2.3 The majority of the top 100 peaks in *A. gambiae* map to unknown regions of the genome. The number of peaks mapping to each chromosome or to unknown regions.
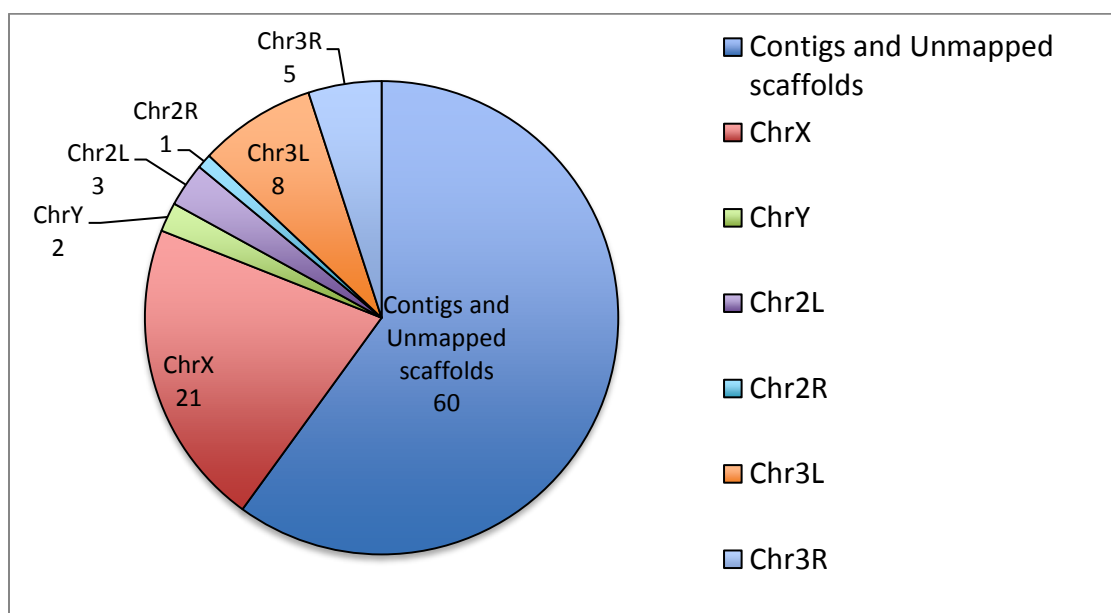
Figure 2.4 The majority of the top peaks in *D. melanogaster* map to contigs and unmapped scaffolds. The number of peaks mapping to each chromosome or to unmapped scaffolds.

2.5. Discussion

It is evident that our method is specifically selecting 2', 3'-cyclic phosphate capped RNAs, as shown by the comparisons of the stably capped U6 snRNA relative to other spliceosomal RNAs that are present in the cell at a 1:1 ratio. This end-modified snRNA is responsible for stability of the spliceosome and assembly of the ribonucleic protein complex. It is the key component of the spliceosome and the absence of the 2', 3'-cyclic phosphate terminus caused by the deficiency of its biogenesis factor directly leads to a human autosomal recessive skin disease. Additionally, instability of the spliceosomal complex can lead to lateral sclerosis and spinal muscular atrophy or be lethal. This further solidifies the importance of this unique RNA terminus.

Of the ncRNAs mentioned previously, some have more obvious methods of avoiding digestion by cellular ribonucleases. For instance, cyclizations of introns by the spliceosome form circRNAs. These are proposed to regulate gene expression and new evidence has emerged suggesting that they may also encode for protein (*82*)! The cyclic structure protects the 3' and 5' ends from degradation. Many proteins also benefit from RNA binding proteins (RBP) to shield themselves from being dismantled. In the case of sisRNAs, there has been little evidence to their structure or source of stability. It has been proposed that these sisRNAs may act as protectors for miRNAs due to their size and hairpin-like structure.

Coding and non-coding RNAs have evolved an assortment of mechanisms to guard them from degradation, some complex while others only require a single cap. As the noncoding RNA world expands further and we gather knowledge of their structures and functions, we will likely meet the familiar 2', 3'-cyclic phosphate RNAs again.

3. COMBINING BIOCHEMICAL SELECTION WITH IN SILICO TECHNIQUES TO DETECT UNIQUE SELF-CLEAVING RIBOZYMES

3.1. Introduction

Since the discovery of self-cleaving ribozymes over 30 years ago, only 9 structurally diverse classes have been identified. These include the hammerhead ribozyme, hairpin ribozyme, human hepatitis delta virus (HDV) ribozyme, *glmS* riboswitch-ribozyme, twister ribozyme, twister sister ribozyme, pistol ribozyme, and hatchet ribozyme. These classes of nucleolytic ribozymes are highly conserved in structure, with few conserved nucleotides. The commonality with all of these ribozymes is their ability to catalyze site-specific intramolecular cleavage of a phosphodiester bond in the backbone of the RNA, which results in products containing 5'-hydroxyl and 2', 3'-cyclic phosphate termini. Some of these ribozymes have also been shown to catalyze the reverse ligation reaction.

Recently, several computational studies have shown these self-cleaving ribozymes to be widely distributed throughout all domains of life (*29, 58, 83*). So abundant in nature that the twister, twister sister, pistol and hatchet ribozyme classes were discovered purely through bioinformatics and data mining in recent years (*58, 60*). By searching for conserved sequences and secondary structure motifs near genes embedded with hammerhead ribozymes, putative ribozyme sequences were identified and activity was verified *in vitro (58, 60).* Although many of the ribozyme candidates were not confirmed to have self-cleavage activity, these studies demonstrated the abundance of novel ribozymes buried within the genomes of a multitude of diverse organisms.

The 2', 3'-cyclic phosphate terminus of self-cleaving ribozyme cleavage products is rather unique when compared to other cellular RNAs that typically terminate with a 3'-hydroxyl. Few RNAs posses these cyclic phosphates caps: the U6 spliceosomal RNA (U6 snRNA), tRNA splicing intermediates, and products of self-cleaving ribozymes (*25, 84, 85*). All of these RNAs are of interest to us. To elaborate, tRNA introns are removed through a non-spliceosomal pathway that generates 2', 3'-cyclic phosphate intermediates before ligation of the exons. tRNA ligases are multifunctional enzymes with the ability to catalyze the ligation of 2', 3'-cyclic phosphate terminal RNAs created by the tRNA splicing endonuclease. These tRNA ligases have been studied at length

and have been shown to possess varying levels of substrate specificity (*86-89*). Previously, it has been demonstrated that the relaxed specificity of *Arabidopsis thaliana* tRNA ligase can be utilized for the ligation of a short adapter oligonucleotide to the cleavage product of the 10-23 deoxyribozyme, which contains a terminal 2', 3'-cyclic phosphate (*61*). This suggests an interesting method for the detection of new ribozymes from total RNA, with the U6 snRNA serving as an internal control. By combining the relaxed specificity of *A. thaliana* tRNA ligase with affinity enrichment, next-generation sequencing, and secondary structure searches, we embarked on a self-cleaving ribozyme hunting expedition.

Aedes aegytpi, the notorious Zika and Dengue virus vector encodes over a thousand copies of the hammerhead ribozyme (*27-29*). It's close relative, the malaria vector, *Anopheles gambiae* harbors at least 9 copies of HDV-like ribozymes (*83*). Several *Drosophila* species have also been shown to carry the R2 ribozyme, a group of HDV-like ribozymes specifically residing within the long terminal repeat R2 retrotransposons (*81, 90, 91*). Using the total RNA from these 3 ribozyme-harboring organisms and building upon previous work on the capture and sequencing of 2', 3'-cyclic phosphate RNAs, we were able to detect annotated ribozymes and new genomic copies of putative hammerhead and HDV-like ribozymes. Additionally, we have compiled extensive lists of novel ribozyme candidates.

3.2. Methods

3.2.1. RNAseq library preparation and next-generation sequencing

cDNA libraries for this RNAseq experiment were prepared as described in Section 2.3.

3.2.2. Computational analysis of RNAseq data

All computational analysis was performed on the Snyder Purdue Community Cluster. Using FASTX-Toolkit and Trimmomatic, the raw reads from the MiSeq 2x150 runs were quality clipped (> Phred Quality Score 20), adapter trimmed, and reads less than 30 nt were removed (*65, 66*). The filtered reads were mapped to the April 2014/June 2017 version of *Aedes aegypti* genome (AaegL3/AaegL5.0), the May 2014 version of *Drosophila melanogaster* genome (BDGP6), or the April 2014 version of *Anopheles gambiae* genome (AgamP4) using Bowtie 2 version 2.3.2 (*67*). The genomic assemblies and annotation files (gff3) were downloaded from Ensembl and

VectorBase databases (*68-70*). An analysis pipeline to create a list of peaks was carried out using a combination of SAMtools, BamTools, BEDTools, and Bioconductor (*71-74*). For our purposes, a peak is defined as any defined region that contains one or more mapped read(s). The output of this analysis pipeline was an ordered list of genomic locations starting with the locations containing the highest coverage per base. Data was visualized using Integrative Genomics Viewer version 2.3.92 and 2.3.3 (*75, 76*). Peak calling and cleavage site prediction was performed manually to search for the distinct drop-off cliff-shaped peaks observed at 2', 3'-cyclic phosphate-to-adapter ligation junctions. The genomic coordinate of the cleavage site was mapped based on the location of the sharp drop-offs. Sequences 200 nt upstream and downstream of a predicted ribozyme cleavage site were extracted for further analysis. To search for the hammerhead ribozyme, HDV ribozyme, and tRNA conserved secondary structure motifs within the dataset, RNArobo toolkit was used (*92*). To cross-reference the information contained in the annotation files, the sequences of the top peaks were fed into a BLAST (Basic Local Alignment Search Tool) (*77*).

3.3. Results

3.3.1. Analysis of RNA sequencing results

Sequencing libraries were created using the total RNA extracted from three species of Diptera: *A. gambiae*, *D. melanogaster*, and *A. aegypti*. The full chromosomally assembled and well-annotated genomes of *A. gambiae and D. melanogaster* resulted in relatively simple paired-end alignment of the reads (Table 2.3 in Section 2.4). The data from these RNA sequencing experiments were analyzed using secondary structure prediction algorithms.

3.3.2. Detection of self-cleaving ribozymes by genomic location and secondary structure motifs

Self-cleaving ribozymes have low sequence identity, but are highly conserved in their secondary structures. In order to perform a secondary structure search within our datasets, sequences 200 nt upstream and downstream of a predicted ribozyme cleavage site were extracted and fed into RNArobo with various descriptor files for HDV ribozyme, the hammerhead ribozyme, and tRNA (Table 2.5, Table 2.6, and Table 2.7). Although not a self-cleaving ribozyme, during tRNA cleavage and maturation, a 2', 3'-cyclic phosphate is formed (*93*). In *A. gambiae*, one full-

length HDV ribozyme motif was identified on the X chromosome; no other annotation information is available (Figure 4A-B). The iconic nested double pseudoknot fold of the HDV ribozyme can be seen in Figure 4B. A type III hammerhead ribozyme-like structure was pinpointed on chromosome 2R in the 5'UTR of the RpL32 gene. This genomic location is particularly interesting because we would expect self-cleaving ribozymes to reside within noncoding regions, such as UTRs and introns. In addition to probing for new examples of HDV-like ribozymes and the hammerhead ribozyme, a search was conducted for previously annotated ribozymes. Two families of HDV-like ribozymes, drz-Agam-1 and drz-Agam-2, were first reported by the Lupták group (*83*). None of the 5 genomic copies were identified by our assay.

Table 3.1 Numerous self-cleaving ribozyme and tRNA structural motifs were identified in *A. gambiae*

RNArobo descriptor	Number of unique matches
Full-length HDV ribozyme	1
HDV ribozyme core	311
HDV ribozyme P1P3P1.1	103
Hammerhead Ribozyme Type I	0
Hammerhead Ribozyme Type II	0
Hammerhead Ribozyme Type III	1
tRNA	30
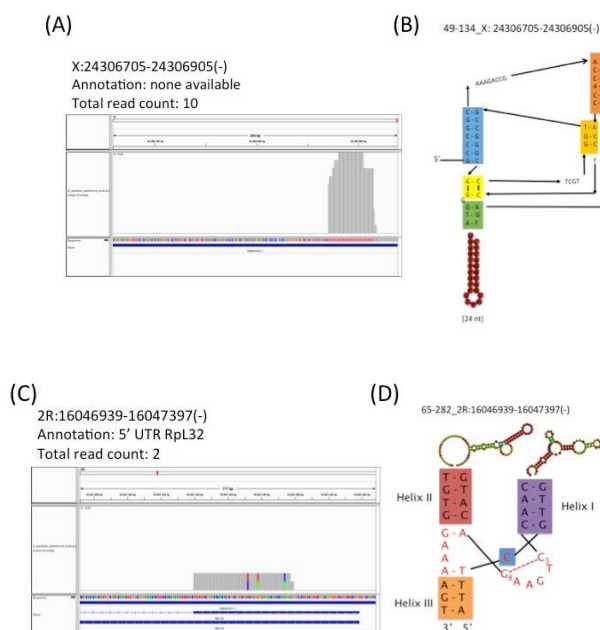
Figure 3.1 RNArobo predicted self-cleaving ribozyme structural motifs in *A. gambiae*. The structures of the non-conserved loops were predicted by Mfold and visualized using VARNA (*94, 95*). Chromosome, genomic coordinates, and +/- strand are listed. (A) Using the full-length HDV ribozyme descriptor, an exact match was found on chromosome X. (B) Predicted secondary structure of the HDV ribozyme-like fold. (C) Type III hammerhead ribozyme motif was detected. (D) Predicted secondary structure of the potential hammerhead ribozyme found by RNArobo.

In *D. melanogaster*, three hammerhead ribozyme motifs were detected, in addition to several abbreviated HDV ribozyme and tRNA structures (Table 2.6). Hammerhead ribozyme-like structures were predicted; a type II on chromosome 3L in the 3'UTR of the FRY-rb gene (Figure 5A-B) and a type III on chromosome X on the 3'UTR of CG42343-RG gene (Figure 5C-D). Again, it's important to note that we are observing these in noncoding regions. Another type III was encoded in the rDNA, within a 28s rRNA pseudogene known as the R2 retrotransposon (Figure 5E-F). In several *Drosophila* species, the 5'UTR of R2 undergoes rapid self-scission to separate the 28s-R2 co-transcript (*91*). The R2 ribozyme is predicted to fold into a double pseudoknot structure, similar to the HDV ribozyme (*81, 91*). It is intriguing that a type III Hammerhead ribozyme fold was identified at this R2 element rather than an HDV ribozyme fold. The confirmation of a bona fide self-cleaving ribozyme at this location would dramatically restructure the family of R2 ribozymes. Exploring near other R2 elements, a handful of were predicted to fold similarly to the core of the HDV ribozyme (Figure 3.2).

Table 3.2 Numerous self-cleaving ribozyme and tRNA structural motifs were identified in
D. melanogaster

| RNArobo descriptor	Number of unique matches
Full-length HDV ribozyme	0
HDV ribozyme core	1007
HDV ribozyme P1P3P1.1	295
Hammerhead Ribozyme Type I	0
Hammerhead Ribozyme Type II	1
Hammerhead Ribozyme Type III	2
tRNA	124
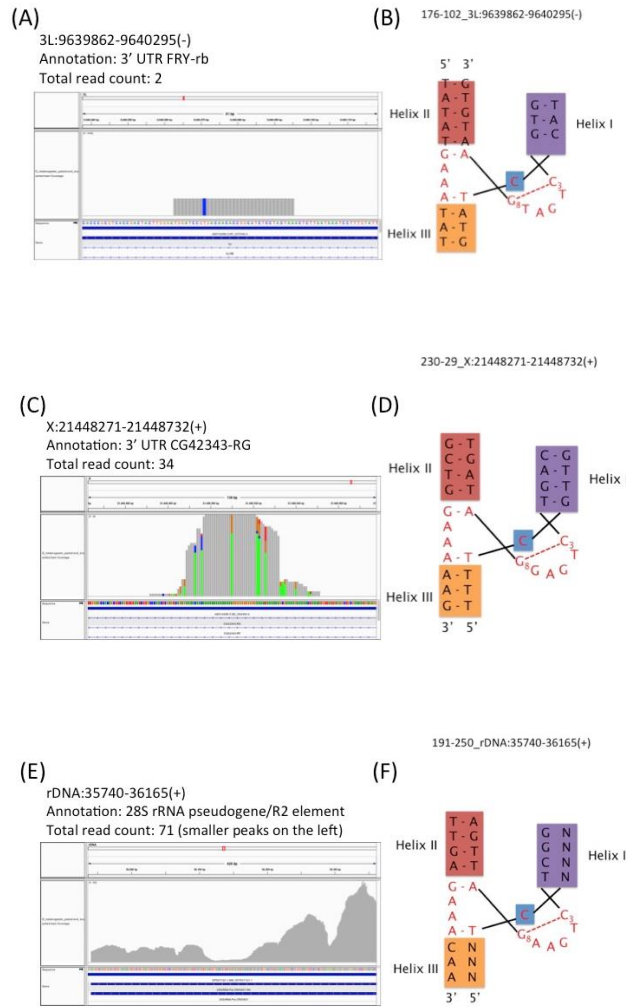
Figure 3.2 RNArobo predicted hammerhead ribozyme structural motifs in *D. melanogaster.* The structures of the non-conserved loops were predicted by Mfold and visualized using VARNA (*94, 95*). A type II (A-B) and type III (C-F) hammerhead ribozyme motifs were detected. Chromosome, genomic coordinates, and +/- strand are listed.
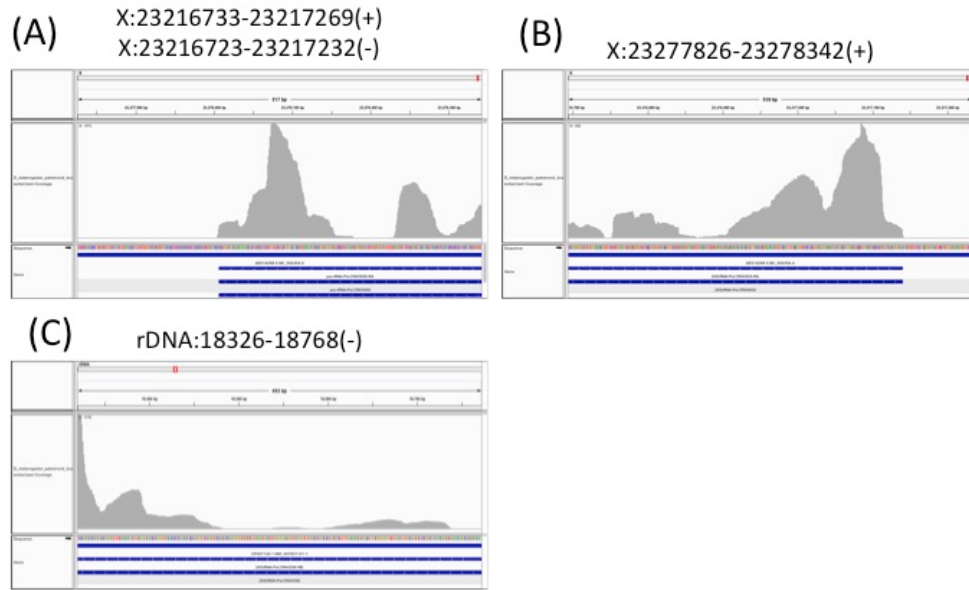
Figure 3.3 RNArobo predicted HDV-like ribozyme motifs near R2 elements in *D. melanogaster.* (A-C) HDV ribozyme core structures were predicted near several R2 elements. Chromosome, genomic coordinates, and +/- strand are listed.

In *A. aegypti*, only abbreviated HDV ribozyme motifs and tRNAs were identified by the structure search (Table 3.1). In the case of this particular mosquito, 1,798 hammerhead ribozyme sequences have been documented (*28*). Sequencing reads from this study were aligned to the genomic coordinates of six hammerhead ribozymes (Table 3.3). The likely explanation for the low detection rate is the prevalence of inactive hammerhead-like RNAs. The vast majority of the documented hammerhead ribozyme sequences were discovered through comparative genomics, with a large sum remaining untested for self-cleavage activity. In contrast, we performed a unique RNAseq experiment where only active ribozymes can be detected. This increases the likelihood of identifying novel bona fide self-cleaving ribozymes by decreasing the number of inactive hits.

Table 3.3 Core HDV ribozyme and tRNA structural motifs were identified in *A. aegypti.* The hammerhead ribozyme is abbreviated here as HH.

RNArobo descriptor	Aae01 (Number of unique matches)	Aae02 (Number of unique matches)
Full-length HDV ribozyme	0	0
HDV ribozyme core	167	486
HDV ribozyme P1P3P1.1	38	119
HH Ribozyme Type I	0	0
HH Ribozyme Type II	0	0
HH Ribozyme Type III	0	0
tRNA	58	135

Table 3.4 Previously described hammerhead ribozymes were found by genomic location in *A. aegypti.* These were not detected by RNArobo. Genomic locations were taken from ref (*28*).

Supercontig	Start of genomic coordinate	End of genomic coordinate
supercont1.322	220141	220196
supercont1.1441	77526	77581
supercont1.25	845230	845285
supercont1.105	458123	458177
supercont1.19	54455	54509
supercont1.1743	18403	18457

3.3.3. Analysis of the top peaks

Peak calling is typically performed computationally to identify genomic areas that are enriched with a higher number of aligned reads compared to the rest of the genome. Because we performed a selective RNAseq experiment and did not observe full genomic coverage after read alignment, manual peak calling was performed. As described earlier, an ordered list of genomic locations containing the highest coverage per base was generated. For *A. gambiae* and *D. melanogaster*, the

annotated locations of the top 100 peaks largely skew towards unassembled regions of the genomic assembly, i.e. UNKN and contigs/unmapped scaffolds (Figure 3.5). As mentioned previously, difficulties in complete genome assembly arise from highly repetitive sequences, which coincidentally, are where self-cleaving ribozymes are likely to reside (*28, 79, 81*). This innate quality presents challenges in extracting full-length sequences to test for *in vitro* self-cleavage activity. Interestingly, in both datasets, chromosome X contained the highest number of hits mapped to a single chromosome. Given that many genes on this chromosome are essential for normal development, our data further supports the notion that self-cleaving ribozymes are differentially expressed at early stages of larval development (*28, 81, 83*). Analysis could not be performed for *A. aegypti*, due to the lack of a chromosome-level genomic assembly with accompanying annotations.

(A)
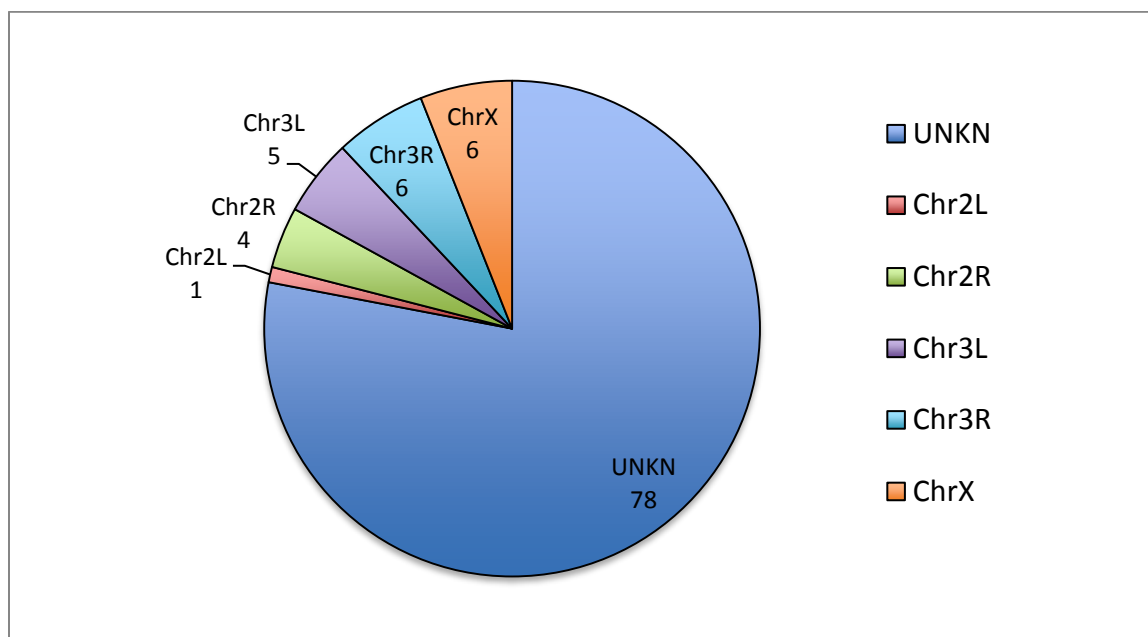
(B)
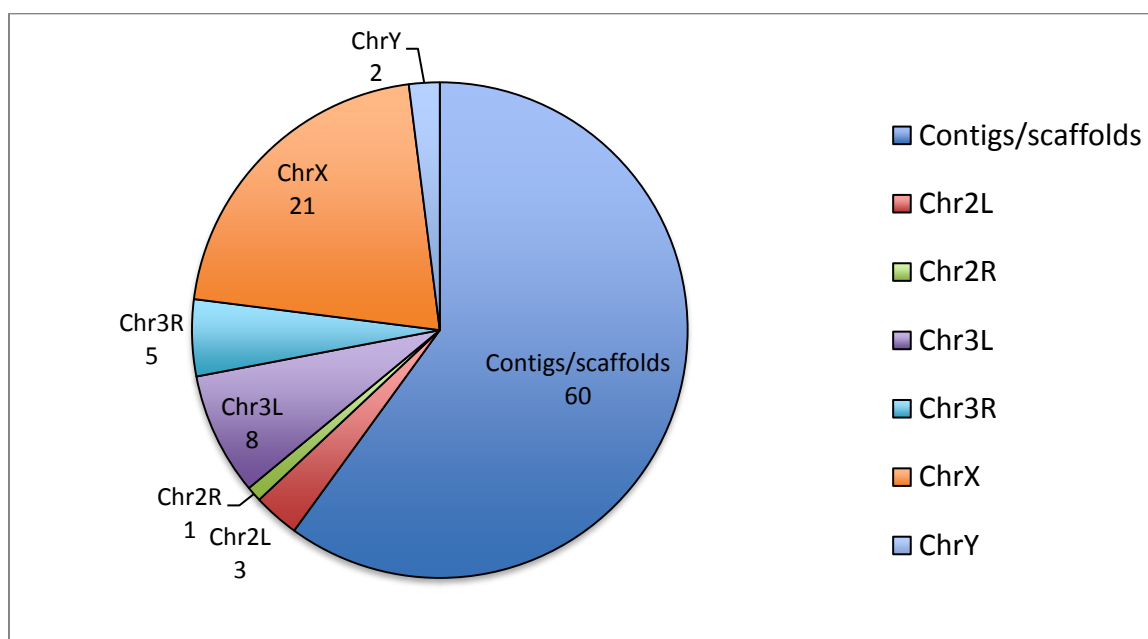
Figure 3.4 The majority of the top 100 peaks in *A. gambiae* and *D. melanogaster* map to contigs, unmapped scaffolds, and unknown regions of the genome. (A) *A. gambiae* (B) *D. melanogaster*

3.4. Discussion

Our data provide strong evidence that our method is specifically selecting 2', 3'-cyclic phosphate capped RNAs, as shown by the comparisons of the stably capped U6 snRNA relative to other spliceosomal RNAs that are present in the cell at equimolar levels (Section 2.4). In addition, we were able to detect secondary structural motifs of hammerhead and HDV-like ribozymes. Most of which will be novel, if shown to have self-cleaving activity.

Few annotated ribozymes were confirmed by this method, which could be caused by low levels of expression or hydrolysis of the 2', 3'-cyclic phosphate prior to capture. Another possibility is the lack of sensitivity or poor genomic mapping accuracy of our method. It would be worthwhile to repeat these experiments with much longer sequencing reads to increase the accuracy of genomic mapping. The caveat to this is that longer reads tend to be more error-prone. Additionally, switching to a model organism that has a high-quality genome assembly and annotation files would help avoid the issues we observed with large numbers of sequencing reads mapping to unmapped scaffolds, rather than chromosomes. On the computational side, we need to work on developing an algorithm for structure prediction and comparative structure similarity.

Although devising a method for the capture and sequencing of 2', 3'-cyclic phosphate terminal RNAs is not an original concept, our method has demonstrated to be the best option. Schutz *et al.* laid the groundwork by proposing cyclic phosphate-capture using tRNA ligase, but saw very few reads mapping to the U6 snRNA. Recently, Honda *et al.* presented a method that utilizes periodate oxidation to negatively select against cyclic phosphate-terminal RNAs. The disadvantage to this method is the lack of specificity. Any RNA with a 3' modification protects the RNA from oxidation, leaving them to be sequenced and subsequently increasing background in the data.

4. CREATING A LIBRARY OF tRNAPyl-BINDING APTAMERS THROUGH SELEX

4.1. Declaration of collaborative work

The work described in this chapter is the result of a collaborative effort in the Golden lab. Samantha Lee performed the work described in this chapter and is the author of the chapter and its figures. However, Golden lab members working on related projects lent their RNA expression constructs and expertise on T-box riboswitches. tRNA^{Gly}, stem I of the *gly*QS T-box riboswitch from *G. kaustophilus*, and the first-generation of STARzyme were cloned by Dr. Ji Chen. tRNA^{Tyr} and stem I of *tyr*S T-box riboswitch from *B. subtillis* were cloned by Rui Gan. I would also like to acknowledge Dr. Andrej Lupták for his insight and guidance with developing and troubleshooting the selection protocol.

4.2. Introduction

The ability to site-specifically incorporate unnatural amino acids into proteins is a powerful tool that has enabled researchers to generate proteins with novel functionalities for use in diverse applications. These applications range from labeling proteins with small organic fluorophores for *in vivo* functional studies, to cross-linking bispecific antibodies for cancer therapies, and generating replication-incompetent virus vaccines (*96-98*).

Introducing an unnatural amino acid into a protein requires an orthogonal animoacyl-tRNA synthetase (aaRS)/tRNA pair: where the tRNA recognizes a unique codon, typically, the amber stop codon. The engineered aminoacyl-tRNA synthetase must recognize a specific tRNA and unnatural amino acid, against all other naturally occurring substrates in the cell or translation system. This defines the orthogonal pair. One method to create orthogonal pairs has been to mutate and engineer aaRSs to alter their substrate specificities (*99*). Although, because protein aaRSs have high specificity for their amino acid substrates, this requires that the unnatural amino acid be similar to the native amino acid substrate while also being distinct enough to be differentiated. A family of *in vitro* selected aminoacyl tRNA synthetase ribozymes, flexizymes (Fx), have overcome the substrate limitations of engineered protein aaRSs (*100*). Although flexizyme is able to utilize

a diverse range of amino acid substrates, it lacks specificity for its tRNA substrate and will indiscriminately charge any tRNA with a CCA-tail (*100*).

To introduce tRNA specificity, our lab has created fusion ribozymes where we've linked a tRNA-binding RNA element, a domain of the T-box riboswitch, to a flexizyme. These <u>s</u>pecific <u>t</u>RNA <u>a</u>minoacylating <u>r</u>ibozymes, STARzymes, have been shown to specifically charge their tRNA with an unnatural amino acid. The first STARzyme prototype fused the *G. kaustophilus gly*QS T-box stem I and stem II to our version of a circularly permutated flexizyme (*101*). The latest prototype uses the *B. subtilis tyr*S T-box stem I (*102*). Data from our lab shows that this STARzyme, created with the *B. subtilis tyr*S T-box, is able to incorporate an unnatural amino acid into a protein when applied in an *E. coli in vitro* translation system. However, we also observed misincorporation of tyrosine, suggesting that this STARzyme/tRNA pair is not orthogonal.

As it happens, an Archaean species, *Methanosarcina barkeri* utilizes a tRNA that uses the same anticodon as the amber suppressor tRNA, tRNA[Pyrrolysine] (tRNA[Pyl]) (*103, 104*). This eliminates the need to mutate the tRNA anticodon as was done previously. Even better, this tRNA has been shown to be orthogonal in eukaryotes and the MS strain of *M. barkeri* expresses a tRNA[Pyl] that has been shown to be orthogonal in both eukaryotes and prokaryotes (*105-107*). tRNA[Pyl] is only expressed in this one species of Archaea and T-box riboswitches are exclusive to gram-positive bacteria, meaning *M. barkeri* tRNA[Pyl] does not have a naturally occurring T-box leader to pair with. The goal of this project is to *in vitro* select T-box leader-like aptamers for tRNA[Pyl]. I will use the *G. kaustophilus gly*QS T-box leader and *B. subtilis tyr*S T-box leader sequences as templates to preserve the overall T-box riboswitch stem I structure, while mutating and randomizing the regions of the T-box stem that make contact with the tRNA. The work described here details the development of a unique *in vitro* selection where both the oligonucleotide library and target molecule are comprised of RNA. The long-term goal of this project is to create a collection of candidate tRNA[Pyl] aptamers to fuse with flexizymes to create new STARzymes.

4.3. Methods

4.3.1. Cloning RNA expression constructs

tRNAs and the stem I of glyQS and tyrS T-box riboswitches were cloned into the pUC19 plasmid. DNA inserts were amplified by extension PCR using the oligonucleotides in Table 2.2 and Phusion High-Fidelity PCR Kit (New England BioLabs). A single cycle of denaturation at 98°C for 30 seconds, annealing at 65°C for 30 seconds, and extension at 72°C for 5 minutes was performed with 8 μM of oligonucleotides 1 and 2 (Table 2.1), and 0.4 mM dNTPs. PCR products were visualized on an agarose gel and the correct fragment size was excised for purification using QIAquick Gel Extraction Kit (QIAGEN). pUC19 and the DNA inserts were double-digested with HindIII and XbaI in Buffer 2.1 (New England BioLabs) at 37°C overnight. Reaction components were removed by phenol:chloroform extraction. Inserts were ligated into pUC19 using T4 DNA ligase then transformed into 5-alpha Competent *E. coli* (New England BioLabs). Transformation products were plated on Luria Broth plates containing 100 μg/mL ampicillin pre-coated with 40 μL of 100 μM isopropyl β-D-1-thiogalactopyranoside (IPTG) and 40 μL of 20 mg/mL 5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside (X-gal). White colonies were selected for Sanger sequencing at the Purdue Genomics Core Facility.

4.3.2. *In vitro* transcription and purification of cloned constructs

A single colony was inoculated into 500 mL of Luria Broth containing 100 μg/mL ampicillin and grown at 37°C overnight with shaking. Cells were harvested by centrifugation and plasmid was purified using the Plasmid Mega Kit (QIAGEN). Plasmid was linearized with BsaI-HF (New England BioLabs) at 37°C overnight. Reaction components were removed by phenol:chloroform extraction. Run-off transcription was performed with 300 μg of linear plasmid in the presence of 2 mM NTPs, 1 μM T7 RNA polymerase, 40 mM Tris-HCl pH 8.0, 25 mM $MgCl_2$, 2 mM dithiothreitol, and 1 mM spermidine. The transcription reaction was incubated in a 37°C water bath for 3 hours. After removing them from the water bath, 30 μM EDTA was added to clear the solution prior to ethanol precipitation. The precipitated products were resuspended in water and run on a 6% polyacrylamide gel containing 7 M urea. The RNA was visualized by UV shadowing and the RNA was excised for purification. The excised gel was crushed though a syringe and frozen in 10 mM Tris-HCl at pH 7.5, 1 mM EDTA, and 250 mM NaCl. The gel mixture was thawed and rocked overnight at 4°C.The gel pieces were filtered for removal, followed by ethanol

precipitation of the eluted RNA. Excess salts and impurities were we removed using 10K or 30K Amicon Ultra-0.5 Centrifugal Filter Units (MilliporeSigma).

4.3.3. 3'-terminus biotin labeling of tRNAs

tRNAs were folded at a 10 µM concentration by heating to 92°C for 3 minutes and cooling to 65°C for 6 minutes with the addition of 5 µM $MgCl_2$ after 5 minutes. The tRNAs were cooled on ice for 10 minutes before proceeding to the biotinylation reaction. Pierce RNA 3' End Biotinylation Kit (Thermo Scientific) was used with a 20 hour incubation at 16°C.

4.3.4. *In vitro* transcription and purification of Round 0 RNA pool

Second strand synthesis of Round 0 DNA templates (Table 2.2) were achieved by 2 cycles of polymerase chain reaction using Phusion High-Fidelity PCR Kit (New England BioLabs). Thermocycling conditions were as follows: initial denaturation at 98°C for 30 seconds, denaturation for 10 seconds, annealing at 65°C or 72°C for 30 seconds, extension at 72° for 30 seconds, and final extension at 72°C for 5 minutes was performed with 50 µM DNA pool, 25 µM of fwd with clone sites and rev oligonucleotides (Table 4.1), and 0.4 mM dNTPs. Before proceeding to *in vitro* transcription, the reaction components were removed using QIAquick PCR Purification Kit (QIAGEN). *In vitro* transcription was performed with PCR product template in the presence of 3 mM NTPs, 1 µM T7 RNA polymerase, 40 mM Tris-HCl pH 8.0, 25 mM $MgCl_2$, 2 mM dithiothreitol, and 1 mM spermidine. The transcription reaction was incubated in a 37°C water bath for 3 hours. After removing from the water bath, 30 µM EDTA was added to clear the solution prior to ethanol precipitation. The precipitated products were resuspended in water and run on a 6% polyacrylamide gel containing 7 M urea. The RNA was visualized by UV shadowing and the RNA was excised for purification. The excised gel was crushed though a syringe and frozen in 10 mM Tris-HCl at pH 7.5, 1 mM EDTA, and 250 mM NaCl. The gel mixture was thawed and rocked overnight at 4°C.The gel pieces were filtered for removal, followed by ethanol precipitation of the eluted RNA. Excess salts and impurities were we removed using 30K Amicon Ultra-0.5 Centrifugal Filter Units (MilliporeSigma).

4.3.5. RNA folding conditions and optimization

Refolding of *in vitro* transcribed RNAs was achieved by denaturing at 92°C for 3 minutes, cooling to 65°C or 70°C for 5 minutes, the addition of 1 or 5 mM $MgCl_2$, an additional minute of

incubation at 65°C or 70°C, and finally bench top or snap cooling on ice for 10 minutes. *G. kaustophilus* T-box stem I and its cognate tRNA^{Gly}, *B. subtilis* T-box stem I and its cognate tRNA^{Tyr}, and STARzyme were folded at a 10 μM concentration at 65°C with 5 mM MgCl$_2$ and snap cooling on ice. Fusaro and MS tRNAPyl were folded at a 5 μM concentration at 65°C with 5 mM MgCl$_2$ and snap cooling on ice. Library 1 and 2 were folded at a 1 μM concentration at 70°C with 1 mM MgCl$_2$ and bench top cooling at room temperature.

4.3.6. Size-exclusion chromatography

Superdex 200 Increase 3.2/300 column (GE Healthcare Life Sciences) was used for all size-exclusion chromatography experiments on the ÄKTA pure chromatography system.

Samples were applied and eluted in 100 μL fractions with 10 mM HEPES pH 7.0, 150 mM KCl, and 1 or 5 mM MgCl$_2$ at 75 μL/min flow rate for 1.5 column volumes. Chromatograms were visualized using UNICORN 7.

4.3.7. Immobilize biotinylated tRNAs to streptavidin-conjugated beads

2 mg of Dynabeads MyOne Streptavidin C1 (Invitrogen) beads were washed according to the manufacturer's RNA manipulation protocol. 100 pmol of tRNAs were immobilized by incubation with gentle rotation for 30 minutes at room temperature. Biotinylated tRNAs were separated with a DynaMag-2 (Invitrogen) for 2 minutes and washed 3 times with 5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, and 1 M NaCl. Beads with immobilized tRNA were resuspended in 10 mM HEPES pH 7, 150 mM KCl, and 5 mM MgCl$_2$.

4.3.8. Round 1 selection: pre-clear, negative, competitive, and positive selection

Prior to each round of selection, the RNA pools were pre-cleared through unconjugated streptavidin beads to remove any species that have high affinity to the magnetic bead or streptavidin. Prior to round 1 of selection, a negative selection and a competitive selection was performed against the pool's cognate tRNA and total *E. coli* tRNAs (Roche). *G. kaustophilus* tRNA^{Gly} was used for Library 1 and B. subtilis tRNA^{Tyr} was used for Library 2 negative selections. These tRNAs were biotinylated and immobilized to streptavidin beads as previously described, then incubated with 1 nmol of RNA pool for 30 minutes at room temperature with gentle rotation. Unbound species were collected and used for competitive selection. 1 nmol of total *E. coli* tRNAs were incubated with the negatively selected RNA pool for 30 minutes before applying the mixture

to immobilized tRNA^{Pyl} and incubating for another 30 minutes. Selected RNA species were separated on a magnet for 2 minutes and washed 3 times with 10 mM HEPES pH 7, 500 mM KCl, and 1 mM MgCl$_2$. RNA was eluted by heating to 90°C for 3 minutes in 95% formamide and 10 mM EDTA, then separating the beads using a magnet. To remove tRNAs and formamide, the elution was concentrated using a 30K Amicon Ultra-0.5 Centrifugal Filter Unit (MilliporeSigma) and buffer exchanged with water. The purified RNAs then reverse transcribed for the next round of selection (see Section 4.3.9).

4.3.9. Round 2-9 selection

Positive selection rounds were carried out as described in Section 2.3.7.3 with a few exceptions. To increase competition between species, the ratio of tRNA:pool, concentration of KCl in the wash buffer, and number of washes were gradually increased with successive rounds.

Round 1 – 1:10 tRNA:pool, 500 mM KCl for 3 washes

Round 2 – 1:100 tRNA:pool, 500 mM KCl for 3 washes

Round 3 – 1:1000 tRNA:pool, 500 mM KCl for 3 washes

Round 4 – 1:1000 tRNA:pool, 750 mM KCl for 3 washes

Round 5 – 1:1000 tRNA:pool, 750 mM KCl for 5 washes

Round 6 – 1:1000 tRNA:pool, 750 mM KCl for 5 washes

Round 7 – 1:5000 tRNA:pool, 750 mM KCl for 5 washes

Round 8 – 1:5000 tRNA:pool, 1 M KCl for 5 washes

Round 9 – 1:5000 tRNA:pool, 1 M KCl for 5 washes

4.3.10. *In vitro* selection by size-exclusion chromatography

Non-labeled RNAs were folded as described in Section 4.3.5. 100 μL of 5 nM tRNA and 50 nM T-box/library (1:10 ratio) were incubated with 10 mM HEPES pH 7.0, 150 mM KCl, and 1 or 5 mM MgCl$_2$ for an hour at room temperature. 50 μL was loaded onto a Superdex 200 Increase 3.2/300 column (GE Healthcare Life Sciences). Size-exclusion chromatography was carried out as described in Section 4.3.6.

4.3.11. Reverse transcription and amplification

2-50 ng of the *in vitro* selected RNAs was mixed with 1 μM RT primer was heated at 95°C for 2 minutes then cooled at room temperature for 10 minutes. TIGRT buffer (20 mM Tris-HCl pH 7.5, 200 μM NaCl, 10 mM $MgCl_2$, and 5 mM DTT) and 1 μL TGIRT-III enzyme (InGex) was gently mixed and incubated at room temperature for 30 minutes. At which point, 1.25 μM dNTPs was added and the reaction mixture was incubated at 60°C for 1 hour (*62-64*). Amplification of cDNA was achieved using Phusion High-Fidelity PCR Kit (New England BioLabs). Thermocycling conditions were as follows: initial denaturation at 98°C for 30 seconds, denaturation for 10 seconds, annealing at 65°C or 72°C for 30 seconds, extension at 72° for 30 seconds, and final extension at 72°C for 5 minutes was performed with 4 μL cDNA, 0.5 μM of fwd with clone sites and rev oligonucleotides (Table 2.1), and 0.2 mM dNTPs. Aliquots were taken every 4 cycles, for 16 cycles and run on a 1.5% agarose gel. The minimum number of cycles required to visualize a product was chosen for further amplification of the remaining cDNA.

4.3.12. Oligonucleotides

All oligonucleotides were synthesized by Integrated DNA Technologies.

Table 4.1 Oligonucleotides used for cloning, reverse transcription, and PCR

	Sequence 5'-3'
Fusaro tRNA^{Pyl} 1	CATTAGTCTAGATAATACGACTCACTATAGGAAACCTGATCATGTAGATCGAATGGACTCTAAATCCGTTCAGCCGGGTT
Fusaro tRNA^{Pyl} 2	ATGCTAAAGCTTGGTCTCATGGCGGAAACCCCGGGAATCTAACCCGGCTGAACGGATTTA
MS tRNA^{Pyl} 1	CATTAGTCTAGATAATACGACTCACTATAGGGAACCTGATCATGTAGATCGAATGGACTCTAAATCCGTTTAGCCGGGTT
MS tRNA^{Pyl} 2	ATGCTAAAGCTTGGTCTCATGGCGGAAACCCCGGGAATCTAACCCGGCTAAACGGATTTA
Library 1 fwd with clone sites	CATTAGTCTAGATAATACGACTCACTATAGGTCGCGATGACGGATC
Library 1 rev with clone sites	ATGCTAAAGCTTGGTCTCNGTCGCTCCGGACTCTAGT
Library 1 rev	GTCGCTCCGGACTCTAGT
Library 1 RT	GTCGCTCCGGACTCTAGTTTCATTAACCG
Library 2 fwd with clone sites	CATTAGTCTAGATAATACGACTCACTATAGGTAAAGATTGAGACAAGTAGAATATCCTTA
Library 2 rev with clone sites	ATGCTAAAGCTTGGTCTCACGGTTCATGAGTCTATTCGATATATCC
Library 2 rev	CGGTTCATGAGTCTATTCGATATATCC
Library 2 RT	CGGTTCATGAGTCTATTCGATATATCCGTCTGTGC

Table 4.2 Round 0 DNA template sequences and hand-mixed ratios for 4% magnesia . Hand-mixed ratios are listed where N1 = 1% of A, 1% of C, 1% of G, 97% of T, N2 = 1% of A, 97% of C, 1% of G, 1% of T, N3 = 1% of A, 1% of C, 97% of G, 1% of T, N4 = 97% of A, 1% of C, 1% of G, 1% of T, and N5 = 25% of A, 25% of C, 25% of G, 25% of T.

	Sequence 5'-3'
Library 1, *glyQS*-based	TAATACGACTCACTATAGG(N1:01010197)(N2:01970101)(N3:01019701)(N2)(N3)(N4:97010101)(N1)(N3)(N4)(N2)(N3)(N3)(N4)(N1)(N2)(N4)(N4)(N1)(N4)(N3)(N1)(N4)(N3)(N1)(N1)(N4)(N4)(N2)(N2)(N2)(N1)(N2)(N1)(N2)(N1)(N1)(N2)(N5:25252525)(N5)(N5)(N5)(N5)(N5)(N5)(N3)(N2)(N2)(N3)(N3)(N3)(N5)(N5)(N5)(N5)(N5)(N5)(N5)(N5)(N5)(N5)(N2)(N2)(N2)(N3)(N3)(N1)(N3)(N4)(N4)(N3)(N4)(N2)(N3)(N3)(N1)(N1)(N4)(N4)(N1)(N3)(N4)(N4)(N4)(N2)TAG(N4)(N3)(N1)(N2)(N2)(N3)(N3)(N4)(N3)(N2)(N3)(N4)C
Library 2, *tyrS*-based	TAATACGACTCACTATA(N3:01019701)(N3)(N1:01010197)(N4:97010101)(N4)(N4)(N3)(N4)(N1)(N1)(N3)(N4)(N3)(N4)(N2:01970101)(N4)(N4)(N3)(N1)(N4)(N3)(N4)(N4)(N1)(N4)(N1)(N2)(N2)(N1)(N1)(N4)(N2)(N3)(N1)(N5:25252525)(N5)(N5)(N5)(N5)(N5)(N5)(N5)(N3)(N2)(N1)(N3)(N4)(N1)(N5)(N5)(N5)(N5)(N5)(N5)(N5)(N5)(N5)(N5)(N4)(N1)(N2)(N4)(N3)(N2)(N4)(N2)(N4)(N3)(N4)(N2)(N3)(N3)(N4)(N1)(N4)(N1)(N4)(N1)(N2)(N3)(N4)(N4)TAG(N4)(N2)(N1)(N2)(N4)(N1)(N3)(N4)(N4)(N2)(N2)G

4.4. Results

4.4.1. Optimal RNA labeling, amplification, and transcription conditions

The first method for selection was capturing the target on paramagnetic beads, where the target of interest is tRNA^{Pyl}. The free 3'-CCA tail was chosen for labeling with biotin to avoid steric disturbance of labeling the 5'-end. The Pierce 3'-end labeling kit utilizes the ligation of a biotinylated cytidine. Ligation efficiency of the folded tRNA averaged 50-55%, where the control RNA consistently displayed complete ligation of the biotinylated cytidine (Figure 4.1). This is likely due to the short length and lack of structure of the control RNA compared to tRNAs. The high molecular weight tRNA band was excised and purified from the gel for use in the selection.
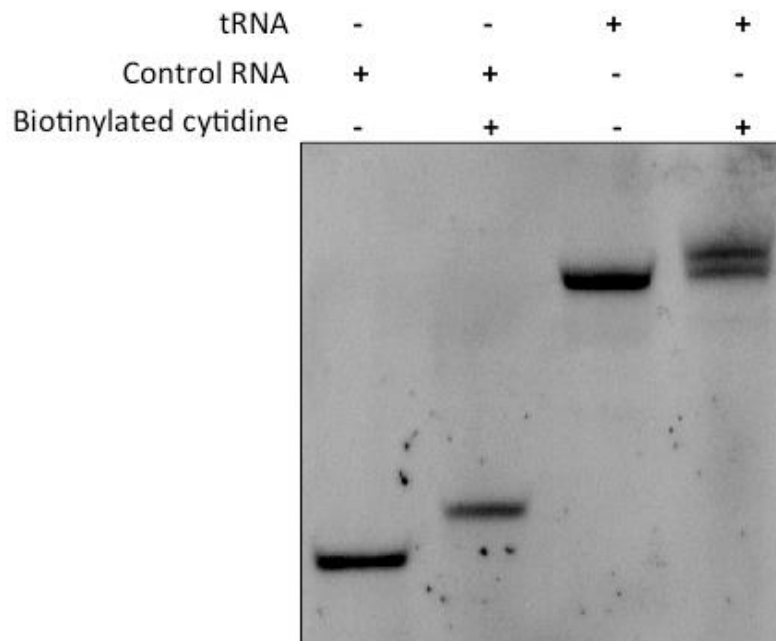
Figure 4.1 Efficiency of the Pierce RNA 3' End Biotinylation Kit. Control RNA was supplied with the system and ligation reactions were run on a 6% polyacrylamide gel containing 7 M urea.

In order to obtain sufficient cDNA template for in vitro transcription, PCR amplification was necessary. It is important to minimize the cycles of PCR as to not decrease the complexity of the pool or skew the population towards particular sequences. To ensure this, aliquots were taken after every 4 cycles of PCR for 16 cycles to determine the minimum number of cycles necessary to

obtain more templates. Excessive cycles of PCR also produces cross amplification, as can be seen by the larger molecular weight smears in the 12-cycle and 16-cycle lanes of the gel (Figure 4.2). It was determined that 4 cycles of PCR was sufficient for the *gly*QS-based library and 8 cycles was optimal for the *tyr*S library.
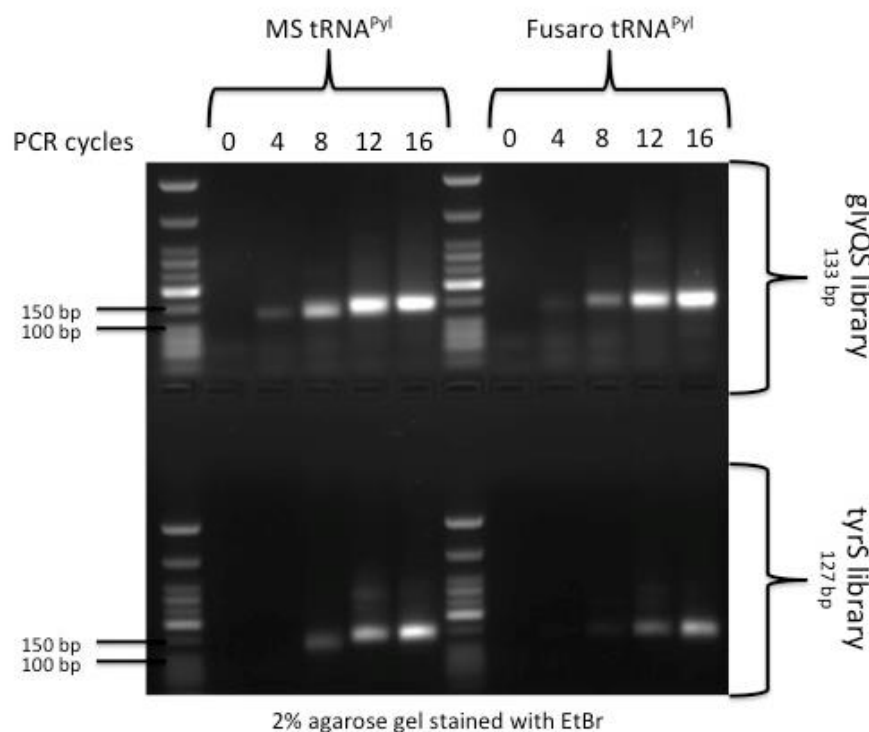
Figure 4.2 PCR amplification of selected cDNA libraries. 10 μL aliquots were taken every 4 cycles and electroporated on a 2% agarose gel stained with ethidium bromide to determine the optimal amplification conditions.

Following minimal PCR amplification of the DNA or cDNA template, RNA pools were *in vitro* transcribed. Figure 4.3 shows the relative molecular weights of the *gly*QS and *tyr*S-based library pools, 102 nucleotides and 98 nucleotides in length. Compared to Round 0 (R0) library transcription, the Round 8 (R8) transcription has much more background. This might be the byproduct of remaining PCR amplification reagents. Although the PCR products are purified using a PCR Purification Kit (QIAGEN), larger primer dimers and template may not have been washed off from the column prior to elution.
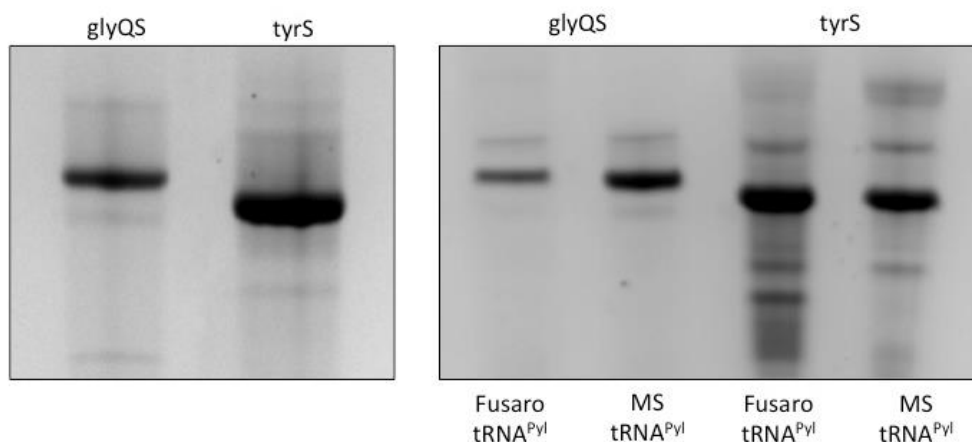
Figure 4.3 *In vitro* transcription of Round 0 RNA library and Round 8 RNA libraries. Aliquots of *in vitro* transcription reactions were electroporated on a 6% polyacrylamide gel containing 7 M urea and stained with Stains-All.

4.4.2. Qualitatively assessing the binding of tRNAs to stem I of T-box leader-like sequences using gel shift assay

Following 8 rounds of selection and exponential enrichment, a gel shift assay was performed to assess binding of the tRNA to the RNA pools. Initially, we observed disappearance of the library band with an increase in tRNA addition (Figure 4.4). This observation paralleled previous results using the native stem I of the glyQS T-box riboswitch, where the tRNA^{Gly} band disappears with the increase of T-box (Figure 4.5). Upon further observation by running samples through a size-exclusion column (data not shown), it was apparent that the multiple bands we observe in the native gel were actually oligomers and not suboptimal folding conformations as we had inferred before.
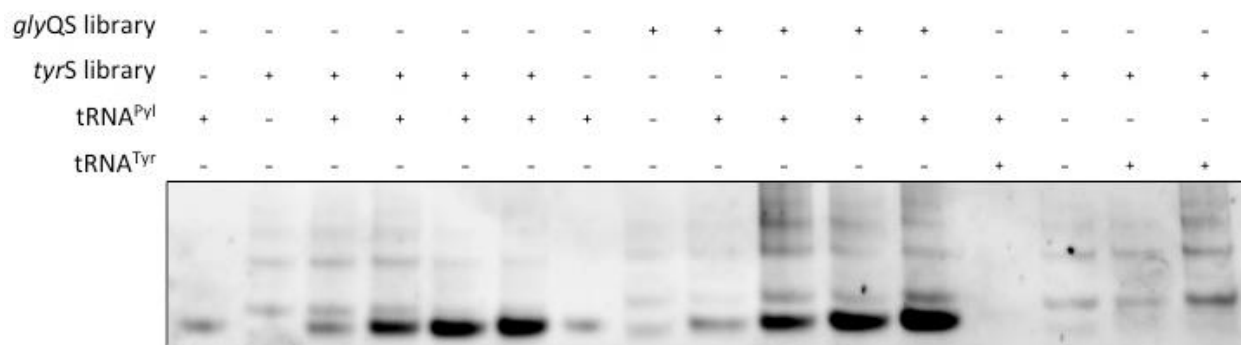
Figure 4.4 Round 4 RNA libraries appear to bind their target tRNA but not tRNA[Tyr]. Samples were run on a 6% polyacrylamide gel containing stained with SYBR Green II.
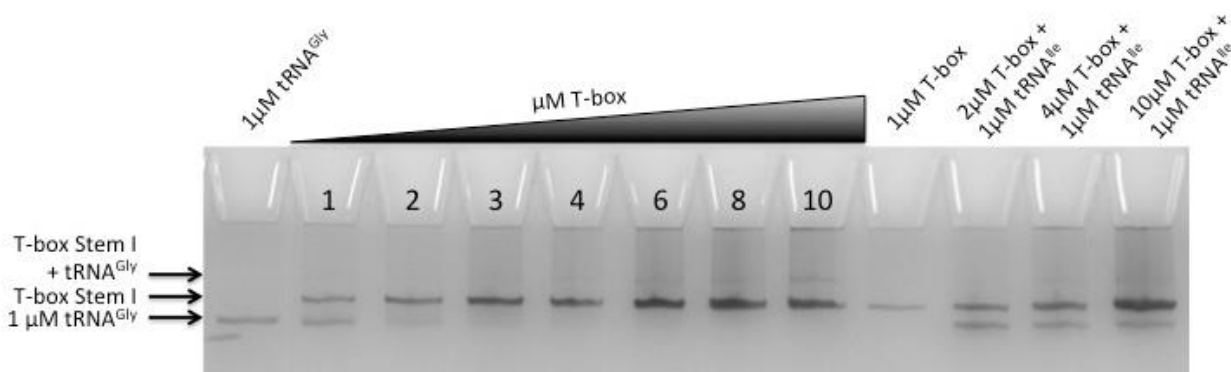
Figure 4.5 Wild-type *gly*QS T-box stem I binds tRNA[Gly]. Samples were run on a 6% polyacrylamide gel containing stained with Stains-All.

Following the optimization of RNA folding to eliminate oligomers, 10 rounds of selection were performed by size-exclusion chromatography. It was determined that monomeric folding of the RNA may have altered the ability for tRNA to bind by folding in a conformation that does not at all resemble a T-box (Figure 4.6). It is also possible that selection by size-exclusion chromatography peak shift was not the optimal method for selection, given that it appears the higher molecular weight complex shifts right, where we would expect smaller molecular weight complexes (Figure 4.7), making fraction selection inaccurate.
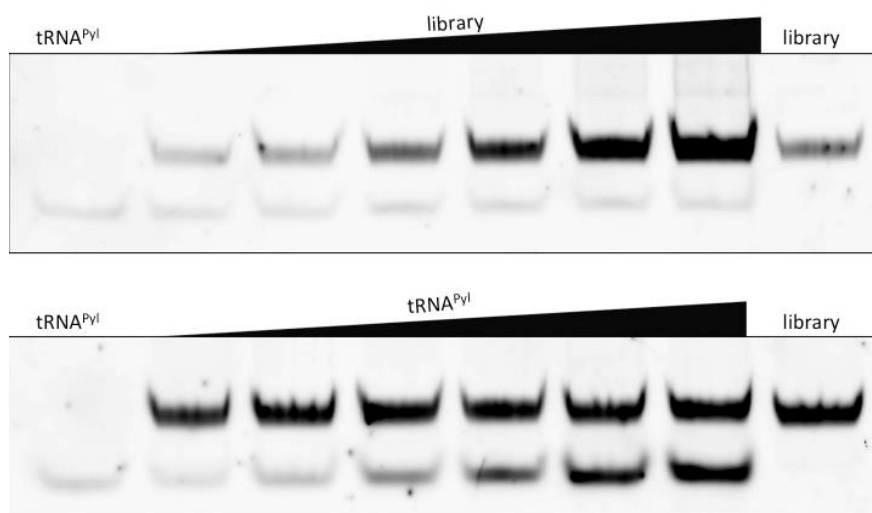
Figure 4.6 Binding of selectively amplified libraries to tRNA^Pyl is not detectable by gel shift assay. Samples were run on a 6% polyacrylamide gel containing stained with SYBR Green II
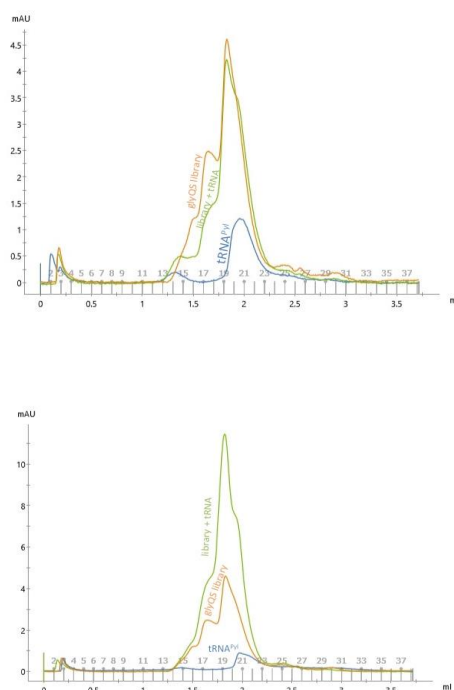
Figure 4.7 *gly*QS library bound to tRNA^Pyl co-migrates with *gly*QS library-only and tRNA-only samples. 1:10 ratio of tRNA: library. Blue: tRNA, Orange: library, Green: tRNA + library.

4.5. Discussion

The major difficulty in this project was pioneering an RNA-binding aptamer library. More commonly, small molecules such as ATP or proteins are used as the bait for a randomized pool of nucleic acids or genomic DNA. The different properties of these molecules greatly benefit their separation and in turn reduce the noise that gets passed on to the next round.

A few areas of improvement that come to mind are devising an efficient method to monitor binding affinity after each successive round of selection, or every few rounds. This way we can track the progress of the selection and have more chances to backup if something looks awry. Another piece of data that would be beneficial in tracking the progress of this experiment would be to sequence the pool after every few rounds of selection. The original method that we chose

involved performing SHAPE-seq on the final library. SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) is a method that reports nucleotide-level structural information of RNAs. When combined with RNAseq, SHAPEseq can gather data on a pool of diverse molecules in a single assay. Although this method is very powerful at reporting structure and sequence in a single reaction, sequencing at the end of our protocol put us at a disadvantage by depriving us of valuable information that could have helped us troubleshoot. This and other alternatives will be further discussed in Section 5.3.

5. THE NEXT GENERATION OF RNA

5.1. Introduction

As with all things science, there is always more work to be done. Reflecting back on the work described here, there is much that was learned and much to still learn. Every failure is a learning opportunity and with every success is accompanied with new questions. Here, I've summarized some thoughts about where I envision these projects heading.

5.2. Navigating big data to identify novel self-cleaving ribozymes

Earlier, we discussed one of the struggles in accurately mapping sequencing reads to genomic assemblies. These difficulties are directly correlated with the quality and state of the genomic assembly, along with corresponding annotation files. When we began this project, we started with *A. aegypti*, *A. gambiae*, and *D. melanogaster*. These were chosen because they've been shown to encode small self-cleaving ribozymes. Some reported to harbor upwards of 1,000 hammerhead ribozymes. We speculated that the abundance of genomic copies would enhance our chances of finding new and uniquely structured self-cleaving ribozymes. These species appeared to be the perfect organisms to validate our assay before moving into other systems. Through trial and error, we have learned that the progress of the genome assembly is key. Going on this, a comprehensive list was created of organisms that have sequenced genomes and have assemblies currently in a chromosome-level of organization (Table 5.1). The number of unplaced scaffolds is also a factor, the fewer unplaced scaffolds, the better the assembly. Finally, of course we'd prefer to work in a system that naturally carries self-cleaving ribozymes in their genomes. The list is topped off with types of ribozymes that have been annotated and if they've been experimentally validated, along with their copy number. Any of these species would be a good place to start, but my personal choice would be *Schistosoma mansoni*. These parasitic blood flukes definitely pose some danger, so hopefully a generous collaborator would be so kind to offer up some RNA.

Table 5.1 Candidate organisms for self-cleaving ribozyme hunting. The ribozymes in green have been experimentally determined to have self-cleavage activity.

Taxonomic name	Common name	Unplaced scaffolds	Annotated ribozymes	Copy no.	Ref.
Anolis carolinensis	lizard	6443	HHRz		(*28*)
Anopheles gambiae str. PEST	mosquito	8144	HDV-like/RTE	9	(*81, 83*)
Apis mellifera	honey bee	5644	HDV-like/R2		(*90, 91*)
			HHRz	8	(*28*)
Arabidopsis thaliana	thale cress	0	HHRz	11	(*27, 108*)
Bos taurus	cattle	0	HHRz	10	(*27, 28*)
			HDV-like/CPEB3	1	(*109*)
Caenorhabditis briggsae	nematode	638	HHRz	1	(*27, 28*)
Canis lupus familiaris	dog	3228	HHRz	6	(*27, 28*)
			HDV-like/CPEB3	1	(*109*)
Chlorocebus sabaeus	green monkey	1432	CoTC		(*110*)
Danio rerio	zebrafish	1035	Twister	10	(*58*)
			HHRz	9	(*27, 28*)
Drosophila melanogaster	fruit fly	1870	HDV-like/R2/Baggins		(*81, 90, 91*)
Drosophila pseudoobscura pseudoobscura	fruit fly	4790	HDV-like/R2		(*90, 91*)
			HHRz	12	(*27, 28*)
Drosophila simulans	fruit fly	10600	HDV-like/R2		(*90, 91*)
Drosophila yakuba	fruit fly	8122	HDV-like/R2		(*91*)
Equus caballus	horse	9604	HHRz	20	(*27, 28*)
			HDV-like/CPEB3	1	(*109*)
			CoTC		(*110*)
Felis catus	cat	4731	HDV-like/CPEB3	1	(*109*)
Gallus gallus domesticus	chicken	15411	HHRz	10	(*27, 28*)
Gorilla gorilla gorilla	western gorilla	46798	CoTC		(*110*)
			HDV-like/CPEB3	1	(*109*)

Table 5.1 continued

Homo sapien	Human	127	HHRz	43	(*27, 28, 111*)
			CoTC		(*110*)
			HDV-like/CPEB3	1	(*109, 112*)
Macaca mulatta	macaque	284705	HHRz	8	(*27, 28*)
			HDV-like/CPEB3	1	(*109*)
Meleagris gallopavo	turkey	221214	HHRz		(*111*)
Microcebus murinus	lemur	2913	HHRz	3	(*27, 28*)
Monodelphis domestica	opossum	0	HHRz	26	(*27, 28*)
Mus musculus	mouse	22	HHRz	41	(*27, 113*)
			HDV-like/CPEB3	1	(*109*)
Nasonia vitripennis	wasp	6169	Twister	47	(*58*)
			HDV-like/R2		(*90, 91*)
			HHRz	18	(*27*)
Neurospora crassa OR74A	fungi	13	VS		(*38*)
			HDV-like	1	(*83*)
Ornithorhynchus anatinus	platypus	200134	HHRz	5	(*28*)
Oryctolagus cuniculus	rabbit	3218	HDV-like/CPEB3	1	(*109*)
			HHRz	2	(*27*)
Oryza sativa Indica	rice	10627	HHRz		(*29*)
Oryza sativa Japonica	rice	55	Twister	8	(*58*)
			HHRz		(*29*)
Pan troglodytes	chimpanzee	5541	HHRz	18	(*27, 28*)
			HDV-like/CPEB3	1	(*109*)
Papio anubis	baboon	63229	CoTC		(*110*)
Pongo abelii	orangutan	8447	CoTC		(*110*)
			HDV-like/CPEB3	1	(*109*)
Populus trichocarpa	tree	2514	Twister	5	(*58*)
Rattus norvegicus	rat	578	HHRz	36	(*27, 28*)
			HDV-like/CPEB3	1	(*109*)
Schistosoma mansoni	flatworm	885	Twister	1051	(*58*)
			HHRz	25370	(*27, 28, 114*)

Table 5.1 continued

			HDV-like/RTE		(*81*)
Sorghum bicolor	grass	1535	Twister	15	(*58*)
Sus scrofa	Pig	4562	HHRz	4	(*27, 28*)
			CoTC		(*110*)
			HDV-like/CPEB3		(*109*)
Taeniopygia guttata	zebra finch	35359	HDV-like/R2	3	(*90*)
			HHRz		(*27, 28*)
Tribolium castaneum	bettle	2148	HDV-like/R2	10	(*90, 91*)
			HHRz	14	(*27, 28*)
Trypanosoma brucei	parasitic kinetoplastid	0	HDV-like		(*115*)
Vitis vinifera	grape	2065	HHRz	7	(*27*)
Xenopus tropicalis	frog	6811	HHRz	215	(*27, 29*)
Yarrowia lipolytica CLIB122	fungi	0	HHRz	14	(*28*)

Another bottleneck of this project has been the ability to computationally sort through the data that we've generated in order to narrow down a reasonable list of candidate sequences to screen for nucleolytic activity. Because small self-cleaving ribozymes have little sequence conservation, the ideal algorithm would have the ability to predict RNA motifs and compare them across the entire dataset. A related challenge is determining the length of the potential ribozyme itself. We are able to accurately map the location of the 2', 3'- cyclic phosphate, but self-cleaving ribozymes cleave at various positions along their sequence. For instance, the HDV ribozyme and related ribozymes cleave at the 5' most end of the ribozyme, keeping the sequence of the ribozyme contiguous. On the other hand, the hammerhead ribozyme cleaves itself internally, breaking the RNA strand into 2 parts. Despite these hurdles, the fast-paced advances of next-generation sequencing and algorithm design ensure these problems can be solved.

5.3. Selecting RNA binding RNAs: combining next-generation sequencing with selection

As touched on in the previous chapter, and a theme throughout this work, coupling existing techniques with next-generation sequencing can create very robust assays. It would be intriguing to observe the results of each round of *in vitro* selection. This would allow us to monitor the evolution of the randomized pool toward an enriched population of molecules with a high degree

of specificity and/or affinity almost in real-time. Comparative analysis of sequencing results has the power to resolve early-stage shifts toward favorable ribonucleotides to determine sequence conservation. These data can influence the subsequent rounds of selection, making it feasible to shift selection conditions towards your desired product. For this project specifically, it would be valuable to track the structure of the RNAs in the randomized T-box pool. The reason for this is because the primary goal is to create STARzymes with the selected aptamers by linking it with a Flexizyme. For this to work, we would need aptamers that don't sterically block the 3' CCA-tail, where the Flexizyme base pairs with the tRNA. Additionally, the aptamers must involve anticodon-specifier sequence binding to preserve the specificity that is the hallmark of STARzymes.

5.4. Final thoughts

The RNA World Hypothesis proposes that RNAs predate protein and even DNA. We can see the remnants of these molecular fossils present in modern ribosomes, RNase P, and other ribozymes. Expansion of the RNA universe underwent a rapid growth spurt in the 80's and 90's before cooling in the recent decade. By continuing journey for new ribozymes, we will not let the RNA World become extinct!

LIST OF REFERENCES

1.   L. T. Chow, J. M. Roberts, J. B. Lewis, T. R. Broker, A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. *Cell* **11**, 819-836 (1977).

2.   S. M. Berget, C. Moore, P. A. Sharp, Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**, 3171-3175 (1977).

3.   T. R. Cech, D. C. Rio, Localization of transcribed regions on extrachromosomal ribosomal RNA genes of Tetrahymena thermophila by R-loop mapping. *Proc Natl Acad Sci U S A* **76**, 5051-5055 (1979).

4.   A. J. Zaug, T. R. Cech, In vitro splicing of the ribosomal RNA precursor in nuclei of Tetrahymena. *Cell* **19**, 331-338 (1980).

5.   P. J. Grabowski, A. J. Zaug, T. R. Cech, The intervening sequence of the ribosomal RNA precursor is converted to a circular RNA in isolated nuclei of Tetrahymena. *Cell* **23**, 467-476 (1981).

6.   A. J. Zaug, T. R. Cech, The intervening sequence excised from the ribosomal RNA precursor of Tetrahymena contains a 5-terminal guanosine residue not encoded by the DNA. *Nucleic Acids Res* **10**, 2823-2838 (1982).

7.   T. R. Cech, A. J. Zaug, P. J. Grabowski, In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27**, 487-496 (1981).

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

9.   S. Altman, The road to RNase P. *Nat Struct Biol* **7**, 827-828 (2000).

10.   F. H. Crick, L. Barnett, S. Brenner, R. J. Watts-Tobin, General nature of the genetic code for proteins. *Nature* **192**, 1227-1232 (1961).

11.   G. Rovera, S. Berman, R. Baserga, Pulse labeling of RNA of mammalian cells. *Proc Natl Acad Sci U S A* **65**, 876-883 (1970).

12.   R. H. Burdon, Ribonucleic acid maturation in animal cells. *Prog Nucleic Acid Res Mol Biol* **11**, 33-79 (1971).

13.   M. W. Taylor, S. A. Volkers, B. K. Choe, J. G. Zeikus, Transfer RNA modifications and synthesis in animal cells. *Cancer Res* **31**, 688-693 (1971).

14.   S. Altman, J. D. Smith, Tyrosine tRNA precursor molecule polynucleotide sequence. *Nat New Biol* **233**, 35-39 (1971).

15.   S. Altman, Isolation of tyrosine tRNA precursor molecules. *Nat New Biol* **229**, 19-21 (1971).

16.   H. D. Robertson, S. Altman, J. D. Smith, Purification and properties of a specific Escherichia coli ribonuclease which cleaves a tyrosine transfer ribonucleic acid presursor. *J Biol Chem* **247**, 5243-5251 (1972).

17.   H. D. Robertson, R. E. Webster, N. D. Zinder, Purification and properties of ribonuclease III from Escherichia coli. *J Biol Chem* **243**, 82-91 (1968).

18.   B. C. Stark, R. Kole, E. J. Bowman, S. Altman, Ribonuclease P: an enzyme with an essential RNA component. *Proc Natl Acad Sci U S A* **75**, 3717-3721 (1978).

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

20. C. L. Peebles *et al.*, A self-splicing RNA excises an intron lariat. *Cell* **44**, 213-223 (1986).
21. B. Ruskin, J. M. Greene, M. R. Green, Cryptic branch point activation allows accurate in vitro splicing of human beta-globin intron mutants. *Cell* **41**, 833-844 (1985).
22. B. Ruskin, A. R. Krainer, T. Maniatis, M. R. Green, Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell* **38**, 317-331 (1984).
23. S. Johansen, V. M. Vogt, An intron in the nuclear ribosomal DNA of Didymium iridis codes for a group I ribozyme and a novel ribozyme that cooperate in self-splicing. *Cell* **76**, 725-734 (1994).
24. G. A. Prody, J. T. Bakos, J. M. Buzayan, I. R. Schneider, G. Bruening, Autolytic processing of dimeric plant virus satellite RNA. *Science* **231**, 1577-1580 (1986).
25. C. J. Hutchins, P. D. Rathjen, A. C. Forster, R. H. Symons, Self-cleavage of plus and minus RNA transcripts of avocado sunblotch viroid. *Nucleic Acids Res* **14**, 3627-3640 (1986).
26. A. C. Forster, R. H. Symons, Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell* **49**, 211-220 (1987).
27. C. Seehafer, A. Kalweit, G. Steger, S. Graf, C. Hammann, in *Rna*. (United States, 2011), vol. 17, pp. 21-26.
28. J. Perreault *et al.*, Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. *PLoS Comput Biol* **7**, e1002031 (2011).
29. C. Hammann, A. Luptak, J. Perreault, M. de la Pena, The ubiquitous hammerhead ribozyme. *Rna* **18**, 871-885 (2012).
30. J. M. Kaper, M. E. Tousignant, G. Steger, Nucleotide sequence predicts circularity and self-cleavage of 300-ribonucleotide satellite of arabis mosaic virus. *Biochem Biophys Res Commun* **154**, 318-325 (1988).
31. L. Rubino, M. E. Tousignant, G. Steger, J. M. Kaper, Nucleotide sequence and structural analysis of two satellite RNAs associated with chicory yellow mottle virus. *J Gen Virol* **71 (Pt 9)**, 1897-1903 (1990).
32. A. Hampel, R. Tritz, RNA catalytic properties of the minimum (-)sTRSV sequence. *Biochemistry* **28**, 4929-4933 (1989).
33. S. Makino *et al.*, Molecular cloning and sequencing of a human hepatitis delta (delta) virus RNA. *Nature* **329**, 343-346 (1987).
34. L. Sharmeen, M. Y. Kuo, G. Dinter-Gottlieb, J. Taylor, Antigenomic RNA of human hepatitis delta virus can undergo self-cleavage. *J Virol* **62**, 2674-2679 (1988).
35. M. Y. Kuo, L. Sharmeen, G. Dinter-Gottlieb, J. Taylor, Characterization of self-cleaving RNA sequences on the genome and antigenome of human hepatitis delta virus. *J Virol* **62**, 4439-4444 (1988).
36. J. H. Chen *et al.*, A 1.9 A crystal structure of the HDV ribozyme precleavage suggests both Lewis acid and general acid mechanisms contribute to phosphodiester cleavage. *Biochemistry* **49**, 6508-6518 (2010).
37. A. R. Ferre-D'Amare, K. Zhou, J. A. Doudna, Crystal structure of a hepatitis delta virus ribozyme. *Nature* **395**, 567-574 (1998).
38. B. J. Saville, R. A. Collins, A site-specific self-cleavage reaction performed by a novel RNA in Neurospora mitochondria. *Cell* **61**, 685-696 (1990).
39. M. T. Kuiper, A. M. Lambowitz, A novel reverse transcriptase activity associated with mitochondrial plasmids of Neurospora. *Cell* **55**, 693-704 (1988).

40.   A. Claude, THE CONSTITUTION OF PROTOPLASM. *Science* **97**, 451-456 (1943).
41.   G. E. Palade, The endoplasmic reticulum. *J Biophys Biochem Cytol* **2**, 85-98 (1956).
42.   G. E. Palade, P. Siekevitz, Liver microsomes; an integrated morphological and biochemical study. *J Biophys Biochem Cytol* **2**, 171-200 (1956).
43.   G. E. Palade, P. Siekevitz, Pancreatic microsomes; an integrated morphological and biochemical study. *J Biophys Biochem Cytol* **2**, 671-690 (1956).
44.   G. E. Palade, A small particulate component of the cytoplasm. *J Biophys Biochem Cytol* **1**, 59-68 (1955).
45.   R. E. Monro, Catalysis of peptide bond formation by 50 S ribosomal subunits from Escherichia coli. *J Mol Biol* **26**, 147-151 (1967).
46.   R. R. Traut, R. E. Monro, THE PUROMYCIN REACTION AND ITS RELATION TO PROTEIN SYNTHESIS. *J Mol Biol* **10**, 63-72 (1964).
47.   P. Nissen, J. Hansen, N. Ban, P. B. Moore, T. A. Steitz, The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920-930 (2000).
48.   N. Ban, P. Nissen, J. Hansen, P. B. Moore, T. A. Steitz, The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **289**, 905-920 (2000).
49.   A. Nahvi *et al.*, Genetic control by a metabolite binding mRNA. *Chem Biol* **9**, 1043 (2002).
50.   W. Winkler, A. Nahvi, R. R. Breaker, Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952-956 (2002).
51.   A. S. Mironov *et al.*, Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111**, 747-756 (2002).
52.   W. C. Winkler, A. Nahvi, A. Roth, J. A. Collins, R. R. Breaker, Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**, 281-286 (2004).
53.   S. Milewski, Glucosamine-6-phosphate synthase--the multi-facets enzyme. *Biochim Biophys Acta* **1597**, 173-192 (2002).
54.   J. A. Collins, I. Irnov, S. Baker, W. C. Winkler, Mechanism of mRNA destabilization by the glmS ribozyme. *Genes Dev* **21**, 3356-3368 (2007).
55.   P. Y. Watson, M. J. Fedor, The glmS riboswitch integrates signals from activating and inhibitory metabolites in vivo. *Nat Struct Mol Biol* **18**, 359-363 (2011).
56.   G. A. Soukup, Core requirements for glmS ribozyme self-cleavage reveal a putative pseudoknot structure. *Nucleic Acids Res* **34**, 968-975 (2006).
57.   J. E. Barrick *et al.*, New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci U S A* **101**, 6421-6426 (2004).
58.   A. Roth *et al.*, A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat Chem Biol* **10**, 56-60 (2014).
59.   Y. Liu, T. J. Wilson, S. A. McPhee, D. M. Lilley, Crystal structure and mechanistic investigation of the twister ribozyme. *Nat Chem Biol* **10**, 739-744 (2014).
60.   Z. Weinberg *et al.*, New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nat Chem Biol* **11**, 606-610 (2015).
61.   K. Schutz, J. R. Hesselberth, S. Fields, Capture and sequence analysis of RNAs with terminal 2',3'-cyclic phosphates. *Rna* **16**, 621-631 (2010).
62.   S. Mohr *et al.*, Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *Rna* **19**, 958-970 (2013).

63. P. J. Enyeart, G. Mohr, A. D. Ellington, A. M. Lambowitz, Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and non-coding RNA analysis. *Mob DNA* **5**, 2 (2014).

64. K. Collins, T. W. Nilsen, Enzyme engineering through evolution: thermostable recombinant group II intron reverse transcriptases provide new tools for RNA research and biotechnology. *Rna* **19**, 1017-1018 (2013).

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

66. D. Blankenberg *et al.*, Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**, 1783-1785 (2010).

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

68. A. Yates *et al.*, Ensembl 2016. *Nucleic Acids Res* **44**, D710-716 (2016).

69. B. L. Aken *et al.*, Ensembl 2017. *Nucleic Acids Res* **45**, D635-d642 (2017).

70. G. I. Giraldo-Calderón *et al.*, VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res* **43**, D707-713 (2015).

71. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).

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

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

74. D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Stromberg, G. T. Marth, BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691-1692 (2011).

75. H. Thorvaldsdottir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192 (2013).

76. J. T. Robinson *et al.*, in *Nat Biotechnol*. (United States, 2011), vol. 29, pp. 24-26.

77. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

78. V. Nene *et al.*, Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* **316**, 1718-1723 (2007).

79. J. K. Biedler, Z. Tu, The Juan non-LTR retrotransposon in mosquitoes: genomic impact, vertical transmission and indications of recent and widespread activity. *BMC Evol Biol* **7**, 112 (2007).

80. K. Licht, J. Medenbach, R. Luhrmann, C. Kambach, A. Bindereif, 3'-cyclic phosphorylation of U6 snRNA leads to recruitment of recycling factor p110 through LSm proteins. *Rna* **14**, 1532-1538 (2008).

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

82. N. R. Pamudurti *et al.*, Translation of CircRNAs. *Mol Cell* **66**, 9-21.e27 (2017).

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

84. E. Lund, J. E. Dahlberg, Cyclic 2',3'-phosphates and nontemplated nucleotides at the 3' end of spliceosomal U6 small nuclear RNA's. *Science* **255**, 327-330 (1992).

85. C. L. Peebles, P. Gegenheimer, J. Abelson, Precise excision of intervening sequences from precursor tRNAs by a membrane-associated yeast endonuclease. *Cell* **32**, 525-536 (1983).

86. M. Englert, H. Beier, Plant tRNA ligases are multifunctional enzymes that have diverged in sequence and substrate specificity from RNA ligases of other phylogenetic origins. *Nucleic Acids Res* **33**, 388-399 (2005).

87. R. C. Schwartz, C. L. Greer, P. Gegenheimer, J. Abelson, Enzymatic mechanism of an RNA ligase from wheat germ. *J Biol Chem* **258**, 8374-8383 (1983).

88. B. L. Apostol, C. L. Greer, Preferential binding of yeast tRNA ligase to pre-tRNA substrates. *Nucleic Acids Res* **19**, 1853-1860 (1991).

89. L. Pick, H. Furneaux, J. Hurwitz, Purification of wheat germ RNA ligase. II. Mechanism of action of wheat germ RNA ligase. *J Biol Chem* **261**, 6694-6704 (1986).

90. D. G. Eickbush, W. D. Burke, T. H. Eickbush, Evolution of the R2 retrotransposon ribozyme and its self-cleavage site. *PLoS One* **8**, e66441 (2013).

91. D. G. Eickbush, T. H. Eickbush, R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA cotranscript. *Mol Cell Biol* **30**, 3142-3150 (2010).

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

93. J. Abelson, C. R. Trotta, H. Li, tRNA splicing. *J Biol Chem* **273**, 12685-12688 (1998).

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

95. K. Darty, A. Denise, Y. Ponty, VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974-1975 (2009).

96. I. Nikic, E. A. Lemke, Genetic code expansion enabled site-specific dual-color protein labeling: superresolution microscopy and beyond. *Curr Opin Chem Biol* **28**, 164-173 (2015).

97. G. Fan, Z. Wang, M. Hao, J. Li, in *J Hematol Oncol*. (London, 2015), vol. 8.

98. L. Si *et al.*, Generation of influenza A viruses as live but replication-incompetent virus vaccines. *Science* **354**, 1170-1173 (2016).

99. L. Wang, A. Brock, B. Herberich, P. G. Schultz, Expanding the genetic code of Escherichia coli. *Science* **292**, 498-500 (2001).

100. H. Murakami, A. Ohta, H. Ashigai, H. Suga, A highly flexible tRNA acylation method for non-natural polypeptide synthesis. *Nat Methods* **3**, 357-359 (2006).

101. J. Chen, G. Manuel, R. Chamberlain, A. Luptak, B. L. Golden. (In preparation).

102. R. Gan, B. L. Golden. (In preparation).

103. G. Srinivasan, C. M. James, J. A. Krzycki, Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science* **296**, 1459-1462 (2002).

104. A. Ambrogelly *et al.*, Pyrrolysine is not hardwired for cotranslational insertion at UAG codons. *Proc Natl Acad Sci U S A* **104**, 3141-3146 (2007).

105. S. Gundllapalli *et al.*, Misacylation of pyrrolysine tRNA in vitro and in vivo. *FEBS Lett* **582**, 3353-3358 (2008).

106. W. T. Li *et al.*, Specificity of pyrrolysyl-tRNA synthetase for pyrrolysine and pyrrolysine analogs. *J Mol Biol* **385**, 1156-1164 (2009).

107. K. Nozawa *et al.*, Pyrrolysyl-tRNA synthetase-tRNA(Pyl) structure reveals the molecular basis of orthogonality. *Nature* **457**, 1163-1167 (2009).
108. R. Przybilski *et al.*, Functional hammerhead ribozymes naturally encoded in the genome of Arabidopsis thaliana. *Plant Cell* **17**, 1877-1885 (2005).
109. C. H. Webb, A. Luptak, HDV-like self-cleaving ribozymes. *RNA Biol* **8**, 719-727 (2011).
110. A. Teixeira *et al.*, Autocatalytic RNA cleavage in the human beta-globin pre-mRNA promotes transcription termination. *Nature* **432**, 526-530 (2004).
111. M. de la Pena, I. Garcia-Robles, Intronic hammerhead ribozymes are ultraconserved in the human genome. *EMBO Rep* **11**, 711-716 (2010).
112. K. Salehi-Ashtiani, A. Luptak, A. Litovchick, J. W. Szostak, A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science* **313**, 1788-1792 (2006).
113. M. Martick, L. H. Horan, H. F. Noller, W. G. Scott, A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. *Nature* **454**, 899-902 (2008).
114. G. Ferbeyre, J. M. Smith, R. Cedergren, Schistosome satellite DNA encodes active hammerhead ribozymes. *Mol Cell Biol* **18**, 3880-3888 (1998).
115. F. J. Sanchez-Luque, M. C. Lopez, F. Macias, C. Alonso, M. C. Thomas, Identification of an hepatitis delta virus-like ribozyme at the mRNA 5'-end of the L1Tc retrotransposon from Trypanosoma cruzi. *Nucleic Acids Res* **39**, 8065-8077 (2011).

VITA

SAMANTHA WAYLIN LEE
DEPARTMENT OF BIOCHEMISTRY | COLLEGE OF AGRICULTURE
PURDUE UNIVERSITY

AREAS OF EXPERTISE AND TECHNICAL SKILLS

1. Developed and executed a customized RNA sequencing assay to identify novel highly structured RNA motifs in *D. melanogaster*, *A. aegypti*, and *A. gambiae*.
   Skills: NGS assay development and troubleshooting, total RNA purification, protein expression and purification, analysis of bioinformatics data, and fluorescent assays.
2. Designed and executed RNA SELEX for the selection of tRNA-binding aptamers.
   Skills: SELEX assay development and troubleshooting, creation of RNA expression constructs, expression and purification of T7 RNA polymerase, *in vitro* transcription and RNA purification, optimization of RNA folding, affinity chromatography, size-exclusion chromatography, and binding kinetics.
3. Performed high-throughput screening for the identification of human telomerase enzyme upregulators using a qPCR-based assay in HDFa, HeLa, and HeK293 human cell lines.
   Skills: maintenance and operation of liquid handling automation, compound library maintenance, cell treatment, cell viability assays, gene expression data analysis, and authorship of reports for distribution to clients.
4. Engineered plant metabolic systems for the uptake and storage of phosphate compounds for the phytoremediation of wastewater.
   Skills: synthetic biology, cloning, *Agrobacterium* transfection, genotype and phenotype characterization, and plant tissue/cell culture.

EDUCATION

2013 – 2019	Ph.D., Biochemistry, Purdue University
	Dr. Barbara Golden, advisor
2010 – 2011	M.S., Biotechnology, University of Nevada, Reno
	Dr. David Shintani, advisor
2009 – 2011	B.S., Biochemistry, University of Nevada, Reno
	Dr. David Shintani, advisor

RESEARCH EXPERIENCE

2014 – 2019	Graduate Research Assistant, Department of Biochemistry, Purdue University
	Dr. Barbara Golden, advisor
2011 – 2013	Research Associate II, Sierra Sciences
	Lancer Brown, supervisor

2010 – 2011	International Genetically Engineered Machines Competition, Massachusetts Institute of Technology Dr. Christie Howard and Dr. David Shintani, advisors
2009 – 2011	Undergraduate and Graduate Research Assistant, Department of Biochemistry and Molecular Biology, University of Nevada, Reno Dr. David Shintani, advisor

PUBLICATIONS

1. Lee S, Mir A, Golden BL. 2019. A method for selective sequencing of cyclic phosphate-terminal RNAs. In preparation.

AWARDS, SCHOLARSHIPS, AND FELLOWSHIPS

2018 – 2019	Maxine and Edward Nichols Fellowship Purdue University
2018	Mentoring Award for Graduate Students The Graduate School, Purdue University
2015 – 2018	Bird Stair Research Fellowship Department of Biochemistry, Purdue University
2013	Incentive Grant The Graduate School, Purdue University
2010	Ronald E. McNair Scholar University of Nevada, Reno
2010	Finlay McDonald Scholarship College of Agriculture, Biotechnology & Natural Resources, University of Nevada, Reno

TEACHING AND MENTORING EXPERIENCE

2019	Graduate Teaching Assistant, Department of Biochemistry, Purdue University Dr. Frederick Gimble, course: Macromolecules
2013 – 2019	Research Advisor, International Genetically Engineered Machines, College of Engineering, Purdue University
2015, 2018	Mentor for Undergraduate Students, NSF Research Experience for Undergraduates, Department of Biochemistry, Purdue University
2017	Graduate Teaching Assistant, Department of Biochemistry, Purdue University Dr. Barbara Golden, course: Macromolecular Machines Dr. Orla Hart, course: Biochemistry
2014 – 2019	Mentor for Undergraduate and Graduate Students, Department of Biochemistry, Purdue University

2017	Peer Mentor, Mentoring@Purdue, College of Agriculture, Purdue University
2012 – 2013	Instructor, Department of Biology, Truckee Meadows Community College Courses: Introduction to Cell and Molecular Biology, General Botany

SERVICE

2017 – 2018	Senator, Purdue Graduate Student Government, Purdue University
2017 – 2018	Life Committee, Purdue Graduate Student Government, Purdue University
2015 – 2017	Chair and Organizer, Hitchhiker's Guide to the Biomolecular Galaxy Symposium, Purdue University
2014 – 2017	Chair and Selection Committee Member, Student-Invited Seminar Committee, Department of Biochemistry, Purdue University

PRESENTATIONS

2017	Poster Presentation: "Ribozyme Hunting", Department of Biochemistry Research Retreat, West Lafayette, IN
2016	Poster Presentation: "Ribozyme Hunting" Department of Biochemistry Research Retreat, West Lafayette, IN
2015	Poster Presentation: "Ribozyme Hunting" Department of Biochemistry Research Retreat, West Lafayette, IN
2014	Poster Presentation: "Peptide synthesis using Stem I of T-box riboswitches and a trans-acting peptidyl-transferase ribozyme", Department of Biochemistry Research Retreat Lafayette, IN
2011	Invited Oral Presentation: "Using the biosynthesis of inositol hexa*kis*phosphate as a method for the bioremediation of wastewater", Nevada Biotechnology Symposium, Reno, NV
2010	Invited Oral Presentation: "Nevada iSense: Development of plant biosensors using *Nicotiana tabacum* as a transgenic plant model", iGEM Jamboree, Cambridge, MA
2010	Invited Oral Presentation: "Using the biosynthesis of inositol hexa*kis*phosphate as a method for the bioremediation of wastewater", McNair Symposium, Berkeley, CA
2010	Invited Oral Presentation: "Using the biosynthesis of inositol hexa*kis*phosphate as a method for the bioremediation of wastewater", Nevada Biotechnology Symposium, Reno, NV