5-2021

Biochemical and Computational Characterization of Small Regulatory RNAs in the Eastern Oyster, Crassostrea virginica

Isabelle Townsend

Biochemical and Computational Characterization of Small Regulatory RNAs in the Eastern Oyster, Crassostrea virginica

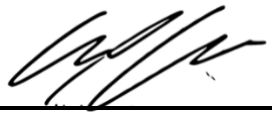
by

Isabelle Townsend

A Thesis
Submitted to the Honors College of
The University of Southern Mississippi
in Partial Fulfillment
of Honors Requirements

May 2021

Approved by:



Alex Flynt, Ph.D., Thesis Advisor,
School of Biological, Environmental and Earth Sciences

Bernd Schroeder, Ph.D., Director,
School of Mathematics and Natural Sciences

Ellen Weinauer, Ph.D., Dean
Honors College

ABSTRACT

RNA interference (RNAi) is a process by which gene expression is regulated using small non-coding RNAs (ncRNAs). Three classes of ncRNAs, including micro-RNA (miRNA), short-interfering RNA (siRNA), and Piwi-interacting RNA (piRNA), are readily distinguishable in eukaryotic systems based on unique characteristics such as read sizes, overlap signatures, and mode of biogenesis. In this study, a method for purification of small RNAs was explored in the eastern oyster, *Crassostrea virginica*. This method involved the use of Sepharose beads for anion exchange chromatography to enhance purification of Argonaute associated small RNAs. Following RNA extraction and purification, small RNA libraries were created and sequenced on the Illumina platform. The results showed that this approach did not accurately reflect the population of ncRNA in the species as observed in the non-bead-treated total RNAs. However, further improvements to the methodology described here could aid in efficiency in recovering small RNA populations that would minimize the presence of other non-Argonaute associated ncRNAs.

Keywords: *miRNA; siRNA; piRNA; Argonaute protein family; anion-exchange chromatography; RNAi; small RNA biogenesis*

DEDICATION

To my little sister, Ella. There is no one on Earth more creative or kind than her. She is an answered prayer, and I am so blessed to have her in my life.

ACKNOWLEDGMENTS

I would like to acknowledge everyone involved in the making of this thesis. Firstly, I want to thank my thesis advisor, Dr. Alex Flynt, for all the effort and encouragement he has given me. I would also like to thank my lab mentor and friend Jacob Peter for taking the time and effort to assist me and teach me throughout the research process. Jay pushed me to be a better researcher and writer, and my time in lab would not have been nearly as enjoyable without him. Thank you to Dr. Flynt's other graduate students, Farid Zia and Sweata Khanal, for always coming to my aid and enjoying all the smiley faces I drew them.

I would like to thank the Honors College for the supportive community it provided. Dr. Weinauer, Dr. Heinhorst, and all the Honors College staff have always been invested in me and encouraged my passion for learning. My time at USM would not have been the grand experience it was without any of you.

I would also like to thank my parents for many ways they have supported me throughout my life. I have accomplished what I have because of their dedication to me and my success. I also want to thank my sister, to whom this thesis is dedicated to, for always uplifting me and pushing me to be a better person. Thank you to all my friends, who have been sources of comfort, fun, and encouragement.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS

RNAi	RNA Interference
ncRNA	Non-Coding RNA
miRNA	microRNA
siRNA	short-interfering RNA
piRNA	Piwi-interacting RNA
RDRP	RNA-Dependent RNA Polymerases
ssRNA	Single-Stranded RNA
pri-miRNA	Primary miRNA
pre-miRNA	Precursor miRNA
RISC	RNA-Induced Silencing Complex
Ago	Argonaute
AGO	Argonaute clade
PIWI	Argonaute clade
TE	Transposable Element
Aub	Aubergine
U	Uracil
A	Adenosine
RPM	Reads Per Million
nt	Nucleotide

CHAPTER I: INTRODUCTION

The Central Dogma of biology is the process by which the genetic information encoded by DNA is transcribed and translated for gene expression. This complex system is vital for life given that all structural and functional requirements for an organism are detailed in their genetic code. Thus, precise regulation of transcription and translation is fundamental to cell growth, development, and response to external stimuli. One important regulator of gene expression are small non-coding RNA (ncRNA) molecules. These short strands of nucleic acids (20 to 31 nucleotides) regulate gene expression through RNA-RNA interactions in a process known as RNA interference (RNAi)[1, 2]. RNAi pathways are present in all multicellular eukaryotic organisms and include both transcriptional and post-transcriptional regulations that involving three major classes of small RNAs, namely microRNA (miRNA), short-interfering RNA (siRNA), and Piwi-interacting RNA (piRNA)[3-5].

While the roles of these molecules have been explored in various model organisms (*Drosophila melanogaster* and *Caenorhabditis elegans*), methods for defining and characterizing these small RNAs are needed to investigate their biology in non-model organisms, including members of phylum Lophotrochozoa[5,6]. This study sought to establish a method for identification and characterization of small RNAs in the eastern oyster, *Crassostrea virginica*, using both biochemical and computational approaches.

Crassostrea virginica and RNAi Factors

Multiple ribonuclease and RNA binding proteins participate in RNAi, such as the type III endoribonucleases, Dicer and Drosha, and proteins of the Argonaute family, such

as AGO and PIWI, as well as RNA-dependent RNA polymerases (RDRP)[7-9]. The combination of these proteins in a particular eukaryotic system determines the capacity for RNAi in an organism. For example, fruit flies possess a dedicated Ago for miRNA and another Ago for siRNA type silencing (Figure 1A)[10]. This combination may not be the same in other clades. Also, while some clades have several homologs of Ago and Piwi, such as in nematodes and humans, other clades only have a few homologs for carrying out RNAi. Similarly, *D. melanogaster* has three Ago proteins and one Piwi, which is very divergent from that seen in humans[10]. Additionally, the presence of an RDRP as an RNAi factor in *C. elegans* makes RNAi pathways in this clade even more different from clades where the protein is lacking, such as flies[6].

Organisms in the Bilateria phylum, Lophotrochozoan, also display specific RNAi factors. In particular, bivalves (*Crassostrea* species) possess a unique combination of RNAi factors when compared to other members of the clade. Species within this phylum have at least one Ago and two Piwi proteins, and most do not have an RDRP (Figure 1B). However, bivalves have a higher number of Ago (2), Piwi (3), and RDRP (2) than most members of the clade (Figure 1B). Given that each of these proteins would serve a specific function in RNAi pathways, this suggests that Bivalves, such as oysters, would have a distinct RNAi pathway compared to other members of the clade.

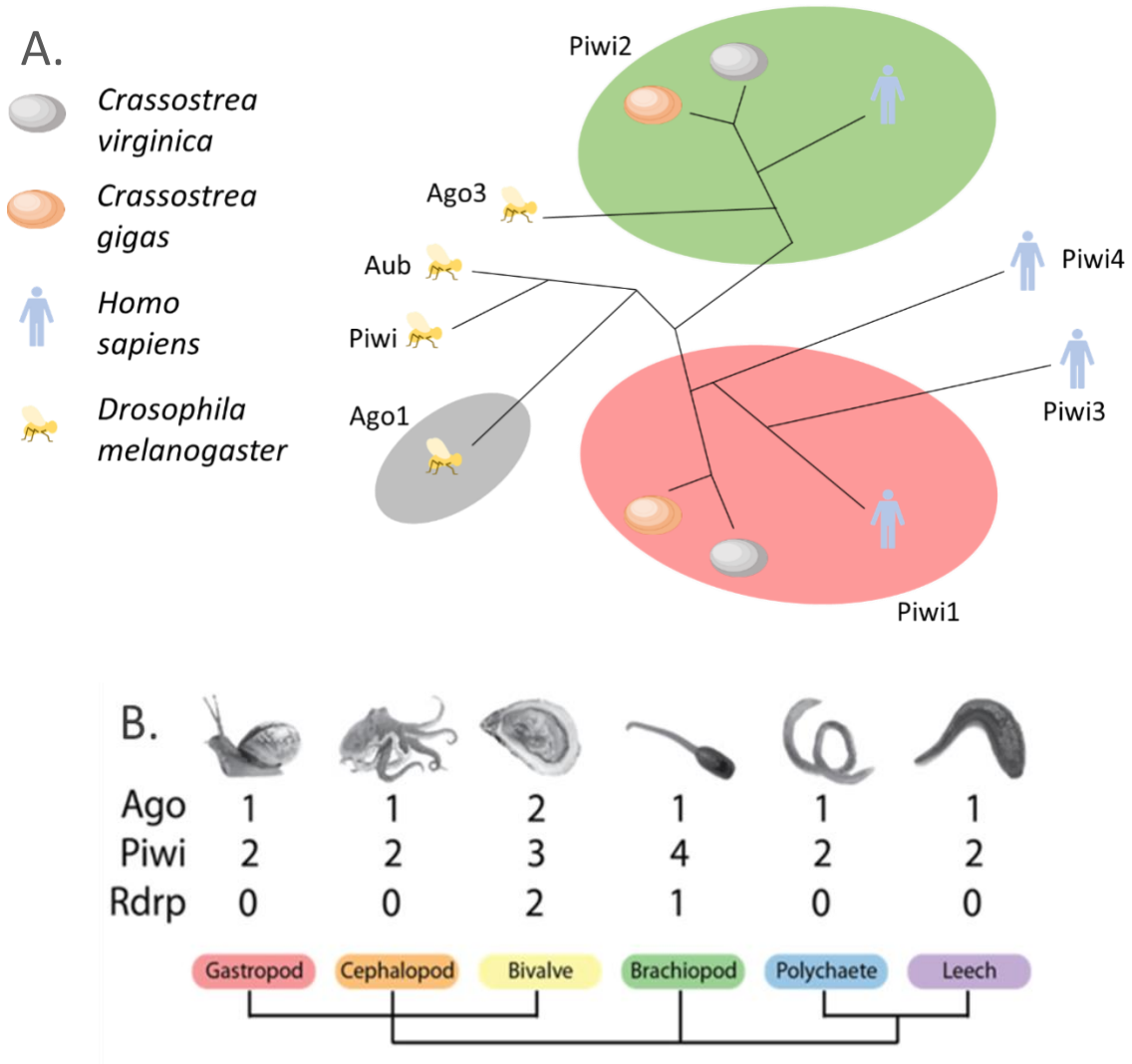


Figure 1: Phylogenetic relationship of Argonaute proteins in C. virginica. (A) shows the relationship in Argonaute proteins between *Crassostrea*, flies, and humans. The green region represents Piwi2 proteins, the pink region represents Piwi1, and the grey represents the miRNA dedicated Ago1 in flies[10]. **(B)** Distribution of RNAi factors across several clades in the lophotrochozoan phylum.

Function of Argonaute Protein Family

Central to the process of RNAi are proteins of the Argonaute family, which are broadly divided into two major subfamilies: AGO and PIWI. While the AGO clade recruits miRNAs and siRNAs, the PIWI clade interacts with piRNAs to regulate transposon activity in germline and somatic cells[3,11-13]. The role of AGO proteins in miRNA function is highly conserved in eukaryotes, while siRNA biology tends to be more species-specific[14]. The proteins' catalytic cycle primarily includes a guide-binding and target-recognition stage, annealing, and target cleavage[15,16]. The four domains, N-terminal, PIWI, MID, and PAZ (PIWI-Ago-Zwille), are crucial for the proteins' mechanisms of action. Ago active sites are located in the PIWI domain, which contains RNase that participate in the cleavage of targeted mRNAs according to the small RNA guide. Meanwhile, the other domains are responsible for the anchoring and unwinding of the guide RNA duplex[3].

Biogenesis of Major Classes of Small RNAs

miRNAs are the most conserved class of small RNAs, and found in virtually all eukaryotes[9,17]. These endogenous, short segments ranging from 21 to 22 nucleotides (nt) in length are instrumental in post-transcriptional regulation of genes[18]. miRNAs are derived from single-stranded RNA (ssRNA) structures in the nucleus that fold back into hairpins (Figure 2)[19]. RNA polymerase II transcribes primary miRNA (pri-miRNA) that are 70 nt long. The transcribed RNA forms a structure containing a terminal loop that is then processed by the RNase III enzyme Drosha to generate a precursor miRNA (pre-miRNA). These pre-miRNAs form hairpin structures that are then transported into the

cytoplasm by the protein Exportin5 through nuclear pores[20]. In the cytoplasm, they are processed by the RNase III enzyme Dicer, which generates miRNA duplexes. The duplex is loaded into AGO to form an RNA-Induced Silencing Complex (RISC), after which strand selection occur, and the strand with the most stable 5' end is retained. This strand then directs the AGO protein to its target[9].

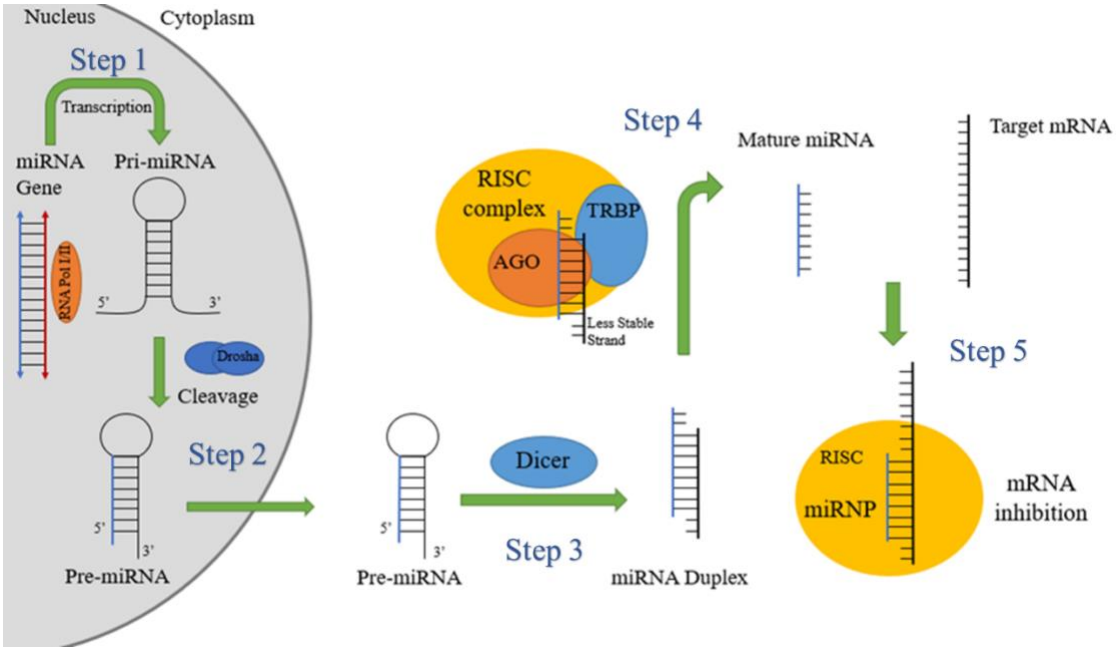


Figure 2: Schematic of miRNA biogenesis and gene silencing pathway. 1) miRNAs originate as pri-miRNA transcripts that are folded into hairpin structures. **2)** These are then processed by the protein drosha into pre-miRNA, which are exported from the nucleus into the cytoplasm. **3)** Dicer cleaves pre-miRNA into miRNA duplexes. **4)** The duplexes are loaded into AGO to form RISC, and the more stable strand is kept as mature miRNA. **5)** miRNA then targets specific mRNA for post-transcriptional inhibition [19,20].

Though similar to miRNA in size, siRNAs are distinct in their biogenesis[21]. siRNAs are composed of 21 to 23 nt long ncRNAs and are derived from dsRNA precursors such as viral dsRNA, transposable elements, dsRNA of exogenous sources, and long hairpin dsRNAs (Figure 3)[22,23]. They can also arise from the replication of RNA viruses, lending them to a role in antiviral immune responses[24]. siRNA biogenesis begins with dsRNA production by the DNA-dependent RNA polymerase II and are then matured by Dicer in the cytoplasm. The resulting short dsRNA are then loaded into AGO proteins to form RISC. These then go on to direct AGO to their specific targets (Figure 3). siRNA can also regulate genes directly by targeting and forming complementary base pairs with genes to disrupt translation[25]. Given this ability to induce gene knockdown, siRNA have been studied as potential tools in disease treatment, such as with viral infections[4,26].

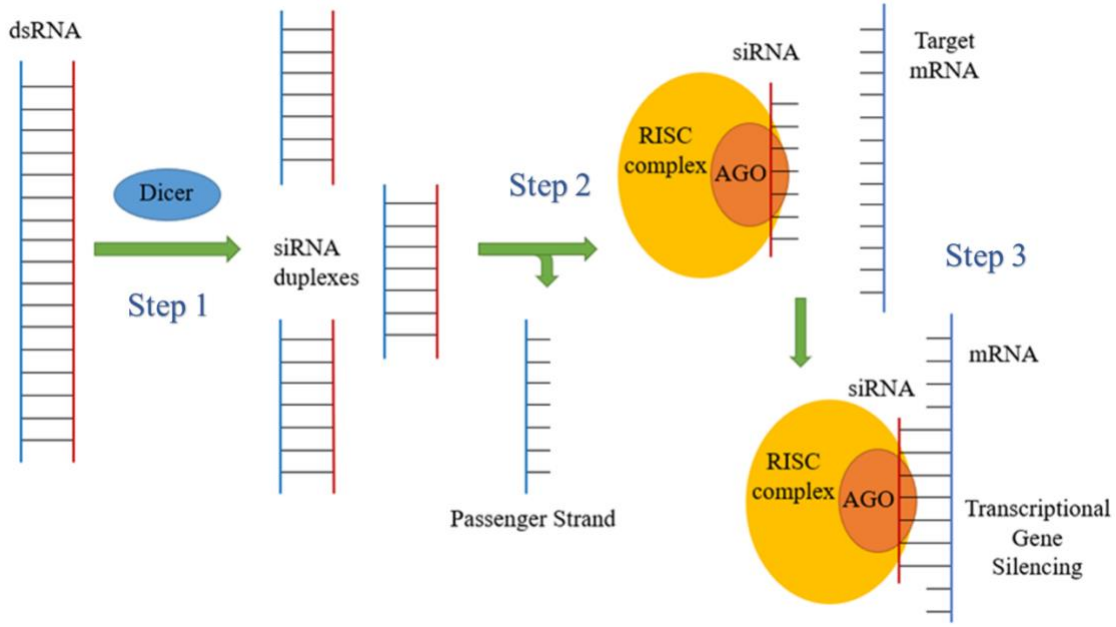


Figure 3: Schematic of siRNA biogenesis and gene silencing pathway. 1) dsRNA precursors are cleaved by Dicer to form siRNA duplexes. **2)** The duplexes are loaded into

AGO to form RISC, and the more stable strand is kept as mature siRNA. **3)** siRNA induces gene silencing via mRNA targeting in RISC or direct mRNA disruption during translation by complementarily binding to mRNA targets[25].

While miRNA and siRNA associate with AGO clade proteins, piRNA are defined by their loading into Piwi-type Argonaute proteins[27]. Piwi proteins were initially thought to be associated specifically in germ cells, suggesting a major role in germline biology[24]. In many organisms, Piwi proteins protect the genome from transposable elements (TEs), which are parasitic genetic elements that can proliferate within genomes and cause insertional mutagenesis that leads to cell dysfunction and death[10,28]. As seen with other small RNA varieties, piRNAs bound to Piwi complementarily base pair with TE transcripts during gametogenesis to suppress their expression and potential for mobilization[27]. To fully understand how it is piRNA accomplish this, it is important to understand how they arise since their biogenesis connects to how they function.

piRNAs arise from single-stranded precursor transcripts in one of two ways, as seen in Figure 4. The first involves the 5' end of a "primary" piRNA being loaded into Piwi after generation by Zucchini [9]. The second "ping pong" method uses piRNA-guided splicing to cleave complementary strands from a secondary piRNA loaded into either Aubergine (Aub) or Ago3[27]. During piRNA production, extant piRNAs initiate cleavage of piRNA cluster transcripts that are rich in TE remnants that then become substrates for Zucchini to processivity generate piRNAs[24,27,29]. These piRNAs then will participate in the self-amplifying Ping Pong cycle, leading to accelerating destruction of TE transcripts. In this, Piwi paralogs loaded with piRNA transcripts recognize TE signatures, cleave the TE fragment, and then generate a novel piRNA from the specific fragment[28]. The cycle is

able to continue due to the consistent structure of the piRNA strands. Characteristic of primary piRNA, there is a 10 nucleotide overhang with a bias for uracil (U) residues on the 5' end and an adenosine (A) present 10 nucleotides from the 5' end[27]. A targeted transposon mRNA complementary to the piRNA generated from piRNA clusters is then drawn into the piRNA/Piwi or Aub complex. The piRNA is cleaved at precisely 10 nucleotides from the 5' end, which activates the target, and a loaded AGO3 complex which then cleaves the complementary target[24]. Trimming of the 3' end marks the production of a mature piRNA[27]. The cleavage can then be copied by AGO3 to make more complementary piRNA that are able to load onto the Aub complexes and silence the activated targets[24]. This silencing method by cleaving the TE fragments and generating piRNA in response to activated targets is entirely self-amplifying and behaves not unlike a ping pong match[28]. This all occurs in germline stem cells in an effort to protect the genome from these mobile elements and can also defend the genome against parasitic nucleic acids. piRNAs and Piwi proteins are also involved in germline development, cell division, and cell maintenance[24].

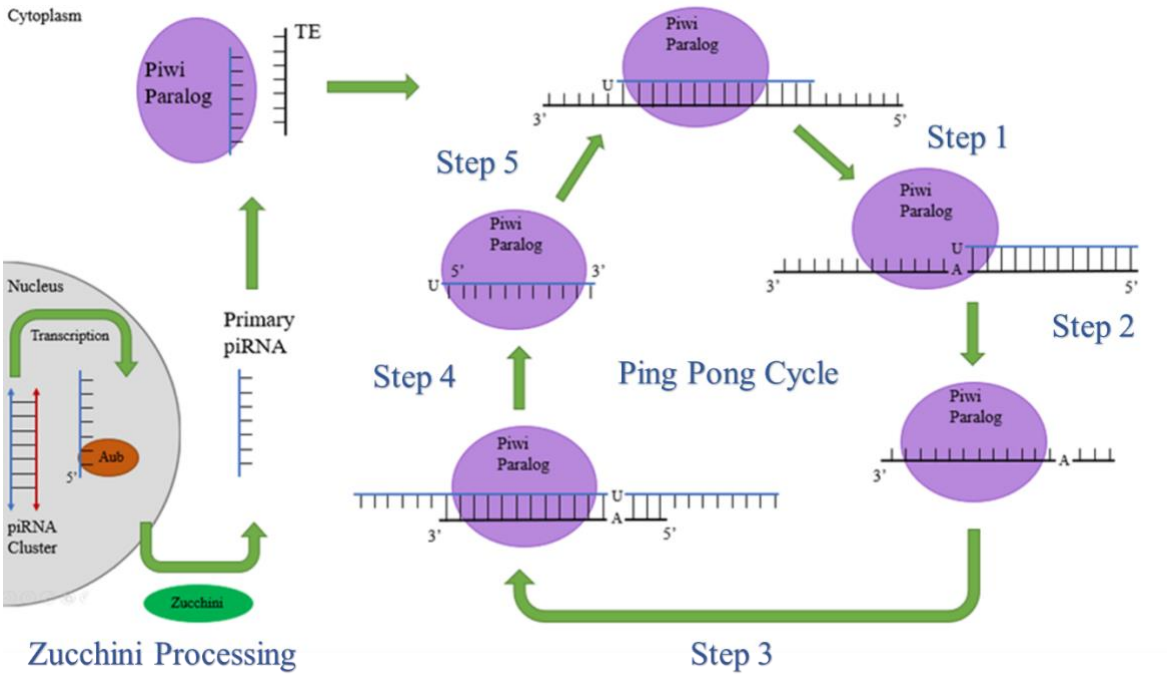


Figure 4: Schematic of piRNA biogenesis (ping pong cycle) and gene silencing pathway. piRNAs are generated either through Zucchini processing following transcription of piRNA clusters (left) or the ping pong cycle (right). In this cycle, **1)** Piwi paralogs loaded with primary piRNAs recognize specific transposable elements (TE), **2)** the TE fragments are cleaved, **3)** novel piRNA complementary to the TE fragment are generated, **4)** the TE fragment is removed, **5)** and the process can begin again, marking it as self-amplifying[9,28].

Previous research has shown that the characteristic 10 nucleotide 3' overhang is highly conserved in piRNAs, even though piRNA sequences themselves are not[27]. It has been noted that, unlike siRNA and miRNA, piRNA biogenesis is Dicer-independent, relying on only proteins in the Argonaute family[24]. Though the pathways used for post-transcriptional regulation of protein-coding genes and TE silencing are known to a certain

degree, the exact mechanisms behind their function and biogenesis are not fully understood[28]. It was believed for some time that their roles were strictly limited to germ cells, but research has shown the piRNA pathway is required by some organisms in their somatic cells, such as for adaptive immunity against viruses[30]. For example, in *Crassostrea gigas* Piwi mRNAs are expressed in the gonads, labial palps, gills, muscle, and mantle, and the piRNA clusters expressed in the germline were not the same as those in the soma[10]. This lends to the idea that piRNA is not conserved in sequence and is tissue specific[28].

Isolated Small RNAs from the Eastern Oyster *C. virginica*

The classes of small RNAs and their biogenesis can be analyzed by read overlaps, nucleotide sequences, and size[14]. In this current study, the focus was on improving Argonaute protein extraction techniques and analyzing the results based on small RNA read sizes. Extraction protocols were assessed by analysis of transcriptomic data. This was done through examining distribution of reads 18 to 32 nt long following genome mapping. Given that siRNA and miRNA range in size from 20 to 22 nt and piRNA range in size from 25 to 32 nt, the three classes were mapped into two size range libraries, maintaining the abundance of each possible length within the ranges. While the typical functions of each class of small RNA is known throughout other organisms, this distribution data for *C. virginica* provides suggestions of which Ago proteins are involved in the function of small RNAs expressed from different regions. With siRNA and miRNA forming complexes with AGO-clade proteins and the close connection

between piRNA and PIWI, the distribution of the ncRNA correlates with the expression of AGO and PIWI proteins.

CHAPTER II: MATERIALS AND METHODS

Q Sepharose and Trizol RNA Extractions

Gonad, gill, and adductor muscle tissues were dissected aseptically from live *C. virginica* obtained from a seafood market in Gulfport. Samples were flash frozen in liquid nitrogen. Frozen samples were then ground with a mortar and pestle into homogenized tissues. The materials from these extractions were subjected first to protein isolation, using Q Sepharose beads, followed by RNA extractions with trizol. The bead protocol was modeled from HiTrap Q Column chromatography techniques for quaternary ammonium anion exchange resin, which has been used in previous studies with other non-model organisms[14]. The process works by separating charged amino acids. Due to Argonaute proteins being highly basic, these riboprotein complexes bind to the Q Sepharose beads and are collected while non-riboproteins bypass the beads and are discarded. As a control, homogenized tissues from the gonads were subjected to protein extractions without the use of Q Sepharose beads followed by a total RNA extraction. Comparisons were made between the control and bead treated samples by observing the resulting populations of RNAs within a 20 nt to 40 nt range. Extractions that yielded small RNAs in this range indicated selective purification of proteins associated with small RNAs, namely those in the Ago family. The exact quantity of beads and specific techniques used for extractions were altered over the course of the study to determine the optimal protocol for selective protein extractions. It is noted the bead extraction protocol is not yet perfected. However, the most promising results were obtained by the methodology reported here.

Six samples were taken for extraction, two from gills, two from gonads, and two from adductor muscle. Crushed tissues, massing between 1 g to 1.1 g, were resuspended in 600 μL of a binding buffer (0.4 mL HEPES, 1 mL 10% glycerol, 333.3 μL 0.1M KOAc, 4 μL 0.2M EDTA, 15 μL 1.5 mM $MgCl_2$, 10 μL 1M DTT, 400 μL 200mM NaCl, 5 μL NP-40, 7.833 mL DI water, 1x Roche Complete EDTA-free protease inhibitor cocktail; pH 8.6) and maintained at 4˚C throughout the extraction procedure. Each tube was gently vortexed before being rocked for 10 minutes. Meanwhile, six tubes of beads were washed with the same buffer to ensure a consistent pH. Eppendorf tubes were filled with 800-1000 μL of resuspended Q Sepharose beads, which were washed twice with 400 μL binding buffer. After washing, 500-600 μL of supernatant was added to new tubes with 1 μL of RNase solution. Next, 400 μL of the protein mix was then transferred to beads and gently mixed. The new tubes were rocked for 15 minutes, and then centrifuged at 5k rpm for 5 minutes. Without removing the tubes from the centrifuge so as not to disturb the spun down beads, 200-300 μL of the supernatant was removed and transferred to new tubes with 300 μL acid phenol chloroform. These were gently shaken, centrifuged at maximum rpm, and 200 μL of the supernatant was transferred to new Eppendorf tubes with 800 μL isopropanol. Samples were then frozen at -20ºC overnight. After the first day of the bead extraction, total RNA was extracted with Trizol from bead purified proteins and crude protein extracts. All six samples and a control were centrifuged at maximum rpm, and the remaining liquid was discarded. The resulting pellets were washed once with 1 mL of 70% ethanol and centrifuged. The ethanol was discarded then pellets were resuspended with 25-30 μL nuclease-free water. Samples were either immediately analyzed for successful purification and RNA quality or stored at -80ºC.

A sample of RNA from extractions was analyzed for quality by gel electrophoresis. Along with a ladder displaying 40 nt and 20 nt lengths, total RNA extractions (the control) were run alongside RNAs resulting from the bead extraction method to compare the purification of the sample. Promising results were then further analyzed for RNA quality. The bead treated samples, along with a ladder and control, were diluted then analyzed with an Agilent Bioanalyzer. RNA quality was determined by observations of peaks located within the small RNA length range (20-40 nt).

Computational Analysis

Following RNA extractions, small RNA libraries were generated from the isolated small RNAs sent for sequencing on the Illumina Next-Generation platform. Computational analysis was completed using Bowtie, Samtools, and Bedtools. In-house scripts were written and used to analyze small RNAs in the sequence libraries (Figure 5)[31]. First, the genome for *C. virginica* (NC_035780.1) was indexed using Bowtie build. Known miRNA reads were then also indexed for later analysis of this small RNA class. The sequencing data sets received from Illumina were pre-processed to remove adapter sequences. For each set of data (gills, gonads, and control), four clipped fastq files were converted into one fq file. For each merged fq, read sizes were analyzed by using an in-house awk command[31]. These categorized read counts were then normalized to the total library size for each library. Counts representing normalized reads were plotted on a line graph. Next, the reads were mapped to indexed miRNAs obtained from a related species on miRbase.org. The unmapped reads were mapped to the genome to discern which reads had at least one alignment. Further distinctions between reads were then made based on

overlap signatures. siRNAs were identified by looking for 20-23 nt long reads that overlap perfectly and possess a 2 nt overhang on their 3' ends which are typical of Dicer processing. Also, piRNAs were identified by searching for 25-30 nt reads that overlap by 10 nt. This overlap signature is typical of ping pong processing. These were achieved by using an in-house script[31]. With the script, several loci believed to be regions of high Dicer and ping pong activity were identified.

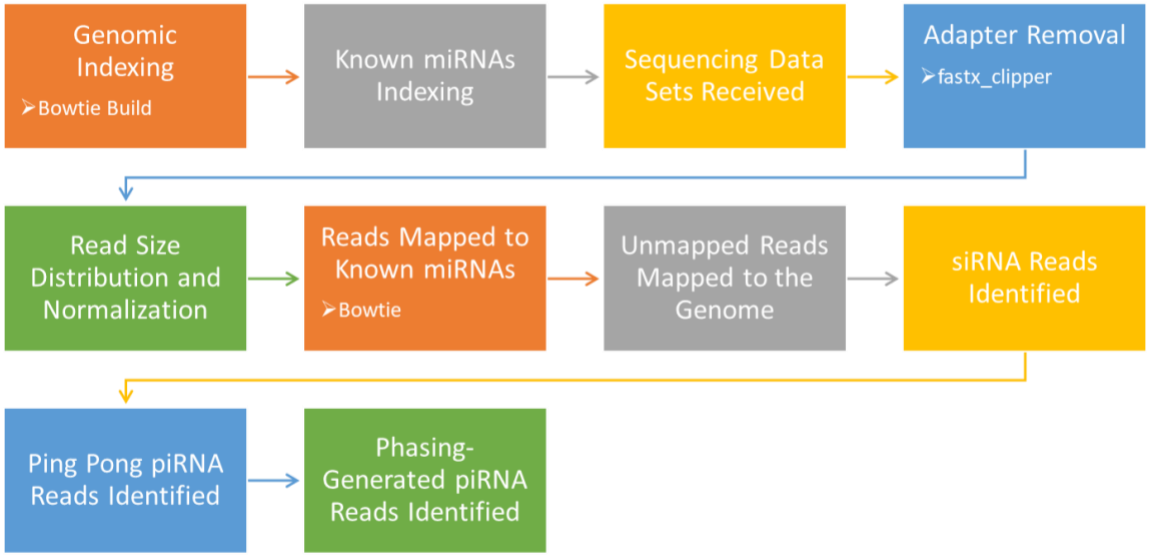


Figure 5: Flow chart of the computational analysis of extraction results using Samtools and Bedtools. The genome for *C. virginica* was indexed followed by indexing of known miRNAs. Sequencing data received from Illumina were recovered and processed to remove adaptors. For each data set (gonads, gills, and control), fastq files were generated and analyzed based on read sizes. These reads were normalized to the total library size for each library. The reads were then mapped to known miRNA reads from a related species. Unmapped reads were then mapped to the genome to determine

percentage of at least one alignment. Characteristic overlap signatures were then used to locate Dicer and ping pong activity in each library.

CHAPTER III: RESULTS

Extractions

Anion exchange chromatography works by separating positively and negatively charged amino acids and was used in this study to isolate basic riboprotein complexes. Since Ago proteins are highly basic, these proteins bind to the Q Sepharose beads and are collected while other proteins are allowed to bypass the beads and are discarded. To optimize this process, bead amounts were varied to determine the ideal ratio of sample to beads. Gel electrophoresis provided an immediate analysis of the success of these purifications. As seen in Figure 6, lanes displaying fragments that were more selective than the non-bead treated total RNA extraction (control) were deemed clean enough samples to proceed with quality analysis. Purification was determined based on comparisons to the ladder used (a mixture of two primers 40 nt and 20 nt in length) and to the control containing extra RNA material. Most successful gels results involved samples from the gills and gonads with bead amounts ranging from 800 μL to 1000 μL of Sepharose beads with tissue mass between 1 g and 1.1 g.

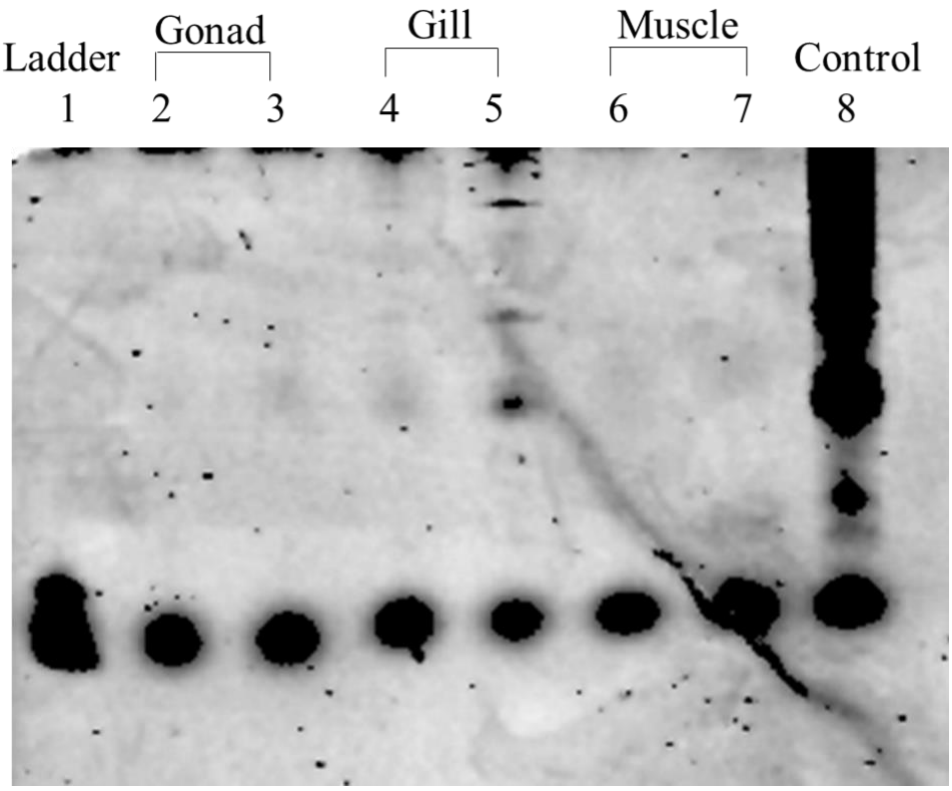


Figure 6: Polyacrylamide gel image showing Sepharose bead-purified RNAs vs non-Sepharose bead purified RNA extraction results. From left to right, Lane 1 is the ladder, and Lane 2 through 7 are RNAs extracted from Sepharose purified proteins. Lanes 2 and 3 are gonad extracted RNAs, Lanes 4 and 5 are gill extracted RNAs, Lanes 6 and 7 are muscle extracted RNAs, and Lane 8 (control) is total RNA extracted from gonad not subjected to Sepharose beads treatment. Lanes 4 and 5 display isolated RNAs.

A quality analysis of the extracted total RNA quality was completed using the Agilent Bioanalyzer. Bioanalyzer results involved a ladder compared against purified and non-purified samples with each containing a marker to align the samples (Figure 7). RNA quality was estimated by observing peaks within the 20-40 nt range, correlating to . Clear peaks in this range were deemed high enough in quality to proceed with sequencing for

further analysis. Results displaying more peaks occurring outside the 20-40 nt range than within the range were deemed lower in quality. These were not sequenced due to extra RNA material interfering with sequencing. The samples with the cleanest reads consisted of those RNAs dissected from the gills and gonads and then purified by bead treatment (Figure 7A and 7B).

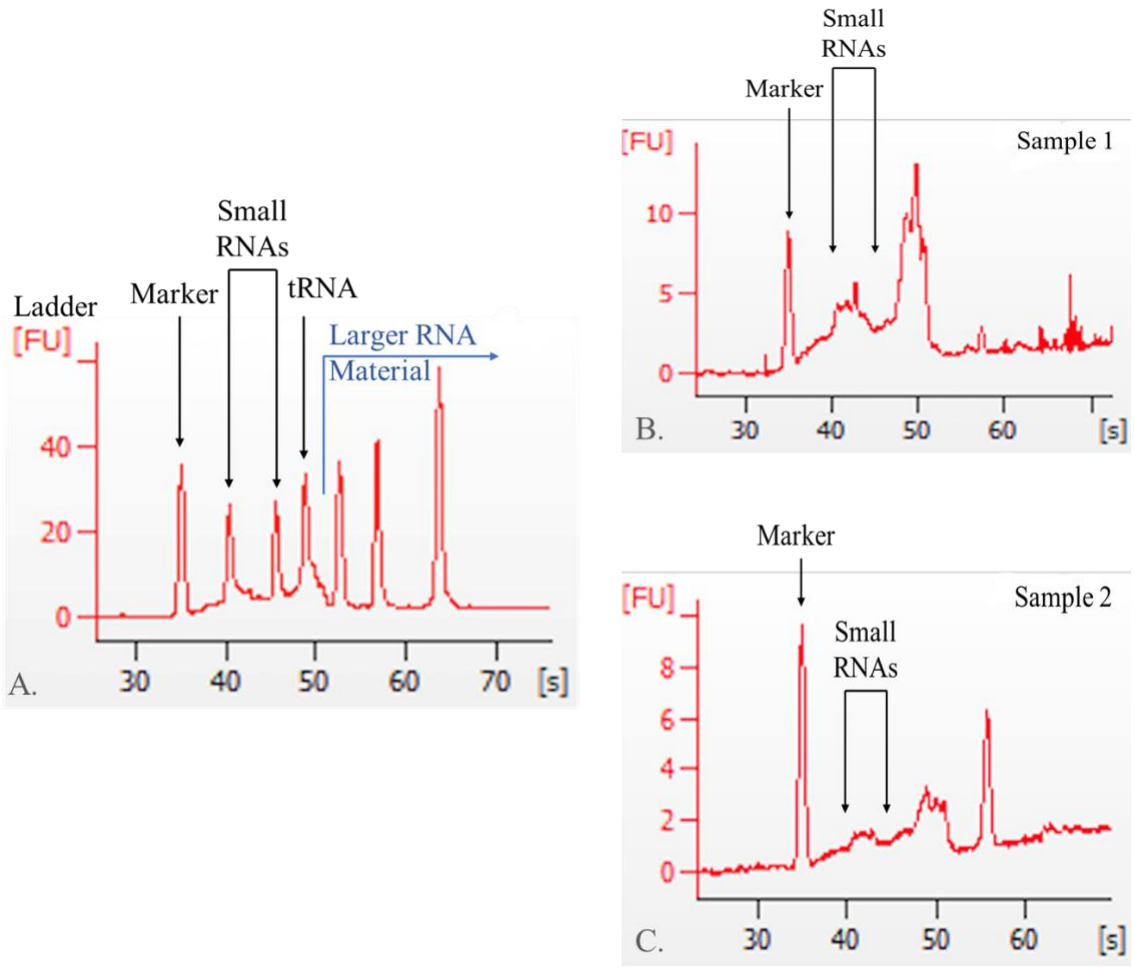


Figure 7: Bioanalyzer images to determine RNA quality. The x-axis represents migration time (s) while the y-axis represents fluorescent units (FU). **A** shows the ladder used, which includes the marker and peaks for read length identification. Desired small

RNAs are located between 40 s and 45 s. **B** shows a sample from gills that was treated with 800 μL of beads. **C** shows a sample from gills that was treated with 1000 μL of beads.

Small RNA Library Analysis

After receiving the libraries from the Illumina sequencing platform, the reads were recovered and analyzed in order to understand the effectiveness of the method employed. The reads in each of the libraries were first categorized and displayed by read sizes on a line graph (Figure 8). As expected, two major peaks corresponding to miRNA/siRNA sized reads and piRNA sized reads were observed. The gonad library contained about 500,000 reads per million (RPM) in the small RNA region, which was unusual situation since we expected to see a higher number of reads in the piRNA region. Also, about 200,000 RPM were recovered from the gill library with very few somatic piRNAs. In contrast to the Sepharose bead-treated libraries, the control library (gonad), which was not subjected to a Sepharose bead treatment, appeared to retain the majority of its piRNAs (about 280,000 RPM (Figure 8). It was also observed that the gill and the gonad Sepharose bead treated libraries contained nearly the same amounts of piRNA sized reads.

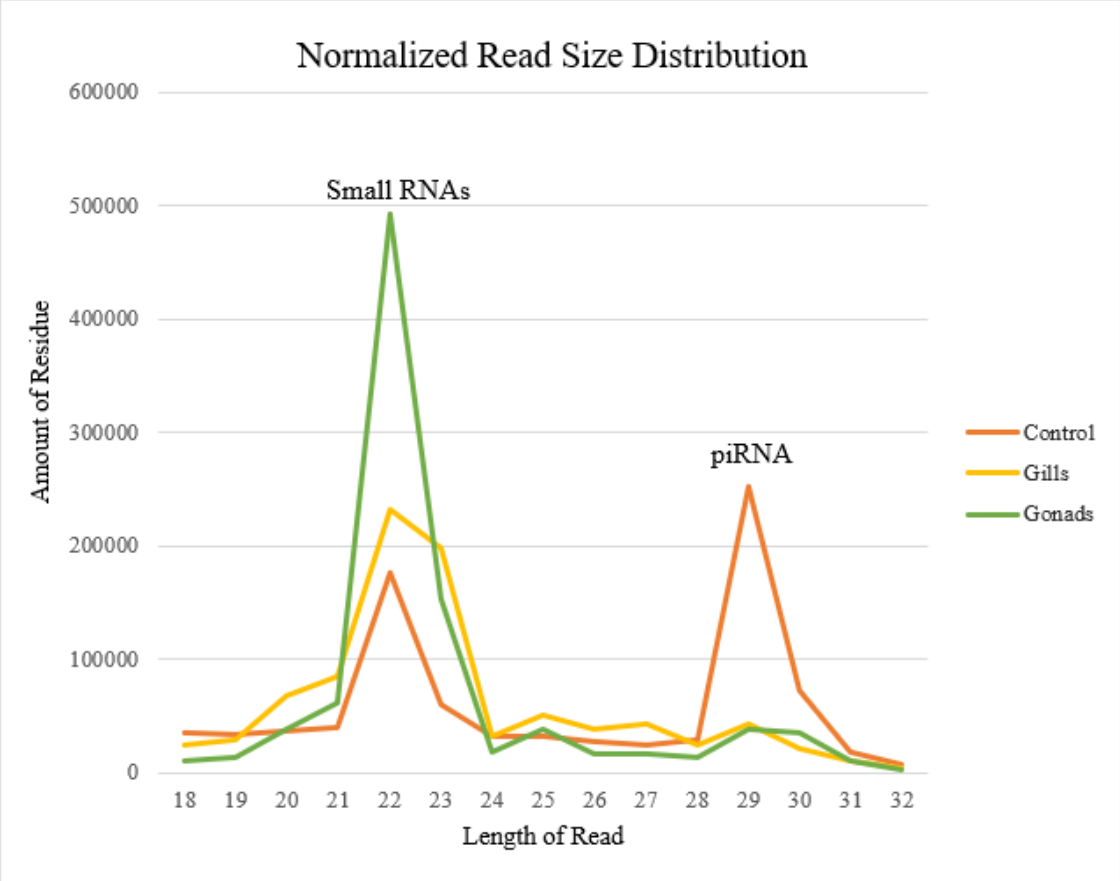


Figure 8: Line graph showing comparisons of read size distribution between libraries. In this figure, the gonads (green) and gills (yellow) represent libraries from Sepharose bead treatment extracted total RNAs, while the control (orange) represents non-Sepharose bead treated extraction (from the gonads). The y-axis represents normalized read count amounts, and the x-axis represents the read lengths.

Reads in each library were counted to estimate the number of reads recovered by the method employed here. For the control data, a total of 170,980,664 reads were identified. The total number of reads were 42,745,166 reads in the control library, 46,279,224 in the gill library and 47,105,592 in the gonad library.

The data sets were then mapped against the genome using a short read aligner (bowtie version one) (Figure 9). It was found that, among the reads mapped to the genome for the control data, 14,050,840 reads were processed in the control library and 97.38% of those reads had at least one alignment to the genome. Of the 27,887,075 reads processed in the gills library, 98.56% had at least one alignment. For the gonads library, of the 35,337,379 reads processed, 98.85% had at least one alignment (Figure 9A). Within each library, the loci were also mapped according to Dicer and ping pong cycle identifiers. For the control library, 34 loci were associated with Dicer (Figure 9B) and 16,080 loci were associated with the ping pong cycle (Figure 9C). For the gills library, 32 loci associated with Dicer (Figure 9B) and 5,160 associated with ping pong (Figure 9C). Lastly, for the gonads library, 8 loci were Dicer associated (Figure 9B) and 11,643 were ping pong associated (Figure 9C).

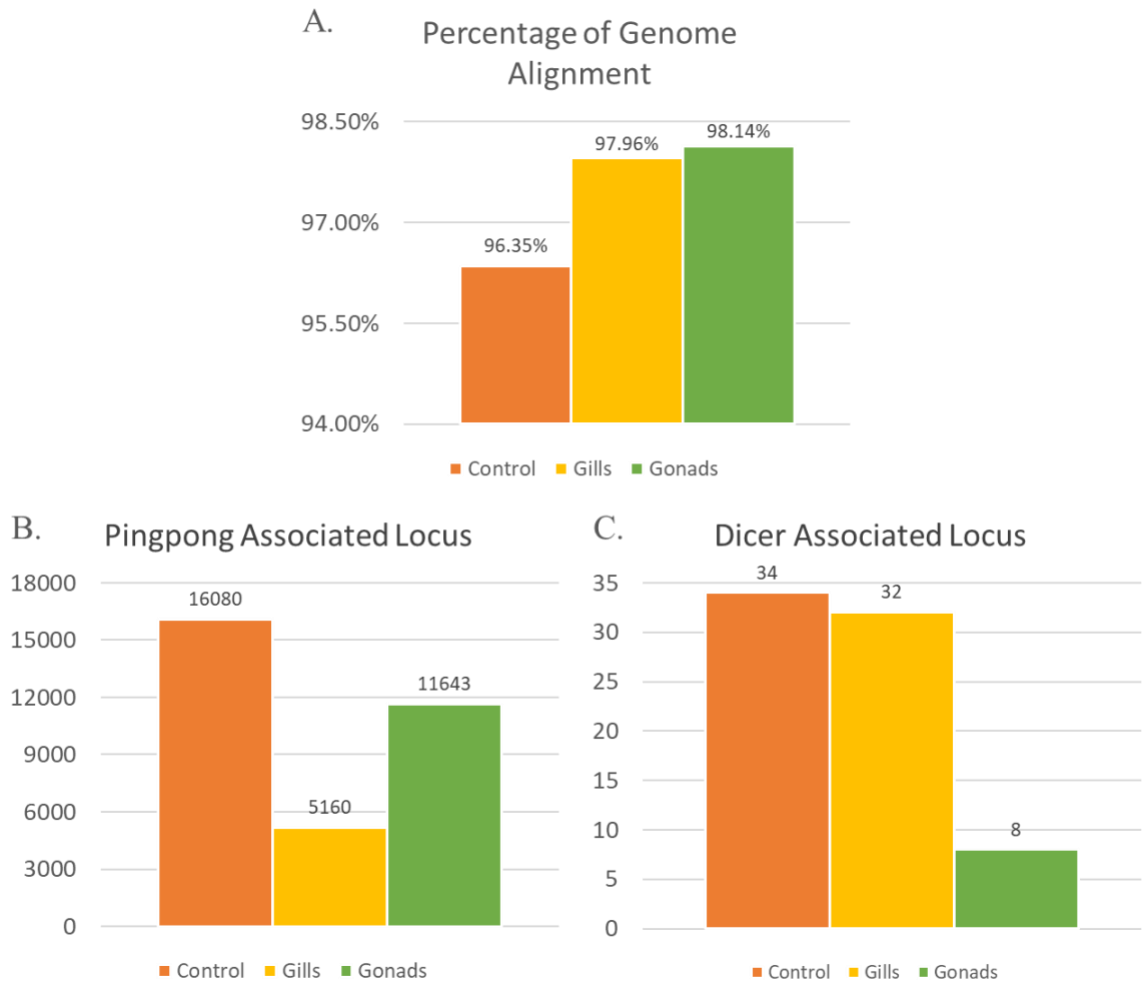


Figure 9: Bar graphs describing reads recovered from sequencing that were computationally analyzed. In this figure, the gonads (green) and gills (yellow) represent small RNA libraries from Sepharose bead treatment extractions while the control (orange) represents non-Sepharose bead treated extractions. **(A)** The chart shows the relationship between each library's alignment to the genome. **(B)** The chart shows the number of loci each library possessed that associated with the ping pong cycle (piRNAs). **(C)** The chart shows the number of loci each library possessed that associated with Dicer (siRNAs and miRNAs).

The identified loci in each library corresponding to Dicer and ping pong activities were then further sorted to obtain loci with the best number of hits within the genome. The top 100 loci were plotted into the bar graphs (Figure 10). In the control library, the highest count of siRNA associated loci with best hits in the genome were about 13 loci, and the highest count of ping pong loci based on loci length were 3400. For the gills library, the highest count of siRNA loci was about 10 and the highest count of ping pong loci based on loci length were 1125. Lastly, in the gonads, the highest count of siRNA loci was 55, and the highest count of ping pong loci based on loci length were 360. Overall, the gonads had higher counts of piRNA loci.

Figure 10: Bar graphs displaying the top 100 hits for Dicer and top 100 hits for ping pong associated loci sizes in the control, gills, and gonads. The y-axis represents the number of hits (nt) for Dicer activity and locus size for ping pong activity while the x-axis denotes the locus.

CHAPTER IV: DISCUSSION

Using the method described in this study, an investigation of ncRNAs in the non-model organism *C. virginica* was completed using biochemical extractions and computational analysis. A technique involving protein extractions with Q Sepharose beads was performed to establish a method for purifying Argonaute-associated proteins from tissue samples. The RNA extractions that preceded this allowed for the analysis of the method's effectiveness. As can be seen in Figure 6, polyacrylamide gels were used for the immediate analysis of the extractions. The observations of the gels showed bead treated samples to be cleaner than the control, which was not treated with beads. The lanes representing bead treated samples showed fewer RNA fragments outside the target nt range in the polyacrylamide gel. This indicated that the Sepharose beads were preventing non-specific materials unassociated with Ago proteins from being bound to them. Artifacts observed in the control group on a polyacrylamide gel confirmed that the bead treatments were more effectively purifying the tissue samples. These results were followed by favorable RNA quality tests via a Bioanalyzer. As shown in Figure 7A and 7B, samples treated with the beads displayed peaks along the 20 nt to 40 nt range. Given that miRNA and siRNA are about 22 nt long and piRNA are 25 nt to 30 nt long, these peaks indicated all three classes were present in the bead treated samples. While the samples were not completely pure, the presence of small RNAs suggested that the bead treated samples, following RNA extractions, were more selective for the Ago proteins associated with these small RNAs than total RNA extractions. With this, it can be said that the beads worked to provide cleaner samples of proteins from tissues and to select for ncRNA-associated proteins, specifically.

When analyzing read size distributions for each of the three RNA libraries, two distinct peaks were seen in the range of 21 to 23 nt (siRNA and miRNA) for all libraries and 28-31 nt (piRNA) for the control library. The gonad library did not show a high enough number of reads corresponding with piRNA. This contradicted previous knowledge of piRNA activity and the control, given that the control samples were also taken from gonad tissue and displayed piRNA reads. This lack of piRNA reads in the gonads indicates that the results of the bead extractions were not a true reflection of the piRNA populations. The presence of reads between 21 and 23 nt long in each library, however, do show that bead treatments did not discriminate against proteins associated with siRNA and miRNA. Because of these mixed results, it can be said the protein extraction did succeed in selecting for proteins associated with some populations of small RNAs. Given that siRNA and miRNA activities are linked with the AGO clade in the Argonaute protein family and that the clade interacting with piRNA is PIWI, it was determined that the protein extractions selected for AGO but not PIWI.

Furthermore, a 97% mapping rate to the genome in all libraries is a reflection of the quality of the sequencing platform used. Since no genome is entirely complete and artifacts such as transcripts from microbial activity can contaminate the library, the high rate of alignment for each library was a favorable result. Of the reads processed, loci associating with both Dicer and ping pong through overlap characteristics were identified to determine both Dicer and ping pong activity. The high levels of ping pong activity, as recovered from top 100 loci in the gonad and control when compared to the gills, is an indication of the important role of piRNAs in germline genome integrity in this organism. The presence of some piRNA loci in the gills region also suggests that piRNAs in the

eastern oyster play a role in somatic transposon regulation or possible control of protein-coding genes. These findings signify a need for further exploration of piRNA pathways in this species. Likewise, gill tissues showed higher levels of Dicer activity, indicating the presence of more siRNA reads in somatic tissues. Since siRNAs participate in gene silencing in somatic cells, this is consistent with previous research[10,30]. The presence of Dicer and ping pong activity in somatic and germline cells is a reflection of the complex RNAi factors in the species. As previously shown in Figure 1B, Bivalves possess a unique combination of RNAi factors compared to other members of the clade. While this study's focus remained on optimizing a protein extraction protocol, the computational analysis of the extraction results does reflect RNAi complexity in a *Crassostrea* species, which warrants further investigation.

As stated, Sepharose beads used in protein purification provided mixed results. Though the gel and bioanalyzer observations showed success in purification, the results of the read size distribution were deemed, overall, an inaccurate representation of small RNAs populations in *C. virginica*. However, alterations to the protein extractions could improve the effectiveness of the method. With the success of anion exchange chromatography in other non-model organisms, such as spider mites[14], the inherent differences between oysters and these organisms likely played a role in the negative result. Salinity in oysters is higher than previously documented species and is considered above what is typically appropriate for this form of protein isolation. This aspect of oysters could cause interference with the beads' activity during purification. It has been shown that high pH ranges and high salinity can cause issues with anion-exchange chromatography. One solution to this issue is to dilute out the salts in the tissues in order

to prevent chromatography interference[32]. Given this, an improvement to the Sepharose bead extraction would be to modify the binding buffer used. Revisions, such as adding more DI water, should be considered in future studies involving small RNA studies in saltwater organisms. However, it is unclear if a decrease in salt concentration would lead to denaturation of small RNA-Argonaute protein complexes. Future studies adapting the reported protocol are encouraged to consider these factors when utilizing Sepharose bead extractions in saltwater organisms.

Nonetheless, this study provided a framework for a prospective method for Ago protein extraction that can be used for further studies into the Ago family and associated small RNAs in this and other species. In addition, the computational work performed in this study provided small RNA libraries for both somatic and germline regions of the non-model organism *C. virginica*. These libraries can be used in future RNAi studies concerning this and other *Crassostrea* species. In this study, the libraries were used for characterization of small RNAs based on size, overlaps, and sequencing, which did show distinct small RNA. The apparent complexity of RNAi pathways in this species was supported here by the small RNAs present in the oysters and the protein activity analyzed in somatic and germline libraries. With the prospective methodology reported in this study and the libraries generated, further intrigue into the unique RNAi factors present in *Crassostrea* species can be conducted. This knowledge will promote a better understanding of the roles of small RNAs and RNAi factors in *C. virginica* and other species in its clade.

APPENDIX A: IRB APPROVAL LETTER

THE UNIVERSITY OF
SOUTHERN MISSISSIPPI

NOTICE OF INSTITUTIONAL REVIEW BOARD ACTION

The project below has been reviewed by The University of Southern Mississippi Institutional Review Board in accordance with Federal Drug Administration regulations (21 CFR 26, 111), Department of Health and Human Services regulations (45 CFR Part 46), and University Policy to ensure:

- The risks to subjects are minimized and reasonable in relation to the anticipated benefits.
- The selection of subjects is equitable.
- Informed consent is adequate and appropriately documented.
- Where appropriate, the research plan makes adequate provisions for monitoring the data collected to ensure the safety of the subjects.
- Where appropriate, there are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of all data.
- Appropriate additional safeguards have been included to protect vulnerable subjects.
- Any unanticipated, serious, or continuing problems encountered involving risks to subjects must be reported immediately. Problems should be reported to ORI via the Incident template on Cayuse IRB.
- The period of approval is twelve months. An application for renewal must be submitted for projects exceeding twelve months.

PROTOCOL NUMBER: 20-1000

SCHOOL/PROGRAM: School of Professional Nursing Practice
RESEARCHER(S): Seymour Eagle, Harvey Golden

IRB COMMITTEE ACTION: Approved
CATEGORY: Expedited (the category listed below is just a sample of one, there are several categories that the protocol could be assigned)

7. Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and
social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human
factors evaluation, or quality assurance methodologies.
PERIOD OF APPROVAL: 10.27.2020 – 10.27.2021

Donald Sacco Jr.

Donald Sacco, Ph.D.
Institutional Review Board Chairperson

REFERENCES

1) Peragine, A., Yoshikawa, M., Wu, G., Albrecht, H. L., & Poethig, R. S. (2004). SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes & development*, *18*(19), 2368-2379.

2) Knauer, S., Holt, A. L., Rubio-Somoza, I., Tucker, E. J., Hinze, A., Pisch, M., Javelle, M., Timmermans, M. C., Tucker, M. R., & Laux, T. (2013). A protodermal miR394 signal defines a region of stem cell competence in the Arabidopsis shoot meristem. *Developmental cell*, *24*(2), 125-132.

3) Tolia, N. H., & Joshua-Tor, L. (2007). Slicer and the argonautes. *Nature chemical biology*, *3*(1), 36-43.

4) Wang, X. H., Aliyari, R., Li, W. X., Li, H. W., Kim, K., Carthew, R., Atkinson, P., & Ding, S. W. (2006). RNA interference directs innate immunity against viruses in adult Drosophila. *Science (New York, N.Y.)*, *312*(5772), 452–454.

5) Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, *128*(6), 1089-1103.

6) Lee, R. C., Hammell, C. M., & Ambros, V. (2006). Interacting endogenous and exogenous RNAi pathways in Caenorhabditis elegans. *RNA*, *12*(4), 589-597.

7) Pinzón, N., Bertrand, S., Subirana, L., Busseau, I., Escrivá, H., & Seitz, H. (2019). Functional lability of RNA-dependent RNA polymerases in animals. *PLoS genetics*, *15*(2), e1007915.

8) Lee, Y. S., Nakahara, K., Pham, J. W., Kim, K., He, Z., Sontheimer, E. J., & Carthew, R. W. (2004). Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*, *117*(1), 69-81.

9) Catalanotto, C., Cogoni, C., & Zardo, G. (2016). MicroRNA in control of gene expression: an overview of nuclear functions. *International journal of molecular sciences*, *17*(10), 1712.

10) Jehn, J., Gebert, D., Pipilescu, F., Stern, S., Kiefer, J. S. T., Hewel, C., & Rosenkranz, D. (2018). PIWI genes and piRNAs are ubiquitously expressed in mollusks and show patterns of lineage-specific adaptation. *Communications biology*, *1*(1), 1-11.

11) Olina, A. V., Kulbachinskiy, A. V., Aravin, A. A., & Esyunina, D. M. (2018). Argonaute proteins and mechanisms of RNA interference in eukaryotes and prokaryotes. *Biochemistry (Moscow)*, *83*(5), 483-497.

12) Song, J. J., Smith, S. K., Hannon, G. J., & Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science*, *305*(5689), 1434-1437.

13) Rivas, F. V., Tolia, N. H., Song, J. J., Aragon, J. P., Liu, J., Hannon, G. J., & Joshua-Tor, L. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. *Nature structural & molecular biology*, *12*(4), 340-349.

14) Mondal, M., Peter, J., Scarbrough, O., & Flynt, A. (2021). Environmental RNAi pathways in the two-spotted spider mite. *BMC genomics*, *22*(1), 1-11.

15) Yang, J., Cho, W. C., & Zheng, Y. (2020). Argonaute proteins: Structural features, functions and emerging roles. *Journal of Advanced Research*.

16) Lisitskaya, L., Aravin, A. A., & Kulbachinskiy, A. (2018). DNA interference and beyond: structure and functions of prokaryotic Argonaute proteins. *Nature communications*, *9*(1), 1-12.

17) Medina, C., da Rocha, M., Magliano, M., Raptopoulo, A., Marteu, N., Lebrigand, K., ... & Jaubert-Possamai, S. (2018). Characterization of siRNAs clusters in Arabidopsis thaliana galls induced by the root-knot nematode Meloidogyne incognita. *BMC genomics*, *19*(1), 1-16.

18) Soreq, L., Salomonis, N., Bronstein, M., Greenberg, D. S., Israel, Z., Bergman, H., & Soreq, H. (2013). Small RNA sequencing-microarray analyses in Parkinson leukocytes reveal deep brain stimulation-induced splicing changes that classify brain region transcriptomes. *Frontiers in molecular neuroscience*, *6*, 10.

19) Rogers, K., & Chen, X. (2013). Biogenesis, turnover, and mode of action of plant microRNAs. *The Plant Cell*, *25*(7), 2383-2399.

20) Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. C*ell*, *116*(2), 281-297.

21) Li, X., Jin, F., Jin, L., Jackson, A., Ma, X., Shu, X., ... & Jin, G. (2015). Characterization and comparative profiling of the small RNA transcriptomes in two phases of flowering in Cymbidium ensifolium. *BMC genomics*, *16*(1), 1-17.

22) Brodersen, P., & Voinnet, O. (2006). The diversity of RNA silencing pathways in plants. *TRENDS in Genetics*, *22*(5), 268-280.

23) Bologna, N. G., & Voinnet, O. (2014). The diversity, biogenesis, and activities of endogenous silencing small RNAs in Arabidopsis. *Annual review of plant biology*, *65*, 473-503.

24) Aravin, A. A., Hannon, G. J., & Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, *318*(5851), 761-764.

25) Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, *127*(6), 1193-1207.

26) Han, Y., He, F., Chen, Y., Liu, Y., & Yu, H. (2018). SiRNA silencing efficacy prediction based on a deep architecture. *BMC genomics*, *19*(7), 59-65.

27) Mohn, F., Handler, D., & Brennecke, J. (2015). piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. *Science*, *348*(6236), 812-817.

28) Hyun, S. (2017). Small RNA pathways that protect the somatic genome. *International journal of molecular sciences*, *18*(5), 912.

29) Roovers, E. F., Rosenkranz, D., Mahdipour, M., Han, C. T., He, N., de Sousa Lopes, S. M. C., & Ketting, R. F. (2015). Piwi proteins and piRNAs in mammalian oocytes and early embryos. *Cell reports*, *10*(12), 2069-2082.

30) Waldron, F. M., Stone, G. N., & Obbard, D. J. (2018). Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes. *PLoS genetics*, *14*(7), e1007533.

31) Peter, J.O., Santos-Ortega, Y., & Flynt, A. Guiding RNAi Through Characterization of Endogenous Small RNA Pathways. Book chapter in RNAi strategies for Pest Management of the MiMB series. 2021.

32) Guignard, C., Jouve, L., Bogéat-Triboulot, M. B., Dreyer, E., Hausman, J. F., & Hoffmann, L. (2005). Analysis of carbohydrates in plants by high-performance anion-exchange chromatography coupled with electrospray mass spectrometry. *Journal of Chromatography A*, *1085*(1), 137-142.