2014

SMALL RNA EXPRESSION DURING PROGRAMMED REARRAGEMENT OF A VERTEBRATE GENOME

Joseph R. Herdy III
University of Kentucky, jrhe238@g.uky.edu

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Joseph R. Herdy III, Student

Dr. Randal Voss, Major Professor

Dr. David Westneat, Director of Graduate Studies

SMALL RNA EXPRESSION DURING PROGRAMMED REARRAGEMENT OF A VERTEBRATE GENOME

THESIS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the College of Arts and Sciences at the University of Kentucky

By

Joseph Roger Herdy III

Lexington, Kentucky

Director: Jeramiah Smith, Professor of Biology

Lexington, Kentucky

2014

ABSTRACT OF THESIS

SMALL RNA EXPRESSION DURING PROGRAMMED REARRAGEMENT OF A VERTEBRATE GENOME

The sea lamprey (*Petromyzon marinus*) undergoes programmed genome rearrangements (PGRs) during embryogenesis that results in the deletion of ~0.5 Gb of germline DNA from the somatic lineage. The underlying mechanism of these rearrangements remains largely unknown. miRNAs (microRNAs) and piRNAs (PIWI interacting RNAs) are two classes of small noncoding RNAs that play important roles in early vertebrate development, including differentiation of cell lineages, modulation of signaling pathways, and clearing of maternal transcripts. Here, I utilized next generation sequencing to determine the temporal expression of miRNAs, piRNAs, and other small noncoding RNAs during the first five days of lamprey embryogenesis, a time series that spans the 24-32 cell stage to the formation of the neural crest. I obtained expression patterns for thousands of miRNA and piRNA species. These studies identified several thousand small RNAs that are expressed immediately before, during, and immediately after PGR. Significant sequence variation was observed at the 3' end of miRNAs, representing template-independent covalent modifications. Patterns observed in lamprey are consistent with expectations that the addition of adenosine and uracil residues plays a role in regulation of miRNA stability during the maternal-zygotic transition. We also identified a conserved motif present in sequences without any known annotation that is expressed exclusively during PGR. This motif is similar to binding motifs of known DNA binding and nuclear export factors, and our data could represent a novel class of small noncoding RNAs operating in lamprey.

KEYWORDS: Small RNAs, PGR, Lamprey, Genomics, Bioinformatics

Joseph Roger Herdy III

August 29, 2014

SMALL RNAS IN PROGRAMMED REARRAGEMENT OF A VERTEBRATE GENOME

By

Joseph Roger Herdy III

Dr. Randal Voss
Director of Thesis

Dr. David Westneat
Director of Graduate Studies

August 29, 2014
Date

TABLE OF CONTENTS

ACKNOWLEDGEMENT

I would like to express the deepest appreciation and thanks to my adviser, Dr. Jeramiah Smith, for his assistance and guidance in the completion of these studies and preparation of this thesis. Jeramiah bestowed unto me a robust method for approaching scientific inquiry that I will draw upon for the rest of my career.

I would also like to thank my committee members Dr. Ann Morris, Dr. Randal Voss, and Dr. Bruce O'Hara for their support and direction while completing my thesis.

In addition, a thank you to Stephanie Bryant and Melissa Keinath: two of the best lab mates a graduate student could ask for, as well as all of my friends and family for their encouragement through this process.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Programmed Genome Rearrangements in Lamprey

The sea lamprey (*Petromyzon marinus*) undergoes developmentally programmed genome rearrangements (PGR) that mediate the loss of 20% (~0.5 Gb) of germline DNA from somatic cells during early embryogenesis. These deletions are reproducible across individuals, and occur 2.5 days post fertilization (corresponding to the mid blastula transition)[1,58]. The rearrangement event occurs when embryos are comprised of thousands of cells, indicating that a coordinated and complex mechanism is required to mediate these eliminations. Further inquiry into the deleted fraction has shown it to be enriched with genes that mediate cell proliferation, oncogenesis, and pluripotency which if misexpressed have the potential to contribute to the development of diseases such as cancer[1]. Thus it is hypothesized that PGR in lamprey is acting as a biological strategy to prevent deleterious misexpression of germline-limited genes. The presence of predictable and extensive reorganizations in lamprey provides the unique opportunity to study the differential genomic requirements of fated somatic and pluripotent germ cells. Owing to the complexity and tight regulation of this event across a large number of cells, I hypothesize that several large scale global regulatory complexes must be acting in tandem to mediate these rearrangements.

The jawless vertebrates (e.g. lampreys and hagfish) are important comparative models for developmental biology, as they represent an ancient offshoot from the vertebrate lineage and can provide insight into the ancestral state of the vertebrate genome. Although most vertebrates undergo several small

local rearrangements during development, such as remodeling of immune receptor loci[2], lamprey undergo large alterations that previously had not been observed in a vertebrate genome. Despite the drastic differences in genome biology, early lamprey development is largely conserved with that of higher vertebrates[61,62]. Although the pattern for loss across all somatic tissues is consistent, essentially nothing is known about the molecular mechanisms that control this programmed loss. Given our current understanding of similar rearrangements occurring in other systems (reviewed below) [3-5], I hypothesize that small RNAs play a role in mediating the fraction that is lost. Only a single study has been performed to characterize lamprey small RNAs[63], and to the best of our knowledge none have been performed on developing embryos. Here, I characterize the dynamics of small RNA expression in early developing lamprey embryos and attempt to correlate changes in the expression of small noncoding RNAs with the rearrangement event.

1.2 NGS and snRNA studies

The introduction of high throughput next generation sequencing (NGS) technologies has revolutionized transcriptomics by allowing RNA analysis through cDNA sequencing at massive scales (RNA-seq). Recently, application of Illumina sequencing by synthesis technology has been used in wide range of organisms to study small non-coding RNA (snRNA) expression profiles[6-8]. This approach has also resulted in the discovery of several new classes of snRNAs[6,20,32]. Direct sequencing also offers the potential to detect variation in mature miRNA length, as well as enzymatic modifications of miRNAs such as nucleotide additions, which provide further insight into the biology and regulation

of miRNAs[40]. With the added depth of sequencing now possible, we have the opportunity to identify low abundance snRNAs or those exhibiting modest expression differences between samples. Next generation snRNA profiling has already been realized in several organisms[3,10,18], and characterizing the small RNA state of the lamprey could help both in developing this organism as a genetic model of vertebrate development, and furthering our understanding of the epigenetic states and molecular pathways that underlie PGR.

1.3 miRNAs

miRNAs were first discovered from the observation in *Caenorhabditis elegans* that RNA from the *lin-4* gene binds to the 3' UTR of lin-14 mRNA and inhibits translation. Since then, thousands of miRNA families in a wide range of taxa have been identified[9] and this number is predicted to increase with the current expansion of sequencing based experiments[10]. miRNAs are short (~22 nt) endogenously transcribed non-coding RNAs that post-transcriptionally control mRNA expression via either disruption of translation or degradation of the target transcript. miRNAs are thought to regulate more than half of all mRNAs in animals[11] and have tightly regulated temporal and spatial expression patterns[12]. The regulatory activities of miRNAs are diverse and critical in controlling processes ranging from apoptosis (miR-25)[56] to glucose homeostasis (miR-103)[64]. Likewise, miRNAs themselves and the cellular machinery that produce them are modulated by a wide variety of effectors[13].

Biogenesis of miRNAs has been reviewed extensively elsewhere[14], but will be summarized briefly here (Fig 1). Genes encoding miRNAs are transcribed into long, primary transcripts (pri-miRNAs) by RNA polymerase II. These pri-miRNAs fold into secondary "hairpin loop" structures which are recognized by the RNAse III enzyme Drosha and cleaved to produce precursor miRNAs (pre-miRNAs). Pre-miRNAs are then transported out of the nucleus, and are processed by the RNAse III enzyme Dicer to yield a mature strand that targets mRNAs and a passenger strand that is normally degraded. The mature strand is loaded into a member of the Argonaut family of proteins within the RNA induced silencing complex (RISC). Through RISC, the mature miRNA acts as a guide strand and confers target specificity using the 2-7 nt "seed" strand and mediates the repression of targets. Because of their functional importance in target binding, miRNAs are grouped into families based on their seed sequences.

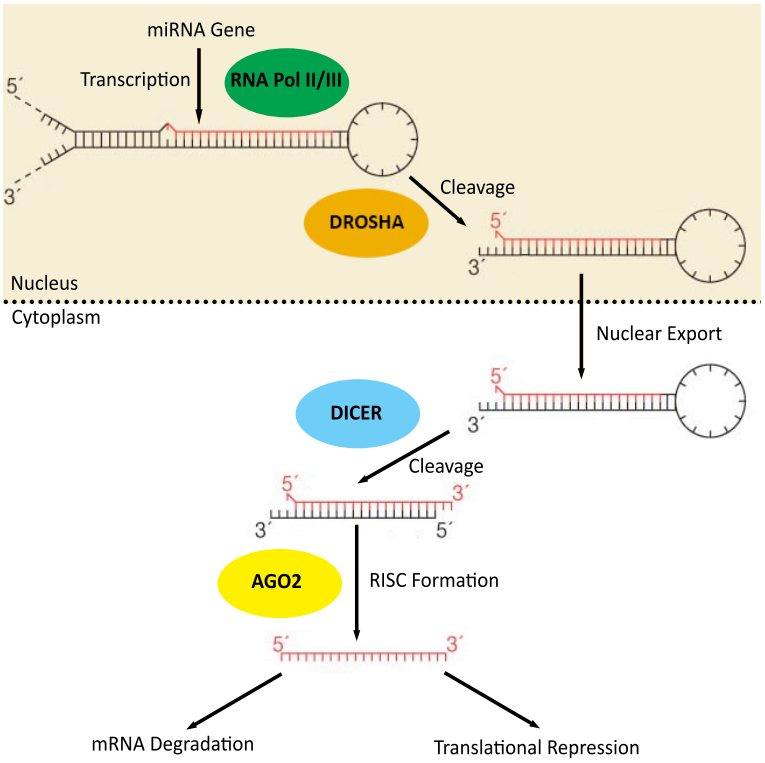

Fig 1: The miRNA biogenesis pathway. miRNAs are transcribed by RNA polymerase II as long primary transcripts (pri-miRNA). These transcripts are then processed in the nucleus by the RNase III enzyme Drosha into pre-miRNAs, which are exported into the cytoplasm by Exportin 5. Pre-miRNAs are processed by Dicer into a duplex consisting of a mature strand (miRNA) and passenger star strand (miRNA*). The mature miRNA is loaded into the RISC and acts as a guide that recognizes target mRNAs based on sequence complementarity. The RISC then represses targets by inhibiting translation or promoting degradation of target mRNAs.

Previous work has shown that miRNAs are vital for cell specification and differentiation in vertebrates[15-17]. It has been documented that different tissues and developmental stages are characterized by their own unique miRNA expression patterns[18]. I reasoned that characterizing the expression patterns of small RNAs through early development of lamprey embryos will give us insight into the roles of small RNA regulation on the development of a basal vertebrate, and might reveal a correlation with PGR.

1.4 miRNAs in PGR

Previous work has shown that miRNAs play crucial roles during vertebrate cell specification including clearing of maternal transcripts[42], modulation of embryonic signaling pathways[43] and promotion of cell differentiation, including germ layer formation and maintenance of pluripotency[44,75]. Owing to the fact that previous analyses[1,65] of the germline limited fraction in lamprey has shown an overrepresentation of gene ontologies associated with pluripotency factors, we hypothesize that miRNAs that contribute to germline pluripotency in lamprey might also be targeted for deletion in a manner similar to coding regions. Studies in human have shown that miRNA misexpression is a hallmark for many types of cancers[20], and it has been proposed that the absence of these genes in soma could serve as an anti-oncogenic strategy. Additionally, as PGR is a complex event requiring precise targeting of sequences for elimination and coordinated regulation across thousands of cells, it seems likely that miRNAs are involved directly or indirectly in mediating the activity of the factors controlling these rearrangements.

1.5 piRNAs

PIWI-interacting RNAs (piRNAs) are another subclass of snRNAs, named for their association with the Argonaute protein PIWI. piRNAs differ from other snRNAs in that they are typically longer (26-32 nt), and are generated in a RNAse III Dicer independent manner[24]. Similar to miRNAs, piRNAs have deep ancestry in Bilateria, and are expressed in a variety of totipotent or pluripotent cell populations to aid in defending genomic integrity from transposable elements[36]. The majority of piRNAs (85-95%) arise from intergeneic repetitive clusters in the genome called piRNA clusters[21]. piRNA clusters can span several hundred kb, and encompass many transposable DNA elements and their remnants. Most piRNAs are antisense to their target transposable element, and can induce silencing by hybridizing with them via their PIWI protein counterparts. Transcripts arising from piRNA clusters are processed into piRNA-like molecules and loaded in PIWI proteins. The factors mediating primary processing of piRNAs remain largely unknown[22]. Additionally, only select PIWI proteins associate with primary piRNA transcripts, and the mechanism underlying the selection of these proteins remains a mystery[22].

Primary piRNAs are then subject to a "Ping Pong" signal amplification cycle in germline to enforce high cellular copy number (Fig 2). In this system, Aubergine or PIWI-Like proteins associate with a primary piRNA transcript and cleave the target RNA via their Slicer activity[23]. This process produces the 5' end of secondary piRNAs, which are then transferred to other members of the PIWI

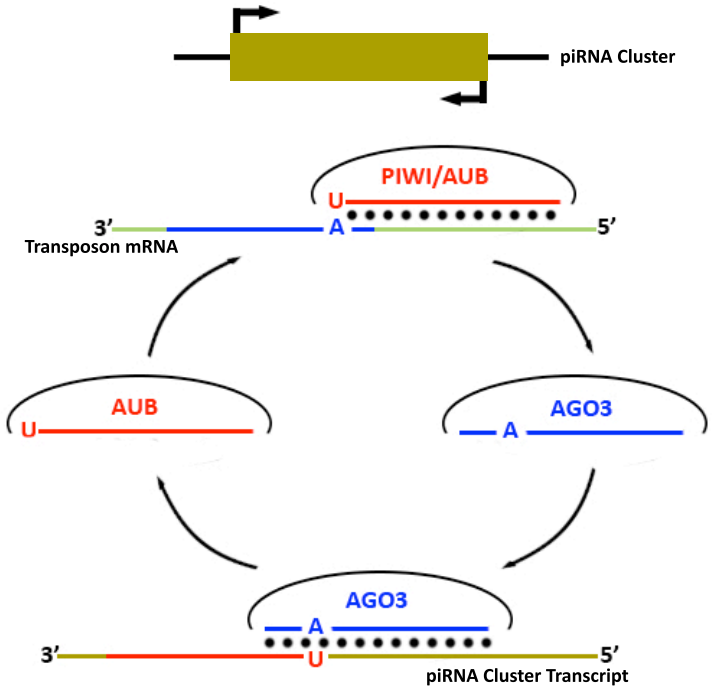

Fig 2: piRNA biogenesis and function. First, primary processing generates primary piRNAs in a Dicer independent manner. These piRNAs then guide PIWI to antisense transcripts from the same piRNA cluster. PIWI proteins use their slicer activity to cleave the target transcript and generate a new 5' end. This 5' end is bound by another PIWI protein. The 3' end of this transcript is trimmed to the length of the mature piRNA, leading a mature secondary piRNA capable of targeting sense transcripts transcribed from the piRNA cluster.

family to be trimmed at the 3' end and give rise to new mature piRNAs. piRNAs produced via the Ping Pong amplification cycle show a bias for Uracil at the 5' end, and Adenosine at the 10th nucleotide from the 5' end[23], which is highly conserved across animal species[21,23-25].

piRNAs are primarily expressed in the ovaries/testes, and knockdowns of PIWI in flies, zebra fish, and mouse causes widespread apoptosis of germ cells and sterility[22,25]. This suggests an essential role of the PIWI-piRNA pathway in maintaining germline integrity. Sequencing of the lamprey genome has revealed homologs of the PIWI and Aubergine proteins, which is unsurprising given their deep ancestry in metazoans[26]. piRNAs have also been implicated in *de novo* methylation of DNA and histone modification in *Drosophila,* zebrafish, and ciliates[27]. Moreover, piRNAs have been implicated as mediators of DNA methylation, histone modification, and PGR in ciliates (reviewed below). Prior to targeted sequencing of lamprey miRNAs, I performed two preliminary studies aimed at understanding the epigenetic state of eliminated DNA in lamprey embryos. First, I performed an analyses of methylation state during lamprey embryogenesis, which suggest the methylated DNA (5-Methylcytosine) may be specifically targeted for elimination during PGR (Fig 3). I have also gathered preliminary immunohistochemistry data which is suggestive of H3K9me3 (Histone H3 Lysine 9 trimethylation: a modification associated with gene silencing) bound DNA being expunged from the nuclei of sectioned rearranging lamprey embryos (Fig 4). Given the multiple complementary functions of piRNAs, I became particularly interested in understanding the dynamics of

piRNA expression during genomic rearrangements in lamprey to attempt to capture any major differential expression of these snRNAs during embryogenesis.

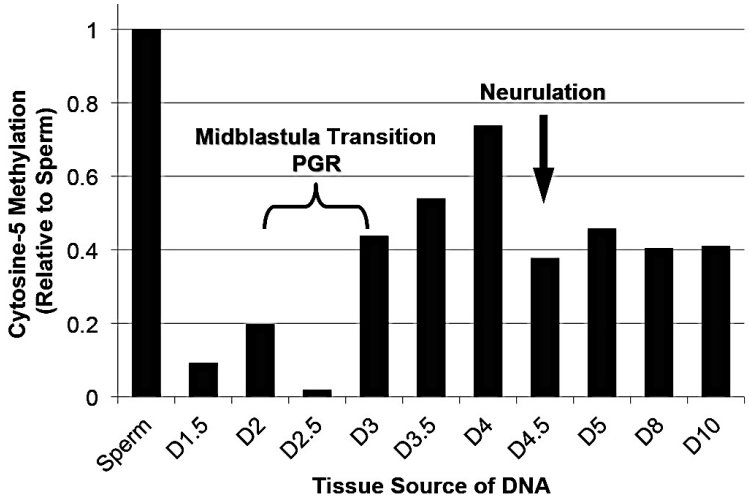

Fig 3: Profile of global 5 cytosine methylation in lamprey, spanning the course of programmed genome rearrangement. Assayed using the MethylFlash assay (Epigentek).

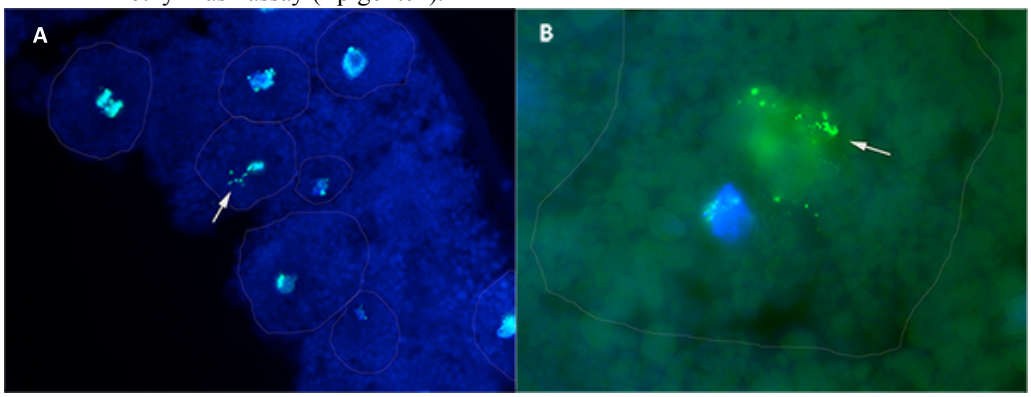

Fig 4: H3K9me3 Staining of Rearranging Lamprey Embryo Sections. PFA fixed and sectioned lamprey embryos (day 2.5 post fertilization) were probed with antibodies for H3K9me3 to observe patterns of repressive histone modification during PGR. Pictured are two examples of a common pattern of modified DNA (green) just outside of nuclei (DAPI). (A) Embryo section under 40X magnification. (B) Embryo section under 100X magnification.

1.6 Small RNAs and PGR in Ciliates

Evidence for programmed genome rearrangements in ciliates dates back to the 1970s[69], and much of the work done to understand PGR has been performed in these organisms. Several studies have demonstrated that small RNAs act as an epigenetic factor controlling programmed genome rearrangements in ciliates[33,69,47]. Ciliates are single celled organisms that exhibit nuclear

dimorphism. They have two distinct genomes: a somatic macronucleus that is polyploidy and actively transcribed, and a transcriptionally silent germline micronucleus that is used only during sexual reproduction. Following sexual conjugation, the newly formed zygotic genome of one of the nuclei is extensively edited to form the new macronucleus. Repetitive elements[35], transposable elements[33], and genes that could be deleterious if misexpressed in adults are all removed[47]; seemingly mirroring the logic in sequences removed from the lamprey somatic genome.

Ciliates employ two functionally distinct mechanisms for somatic DNA elimination, both of which make use of piRNAs in unique ways [34] (Fig 5). In *Tetrahymena*, piRNAs provide sequence-specificity to target germline-restricted regions in the developing somatic macronucleus for elimination. Targeted regions are then flagged with a histone modification that groups these sequences for collective deletion. Long non-coding RNA transcribed from the paternal somatic macronucleus acts as a sponge against germline piRNAs. Germline derived piRNAs that are not "absorbed" by the parental somatic long non-coding RNA then specifically mark only non-somatic regions in the newly formed zygotic macronucleus for elimination. Conversely, *Oxytrichia* use piRNAs and a PIWI homolog to mark sequences for retention in the somatic macronucleus. piRNAs also impart epigenetic heritability in ciliates, demonstrated by experiments where injection of synthetic piRNA mimics that hybridize to genes marked for deletion resulted in the retention of these genes for several sexual generations[35]. As my preliminary analyses suggest the possibility that eliminated DNA in lamprey is marked by repressive DNA and histone modifications, it seems possible that piRNAs are involved in marking sequences for elimination by recruiting

chromatin remodeling proteins in a manner similar to that employed by *Tetrahymena*.

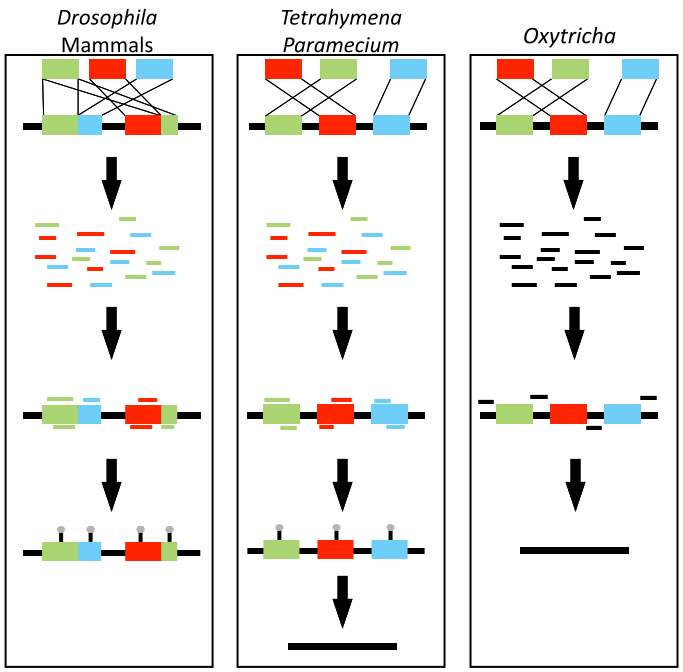

Fig 5: Small RNA-Mediated Genome Defense. Foreign transposable and repetitive elements are represented by blue, green, and red. Mammals and *Drosophila* use piRNAs to silence transposable elements by DNA methylation or repressive histone modifications (gray circles). In *Tetrahymena* and *Paramecium*, piRNA homologs recognize IES sequences, and guide genomic excision by marking them with repressive histone modifications. *Oxytricha* eliminates IES sequences using an orthogonal mechanism; piRNAs correspond to retained sequences.

Further justification for exploring small RNAs as a modulator of lamprey PGR comes from evidence gathered in my pilot study of small RNAs in D2.5 and D3 embryos (Fig 6). Additionally, several recent studies utilizing NGS have been successful in elucidating novel small RNA species[70]. I identified a distribution of

reads between ~30-47 nucleotides that are not explained by any known RNA species, and contain a conserved sequence motif (discussed below). Given the presence of this potentially novel class of small RNAs in actively rearranging embryos, we sought to further characterize this fraction and assess its presence in embryonic time points immediately before and after the rearrangement event.

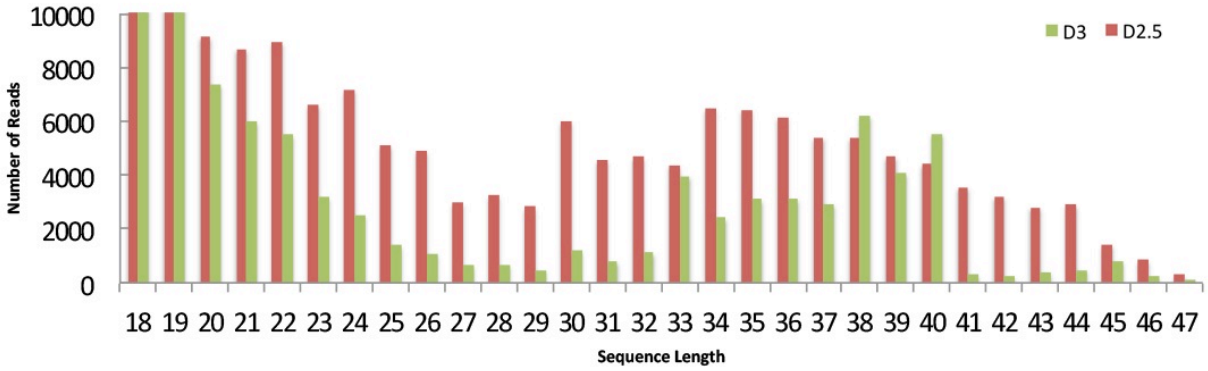

Fig 6: Size Distribution of Unexplained small RNAs. Non-annotated RNA reads from D2.5 and D3 embryos are indicated by color. The size distribution and abundance of the reads from each stage are indicated.

1.7 Scope of Thesis:

In this study, I sought to perform initial characterization of the small RNA complement of lamprey embryos temporally across embryogenesis including PGR. It is my hypothesis that I will observe a shift in small RNA expression spanning the PGR event, and that some small RNAs will be eliminated from the somatic genome and maintained exclusively in the germline. Understanding the dynamics of small RNA expression in early lamprey embryos could shed light into regulation of PGR, and provide comparative perspective on the basal state of small RNA biology during the initial stages of vertebrate embryogenesis.

To accomplish this, I collected Illumina deep sequencing data from total RNA isolated from embryos day 1, 2, 2.5, 3, 4 and 5 (D1-D5) post fertilization. The first aim was to assess the dynamics of known small RNA expression using databases of known miRNAs from miRBase, known piRNAs using piRNABank and Repbase, and other known RNA species using the Rfam database. Expression levels of annotated families of the small RNAs can be tracked across early development and provide us with resolution into the dynamics of small RNA expression through embryogenesis. The second aim of the study was to identify novel miRNAs and piRNAs and their covalent modifications using several existing software pipelines. These analyses reveal previously non-annotated families of piRNAs and miRNAs in gnathostomes, and contribute to our understanding of the diversity of small RNA species present in the vertebrate clade. The third aim of the study was to characterize a distribution of reads between 30-45 nucleotides sampled in my pilot study of D2.5 and D3 embryos. These reads are not explained by any known small RNA species, but contain a specific motif that is similar to known binding domains of transcription factors. The fourth aim of my study was to identify any small RNAs that are deleted during PGR and only present in the germline genome. Initial candidates were identified computationally using the somatic and germline assemblies as background for prediction purposes, then a subset of these were verified for germline specificity using PCR on DNA derived from germline and somatic tissues. Altogether, these studies will further our understanding of the endogenous small RNA classes present in early embryogenesis of a basal vertebrate, and reveal the dynamics of small RNA expression associated with genomic rearrangements in lamprey.

CHAPTER 2

MATERIALS AND METHODS

2 Animals

All animals were obtained from the Lake Michigan population, via the Great Lakes Fisheries Commission (GLFC) and maintained under University of Kentucky IACUC protocol number 2011-0848. Animals were euthanized by immersion in MS-222 (150 μg/mL), dissected, and tissues immediately were snap-frozen for the isolation of DNA from adult germline and somatic tissues.

2.1 Lamprey Embryos

In vitro fertilizations were performed with adult animals provided by the GLFC. Fertilizations were performed in vitro using 500 mL crystallization dishes. Females were processed to extract ~8,000 embryos per spawn by applying gentle pressure to the abdominal cavity and extruding eggs into artificial spring water ([NaCl: 5.9×10^{-3}M, $NaHCO_3$: 7.75×10^{-4}M, KCl: 3.3×10^{-4}M, $MgSO_4$: 1.7×10^{-4}M, $CaCl_2$: 1.8×10^{-4}M, Amquel: 4.5×10^{-4}M, Novaqua: 8.4×10^{-4}M). Embryos were then fertilized by extracting ~1 mL of sperm from 1-3 males by applying gentle pressure to the abdominal cavity. The mixture was then gently swirled to disperse sperm over the eggs and allowed to sit at room temperature for 10 minutes. Following the 10 minute incubation embryos were washed once with DI water in order to remove remaining sperm, then immediately replaced with fresh spring water. These embryos were then incubated at 18C to the appropriate developmental stage. Embryos were collected in 1.7 mL centrifuge tubes and snap

frozen for subsequent RNA extractions. RNA was extracted from ~100 flash frozen lamprey embryos from D2 (initiation of mid-blastula transition) D2.5 (mid-blastula transition), D3 (dorsal cone formation), D4 (beginning gastrulation), and D5 (neural groove formation), and ~200 embryos from D1 (24-32 cell stage)[58].

2.2 Small RNA Sequencing

Total RNA from sampled embryonic time points was isolated using Trizol. RNAs were quality controlled for RIN (Agilent RNA 6000 Nano) and samples with RIN > 8 were sent to the Hudson Alpha Genomic Services Lab (HudsonAlpha, Huntsville AL) for sequencing on an Illumina platform. Initially, small RNAs (<200 nt) were purified from total RNA using the miRNeasy Mini Kit (Qiagen), which includes RNAs from approximately 18 nt and upwards. Libraries for Illumina sequencing were then prepared following the NEB Small RNA protocol (New England Biolabs). Finally, Pippin prep (Sage Science) was utilized to select fragments between 110-140 bp (including sequencing adapters, or 1–50 excluding adapters) for the pilot study, and between 110-190 bp in the follow up study. Generated libraries were purified by QiAquick column (Qiagen). Samples from the pilot study were submitted to a MiSeq flow cell, whereas the follow up study was sequenced on a HiSeq platform. Samples from the pilot study on D2.5 and D3 embryos produced 20.9 million and 17.4 million reads, respectively, and the following numbers of sequencing reads were obtained for each time point in the follow up study: D1: 17.8 million , D2 zero, D2.5: 32.7 million , D3: 36.7 million, and D4: 22.7 million, D5 zero. Time points yielding zero reads apparently

represent errors in library preparation, sequencing, or library pooling. These issues are the subject of ongoing investigations.

2.3 Annotation of known miRNAs/piRNAs

Initial reads were processed to remove adaptor sequences (Illumina miRNA 3'RC adapter: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT) and filter for high quality reads (qscore >40) using the fastx toolkit[39].

All small RNA alignments performed in this study done using the short read alignment algorithm Bowtie[44]. The output was composed of the first valid alignment bowtie encounters for a given sequence allowing for up to two mismatches in order to increase sensitivity. Initially, reads were processed to remove all degraded mRNAs, rRNAs, tRNAs, Rfam encoded snRNAs, mitochondrial reads, and bacterial contamination by alignment to these known classes. Next, miRNA reads were annotated based on a perfect match to the pre-microRNA or mature miRNA sequence as reported on miRBase[9]. Small reads perfectly matching the stem regions of these hairpins or overlapping with known miRNA mature and passenger strands were retrieved using Bowtie with default parameters[41]. RNA secondary structures were predicted using RNAfold[52]. To identify miRNA reads with non-template directed nucleotide additions, reads with exactly one mismatch at the 3' end were extracted from the dataset.

To identify piRNAs, consensus sequences for vertebrate repetitive elements were retrieved from Repbase[48], and consensus piRNA sequences were retrieved from piRNAbank[49]. Small RNAs mapping perfectly to these consensus

sequences and their genomic flanking regions were retrieved using Bowtie with default parameters, and sorted into piRNA libraries.

2.4 Novel micro and PIWI Interacting RNA Prediction

Predicted miRNAs were determined computationally with the algorithm miRDeep2[55], using filtered sequence data. To identify novel miRNAs, samples from all four developmental stages were submitted to miRDeep analysis with a cutoff score of 4. miRDeep also estimates false positives by random permutation of the signature and structure-pairings in the input dataset to test the hypothesis that the structure (hairpin) of true miRNAs is recognized by Dicer and causes the signature. In my study, the estimated incidence of false positives was below 4%, indicating that the vast majority of predicted miRNAs had corresponding sequences characteristic of Drosha/Dicer processing. Predicted miRNAs by miRDeep were further filtered by aligning to databases of known miRNAs, tRNAs, rRNAs, transposable elements, and other known small RNAs.

Predicted piRNAs were determined computationally with the algorithm piRNApredictor using default parameters[50]. piRNApredictor makes use of positional nucleotide bias resulting from piRNA biogenesis to annotate sequences as piRNAs. For prediction analyses, RNAs shorter than 25 nucleotides were excluded to increase accuracy as piRNAs are typically 26-32 nt in length. Predicted piRNAs by piRNApredictor were further filtered by aligning to known miRNAs, tRNAs, rRNAs, snRNAs, and other known small RNAs.

2.5 Screening for Germline limited snRNAs

Oligonucleotide primer pairs were designed to amplify miRNA or piRNA containing genomic regions that were present in the latest lamprey germline assembly but had no corresponding bowtie alignment to any sequence in the WGS dataset. Oligonucleotides were designed using Primer3[56] and used to prime PCR reactions under standard amplification conditions [2 ng of DNA, 50 ng of each primer, .3 U Taq polymerase, 1X PCR buffer, and 200 mM each of dATP, dCTP, dGTP, and dTTP; termal cycling at 94 C for 4 min; 33 cycles of 94 C for 15 s, 60 C for 15 s, 72 C for 15 s; and 72 C for 5 min]. These reactions also included a second pair of oligonucleotides at ½ the concentration of target primers (25 nM) that were designed to amplify a single (gene-coding) region of the lamprey somatic genome [*P. marinus* Scaffold_ 256 (NCBI accession number GL476584): F, TCAACACCTACGGTTCACCA; R, CCTTAAAGGCAGCGC TATTG]. The DNAs used in these reactions were extracted from multiple tissues (testes, blood, liver, kidney, muscle, and tail fin) that were collected from two individuals (Animals 12 and 13), using standard phenol chloroform extraction[57].

2.6 Motif Identification

Motif discovery on the non-annotated fraction of reads between 30-49 nucleotides was performed using the comparative algorithm MEME: Multiple Em for Motif Elicitation[37]. Motif identification for the unexplained distribution was performed on the non-redundant set of reads to avoid calling high copy number sequences as distinct motifs. Briefly, sequences were divided in the target and background groups for each application of the algorithm. Background sequences were derived from the germline genome assembly, and were selectively weighed

to equalize the distribution of G+C content to avoid comparing sequences of different general sequence content. Motifs between lengths of 8 and 50 nucleotides were identified exhaustively by screening all oligonucleotides present in the dataset for enrichment in the target set compared to the background set using Bayes optimal classifier to score enrichment. Up to 2 mismatches were allowed to increase sensitivity of the method. Motifs representing >15% enrichment in the unexplained fraction compared to background were considered for further inquiry. Sequence logos were generated using WebLOGO[38].

CHAPTER 3

RESULTS

3 Small RNA Sequencing

I isolated total RNA from lamprey embryos at six different developmental stages (1, 2, 2.5, 3, 4 and 5 days post fertilization) expanding on an experiment performed in D2.5 and D3 embryos. These time points correspond to the 24-32 cell stage, the mid blastula transition, dorsal cone formation, beginning gastrulation, and development of the neural crest at day 5[58]. We chose these stages because they represent key changes in expression patterns for vertebrate development, and encompass the rearrangement event which occurs during the mid blastula transition[1]. To examine gene expression patterns during these stages, small RNA libraries (15-50 nt) were prepared and sequenced with 50 bp single end sequencing on an Illumina HiSeq platform. Barcoded libraries were constructed and sequenced by the Genome Services Lab at HudsonAlpha Institute for Biotechnology. Sampled from day 2 and day 5 yielded no reads, but

the remaining libraries yielded large numbers of reads that passed quality and vector filtering (Table 1) and capture critical time points that include and flank PGR.

Table 1: Number of reads obtained from each library post-trimming.

D1	D2.5	D3	D4
17.8 million	32.7 million	36.7 million	22.7 million
PILOT STUDY	20.9 million	17.4 million	

Analysis of the size distribution of all the non-redundant reads within our libraries revealed at least two classes of snRNAs between 15-40 nucleotides. One size class

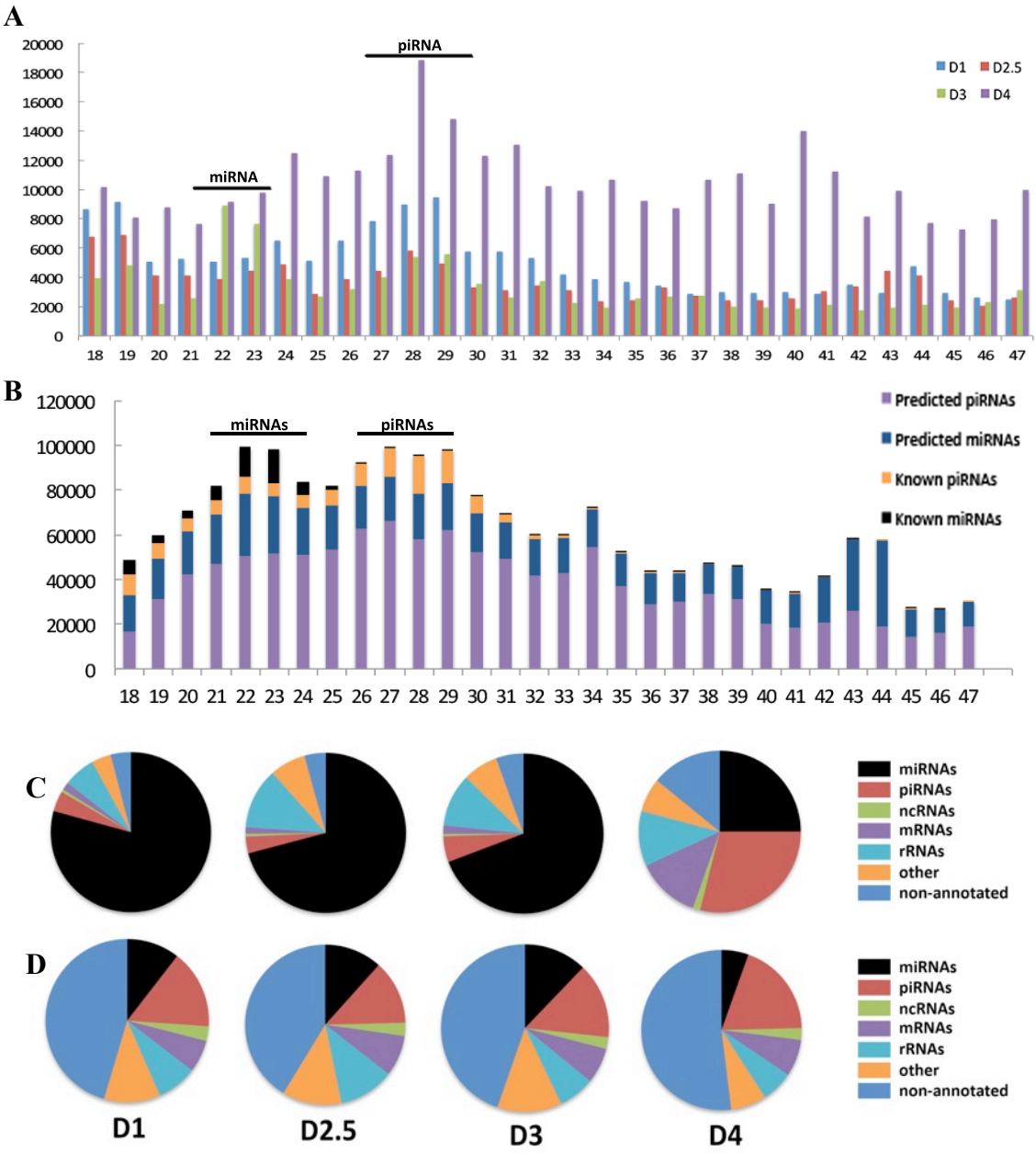

Fig 7: Summary of short read sequences generated under this study: (A) Size distribution of all small RNAs between 18 and 47 nt. RNA reads derived from 4 developmental stages are indicated in different colors. The size distribution and abundance of the reads from each stage are as indicated. (B) Length distribution of reads annotated as miRNAs and piRNAs. (C) Read frequency for all sequences. The annotation and frequency for all small RNA reads from different developmental time points are as indicated. (D) Read frequency for unique sequences. Small RNAs composed of an identical sequence were grouped together as a single subset. This analysis shows the diversity of

peaked at ~22 nucleotides and a second at ~28 nucleotides, consistent with lengths of miRNAs and piRNA respectively (Fig 7).

3.1 miRNA expression analysis

Analysis of small RNA reads identified 22,110 known miRNAs, which could be grouped into 9622 distinct families (Fig 7). Most of the reads detected were derived from the mature binding strand, with the star or passenger strands appearing at a much lower frequency. These observations are consistent with the standard miRNA biogenesis model[21] (Fig 1). Previous studies have shown that miRNA read numbers reflect actual miRNA levels, although with a small amount of bias introduced by secondary structures and post-transcriptional modifications[51]. The read counts of miRNAs are also highly correlated with their activity in vivo, as miRNAs are simply the targeting mechanism of the endogenous RISC present in most eukaryotes. Changes in miRNA abundance across development were estimated by direct comparison of read counts for a given miRNA across sampled stages, a strategy that has proven successful in previous studies[7,8,18]. As an increase in the general amount of transcription occurs as development proceeds, the miRNA read numbers from each sampled time point were standardized to the read counts of *let-7,* which was expressed at levels near the average miRNA read count in all four libraries, divided by the total number of reads annotated as miRNAs in each library.

In sequences generated from embryos at day 1 post-fertilization, I identified 9101 miRNAs belonging to 2401 families (Fig 8, 9). Because this early embryonic time point was sampled prior to the onset of zygotic transcription, it can be inferred that these miRNAs are maternally deposited. It is interesting to note that maternally deposited miRNAs are generally present at relatively high levels in lamprey embryos, suggesting they may play an important role during the

earliest stages of development[17.] I also observed that miR-430 was overwhelmingly the most abundant miRNA species expressed from D2.5 –D4. As miR-430 has been associated with clearing of maternally deposited transcripts in other species[42,66], we believe the robust expression of miR-430 reflects the onset of the maternal-zygotic shift in lamprey embryogenesis.

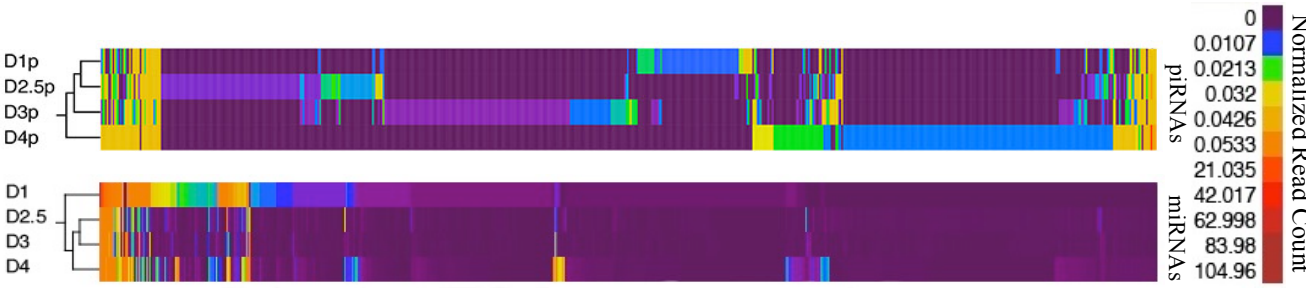

Fig 8: snRNA Expression Profiles. miRNA (22,110 sequences) and piRNA (30,078 sequences) expression was normalized to a moderately abundant species present in all libraries and compared across four developmental stages. Clustering was performed using JMP hierarchical clustering with default parameters. Red indicates high levels of expression and purple indicates low levels of expression

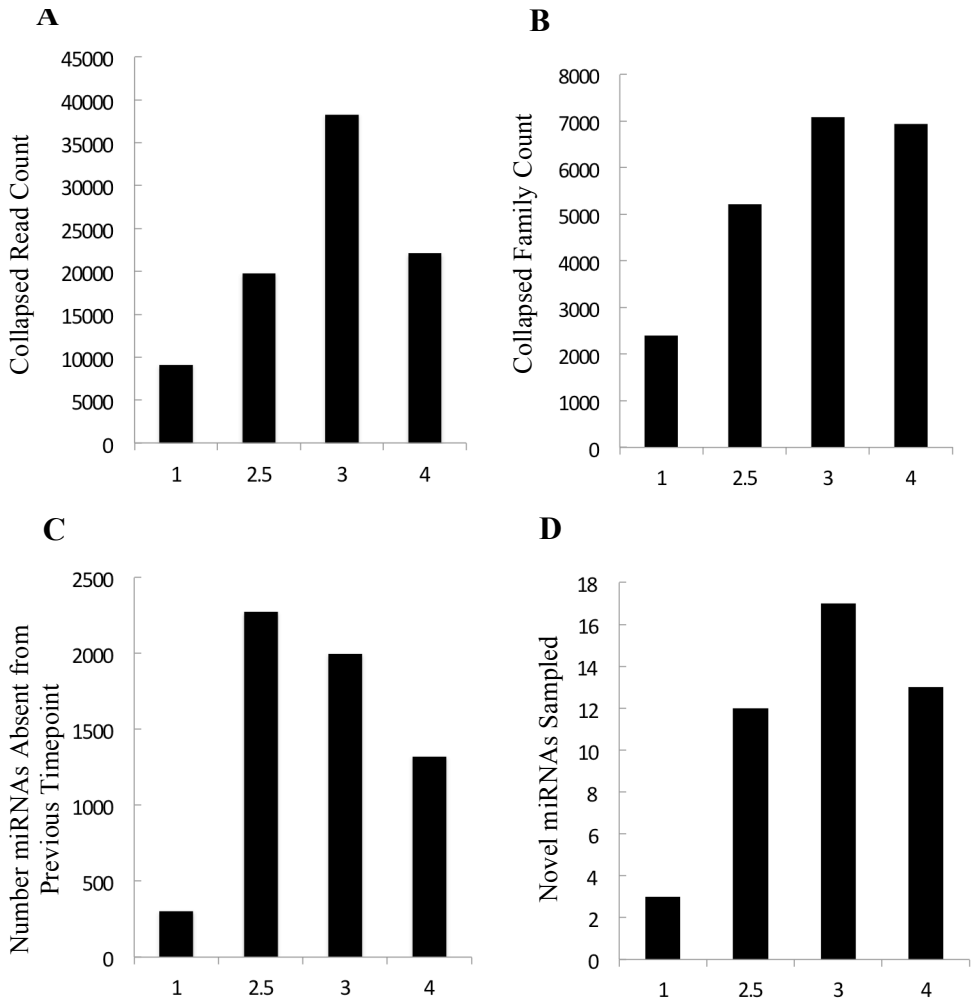

Fig 9: miRNA Expression Dynamics. (A) Read frequency of unique miRNA families and family variants previously reported on miRBase across four developmental stages. (B) Read frequencies for miRNA families reported on miRBase. All mature sequences containing the same 5' seed sequence were grouped into a single family subset. (C) Abundance of miRNA species absent from the previous libraries across four developmental stages. (D) Number of novel miRNAs annotated by predictive software miRDeep2 across development. Predicted miRNAs had to be sampled at a minimum of twice per library, had a miRDeep cutoff score of 4, and aligned to no known RNA species present in any queried database.

3.2 Identification of Novel miRNAs

We next sought to identify and annotate novel miRNAs present in our raw sequencing data. To accomplish this, we analyzed all small RNA reads from each time point using the miRDeep2 algorithm[55]. The algorithm identified a total of 398 candidate novel miRNAs. These predicted miRNAs were aligned to known metazoan miRNA sequences deposited on miRBase[9], to remove predicted sequences that match previously annotated miRNAs. The set of putatively novel miRNAs was further filtered to remove tRNAs, rRNAs, transposable elements, other known small RNAs and reads that were not resampled in at least one

library. I also discarded 68 sequences where the mature form of the candidates was shorter than 18 nt and longer than 26 nt, as mature miRNAs are unlikely to deviate from that size distribution[14]. After filtering, 45 sequences were retained that may represent novel miRNAs in lamprey (Fig 10).

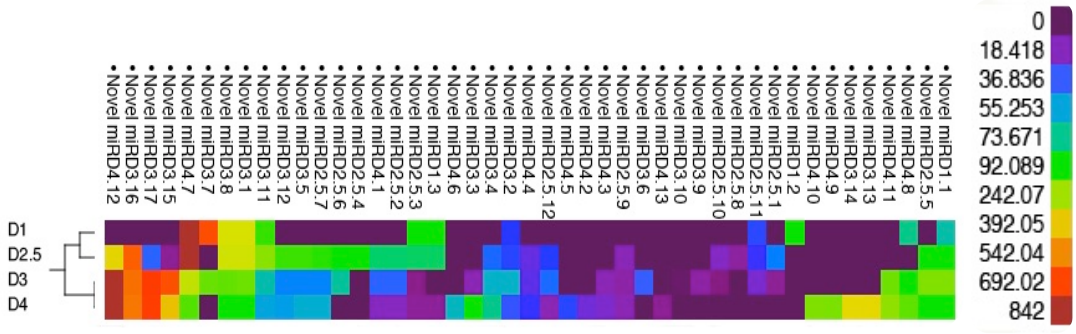

Fig 10: Novel miRNA Expression Profile. Read counts of miRNAs identified as novel by miRDeep2 were compared across four developmental stages. Red indicates high levels of expression and purple indicates low levels of expression. Many novel miRNAs are expressed highly during the earliest stages of development and are likely maternally deposited.

3.3 miRNA Sequence Variation

Sequence variations at the 3' end of miRNAs have been reported following deep sequencing[51,53,54]. The type of nucleic acid added to a 3' modification can result in the stabilization or degradation of the target miRNA[42]. A notable proportion of reads contained 3' miRNA modification at all sampled stages of development, with 2.7% at D1, 3.8% at D2.5, 6.9% at D3, and 4.7% at D4. Variations were identified by mismatches between the reads and their corresponding genomic loci, with the majority of variations occurring at the 3' end. In 56 cases out of 4,644, the additional nucleotides matched the sequence of the pre-miRNA, revealing a small degree of variation (~1%) in the precise location of cleavage sites during miRNA processing.

The majority of sequence variation detected (96% averaged across sampled time points) resulted from the addition of uracil and adenine. Only a few families of miRNAs (18) had cytosine or guanine addition, and these modifications were never

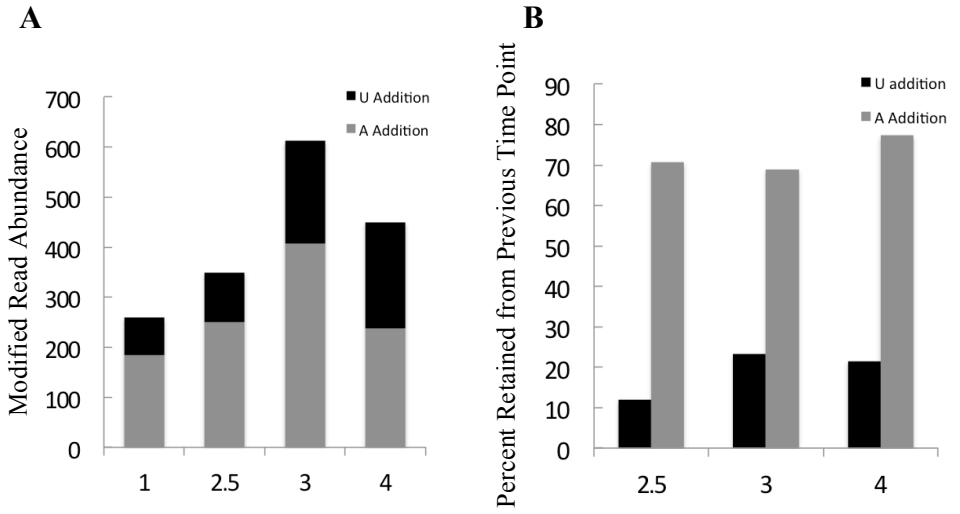

Fig 11: miRNA Sequence Heterogeneity. (A) Abundance of reads with Uridylated or Adenylated 3' ends from four sampled time points. (B) The ratio of A or U tailed reads to the total number of reads containing those modifications at the previous sampled developmental stage.

observed in more than one library. Our data shows increasing modification of a variety of families of miRNAs as embryogenesis proceeds (Fig 9), suggesting the activities of these miRNAs are being actively modulated. To further evaluate the effects of 3' modifications, we assessed the abundance of uridylated and adenylated transcripts in the following time point. As shown in (Fig 13) miRNAs that have been uridylated are much less likely to be sampled in the following time point compared to miRNAs that have been adenylated. This observation is consistent with the idea that uridylation acts as a signal for destabilization whereas adenylation results in retention, and suggests that lamprey employ the same regulatory logic observed in other vertebrates[7,67].

Increased levels of uridylation were observed as development proceeded, suggesting a clearing of maternally deposited miRNAs. Interestingly, if we examine all the modifications of miRNAs across all sampled time points, we see the highest amount in D3, followed by D4 and D2.5. These data are consistent with the idea that there is active stabilization/destabilization of maternally deposited transcripts as zygotic expression initiates, as well as suggesting a shift in general modification levels associated with the midblastula transition and completion of genome rearrangement.

3.4 Lamprey piRNA expression

We observed a distinct distribution of reads between ~25-31 nt that map to either repetitive or unique genomic loci. Based on the size distribution of these reads, I inferred that many of these reads likely represent piRNAs. A subset of these reads mapped to known piRNAs as made available on piRNAbank. Further evidence that these are bona fide piRNAs is that 114,370 sequences mapped to known repetitive elements deposited on Repbase. Finally, piRNA prediction from this subset of reads was performed using the algorithm piRNA predictor, which makes use of a kmer scheme to call reads as potential piRNAs[50]. From among the 24,535 piRNAs predicted by piRNA predictor, 9,077 had sufficient evidence to be annotated as piRNAs in lamprey. For the reads predicted to be derived from bona-fide piRNAs, there was a strong preference for an A at position 10, and a U at position 1 (Fig 13). These data are consistent with the expectations given the parameters of the search algorithm and the canonical piRNA Ping-Pong biogenesis pathway (Fig 2). In general, piRNA family diversity increased through embryogenesis to a peak at D3, after which diversity decreased in D4 embryos

(Fig 12). Interestingly, piRNA abundance peaked at D4 (Fig 6). This suggests that the overall composition of piRNAs at D4 is more homogenous but total (normalized) transcription is increasing, possibly due to the increased diversity of cell types present at D4. As genomic rearrangements complete during D3, this signal of a drop in piRNA diversity may reflect a change in the small RNA state induced by the completion of genomic rearrangements. Alternatively, if PGR is targeting piRNAs, the observed drop could also reflect a decrease in piRNA diversity initiated by elimination from the somatic genome.

3.5 Germline Specificity

To test germline specificity of snRNAs, I performed alignments of miRNA and piRNA reads to sequences derived from sperm DNA and a large dataset of WGS sequences that were derived from the liver of the individual that was the subject of the current lamprey genome

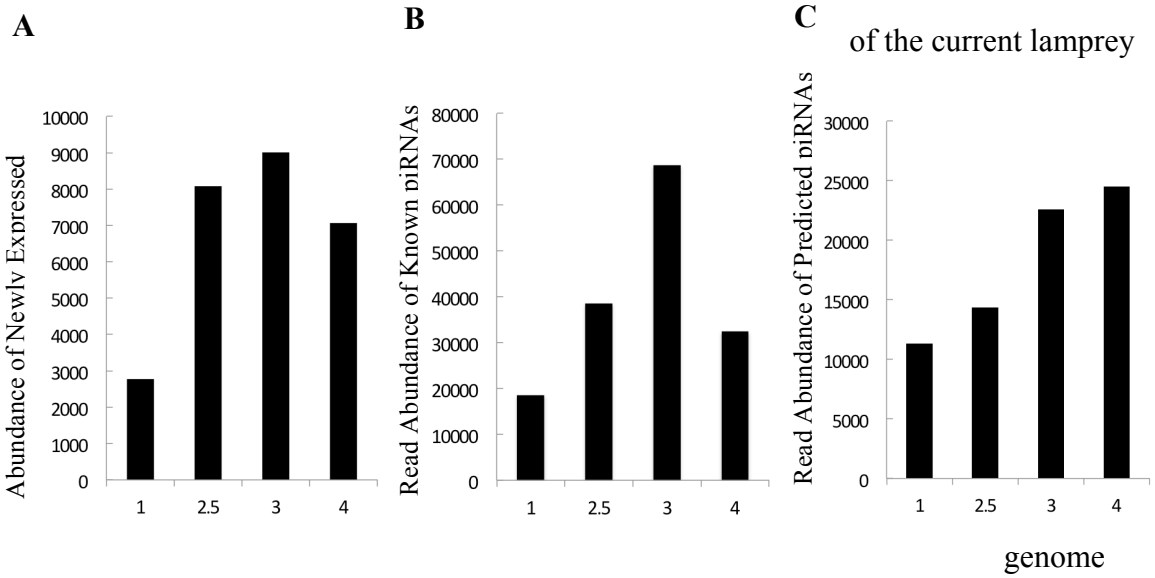

Fig 12: piRNA Expression Dynamics. (A) Abundance of piRNA species absent from the previous library across four developmental stages. (B) Read frequency of unique piRNA species previously reported on piRNAbank across four developmental stages. (C) Read frequency of sequences annotated as piRNAs by predictive software piRNApredictor.

assembly[68]. These analyses identified a total of 4,664 germline-specific candidate miRNAs and 6,834 piRNAs (Table 2).

Polymerase chain reaction primers were developed for 24 miRNA loci and 12 piRNA loci that mapped to sperm derived reads but showed no strong alignment to the WGS database. One miRNA primer pair and one piRNA primer pair produced reproducible and correctly sized amplicons in germline (sperm or testes) but produced no amplicon in somatic tissues (blood, liver, kidney, muscle, or tail fin) from two different animals (Fig 15). The other primer pairs yielded a single strong fragment of expected size in DNA from all tissues, suggesting that these sequences were not germline specific. The lack of alignment to the somatic WGS dataset presumably reflects the fact that these sequences lie in gaps of coverage in somatic dataset. The validated germline specific miRNA is a member of the mir25-a family, which is known to be upregulated in several types of cancer[59]. The validated germline specific piRNA maps back to an RTE retrotransposon present in the lamprey genome. The logic underlying its removal from the somatic genome is less clear than mir25a.

Table 2: Numbers of predicted germline specific miRNAs and piRNAs sampled in each library.

snRNA	D1	D2.5	D3	D4
miRNAs	641	911	2,006	1,106
piRNAs	1,349	2,193	2,152	1,139

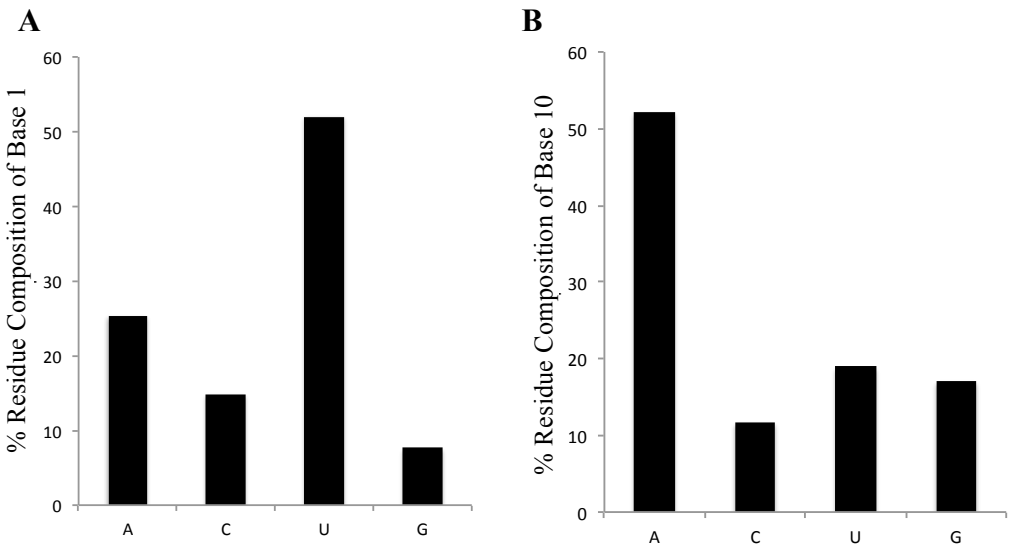

Fig 13: Residue Composition of Predicted piRNAs. (A) % Composition of nucleotide residues present at the first bp from the 5' end in piRNAs annotated computationally with piRNApredictor. (B) % composition of nucleotide residues present at the 10th bp from the 5' end in piRNAs annotated computationally with piRNApredictor. These patterns of nucleotide enrichment are consistent with the algorithm and previously described Ping-Pong biogenesis pathway.

3.6 Unexplained fraction

When plotting the entire distribution of unique reads for a given length in our pilot study, we noticed a humped distribution of reads between 30-40 nucleotides (Fig 6). These reads are not explained by tRNA, rRNA, degraded mRNA, mitochondrial RNA, bacterial RNA, miRNAs, piRNAs, or any other annotated snRNA species. As such, we sought to characterize this fraction in an independent set of libraries (derived from different embryos and using an "updated" method for library preparation) to determine if this population of sequences was reproducible and if there were any shared sequence signatures among this potentially new class of small RNAs .

Our follow up study did not recapitulate the humped distribution observed in our pilot study (Fig 12). However, these samples did contain a large fraction of reads in the 30-45 nt range and a large fraction of non-annotated reads (Table 3, Fig

14). In retrospect, it seems likely that the humped shape of the distribution observed in the pilot study was an artifact due to the library preparation and size selection method, which should have selected against the sequencing or RNAs greater that ~40bp.

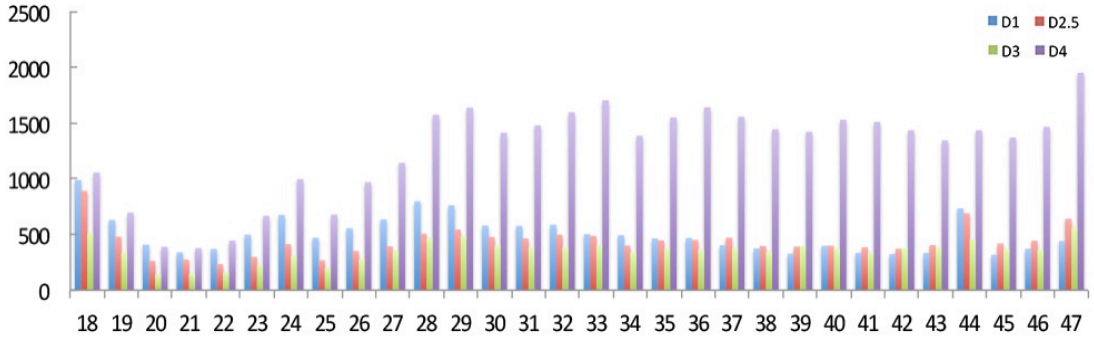

Fig 14: Size Distribution of Unexplained Fraction. All non-annotated small RNAs between 18 and 47 nt. RNA reads derived from 4 developmental stages are indicated in different colors. The size distribution and abundance of the reads from each stage are as indicated.

Table 3: Ratio of non-annotated reads between 30-45 nt to total non-annotated reads in the pilot and follow up study.

Library	D2.5	D3
Pilot Study	23%	22%
Follow Up	21%	19%

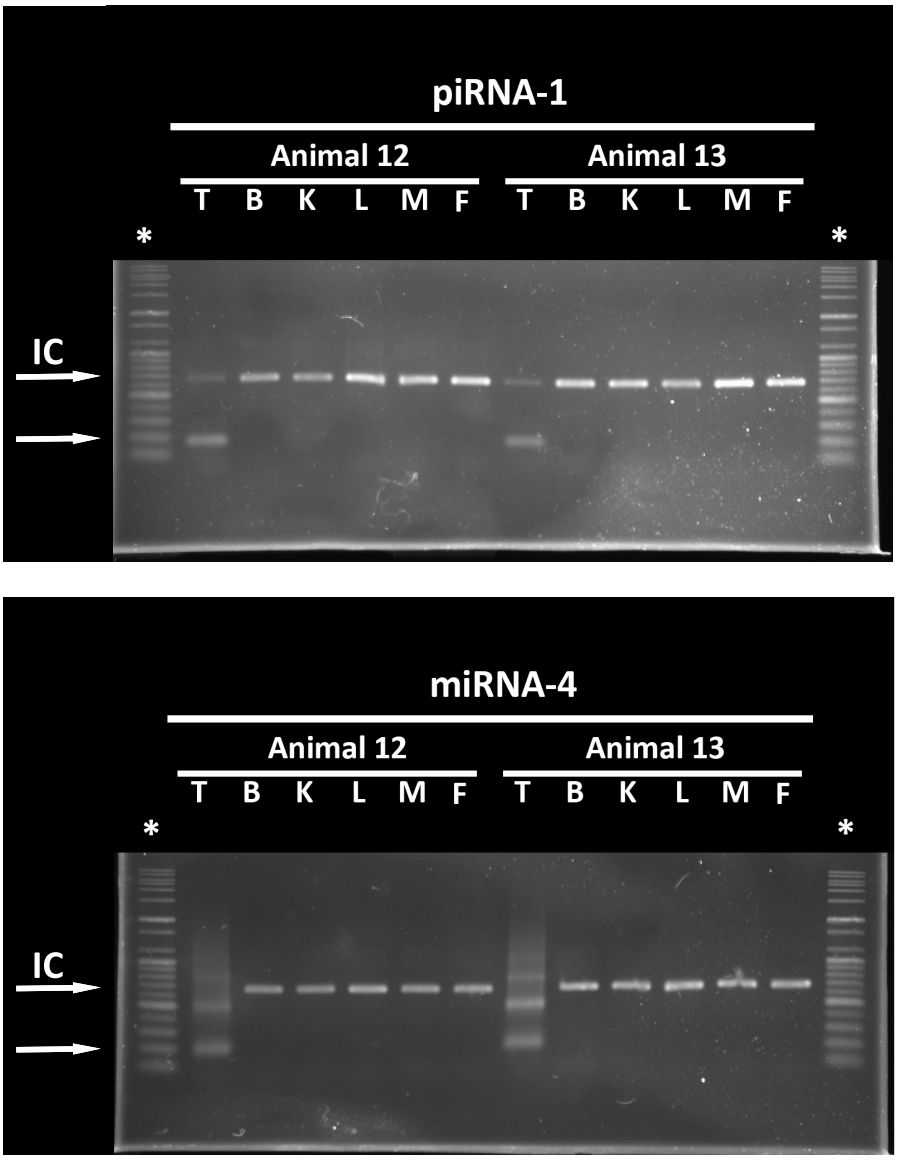

Fig 15: Polymerase chain reaction validation for two germline-limited snRNAs that were identified in computational comparison between sperm and blood. Fragments were amplified from the DNA of several tissues [testes (T), blood (B), kidney (K), liver (L), muscle (M), and tail fin (F)] that were collected from two animals (Animals 12 and 13). Primers for internal control (IC) were included at ½ the concentration of target primers. The expected sizes of target and IC fragments are marked on the left by arrows. Presence of primers for miRNA-4 appears to produce fragments, which hybridize with IC fragments. However, no miRNA-4 fragments were seen in DNA sampled from somatic tissues, and miRNA-4 primer pairs run without an IC produce a single fragment of the expected size. Amplified fragments are flanked by size standards (asterisk, 100 bp plus DNA ladder, Invitrogen).

3.7 Motif Analysis

Motif discovery in the unexplained fraction was achieved using the MEME Suite. My dataset of unexplained reads was >100,000 sequences in each individual libraries, and as MEME employs an all by all comparison method, the

computational demands for searching the entire dataset as a whole were substantial. In order to reduce the machine run time to reasonable levels, all reads from the unexplained fraction at each time point were randomly subsampled into fifths and queried for conserved motifs using default parameters. All fractions from unexplained sequences contained in our D2.5 and D3 libraries returned a motif similar to one observed in a pilot study of the same time points (Fig 14). Interestingly, D1 returned a compositionally distinct motif, and only motifs associated with polyadenylation were detected at D4. As I was only able to detect this motif in sequences from the mid-blastula transition, I will notate it as Mid-Blastula Limited Sequence (MBLS) for the purposes of this thesis. Although sequences containing MBLS motif were present at D1 and D4, the abundance of this motif in D1 and D4 was not significantly different from background sequence (<15% enrichment at both D1 and D4, Fig 17). The MBLS motif was queried against the Uniprot database of known DNA binding motifs to identify potential analogs. Our search indicates that MBLS is similar to a binding motif of Interferon regulatory factor 3 (IRF3), which contains several functional domains including a DNA binding domain and nuclear export signal[68] (Fig 16). As small RNAs are responsible for guiding excisional machinery to the sequences eliminated in ciliate PGR, it is interesting to speculate that sequences containing this motif are mediating the nuclear export of material excised from the somatic genome using the nuclear export signal contained in IRF3.

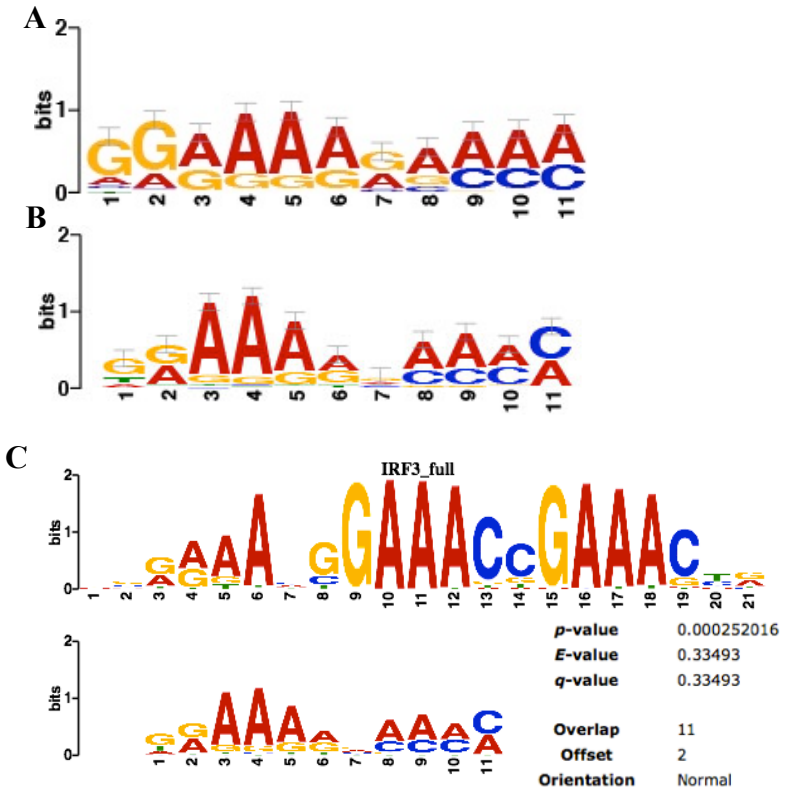

Fig 16: Computationally Identified Sequence Motifs Enriched in Non-Annotated Reads. (A) Consensus motif computed from the averages of PSPM matrices found in non-annotated sequences sampled in D2.5 and D3 of pilot study. (B) Consensus motif computed from the average of PSPM matrices of D2.5 and D3 samples of independently prepared libraries (the MBLS sequence). (C) Consensus motif in (B) was queried for homolog to known motifs in the JASPAR and UniPROBE vertebrate motif database using TOMTOM. The most significant hit returned matches a known DNA binding domain in IRF3.

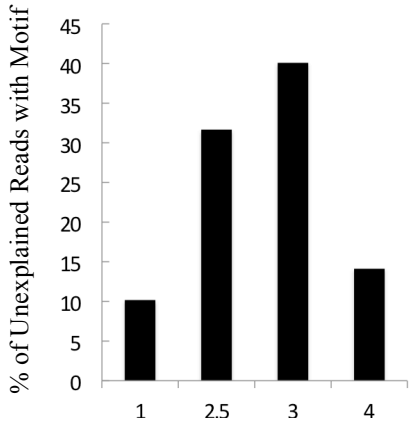

Fig 17: Motif Distribution in Unexplained Sequences. The ratio of sequences containing the MBLS consensus motif relative to the total non-annotated reads in a given library.

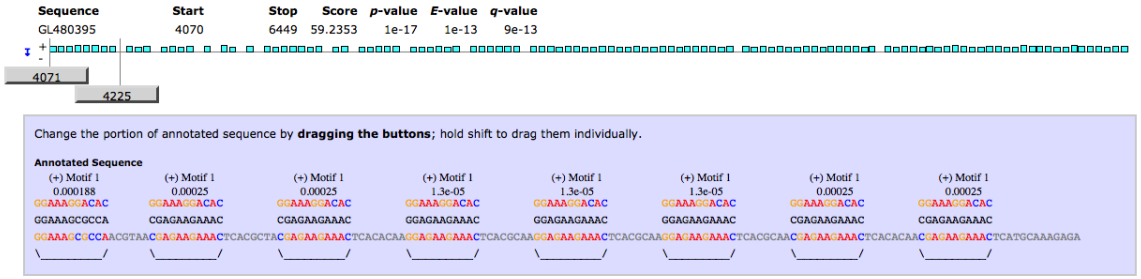

Fig 18: Representative MCAST Output. The lamprey germline genome was searched for statistically significant non-overlapping hits to the MBLS consensus motif. Only alignments that produced a *p*-value <0.0005 were considered. The maximum distance allowed between the hits was 50 bp. This query identified 246 discreet loci that contained >100 instances of the consensus motif.

3.7.1 Clustering in the genome

We next sought to characterize the distribution of MBLS motifs in the genome. We used FIMO[60] to search for individual motif occurrences in both the lamprey somatic and germline assembly. Our search showed no bias in motif content of the somatic compared to the germline assembly, suggesting these sequences occur with equal frequency in both. As piRNAs and other snRNAs tend to cluster in the genome, we next queried our motif against the lamprey ENSEMBL genome using MCAST, a motif cluster identification software available through the MEME suite[37]. Our analyses showed that the MBLS motif clusters quite heavily in the genome, with 246 discreet loci with fewer than 50 bp between each occurrence of the motif (Fig 18). This organization suggests that large numbers of sequences with this motif can be transcribed en masse, which might be an important biological function for these RNAs during PGR. As robust expression of clustered snRNAs is vital for organizing PGR in ciliates, these sequences provide an enticing candidate for a similar regulatory mechanism operating in lamprey rearrangements.

CHAPTER 4

DISCUSSION

In this study I characterized the small RNA profile of early *P. marinus* embryogenesis and analyze the temporal dynamics of snRNAs with high throughput sequencing. I isolated total RNA from D1, D2, D2.5, D3, D4, and D5 lamprey embryos corresponding to the 24-32 cell stage through the mid blastula transition, dorsal cone formation, beginning gastrulation, and neural crest formation. To examine gene expression patterns during these stages, small RNA libraries were prepared for high throughput sequencing using the Illumina platform. In total, these RNAseq experiments yielded more than 100 million (109,897,951) sequence reads for our discreet embryonic stages. Analyses of these datasets reveal several salient aspects of small RNA biology during early embryogenesis and programmed genome rearrangement including 3' modification of miRNAs, a novel class of sequences, and germline specific snRNAs; these are discussed in more detail below. Finally, I speculate as to further lines of research that might advance our understanding of small RNA biology in the context of PGR and lamprey embryogenesis.

4 A novel class of sequence upregulated during PGR

Perhaps the most surprising (and perhaps mechanistically relevant) finding of the current study is the discovery of a novel class of small RNAs that is highly over-represented in embryos that are actively undergoing programmed genome rearrangement. Sequencing of D2.5 and D3 embryos from a pilot study discovered a distribution of reads between 30-45 nucleotides that are not

explained by any known RNA species. Analysis of the composition of these sequences revealed a conserved motif that is similar to known DNA binding domains and clusters in large (several kb) stretches of the genome. Although our replicate study did not reproduce a comparable humped distribution of reads, motif analysis of 30-45 nt sequences revealed enrichment for a motif similar in composition, genomic clustering, and homology to known DNA binding motifs. Interestingly, this motif (MBLS) was only significantly enriched relative to background in libraries derived from D2.5 and D3 embryos. Over 30% of the sequences in D2.5 and D3 contain the MBLS motif, which corresponds to thousands of individual transcripts (19,225 and 20,047 respectively: Fig 15). The lack of significant enrichment after D3 suggests the intriguing possibility that sequences containing it could be related to the rearrangement process. As other taxa that undergo PGR use snRNAs to mediate targeting of rearrangements, the replication of a conserved DNA binding RNA motif expressed exclusively during PGR is enticing as a potential factor involved in lamprey PGR.

4.1 miRNAs and piRNAs are germline specific

Genomic DNAs from all somatic tissues surveyed showed a consistent pattern of loss for miRNA-25a and a piRNA targeting an RTE retrotransposon. These data are consistent with patterns of loss previously verified for the germline limited sequence *Germ1*[1] and a handful of protein coding genes[65], and confirms that small RNAs are also targets of elimination during PGR. The primer pair miRNA-4 corresponds to a lamprey homolog to miR-25a. As miR-25a desensitizes cells to induced apoptosis[59], I hypothesize that its removal during PGR functions as a method for limiting the proliferation potential of somatic tissues compared to

totipotent germ cells. The biological function of the removal of a piRNA locus from the somatic genome is less clear, but mirrors patterns of loss in other systems that undergo PGR[3,34,47], and as is the case in the *Oxytricha*, could indicate a potential role of piRNAs in mediating PGR[35]. Alternatively, as piRNAs typically lie in repeat rich portions of the genome, piRNA-1's absence in soma could be the result of the removal of its target repetitive sequences or an alternate regulatory role for this piRNA. Generally, validation of the elimination of snRNA sequences from somatic tissues could give us insight into the selection criteria for removal of germline-limited sequences from soma.

4.2 Dynamics of small RNA expression across time points.

4.2.1 miRNA expression

miRNAs are thought to play a critical role during early development, where embryos must reprogram their transcriptional landscape as they continue to differentiate. Using deep sequencing, we obtained expression patterns for 9,622 distinct miRNAs and identified 45 novel miRNAs expressed during early development (Figs 9, 10). Previous studies suggest that miRNA expression patterns become increasingly complex as development proceeds. Our results show that a large number of miRNAs are present at the earliest stages of development before zygotic transcription has begun (Fig 8). These miRNAs are likely maternally deposited and may either regulate the earliest stages of embryonic development or function similar to mir-430 in clearing maternal transcripts for degradation during the maternal-zygotic shift. We found that the most abundant miRNA species present are derived from the mir-430 family, highlighting the

importance and deep conservation of this family's function in early embryogenesis (Fig 19).

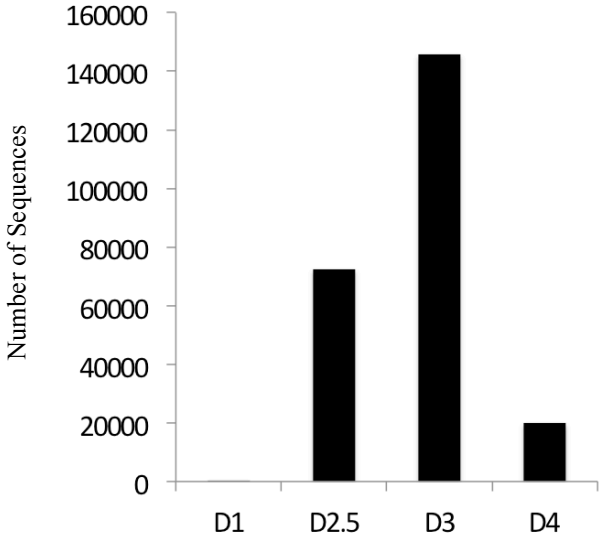

Fig 19: miR-430 Expression. Read counts of miR-430 from D1-D4 of development. miR-430 was the most abundant miRNA family at all sampled time points.

Interestingly, miRNA diversity increases as development proceeds, peaking at D3, after which both diversity and overall read count for miRNAs drops. Previous studies in vertebrates have reported an increase in the complexity of miRNA expression proceeding through equivalent stages of embryogenesis[11,17,57]. As the PGR event spans D2.5-D3 of development, it is possible that the discrepancy we see from this pattern is due to miRNAs being expressed to regulate activity of the machinery responsible for orchestrating genomic rearrangements. Alternatively, this incongruity could be explained by differences in sample/library preparation between D3 and D4; although it should be noted that these samples were prepared and sequenced at the same time and by the same facility. Based on the verification of a germline limited miRNA, and computational identification of thousands more, another possibility is that PGR could limit the number of miRNA species

available for somatic expression, and result in an embryo wide reduction in miRNA expression. Future sequencing replicates of these time points should resolve the true dynamics of miRNA expression through these developmental stages. Additionally, sampling miRNA expression at stages later in embryogenesis could provide context for the reduction I see at D4, and give a more complete view of the small RNA complement of lamprey embryogenesis.

4.2.2 piRNA expression

Studies have reported several overlapping roles for piRNAs: first they are specifically expressed in the germline as a means of maintaining genomic integrity by destabilizing the transcripts of endogenous retroelements. Second, piRNAs can serve as an epigenetic signal of self vs non-self transcripts. Finally, they can act as mediators of genomic rearrangements in other model systems. We identified many piRNAs present in D1 embryos, which is consistent with maternal inheritance of this subset of RNAs and/or their expression in the maintenance of the genome integrity of oocytes. Sequences annotated as piRNAs by both alignment to known databases and using predictive software show a uniform bias for uracil at position one and adenine at position 10 (Fig 11). This is consistent with accepted models of piRNA biogenesis, and suggests that lamprey employ the Ping-Pong model described in other organisms. Temporally across embryogenesis, overall piRNA diversity peaked at D3, but the percentage of total reads attributed to piRNAs remained static from D1-D3 (Fig 6). During D4, however, we observed a dramatic expansion of the percent composition of reads annotated as piRNAs compared to other RNA species. The exact nature of this expansion is unclear, but it could reflect the endogenous piRNA/PIWI response to

increased transposable element activity as the germline differentiates. Another intriguing possibility is that piRNA expression is increased in a response to the completion of PGR, the function of which is uncertain. Future studies characterizing adult and later embryonic levels of piRNA expression should shed light on the nature of the expansion seen during D4 of embryogenesis.

4.3 3' end modifications associated with retention or clearing of miRNAs

miRNA 3' tailing and trimming has been previously reported in flies, mouse, human, and zebrafish cells[40,44,45]. As in other species, modifications detected in lamprey were primarily adenylation and uridylation, at the 3' ends. The temporal patterns of adenylation versus uridylation addition suggests that the addition of adenosine residues might stabilize miRNA half-life, whereas uracil addition may promote degradation and clearing of transcripts. We detect the highest levels of modification during the D3, after which modification levels begin to drop. It seems likely that these modifications are related to mechanisms regulating miRNA half life during the maternal to zygotic transition and the differentiation of somatic cells types as is known for other species[67,7].

4.4 Future Directions

A major question that goes unanswered is: What purpose do these DNA deletions serve in the lamprey? Discerning how lampreys regulate rearrangements will be fundamental in understanding the biological role that PGR plays in embryonic and adult lamprey. As ciliates use snRNAs to direct targeting of their rearrangements, characterizing the small RNA profile of rearranging embryos

provides a potential avenue to discover key processes associated with lamprey PGR. The studies presented here have uncovered several interesting patterns of small RNA expression correlated with PGR. Replication of these results in separately generated libraries will be critical to verify their relevance in the rearrangement process. Characterizing small RNA expression in an adult animal and comparing it to expression during embryogenesis could also present an interesting comparative framework for differential small RNA expression. A thorough examination of the adult snRNA complement would give us the ability to understand which genes are upregulated during development compared to adult expression, and allow us to more accurately discern which snRNAs, if any, are highly expressed solely during PGR. As I have identified a novel distribution of reads during early development, it would be interesting to investigate the expression of these sequences in adult lamprey. Germline specificity of sequences containing MBLS in adults would give substantial evidence to the possibility that the activity of these sequences is associated with germline biology.

Another key study will be to determine if there is variation in the genomic snRNA content between somatic tissues resulting from lineage-specific rearrangement events. Cryptic variation between the genomic content of different tissues is still quite possible, and revealing these differences could be critical in elucidating the true regulatory function of PGR in lamprey. As many miRNAs show prominent tissue specificity in their expression[73,74] and are targets of deletion during PGR, they provide an attractive avenue of study for identifying genomic tissue variation resulting from lineage specific rearrangements. Finally, future studies should seek to more thoroughly describe the novel class of small RNAs sampled from D2.5 and D3 embryos in both our pilot and follow up

study. Sampling of snRNAs from a wider temporal scale through development could permit higher resolution on the expression patterns of these sequences following key developmental stages. As mentioned above, understanding their expression patterns in adults would be critical in determining if these unexplained RNAs serve a biological function outside of PGR. As expression of these sequences could be germline limited (*a la* piRNAs), Locked Nucleic Acid probes could presumably be employed to track their expression *in situ* and ascertain expression patterns in different cell populations of rearranging embryos. Verification of a novel, germline limited class of snRNA expressed highly during PGR would be interesting considering regulation of PGR in the ciliate systems.

Ultimately the work presented here provides an initial glimpse into the regulatory framework surrounding lamprey PGR. Although much work remains to be done, these approaches can improve our understanding of small RNAs in a basal vertebrate, and could reveal the purpose or method behind lamprey's strategy of programmed genome rearrangement.

References:

1. Smith, J. J., et al. (2009). "Programmed loss of millions of base pairs from a vertebrate genome." *Proceedings of the National Academy of Sciences* 106(27): 11212-11217.
2. Cooper, M. D. and M. N. Alder (2006). "The evolution of adaptive immune systems." *Cell* 124(4): 815-822.
3. Mochizuki, K., et al. (2002). "Analysis of a piwi-Related Gene Implicates Small RNAs in Genome Rearrangement in *Tetrahymena*." *Cell* 110(6): 689-699.
4. Nowacki, M., et al. (2007). "RNA-mediated epigenetic programming of a genome-rearrangement pathway." *Nature* 450(7172).
5. Wang, J., et al. (2011). "Deep small RNA sequencing from the nematode *Ascaris* reveals conservation, functional diversification, and novel developmental profiles." *Genome research* 21(9): 1462-1477.
6. Morin, R. D., et al. (2008). "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells." *Genome research* 18(4): 610-621.
7. Wei, C., et al. (2012). "Transcriptome-wide analysis of small RNA expression in early zebrafish development." *RNA* 18(5): 915-929.
8. Graveley, B. R., et al. (2011). "The developmental transcriptome of *Drosophila melanogaster*." *Nature* 471(7339): 473-479.
9. Griffiths-Jones, S., et al. (2006). "miRBase: microRNA sequences, targets and gene nomenclature." *Nucleic acids research* 34(suppl 1): D140-D144.
10. Lewis, B. P., et al. (2003). "Prediction of mammalian microRNA targets." *Cell* 115(7): 787-798.
11. Fabian, M. R., et al. (2010). "Regulation of mRNA translation and stability by microRNAs." *Annu Rev Biochem* 79: 351-379.
12. Biemar, F., et al. (2005). "Spatial regulation of microRNA gene expression in the Drosophila embryo." *Proceedings of the National Academy of Sciences of the United States of America* 102(44): 15907-15911.
13. Cai, Y., et al. (2009). "A brief review on the mechanisms of miRNA regulation." *Genomics, Proteomics & Bioinformatics* 7(4): 147-154.
14. Winter, J., et al. (2009). "Many roads to maturity: microRNA biogenesis pathways and their regulation." *Nat Cell Biol* 11(3): 228-234.
15. Chen, C.-Z., et al. (2004). "MicroRNAs modulate hematopoietic lineage differentiation." *Science* 303(5654): 83-86.
16. Yi, R., et al. (2008). "A skin microRNA promotes differentiation by repressing 'stemness'." *Nature* 452(7184): 225-229.
17. Wienholds, E. and R. H. Plasterk (2005). "MicroRNA function in animal development." *FEBS letters* 579(26): 5911-5922.
18. Wienholds, E., et al. (2005). "MicroRNA expression in zebrafish embryonic development." *Science* 309(5732): 310-311.
19. Judson, R. L., et al. (2009). "Embryonic stem cell–specific microRNAs promote induced pluripotency." *Nat Biotechnol* 27(5): 459-461.
20. Esquela-Kerscher, A. and F. J. Slack (2006). "Oncomirs—microRNAs with a role in cancer." *Nature Reviews Cancer* 6(4): 259-269.
21. Kim, V. N., et al. (2009). "Biogenesis of small RNAs in animals." *Nature reviews Molecular cell biology* 10(2): 126-139.

22. Guzzardo, P. M., et al. (2013). "The piRNA pathway in flies: highlights and future directions." *Current opinion in genetics & development* 23(1): 44-52.
23. Aravin, A. A., et al. (2007). "The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race." *Science* 318(5851): 761-764.
24. Beyret, E., et al. (2012). "piRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism." *Cell research* 22(10): 1429-1439.
25. Houwing, S., et al. (2008). "Zili is required for germ cell differentiation and meiosis in zebrafish." *The EMBO journal* 27(20): 2702-2711.
26. Grimson, A., et al. (2008). "Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals." *Nature* 455(7217): 1193-1197.
27. Aravin, A. A., et al. (2008). "A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice." *Molecular cell* 31(6): 785-799.
28. Zamore, Phillip D., et al. "RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals." *Cell* 101.1 (2000): 25-33.
29. Haasnoot, Joost, Daniel Cupac, and Ben Berkhout. "Inhibition of virus replication by RNA interference." *Journal of biomedical science* 10.6 (2003): 607-616.
30. Lipardi, Concetta, and Bruce M. Paterson. "Identification of an RNA-dependent RNA polymerase in *Drosophila* involved in RNAi and transposon suppression." *Proceedings of the National Academy of Sciences* 106.37 (2009): 15645-15650.
31. Shirayama, Masaki, et al. "piRNAs Initiate an Epigenetic Memory of Nonself RNA in the *C. elegans* Germline." *Cell* 150.1 (2012): 65-77.
32. Pal-Bhadra, Manika, Utpal Bhadra, and James A. Birchler. "RNAi Related Mechanisms Affect Both Transcriptional and Posttranscriptional Transgene Silencing in *Drosophila*." *Molecular cell* 9.2 (2002): 315-327.
33. Mochizuki, K., et al. (2002). "Analysis of a *piwi*-Related Gene Implicates Small RNAs in Genome Rearrangement in *Tetrahymena*." *Cell* 110(6): 689-699.
34. Jahn, Carolyn L., and Lawrence A. Klobutcher. "Genome remodeling in ciliated protozoa." *Annual Reviews in Microbiology* 56.1 (2002): 489-520.
35. Fang, Wenwen, et al. "Piwi-Interacting RNAs Protect DNA against Loss during< i> Oxytricha</i> Genome Rearrangement." *Cell* 151.6 (2012): 1243-1255.
36. Juliano, Celina, Jianquan Wang, and Haifan Lin. "Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms." *Annual review of genetics* 45 (2011).
37. Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", *Nucleic Acids Research*, 37:W202-W208, 2009.
38. Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190, (2004).
39. Pearson, William R., et al. "Comparison of DNA sequences with protein sequences." *Genomics* 46.1 (1997): 24-36.
40. Morin, Ryan D., et al. "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells." *Genome research* 18.4 (2008): 610-621.
41. Langmead, Ben, et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome Biol* 10.3 (2009): R25.
42. Giraldez, Antonio J., et al. "Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs." *Science* 312.5770 (2006): 75-79.

43. Martello, Graziano, et al. "MicroRNA control of Nodal signalling." *Nature* 449.7159 (2007): 183-188.
44. Rosa, Alessandro, Francesca M. Spagnoli, and Ali H. Brivanlou. "The miR-430/427/302 family controls mesendodermal fate specification via species-specific target selection." *Developmental cell* 16.4 (2009): 517-527.
45. Gurtan, Allan M., and Phillip A. Sharp. "The role of miRNAs in regulating gene expression networks." *Journal of molecular biology* 425.19 (2013): 3582-3600.
46. Meister, Gunter. "Argonaute proteins: functional insights and emerging roles." *Nature Reviews Genetics* 14.7 (2013): 447-459.
47. Sabin, Leah R., M. Joaquina Delás, and Gregory J. Hannon. "Dogma derailed: the many influences of RNA on the genome." *Molecular cell* 49.5 (2013): 783-794.
48. Jurka, Jerzy. "Repbase update: a database and an electronic journal of repetitive elements." *Trends in genetics* 16.9 (2000): 418-420.
49. Lakshmi, S. Sai, and Shipra Agrawal. "piRNABank: a web resource on classified and clustered Piwi-interacting RNAs." *Nucleic acids research* 36.suppl 1 (2008): D173-D177.
50. Zhang, Yi, Xianhui Wang, and Le Kang. "A k-mer scheme to predict piRNAs and characterize locust piRNAs." *Bioinformatics* 27.6 (2011): 771-776.
51. Landgraf, Pablo, et al. "A mammalian microRNA expression atlas based on small RNA library sequencing." *Cell* 129.7 (2007): 1401-1414.
52. Denman, R. B. "Using RNAFOLD to predict the activity of small catalytic RNAs." *Biotechniques* 15.6 (1993): 1090-1095.
53. Kamminga, Leonie M., et al. "Hen1 is required for oocyte development and piRNA stability in zebrafish." *The EMBO journal* 29.21 (2010): 3688-3700.
54. Lehrbach, Nicolas J., et al. "LIN-28 and the poly (U) polymerase PUP-2 regulate let-7 microRNA processing in Caenorhabditis elegans." *Nature structural & molecular biology* 16.10 (2009): 1016-1020.
55. Friedländer, Marc R., et al. "Discovering microRNAs from deep sequencing data using miRDeep." *Nature biotechnology* 26.4 (2008): 407-415.
56. Untergasser, Andreas, et al. "Primer3Plus, an enhanced web interface to Primer3." *Nucleic acids research* 35.suppl 2 (2007): W71-W74.
57. Sambrook, Joseph, and W. Russell David. *Molecular cloning: a laboratory manual*. *Vol. 3*. Cold spring harbor laboratory press, 1989.
58. Tahara, Yutaka. "Normal stages of development in the lamprey, Lampetra reissneri (Dybowski)." *Zoological science* 5.1 (1988): 109-118.
59. Zhang, Haiyan, et al. "MiR-25 regulates apoptosis by targeting Bim in human ovarian cancer." *Oncology reports* 27.2 (2012): 594.
60. Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. "FIMO: scanning for occurrences of a given motif." *Bioinformatics* 27.7 (2011): 1017-1018.
61. Gilbert, S. F. "Early vertebrate development: mesoderm and endoderm." *Developmental Biology. 5th ed. Sinauer Assoc. Inc., Sunderland, MA* (1997): 341-357.
62. Snell, George D., and Leroy C. Stevens. "Early embryology." *Biology of the laboratory mouse* 2 (1966): 205-245.
63. Heimberg, Alysha M., et al. "microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate." *Proceedings of the National Academy of Sciences* 107.45 (2010): 19379-19383.

64. Rottiers, Veerle, and Anders M. Näär. "MicroRNAs in metabolism and metabolic disorders." *Nature Reviews Molecular Cell Biology* 13.4 (2012): 239-250.
65. Smith, Jeramiah J., et al. "Genetic consequences of programmed genome rearrangement." *Current Biology* 22.16 (2012): 1524-1529.
66. Lund, Elsebet, et al. "Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos." *Rna* 15.12 (2009): 2351-2363.
67. Burroughs, A. Maxwell, et al. "A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness." *Genome research* 20.10 (2010): 1398-1410.
68. Hiscott, John, et al. "Triggering the interferon response: the role of IRF-3 transcription factor." *Journal of interferon & cytokine research* 19.1 (1999): 1-13.
69. Gorovsky, Martin A. "Genome organization and reorganization in Tetrahymena." *Annual review of genetics* 14.1 (1980): 203-239.
70. Harding, J. L., et al. (2014). "Small RNA profiling of Xenopus embryos reveals novel miRNAs and a new class of small RNAs derived from intronic transposable elements." *Genome Res* 24(1): 96-106.
71. Röther, Susanne, and Gunter Meister. "Small RNAs derived from longer non-coding RNAs." *Biochimie* 93.11 (2011): 1905-1915.
72. Haussecker, Dirk, et al. "Human tRNA-derived small RNAs in the global regulation of RNA silencing." *Rna* 16.4 (2010): 673-695.
73. Lagos-Quintana, Mariana, et al. "Identification of tissue-specific microRNAs from mouse." *Current Biology* 12.9 (2002): 735-739.
74. Sood, Pranidhi, et al. "Cell-type-specific signatures of microRNAs on target mRNA expression." *Proceedings of the National Academy of Sciences of the United States of America* 103.8 (2006): 2746-2751.
75. Judson, Robert L., et al. "Embryonic stem cell–specific microRNAs promote induced pluripotency." *Nature biotechnology* 27.5 (2009): 459-461.

Vita

Author's Name: Joseph Herdy

Birthplace: Winchester, TN

Birthdate: July 22, 1990

Education:

Master of Science in Molecular Biology
University of Kentucky
August – 2014

Bachelor of Science in Biology
University of Alabama in Huntsville
May 2012

Research Experience:

University of Kentucky
Lexington, KY
May 2012 – June 2014
Graduate Research Assistant

University of Alabama in Huntsville
Huntsville, AL
July 2009 – May 2012
Undergraduate Researcher

Peer Reviewed Publications:

"A Sequence-based Genotyping Method to Distinguish Oplismenus hirtellus ss; undulatifolius from U.S. Native and Ornamental Species of Oplismenus", Leland J. Cseke, Shiloh R. McCollum, Andrew B. Adrian, Derek Thacker, Joseph R. Herdy, Sharon M. Talley. JoVE (2013).

Leland Cseke and Joseph Herdy, "Extraction/Characterization of DNA" Cell Biology Protocols Module, Methods Navigator, Ed. P. Michael Conn, Elsevier (2012).

Published Abstracts:

Joseph R. Herdy and Jeramiah Smith. Molecular Correlates of Programmed Genome Rearrangement. Plant and Animal Genome Conference XXI, San Diego CA (Jan 14, 2013).

Joseph R. Herdy and Jeramiah Smith. Small RNAs in Programmed Genome Rearrangement of a Vertebrate. Plant and Animal Genome Conference XXII, San Diego CA (Jan 14, 2014).