miRNA Regulation in Development

Sabah Kadri

January 2012
CMU-CB-12-100

Lane Center for Computational Biology
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Committee Members:
Veronica Hinman, Advisor
Panayiotis Benos, Advisor
Russell Schwartz
Kausik Chakrabarti
Javier Lopez

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This research was sponsored by the National Science Foundation under grant No. IOS 1024811 and the National Institutes of Health under grant No. R01LM007994, and R01LM009657. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, or any sponsoring institution, the U.S. government, or any other entity.

Keywords: microRNAs, echinoderms, HHMMiR, hierarchical hidden Markov model, NGS, *Dicer*, *Argonaute*, sea urchin, sea star, *Strongylocentrotus purpuratus*, *Patiria miniata*

To *Zakia & Shoeb Kadri.* I would not be here without you.

Abstract

microRNAs (miRNAs) are small (20-23 nt), non-coding single stranded RNA molecules that play an important role in post-transcriptional regulation of protein-coding genes. miRNAs have been found in all animal lineages, and have been implicated as critical regulators during development in multiple species. The echinoderms, *Strongylocentrotus purpuratus* (sea urchin) and *Patiria miniata* (sea star) are excellent model organisms for studying development due to their well-characterized transcriptional gene networks, ease of working with their embryos in the laboratory and phylogenetic position as invertebrate deuterostomes. Literature on miRNAs in echinoderm embryogenesis is limited. It has been shown that RNAi genes are developmentally expressed and regulated in sea urchin embryos, but no study in the sea urchin has examined the expression of miRNAs.

The goal of my work has been to study miRNA regulation in echinoderm developmental gene networks. I have identified developmentally regulated miRNAs in sea urchin and sea star embryos, using a combination of computational and wet lab experimental techniques. I developed a probabilistic model (named HHMMiR) based on hierarchical hidden Markov models (HHMMs) to classify genomic hairpins into miRNA precursors and random stem-loop structures. I then extended this model to make an efficient decoder by introduction of explicit state duration densities. We used the Illumina Genome Analyzer to sequence small RNA

libraries in mixed stage population of embryos from one to three days after fertilization of *S. purpuratus* and *P. miniata*. We developed a computational pipeline for analysis of these miRNA-seq data to reveal the miRNA populations in both species, and study their differential expression. We also used northern blots and whole mount in situ hybridization experimental techniques to study the temporal and spatial expression patterns of some of these miRNAs in sea urchin embryos. By knocking down the major components of the miRNA biogenesis pathway, we studied the global effects of miRNAs on embryo morphology and differentiation genes. The biogenesis genes selected for this purpose are the RNAse III enzyme, *Dicer* and *Argonaute*. *Dicer* is necessary for the processing of mature miRNAs from hairpin structures while *Ago* is a necessary part of the RISC (RNA interference silencing complex) assembly, which is required for the miRNA to hybridize to its target mRNA site. Knocking down these genes hinders normal development of the sea urchin embryo and leads to loss of the larval skeleton, a novel phenotype not seen in sea stars, as well as abnormal gastrulation. Comparison of differentiation gene marker expression between control and *Ago* knocked down sea urchin embryos shows interesting patterns of expansion and suppression of adjoining some embryonic territories, while ingression of larval skeletogenesis progenitors does not occur.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Firstly, I cannot overstate my gratitude to my advisors, Veronica Hinman and Panayiotis Benos, who have provided continuous support throughout the course of my graduate life. With their enthusiasm and support, I was able to take on and sustain a very challenging project. They have encouraged my research interests since I joined the PhD program, and have taught me how to shape my thoughts effectively and productively.

I would like to thank my thesis committee members, Russell Schwartz, Kausik Chakrabarti and Javier Lopez for their insightful feedback and support. I would especially like to thank Kausik for patiently guiding me with all my biochemical work, and motivating me to take up the immunoprecipitation part of my work. I have enjoyed all the discussions I have had with him. Russell has been a source of inspiration. He has guided me through my graduate life at CMU, and is one of the most caring teachers I am encountered.

I am truly grateful to my summer internship mentors, Erik Sassaman and Joseph Szustakowski. They have provided me with invaluable support and advice during and after my internship.

I would also like to acknowledge the members of the two labs I have been a part of for the past 4.5 years – the Hinman lab and the Benos lab. The Hinman lab members - Kristen Yankura, Brenna McCauley, Alys Cheatle, Stephanie Hughes, Walter Lewis, our alumni and our army of undergraduates, especially Laura Filliger and Sohee Jeon, I have enjoyed my time

working here and learnt a lot. Thanks to Kristen to carry out sea star library screens and to Brenna for sharing her differentiation gene probes. I would like to thank the Benos lab members – Rachel Brower-Sinning, Grace Huang, Claudia Coronnello, Tridib Dutta, Harry Athanassiou, Lucas Santana dos Santana, Abha Bais and Arshi Arora, for insightful discussions and fun times.

I would like to thank Calen Nicols at Wistar Institute for answering all my questions about Illumina sequencing, and for efficiently carrying out small RNA library preparations and sequencing. Zissimos Mourelatos has been extremely generous in sending us their human anti-Ago2 antibody from UPenn. I cannot thank Haibing and Lauren Ernst at MBIC enough for teaching me confocal microscopy and fluorometry, respectively. Their enthusiasm to share knowledge is infectious. Peter Yau at University of Illinois deserves gratitude for showing me how to analyze mass spectrometry results.

I would like to extend a special thanks to my colleagues and friends in the Joint CMU-Pitt PhD program in Computational Biology, for crazy fun times and deep conversations, and for being there when I needed them. I would not have made it without some of you.

Finally, and most importantly, thank you to my dearest family & friends – Mom & Dad my inspiration and my dearest baby brother, Bobby! My support system – my friends, both local and internationally, never left my side, in the best or worst of times. I love them.

1.0 INTRODUCTION: DEVELOPMENTAL BIOLOGY OF MICRORNAS AND ECHINODERMS

Transcription factors interact with *cis*-regulatory modules that control gene expression, to form gene regulatory networks (GRNs). A GRN can be viewed as a set of modules or sub-circuits that communicate with each other, through regulatory signals. In the context of GRNs, a module is defined as a set of transcription factors (nodes) that interact with each other, to execute a common function. The module has defined input and output nodes that control how it interacts with the rest of the network, but the genes of each module do not associate significantly with genes of other modules (Alon 2007). A complete developmental GRN specifies all interactions required to generate a cell type/fate. During development, most of the transcriptome becomes active, and modules within the network contribute to the many pathways that lead to differentiation of various cell lines. GRNs are typically studied at the transcriptional level of regulation. The developmental program, the process that creates a multi-cellular organism from a single cell, involves gene regulation at various levels – transcriptional, post-transcriptional and post-translational. That is, there are mechanisms of regulation in the cell, other than transcriptional, that affect the abundance and activity of the final gene products. *Post-transcriptional regulation of gene expression is one such class of regulation, affecting mRNA and protein levels in the cell.*

Recently, many classes of small RNAs working at the post-transcriptional level have been identified. These RNAs have been shown to have a range of functions, including fine-tuning gene expression, transposon silencing, and regulation of protein translation (Bushati & Cohen 2007). The main classes of these small RNAs are: microRNAs (miRNAs), endogenous small interfering RNAs (endo-siRNAs), and piwi-interacting RNAs (piRNAs). The distinction is mainly based on size, biogenesis pathways, and the particular *Argonaute* (*Ago*) protein with which they are associated (V. N. Kim et al. 2009). However, miRNAs are the most widely studied post-transcriptional regulators of gene expression, by far.

For this thesis, I will focus on miRNA-based regulation of transcription factors. One primary goal of this work is to investigate the role of the miRNA pathway in development. The goal of studying development along a timeline in a range of model organisms has been to extrapolate how homologous genes (coding or non-coding) might regulate other developmental events temporally, in a variety of other organisms. It is also interesting to study how differences in these genes, either spatially or temporally, may explain the variation between the species.

Two developmentally similar echinoderms - sea urchins and sea stars have been selected for this study. These species are particularly suited for the study and comparison of transcription-based regulation with post-transcriptional regulation, due to their phylogenetic position, well-characterized TF networks, and the ease of working with their embryos in the laboratory. See **Section 1.2**. Echinoderms and chordates make up the two major phyla under deuterostomia. Thus, these invertebrates diverged from the chordates much later than other invertebrate model organisms such as, arthropods and nematodes. Much is known about how their GRNs have evolved, and the consequences of this for their development. This will make the study of

miRNA-mediated regulation in context of a transcriptional GRN, more insightful and informative.

There has been much speculation about the role of miRNAs in development. Some may act as genetic switches, enforcing strong repression of one or few important targets and, thus, have a major impact on a biological pathway or process. Others may exhibit subtle effects by maintaining transcript and/or protein levels below a threshold. It is also possible that many miRNAs cause minor changes in the expression level of the same gene, but their cumulative effect has a stronger impact on cell fate. A well-characterized GRN will help de-convolute these paradigms of regulation.

I hypothesize that miRNAs play a crucial role as post-transcriptional regulators of gene expression in the development of echinoderms. Our goal has been to study if and how miRNAs interact with various pathways to determine cell fate in developing embryos. The study of evolution of these genes between two species under the same phylum is very insightful. It has been shown that miRNA target sites are gained and lost during evolution (K. Chen & Rajewsky 2006), thus, implying flexibility in the layer of miRNA regulation. This will be the first study that studies the development of miRNAs within the context of well-established pathways in development. It will, therefore, contribute to an understanding of the conservation of miRNA function, and the role miRNAs may play in shaping development.

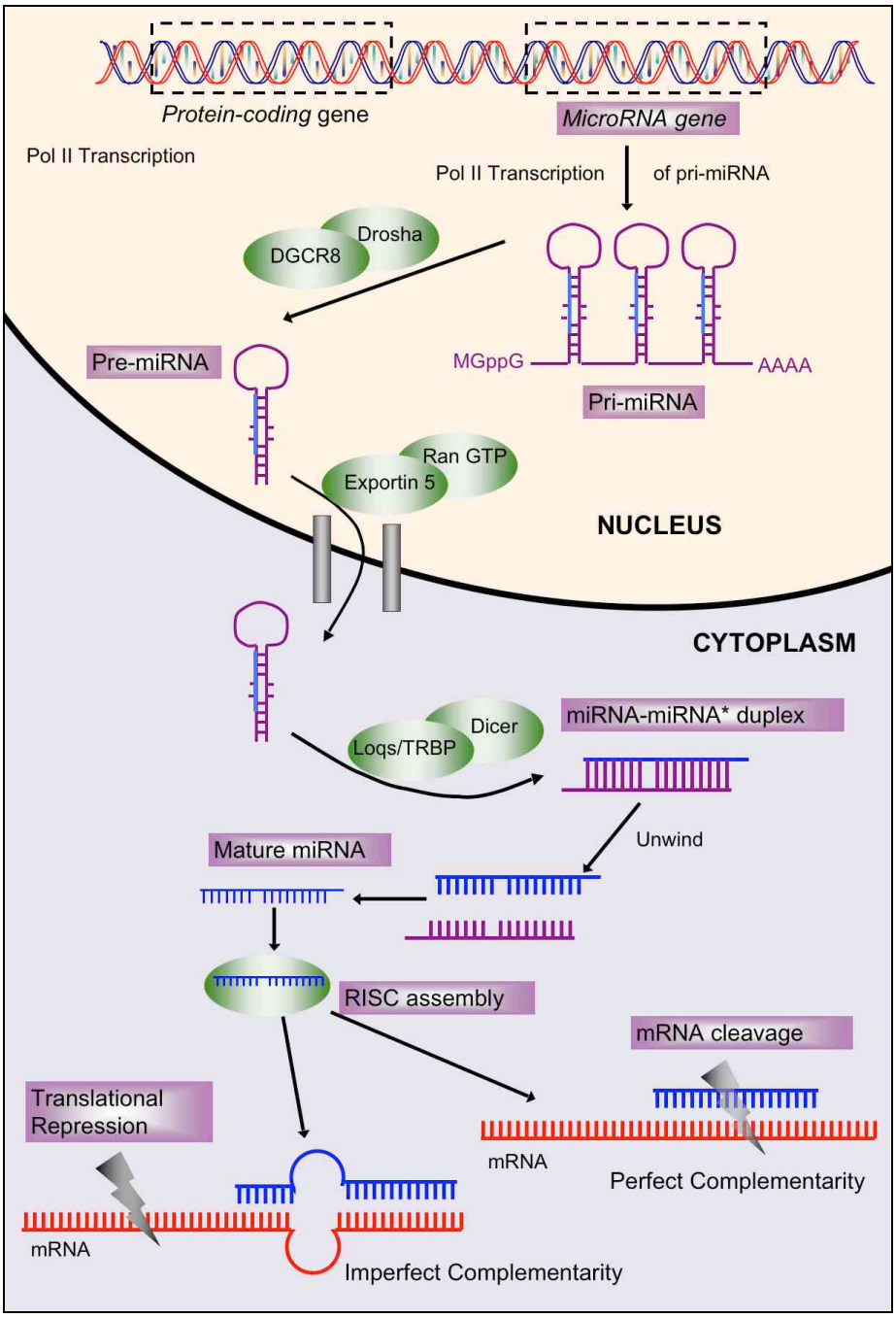


Figure 1.1: Biogenesis of miRNAs (Figure from (Kadri et al. 2009)): miRNA genes are transcribed in the nucleus, where they undergo processing by *DGCR8/Pasha* and the RNAse III family enzyme, *Drosha*. The pre-miRNA is then transported into the cytoplasm where it is processed by *Dicer*, and the cofactor *TRBP* to generate a ~22 nt miRNA: miRNA* duplex. After unwinding, the miRNA forms part of the RISC assembly and causes mRNA degradation or translational repression.

1.1 miRNAs

miRNAs belong to a class of small (~22 nts), non-coding RNA molecules that regulate protein coding gene expression post-transcriptionally.

1.1.1 Biogenesis Pathway

miRNAs gene are generally transcribed by RNA Polymerase II into a primary transcript (**Figure 1.1**) although some repeat-associated miRNAs are transcribed by RNA Polymerase III (Borchert et al. 2006), by their own promoters (Marson et al. 2008; Corcoran et al. 2009), or as parts of introns of protein coding genes (Baskerville & D. P. Bartel 2005; Ruby et al. 2007; Y.-K. Kim & V. N. Kim 2007). These primary transcripts (pri-miRNAs) can be several kilobases (kb) long, and contain local stem loop structures. The primary transcript is cleaved at the stem of the secondary structure to release the characteristic RNA hairpin structure by a complex composed of the RNase III-type protein Drosha and DiGeorge syndrome critical region gene 8 (DGCR8) (Pasha in Drosophila and *C. elegans*) proteins (V. N. Kim 2005). The product processed by the Drosha-DGCR8 complex, also called the Microprocessor complex is called the pre-miRNA, which is the precursor of the mature miRNA (**Figure 1.1**).

The ~70 nucleotide (nt) long pre-miRNA is exported to the cytoplasm by Exportin-5 via a Ran-GTP dependent mechanism, where an RNase III enzyme, Dicer, processes the pre-miRNA into a ~22 nt long duplex with 2nt overhangs at the 3'end (Hutvagner 2001; R F Ketting et al. 2001; D. P. Bartel 2004). This duplex contains the guide strand (mature miRNA) and the passenger strand (miRNA*). The mature miRNA typically has relatively higher steady-state levels than the corresponding miRNA*. However some miRNA* reach substantial levels and are

known to have regulatory roles (J.-S. Yang et al. 2010). *Dicer* is highly conserved and found in most eukaryotes. This gene has homologs in different species. *D. melanogaster* Dicer1 interacts with Loquacious (LOQS) whereas human Dicer interacts with TAR RNA-binding protein (TRBP) and PACT.

The mature miRNA is loaded onto an Argonaute (AGO) protein complex RNA-Induced Silencing Complex (RISC), while the miRNA* strand usually degrades (see **Figure 1.1B**). It has been shown that the selection of the strand is determined by its thermodynamic stability (Khvorova et al. 2003). Some organisms can have multiple AGO proteins, for example, humans have AGO1-4. Of these, only AGO2 has slicer activity.

1.1.2 miRNA Targeting mechanisms

The exact mechanism by which miRNAs regulate their targets is still unclear. In plants, most (but not all) miRNA regulation is done by mRNA cleavage whereas in animals, translational repression is more common (Millar & Waterhouse 2005). Animal miRNAs typically target 3' untranslated regions (UTRs) of protein coding genes, and usually down-regulate their expression by affecting their protein levels (Selbach et al. 2008), either by inhibiting mRNA translation, or by increasing its degradation rate (D. P. Bartel 2009; Chendrimada et al. 2005). Studies have recently shown miRNA-binding sites to be present in the 5' UTR and coding regions of genes as well (I. Lee et al. 2009; Lytle et al. 2007). There is mechanistic diversity in target regulation by miRNAs, including translational repression of target mRNAs, and regulation of mRNA decay (Millar & Waterhouse 2005). It has been predicted that 10-30% of protein-coding genes are regulated by miRNAs (Grün et al. 2005; K. Chen & Rajewsky 2006). Each individual miRNA can potentially target 200 or more transcripts (Krek et

al. 2005). Identification of miRNA targets in animals is a standing problem due to the degree of imperfect complementarity as well as little understanding of how the targets are identified.

Recent research has established the role of miRNAs in disease and developmental processes (Alvarez-Garcia & E. A. Miska 2005). As of December 2012, there are 295 human diseases associated with miRNAs in the human miRNA associated disease database (http://cmbi.bjmu.edu.cn/hmdd). In development, miRNAs that act as developmental switches of important TFs, can cause strong phenotypic changes, when knocked down or deleted (R. Lee 1993; B J Reinhart et al. 2000). Some miRNAs that cause drastic phenotypic effects do not act as switches, but their targets can act as switching genes (Bushati & Cohen 2007). (A switch target is one whose expression can be reduced to a level at which it loses its function, that is, it is switched off (D. P. Bartel & C.-Z. Chen 2004).) Other miRNAs stabilize the transcript and/or protein abundance thus fine-tuning the developmental programs. These miRNAs will have very subtle phenotypic changes, if at all (Giraldez et al. 2006). This function can reflect how miRNAs smooth out fluctuations in gene expression in the cells, or make sure that the expression levels of their targets are suitable to the conditions of the cell.

TFs are similar to miRNAs in many ways: both are developmentally regulated, act in *trans* to bind a specific *cis* target sequence, and act combinatorially as well as pleiotropically.

1.1.3 Evolution of miRNAs

Genes involved in the RNA interference (RNAi) pathways have been present in eukaryotes since early eukaryotic evolution (Cerutti & Casas-Mollano 2006). Since the discovery of the first miRNAs, lin-4 and let-7 in *C. elegans* (R. Lee 1993; B J Reinhart et al. 2000) , miRNAs have been identified in plants, animals, and viruses (Millar & Waterhouse 2005; Pfeffer

et al. 2005). Thus, miRNAs may have been regulating gene expression since early evolution. Some miRNA families are conserved throughout Bilaterians (Prochnik et al. 2007). However, miRNA evolution is dynamic and rapid. miRNAs evolve such that there are extensive lineage specific expansions. New miRNAs are continually discovered in various lineages. For example, 40% of primate miRNAs are specific and not found in other mammals (Sempere et al. 2006). Once gained, new miRNAs are usually maintained within the descendants of that lineage (B. M. Wheeler et al. 2009). But, do conserved miRNAs have conserved expression patterns across different species? Comparisons of the expression patterns of miRNAs among zebrafish, medaka fish, chick, and mouse, show conservation of expression of miRNA orthologs (Ason et al. 2006; Wienholds & R. H. A. Plasterk 2005; Gajewski et al. 2006; Christodoulou et al. 2010).

It is known from TF-based regulation that orthology of genes does not necessarily imply conservation of function. There is very little functional data available for such comparisons between orthologous miRNAs. Chen & Rajewsky (K. Chen & Rajewsky 2006) used computational target predictions to study the conservation of miRNA-mRNA interactions between three species - human, *C. elegans* and Drosophila. They showed ~10% of predicted targets of orthologous miRNAs to be conserved between humans and either *C. elegans* or Drosophila, but only 0.7% conserved between all three. Thus, they showed that despite conservation in miRNAs and 3'UTRs, the divergence in miRNA targets is very rapid. The few deeply conserved targets across human, *C. elegans* and Drosophila are enriched for essential developmental processes (K. Chen & Rajewsky 2006).

1.1.4 Developmental Role of miRNAs across multiple species

It has been suggested that miRNAs are largely involved in embryonic development and this hypothesis is supported by data in mouse and Drosophila (Yu et al. 2007). Various studies have been carried out to study the function of miRNAs in animal development (Alvarez-Garcia & E. A. Miska 2005; V. Ambros 2004; Bushati & Cohen 2007; Wienholds & R. H. A. Plasterk 2005). The role of miRNAs, *lin-4* & *let-7*) in **developmental timing** is already well established in *C. elegans* (R. Lee 1993; B J Reinhart et al. 2000). Even in the plant Arabidopsis, *Dicer-Like1* mutants have determined miRNAs as key regulators of embryo maturation (Willmann et al. 2011). Dicer is also essential for mouse oocyte maturation(Murchison et al. 2007).

Dicer mutant experiments have been carried out in various species, establishing the importance of the RNAi pathway in development. For example, in mice, loss of *Dicer1* is lethal to early mouse development and leads to depletion of embryonic stem cells (Emily Bernstein et al. 2003). In zebrafish, loss of *Dicer1* leads to arrested development (Wienholds et al. 2003). Loss of *Dicer* homolog, CARPEL FACTORY has pointed to critical roles for miRNAs during Arabidopsis development (Brenda J Reinhart et al. 2002).

miRNAs have also been established as key regulators of cell fate and differentiation (eg. neurogenesis, muscle differentiation etc.) (Ivey & Srivastava 2010). In zebrafish, *miR-430* has multiple roles including brain morphogenesis (Giraldez et al. 2006; Giraldez et al. 2005). On the other hand, studies have shown that miRNAs in zebrafish may not be involved in cell specification but maintenance of tissue identity instead, in later development (Wienholds et al. 2005). *miR-15* and *miR-16* miRNAs are involved in regulation of patterning by interaction with signaling cascades (Martello et al. 2007).

Studies have also shown that miRNAs play an important role in regulation of cell proliferation and cell death during development in Drosophila – for example, *bantam* targets *hid* in response to developmental signals (Brennecke et al. 2005) and *miR-14* targets *Drice* due to stress response (Xu et al. 2003).

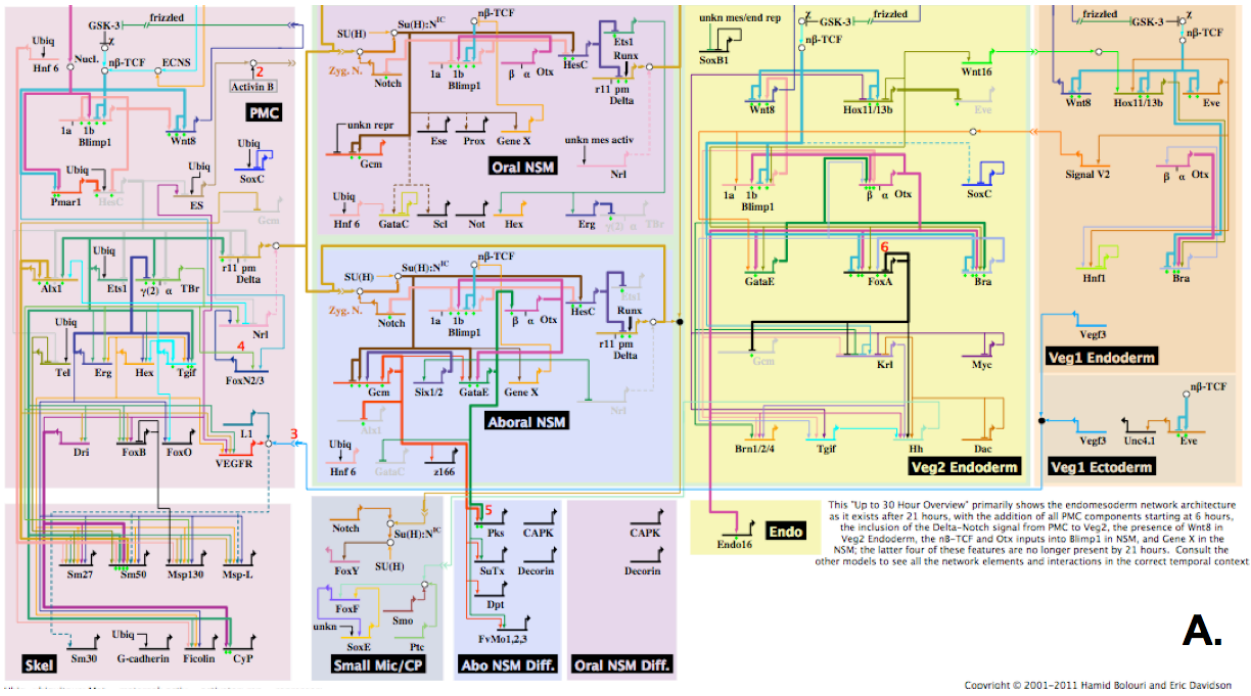
1.2 Developmental Biology of Echinoderms

1.2.1 Echinoderms are an excellent model system in developmental biology

The sea urchin, *Strongylocentrotus purpuratus* and the sea star, *Patiria miniata* are used as model organisms for developmental and evolutionary studies – from fertilization to morphogenesis. The reasons for this include their phylogenetic position (invertebrate deuterostomes), and their well-characterized transcription factor gene networks.

Echinoderms are basal deuterostomes that share a common ancestor with the chordates, and have an endoskeleton. The echinoderms are invertebrates in the lineage that leads to chordates, and are thus, the closest invertebrate outgroup to the chordates. This phylogenetic position makes this phylum extremely important from an evolutionary point of view.

The gene regulatory networks (GRNs) in the sea urchin continue to get increasingly detailed, with a complexity unmatched in other developmental model systems (**Figure 1.2A**). Much work has been done to identify regulatory modules that drive cell fate determination and specification in embryonic development of the sea urchin and sea star (D. R. McClay 2011; Eric H Davidson et al. 2002; V. F. Hinman et al. 2007; V. F. Hinman & Eric H Davidson 2007; Oliveri et al. 2002; Su 2009).



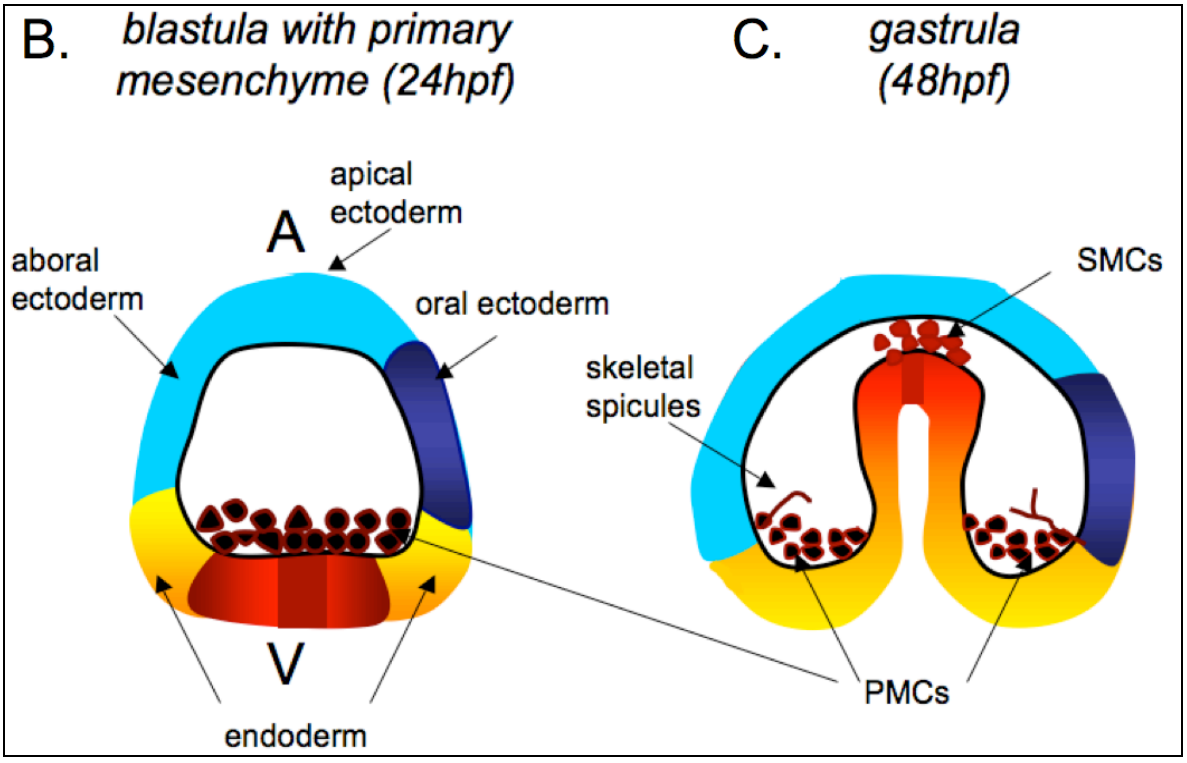


Figure 1.2: Complexity of GRNs in sea urchin development demonstrated using the endomesoderm network(A) and cartoon representation of the cell fate map during developmental stages of the sea urchin embryo (B-D) based on (Gilbert 2000). (B-C) Embryonic territories are color-coded similar to (Gilbert 2000). [hpf: hours post fertilization; A: Animal pole; V: Vegetal pole]

1.2.2 Morphology of the developing sea urchin

The sea urchin is an indirectly developing echinoderm, with larval stages and then morphogenesis into an adult. The developmental stages of its embryo are well defined. Sea urchin embryos consist of 10-15 cell types. **Figure 1.2B & C** shows a cartoon representation of the fate map of the sea urchin embryos for selected embryonic stages, derived from (Gilbert 2000). Most of the cell fates are specified by the 60-cell stage.

The blastula stage begins at the 128-cell stage, when cells form a hollow sphere surrounding a central cavity (blastocoel). Cells at the vegetal half of the blastula thicken and form the vegetal pole. The cells at the animal half secrete the hatching enzyme that digests the fertilization envelope. This stage is the *hatch blastula*.

At the mesenchyme blastula stage (**Figure 1.2B**), the primary mesenchyme cells (PMCs) ingress into the blastocoel (R. E. Peterson & David R McClay 2003; Wu et al. 2007). These micromeres undergo autonomous specification and later become skeletogenic mesenchyme. They induce the adjacent cells to become veg2 cells (Sherwood & D R McClay 1997), while veg2 cells induce the cells adjacent to them to become veg1 cells. As the PMCs ingress, cells from the vegetal plate invaginate into the blastocoel, thus, forming the archenteron (primitive gut). These endodermal cells adjacent to the PMCs invaginate the farthest and become the foregut. The next layer becomes the midgut and the last layer to invaginate becomes the hindgut (**Figure 1.2C**). The animal half gives rise to ectoderm – larval skin and neurons. The apical plate (apical ectoderm **Figure 1.2B-C**) develops within thickened epithelium at the animal pole of the embryo (Su 2009).

1.3 miRNAs in Echinoderms

Despite the intense research that has been devoted to their developmental transcriptional pathways (Oliveri et al. 2002; Eric H Davidson 2009; V. F. Hinman, A. T. Nguyen & Eric H Davidson 2003; V. F. Hinman, A. T. Nguyen, R Andrew Cameron, et al. 2003), little has been known about miRNA expression in these two organisms, especially during their early developmental stages. In early work, Pasquinelli *et al.* (Pasquinelli et al. 2000) examined the expression of the highly conserved *let-7* miRNA in 14 species from 8 phyla, and found that only sea urchin embryos lacked mature transcripts for the miRNA. More recently, Song *et al.* (Song & Wessel 2007) showed that the main genes involved in the RNAi pathway are expressed in sea urchin embryos, and Wheeler *et al.* (B. M. Wheeler et al. 2009) found 45 miRNAs to be expressed in the adult sea urchin using 454 sequencing. They also sequenced a species of sea star, *H. sanguinolenta* and found 42 miRNAs in this sea star adult. miRBase (v. 17, April 2011) contains 64 entries for *S. purpuratus* miRNAs (including miRNA* species) (B. M. Wheeler et al. 2009; Campo-Paysaa et al. 2011). Since developmental transcription factor gene networks are very detailed in these organisms (more than in any other echinoderm species), a systematic overlay of miRNA level regulation will provide invaluable insight into the cumulative effects of transcriptional and post-transcriptional regulation on developmental wiring.

1.4 Specific Aims

The goal of this project is to study the role of the miRNA layer of regulation involved in development, using echinoderms as a model system. To achieve this goal, we used probabilistic

modeling to improve miRNA precursor prediction and build a classifier without the requirement of conservation information. We also sequenced small RNA libraries to discover the miRNA populations that are developmentally expressed. Finally, we knocked down two key components of the miRNA biogenesis pathway and studied the effects on sea urchin development.

1.4.1 Improving miRNA precursor predictions using a probabilistic framework without requirement of conservation data

Most miRNA prediction methods require sequence conservation from another closely related species. We developed a new classification method to predict miRNA genes (pre-miRNAs) from the single-loop hairpins of the genome, without use of evolutionary information. The miRNA gene classification approach was based on a probabilistic framework, using hierarchical hidden Markov models (HHMMs) (Fine et al. 1998). Our method is called HHMMiR. The distinct regions of the secondary RNA structure of the miRNA precursor were used to define the states of HHMMiR. We added explicit state duration densities to improve decoding efficiency of the model. See **Chapter 2.0** for a detailed description.

1.4.2 Identification and analysis of the small RNA populations involved in development of sea urchin and sea star embryos.

We examined the subset of miRNAs expressed during the developmental stages of sea urchin and sea star, and their relative abundance, by processing and analyzing deep sequencing reads from small RNA libraries. We wanted to understand which miRNAs are expressed in developing

embryos, and the spatial and/or temporal profiles of the most important miRNAs. This is the first step towards studying the function of this class of small RNAs in development of echinoderms.

We developed a computational pipeline for conserved and novel miRNA discovery from the Illumina reads. This will involved pre-processing steps to filter out spurious reads and noise, map the reads to genome sequence (if available), and remove other noncoding RNAs (tRNAs, rRNAs etc.) and degradation products from the filtered reads. We used BLAST (Altschul et al. 1990) to search for conserved miRNAs (with stringent parameters), and the secondary structure of the genomic region flanking the read, was checked for pre-miRNA structural characteristics. We used miRDeep (M. R. Friedländer et al. 2008) to discover novel miRNAs in the reads. Thus, the pipeline will be able to discover not only conserved, but also novel miRNAs. Some miRNAs were also selected for experimental validation. See **Chapter 3.0** for more details.

1.4.3 Study the effects of miRNA function knock down during sea urchin development.

We knocked down two key components of the miRNA biogenesis pathway in the sea urchin – *Dicer* and *Argonaute*, to study their effects in development. Normal development of the sea urchin is hindered with suppressed miRNA function. Morphologically, larval skeletogenesis is blocked and the gut formation is abnormal. We also used differentiation gene markers to study known pathways that might be downstream of the miRNA pathway.

For the future, one may use a high throughput approach, called HITS-CLIP to immunoprecipitate the RISC complex and sequence the interacting miRNA and mRNA populations. Preliminary work towards testing an antibody for this purpose is presented.

2.0 EFFICIENT DE NOVO PREDICTION OF MIRNAS USING HIERARCHICAL HIDDEN MARKOV MODELS

The first animal miRNA genes, let-7 and lin-4, were discovered in *Caenorhabditis elegans* by forward genetics (R. Lee 1993; Wightman et al. 1993; B J Reinhart et al. 2000). But this method is relatively inefficient for recognition of miRNAs on a genome-wide scale. Currently, miRNA genes are biochemically identified by cloning and sequencing size-fractionated cDNA libraries. This method has limitations as well, because some miRNAs may be expressed at very low levels in a particular cell type or developmental stage; they may also be difficult to clone. Deep sequencing is being used on a large scale to identify small non-coding RNAs in the genome, but this is a relatively expensive method (although the cost is reduced as the technology advances) and can only identify miRNAs expressed in a single cell type or under a given condition. Computational methods are fast and inexpensive and a number of approaches have been developed to predict miRNA genes, genome-wide.

The identification of miRNA genes in newly sequenced organisms is still based, to a large degree, on extensive use of evolutionary conservation, which is not always available. We have developed HHMMiR, a novel approach for *de novo* miRNA hairpin prediction in the absence of evolutionary conservation. Our method implements a Hierarchical Hidden Markov Model (HHMM) that utilizes region-based structural as well as sequence information of miRNA precursors. We first established a template for the structure of a typical miRNA hairpin by

summarizing data from publicly available databases. We then used this template to develop the HHMM topology.

Our algorithm achieved average sensitivity of 84% and specificity of 88%, on 10-fold cross-validation of human miRNA precursor data. We also show that this model, trained on human sequences, works well on hairpins from other vertebrate as well as invertebrate species. Furthermore, the human trained model was able to correctly classify ~97% of plant miRNA precursors. The success of this approach in such a diverse set of species indicates that sequence conservation is not necessary for miRNA prediction. This may lead to efficient prediction of miRNA genes in virtually any organism. By adding explicit state duration densities to the internal states of HHMMiR, we were able to improve the efficiency of the model as a decoder of the hairpins. Most of the HHMMiR part of this chapter is taken from (Kadri et al. 2009).

2.1 Introduction

2.1.1 Previous computational miRNA prediction methods

Most computational approaches for miRNA prediction depend heavily on conservation of hairpins in closely related species (Lim et al. 2003; Ohler et al. 2004; Grad et al. 2003; Eric C Lai et al. 2003). Some methods have used clustering or profiling to identify miRNAs, (Sewer et al. 2005; Legendre et al. 2005; Ohler et al. 2004). The approach by Bentwich *et al.* (I. Bentwich et al. 2005) is interesting in that the whole genome is folded and scores are assigned to hairpins based on various features, including hairpin structural features and folding stability (no use of evolutionary conservation).

Machine learning approaches in the past have used support vector machines with high dimensional basis functions for classification of genomics hairpins (Sewer et al. 2005; Pfeffer et al. 2005; Xue et al. 2005). Some of these methods depend on cross-species conservation for classification, while others do motif finding using multiple alignments. See **Section 2.2.5** for further comparisons.

2.1.2 Hierarchical Hidden Markov Models

Hierarchical Hidden Markov Models (HHMMs) constitute a generalization of Hidden Markov Models (HMMs). They have been successfully used for modelling stochastic levels and length scales (Fine et al. 1998). An HHMM has two types of states: internal states and production states. Each internal state has its own HHMM but cannot emit symbols by itself. It can activate a sub-state by a vertical transition. Sub-states can also make vertical transitions, until the lowest level in the hierarchy (production state) is reached. Production states are the only states that can emit symbols from the alphabet via their own probability distributions. Sub-states at the same level of hierarchy will be activated through horizontal transitions till an "end state" is reached. Every level has only one "end state" for each parent state that shifts control back to the parent. Thus, each internal state can emit sequences instead of single symbols. The node at the highest level of the hierarchy is called the "root" node while the leaf nodes are the productions states. Please refer to *Methods* for information about HHMM parameters and their estimation.

2.1.3 Data summarization

We consider the hairpin stem-loop for predictions since it is structurally, the most prominent feature during biogenesis (**Figure 1.1**). miRNA genes can be divided into four regions depicted in **Figure 2.1a**. After transcription, the RNA strand folds to form the hairpin precursor (**Figure 1.1** and **Figure 2.1a**). The "loop" is the bulged end of the hairpin. The "miRNA" region defines the miRNA-miRNA* duplex (sans the 3' overhangs) that is processed by Dicer and further unwound. The region of the precursor extending from the end of the loop to the "miRNA" region is called the "extension". This region can be of variable length. The part of the hairpin sequence beyond the "miRNA" region may be part of the pri-miRNA in the nucleus and processed by *Drosha*. Thus, it has been named as "pri-extension", as suggested in Saetrom et al. (Saetrom et al. 2006).

The results presented in **Table 2.1** and **Figure 2.4** show that the differences that exist between vertebrate and invertebrate miRNA genes are rather small. So, a probabilistic method trained in data from one organism is likely to be able to perform well in another organism. As evident from the results in **Table 2.1**, the differences between length distributions of plant and animal precursors are relatively drastic, with the former having longer extension regions. However, overall, the length distributions in each of these regions for representative species (vertebrate, nematode, insect & plant) seem to be similar (**Figure 2.7**). The lengths of miRNAs are however, conserved across the kingdoms and so are those of the loops. More information about species-specific differences is provided as **Appendix B** and in **Figure 2.7**. These genomes constitute an excellent test set for our algorithm in that they span various kingdoms, with different miRNA characteristics. Thus, it will be very useful to see how well an HMMM trained

on (say) human sequences will be able to predict miRNA stem-loops in another vertebrate or invertebrate species and plants.

Table 2.1: Characteristics of miRNA hairpins in various taxa. HP: Hairpin length; LP: Loop length; MIR: MiRNA length; EXT: Distance of miRNA duplex from end of loop; PRI: Length of extension from end of miRNA to end of precursor. A list of organisms used for this Table is provided in **Appendix B**.

	HP	LP	MIR	EXT	PRI
Mean					
Vertebrates	86.7	7.3	22.0	5.0	12.6
Invertebrates	91.8	7.9	22.2	5.8	13.8
Plants	125.4	6.5	21.2	25.5	12.8
Std. Dev.					
Vertebrates	13.8	3.5	0.9	3.4	7.0
Invertebrates	13.1	3.9	1.3	4.5	5.9
Plants	43.2	3.6	1.0	18.4	10.3
Minimum					
Vertebrates	55	3	16	0	0
Invertebrates	54	3	18	0	0
Plants	58	3	17	0	0
Maximum					
Vertebrates	153	22	26	34	50
Invertebrates	215	30	28	55	32
Plants	545	35	24	102	78

2.2 HHMMiR

2.2.1 The HHMM model

HHMMiR is built around the miRNA precursor template illustrated in **Figure 2.1a**. The figure presents the four characteristic regions of stem-loop of a typical miRNA gene as described above. The length distributions of each of these regions are derived from **Table 2.1**. Each region, except the loop itself has three states: *match*, *mismatch*, and *insertion/deletion* (*indel*). *Match* means a base pairing at that position in the stem-loop, while *mismatch* means bulges on both arms at that position in the folded hairpin. The loop will only have the *indel* state. Examples of these states are presented in **Figure 2.1a**.

The HHMM resulting from this scheme has three levels (**Figure 2.2**). *Hairpin* is the root node and can vertically transition to its *Loop* sub-state only. In our model, every hairpin begins with a loop. The four internal states at the second level correspond to the four main regions of the hairpin from **Figure 2.1a**. This level also has an *End* (L_{end}) state to transfer control back to the *Hairpin*. Each internal state has a probabilistic model at the next lower level. A *Loop* cannot have base pairs and thus, has only one sub-state: I (*Indel*). The *Extension* state can only emit an M (*match*) state, when entered, since a mismatch or indel would become a part of the loop. The *miRNA* and *Pri-Ext* states can begin with a match, mismatch or indel. Each of these states has an End state (L_{end}, R_{end}, P_{end} respectively) (see **Figure 2.2**).

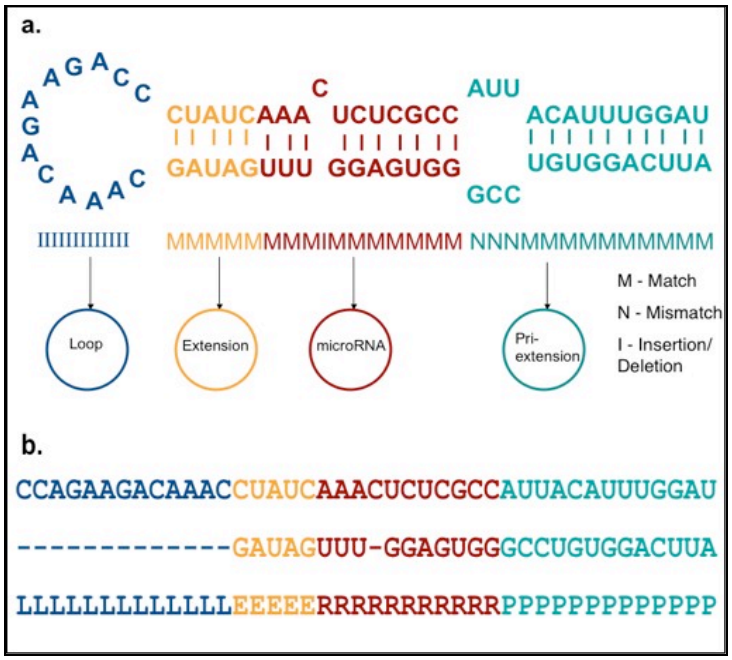


Figure 2.1: The miRNA hairpin: (a) *Template:* In our model, the miRNA precursor has four regions- "Loop" is the bulge and outputs *indels* only; "Extension" is a variable length region between the miRNA duplex and the loop; "microRNA" represents the duplex, without 3' overhangs; "Pri-extension" is the rest of the hairpin. The latter three regions can output *matches, mismatches* and *indels*. (The nucleotides distribution and lengths are not to scale) (b) *Labeled precursor*: The precursor shown in (a) is labeled according to the regions it represents. This is the input format of training data for HHMMiR. L: Loop; E: Extension; R: MiRNA; P: Pri-miRNA.

2.2.2 Datasets and Alphabet Selection

The training dataset contained a total 527 human miRNA precursors (positive dataset) and ~500 random hairpins (negative dataset), based on criteria derived from summarization (see **Methods**). The *RNAfold* program from Vienna Package (Hofacker 2003) was used to obtain the secondary structure of these hairpins with the minimum fold energy (*mfe*). The parameters of the model were estimated using a modified Baum-Welch algorithm (see **Data Collection and**

Processing for details on data sets and algorithms). All tests were conducted with 10-fold cross validation with random sampling.

We tested our model on two alphabets: $\sum_1$ with *matches* M = {AU, GC, GU}, *indels* I = {A-, G-, C-, U-} and *mismatches* N_1 = {AA, GG, CC, UU, AC, AG, CU}; and $\sum_2$, which is similar to $\sum_1$ except that the mismatch set is more concise: N_2 = {XX, XY}, where XX stands for one of {AA, GG, CC, UU} and XY stands for one of {AC, AG, CU}. In our alphabet, a *match*, say, AU has the same probability as UA, that is, an 'A' on either stem base paired with 'U' on the other stem. Thus, $\sum_1$ uses sequence as well as structure information for mismatches, where as $\sum_2$ eliminates sequence information for mismatches and only considers structure information. Cross-validation tests using MLE showed that the model with alphabet $\sum_1$ performed substantially better, both in terms of sensitivity and specificity (**Table 2.2**) (see **Parameter Estimation & Testing** for more details on these calculations).

It is surprising that $\sum_1$ performs better than $\sum_2$, because one would expected that mismatches in the stem-loop would not be characteristic of the miRNA sequence, since they do not contribute to the base pairing of the stem and thus the overall folding energy, on which other algorithms are based (I. Bentwich et al. 2005). Furthermore, $\sum_1$ alphabet has more parameters. In order to rule out that the better performance is due to parameter over fitting, we repeated training with multiple datasets of different sizes and the results remained the same (data not shown). In the remaining of this chapter, we use the $\sum_1$ alphabet.

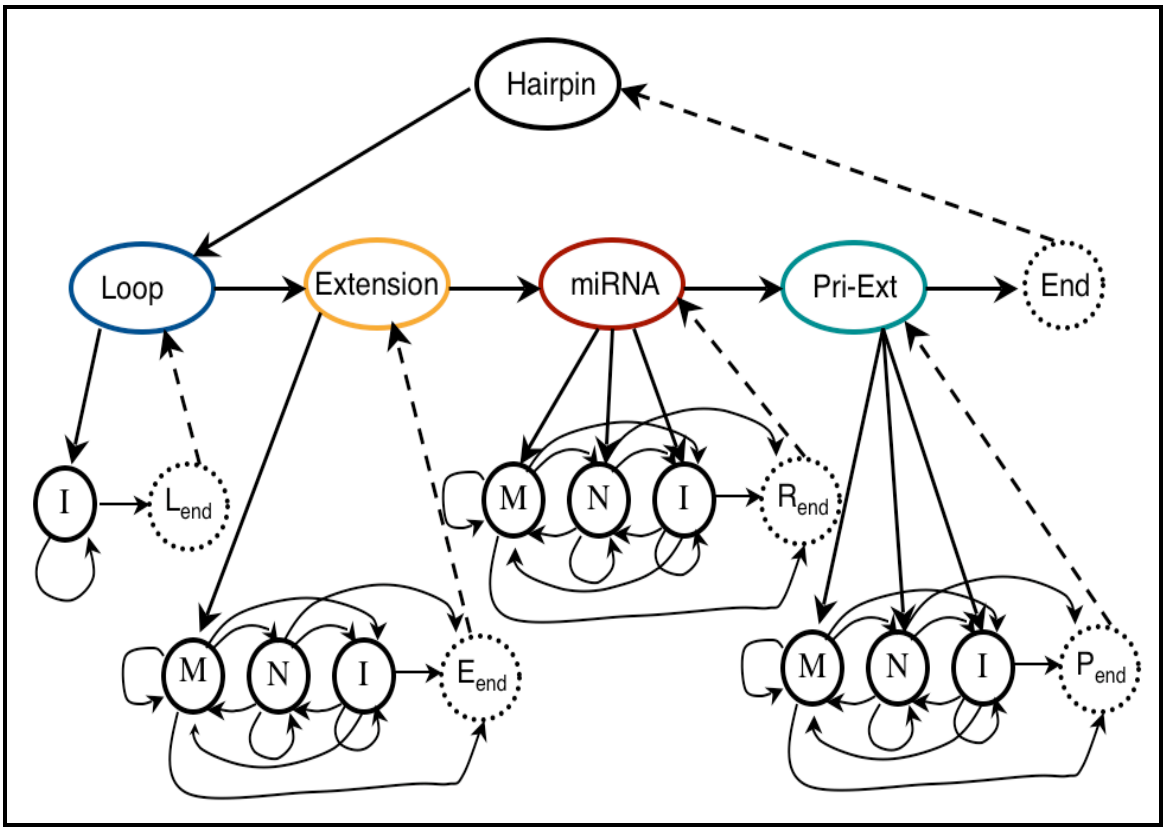


Figure 2.2: The HHMM state model (based on the microRNA hairpin template): The oval shaped nodes represent the *internal states*. The colors correspond to the biological region presented in Figure 2.1a. The circular solid lined nodes correspond to the *production states*. The dotted lined states correspond to the silent end states. M: *Match* states, N: *Mismatch* states, I: *Indel* states, L_{end}: Loop end state, R_{end}: miRNA end state, P_{end}: pri-extension end state.

Table 2.2: Results for different alphabet sizes: $\sum_1$ (larger alphabet) shows better accuracy than $\sum_2$ (smaller alphabet); Sn: Sensitivity; Sp: Specificity; FDR: False Discovery rate. All numbers are in percentages.

Alphabet	Sn	Sp	FDR
$\sum_1$	74.5	94.1	15.8
$\sum_2$	55.0	48.5	51.0

2.2.3 Training Algorithms: Performance Evaluation

We implemented and compared variations of two existing algorithms for parameter estimation: Baum-Welch and Maximum Likelihood Estimate (MLE). The positive model was trained using MLE since training data (stem-loop hairpins) can be labeled as *loop, extension, miRNA* and *pri-extension* (**Figure 2.1b**) using existing annotations. Negative data on the other hand, are obviously unlabelled, so both algorithms were compared for training this dataset. We will call the MLE trained model, MLE-HHMMiR whereas the Baum Welch trained model will be called BW-HHMMiR for this evaluation. For MLE-HHMMiR, we used length distributions from database summarization (**Table 2.1**) to perform *random labeling* of the four regions on the negative datasets. Overall, we found Baum-Welch performed similar to MLE (and slightly better). The area under the ROC curve for the MLE-HHMMiR is 0.912 whereas for BW-HHMMiR is 0.920 (**Figure 2.3**). The ratio of the log-likelihoods output by the two models decides the fate of the test hairpin. In order to decide a threshold for this ratio, the trade-off between sensitivity and specificity was considered by calculating the *Mathews correlation coefficient* (**Table 2.2**).

The highest MCC value was 0.73 for BW-HHMMiR and 0.71 for MLE-HHMMiR, and thus, these ratios were fixed at 0.71 and 0.99, respectively. An average 84% sensitivity and 88% specificity was achieved. Even though, the difference between the two algorithms is not much, we choose BW-HHMMiR for further tests. This is because MLE-HHMMiR depends on *random labeling* of hairpins and thus, performance will vary according to the labeling. In order to account for the absence of certain base pairs or *indels* in a certain sequence while using Baum-Welch, we introduce pseudo-counts to correct for the same.

Table 2.3: Results for cross-validation using different algorithms: FDR: False Discovery Rate; SD: Standard Deviation. All numbers are in percentages.

Method	Sensitivity		Specificity		FDR	
	Mean	SD	Mean	SD	Mean	SD
Baum-Welch	84.0	18.6	88.0	6.6	11.8	5.6
MLE	74.5	13.7	94.1	2.7	15.9	8.0

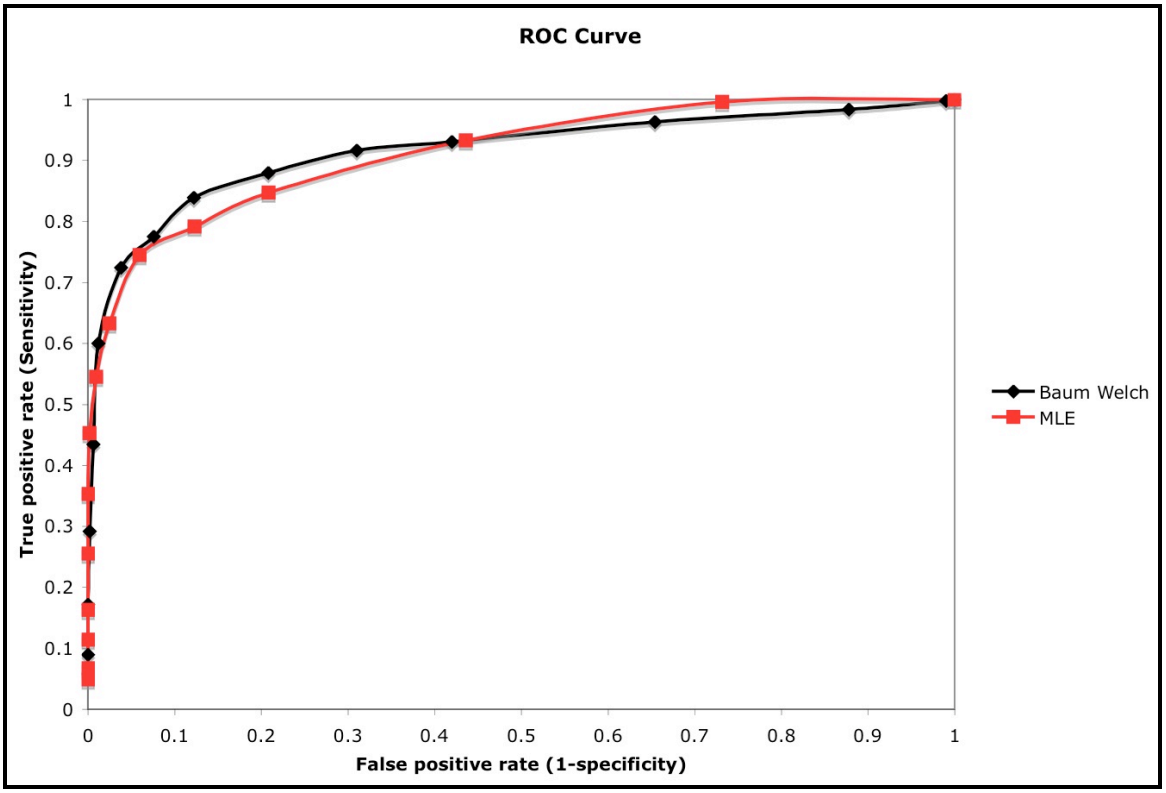


Figure 2.3: ROC curves for Baum-Welch and MLE training on the negative model: 10-fold cross-validations used with Baum-Welch (*black curve*) and MLE (*red curve*) for training the negative model. Positive model was trained using MLE in both cases.

2.2.4 Testing prediction efficiency in other organisms

Next, we examined how well our model trained on human sequences could predict known miRNAs in other species. The selected species were chosen as representatives of their respective phyla. In particular, HHMMiR was tested on the following: *M. musculus* (mammal), *G. gallus* (birds), *D. rerio* (fish), *C. elegans* (worms), *D. melanogaster* (flies), *A. thaliana* and *O. sativa* (plants). All these species are well studied and annotated. The results are shown in **Table** 2.4. HHMMiR is able to predict 85% of most animal precursors. Its overall sensitivity was also about 85%. What is more surprising, however, is the higher performance we observe in prediction of plant precursors, given the differences in length distributions of the miRNA stem-loops between plants and animals (**Table 2.1**). The fact that mouse miRNAs are predicted at lower rate probably reflects the larger number of hairpins known for this species. The specificity over the mouse data is also very high (84%) and remains surprisingly high in the two invertebrate species (~75%).

The reason for a good cross-species performance could be attributed to similar trends seen in the trained parameters across species at varying evolutionary distances (**Figure 2.4**). We see similar trends in the emission probabilities across human, insect, worm and plant data (representative species shown in **Figure 2.4**), with a preference for As and Us in the *indels*, limited G:U wobble, and U:U, A:C and A:G are more probable as the *mismatches* across all regions of the hairpin. However, some differences can also be seen. For example, G:C base pairs are more probable than A:U base pairs in humans than the other three species. Some subtle differences can be seen across the four regions, with some base pairs being more probable in the miRNA region than in the other regions.

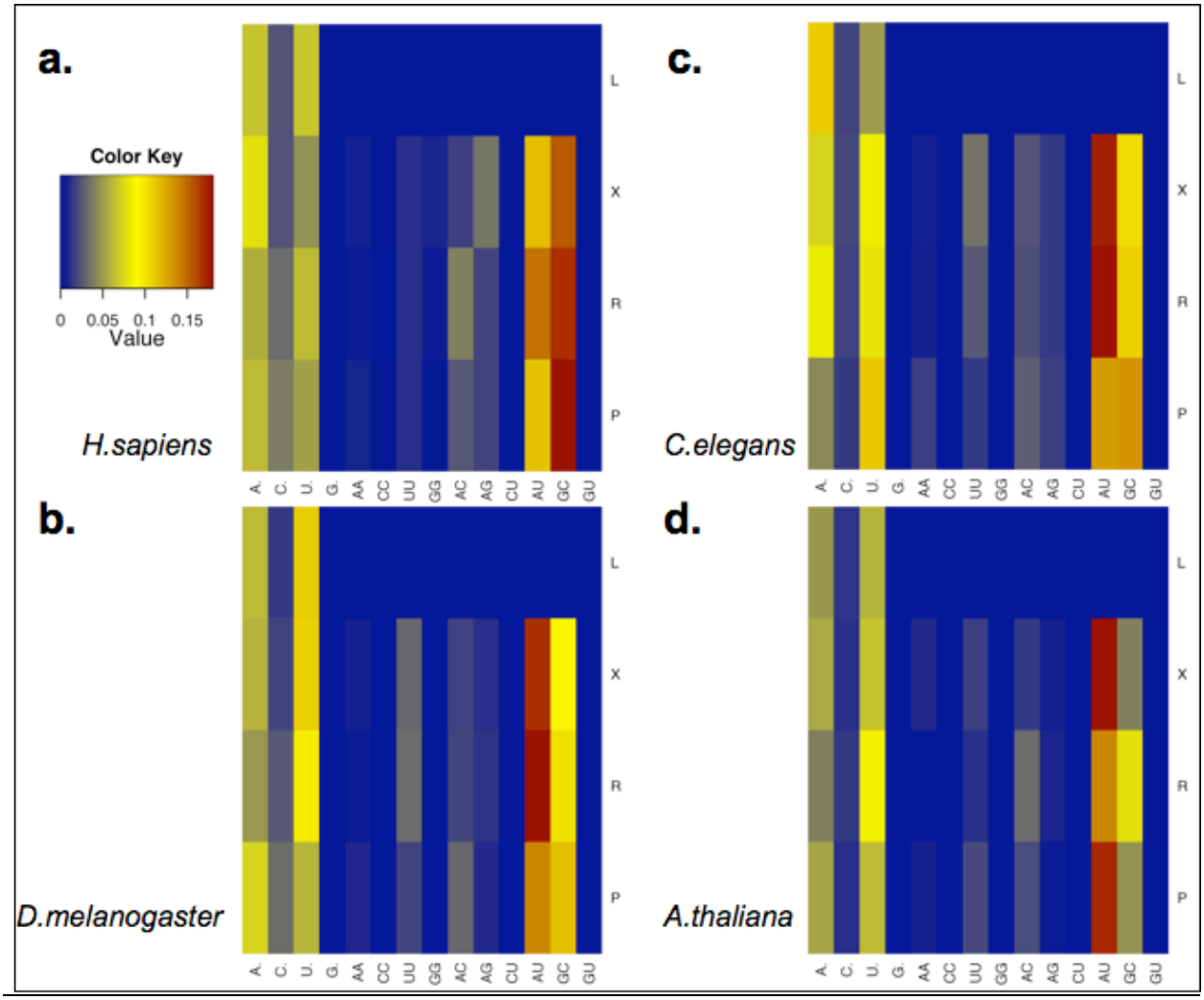


Figure 2.4: Emission probabilities across multiple species: A heat map of the emission probabilities for each base pair in the alphabet is shown for (a) human (b) an insect (*D. melanogaster*) (c) a nematode (*C. elegans*) and (d) a plant (*A. thaliana*). The rows represent the distinct regions of a typical miRNA hairpin as shown in **Figure 2.1**. *L: Loop, X: Extension, R:miRNA, P:Pri-extension.* (A., C., U., G. are *indels*; AA, CC, UU, GG, AC, AG, CU are *mismatches*; AU, GC, GU are *matches*)

Table 2.4: Results of tests on other species.

Organism	Total hairpins	% correctly predicted
M. musculus	422	74.7
G. gallus	147	89.1
D. rerio	334	88.3
C. elegans	131	85.5
D. melanogaster	143	93.0
A. thaliana	114	97.4
O. sativa	188	85.7
Total	**1479**	**85.1**

Table 2.5: Results for comparison between two precursor prediction methods: The percentages represent the ratio of hairpins correctly predicted.

Test set		Total hairpins	% correctly predicted
Positive Sets			
New human hairpins in registry at the time.	39	92.3	97.4
M. musculus	36	94.4	88.9
R. norvegicus	25	80.0	84.0
G. gallus	13	84.6	100
D. rerio	6	66.7	100
C. elegans	110	86.4	90.9
C. briggsae	73	95.9	95.9
D. melanogaster	71	91.6	95.8
D. pseudoobscura	71	90.1	98.6
A. thaliana	75	92.0	97.3
O. sativa	96	94.8	86.5
Epstein Barr virus	5	100	80.0
TOTAL	**620**	**91**	**93.2**
Negative Sets			
Folded genome hairpins from Chromosome 19	2444	89	88.6
Negative hairpin Set	1000	88.1	89.4
TOTAL	**3444**	**88.7**	**88.8**

2.2.5 Comparison with other approaches

As described earlier, there are very few machine learning methods that do not require evolutionary information to predict miRNAs. For example, Nam *et al.* (Nam et al. 2005) presented another probabilistic model, which is a motif finding method for mature miRNA region prediction. An SVM-based approach has been proposed (Xue et al. 2005) that parses the *mfe* structure in "triplets": structural information about the pairing states of every three nucleotides, represented using dot-bracket notation. This method showed an accuracy of ~90% using the data available in the registry at the time. We used the training and test sets used by the "triplet SVM" to train and test our model, HHMMiR, and we found it to perform better in almost all datasets (**Table 2.5**). The only exceptions are the mouse (but not rat) and Arabidopsis (but not rice). Also, their model was able to predict all the five then known miRNAs from Epstein-Barr virus, whereas HHMMiR predicted four. *Overall, HHMMiR exhibits sensitivity of 93.2% and specificity of 89% in these datasets.*

2.2.6 Methods

2.2.6.1 Data Collection and Processing

miRNA genes were obtained from the microRNA registry, version 10.1 (December 2007) (Griffiths-Jones 2006), which contains 3265 miRNAs from animals and 870 from plants. For training HHMMiR, we used the residual 527 human hairpins, after filtering out precursor genes with multiple loops. Each gene was folded with the RNAfold program, which is part of the Vienna package (Hofacker 2003), using the default parameters to obtain the secondary structure

with minimum fold energy. The negative set consists of coding regions and random genomic segments from the human genome that were obtained using the UCSC genome browser(Kent et al. 2002). These regions were folded and processed as described below.

Genomic sequences were folded in windows of 1 kb, 500nts and 300nts with an overlap of 150nts between consecutive windows. Nodes from the TeraGrid project (Catlett et al. 2007) were used for the genome folding. We tested the various window sizes on the relatively small C. elegans genome. We discovered that 500nts windows cover most known miRNA hairpins. Windows of 300nts exhibited high degree of redundancy without adding more hairpins to those of the 500nts windows, while 1kb windows missed a higher percentage of known miRNAs (data not shown). For this study, we used hairpins extracted from 500nts windows. We were able to recover ~92% of the known miRNAs from *C. elegans* in this way. The remaining 8% may have been accounted for by existence of multiple loops or specificity of the parameters used. The hairpins were extracted from these folded windows using the following parameters: each hairpin has at least 10 base pairs, has a maximum length of 20 bases for the loop, and a minimum length of 50 nucleotides. The data flow of this process is presented in **Figure 2.5**.

After the hairpins are extracted, we process them to an input format representing the hairpin's secondary structure (**Figure 2.1b** & **Figure 2.5**) to be compatible with the HHMM shown in **Figure 2.2**.

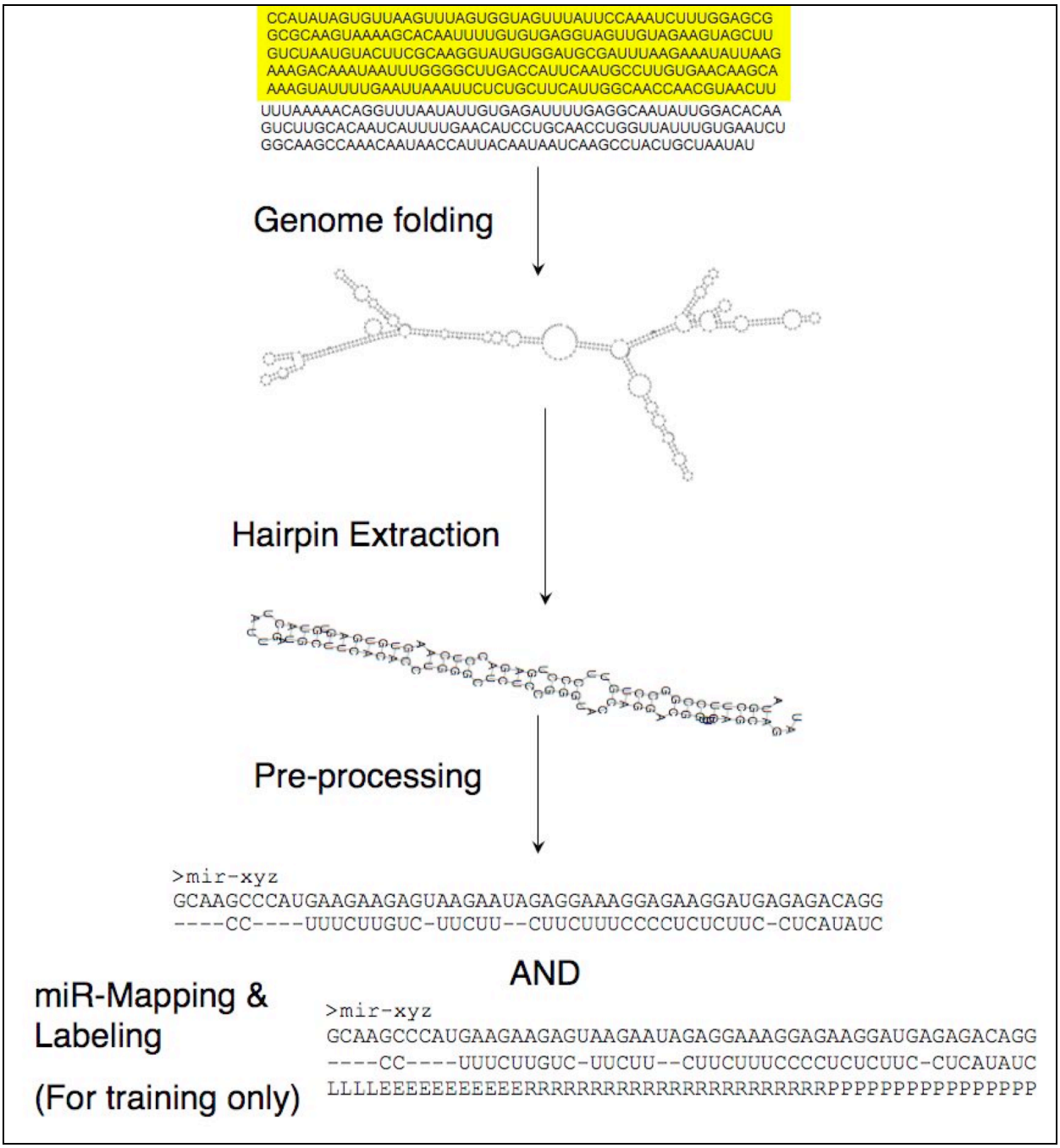


Figure 2.5: Data flow for hairpin extraction from the genome: The genome is first folded using windows of 500 nts with 150 nts overlap between consecutive windows. Hairpins are then extracted from the folded windows using the parameters described in the text. Hairpins are pre-processed into a suitable format for training/testing using the states shown in **Figure 2.2** (L: Loop; E: Extension; R: miRNA; P: pri-miRNA extension). For the purpose of testing, the folded sequence is pre-processed into 2 lines of input representing the 2 stems of the hairpin. An example is given in **Figure 2.1b**.

The labeling is done only for training data. For the purpose of labeling, the miRNA is first mapped to the folded hairpin (on either or both arms), and then the region representing the miRNA is labeled as the duplex miRNA (R) region. Our method does not consider the 3' overhangs generated during Dicer processing. The main bulge is labeled as the loop (L), whereas the remaining region between loop and miRNA is represented as the extension (E). The rest of the hairpin beyond the miRNA is labeled as pri-extension (P).

2.2.6.2 Parameter Estimation & Testing

(a) Parameter Estimation

Two separate HHMM models are trained, one on positive data set (miRNAs and their corresponding hairpins) and the other on negative data set (hairpins, randomly chosen from the coding parts of the genome). The hairpins are pre-processed and labeled (if needed) before parameter estimation. Baum-Welch requires no labeling, but for MLE, we applied random labeling, as described above (**Figure 2.1a**).

The parameter set of an HHMM is denoted by:

$$\lambda = \left\{\lambda^{q^d}\right\}_{d\in\{1,\ldots,D\}} = \begin{Bmatrix} \left\{A\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}, \\ \left\{\prod\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}, \\ \left\{E\left(q^D\right)\right\} \end{Bmatrix} \quad \text{(Eq. 2.1)}$$

The alphabet will be denoted by $\sum = \{\sigma_i\}$ and the observed finite string $O = o_1 o_2 \ldots o_N$. The i^{th} state at hierarchical level d is denoted as q_i^d (denoted as q^d in absence of ambiguity). The highest level of hierarchy (of the root) is 1 while the lowest (of the production states) is D.

The number of sub-states of q_i^d is denoted by $\left|q_i^d\right|$. Each internal state q^d $\left(d \in 1,2... D-1\right)$ has an initial distribution vector

$$\prod\left(q^d\right) = \left\{\pi\left(q_j^{d+1}\middle|q^d\right)\right\} = \left\{P\left(q_j^{d+1}\middle|q^d\right)\right\} \quad \text{.....(Eq. 2.2)}$$

that is the probability that q^d will make a vertical transition to its j^{th} sub-state at level $d+1$. Similarly, each internal state will also have a transition matrix,

$$A\left(q^d\right) = \left(a_{jk}^{q^d}\right) = P\left(q_k^{d+1}\middle|q_j^{d+1}\right) \text{.....(Eq. 2.3)}$$

where each $a_{jk}^{q^d}$ is the probability that the j^{th} sub-state of q^d will transition to the k^{th} sub-state. The production states q^D will have emission probability vector,

$$E\left(q^D, q^{D-1}\right) = \left\{e\left(\sigma_l\middle|q^D, q^{D-1}\right)\right\} = \left\{P\left(\sigma_l\middle|q^D, q^{D-1}\right)\right\} \text{.....(Eq. 2.4)}$$

where is the probability $e\left(\sigma_l\middle|q^D, q^{D-1}\right)$ is the probability that production state q^D whose parent state is q^{D-1} will emit symbol $\sigma_l \in \sum$.

Now we will define the various probabilities that are required to be calculated for parameter estimation.

(i) $\alpha\left(t, t+k, q_i^{d+1}, q^d\right) = P(o_t \cdots o_{t+k}, q_i^{d+1}$ finished at $o_{t+k}\middle|q^d$ started at $o_t)$ is the forward probability of emitting the substring $o_t \cdots o_{t+k}$ when the parent state q^d was entered at o_t and the subsequence ended at sub-state q_i^{d+1}.

(ii) $\chi\left(t, q_i^{d+1}, q^d\right)$ is the probability of making a vertical transition from parent q^d to q_i^{d+1} just before the emission of o_t.

(iii) $\xi\left(t,q_i^{d+1},q_j^{d+1},q^d\right) = P(o_1 \cdots o_t, q_i^{d+1} \rightarrow q_j^{d+1},\ o_{t+1} \cdots o_N | \lambda)$ is the probability of making a horizontal transition from q_i^{d+1} to q_j^{d+1} where both are sub-states of q^d after the emission of o_t and before the emission o_{t+1}.

(iv) $\gamma_{in}\left(t,q_i^{d+1},q^d\right) = \sum_{k=1}^{|q^d|} \xi\left(t-1,q_k^{d+1},q_i^{d+1},q^d\right)$ is the probability of performing a horizontal transition to q_i^{d+1} which is sub-state of q^d before o_t is emitted. Further details on the algorithms are given in .

The parameters are estimated as follows:

$$\hat{\pi}\left(q_i^2 \middle| q^1\right) = \chi\left(t,q_i^2,q^1\right)$$

$$\hat{\pi}\left(q_i^{d+1} \middle| q^d\right) = \frac{\sum_{t=1}^{T} \chi\left(t,q_i^{d+1},q^d\right)}{\sum_{i=1}^{|q^d|} \sum_{t=1}^{T} \chi\left(t,q_i^{d+1},q^d\right)} \quad (1 < d < D-1) \qquad \text{(Eq. 2.5)}$$

$$\hat{a}_{jk}^{q^d} = \frac{\sum_{t=1}^{T} \xi\left(t,q_i^{d+1},q_j^{d+1},q^d\right)}{\sum_{k=1}^{|q^d|} \sum_{t=1}^{T} \xi\left(t,q_i^{d+1},q_k^{d+1},q^d\right)} \qquad \text{(Eq. 2.6)}$$

$$\hat{e}\left(\sigma_l \middle| q^D, q^{D-1}\right) = \left(\sum_{o_t=\sigma_l} \chi\left(t,q_i^D,q^{D-1}\right) + \sum_{t>1,o_t=\sigma_l} \gamma_{in}\left(t,q_i^D,q^{D-1}\right)\right) / \left(\sum_{t=1}^{T} \chi\left(t,q_i^D,q^{D-1}\right) + \sum_{t=2}^{T} \gamma_{in}\left(t,q_i^D,q^{D-1}\right)\right) \qquad \text{(Eq. 2.7)}$$

(b) Testing

As described above, classification of test hairpins depends on the ratio of the log-likelihoods generated by the positive and negative models. For each hairpin, the probability that a certain model emitted the hairpin is given by:

$$P(O|\lambda) = \sum_{i=1}^{|q^1|} \alpha(1, T, q_i^2, q^1) \quad \text{(Eq. 2.8)}$$

(c) Measures of Accuracy

The different terms and measures used to calculate the efficiency of HHMMiR are listed in the **Table 2.6**.

Table 2.6: Measures for accuracy calculation: TP: *True Positives*; TN: *True Negatives*; FP: *False Positives*; FN: *False Negatives*.

Measure	*Calculation*
Sensitivity (Sn)	$Sn = TP / TP + FN$
Specificity (Sp)	$Sp = TN / TN + FP$
False Discovery Rate (FDR)	$FDR = FP / TP + FP$
Matthew's Correlation Coefficient (MCC)	$MCC = (TP \cdot TN - FP \cdot FN) / \left(\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}\right)$

2.3 HESD-HMM

2.3.1 Need for explicit state duration densities for decoding

Although HHMMiR is a good classifier for miRNA precursors, its performance is not as good when decoding potential miRNA hairpins, using a modified Viterbi algorithm. The performance evaluating parameters were (i) difference in the start position of the miRNA region and (ii) difference in the length of the real and predicted miRNA sequence. As shown in **Figure 2.6a**, HHMMiR can miss the correct length of the miRNA by almost 20-30 bases for some samples. Although, a human (HSA) trained model performs fairly on human and Drosophila (DME) datasets, it can perform very poorly on *C. elegans* (CEL) data. A model trained on DME data is unable to predict the correct size of the miRNA in HSA and CEL efficiently (**Figure 2.6a**).

We investigated the reason for the poor performance of HHMMiR as a decoder. State duration modelling is a known weakness of standard HMMs. In a typical HMM state q_i, the probability of being in that state for duration k follows an exponential distribution.

Consider observation sequence $O = \left\{ \underset{1}{o_{q_i}}, \underset{2}{o_{q_i}}, \underset{3}{o_{q_i}}, ..., \underset{k}{o_{q_i}}, \underset{k+1}{o_{q_j} \neq o_{q_i}} \right\}$

$$\Pr\left(O \middle| \text{model}, q_1 = o_{q_i}\right) = \left(a_{ii}\right)^{k-1}\left(1 - a_{ii}\right)$$
$$= p_i(k) = \Pr(k \text{ consecutive observations in state } q_i)$$

$p_i(k)$ is a probability density distribution (pdf) of duration k in state q_i. This pdf is exponential for a typical Markov chain (**Figure 2.8**). However, empirical data shows that the real

duration densities of the four regions of a miRNA hairpin cannot be modelled by an exponential distribution (**Figure 2.7**).

Thus, we introduced explicit state duration densities to all internal states of the hierarchical HMM to model these empirical distributions (**Figure 2.8b**).

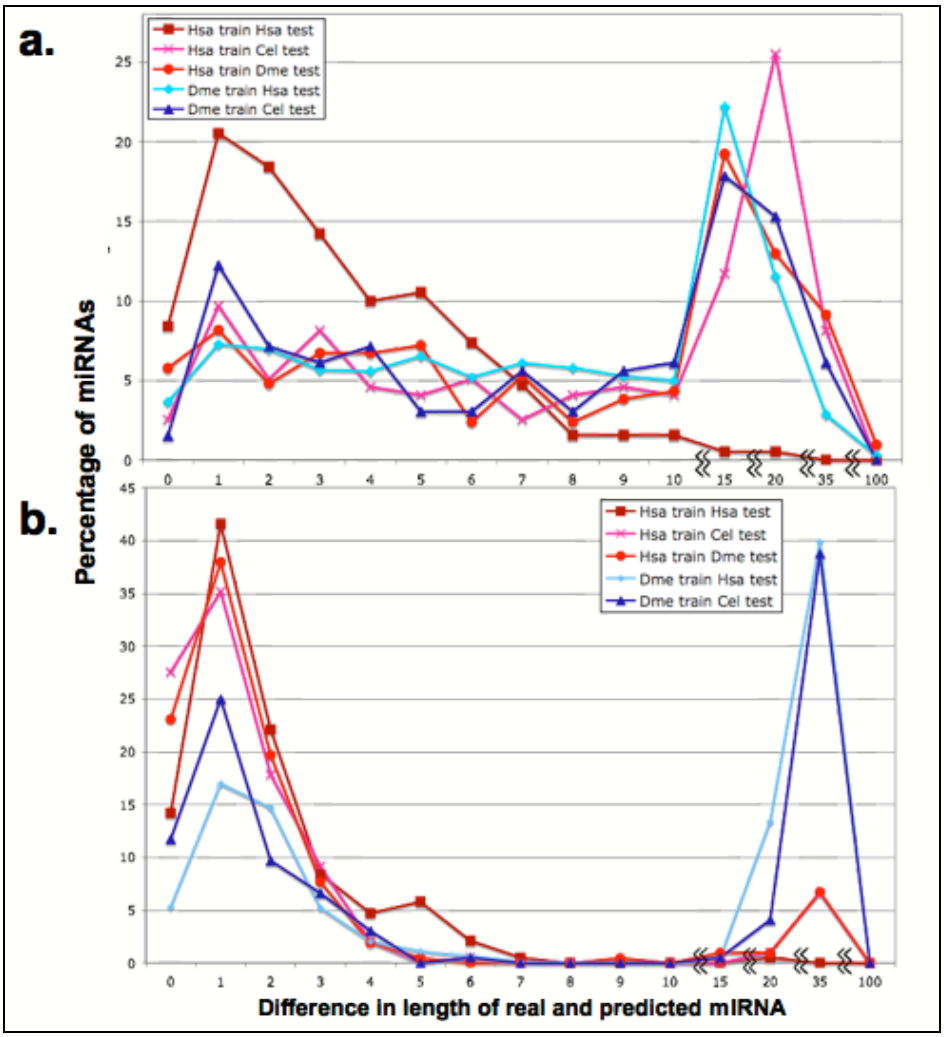


Figure 2.6: Decoding Error in miRNA predictions using HHMMiR: Difference in lengths of real vs. predicted miRNAs using HHMMiR. Each line represents a cross-species run where HHMMiR parameters were trained used miRNA data from one species and tested on another species. For "Hsa train Hsa test", the model was trained on randomly sampled 2/3rd of the dataset and tested on the remaining 1/3rd dataset, using random sampling (*Hsa:* human; *Cel: C. elegans*; *Dme*: *D. melanogaster*).

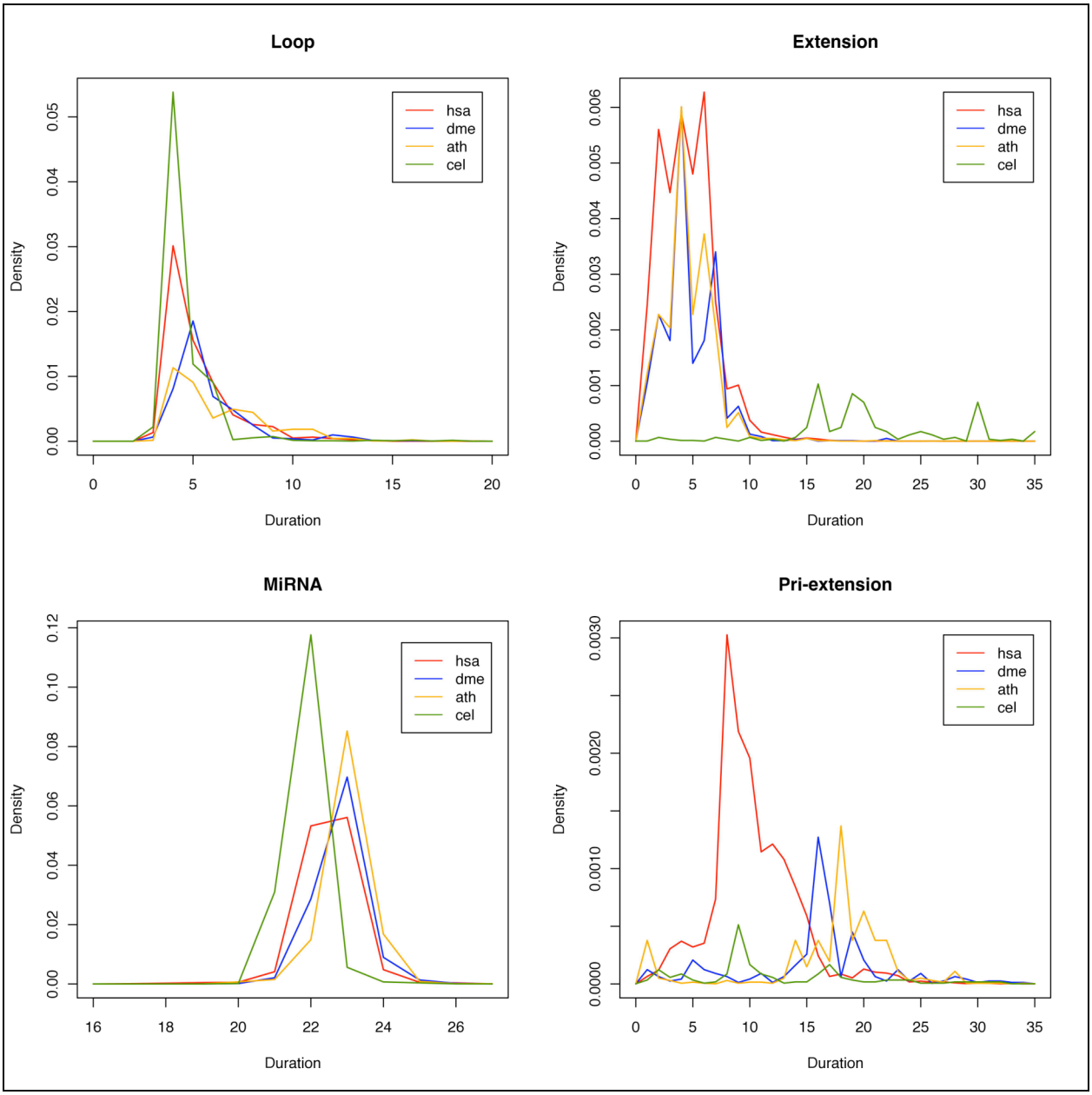


Figure 2.7: Real duration densities at the internal states: The distributions are truncated at 35bps for extension and pri-extension for better visibility. (hsa- *H. sapiens*; dme- *D. melanogaster*; ath- *A. thaliana*; cel- *C. elegans*)

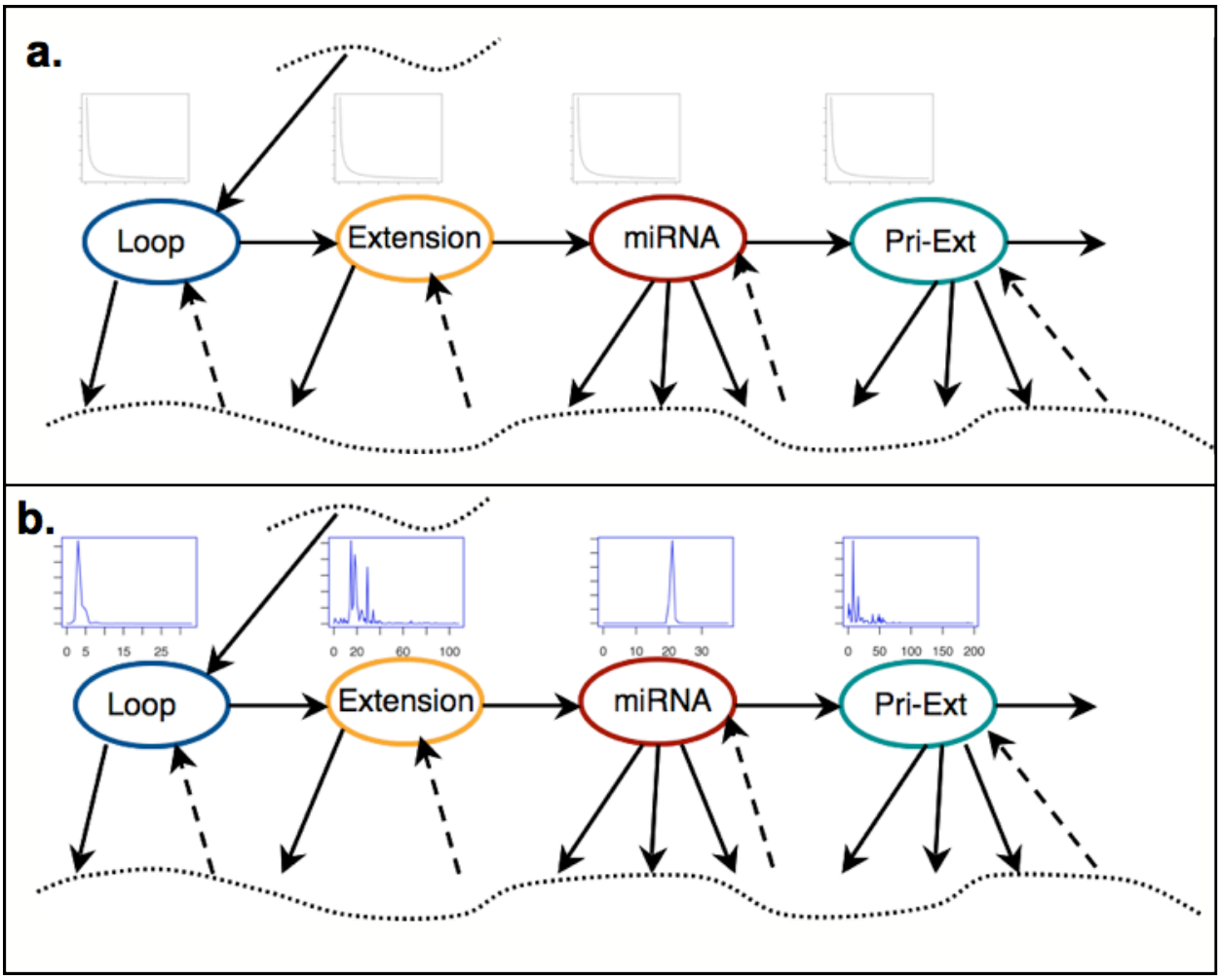


Figure 2.8: Cartoon representation of HESD-HMM model with explicit state duration densities: (a) Probability density function (pdf) for internal states of HHMMiR is exponential. (b) After adding explicit state duration densities to internal states, HESD-HMM learns the pdfs using MLE. A transition is only made once an appropriate number of observations have occurred in that state.

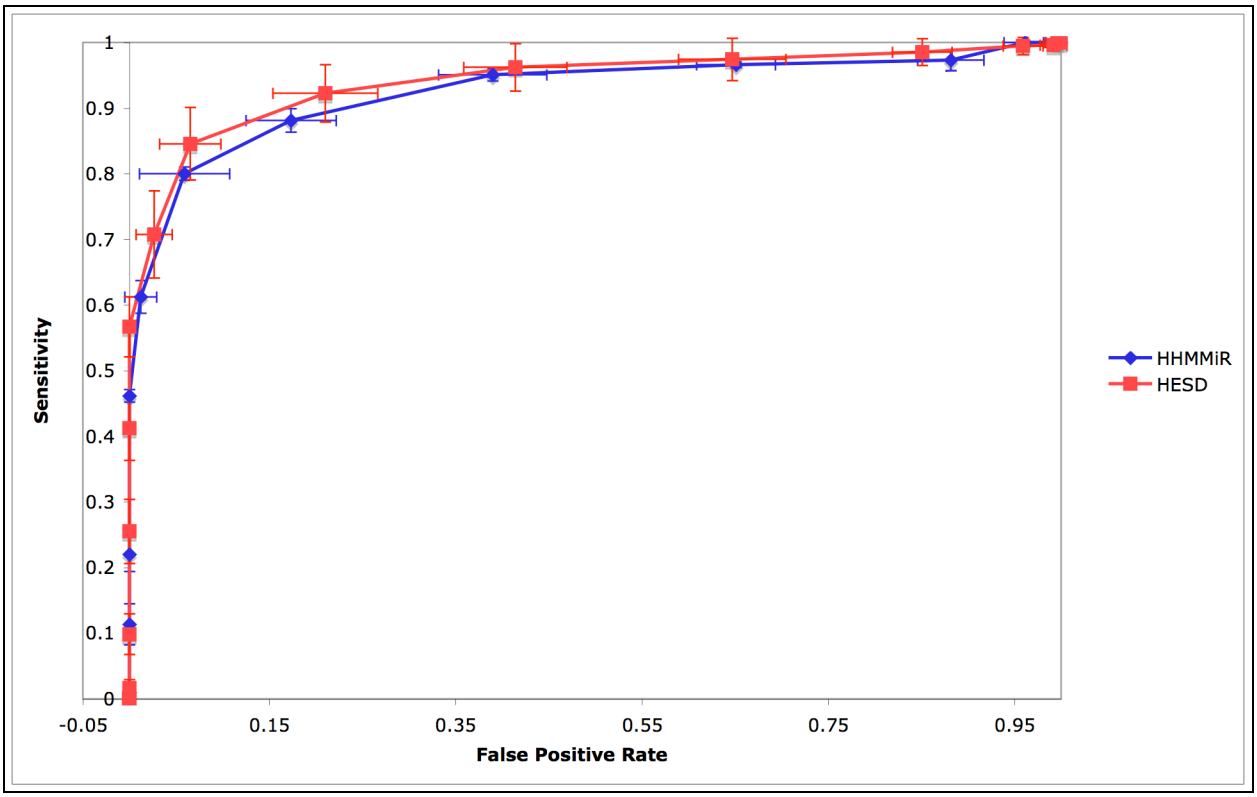


Figure 2.9: ROC curve on human data to compare classification performance before and after adding explicit state duration densities: 10-fold cross-validations used with HHMMiR (*blue curve*) and HESD-HMM (*red curve*). Positive and negative models were trained using MLE in both cases.

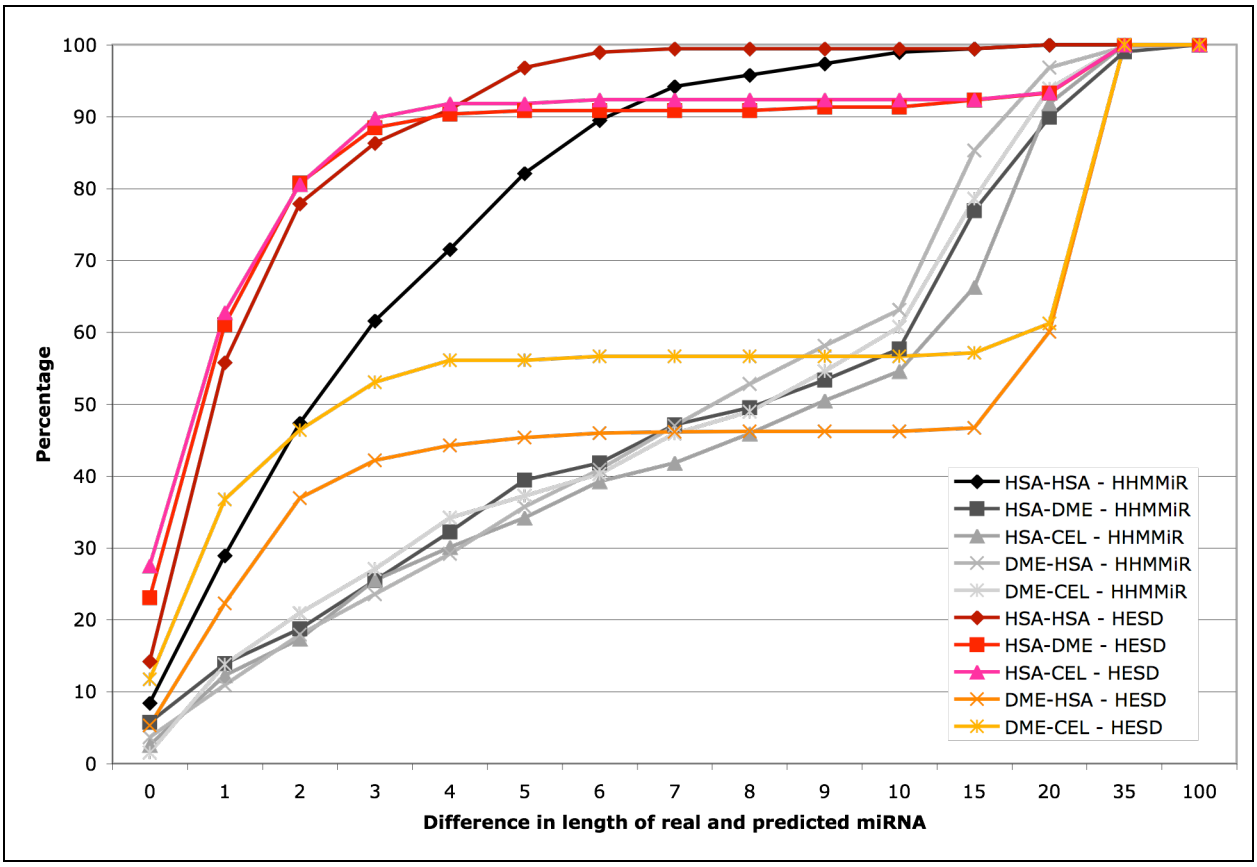


Figure 2.10: Cumulative distribution in decoding error between HHMMiR and HESD-HMM: Line plots in black and shades of gray represent cumulative density distributions (cdfs) of test set prediction error in miRNA length using HHMMiR for decoding, whereas line plots in other colors represent the same cdfs using HESD-HMM for decoding. A drastic improvement in decoding can be seen with the human trained models.

2.3.2 Testing efficiency of classification and decoding

On 10-fold cross validation of human miRNA data, classification improves only slightly with explicit state duration density parameters added to the internal states of the model, as shown by the *red* plotted line in **Figure 2.9**. On decoding human miRNA hairpins with a model trained on human data (self-test) ($2/3^{rd}$ of the dataset -randomly sampled- for training the model, and the remaining $1/3^{rd}$ dataset was used for testing purpose). We found that ~80% of the hairpins are decoded with a difference in length of at most 2 bases. Whereas, only ~47% of the hairpins are decoded within at most 2 bases with HHMMiR (compare *dark red* and *black* line plots in **Figure 2.10**). The improvement in predicted lengths is expected with the explicit state duration densities.

2.3.2.1 Cross species decoding performance

When the models were trained on data from one species and tested on another species, we saw variability in performance (**Figure 2.6** & **Figure 2.10**), depending on the training dataset. As seen in **Figure 2.10**, models trained on human miRNAs, perform drastically better with explicit state durations, than models trained on Drosophila data. The improvement in performance with human data is expected since the miRNA characteristics in humans are typical of overall animal miRNAs, whereas Drosophila characteristics are slightly different from other animal species (this difference in trends can be see in **Figure 2.7**).

2.3.3 Methods

2.3.3.1 Datasets and Processing

miRNA genes were obtained from the microRNA registry, version 17 (April 2011) (Griffiths-Jones 2006). For training HESD-HMM, we used the ~1300 human hairpins, after filtering out precursor genes with multiple loops. Each gene was folded with the RNAfold program, which is part of the Vienna package (Hofacker 2003), using the default parameters to obtain the secondary structure with minimum fold energy. The negative set is the same used for HHMMiR mentioned in **Section 2.2.6.1**. All processing steps are shown in **Figure 2.5**.

2.3.3.2 Modification to the Viterbi algorithm

The key to understanding the modified Viterbi algorithm for its two conditions is that there can be two cases for the length of the sequence for which the forward probability is being calculated:

For $o_t o_{t+1} \ldots o_{t+k}$ and Z is the maximum duration of the state in question,

- When $k < Z$ and thus, length of the sequence $o_t o_{t+1} \ldots o_{t+k}$ is $k+1$

  Consider,

  - The probabilities for q_i^d emitting the entire sequence, which is a product of the
    - Probability that the state is activated by its parent, $\pi^{q^{d-1}}\left(q_i^d\right)$
    - Explicit duration density for $k+1$ $p^{q_i^d}\left(k+1\right)$
    - Probability of emitting the sequence

      $\sum_{s=1}^{\left|q_i^d\right|} \alpha\left(t, t+k, q_s^{d+1}, q_i^d\right) \cdot a_{s\,\text{end}}^{q_i^d}$ for internal states

$\prod_{s=0}^{k} b^{q_i^D}(o_{t+s})$ for production state

- For each subsequence with lengths 0,1,…,k-1,k, and thus, $\sum_{l=0}^{k-1}$ the subsequence ended at a state (at same level as q_i^d) and transitioned to q_i^d

  $\sum_{\substack{j=1 \\ j \neq i}}^{|q_i^{d-1}|} \alpha\left(t, t+l, q_j^d, q^{d-1}\right) \cdot a_{ji}^{q^{d-1}}$

  then q_i^d emitted the remaining subsequence

  $p^{q_i^d}(k-l) \cdot \sum_{s=1}^{|q_i^d|} \alpha\left(t+l+1, t+k, q_s^{d+1}, q_i^d\right) \cdot a_{s\,end}^{q_i^d}$ for internal states.

- When $k \geq Z$

  We will calculate this probability by taking into account that q_i^d emitted a subsequence of length $z = 1...Z$, multiplied by the forward probability of another sibling state of q_i^d emitted the remaining subsequence. Thus, $o_t o_{t+1} ... o_{t+k-z}$ was emitted by some q_j^d, then a transition was made from q_j^d to q_i^d and q_i^d emitted the remaining $o_{t+k-z+1} ... o_{t+k}$

For the detailed algorithm, refer to **Appendix A.1**.

2.4 Conclusions

MiRNA genes constitute a highly conserved mechanism for gene regulation across all animal and plant species. The characteristics of the precursor miRNA stem-loops are well conserved in both vertebrate and invertebrate animals and fairly conserved between animals and plants. As seen in **Table 2.1**, plant hairpins tend to be generally longer than those in animals,

while vertebrates have shorter precursors than invertebrates. Although, the "extension" and "pri-extension" regions may vary in length between animals and plants (much longer in plants), the lengths of the "miRNA" and "loop" regions are more similar in length. Thus, even across evolutionary time, the basic characteristics of miRNAs have not changed dramatically.

We designed a template for a typical precursor miRNA stem-loop and we built an HHMM based on it. HHMMiR was able to attain an average sensitivity of 84% and specificity of 88% on 10-fold cross validation of human data. We trained HHMMiR on human sequences and the resulting model was able to successfully identify a large percentage of not only mouse, but also invertebrate, plant and virus miRNAs (**Table** 2.4). This is an encouraging result showing that HHMMiR may be very useful in predicting which genomic stemloops contain miRNAs across long evolutionary distances without the requirement for evolutionary conservation of the sequences. This would be very beneficial for identification of miRNA-containing hairpins in organisms that do not have a closely related species sequenced, such as *Strongylocentrotus purpuratus* (sea urchin) and *Ornithorhynchus anatinus* (platypus) (Samollow 2008).

Nam *et al.* (Nam et al. 2005) previously applied probabilistic learning for identifying the miRNA pattern/motif in hairpins. The advantage of the hierarchy used by HHMMiR is that it parses each hairpin into four distinct regions and processes each of them separately. This represents the biological role of each region better, which is reflected in the distinct length distributions and neighborhood base-pairing characteristic of that region. Furthermore, the underlying HHMM provides an intuitive modelling of these regions. We compared two modifications of the MLE and Baum-Welch algorithms for modelling the negative datasets, and we found them to perform similarly. Baum-Welch was selected for this study, since it does not require (random) labeling of the negative set.

The drawback of HHMMiR is that it depends on the *mfe* structure the *RNAfold* program returns (Hofacker 2003). In the future, different folding algorithms can be tested. Alternatively, the probability distribution of a number of top scoring folding energy structures returned by this package can be used to consider the entire space of secondary RNA structures.

Adding explicit state duration densities increases the decoding power of HHMMiR, as seen in **Figure 2.6** and **Figure 2.10**. This is because the state duration of a typical HMM, which is an exponential family distribution, cannot model the empirical duration densities. Thus, modelling the state durations with empirical data can cause increased decoding power, in same species as well as across species datasets.

The success of our approach shows that the conservation of the miRNA mechanism may be at a much deeper level than expected. Further developments of the HHMMiR algorithm include the extension of the precursor template model (**Figure 2.2**) to be able to predict pri-miRNA genes with multiple stem-loops. Another extension would be to train a model to decode all HHMMiR predicted hairpins to identify the miRNA genes in them.

3.0 MIRNA DISCOVERY IN ECHINODERM EMBRYOS USING NEXT GENERATION SEQUENCING

3.1 Introduction

The first miRNAs, *lin-4* and *let-7* were discovered in *C. elegans*, as regulators of developmental timing (R. Lee 1993; B J Reinhart et al. 2000), and since then, miRNAs have been implicated in many developmental and tissue differentiation processes (Kloosterman & R. H. A. Plasterk 2006; V. Ambros 2004). miRNAs have been found in all animal lineages, although specific miRNAs have been lost and gained during evolution (Berezikov et al. 2010; Sempere et al. 2006). Some orthologous miRNAs are associated with conserved expression in similar tissues, which may suggest conservation of function (Christodoulou et al. 2010).

In this chapter, we present for the first time, concrete evidence that many small non-coding RNA genes (including miRNAs) are expressed in high-numbers in the early developmental stages of two distantly related species, *S. purpuratus* and *P. miniata*, which last shared a common ancestor almost 500 million years ago (MYA) (Wada & Satoh 1994). The goal of this study is to determine the pool of miRNAs involved in development of these two echinoderm species. We sequenced small RNA libraries of mixed population embryos from each of these echinoderms using Illumina Genome Analyzer (Illumina, Inc.), which provides a better depth of sequencing compared to 454. In the future, it will be extremely interesting to study

stage-specific expression of these miRNAs. Comparison of the two sequenced datasets showed that a large number of miRNAs are expressed during development in the two species. Most of the identified miRNAs have homologs in other species, but a number of novel (echinoderm-specific) miRNAs were also identified. The data reported here will provide a valuable resource for evolutionary comparisons across a broader distance in the phylogenetic branch of deuterostomes, and this can help complete the puzzle of developmental gene regulatory networks in these two model organisms. Most of this chapter is from (Kadri et al. 2011).

Table 3.1: : Summary statistics of sea urchin and sea star deep sequencing data, and annotations. Note that the number of reads for non-coding RNAs, such as tRNAs, rRNAs, snRNAs, snoRNAs and miRNAs, are for the length range 17-26nts. For discovery of conserved miRNAs in the libraries, only tags with more than 2 reads were used, whereas, for potential novel predictions, tags with more than 5 reads were used.

	S. purpuratus	*P. miniata*
Genome	800 Mb	500Mb
Total number of reads	12,907,171	9,760,097
Reads mapped to genome	9,401,944	N/A
Tags (collapsed reads)"	2,486,028	2,513,198
Reads mapped to:		
tRNAs	7,550	33,551
rRNAs	288,036	319,035
snRNAs & snoRNAs	6,217	1,805
miRNAs (conserved)	376,007	48,320
miRNAs (potentially novel)	5,834	281
Number of conserved miRNAs	47	38
Potentially novel miRNAs (miRDeep)	11	3

3.2 Results

3.2.1 A rich population of non-coding RNAs is expressed in sea urchin and sea star embryos.

High-throughput sequencing data (Illumina Genome Analyzer, Illumina, Inc.) corresponding to small RNAs were collected from a mixed embryonic population, individually from *S. purpuratus* (sea urchin) and *P. miniata* (sea star) as described in Methods. According to the Illumina protocol, the method specifically targets small RNAs with 3' hydroxyl group, so the RNAs processed by *Dicer* and other RNA processing enzymes are preferentially sequenced with this method. A collection of publicly available programs and in-house made scripts were used to parse the Illumina reads, and quantify known and novel miRNA gene expression (see **Computational analysis procedure and pipeline**).

Illumina sequencing of the small RNA libraries returned ~13 million reads for sea urchin and ~9.8 million reads for sea star embryos (**Table 3.1**). After removal of low quality 3' ends and linker sequences, the remaining reads (~11.6 and ~9.01 million reads from sea urchin and sea star, respectively) were collapsed into "tags" based on sequence identity (see **Materials & Methods**). This process resulted in a total of ~2.5 million tags from each species (**Figure 3.12**).

We focused on sequences of length 17-26 nts, since this is the typical size class expected for miRNAs. The histograms of the corresponding length distributions of reads and tags show similar trends between the two species (**Figure 3.2**). In the sea urchin reads, there is a peak of relatively highly expressed sequences at 22-23 nts (corresponding to the typical length of a miRNA) (**Figure 3.2**). The quality of the RNA was checked using a Bioanalyzer (**Figure 3.1**),

before and after adapter ligation, and indicated that the RNA was preserved (For more details, see **Materials & Methods**).

Presently, *S. purpuratus* is the only echinoderm with a sequenced genome (Sea Urchin Genome Sequencing Consortium et al. 2006). The sea urchin genome is not assembled into chromosomes yet, due to being highly polymorphic and having highly repetitive sequences. About 62% of the 17-26 nt sea urchin reads mapped to the genome (**Figure 3.4a**) (81%, if reads of all lengths are considered). The reads that do not map to the genome could be the result of sequencing errors or genome quality. Since the sea star genome is unavailable, we assign all unmapped reads to the "unknown" category (**Figure 3.4d**). Similarity searches against miRNAs and other known RNAs (coding and non-coding genes) were performed (see **Materials & Methods**). Approximately one quarter of the 17-26 nt long reads map to non-coding RNAs (14% to miRNAs and 10% to other non-coding RNAs), another quarter are mRNA degradation products, while 13% of reads map to the genome, but do not map to any annotated regions (**Figure 3.4a**). **Figure 3.4c&d** show the RNA composition of individual lengths in this size range in the sea urchin and sea star respectively. The 22nt long sea urchin reads were most enriched for miRNAs, while this trend was not seen in the sea star library. All the size classes show an almost uniform distribution of mRNA and rRNA partial reads. The un-annotated reads could be attributed to the relatively poor annotation quality of the sea urchin genome, or to large-scale transcription as it has been observed in other species (Preker et al. 2008; Taft et al. 2009; Anon 2004). For example, a recent report showed that most intergenic reads are found near transcription start or termination sites (van Bakel et al. 2010).

The relative abundance of the reads and tags that map to various non-coding RNAs varies substantially between sea urchin and sea star (**Figure 3.4b**). This is particularly true for

miRNAs, where 61.4% of the sea urchin reads (17-26 nts) map to miRNA sequences compared to 12.6% of sea star reads. For sea urchin embryos, the miRNA reads collapse to ~1,000 tags (that correspond to 42 miRNA genes), indicating a high expression of the miRNA genes (reads/gene average: 3,800; median: 413; 14 genes have >1,000 reads). By contrast, we found that a relatively higher number of sea star embryonic reads are mapped to (parts of) tRNA and rRNA genes (1.5% compared to 0.001%) (7.7% reads to tRNA and 77.9% reads to rRNA compared to 0.8% and 37% respectively) (**Figure 3.4b**). This may reflect a sampling bias, or may indicate that fewer miRNAs are expressed in sea star embryos compared to sea urchin embryos. We found miRNA* species for most miRNAs, and in some cases, the miRNA* was more abundant than the miRNA itself (for example, *miR-200, miR-2008, miR-219, miR-2011*) (**Figure 3.3**).

In summary, the sea urchin and sea star samples showed differences in the distribution of annotated small RNA classes, with the most striking difference being the relative higher enrichment of miRNAs in sea urchin embryos.

3.2.2 Conservation of developmental miRNA gene expression in echinoderms.

We used sequence homology as well as information about the secondary stem-loop structure of precursor sequence to search for conserved and novel miRNAs in sea urchin and sea star embryonic libraries (see **Small RNA library preparation**). We found a total of 47 sea urchin and 38 sea star miRNAs mapping to known sequences in the miRBase registry (v. 17, April 2011) (Griffiths-Jones 2006) (**Table 3.1**). **Figure 3.5a** shows the overlap between miRNAs found expressed in the two embryonic libraries as well as adult sea urchins (B. M.

Wheeler et al. 2009). Overall, 53 miRNAs are expressed in the embryonic stages of one or both species, whereas, 31 are expressed in sea urchin adults as well as in the embryonic stages of both species (**Figure 3.5a**). This figure does not include the miRNA* species. When comparing miRNA expression between the two species, 25 are present in sea urchin only, 4 in sea star only (*miR-92d, miR-1692, miR-100, miR-4171*) and 34 in both species (**Figure 3.5a**). The common hits are considered as putative candidates for phylum specific miRNAs. *miR-100* is considered a sea star specific miRNA in **Figure 3.5** as it was absent in our sea urchin embryonic library and Wheeler *et al.* did not find this miRNA in the sea urchin adult by 454 sequencing (B. M. Wheeler et al. 2009).

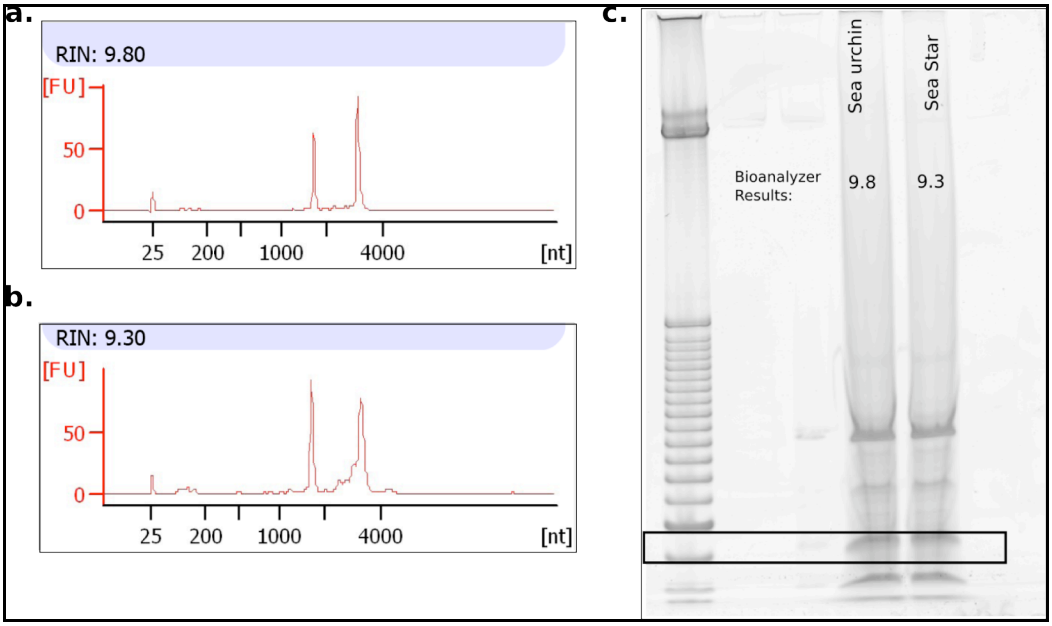


Figure 3.1: The RNA quality was checked using the BioAnalyzer before (a,b) and after (c) adapter ligation. **(a)** Distribution of lengths of the RNA sample from sea urchin before adapters were ligated. The first peak (~20-25 nt) corresponds to the small RNA population. **(b)** Length distribution of sea star RNA sample before adapter ligation. **(c)** The adapter-ligated RNA was run on a gel and size-selected for small RNAs.

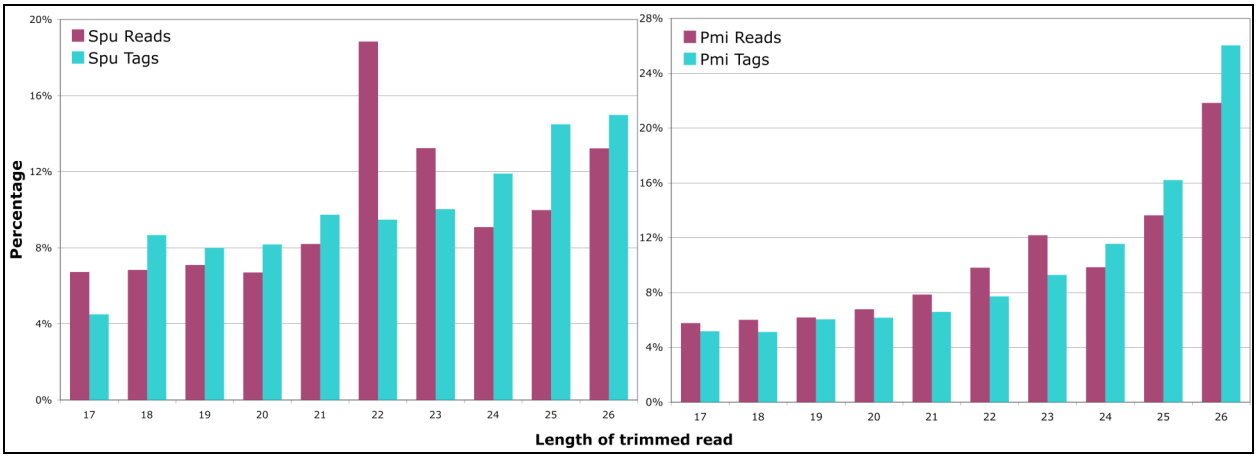


Figure 3.2: Length distributions of sea urchin and sea star reads. Histogram of length distribution of reads and tags in sea urchin and sea star small RNA Illumina libraries. The peak corresponding to the typical length of a miRNA is seen at 22nts in sea urchin, but this peak is not as enhanced in the sea star library. *Spu: Strongylocentrotus purpuratus; Pmi: Patiria miniata*

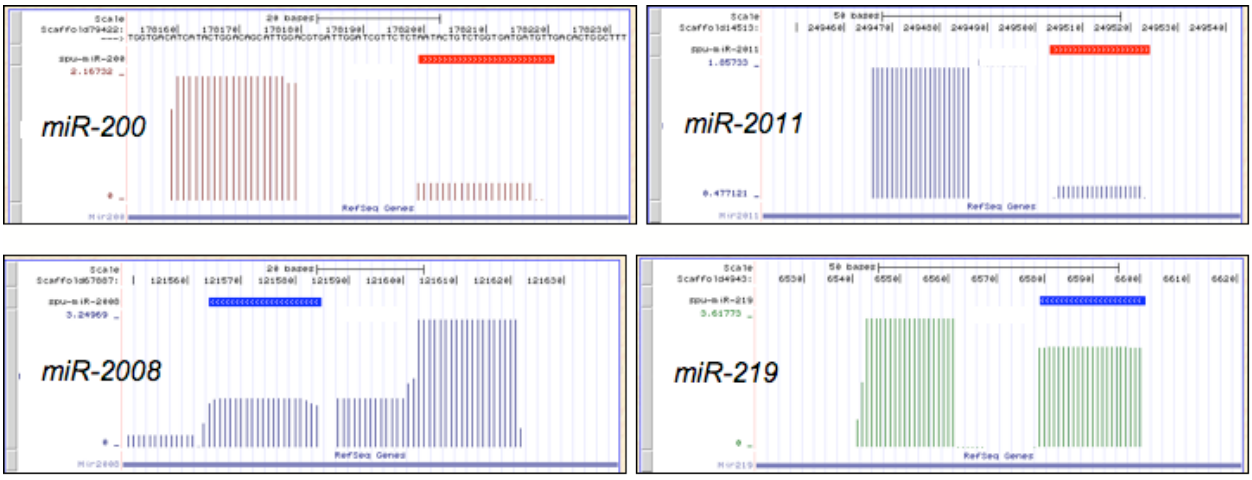


Figure 3.3: Reads for mature miRNA and miRNA* in UCSC genome browser for the sea urchin. Reads (logarithm scale) for miRNA and miRNA* for cases in which the miRNA* is more abundant than miRNA.

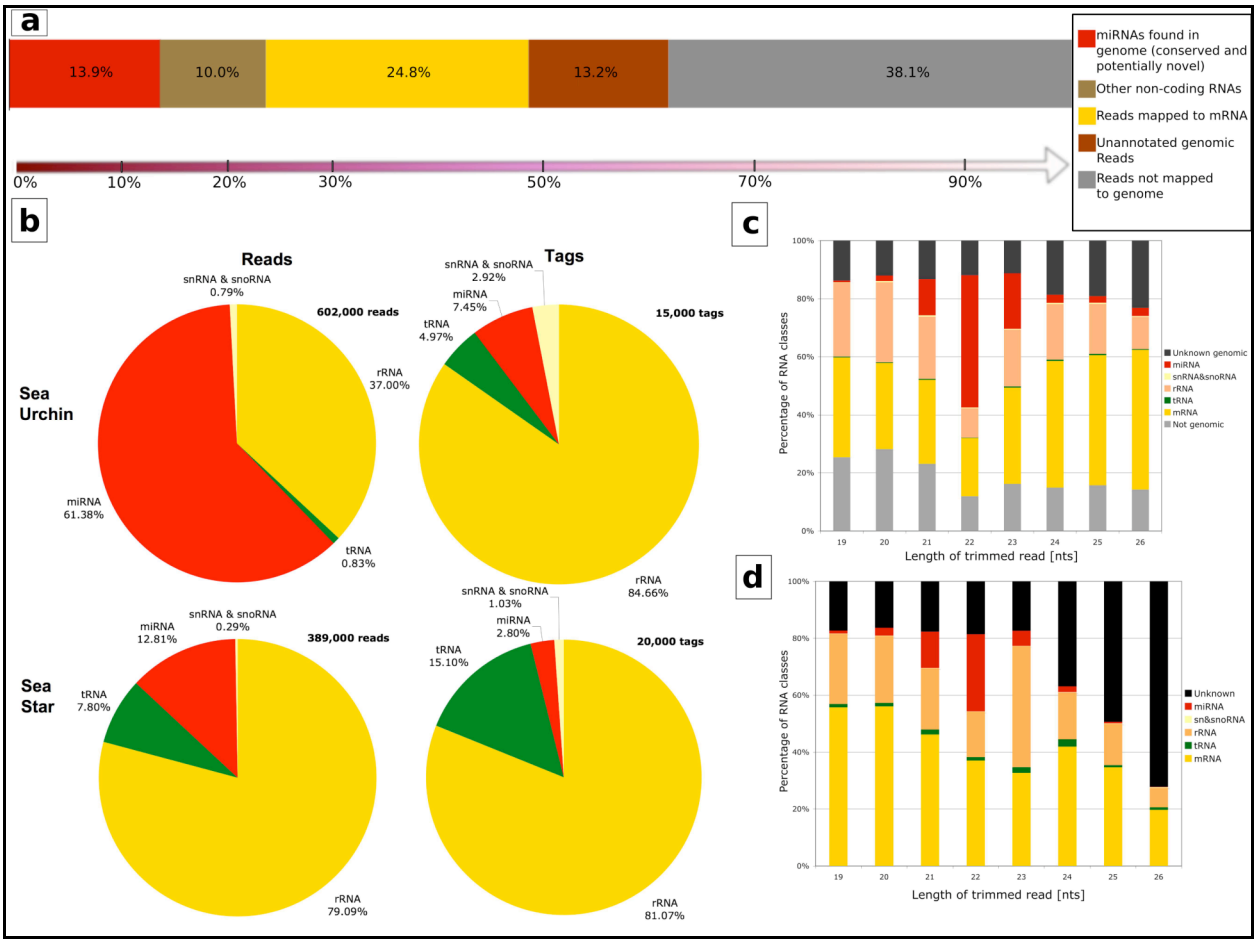


Figure 3.4: Distribution of annotated reads in small RNA libraries. (a) Bar showing the distribution of annotated reads 17 to 26 nts in length, for sea urchin. **(b)** Fractional distribution of non-coding RNAs in sea urchin and sea star embryonic small RNA libraries. Mapping of the annotated classes to reads and tags, shows the relative abundance (frequency) of each class per tag. All classes of non-coding RNAs compared were mapped to reads of lengths 17 to 26 nts. *Spu: Strongylocentrotus purpuratus; Pmi: Patiria miniata*

The flanking genomic regions of the conserved miRNAs were folded into their mfe secondary RNA structure using the *RNAfold* program from Vienna Package (Hofacker 2003). These secondary structures were checked for the typical stem-loop structures characteristic of *Dicer* processing (See **APPENDIX G** for the stem loop structures of some of the high abundance *S. purpuratus* miRNAs). Additionally, the current version of the sea urchin genome (version 2.1, UCSC Genome Browser (Kent et al. 2002)) lacks *miR-100* sequence as well. However, northern blot analysis previously showed that *miR-100* is present in sea urchin adult (coelomycytes and mesenchyme) (Sempere et al. 2006). It will be interesting to verify whether the adult tissue in sea urchin expresses it or not, thus, deciding its position as a species specific or phylum-conserved miRNA.

3.2.2.1 Novel miRNAs

We used miRDeep to identify potentially novel miRNAs in sea urchin (M. R. Friedländer et al. 2008) (we were not able to use miRDeep on the sea star dataset, because of the lack of the genomic sequence in this species.) Of the 11 novel predictions, 8 genes (5,183 reads) have seed sequences (positions 2-8) similar to known miRNAs in the registry (**Figure 3.6a**), while 3 are novel sequences with a total ~400 reads. Each of the potentially novel sea urchin predictions is part of stem-loop genomic hairpins, characteristic of *Dicer* processing (**Figure 3.6**). The novel sea urchin predictions were also matched to sea star reads. Three out of the 11 predictions were found in sea star (**Figure 3.12**). These three tags may therefore, represent echinoderm specific miRNAs. The other 10 tags may represent genes that have evolved after the divergence of the sea star and sea urchin lineages, although the sea star genome sequence is required before we make a definite assessment of this fact.

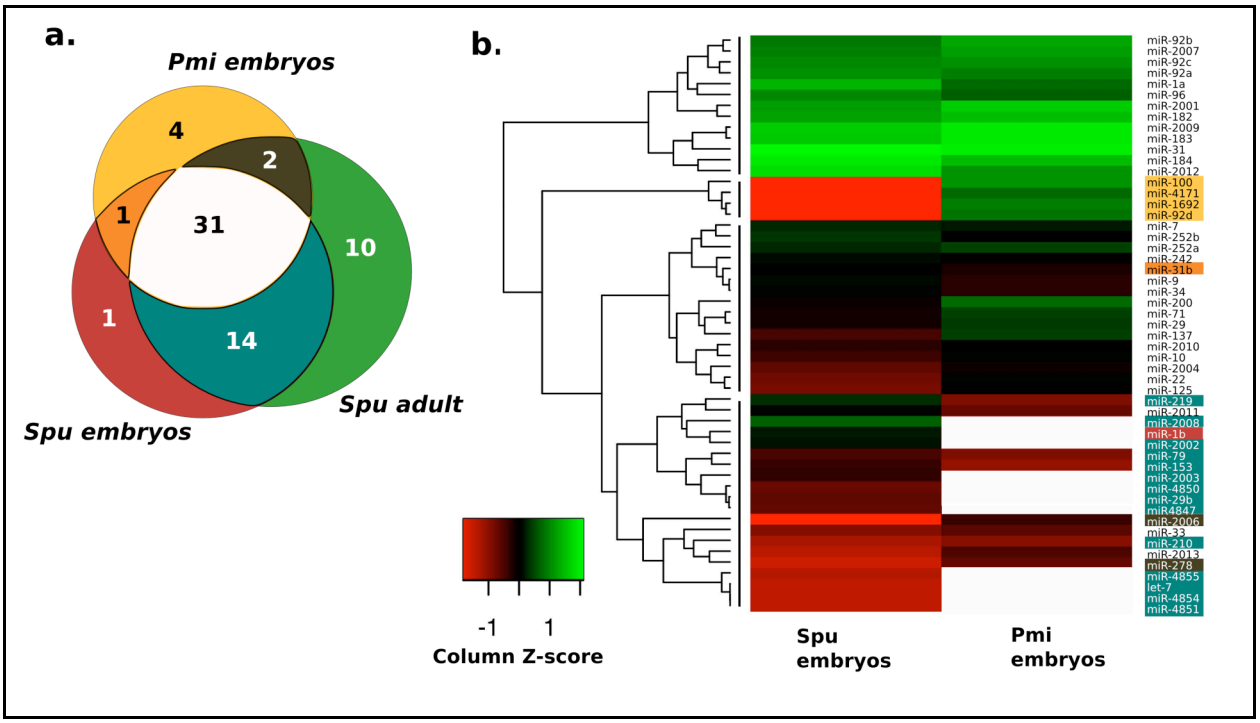


Figure 3.5: (a) Venn Diagram showing overlap between conserved miRNAs in sea urchin and sea embryos, and sea urchin adult (miRBase (Griffiths-Jones 2006)). Only Illumina tags >2 reads were treated as potential true miRNAs. This figure does not include the miRNA* species. **(b)** Heat map showing the relative miRNA expression between sea urchin and sea star embryos ($\log_2$ transformed relative expression values). Average linkage clustering using Euclidean distance as the distance metric was used to generate the heat map (**Materials & Methods**). Since the genome sequence for sea star is unavailable, absence of certain miRNAs from the small RNA library in sea star, but its presence in sea urchin is treated as missing values for sea star. Missing values for sea star are indicated by the background color. Only miRNAs with zero reads are treated as missing values, whereas miRNAs with 1 or 2 reads are shown in the heat map.

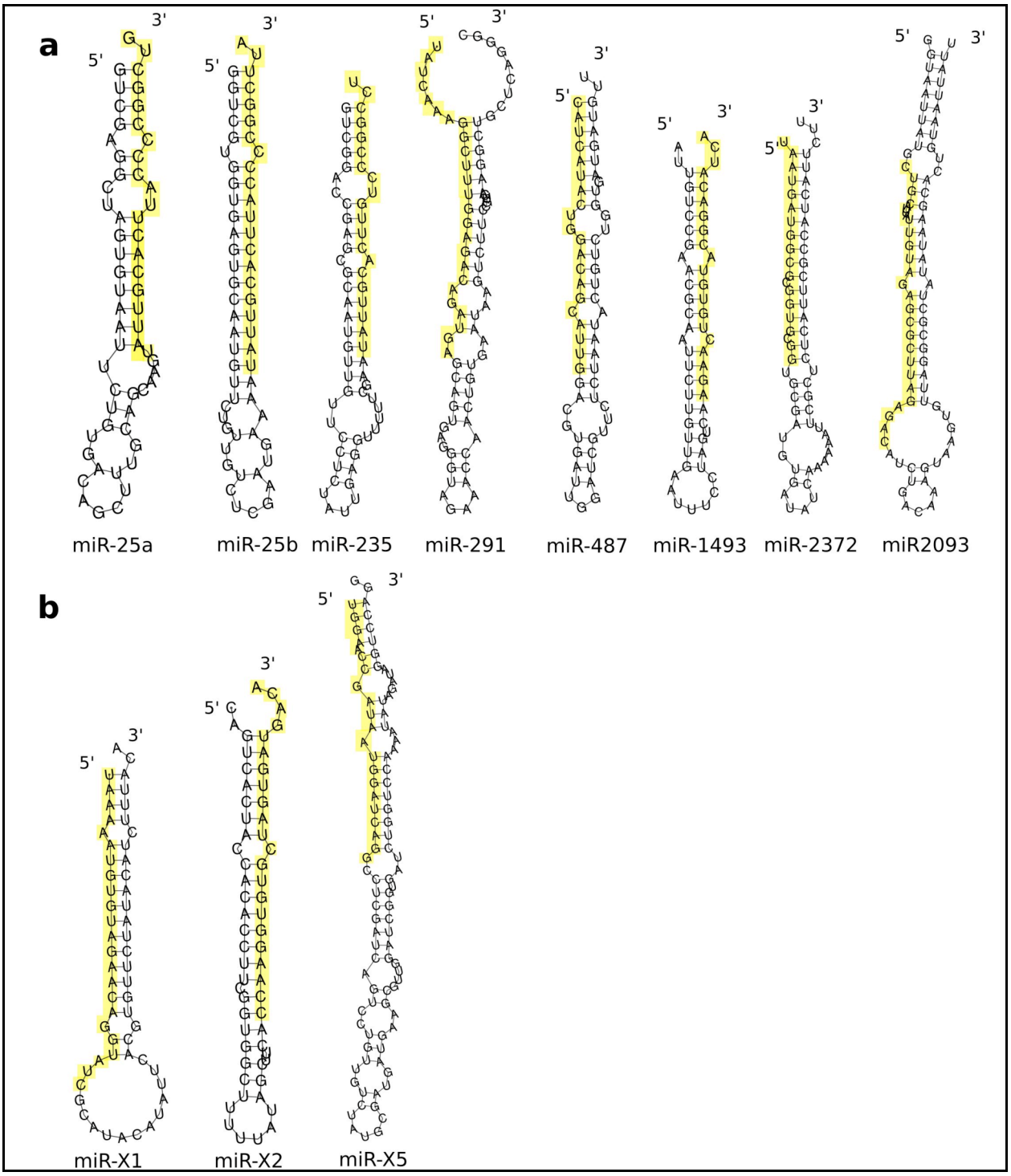


Figure 3.6: Stem-loop structures of the novel miRNA miRDeep (M. R. Friedländer et al. 2008) predictions in sea urchin. (a) miRNAs that share their seeds with known miRNAs. The temporary labels are the names of miRNA **(b)** Precursors of novel miRNAs without any seed conservation.

miRBase Release 17 (April 2011) (Griffiths-Jones 2006) currently contains 64 sea urchin gene entries, all obtained from adult tissue by 454 sequencing (B. M. Wheeler et al. 2009; Campo-Paysaa et al. 2011), including miRNA* species. No sea star miRNA genes are present in miRBase. Our embryonic libraries add 16 new sea urchin miRNA genes to this pool (2 conserved, 11 potentially novel and 3 miRNA*s); and 41 sea star miRNA genes (38 conserved, 3 potentially novel).

3.2.2.2 Comparison of miRNA genes expressed in embryos and adults

Most of the sea urchin miRNAs (45 out of 59) are expressed both in embryos (our dataset) and adults (miRBase registry) (**Figure 3.5a**). However, twelve miRNAs are present in the adult sea urchin only, but not in the embryonic stages considered. These may correspond to adult-specific miRNAs with no role in development, or might have developmental roles outside of the embryonic stages considered for this study. On the other hand, *miR-31b* and *miR-1b* were found to be early development specific for the sea urchin, with no expression in the adult (**Figure 3.5**). The most surprising result was *let-7* reads in the sea urchin embryos. Pasquinelli *et al.* (Pasquinelli et al. 2000), using northern blots, had shown that *S. purpuratus* embryos contain the *let-7* precursors, but not the mature *let-7* miRNA. We found 16 high-quality reads corresponding to this miRNA in our sample. We suspect that the relatively low abundance of this gene made it undetectable to northern blots. **Appendix C** shows the differences in sequence of *S. purpuratus* mature miRNAs between embryonic (Illumina sequencing) data and the adult 454 sequencing data. Most sequences are the same and few differences are seen at the 5' or 3' end. However, *miR-31b* shows a difference of one base at position 11.

There is no adult miRNA data for the *P. miniata* (PMI). However, Wheeler *et al.* (B. M. Wheeler et al. 2009) sequenced a species of sea star, *H. sanguinolenta* (HSN). On comparison of the PMI embryo data with the HSN adult data, 34 miRNAs were found in both species, 13 were found in HSN only and 8 were found in PMI only (**Appendix D**). Some changes are seen between the sequences of the same miRNA (indicated by *bold letters* in **Appendix D**) but most of these are at the 3' end of the miRNA and could be due to some disparity in the results from different sequencing platforms or due to sequencing errors. The presence or absence of miRNAs between the two datasets might be due to different developmental stages, and might not represent species level changes.

In summary, we find that the pool of miRNAs is more or less conserved between embryonic and adult sea urchin. When we compared the developmentally expressed miRNAs between the two species we find that majority of them were conserved, although some relatively highly abundant miRNAs in sea urchin embryos did not have any reads in sea star embryos (for example, *miR-2008*) (**Figure 3.5b**). The overall conservation of miRNA genes may imply that possible differences in miRNA function may be due to differences in their spatial expression or their expression levels, or differences in their target genes and their expression.

3.2.3 miRNA gene expression shows similar trends between the two echinoderm embryos

Figure 3.5b shows a heat map corresponding to relative abundance of overlapping miRNAs between the sea urchin and sea star embryos. The miRNAs can be classified into 4 main groups based on their expression trends, (**1)** relatively high abundance in both species, **(2)**

relatively high abundance in sea star embryos, but lower abundance in sea urchin embryos, **(3)** relatively high abundance in sea urchin embryos, but low abundance in sea star embryos, and **(4)** medium to low abundance in both species. Overall, we found that most miRNAs show similar patterns of expression in the two species. This indicates that the two echinoderms may share many features of their regulatory programs. However, some differences are also become apparent. Out of the 14 highly expressed sea urchin miRNAs, 11 are also relatively highly expressed in sea star, which may indicate possible overlap in the post-transcriptional gene regulatory mechanisms. From the remaining three, two (*miR-183* and *miR-1a*) are of medium abundance in sea star, while *miR-2008* has a single read in sea star library (**Figure 3.5b**). On the other hand, three highly expressed and one moderately expressed miRNA in sea star (*miR-1692*, *miR-100*, and *miR-92d*; and *miR-4171*, respectively) have no reads in the sea urchin library (**Figure 3.5b**). These differentially expressed miRNAs are probably indicative of the differences between the two developmental programs. We note, however, that this is the first attempt to map the developmental post-transcriptional regulome in echinoderms, and spatial as well as temporal expression may vary even between the miRNAs that appear to be abundant in both species.

Since the embryonic libraries were made from a mixed population sample (i.e., different developmental stages sequenced together), we used northern blots of a few miRNAs in various early developmental stages of sea urchin and sea star embryos were used to study the spatial expression patterns of some conserved miRNAs (**Figure 3.7**). *miR-2009* was found in 1day, 2day and 3day old embryos in both species. *miR-31* and *miR-10* was found in all stages considered in sea urchin and sea star respectively. *miR-184* was only barely visible on the 3day old embryos of sea urchin with undetectable levels in 1day and 2day old embryos, and might be more development specific than the other miRNAs. However, the signal levels for sea star were

undetectable. This might be due to the low sensitivity of the protocol (See **Materials & Methods**). It will be interesting to use whole mount in situ hybridization to compare the spatial and temporal patterns of these miRNAs (See **3.2.5**).

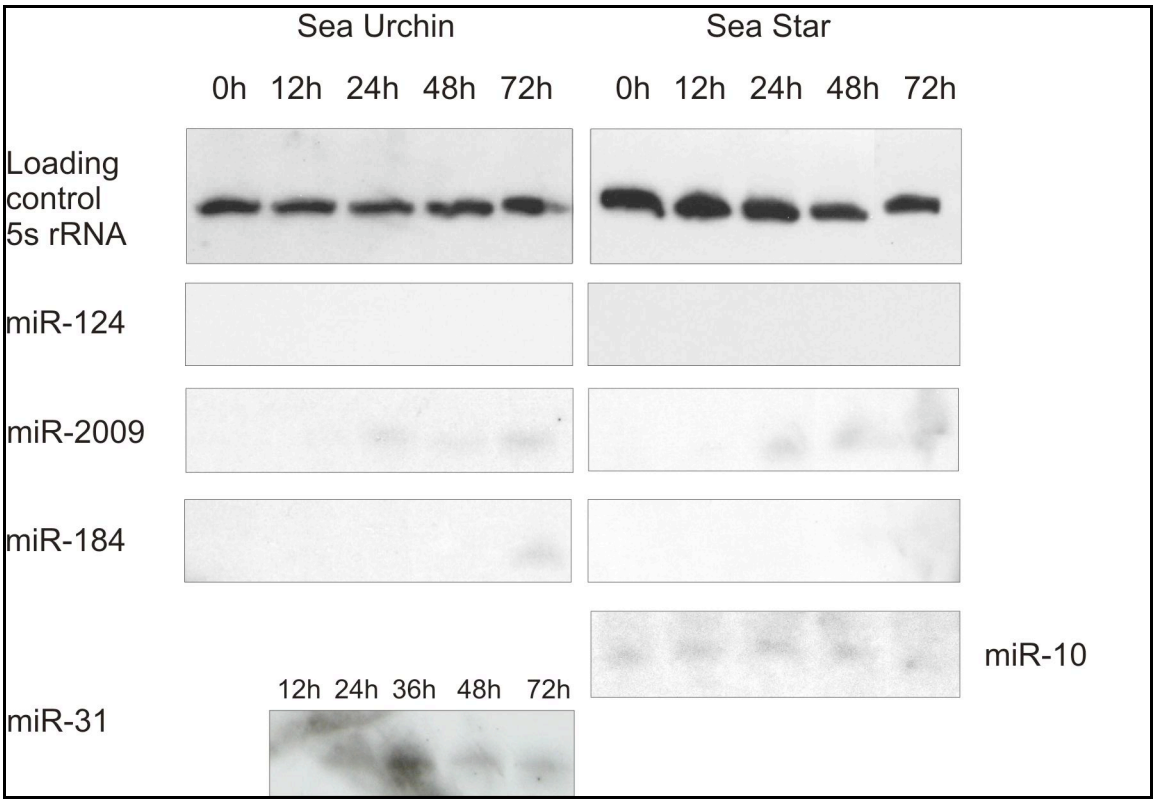


Figure 3.7: Northern Blot showing the expression of a few conserved miRNAs in *S. purpuratus* (sea urchin) and *P. miniata* (sea star) embryos. 5S rRNA is used as the loading control while *miR-124* is used as the negative control.

3.2.4 Evolution of miRNA sequences in the echinoderm animal lineage

miRNA families have been found in all analyzed animal lineages. It has been shown that evolutionary trends across metazoans show rare substitutions in mature miRNA sequence (B. M. Wheeler et al. 2009). We found that about half of the miRNAs in sea urchin and sea star are identical in sequence, and the rest have acquired single or multiple mutations. All alignments between the three species are listed in **Appendix E**. Many of these differences are at the 3' end of the miRNA, and represent the addition or loss of two or more bases. A mutation at the last base of the miRNA between two species is not treated as a change, as this might be a sequencing error and in any case it is not expected to affect the function of the mature miRNA. Differences at the 3' end may be due to differences in the processing of the miRNA precursors between the two species. Striking differences are seen in abundant miRNAs such as, *miR-2001*, *miR-182, miR-183, miR-2007* and *miR-92b*, where the mutation(s) occurs in the middle of the sequence (**Appendix E**). **Figure 3.9** shows the comparative analysis of mutations in miRNAs between the two echinoderms, using the hemichordate, acorn worm, *Saccoglossus kowalevskii* as an outgroup. The miRNAs can be grouped in several clusters based on the mutations across evolutionarily divergent species (**Figure 3.9**). Only ten of the 28 miRNAs that are present in all three species (**Figure 3.9**, categories A, B, and C) are identical in all of them; seven seem to have acquired mutations in the *S. kowalevskii* lineage (or in the echinoderm ancestor), five in the sea urchin lineage and only two in the sea star lineage. The remaining four miRNAs have differences in all three species (**Figure 3.9**, category B). It will be very interesting to further investigate the effects of these mutations on the loss or gain of target sequences between the two echinoderms.

A very interesting observation was seen with *miR-2008*, which seemed to present in *S. purpuratus* and *S.kowalevskii*, but not in *P. miniata* based on our library data. Whole mount in situ hybridization on late stage sea star embryos showed that *miR-2008* is indeed present in sea star, but is not expressed in the early stage embryos considered for our library preparation (**Figure 3.8**).

We, thus, anticipate that our dataset will provide a rich source for future evolutionary studies, as both the miRNA and target sites may have evolved quite rapidly to facilitate new regulatory interactions.

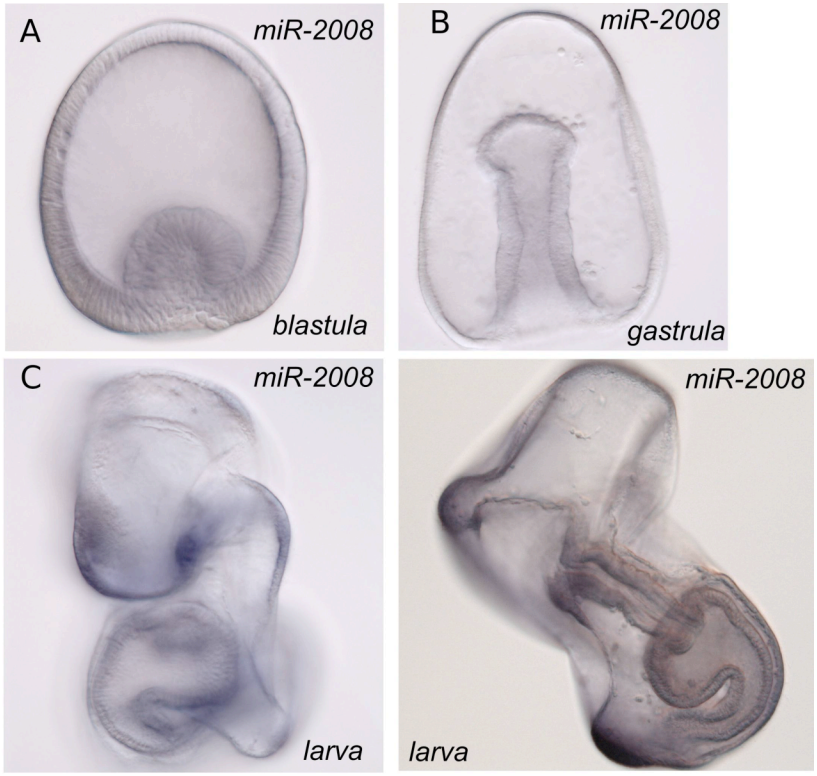


Figure 3.8: Whole mount in situ hybridization of *P. miniata* embryos using LNA probes antisense to *miR-2008*. Blastula and gastrula stages do not show any expression for this miRNA, consistent with the embryonic small RNA library. However, we see expression of *miR-2008* in late stage larvae.

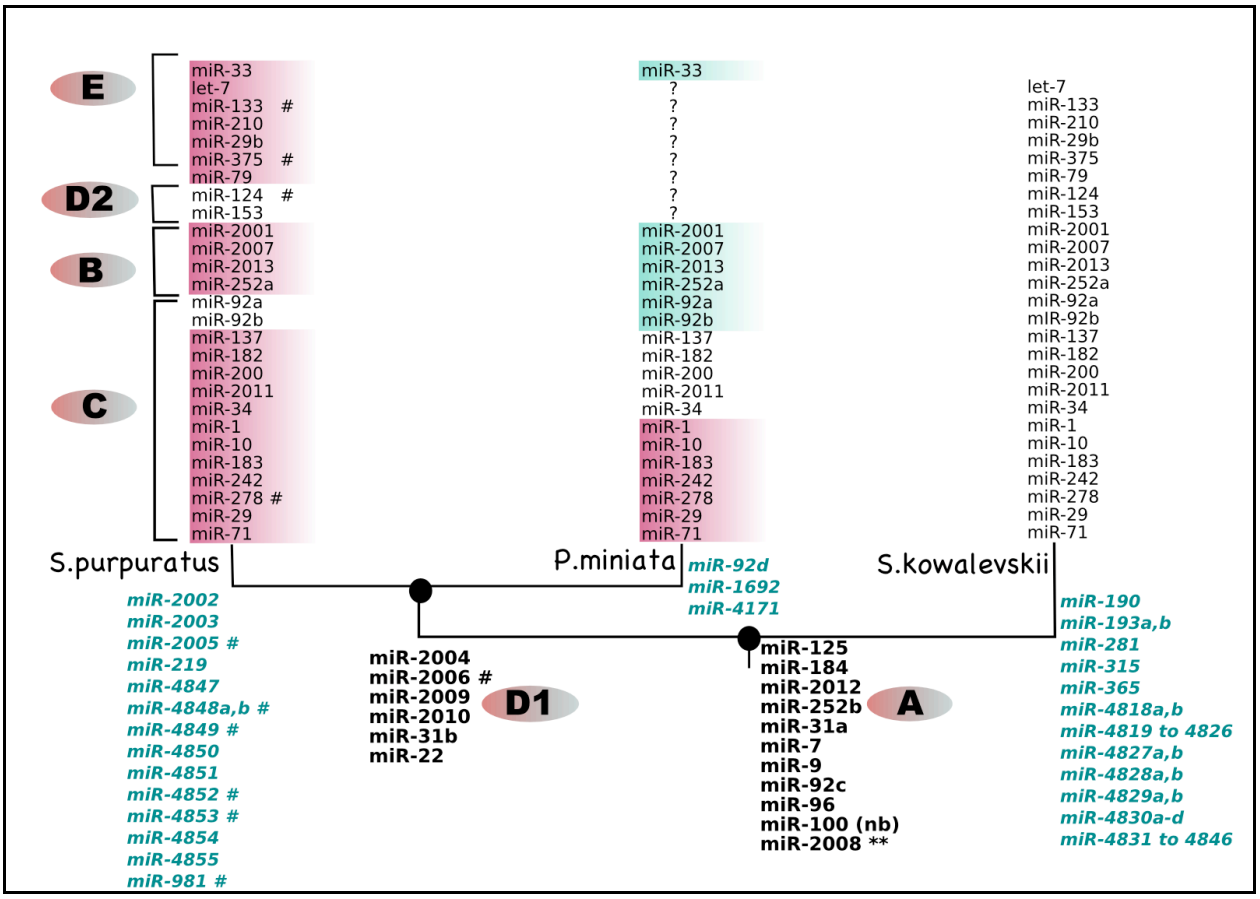


Figure 3.9: Phylogenetic comparison of sequence similarities between sea urchin, *S. purpuratus* and sea star, *P. miniata*. The hemichordate, S. kowalevskii has been used as the outgroup and the sequences in that species are used as the reference sequences. miRNA sequences in S. purpuratus or P. miniata that differ from the reference sequence are colored. Same color represents identical sequences. Absence of a miRNA from a species (represented by a blank) indicates absence of that miRNA from the reads and the registry. The miRNAs can be classified into 6 groups: **(A)** identical sequence and present in all three species; **(B)** present in all three species, but the sequence differences in all miRNAs; **(C)** present in all three species, but one or more species show mutations; **(D1)** identical sequence and present in S. purpuratus and P. miniata; **(D2)** identical sequence and present in S. purpuratus and S.kowalevskii; **(E)** present in two species with difference(s) in sequence; **(F)** the gene gained in a single species or lost in other two species. Group F is represented by the blue miRNAs at the node for the specific species; # : miRNA is in the registry but has ≤2 read frequency in the embryonic reads; nb: miRNA was shown to be present in adult tissue by northern blot (Sempere et al. 2006) but is not present in registry. **: miR-2008 was found in late sea star embryos by whole mount in situ hybridization but not in early embryos (**Figure 3.8**)

3.2.5 Localization of miRNA expression using whole mount in situ hybridization in sea urchin embryos

Once the list of miRNAs in the developing sea urchin is obtained, their spatial localization can be determined using whole mount in situ hybridization (WMISH). Spatial information does not only validate the presence of the selected miRNAs, but also serve as the first step to de-convolute the set of potential direct targets, and the function of miRNAs. Selection of miRNAs for this step was based on several criteria, such as, conservation across the lineages, relative abundance, and evolutionary and/or functional information in other species (from **Figure 3.5** & **Figure 3.9**). At the same time, northern blots, as described above, were used to obtain temporal information about the miRNAs, while providing additional validation (**Figure 3.7**).

Locked nucleic acid (LNA) probes were used for WMISH due to the short size of miRNAs and the high specificity of LNA compared to oligonucleotide probes. LNA probes have become the standard for miRNA localization studies, and have been broadly used in zebrafish and mouse (Wienholds & R. H. A. Plasterk 2005; Kloosterman et al. 2006). We faced challenges with background staining in my experiments with single (3') DIGoxigenin (DIG) labeled probes. Thus, we used double DIG labeled probes from Exiqon and high Tween buffer washes between multiple color reactions to eliminate/reduce background. See **Materials & Methods** for details.

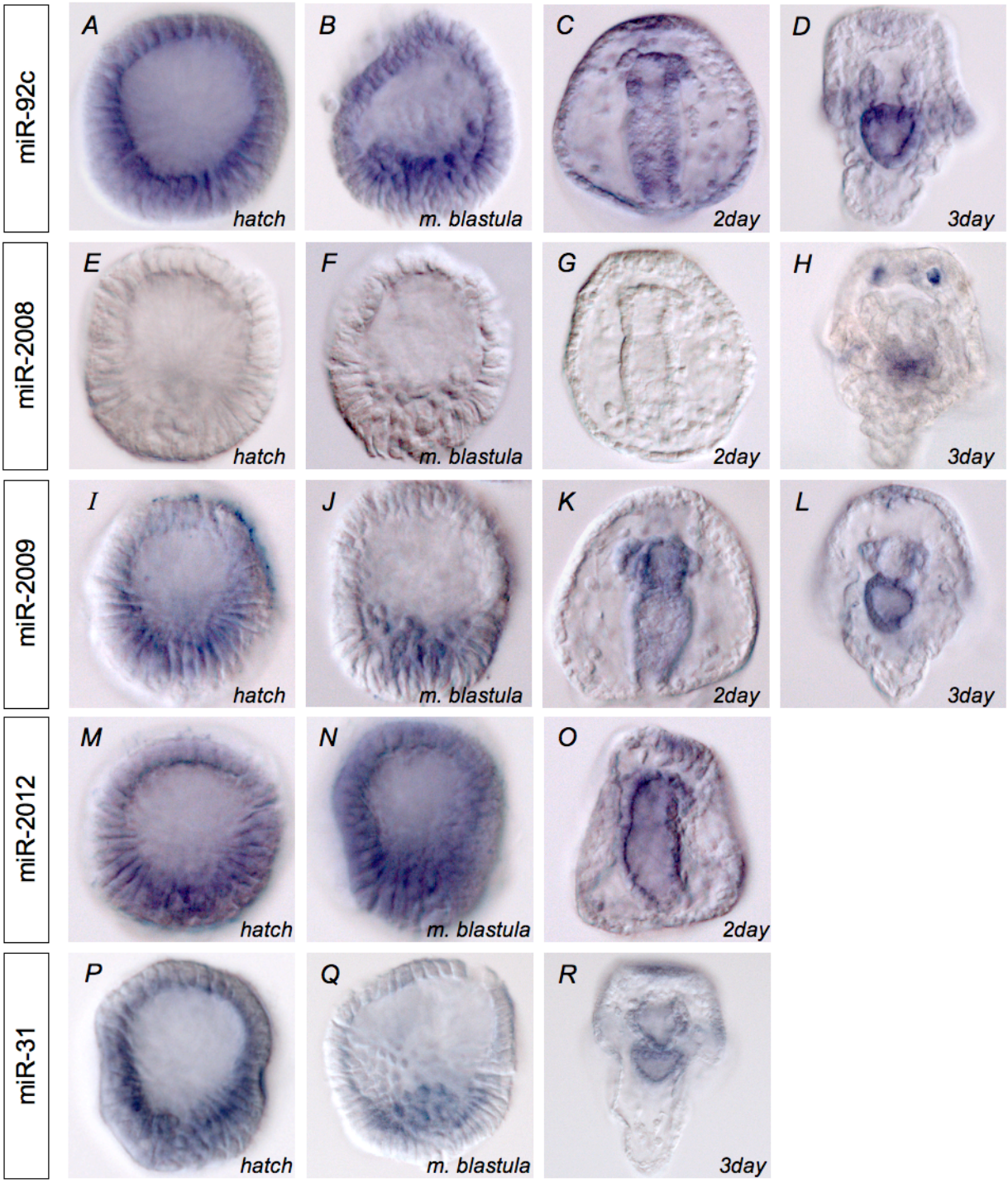


Figure 3.10: WMISH using LNA probes of selected miRNAs in sea urchin embryos: WMISH was performed for four miRNAs found in cluster 1 (highly abundant in sea urchin and sea star) (data from **Figure 3.5b**) (*miR-92c* in A-D; *miR-2009* in I-L; *miR-2012* in M-O; *miR-31* in P-R) and for a miRNA found in the sea urchin but not sea star library (*miR-2008* (E-H). *miR-2009* (I-L) is an echinoderm specific miRNA whereas the other four are highly conserved in multiple species (**Figure 3.9**).

With the exception of *miR-2008*, all miRNAs investigated were found to be ubiquitous during the hatch blastula stage (**Figure 3.10A,I,M,P**), after which *miR-92c* and *miR-2012* continue to be ubiquitous during mesenchyme blastula (**Figure 3.10B,N**) while *miR-2009* and *miR-31* are localized to the Primary Mesenchymal Cells (PMCs) (**Figure 3.10J,Q**). It is important to note that although these assays were reproducible, the quality of the color staining was not as good as those seen with oligonucleotide probes for protein-coding genes. This could be an effect of hybridization of LNA probes in the sea urchin system or the result of non-specific binding.

miR-2009, an echinoderm-specific miRNA (**Figure 3.9**), was found to be localized in the foregut and midgut at 48hrs of development (**Figure 3.10K**). This expression expands to gut, ciliary band and apical plate at 3 days of development (**Figure 3.10L**). *miR*-31, the most abundant miRNA in the sea urchin library has the same expression pattern at this developmental stage. As explained in **Section 3.2.4**, *miR-2008* is expressed late in sea star development, and was not found in the early embryonic libraries. This miRNA was not found in the hatch, blastula or gastrula stages of development in the sea urchin (**Figure 3.10E-G**), but came up at 72hrs of development - the pluteus stage (**Figure 3.10H**), in two cells at the animal pole and in some part of the gut.

All five miRNAs studied are expressed in the gut region (**Figure 3.10D,H,L,R**), with differences seen in expression in the ectodermal region. As noted above, *miR-2009* and *miR31* are expressed in the ciliary band and apical plate, *miR-92c* is only seen in the ciliary band, whereas *miR-2012* is only seen in the apical plate region. *miR-2008* is not found in either of these ectodermal regions, but found in two animal pole cells, as explained above.

LNA probes are extremely sensitive to temperature and concentration. For example, a difference of 2°C in the hybridization temperature can drastically affect the result, as observed with hybridization of *miR-2009* probe at 46 and 48 degrees Celsius (data not shown). We suspect that some of the stained embryos have relatively more background than the others, and more optimization can yield better results.

We also made some preliminary attempts at WMISH using the primary transcripts of the miRNAs. Based on the small RNA sequencing data and genomic locations, we identified miRNA clusters for *S. purpuratus*, that is miRNAs that are located within 5kb (for our purpose) of each other, and probably part of the same primary transcript. We amplified parts of the primary transcript between two miRNAs and synthesized DIG labeled probes using the amplicon. This approach has the advantage of a longer probe (~1kb compared to 22nts of the LNA probe). Probes were made and are ready for future use.

3.2.6 Data visualization

We created custom tracks on the UCSC genome browser (Kent et al. 2002) to visualize the frequencies of mapped reads for the sea urchin genome (**Figure 3.11**). We added our annotated miRNA names as part of this track. This makes it convenient to study annotated regions of the genome in context of their expression in the small RNA library. An example of this visualization can be seen in **Figure 3.11**, where only regions around the miRNAs in miRNA clusters are expressed whereas the intergenic regions between the miRNAs are not expressed. This also provides further proof that the miRNAs we identified are real, functional, developmentally expressed genes.

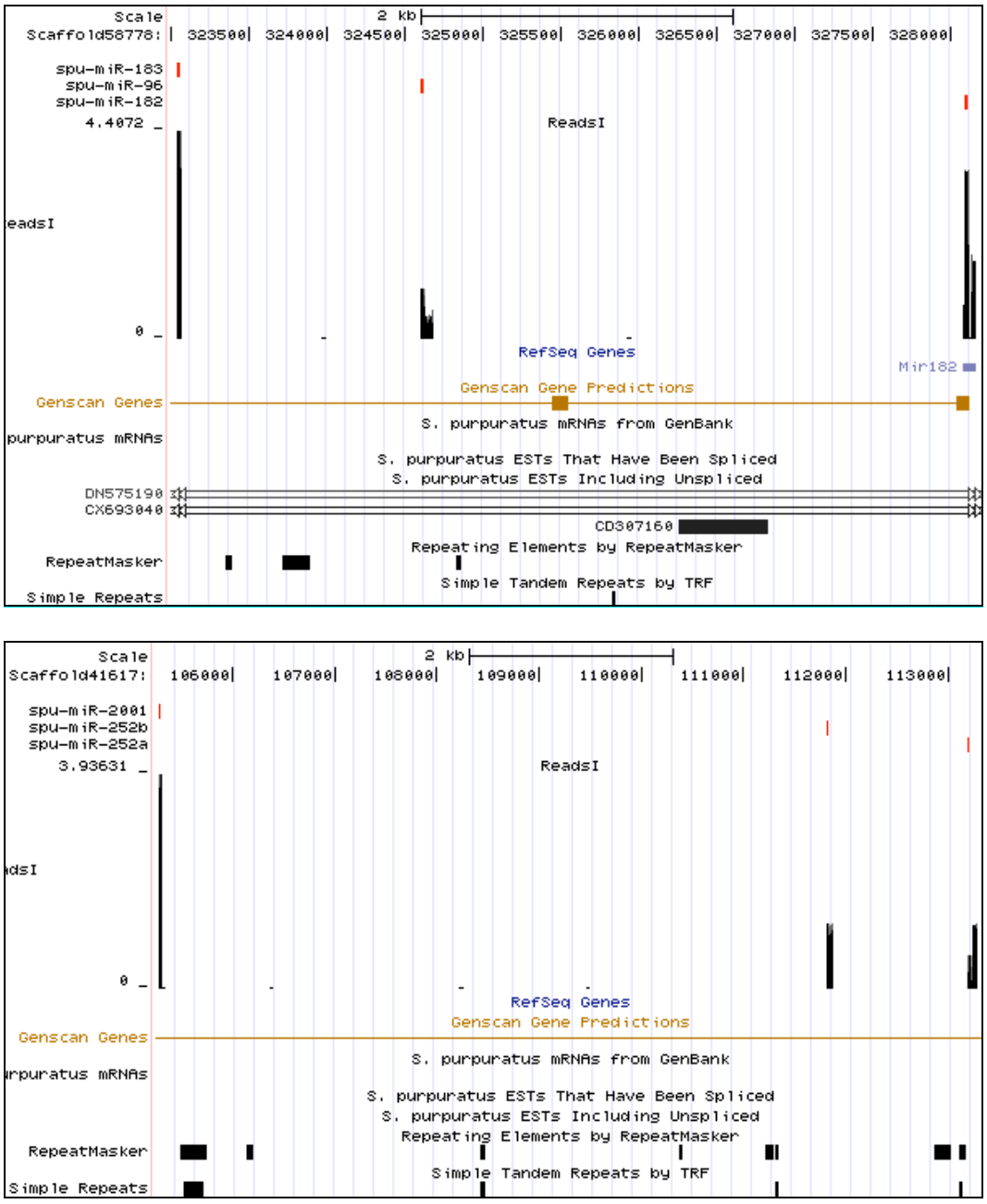


Figure 3.11 Data visualization using the UCSC genome browser: Read frequencies for clusters of *miR-183*, *miR-96*, *miR-182* (top panel) and *miR-2001*, *miR-252a*, *miR-252b* (bottom panel) as seen in the custom tracks made for data visualization in the UCSC genome browser (Kent et al. 2002).

3.3 Materials & Methods

3.3.1 Small RNA library preparation

Sea urchins and sea stars were collected by Marinus Scientific LLC in Southern California (http://www.marinusscientific.com/) and purchased by us. Total RNA was extracted from embryos at 24h, 48h and 72h after fertilization using miRVana RNA isolation kit (Ambion). Embryo populations were combined, separately for each species, and the mixed population samples were sent for small RNA library preparation and sequencing to the Genomics & Microarray Facility at Wistar Institute, Philadelphia. Prior to library preparation, RNA quality was checked using the Bioanalyzer and was found to be very good with very little degradation (**Figure 3.1**).

The Bioanalyzer profiles for the total RNA of the sea urchin and sea star embryos are presented in **Figure 3.1a** and **Figure 3.1b**. These results indicate good RNA quality. There are three prominent peaks – the smallest peak corresponds to 5s rRNA and small RNAs in the sample, while the other peaks correspond to 18s and 28s rRNAs respectively. Although the rRNA profile is not a definitive indication of RNA integrity, the profiles show that smaller RNAs are not hidden by degraded products of different sizes. The gel in **Figure 3.1c** shows the RNA after adapter ligation. The highlighted band was excised to run on the Illumina Genome Analyzer (Illumina, Inc.).

Illumina adapters were ligated to the 5' and 3' ends of RNA, as described in the Illumina v1.5 protocol for small RNA sequencing samples. Small RNA molecules were size selected (**Figure 3.1**), and RT-PCR amplification was used to generate the cDNA libraries for both

species. The 36bp run on the Illumina Genome Analyzer (Illumina, Inc.) was used for sequencing these cDNAs.

3.3.2 Computational analysis procedure and pipeline

Base calling was performed by the Bioinformatics facility at Wistar Institute. The resulting sequences were subjected to our computational pipeline, which consists of a number of in-house made scripts (**Figure 3.12**). Briefly, pre-processing steps involve low quality read filtering, 3' adapter removal, and minimum length filtering (for us, *n*=17 nts). First, we performed quality filtering by converting the Illumina quality codes for each base to its Phred quality score, and trimming the low quality 3' ends of the reads. A cut-off of 20 was selected based on the histogram of qualities for all reads (data not shown). 3' adapters were trimmed using the *novoalign* program (*www.novocraft.com*). This program uses ungapped semi-global alignment of adapter sequence against the read using a weight matrix from read and base qualities, and trimming is performed from start of the optimum alignment. 5' adapter sequence was trimmed based on perfect sequence match of more than 10 nts at the 5' end. All reads shorter than 17nts were removed from this dataset using the minimum length filter. Reads shorter than 17nts can give many non-specific hits in subsequent mappings. A total of 7.8% sea urchin and 6.8% sea star reads were discarded in these steps (**Figure 3.12**). The remaining reads are aligned to produce *tags* of genes and calculate their expression as number of independent reads each tag has. Reads with 100% sequence identity and length difference of 2 nts or less were collapsed. All sequences matching other non-coding RNAs – tRNAs, rRNAs, snRNAs, snoRNAs – were excluded from further analysis (*S. purpuratus:* 16,727 tags; 301,803 reads; *P. miniata:* 22,033 tags; 354,391 reads) (**Table 3.1** and **Figure 3.12**). Also, similarity to known

miRNA genes is used to identify evolutionary conserved miRNAs. If a genome is available (i.e., sea urchin, in our case) the reads are mapped to the genome and novel miRNA genes are discovered using miRDeep (M. R. Friedländer et al. 2008), following the authors' instructions.

Sea urchin tRNA sequences were obtained from UCSC (http://gtrnadb.ucsc.edu/); and snRNA and snoRNA sequences from NCBI. rRNA sequences were gathered from a variety of sources for three sea urchin species (*S. purpuratus, P. lividus, L. variegatus)*, including UCSC genome browser (Kent et al. 2002) and EBI databases (http://www.ebi.ac.uk/Databases/). Since there is no tRNA, snoRNA or snRNA data publicly available for the sea star, the sequences from sea urchin were used for the search in sea star. For sequence similarity match we used BLAST (Altschul et al. 1990). The parameters used to map miRNAs to Illumina reads were -e 0.01 -p 100 -W 8. The word size chosen was based on the size of the miRNA seed region. For mapping reads to the genome and other conserved sequences, parameters used were -W 12 -p 80. All hits with length less than 85% of the length of the query sequence were ignored. mRNA sequences for the sea urchin and sea star were compiled using NCBI predicted genes (D. A. Benson et al. 2008) and the SpBase (http://spbase.org) database (Preker et al. 2008; Taft et al. 2009) was also used for *S. purpuratus*.

The computational pipeline to analyze Illumina reads is available at http://www.benoslab.pitt.edu/services.html

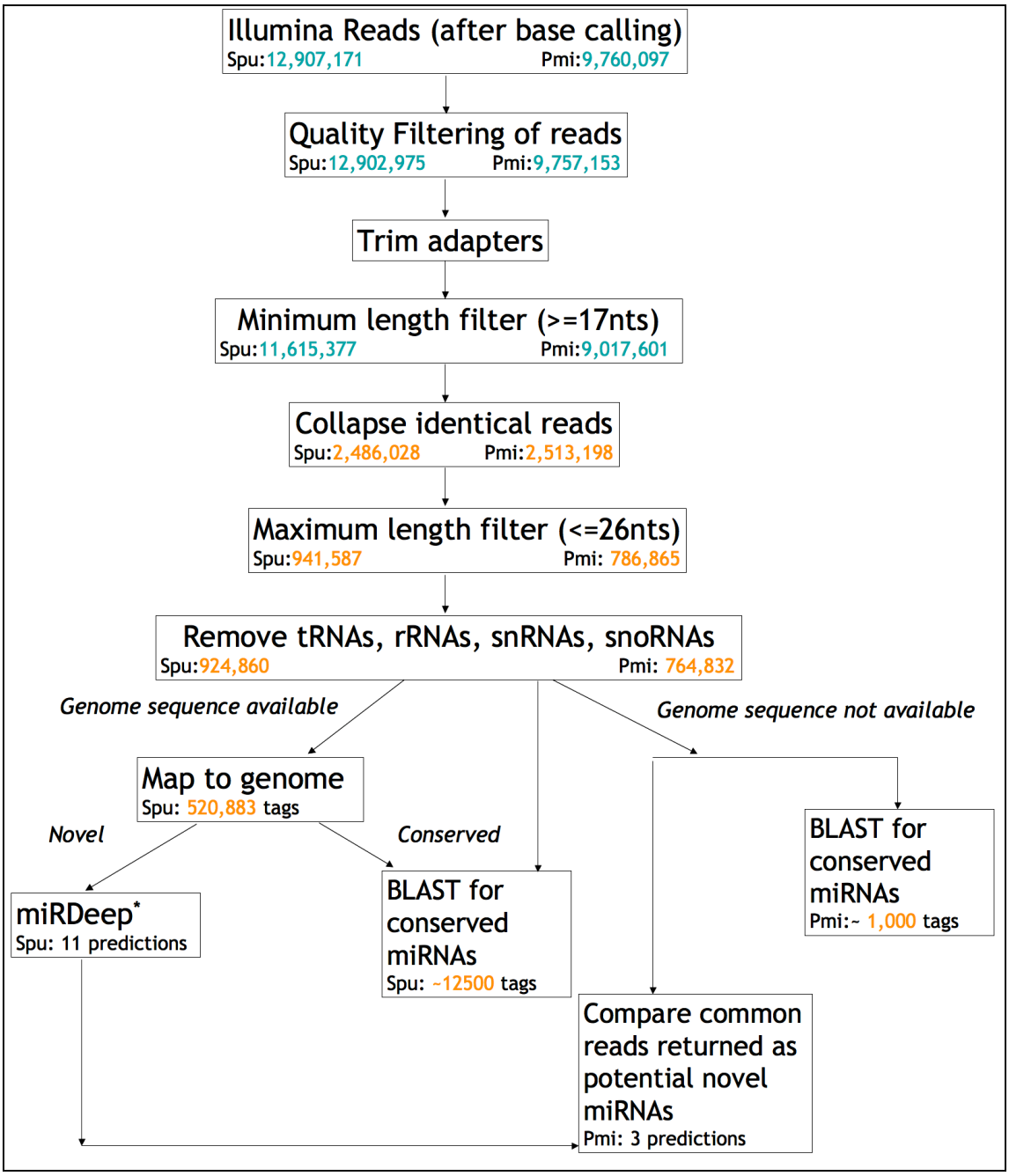


Figure 3.12: Computational pipeline for analysis of deep sequencing libraries for discovery of small non-coding RNAs. Illumina reads undergo numerous filtering steps based on quality and length. The pipeline has two branches: for a species with genome sequence, and for a species without a sequenced genome, but a closely related sequenced species. *Spu: Strongylocentrotus purpuratus; Pmi: Patiria miniata.* miRDeep **(M. R. Friedländer et al. 2008)**; BLAST (Altschul et al. 1990) *Green color:* Reads *Orange:* Tags

3.3.3 Hierarchical clustering of gene expression values

The relative abundance of each miRNA in each sample was log2 transformed for better visualization of the data. Average linkage hierarchical clustering was performed using Euclidean distance as the distance metric. The distance between two clusters X and Y is given by:

$$D(X,Y) = \frac{\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} ed\left(x_i, y_j\right)}{n_X n_Y},\ x_i \in X, y_j \in Y,$$

where x_i is the vector of log2 transformed relative abundances of miRNA *i*, y_j is the vector of log2 transformed relative abundances of miRNA *j*, $ed\left(x_i, y_j\right)$ is the Euclidean distance between $x_i \in X$ and $y_j \in Y$, n_X is the number of samples in cluster X, n_Y is the number of samples in cluster Y.

3.3.4 Whole mount in situ hybridization

We followed our lab protocol (V. F. Hinman, A. T. Nguyen, R Andrew Cameron, et al. 2003; V. F. Hinman, A. T. Nguyen & Eric H Davidson 2003) except we used an antisense double DIG labeled locked nucleic acid (LNA) probe (Exiqon Inc.) at concentrations of between 2pmol to 4pmol per 100ul of hybridization solution and at 20-22°C below the melting temperature of the probe as recommended by the supplier. The proteinase-K treatment was used for the *miR-31* probe. Hybridization was carried out at 42°C at 0.03pmol/µl. Hybridization for *miR-2008*, *miR-2012*, and *miR-2009* was carried out at 46°C at 0.02pmol/µl. *miR-92c* probe showed staining at 0.01pmol/µl and a high hybridization temperature of 58.5°C.

3.3.5 Northern Blot

We extracted total RNA from sea urchin and starfish embryos using the miRVana kit by Ambion. Standard northern blot protocols were performed using 10-15µg of total RNA. The RNA was run on a 15% polyacrylamide gel and transferred to a membrane. Decade(Ambion) markers were labeled with γ- P^{32}, according to manufacturer's protocol. These size markers produce a 10 nt RNA size ladder from 10 to 100 nt, and are used to estimate size. Antisense miRNA Starfire™ (IDT) α-P^{32} oligonucleotide labeled probes were hybridized to the membrane, and the exposed film was observed for bands of correct size, corresponding to the miRNA being tested. See **APPENDIX F1** for details on the Northern blot primer sequences.

4.0 MIRNA PATHWAY IS NECESSARY FOR NORMAL DEVELOPMENT OF SEA URCHIN EMBRYOS

The miRNA pathway has important roles in normal embryonic development of a variety of animals like mouse, mammals, zebrafish, chicken, *C. elegans,* Drosophila and plants like Arabidopsis as explained in **Chapter 1.0** , as well as stem cell differentiation, embryogenesis, and developmental timing (Emily Bernstein et al. 2003; W. J. Yang 2004; Zhao et al. 2008; V. Ambros 2003; Alvarez-Garcia & E. A. Miska 2005; Willmann et al. 2011; Blakaj & H. Lin 2008; Bannister et al. 2009; Suh & Blelloch 2011; Kloosterman et al. 2007; Prather et al. 2009). See **Section 1.1.4** for more details.

The transcription factor gene regulatory networks (GRNs) in the sea urchin continue to get increasingly detailed, with a complexity unmatched in any other developmental model system (**Figure 1.2A**), but there is no information on the post-transcriptional layer of gene regulation in this system. Approximately 80% transcription factors identified in the sea urchin genome, are expressed during embryogenesis (Howardashby et al. 2006), along with a rich miRNA population (**Chapter 3.0**). This does not imply how miRNAs are used in development, but it does highlight the complexity of sea urchin gene-gene interaction networks.

The only study of miRNA pathway genes in the sea urchin has been in 2007 by Song & Wessel (Song & Wessel 2007). No follow up studies exist to study the function of this pathway in this model system. Here, we knocked down some key components of the miRNA biogenesis

pathway in early sea urchin embryos, and studied the effects on embryo morphology as well as expression patterns of differentiation gene markers that may be downstream of this pathway. We suggest a future high-throughput direction to this project, and show preliminary work for the same.

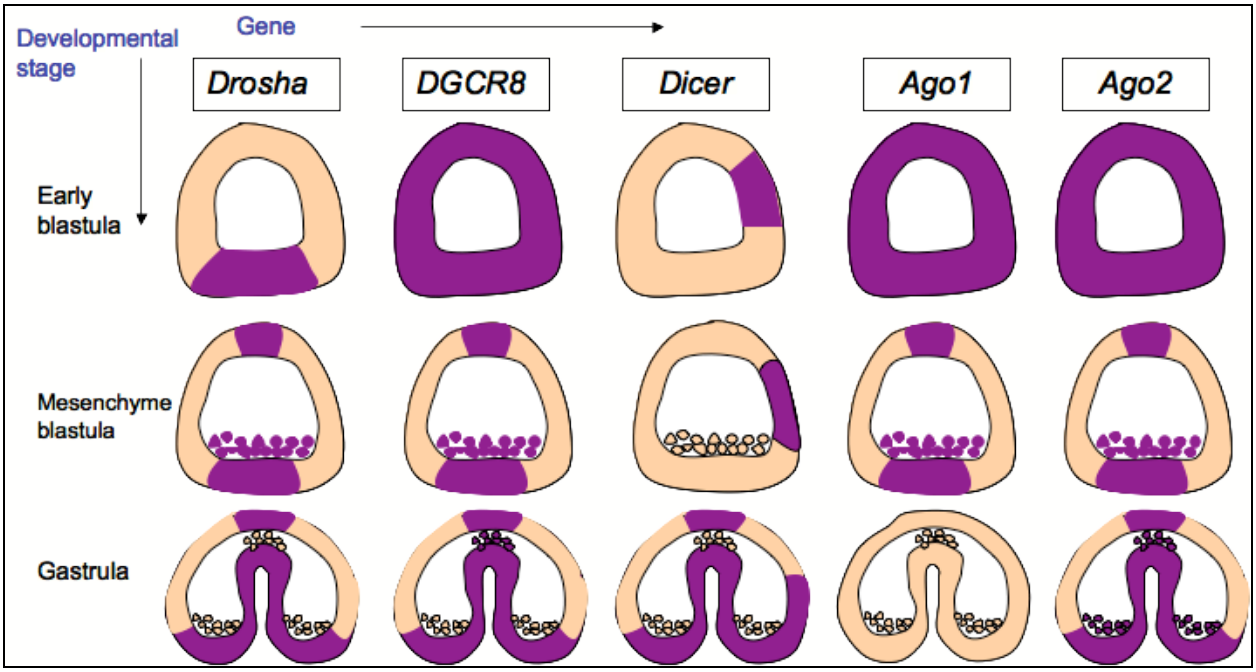


Figure 4.1: Dynamic expression of miRNA biogenesis genes in sea urchin embryos: The expression patterns are based on the conclusions in (Song & Wessel 2007). The rows represent the early developmental stages of the sea urchin embryos, while the columns represent expression of a particular miRNA biogenesis pathway gene. The purple color represents the expression of the specific gene at the specific developmental time-point.

4.1 Introduction

4.1.1 Biogenesis genes are expressed in sea urchin embryos

Song & Wessel (Song & Wessel 2007) did bioinformatics searches in *S. purpuratus* and found homologs of protein-coding genes of the miRNA biogenesis pathway. They used quantitative real time polymerase chain reaction (QRT-PCR) and whole mount in situ hybridization (WMISH) to study spatial and temporal expression patterns of these genes involved in the RNAi pathway, in developing sea urchin embryos. They concluded that dynamic expressions suggest that sea urchin embryos use the RNAi pathway selectively during development.

Figure 4.1 shows a cartoon representation of the expression patterns of some of the genes crucial for miRNA biogenesis in early sea urchin development. As shown in **Figure 1.1**, Drosha & DGCR8 are part of a complex that processes primary transcripts into miRNA precursors, while Dicer processes the precursors into the mature miRNA duplex. The sea urchin has 2 Argonaute (AGO) proteins that are necessary components of the protein-RNA complex that helps target miRNAs to their target sites. The problem with some expression patterns seen in **Figure 4.1** is the *lack of overlap in the embryonic domains* where the genes are expressed. We expect that all genes in the pathway should have some overlapping territories, except if they have other functions in another territory.

According to their results, some of the genes involved in this pathway have mutually exclusive territories. For example *Drosha* and *Dicer* in top panel of **Figure 4.1**, are seen to be expressed in mutually exclusive territories. In the mesenchyme blastula stage, all genes shown have the same expression pattern in the vegetal plate, apical ectoderm and primary mesenchyme

cells, except *Dicer*. By the gastrula stage, all genes except *Ago1* are expressed in the endoderm and apical ectoderm. The expression of *Ago1* is cleared after early gastrula. Due to poor quality of the in situ hybridizations in the paper, and lack of information on protein localization, it is difficult to speculate on the use of this pathway or its expression.

To confirm the time-points at which sea urchin embryos express *Ago1* during development, I designed RT-PCR primers around a region that differs between *Ago1* and *Ago2* (See **APPENDIX F**). RT-PCR indicated that the sea urchin embryos express *Ago1* during development from egg through 48 hours post fertilization (hpf) (See **Figure 4.2b**). We also repeated some QRT-PCR assays, similar to those by Song &Wessel (Song & Wessel 2007) (**Figure 4.2c**), and we found some slightly different trends. We saw that both *Dicer* and *Ago1* (and not just *Ago1*) have slightly higher expression than other stages at 24hpf. Overall, we saw that all three genes investigated are expressed in the developmental stages in question.

4.1.2 Biogenesis genes expressed in sea star embryos

To check for the existence of RNAi genes in sea star, Kristen Yankura in the Hinman Lab carried out library screens for the cytoplasmic RNase III enzyme, *Dicer*, and the nuclear microprocessor complex component, *DGCR8*. *Dicer* is essential for processing of mature miRNAs and DGCR8 has been shown to provide stability to the *Drosha*: pri-miRNA complex (V. N. Kim et al. 2009). Clones obtained from the library screens were used to design primers for quantitative reverse transcriptase- polymerase chain reaction (QRT-PCR). It was observed that the QRT-PCR cycle thresholds are relative to those obtained for some important regulatory genes in sea stars (V. F. Hinman, A. T. Nguyen, R Andrew Cameron, et al. 2003) (**Figure 4.2d**). Thus, transcripts of two

of the major RNAi pathway genes are present in sea star embryos during developmental stages in question.

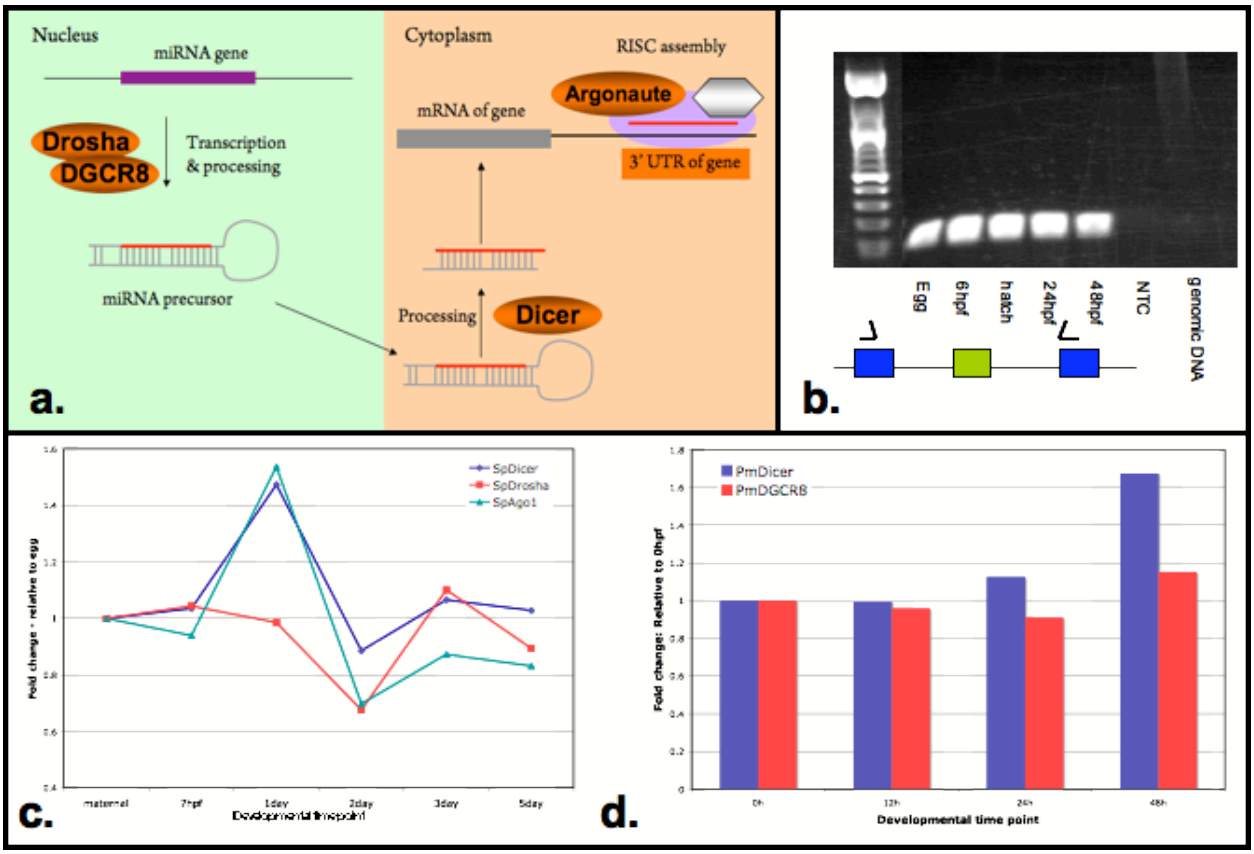


Figure 4.2: miRNA biogenesis genes: (a) Cartoon representation of key genes involved in miRNA biogenesis. *Drosha* and *DGCR8* process the primary transcript in the nucleus into the miRNA precursor, which is then processed by *Dicer* into the mature miRNA. *Argonaute* is a critical component of the protein:RNA complex that is necessary for the miRNA to bind to its target sequence. **(b)** RT-PCR showing the presence of *Ago1* in sea urchin from Egg through 48hpf. NTC: No template control. Cartoon below the gel shows the location of the primers in the *Ago1* gene. The reverse primer was designed two exons downstream of the forward primer exon. **(c)** QRT-PCR results miRNA biogenesis genes in sea urchin (Sp). The y-axis represents the fold change relative to the Egg (maternal). The number of transcripts estimated in egg was 910 for *SpDicer,* 350 for *SpDrosha* and 670 for *SpAgo1*. **(d)** QRT-PCR results two miRNA biogenesis genes in sea star (Pm). The y-axis represents the fold change relative to the 0hpf (embryo immediately after fertilization).

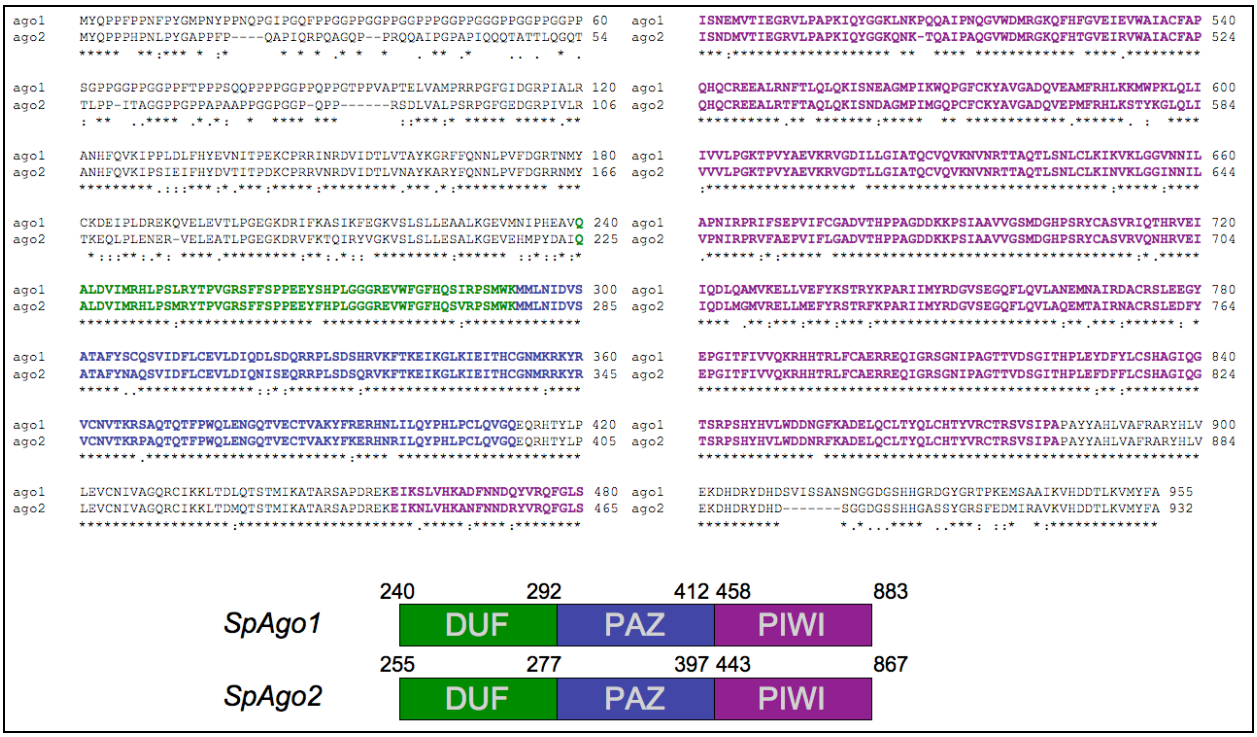


Figure 4.3: Alignment of the sea urchin *Argonaute* proteins: ClustalW (Goujon et al. 2010; Larkin et al. 2007) alignment of protein sequences of the two *S. purpuratus* Argonautes. ago1: *SpAgo1*; ago2: *SpAgo2.* The three colored domains are based on NCBI domain predictions.

4.2 Methods & Results

4.2.1 Knockdown of miRNA biogenesis genes is performed using morpholino antisense oligonucleotide technology

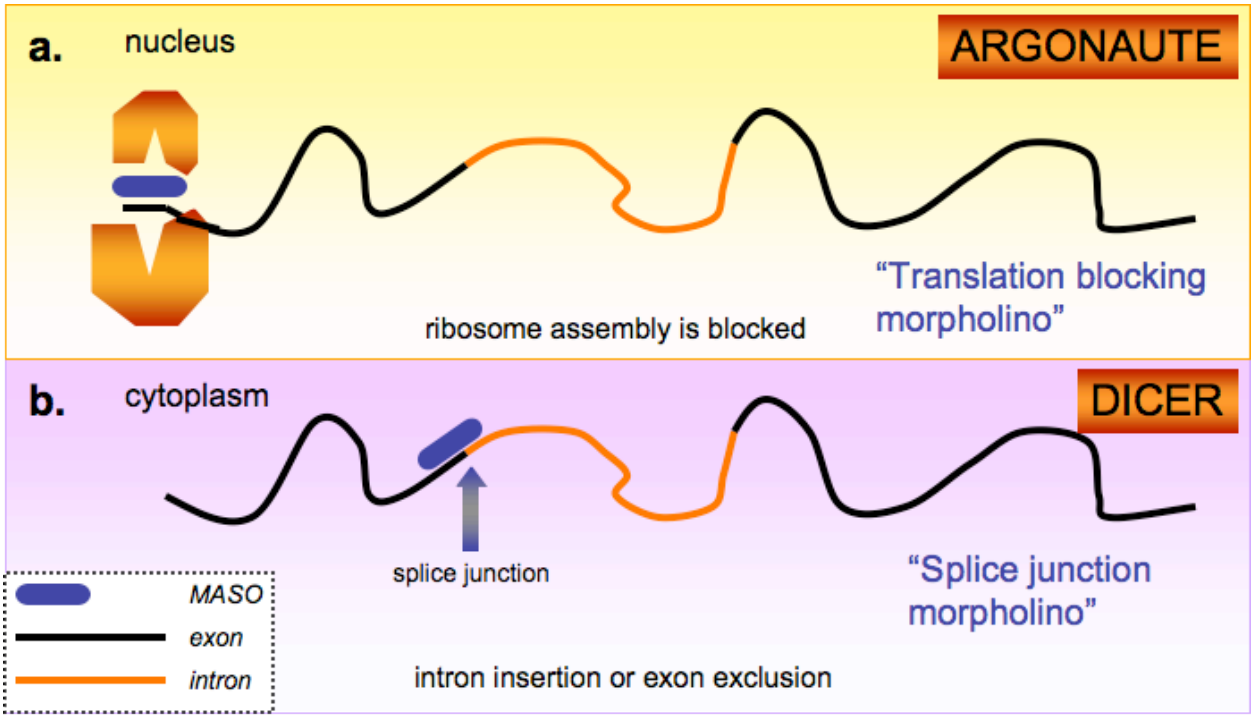


Figure 4.4: Morpholino AntiSense Oligonucleotides (MASOs) change gene expression using steric blocking: a. Translation blocking MASO – is complementary to a site between the 5' cap and start codon. It blocks the ribosome assembly, and thus, prevents translation of the protein. **b.** Splice junction MASO – is complementary to a splice junction site and causes intron insertion or exon exclusion, depending on its location.

Morpholino AntiSense Oligonucleotides (MASOs) are usually 25mers, designed with a morpholine backbone. MASOs bind to a complementary sequence on a RNA and operate via a mechanism called steric blocking. That is, they bind to their complementary RNA due to higher RNA-binding affinity, and block access of the ribosome initiation complex or splicing proteins. The blocking by a morpholino can occur at the 5'end of a gene or at a slice site, depending on the type of MASO.

There are two main types of MASOs: (*i*) Translation blocking MASOs: These MASOs block the progression of the translation initiation complex, preventing assembly of the ribosome, and thus, the translation of the protein (**Figure 4.4A**). (*ii*) Splice junction MASOs: These MASOs modify the targeted mRNA by either intron insertion or exon exclusion. Usually, targeting the splice junction of an internal exon causes the deletion of the targeted exon, whereas, targeting the splice junction of a flanking exon often leads to intron insertion (**Figure 4.4B**).

MASOs have been used successfully in the sea urchin model system (L M Angerer et al. 2001; Eric H Davidson et al. 2002). I designed the MASOs that specifically bind the splice junctions of an intron-exon and an exon-intron boundary of the sea urchin *Dicer* (**Figure 4.5** – top panel). Out of the two splice-junction MASOs, only the latter seemed to affect the mRNA (courtesy, Brenna McCauley). This MASO was used to knockdown the function of *Dicer* in early sea urchin embryos.

The *SpAgo1* protein-coding sequence was mapped to the sea urchin genome, and the candidate 5' UTR genomic sequence was obtained. Using flanking primers, we amplified the region, 5' of the start codon, and then cloned and sequenced it. We then designed a translation blocking MASO (**Figure 4.5** – middle panel) using GeneTools (http://www.gene-tools.com/) to knockdown the function of this gene in *S. purpuratus*. Due to lack of genomic sequence for

SpAgo2, we performed 5' RACE (random amplification of cDNA ends) with gene specific primers, and successfully amplified and sequenced the RACE product.

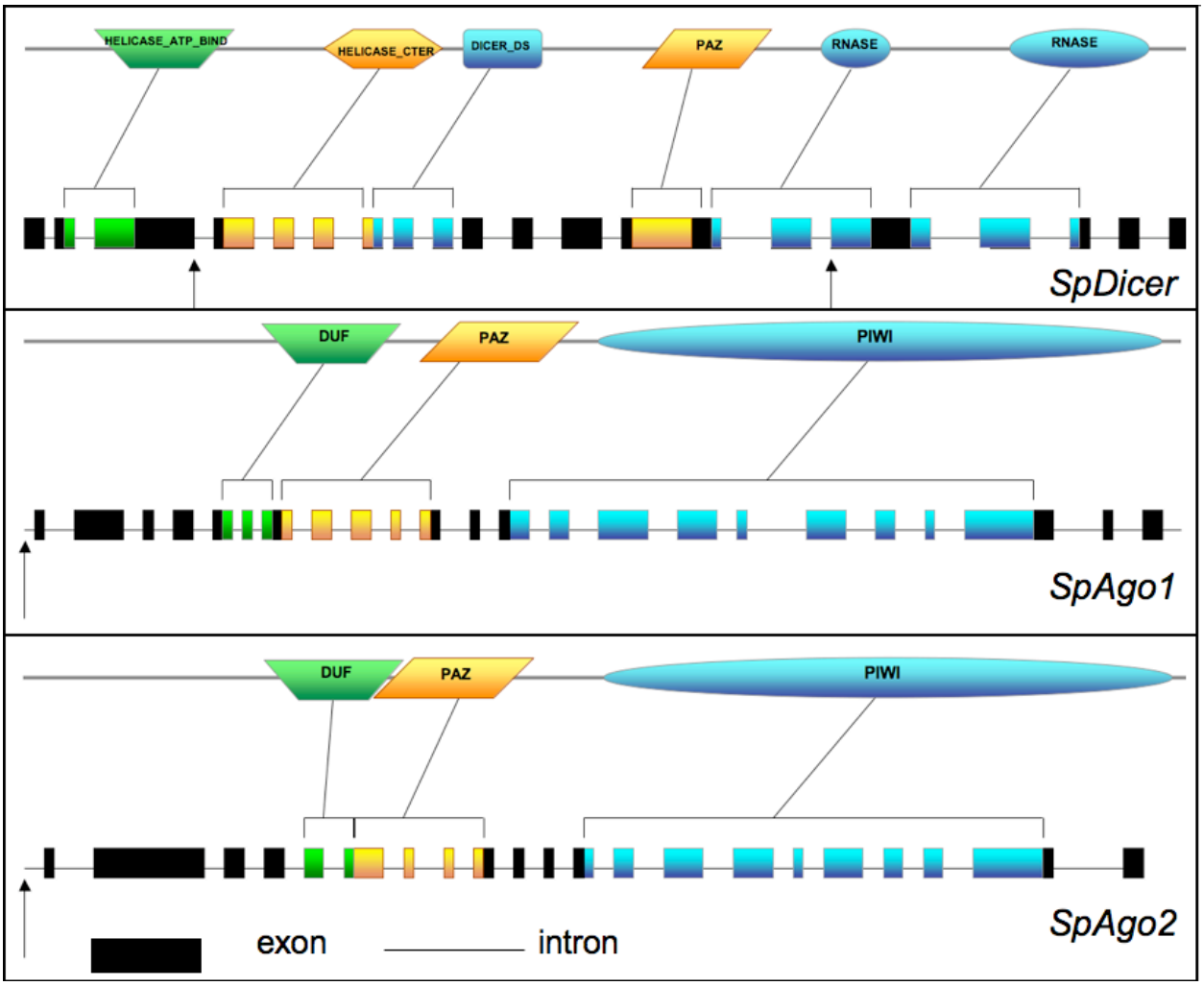


Figure 4.5: Cartoon representation of gene structures and locations of sea urchin MASOs: The protein domain mapped to the exon organization for *SpDicer, SpAgo1* and *SpAgo2*. The black arrows represent the location of the MASOs. The splice junction MASOs for *SpDicer* were designed upstream of a helicase domain at the N-terminus of the gene, and in the middle of the first RNase domain near the C-terminus of the gene. Translation blocking MASOs were designed for *SpAgo1* & *SpAgo2*, indicated by arrows at the 5'end of the genes. (The drawings are not made to scale.)

Since *SpAgo2* genomic sequence did not have sequence upstream of the start codon, we performed 5' RACE, and used the cloned sequence to design a translation blocking MASO for *SpAgo2* (**Figure 4.5** – bottom panel). For more details, see **Detailed Materials &** Methods.

We also made attempts to clone full-length *PmDicer* so that we could design a MASO targeting this transcript using multiple techniques, but they were unfortunately unsuccessful (data not shown). (*i*) **5' RACE:** We had part of the *PmDicer* gene that was sequenced from a library screen clone (courtesy: Kristen Yankura). This sequence is approximately 3kb downstream from the start codon. Numerous optimizations for the 5' RACE with different gene specific primers, and PCR conditions were unsuccessful in obtaining a specific band due to non-specific amplification. (*ii*) **Degenerate PCR:** We used degenerate PCR to amplify the N-terminus of the gene for a better chance at 5' RACE primer design. The sequenced bands did not map to Dicer, and thus, degenerate PCR picked up non-specific products. (*iii*) **Library Screen:** Species-specific probes were designed for *PmDicer* and *PmDGCR8* to perform a library screen on the *P. miniata* 3-day cDNA library filters. Multiple colonies with low signal were cloned and sequenced. However, the sequences did not map to *Dicer* or *DGCR8*.

4.2.2 Distinct morphological changes when *Dicer* and *Argonaute* functions are perturbed in sea urchin embryos.

I carried out titration experiments to determine the most effective concentration of *SpAgo1* & *SpAgo2* MASOs to be injected into the *S. purpuratus* embryos. The titration experiments for *SpDicer* was carried out by Brenna McCauley in the Hinman Lab.

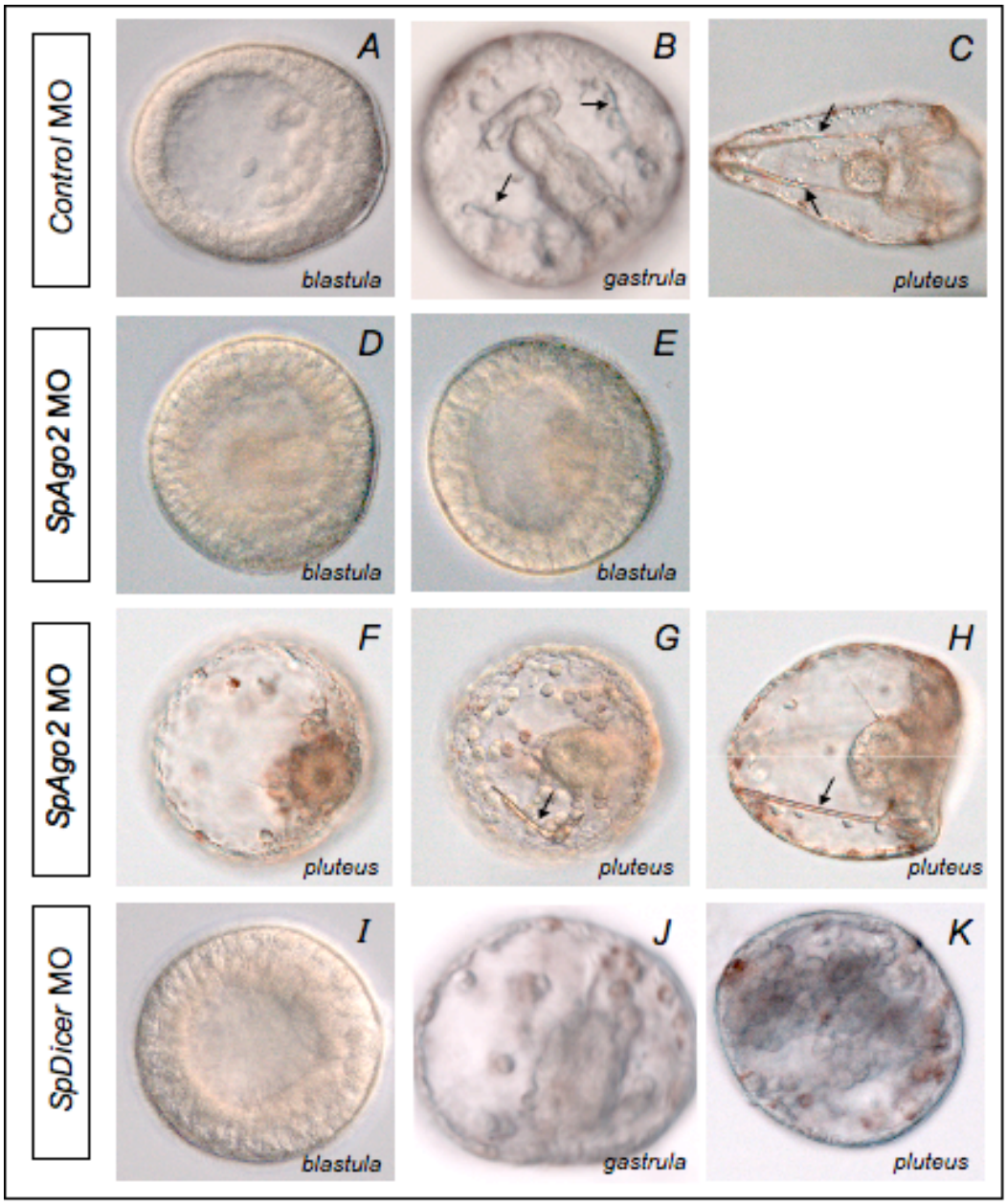


Figure 4.6: *SpDicer* and *SpAgo2* hinder normal development of the sea urchin embryo: (A-C) Blastula through pluteus stages of control MASO injected embryos. **(D-H)** *SpAgo2* knocked down sea urchin embryos. D & E are the blastula stage, with decreased volume of MASO from D to E. Some PMCs can be seen in E. F-G represent the pluteus stage with decreased volume of injected MASO from F through H. **(I-K)** Blastula through pluteus developmental stages of *SpDicer* knocked down embryos. (*Black arrows indicate the larval skeleton. All blastula stage embryos are aligned with the animal-vegetal axis along the horizontal axis of the image.*)

MASO for *SpAgo1* was injected for titration at 100μM, 200μM, 400μM, 600μM and 800μM concentrations. Even at the lowest concentration, the MASO consistently produced a toxic phenotype, where the blastocoel was filled with cells at the blastula stage of development or development was arrested at late cleavage stages (data not shown). Altering the concentration as well as the volume of the injected morpholino did not prevent the toxicity effect. Thus, either this particular morpholino non-specifically blocks some critical genes or *Ago1* knock down affects the regulation some developmentally necessary genes.

Ago2 knock down had more specific effects on the morphology of the embryo. Knocking down this protein in the sea urchin embryo leads to loss of the larval skeleton, one of the evolutionary novelties not seen in sea stars (**Figure 4.6D-H**). The larval sea urchin has at least two mesodermal cell types that are absent in larval sea star – the skeletogenic mesenchyme and pigment cells. They are considered novel phenotypes because these cell types are also absent from the larvae of other evolutionary groups of echninoderms. Depending on the volume of the titrated concentration injected, the embryos show a range of phenotypes from no larval skeleton in increased volumes and most times, tiny spicules to an abnormal larval skeleton at the lowest volume (**Figure 4.6F-H**). The blastula stage embryos show a similar range of morphology from ingression of none to few PMCs with increased to decreased volumes of morpholino (**Figure 4.6D & E**). Thus, knockdown of SpAgo2 protein blocks PMC ingression. As described in **Chapter 1.0** , the PMCs ingress after ~24hpf from the vegetal plate into the blastocoel, and are responsible for formation of an extracellular matrix of proteins in which precipitation of calcium carbonate occurs and spicules are formed (R. E. Peterson & David R McClay 2003; Wu et al. 2007; S. Benson 1987; S. C. Benson et al. 1986). The PMCs also interact with the ectoderm and this affects the size of the skeletal rods. We expect that since PMC ingression is blocked, and not

delayed, downstream of the *SpAgo2* pathway, that miRNAs interact with the PMC pathway or its progenitors upstream of PMC ingression. A cascade of signals originating in the PMCs cause cell fate specification in the sea urchin embryo. It will be interesting to study the downstream effects on other cell types of the miRNA interactions within these cells.

The other phenotypic effect observed with *SpAgo2* knockdown is abnormal gut development (**Figure 4.6F-H**). Since it is known that signals initiating in the micromeres induce the veg2 cells to become the archenteron (**Section 1.2.2**), it might be possible that miRNA interactions that might be upstream of the gut formation, might occur in the PMCs or after induction of veg2 cells. It is important to note that that abnormal gut formation can be a toxic effect of MASOs, thus, further validation is important.

The defects seen in *SpDicer* knocked down embryos were similar (**Figure 4.6I-K**), but the embryos are look sicker and the morphological effects are not as specific as embryos in which SpAgo2 function is perturbed. Thus, we selected the MASO for *SpAgo2* for further experiments.

4.2.3 Markers for distinct embryonic domains are used to compare effects of gene knockdown on various territories

We used whole mount in situ hybridization of gene transcripts to isolate downstream targets in order to characterize the role of the miRNA pathway in the development of *S. purpuratus*. We selected four genes that are expressed in four distinct embryonic territories and set out to answer the question: is this gene or the pathway it represents downstream of *SpAgo2*? We used a MASO directed against *SpAgo2* that reduces the expression of the SpAgo2 protein.

I collected *SpAgo2* MASO injected embryos at 30 hours post fertilization (*hpf*) time-point and fixed some of them using the paraformaldehyde fixation protocol, standard to the lab (V. F. Hinman, A. T. Nguyen, R Andrew Cameron, et al. 2003; V. F. Hinman, A. T. Nguyen & Eric H Davidson 2003). These fixed embryos were used to perform WMISH using existing probes.

We selected *Sm50* as a marker for PMCs, since PMC ingression is blocked in *SpAgo2* knocked down embryos (**Figure 4.6D-E**), and *Endo16* as the endoderm-specific marker as gut formation is affected (**Figure 4.6-F-H**). We also selected *Pks* and *Rsh* as markers of pigment cell and ectoderm differentiation.

4.2.3.1 PMC-specific marker *SpSm50* expression disappears in *Ago2* knocked down embryos.

Sm50 is a PMC specific spicule matrix protein expressed in the PMCs (S. C. Benson et al. 1986) (**Figure 4.7 A)**. As shown in **Figure 4.6D-E,** embryos with inhibited *Ago2* function show little or no ingression of PMCs into the blastocoel. Since *Sm50* gene is expressed exclusively in the PMCs at the mesenchyme blastula stage of development, expression of *Sm50* is also repressed in embryos with perturbed *Ago2* function (**Figure 4.7 E)**. The embryo in **Figure 4.7 E** did not show any PMC ingression, and has no *Sm50* expression.

4.2.3.2 Ectoderm territory of *SpRsh* expands in *Ago2* knocked down embryos.

The ectoderm territory of the embryo consists of the oral, aboral ectoderm, ciliary band and apical tuft (**Figure 1.2B-D**). We did not see any obvious morphological differences in the ectoderm of the *SpAgo2* knocked down embryos (**Figure 4.6D-H**). We selected *SpRsh* as the differentiation gene marker for a subset of this territory.

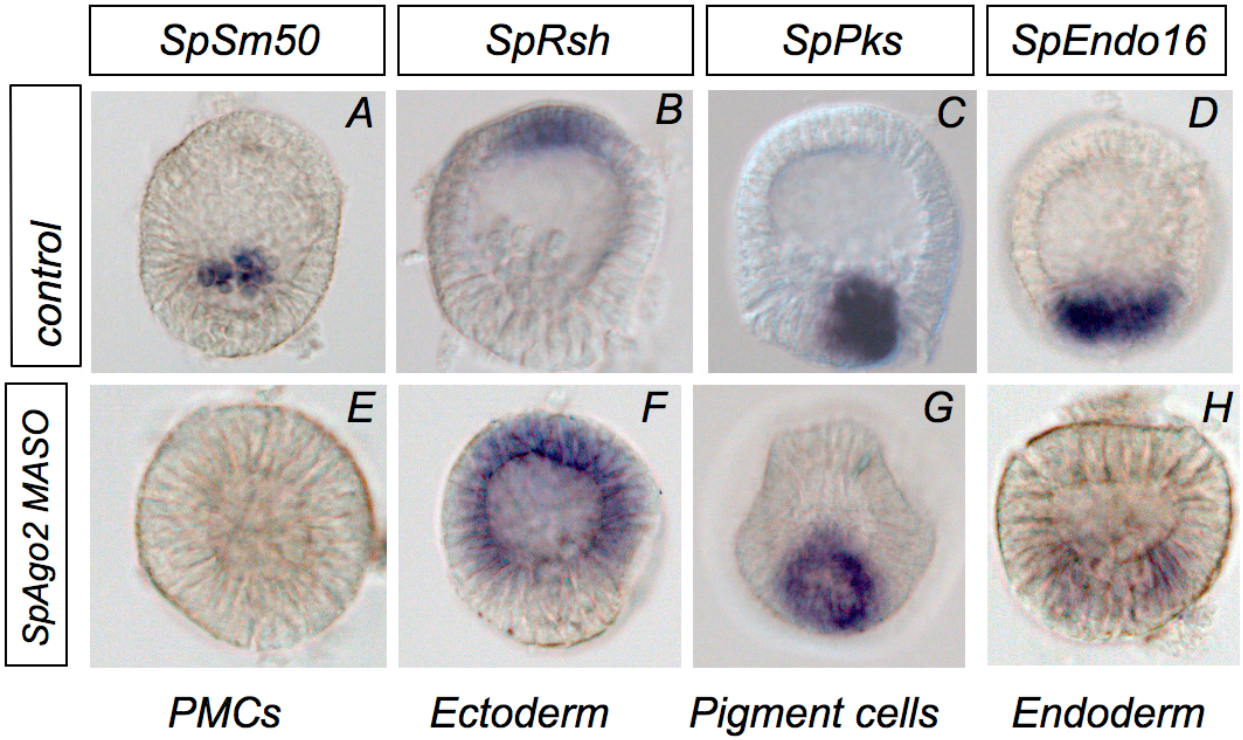


Figure 4.7: Expression of differentiation gene markers in control and *SpAgo2* knocked down embryos: DIG-labeled probes for *SpSm50* (A,E,I), *SpRsh* (B,F,J), *SpPks* (C,G,K) and *SpEndo16* (D,H,L). **(A-D)** Normal expression of the differentiation markers in control embryos. These territories are represented in cartoon form in **(I-L)**. Sm50 is a PMC-specific matrix protein found in spicules, and is found in the PMCs. *Rsh* is a cilia gene expressed in the apical ectoderm of the developing sea urchin embryo. *Pks* is a pigment cell marker expressed in veg2 cells in one half of the embryo. *Endo16* is an endoderm-specific marker expressed in the vegetal plate. **(E-H)** Expression of the respective differentiation markers in *SpAgo2* knocked down embryos. All embryos were fixed at 30hpf.

SpRsh is expressed in the regions that form cilia in the sea urchin embryos (Dunn et al. 2007). This gene is expressed in the apical region of the blastula stage embryos after hatching (**Figure 4.7 B)**. In embryos with perturbed *SpAgo2* function, transcripts of *SpRsh* are expressed ectopically throughout the oral and aboral ectoderm (**Figure 4.7F**).

4.2.3.3 The *SpPks* gene does not show any drastic changes in expression

At the mesenchyme blastula stage, polyketide synthase (*Pks*) is expressed in a subset of the SMC precursors that are specified into pigment cells at gastrulation (Calestani et al. 2003). The gene is expressed in a hollow ring-like pattern in the vegetal plate of embryos with miRNA function inhibition. However, the difference in expression was not as drastic as those described above and might be due to delays caused during development, and not the knock down. **APPENDIX I** show in situ hybridization of *Pks* in sea urchin embryos 3 days after fertilization.

4.2.3.4 Reduced expression of the endomesodermal marker, *SpEndo16* after *SpAgo2* knockdown.

Micromeres induce the expression of the endoderm-specific marker *SpEndo16* (Romano & Wray 2003) (**Figure 4.7 D&L)**. We saw that in *SpAgo2* knockdown, with the reduction of PMCs in the mesenchymal blastula, we also see repression of *SpEndo16* expression (**Figure 4.7H).** Recent work by Song et a. (Song et al. 2011) also showed a reduction in *Endo16* signal when *Dicer* function was perturbed in sea urchin embryos.

4.3 Future direction: High throughput network reconstruction

It will be very useful in the future to use a high-throughput experimental set-up to immunoprecipitate a component of the RISC assembly, in order to recover miRNA-mRNA interactions during various time-points of development. The function of individual miRNAs can then be blocked, and assays for transcript abundance changes with particular focus on genes with known developmental functions during embryogenesis can be carried out. miRNA target prediction methods were combined using a supervised learning method along. Green Fluorescent Protein (GFP) reporter assays can be used to deduce direct miRNA-target interactions. Some preliminary work was done for this. A high-throughput approach will provide a thorough understanding of the global roles of miRNAs during embryogenesis.

It has been postulated that miRNAs may function to fine-tune levels of their target proteins to some biologically relevant levels (tuning targets) (D. P. Bartel & C.-Z. Chen 2004), and many miRNA mutants do not individually produce obvious phenotypes (E. A. Miska et al. 2007). Other studies have shown that some miRNAs produce drastic phenotypic changes (R. Lee 1993; B J Reinhart et al. 2000). Thus, it is possible that most miRNAs in sea urchin embryos fine-tune expression and a few (or none) are involved in the developmental effects described above.

4.3.1 HITS-CLIP

We would like to immunoprecipitate the cytoplasmic protein:RNA complex (RISC) involved in miRNA function, and then sequence the miRNA and mRNA populations involved. The target protein should be an integral component of the RISC assembly and, thus, we should be able to recover the miRNAs and mRNAs interacting in the cells at a certain developmental stage. After

obtaining the pool of miRNAs and mRNAs interacting in the embryo at a specific developmental stage, half of this RNA pool can be hybridized to a custom tiling array covering the genome [29] to detect parts of the transcriptome that potentially interact with miRNAs. The remaining RNA pool can be sequenced using Illumina or 454 sequencing to obtain sequence data for the miRNAs and mRNAs. Using the results from **Chapter 3.0** , we can identify the miRNAs in the sequences and treat the remaining high quality reads as potential mRNA targets.

The first step for this set-up is to obtain the antibody to immunoprecipitate a RISC assembly protein. I have performed preliminary work towards this goal as shown in **4.3.1.2**.

4.3.1.1 The HITS-CLIP technology

High-throughput sequencing of RNA isolated by crosslinking followed by immunoprecipitation (**HITS-CLIP**) is an in vivo approach to perform genome-wide mapping protein–RNA binding sites (Licatalosi et al. 2008). This method has been used to study populations of interacting miRNAs and mRNAs by covalently cross-linking native *Argonaute* protein-RNA complexes, followed by immunoprecipitation of the complex (Chi et al. 2009). The RNA populations are isolated from the complex, and sequenced using any RNA-seq.

4.3.1.2 A human antibody against SpAgo was tested in sea urchin samples

RISC contains one of multiple Ago proteins in animals (sea urchin has 4 Ago-like proteins (Song & Wessel 2007), out of which two seem to be involved in miRNA function). I used the sequence of these two AGO proteins to look for existing antibodies that might be able to target conserved parts of the protein. Human *Ago* shares 69% identity with the sea urchin *Ago*.

Specifically the human antibody was raised against the PIWI domain of the protein (**Figure 4.9**). This domain has 73% identity with *SpAgo1* and 79% identity with *SpAgo2*.

The protein structures of SpAGO1 and SpAGO2 were predicted using SWISS-MODEL (Arnold et al. 2005; Kiefer et al. 2009; Peitsch 1995). See **Detailed Materials &** Methods for more details. The pdb files were visualized using VMD (http://www.ks.uiuc.edu/Research/vmd/) (Humphrey et al. 1996). As shown in **Figure 4.8**, the structure of SpAGO1 is very similar to a human AGO2 crystal structure. The structure of *SpAgo2* could not be fully modelled using homology modelling. Partial structures of certain domains can be seen **Figure 4.8**. The PIWI domain is a well-conserved domain of the Argonaute family of proteins (**Figure 4.9**). The crystal structure of the proteins indicates that this domain can be easily accessible to an antibody.

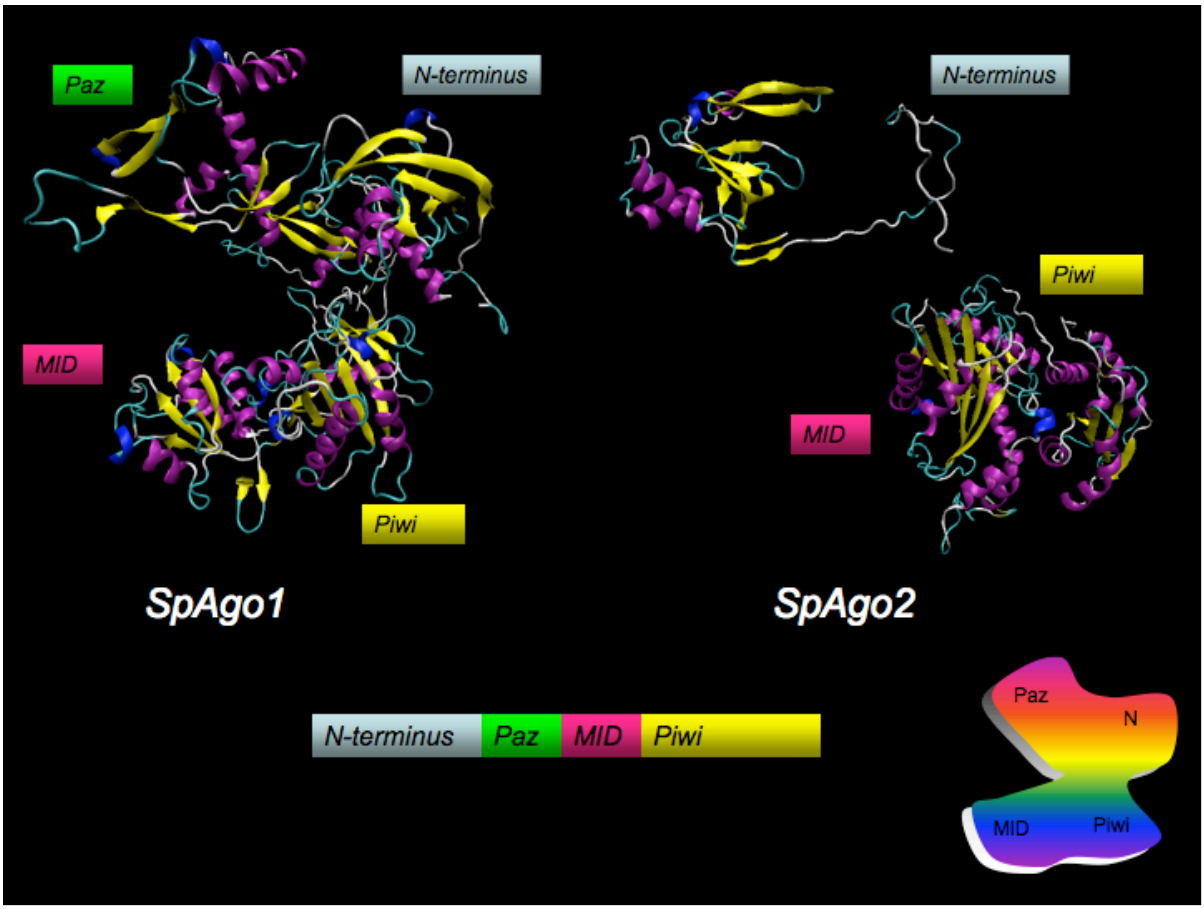


Figure 4.8: Protein structure of sea urchin Agos: The Argonaute protein has four main domains in humans (template on which these structures are predicted), which are labeled on the proteins, SpAGO1 & SpAGO2. The structures were predicted using homology modelling in SWISS-MODEL (See **Detailed Materials & Methods**), and visualized using VMD (http://www.ks.uiuc.edu/Research/vmd/) (Humphrey et al. 1996). The structure is colored by secondary structure of the protein. Only partial structures for parts of SpAGO2 were predicted. The cartoon at the bottom right corner represents the structural positioning of the four protein domains in an *Argonaute* protein.

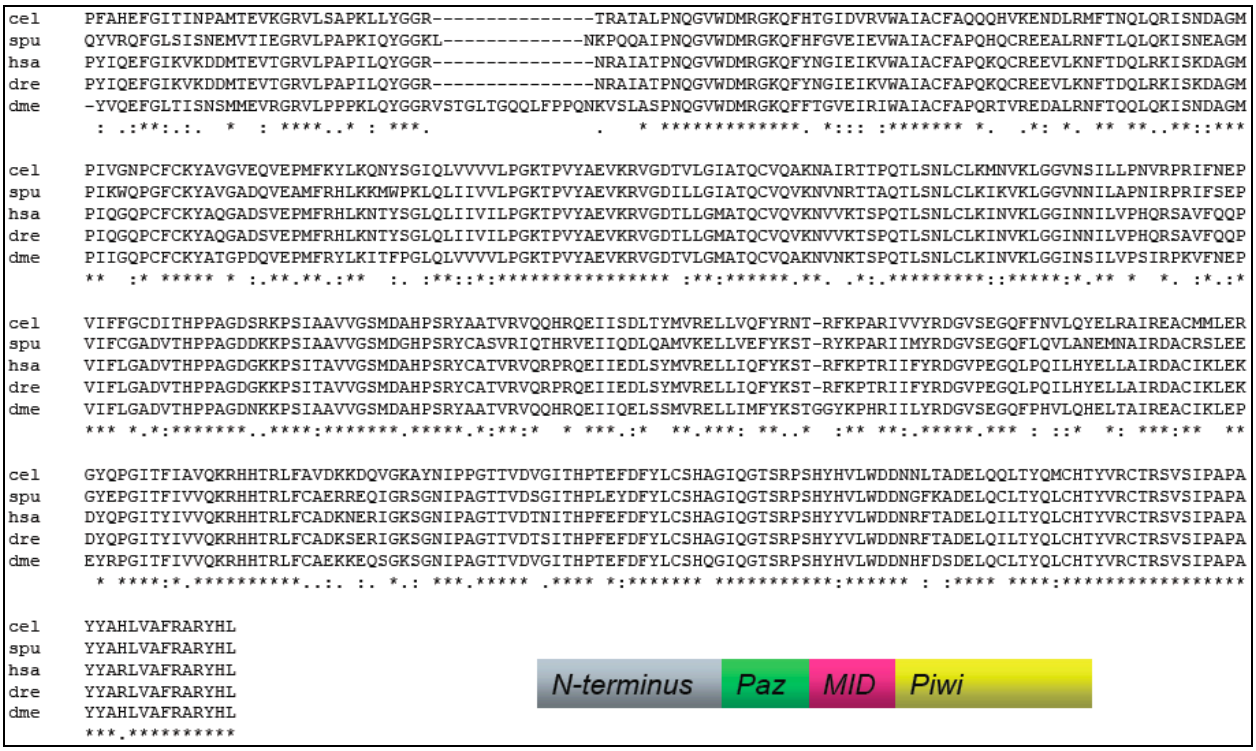


Figure 4.9: Conservation of the PIWI domain across multiple species: The PIWI domain is a very conserved domain across multiple Argonaute proteins. Here, species were selected as representative of various clades. The alignment was performed using ClustalW (Goujon et al. 2010; Larkin et al. 2007). (cel: *C. elegans;* spu: *S. purpuratus*; hsa: *H. sapiens*; dre: *D.rerio*; dme:*D. melanogaster*.)

(a) Western Blot shows cross-reactivity with sea urchin bands

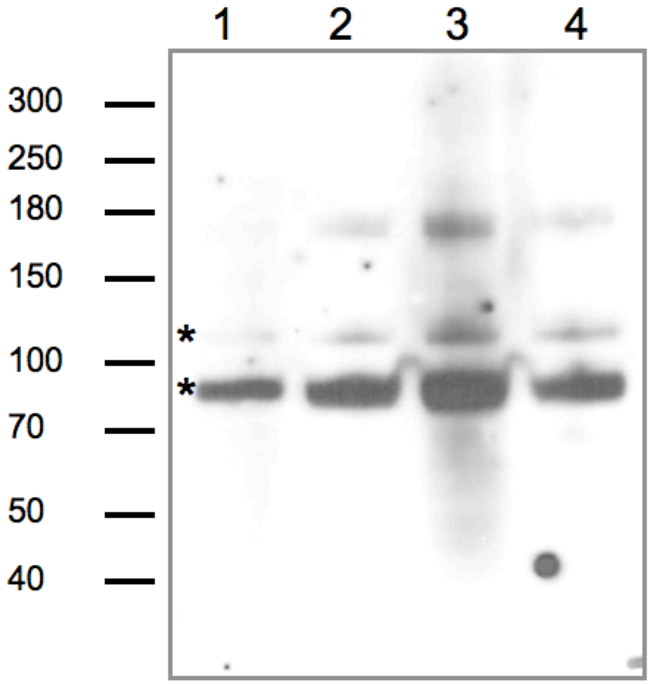


Figure 4.10: Western Blot with 2A8 antibody: The size markers on the left side of the figure indicate the mass in kDa. The lanes have protein extract prepared from sea urchin embryos 24hpf (See **Detailed Materials & Methods**) in increasing amounts, with lane 3 having the maximum protein. Lane 1 has extract from 100 embryos, 200 in lane 2, 500 embryos in lane 3 and 400 embryos in lane 4.* represent the two lanes close to the predicted size of the protein.

I tested two human Argonaute antibodies on *S. purpuratus* whole protein extracts using western blots: (i) a monoclonal antibody, 2A8 developed by the Mourelatos lab at the University of Pennsylvania (Nelson et al. 2007). This antibody was raised against the extremely conserved PIWI domain of the protein but the epitope against which it was raised is unknown. (ii) A polyclonal peptide antibody from Millipore, raised against the not-very-conserved N-terminus of the protein.

The 2A8 antibody were chosen due to the conserved nature of the PIWI domain across *Ago* proteins in multiple species (**Figure 4.9**), whereas the Millipore antibody was chosen due to conservation of the peptide in *S. purpuratus* SpAGOs, respectively. Western blots did not show any cross-reactivity with the Millipore peptide antibody (data not shown). However, 2A8 showed cross-reactivity with at least 3 bands on the protein gel, two of which are close in size to the predicted size of the SpAGO proteins (106 kDa) – at ~90kDa and ~110kDa respectively (indicated by * in **Figure 4.10**). See **Detailed Materials &** Methods for details on the Western blot protocol.

On the other hand, the same protocol when carried out for *P. miniata* (sea star) protein extracts showed a lot of background and did not show any of the enriched bands that were seen in *S. purpuratus* western blots (**APPENDIX I**).

(b) Immunoprecipitated bands were tested using mass spectrometry

In order to confirm whether the enriched band(s) on the Western blot correspond to the protein of interest, AGO, I performed immunoprecipitations without cross-linking to pull down the protein with different salt concentrations. The immunoprecipitated samples showed 3 bands of pulled down proteins. Interestingly, the most enriched band on the western blot (at ~90kDa) (**Figure 4.10)** did not immunoprecipitate, but the portion of the gel corresponding to this size was sent for mass spectrometry, along with the most enriched ~110kDa band (**Figure 4.11**) and the less enriched ~140kDa band to the University of Illinois (Dr. Peter Yau). Analysis of the LC/MS mass spectrometry results showed inconclusive results for the peptide sequences that were returned. The conclusion was insufficient starting amount of the protein. It has been suggested that we scale up the sample volume 4X and repeat the immunoprecipitation.

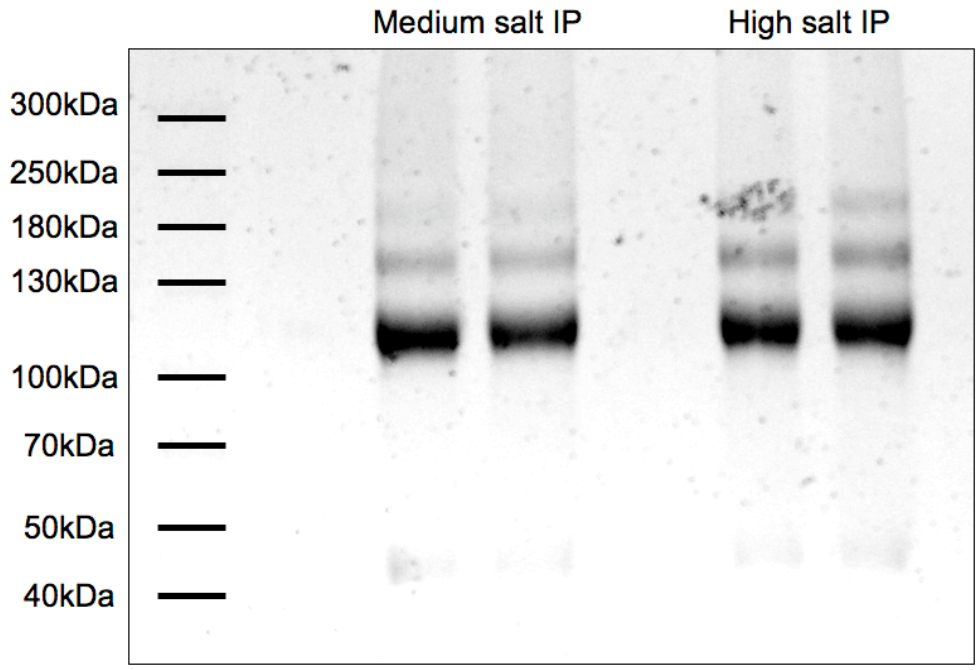


Figure 4.11: Silver stained gel with immunoprecipitated sea urchin extract: The left lanes show the samples immunoprecipitated with medium salt buffers, while right lanes show samples immunoprecipitated with high salt buffers.

Another experiment that can be performed to verify whether any of the bands seen in western blots correspond to AGO, is western blot on protein extracts from control and *Ago* knocked down embryos, using the MASO described in **Section 4.2.1**. If a band of the expected size diminishes or vanishes in the *Ago*-knocked down protein extracts, the band probably corresponds to SpAGO. However, if the knockdown is for a single *Ago* gene and the signal is not lighter, it could correspond to cross reactivity with the other *Ago* gene.

4.3.2 Predicted miRNA-mRNA gene networks

For the computational predictions, I used a combinatorial approach using existing miRNA target prediction algorithms (RNAhybrid (Rehmsmeier 2004), miRanda (Enright et al. 2003), TargetScan (Lewis et al. 2005)). These algorithms consider features like thermodynamics, RNA:RNA duplex energy and site accessibility. Previous studies have shown that there is very small overlap between the predictions of various methods (Rajewsky 2006). Recent years have seen an increase (although small) in the number of validated targets of miRNAs. We used available validated data of interacting RNAs from immunoprecipitation data in *C. elegans* (Zhang et al. 2007), and developed a supervised learning approach to combine the weak classifiers into a strong classifier using boosting. We used this approach to predict miRNA targets in sea urchin genes. Thus, the question underlying this approach was, given a UTR, and a set of miRNAs expressed in a sample, is the UTR under regulation by a subset of miRNAs?

One of the advantages here is that this method considers combinatorial effect of the miRNAs, is more likely to be important for correct target regulation (Krek et al. 2005). RNA abundance data from deep sequencing, if available, is also considered.

The top ten most abundant miRNAs discovered in **Chapter 3.0** were used for the computational predictions. Manoj Samanta provided us with 3'UTR data (personal communication) obtained from the tiling array (Samanta et al. 2006) along with mappings of the UTR IDs to their Genbank IDs. I have also hand-curated 3' and 5' UTRs of few developmentally important genes. An example of miRNA target predictions on the PMC network in early sea urchin development is shown in **Figure 4.12**.

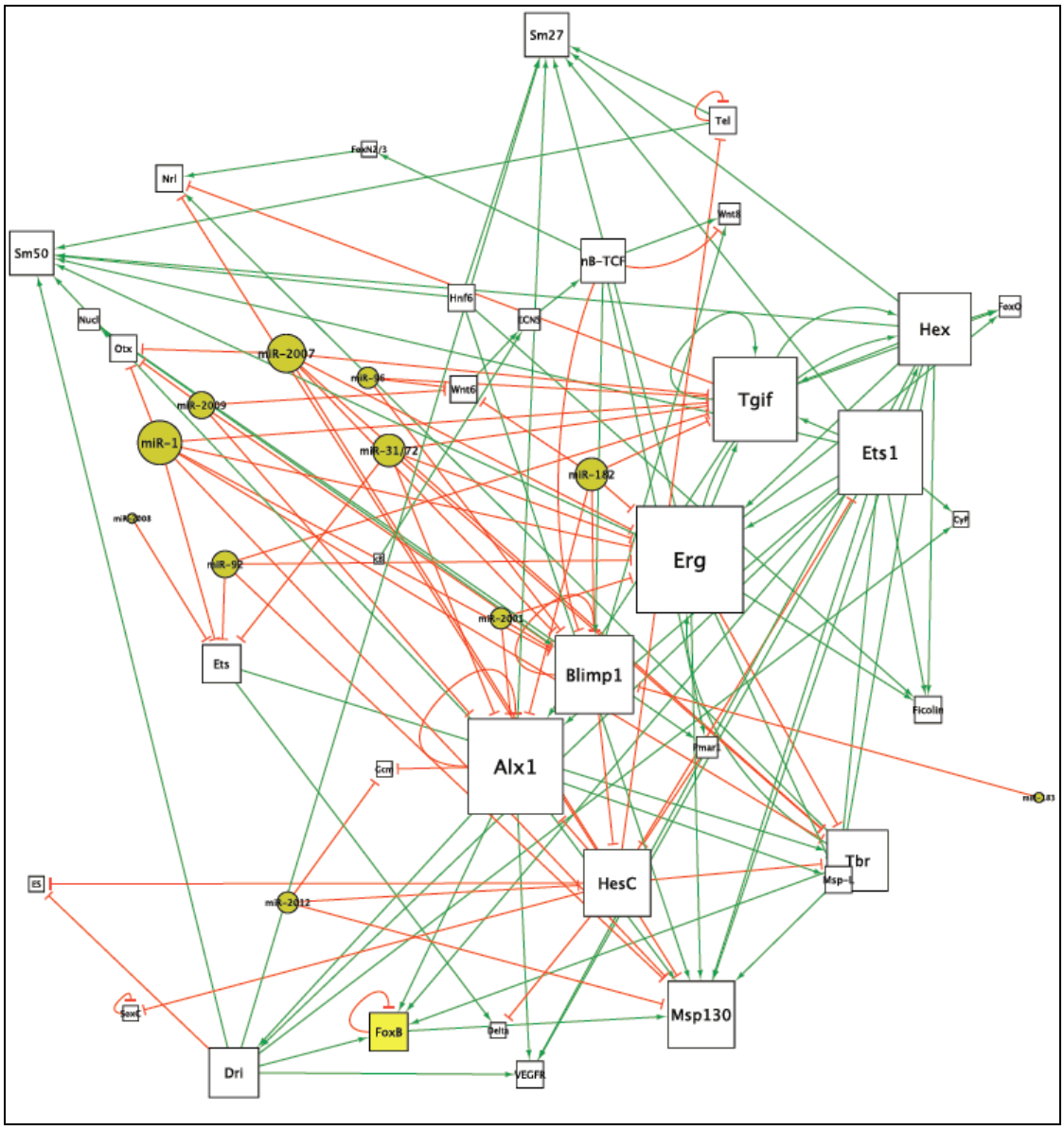


Figure 4.12: The early sea urchin PMC network overlaid with miRNA predictions: The top 10 most abundant miRNAs in the sea urchin libraries were used for target predictions. Visualization was done using Cytoscape (Cline et al. 2007; Shannon 2003). The size of the node is proportional to the sum of the indegree and outdegress. Thus, the more densely connected nodes are the largest in the network. *green line* – up-regulation; *red line* – down-regulation.

4.3.2.1 Validation methods: GFP reporter assays

The computational prediction of multiple binding sites of a miRNA(s) in the UTR of a candidate target, in a pathway that is downstream of *Argonaute,* will qualify the gene for experimental validation. MASOs can be used to knock down *Dicer* and *Argonaute* as described in **Section 4.2.1**. A GFP reporter assay can be used to deduce direct interaction by site mutation experiments. In this assay, constructs containing the 3' UTR of the target gene in question, and the GFP gene are injected into the embryos (with gene-specific and control MASOs). Expression of the GFP protein is then compared between embryos with knocked down miRNA pathway expression and control embryos. A difference in fluorescence of the GFP 3'UTR construct between control and *SpDicer* or *SpAgo2* knocked down embryos will indicate direct regulation of the 3' UTR by miRNAs.

For preliminary work, ten TFs were selected from the predictions shown above for further validation.

From the ten UTRs started with, at least one construct was successfully made - *SpSoxB1*. Details on primers are in **APPENDIX F.3**. In the network, SoxB1 has an unknown repressor that affects its protein in the vegetal plate. The GFP SoxB1 construct was injected in control and Dicer knocked down sea urchin embryos. If miRNAs regulate the 3' UTR of SoxB1, a difference should be seen in the fluorescence of the two sets of embryos, as described earlier. No visual difference was seen for the SoxB1 construct. Another method that can be used to detect differences in fluorescence (that are not visible under the microscope) is the use of a fluorometer to make quantitative measurements, and compare them between the two conditions. However, calculation of the fluorescence is not straightforward, as there are unknowns like GFP protein turnover, and the time point at which SoxB1 might come under miRNA regulation. The

difference of the fluorescence reading and the reading above background, normalized to RFP levels can be used. RFP is injected for the purpose of normalization against variability in injection volumes on the part of the injector.

4.4 Detailed Materials & Methods

4.4.1 Embryo cultures

Adult sea urchins, were maintained in seawater at 10 °C. Cultures was started in artificial seawater at 15 °C using standard methods as described in (Ettensohn et al. 2004).

4.4.2 Morpholino design

We cloned and sequenced the 5'RACE product for *SpAgo2*. The genomic region upstream to the first start codon of *SpAgo1* was sequenced using RT-PCR. The sequences obtained from multiple high quality colonies were submitted to GeneTools (http://www.gene-tools.com/), and translation-blocking morpholinos specific to *SpAgo1* and *SpAgo2* were designed (**Figure 4.5** - middle and bottom panels).

For *SpDicer*, two splice junction MASOs were designed by amplifying an intron-exon boundary (*SpDicer-1*), and an exon-intron boundary (*SpDicer-2*). *SpDicer-1* is found upstream of the helicase domain whereas *SpDicer-2* is found in the middle of the first RNAse domain (**Figure 4.5** – top panel). More sequence information about the individual morpholino sequences can be found in **APPENDIX F.2**.

4.4.3 Cloning and RACE

The primers used for cloning and RACE are indicated in **APPENDIX F.** The cDNA clones were sequenced at Yale sequencing Center. 5' RACE was performed using the GeneRacer Kit (Invitrogen) to get additional 5' sequence information.

4.4.4 Protein structure modelling

The automated-mode SWISS-MODEL structure modelling web server (http://swissmodel.expasy.org) was used to predict the *SpAgo* protein structures. SWISS-MODEL (Arnold et al. 2005; Kiefer et al. 2009; Peitsch 1995) is a protein structure homology-modelling server in the automated mode since we expected the target–template similarity to be sufficiently high to allow for fully automated modelling, since *Argonaute* proteins are quite conserved across species. In the automated mode, SWISS-MODEL identifies suitable templates based on a Blast *E*-value limit of structural templates, aligns the target sequence with the template structures, and builds the model based on the alignment. Some quality evaluations are also performed on the model.

4.4.5 Western Blots & Immunoprecipitation

Western blots were performed using protocols from the Chakrabarti lab using the ECL chemiluminescence kit. For the 2A8, we used 1:100–1:200 dilution. For immunoprecipitation using 2A8, 10μL of 2A8 was used with 1 mL of protein-G agarose (Invitrogen) and lysate from cells or tissues. Cells or tissues were lysed in lysis buffer using 300mM NaCl and 500mM NaCl

for medium and high salt buffers respectively. Samples were run on 4%-12% TrisCl SDS page gels.

4.4.6 Embryo injections and fixation

Eggs were de-jellied by incubating in acid sea water (pH 4.02) for ~1 min pipetted in rows on a plastic culture dish coated with 1% protamine sulfate, containing 15mg/ml PABA sea water. Morpholinos were then injected into them immediately after fertilization. For *SpDicer-2* as well as *SpAgo2*, we used 0.6mM of morpholino concentration. The control morpholino was injected at 0.4mM concentration.

We used a 4% paraformaldehyde in 100 mM MOPS buffer fixative. The embryos, at the desired developmental time point, were fixed for an hour at room temperature and then overnight at 4°C and stored in 70% ethanol until use.

4.4.7 Whole mount in situ hybridization (WMISH)

The standard lab protocols were used for WMISH (V. F. Hinman, A. T. Nguyen, R Andrew Cameron, et al. 2003; V. F. Hinman, A. T. Nguyen & Eric H Davidson 2003). The embryos were rehydrated by three washes with 1X MOPS buffer. The embryos were pre-hybridized for 30min-2h in the hybridization buffer containing 70% formamide at 58°C. Hybridization with hybridization buffer containing the DIG-labeled probe was carried out for 5 days. Post-hybridization washes were carried out with hybridization buffer twice at 58°C and with MAB buffer three times at room temperature. The embryos were incubated in a block of MAB buffer and 2% Roche block for 30 min at room temperature, followed by incubation in the MOPS

Roche block and the Anti-DIG AP antibody at room temperature. Excess antibody was washed off using MAB washes. Embryos were washes with the AP buffer followed by color staining reaction in AP buffer with NBT/BCIP.

4.5 Conclusions

This is the first known study investigating the role of the miRNA pathway in cell specification pathways in early echinoderm development.

Knocking down key components of the sea urchin miRNA biogenesis pathway showed hindered development of the embryos. The most striking phenotypic effects were seen on the larval skeleton and gut formation. Whole mount in situ hybridization with differentiation gene markers showed repression of the endoderm-specific and PMC-specific markers, while an expansion of the apical ectoderm marker was seen. We expect that the precursors of these cell types have miRNA targets upstream in the pathway, but downstream of the miRNA pathway.

In a paper published one month ago (Song et al. 2011), the authors performed loss of function assays using *Dicer* and *Drosha* in sea urchin embryos, and saw effects like gastrulation failure and embryonic lethality. They also saw reduction in *Endo16* expression with *Dicer* function inhibition; similar to what we see when *Ago2* function is inhibited. They do not comment on effects on the larval skeleton.

It will be extremely interesting to see whether majority of developmentally regulated miRNAs in the sea urchin have few targets with major cell fate consequences, or they have many targets with subtle effects that lead to strong developmental effect. It will also be interesting to

see if miRNAs are present as critical regulators of cell differentiation or developmental timing or are mere fine-tuning instruments for protein levels.

APPENDIX A

MODIFIED ALGORITHMS FOR HIERARCHICAL HIDDEN MARKOV MODELS

Finite alphabet: $\sum$

Observed string: $O = o_1 o_2 \ldots o_N$ such that $o_i \in \sum$

Highest level of hierarchy (root): 1

Lowest level of hierarchy (leaves): D

Depth of hierarchy: $d \in \{1,\ldots,D\}$

i^{th} state at hierarchical level d: q_i^d

Number of sub-states of q_i^d: $\left|q_i^d\right|$

Parameters of HHMM:

$$\lambda = \left\{\lambda^{q^d}\right\}_{d\in\{1,\ldots,D\}} = \begin{Bmatrix} \left\{A\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}, \\ \left\{\prod\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}, \\ \left\{E\left(q^D\right)\right\} \end{Bmatrix}$$

1. **Sub-state Transition Matrix:**

$\left\{A\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}$ such that $A\left(q^d\right)=\left(a_{jk}^{q^d}\right)=P\left(q_k^{d+1}\middle|q_j^{d+1}\right)$

$a_{jk}^{q^d}$ is the probability that the j^{th} sub-state of q^d will transition to its k^{th} sub-state.

2. **Initial Sub-state distribution:**

$\left\{\prod\left(q^d\right)\right\}_{d\in\{1,\ldots,D\}}$ such that $\prod\left(q^d\right)=\left\{\pi\left(q_j^{d+1}\middle|q^d\right)\right\}=\left\{P\left(q_j^{d+1}\middle|q^d\right)\right\}$

$\pi\left(q_j^{d+1}\middle|q^d\right)$ is the probability that q^d will make a vertical transition to its j^{th} sub-state at level d+1.

3. **Output probability distribution:**

$\left\{E\left(q^D\right)\right\}$ such that $E\left(q^D,q^{D-1}\right)=\left\{e\left(\sigma_l\middle|q^D,q^{D-1}\right)\right\}=\left\{P\left(\sigma_l\middle|q^D,q^{D-1}\right)\right\}$

$e\left(\sigma_l\middle|q^D,q^{D-1}\right)$ is the probability that production state q^D will emit symbol $\sigma_l\in\sum$.

Modified Baum Welch algorithm:

Calculate the following probabilities:

1. **Forward Probabilities**

$\alpha\left(t,t+k,q_i^{d+1},q^d\right)=P(o_t\cdots o_{t+k},q_i^{d+1}$ finished at $o_{t+k}\middle|q^d$ started at $o_t)$

Initialization:

Production states:

$\alpha\left(t,t,q_i^D,q^{D-1}\right)=\pi\left(q_i^D\middle|q^{D-1}\right)e\left(o_t\middle|q_i^D,q^{D-1}\right)$

Internal States:

$$\alpha\left(t,t,q_i^d,q^{d-1}\right)=\pi\left(q_i^d\middle|q^{d-1}\right)\left[\sum_{j=1}^{\left|q_i^d\right|}\alpha\left(t,t,q_j^{d+1},q_i^d\right)\cdot a_{j\,end}^{q_i^d}\right]$$

Iteration:

Production states:

$$\alpha\left(t,t+k,q_i^D,q^{D-1}\right)=\left[\sum_{j=1}^{\left|q^{D-1}\right|}\alpha\left(t,t+k-1,q_j^D,q^{D-1}\right)\right]e\left(o_{t+k}\middle|q_i^D,q^{D-1}\right)$$

Internal States:

$$\alpha\left(t,t+k,q_i^d,q^{d-1}\right)=\sum_{l=0}^{k-1}\left[\sum_{j=1}^{\left|q^{d-1}\right|}\alpha\left(t,t+l,q_j^d,q^{d-1}\right)\cdot a_{ji}^{q^{d-1}}\right]\cdot \left[\sum_{s=1}^{\left|q_i^d\right|}\alpha\left(t+l+1,t+k,q_s^{d+1},q_i^d\right)\cdot a_{s\,end}^{q_i^d}\right] + \pi\left(q_i^d\middle|q^{d-1}\right)\left[\sum_{j=1}^{\left|q_i^d\right|}\alpha\left(t,t+k,q_j^{d+1},q_i^d\right)\cdot a_{j\,end}^{q_i^d}\right]$$

2. **Backward Probabilities**

$$\beta\left(t,t+k,q_i^d,q^{d-1}\right)=P\left(o_t\cdots o_{t+k}\middle|q_i^d \text{ started at } o_t,\ q^{d-1} \text{ finished at } o_{t+k}\right)$$

Initialization:

Production states:

$$\beta\left(t,t,q_i^D,q^{D-1}\right)=e\left(o_t\middle|q_i^D,q^{D-1}\right)\cdot a_{i\,end}^{q^{D-1}}$$

Internal States:

$$\beta\left(t,t,q_i^d,q^{d-1}\right)=\left[\sum_{j=1}^{\left|q_i^d\right|}\pi\left(q_j^{d+1}\middle|q_i^d\right)\cdot\beta\left(t,t,q_j^{d+1},q_i^d\right)\right]a_{i\,end}^{q^{d-1}}$$

Iteration:

Production states:

$$\beta\left(t,t+k,q_i^D,q^{D-1}\right)=e\left(o_t\middle|q_i^D,q^{D-1}\right)\left[\sum_{j\neq end}^{\left|q^{D-1}\right|}a_{ij}^{q^{D-1}}\cdot\beta\left(t+1,t+k,q_j^D,q^{D-1}\right)\right]$$

Internal States:

$$\beta\left(t,t+k,q_i^d,q^{d-1}\right)=\sum_{l=0}^{k-1}\left[\sum_{j=1}^{\left|q_i^d\right|}\pi\left(q_j^{d+1}\middle|q_i^d\right)\beta\left(t,t+l,q_j^{d+1},q_i^d\right)\right]\cdot \left[\sum_{s=1}^{\left|q^{d-1}\right|}a_{ij}^{q^{d-1}}\cdot\beta\left(t+l+1,t+k,q_s^d,q^{d-1}\right)\right] + \left[\sum_{j=1}^{\left|q_i^d\right|}\pi\left(q_j^{d+1}\middle|q_i^d\right)\cdot\beta\left(t,t+k,q_j^{d+1},q_i^d\right)\cdot a_{i\,end}^{q^{d-1}}\right]$$

3. **Auxiliary variables:**

A. $\eta_{in}\left(t,q_i^d,q^{d-1}\right)=P\left(o_1\cdots o_{t-1},q_i^d \text{ started at } o_t\middle|\lambda\right)$

Initialization:

$$\eta_{in}\left(1,q_i^2,q^1\right)=\pi\left(q_i^2\middle|q^1\right)$$

$$\eta_{in}\left(1,q_i^d,q_j^{d-1}\right)=\eta_{in}\left(1,q_j^{d-1},q^{d-2}\right)\cdot\pi\left(q_i^d\middle|q_j^{d-1}\right)$$

Iteration:

For $1<t$

$$\eta_{in}\left(t,q_i^2,q^1\right)=\sum_{j=1}^{\left|q^1\right|}\alpha\left(1,t-1,q_j^2,q^1\right)a_{ji}^{q^1}$$

$$\eta_{in}\left(t,q_i^d,q_j^{d-1}\right)=\sum_{s=1}^{t-1}\eta_{in}\left(s,q_j^{d-1},q^{d-2}\right)\left[\sum_{l=1}^{\left|q_j^{d-1}\right|}\alpha\left(s,t-1,q_l^d,q_j^{d-1}\right)a_{li}^{q_j^{d-1}}\right] + \eta_{in}\left(t,q_j^{d-1},q^{d-2}\right)\cdot\pi\left(q_i^d\middle|q_j^{d-1}\right)$$

B. $\eta_{out}\left(t,q_i^d,q^{d-1}\right)=P\left(q_i^d \text{ finished at } o_t,o_{1t+1}\cdots o_N\middle|\lambda\right)$

Initialization:

For $t<N$

$$\eta_{out}\left(t,q_i^2,q^1\right)=\sum_{j=1}^{\left|q^1\right|}a_{ij}^{q^1}\cdot\beta\left(t+1,N,q_j^2,q^1\right)$$

Iteration:

For $t<N$

$$\eta_{in}\left(t,q_i^d,q_j^{d-1}\right)=\sum_{k=t+1}^{N}\left[\sum_{l=1}^{\left|q_j^{d-1}\right|}a_{il}^{q_j^{d-1}}\beta\left(t+1,N,q_l^d,q_j^{d-1}\right)\right]\eta_{out}\left(k,q_j^{d-1},q^{d-2}\right)$$
$$+a_{i\,end}^{q_j^{d-1}}\cdot\eta_{out}\left(t,q_j^{d-1},q^{d-2}\right)$$

$$\eta_{out}\left(N,q_i^d,q_j^{d-1}\right)=a_{j\,end}^{q_j^{d-1}}\cdot\eta_{out}\left(N,q_j^{d-1},q^{d-2}\right)$$

4. **Horizontal Transition Probabilities**

$$\xi\left(t,q_i^{d+1},q_j^{d+1},q^d\right)=P(o_1\cdots o_t,q_i^{d+1}\rightarrow q_j^{d+1},o_{t+1}\cdots o_N|\lambda)$$

Estimation:

\

$$\xi\left(t,q_i^2,q_j^2,q^1\right)=\frac{\alpha\left(1,t,q_i^2,q^1\right)\cdot a_{ij}^{q^1}\cdot\beta\left(t+1,N,q_j^2,q^1\right)}{P\left(O|\lambda\right)}$$

$$\xi\left(N,q_i^2,q_j^2,q^1\right)=\frac{\alpha\left(1,N,q_i^2,q^1\right)\cdot a_{ij}^{q^1}}{P\left(O|\lambda\right)}$$

For $t<N$

$$\xi\left(t,q_i^d,q_j^d,q_l^{d-1}\right)=\frac{1}{P\left(O|\lambda\right)}\left[\sum_{s=1}^{t}\eta_{in}\left(s,q_l^{d-1},q^{d-2}\right)\cdot\alpha\left(s,t,q_i^d,q_l^{d-1}\right)\right]a_{ij}^{q_l^{d-1}}$$
$$+\left[\sum_{k=t+1}^{N}\beta\left(t+1,k|q_j^d,q_l^{d-1}\right)\cdot\eta_{out}\left(k,q_l^{d-1},q^{d-2}\right)\right]$$

$$\xi\left(t,q_i^d,q_{end}^d,q_j^{d-1}\right)=\frac{1}{P\left(O|\lambda\right)}\left[\sum_{s=1}^{t}\eta_{in}\left(s,q_j^{d-1},q^{d-2}\right)\cdot\alpha\left(s,t,q_i^d,q_j^{d-1}\right)\right]$$
$$a_{i\,end}^{q_j^{d-1}}\cdot\eta_{out}\left(t,q_j^{d-1},q^{d-2}\right)$$

5. **Vertical Transition Probabilities**

$$\chi\left(t,q_i^d,q^{d-1}\right)=P\left(q_i^d \text{ started at } t|\lambda,O\right)$$
$$=P(o_1\cdots o_{t-1},\ \underset{q_i^d}{\overset{q^{d-1}}{\downarrow}}\ ,o_t\cdots o_N|\lambda,O)$$

Initiation:

$$\chi\left(1,q_i^2,q^1\right)=\frac{\pi\left(q_i^2|q^1\right)\cdot\beta\left(1,N,q_i^2,q^1\right)}{P\left(O|\lambda\right)}$$

Iteration:

For $2<d$

$$\chi\left(t,q_i^d,q_j^{d-1}\right)=\frac{\eta_{in}\left(t,q_j^{d-1},q^{d-2}\right)\cdot\pi\left(q_i^d|q_j^{d-1}\right)}{P\left(O|\lambda\right)}$$
$$\left[\sum_{k=t}^{N}\beta\left(t,k,q_i^d,q_j^{d-1}\right)\cdot\eta_{out}\left(k,q_j^{d-1},q^{d-2}\right)\right]$$

Parameter Estimation:

1. $\gamma_{in}\left(t,q_i^{d+1},q^d\right)$ is the probability of performing a horizontal transition to q_i^{d+1} which is sub-state of q^d before o_t is emitted

$$\gamma_{in}\left(t,q_i^{d+1},q^d\right)=\sum_{k=1}^{|q^d|}\xi\left(t-1,q_k^{d+1},q_i^{d+1},q^d\right)$$

2. $\gamma_{out}\left(t,q_i^{d+1},q^d\right)$ is the probability of performing a horizontal transition from q_i^{d+1} which is sub-state of q^d to any of the other sub-states of q^d after o_t is emitted

$$\gamma_{out}\left(t,q_i^{d+1},q^d\right)=\sum_{k=1}^{\left|q^d\right|}\xi\left(t,q_i^{d+1},q_k^{d+1},q^d\right)$$

Thus,

$$\hat{\pi}\left(q_i^2\middle|q^1\right)=\chi\left(t,q_i^2,q^1\right)$$

$$\hat{\pi}\left(q_i^{d+1}\middle|q^d\right)=\frac{\sum_{t=1}^{T}\chi\left(t,q_i^{d+1},q^d\right)}{\sum_{i=1}^{\left|q^d\right|}\sum_{t=1}^{T}\chi\left(t,q_i^{d+1},q^d\right)}\quad\left(1<d<D-1\right)$$

$$\hat{a}_{jk}^{q^d}=\frac{\sum_{t=1}^{N}\xi\left(t,q_i^{d+1},q_j^{d+1},q^d\right)}{\sum_{k=1}^{\left|q^d\right|}\sum_{t=1}^{N}\xi\left(t,q_i^{d+1},q_k^{d+1},q^d\right)}=\frac{\sum_{t=1}^{N}\xi\left(t,q_i^{d+1},q_j^{d+1},q^d\right)}{\sum_{t=1}^{N}\gamma_{out}\left(t,q_i^{d+1},q^d\right)}$$

$$\begin{aligned}\hat{e}\left(\sigma_l\middle|q^D,q^{D-1}\right)=&\Big(\sum_{o_t=\sigma_l}\chi\left(t,q_i^D,q^{D-1}\right)\\&+\sum_{t>1,o_t=\sigma_l}\gamma_{in}\left(t,q_i^D,q^{D-1}\right)\Big)/\Big(\sum_{t=1}^{T}\chi\left(t,q_i^D,q^{D-1}\right)\\&+\sum_{t=2}^{T}\gamma_{in}\left(t,q_i^D,q^{D-1}\right)\Big)\end{aligned}$$

A.1 MODIFIED VITERBI ALGORITHM WITH EXPLICIT STATE DURATION DENSITIES IN INTERNAL STATES OF HHMM

Every pair of states (state, parent) has three variables:

- $\delta\left(t,t+k,q_i^d,q^{d-1}\right)$: Likelihood of the most probable state sequence generating $o_t \ldots o_{t+k}$ assuming it was solely generated by a recursive activation that started at time step *t* from state q^{d-1} that ended at q_i^d and returned to q^{d-1} at time step $t+k$.
- $\psi\left(t,t+k,q_i^d,q^{d-1}\right)$: Index of the most probable state to be activated by q^{d-1} before q_i^d. If such a state does not exist ($o_t \ldots o_{t+k-z}$ was solely generated by q_i^d), we set $\psi\left(t,t+k,q_i^d,q^{d-1}\right)=-1$.
- $\tau\left(t,t+k,q_i^d,q^{d-1}\right)$: Time step at which q_i^d was most probable to be called by q^{d-1}. If q_i^d generated the entire subsequence, $\tau\left(t,t+k,q_i^d,q^{d-1}\right)=t$.

$$MAX_{l\in S}\left\{f(l)\right\}=\max_{l\in S},\arg\max_{l\in S}\left\{f(l)\right\}$$

Production State:

Initialization:

$$\begin{aligned}\delta\left(t,t,q_i^D,q^{D-1}\right)&=\Pr\left(q^{D-1}\text{activated } q_i^D\right)\cdot\Pr\left(q_i^D \text{ emitted } o_t\right)\\&=\pi^{q^{D-1}}\left(q_i^D\right)\cdot b^{q_i^D}\left(o_t\right)\end{aligned}$$

$$\psi\left(t,t,q_i^D,q^{D-1}\right)=-1$$
$$\tau\left(t,t,q_i^D,q^{D-1}\right)=t$$

Recursion:

$$\left(\delta\left(t,t+k,q_i^D,q^{D-1}\right),\psi\left(t,t+k,q_i^D,q^{D-1}\right)\right)=MAX_{\substack{1\le j\le\left|q^{D-1}\right| \\ j\neq i}}\left\{\begin{matrix}\delta\left(t,t+k-1,q_j^D,q^{D-1}\right) \\ \cdot a_{ji}^{q_j^{D-1}}\cdot b^{q_i^D}\left(o_{t+k}\right)\end{matrix}\right\}$$
$$\tau\left(t,t+k,q_i^D,q^{D-1}\right)=t+k$$

Internal State:

Initialization:

$$\delta\left(t,t,q_i^d,q^{d-1}\right)=\Pr\left(q^{d-1}\text{activated } q_i^d\right)\cdot\Pr\left(q_i^d\text{ emitted } o_t\right)$$
$$=MAX_{1\le s\le\left|q_i^d\right|}\left\{\begin{matrix}\pi^{q^{d-1}}\left(q_i^d\right)\cdot\left[\delta\left(t,t,q_s^{d+1},q_i^d\right)\right] \\ \cdot a_{s\,end}^{q_i^d}\cdot p^{q_i^d}(1)\end{matrix}\right\}$$

$$\psi\left(t,t,q_i^d,q^{d-1}\right)=-1$$
$$\tau\left(t,t,q_i^d,q^{d-1}\right)=t$$

Recursion:

$k<Z$

where Z is the maximum duration of q_i^d and $k>0$

For $t'=t+1,...,t+k$

$$\mathrm{X} = MAX_{1\le s\le \left|q_i^d\right|}\left\{\delta\left(t', t+k, q_s^{d+1}, q_i^d\right)\cdot a_{s\,end}^{q_i^d}\right\}$$

$$\left(\Delta(t'),\Psi(t')\right) = MAX_{\substack{1\le j\le \left|q^{d-1}\right| \\ j\ne i}}\left\{\begin{matrix}\delta\left(t,t'-1,q_j^d,q^{d-1}\right)\cdot a_{ji}^{q^{d-1}} \\ \cdot p^{q_i^d}\left(k-\left(t'-t\right)\right)\cdot \mathrm{X}\end{matrix}\right\}$$

For t,

$$\Delta(t) = \pi^{q^{d-1}}\left(q_i^d\right)\cdot MAX_{1\le s\le \left|q_i^d\right|}\left\{\delta\left(t,t+k,q_s^{d+1},q_i^d\right)\cdot a_{s\,end}^{q_i^d}\right\}\cdot p^{q_i^d}(k+1)$$

$$\Psi(t) = -1$$

Most probable switching time:

$$\delta\left(t,t+k,q_i^d,q^{d-1}\right),\tau\left(t,t+k,q_i^d,q^{d-1}\right) = MAX_{t\le t'\le t+k}\Delta(t')$$
$$\psi\left(t,t+k,q_i^d,q^{d-1}\right) = \Psi\left(\tau\left(t,t+k,q_i^d,q^{d-1}\right)\right)$$

$k \ge Z$ and $k > 0$

For $t' = t+k-Z+1,\ldots,t+k$

$$\mathrm{X} = MAX_{1\le s\le \left|q_i^d\right|}\left\{\delta\left(t', t+k, q_s^{d+1}, q_i^d\right)\cdot a_{s\,end}^{q_i^d}\right\}$$

$$\left(\Delta(t'),\Psi(t')\right) = MAX_{\substack{1\le j\le \left|q^{d-1}\right| \\ j\ne i}}\left\{\begin{matrix}\delta\left(t,t'-1,q_j^d,q^{d-1}\right)\cdot a_{ji}^{q^{d-1}} \\ \cdot p^{q_i^d}\left(k-\left(t'-t\right)\right)\cdot \mathrm{X}\end{matrix}\right\}$$

Most probable switching time:

$$\delta\left(t,t+k,q_i^D,q^{D-1}\right),\tau\left(t,t+k,q_i^D,q^{D-1}\right) = MAX_{t+k-Z+1\le t'\le t+k}\Delta(t')$$
$$\psi\left(t,t+k,q_i^D,q^{D-1}\right) = \Psi\left(\tau\left(t,t+k,q_i^D,q^{D-1}\right)\right)$$

APPENDIX B

SUMMARIZATION OF CHARACTERISTICS OF MIRNA HAIRPIN STRUCTURES OF VARIOUS SPECIES

HP: Hairpin Length

LP: Loop Length

MIR: miRNA Length

EXT: Extension Length

PRI: Primary extension Length

Invertebrates:

	HP	LP	MIR	EXT	PRI
Mean					
Anopheles Gambiae	89.8	7.4	22.4	5.7	13.3
Apis mellifera	94.4	9.0	22.3	3.9	15.6
Bombyx mori	91.0	8.0	22.1	4.6	15.0
Drosophila melanogaster	89.1	8.0	22.3	6.0	12.4
Drosophila pseudoobscura	87.0	8.3	22.5	5.7	11.3
Schmidtea mediterranea	88.4	7.2	21.7	5.9	13.0
Caenorhabditis briggsae	96.4	7.5	22.3	6.3	16.3
Caenorhabditis elegans	95.6	7.9	22.2	6.1	14.9
Standard Deviation					
Anopheles Gambiae	8.7	3.5	1.4	2.5	5.1
Apis mellifera	7.4	4.1	1.6	2.7	4.9
Bombyx mori	10.1	3.7	1.7	4.0	3.3
Drosophila melanogaster	18.5	4.1	1.2	6.9	6.6
Drosophila pseudoobscura	11.1	3.9	1.3	3.8	5.9
Schmidtea mediterranea	12.7	4.3	1.2	3.9	4.9
Caenorhabditis briggsae	11.0	3.4	1.1	3.3	6.9
Caenorhabditis elegans	9.2	3.9	1.2	3.3	4.6
Minimum					
Anopheles Gambiae	69.0	3.0	20.0	1.0	2.0
Apis mellifera	77.0	3.0	19.0	0.0	4.0
Bombyx mori	80.0	4.0	19.0	1.0	9.0
Drosophila melanogaster	54.0	3.0	20.0	0.0	0.0
Drosophila pseudoobscura	62.0	3.0	20.0	1.0	0.0
Schmidtea mediterranea	59.0	3.0	18.0	0.0	2.0
Caenorhabditis briggsae	67.0	3.0	19.0	1.0	4.0
Caenorhabditis elegans	56.0	3.0	19.0	0.0	0.0
Maximum					
Anopheles Gambiae	112.0	16.0	28.0	14.0	29.0
Apis mellifera	107.0	18.0	28.0	11.0	30.0
Bombyx mori	122.0	15.0	27.0	16.0	20.0
Drosophila melanogaster	215.0	22.0	28.0	55.0	31.0
Drosophila pseudoobscura	110.0	17.0	28.0	18.0	28.0
Schmidtea mediterranea	128.0	26.0	24.0	18.0	26.0
Caenorhabditis briggsae	116.0	16.0	26.0	21.0	32.0
Caenorhabditis elegans	110.0	30.0	25.0	20.0	27.0

Vertebrates:

	HP	LP	MIR	EXT	PRI
Mean					
Danio rerio	94.1	7.6	22.1	4.8	16.4
Fugu rubripes	80.8	7.9	22.1	4.7	9.5
Tetraodon nigroviridis	80.3	7.8	22.1	4.7	9.0
Xenopus tropicalis	82.9	7.8	21.9	4.6	11.0
Gallus gallus	86.7	7.2	21.9	4.7	12.9
Bos taurus	84.8	7.0	22.2	4.7	11.8
Canis familiaris	95.8	7.2	22.0	7.7	15.5
Homo sapiens	88.4	6.9	21.9	5.4	13.0
Monodelphis domestica	80.9	8.4	21.7	4.5	10.0
Macaca mulatta	87.6	7.1	21.8	5.1	13.1
Mus musculus	85.6	7.0	21.9	5.5	11.8
Pan troglodytes	88.1	7.3	21.8	4.6	14.2
Rattus norvegicus	86.4	7.0	22.0	4.6	13.0
Standard Deviation					
Danio rerio	18.1	3.6	0.6	3.3	9.2
Fugu rubripes	10.0	4.1	0.6	2.5	5.2
Tetraodon nigroviridis	10.6	4.0	0.6	2.7	5.8
Xenopus tropicalis	10.5	3.7	1.0	2.7	5.5
Gallus gallus	11.4	3.6	1.1	3.1	6.3
Bos taurus	12.5	3.5	0.9	2.9	6.3
Canis familiaris	11.6	3.4	0.0	3.8	4.6
Homo sapiens	12.4	3.0	1.0	3.8	6.1
Monodelphis domestica	12.0	4.4	1.3	2.7	5.6
Macaca mulatta	13.4	2.7	0.9	3.5	7.3
Mus musculus	14.3	3.2	1.0	4.7	7.3
Pan troglodytes	13.5	3.0	1.0	2.6	7.2
Rattus norvegicus	11.3	3.2	0.9	2.7	5.6
Minimum					
Danio rerio	63.0	3.0	18.0	0.0	0.0
Fugu rubripes	63.0	3.0	20.0	0.0	0.0
Tetraodon nigroviridis	62.0	3.0	20.0	0.0	0.0
Xenopus tropicalis	60.0	3.0	17.0	0.0	0.0
Gallus gallus	63.0	3.0	17.0	0.0	0.0
Bos taurus	59.0	3.0	20.0	0.0	0.0
Canis familiaris	85.0	3.0	22.0	2.0	10.0
Homo sapiens	55.0	3.0	19.0	0.0	0.0
Monodelphis domestica	57.0	3.0	16.0	0.0	2.0
Macaca mulatta	64.0	4.0	19.0	0.0	0.0
Mus musculus	61.0	3.0	18.0	0.0	0.0
Pan troglodytes	62.0	3.0	19.0	0.0	0.0
Rattus norvegicus	59.0	3.0	19.0	0.0	0.0
Maximum					

Danio rerio	153.0	18.0	24.0	22.0	50.0
Fugu rubripes	112.0	22.0	25.0	13.0	27.0
Tetraodon nigroviridis	122.0	22.0	25.0	13.0	38.0
Xenopus tropicalis	109.0	20.0	24.0	16.0	30.0
Gallus gallus	112.0	20.0	25.0	18.0	31.0
Bos taurus	119.0	16.0	25.0	17.0	25.0
Canis familiaris	111.0	12.0	22.0	13.0	22.0
Homo sapiens	137.0	16.0	25.0	26.0	35.0
Monodelphis domestica	111.0	22.0	24.0	18.0	25.0
Macaca mulatta	112.0	14.0	24.0	25.0	30.0
Mus musculus	128.0	17.0	26.0	34.0	37.0
Pan troglodytes	119.0	14.0	24.0	14.0	35.0
Rattus norvegicus	112.0	21.0	25.0	17.0	32.0

Plants:

	HP	LP	MIR	EXT	PRI
Mean					
Triticum aestivum	154.7	11.4	22.3	29.6	19.5
Zea mays	129.0	5.8	20.7	25.3	15.4
Arabidopsis thaliana	138.2	6.0	21.1	25.7	19.5
Populus trichocarpa	110.4	6.6	21.3	22.8	8.1
Vitis vinifera	109.4	7.7	21.0	19.7	10.3
Oryza sativa	136.0	6.2	21.4	30.3	13.3
Standard Deviation					
Triticum aestivum	34.1	7.9	1.8	21.9	11.2
Zea mays	34.2	2.5	0.5	14.8	7.2
Arabidopsis thaliana	50.1	3.5	0.8	17.6	17.8
Populus trichocarpa	31.2	3.6	1.1	15.3	2.8
Vitis vinifera	27.3	4.4	0.5	14.2	0.8
Oryza sativa	50.9	2.5	1.0	22.4	9.8
Minimum					
Triticum aestivum	107.0	4.0	19.0	2.0	2.0
Zea mays	74.0	3.0	20.0	8.0	1.0
Arabidopsis thaliana	73.0	3.0	20.0	2.0	0.0
Populus trichocarpa	58.0	3.0	17.0	0.0	0.0
Vitis vinifera	83.0	3.0	20.0	1.0	8.0
Oryza sativa	60.0	3.0	20.0	0.0	0.0
Maximum					
Triticum aestivum	218.0	35.0	24.0	77.0	39.0
Zea mays	221.0	14.0	22.0	68.0	20.0
Arabidopsis thaliana	337.0	18.0	24.0	89.0	78.0
Populus trichocarpa	226.0	28.0	24.0	79.0	11.0
Vitis vinifera	222.0	25.0	22.0	83.0	13.0
Oryza sativa	312.0	18.0	24.0	102.0	51.0

APPENDIX C

COMPARISON OF MATURE MIRNA SEQUENCES BETWEEN S. PURPURATUS ADULT AND EMBRYO DATA

Adult data is from (B. M. Wheeler et al. 2009). Differences are highlighted in bold. **E:** Embryonic data from Illumina platform; **A:** Adult data from 454 sequencing platform.

miRNA	Sequence
spu-let-7	E: TGAGGTAGTAGGTTATATAGTT A: TGAGGTAGTAGGTTATATAGTT
spu-miR-1	E: TGGAATGTAAAGAAGTATGTAT A: TGGAATGTAAAGAAGTATGTAT
spu-miR-1b	E: TGGAATGTAAAGAAGTATGTAC
spu-miR-10	E: AACCCTGTAGATCCGAATTTGTG A: AACCCTGTAGATCCGAATTTGTG
spu-miR-125	E: TCCCTGAGACCCTAACTTGTGA A: TCCCTGAGACCCTAACTTGTGA
spu-miR-137	E: **T**TATTGCTTGAGAATACACGTA- A: -TATTGCTTGAGAATACACGTA**G**
spu-miR-153	E: TTGCATAGTCACAAAAGTGATT A: TTGCATAGTCACAAAAGTGATT
spu-miR-182	E: TTTGGCAATTGATAGAATTCACACT A: TTTGGCAATTGATAGAATTCACACT
spu-miR-183	E: TATGGCACTATAGAATTCACTG A: TATGGCACTATAGAATTCACTG
spu-miR-184	E: TGGACGGAGAACTGATAAGGGC A: TGGACGGAGAACTGATAAGGGC

spu-miR-200	E: TAATACTGTCTGGTGATGATGTT A: TAATACTGTCTGGTGATGATGTT
spu-miR-2001	E: ATGTGACCGATATAATGGGCAT A: ATGTGACCGATATAATGGGCAT
spu-miR-2002	E: TGAATACATCTGCTGGTTTTTAT A: TGAATACATCTGCTGGTTTTTAT
spu-miR-2003	E: AACCCGTAAGGTCTTAACTTGTG A: AACCCGTAAGGTCTTAACTTGTG
spu-miR-2004	E: TCACACACAACCACAGGAAGTT A: TCACACACAACCACAGGAAGTT
spu-miR-2007	E: TATTTCAGGCAGTATACTGGTAA A: TATTTCAGGCAGTATACTGGTAA
spu-miR-2008	E: ATCAGCCTCGCTGTCAATACG**A** A: ATCAGCCTCGCTGTCAATACG-
spu-miR-2009	E: TGAGTTGTCCCACAAAGAACAC- A: TGAGTTGTCCCACAAAGAACAC**A**
spu-miR-2010	E: TTACTGTTGATGTCAGCCCCTT A: TTACTGTTGATGTCAGCCCCTT
spu-miR-2011	E: ACCAAGGTGTGCTAGTGATGAC A: ACCAAGGTGTGCTAGTGATGAC
spu-miR-2012	E: TAGTACTGGCATATGGACATTG A: TAGTACTGGCATATGGACATTG
spu-miR-2013	E: TGCAGCATGATGTAGTGGTGT**A** A: TGCAGCATGATGTAGTGGTGT-
spu-miR-210	E: TTGTGCGTGCGACAGCGACTGA A: TTGTGCGTGCGACAGCGACTGA
spu-miR-219	E: TGATTGTCCGAACGCAATTCTTG A: TGATTGTCCGAACGCAATTCTTG
spu-miR-22	E: **T**CAGCTGCCCGGTGAAGTGTATA A: -CAGCTGCCCGGTGAAGTGTATA
spu-miR-242	E: TTGCGTAGGCGTTGTGCACAGT A: TTGCGTAGGCGTTGTGCACAGT
spu-miR-252a	E: CTAAGTACTAGTGCCGTAGGTT A: CTAAGTACTAGTGCCGTAGGTT
spu-miR-252b	E: CTAAGTAGTAGTGCCGCAGGTA A: CTAAGTAGTAGTGCCGCAGGTA
spu-miR-29	E: AAGCACCAGTTGAAATCAGAGC A: AAGCACCAGTTGAAATCAGAGC
Spu-miR-29b	E: TAGCACCATGAGAAAGCAGTAT A: TAGCACCATGAGAAAGCAGTAT

spu-miR-31a	E: AGGCAAGATGTTGGCATAGCT**G** A: AGGCAAGATGTTGGCATAGCT-
spu-miR-31b	E: AGGCAAGATGCTGGCATAGCT
spu-miR-33	E: GTGCATTGTCGTTGCATTGCAT A: GTGCATTGTCGTTGCATTGCAT
spu-miR-34	E: CGGCAGTGTAGTTAGCTGGTTG A: CGGCAGTGTAGTTAGCTGGTTG
spu-miR-4847	E: TAATGATGGCGCGGTGCGGTGC A: TAATGATGGCGCGGTGCGGTGC
spu-miR-4850	E: TTATCATGACTGTAAACAGGAGG A: TTATCATGACTGTAAACAGGAGG
spu-miR-4851	E: TGATTACTTGCTTTGGAGTTCTT A: TGATTACTTGCTTTGGAGTTCTT
spu-miR-4854	E: TGTTGCAGTGACGACTTCGCGC A: TGTTGCAGTGACGACTTCGCGC
spu-miR-4855	E: TGTGTAACATCTCATTCAGTGGGT A: TGTGTAACATCTCATTCAGTGGGT
spu-miR-7	E: TGGAAGACTAGTGATTTTGTTGT A: TGGAAGACTAGTGATTTTGTTGT
spu-miR-71	E: TGAAAGACATGGGTAGTGAGATT A: TGAAAGACATGGGTAGTGAGATT
spu-miR-79	E: ATAAAGCTAGGTTACCAAAGAT**A** A: ATAAAGCTAGGTTACCAAAGAT-
spu-miR-9	E: TCTTTGGTTATCTAGCTGTATG A: TCTTTGGTTATCTAGCTGTATG
spu-miR-92a	E: TATTGCACTTGTCCCGGCCTAC- A: TATTGCACTTGTCCCGGCCTAC**T**
spu-miR-92b	E: TATTGCACTTGTCCCGGCCTGC A: TATTGCACTTGTCCCGGCCTGC
spu-miR-92c	E: TATTGCACTCGTCCCGGCCTGC A: TATTGCACTCGTCCCGGCCTGC
spu-miR-96	E: TTTGGCACTAGCACATTTTGC A: TTTGGCACTAGCACATTTTGC

APPENDIX D

COMPARISON OF MATURE MIRNA SEQUENCE DATA BETWEEN TWO SEA STAR SPECIES

Comparison of mature miRNA sequences between *H.sanguinolenta* adult data (B. M. Wheeler et al. 2009) and *P. miniata* embryonic data. Differences are highlighted in bold. *Pmi: P. miniata; Hsn: H. sanguinolenta.*

miR-1	Pmi TGGAATGTAAAGAAGTATGTAT Hsn TGGAATGTAAAGAAGTATGTAT
miR-7	Pmi TGGAAGACTAGTGATTTTGTTGT Hsn TGGAAGACTAGTGATTTTGTTGT
miR-8, 141, 200	Pmi TAATACTGTCTGGTAATGATGTT Hsn R1 TAATACTGTCTGGTAATGATGT-
miR-9	Pmi TCTTTGGTTATCTAGCTGTATGA Hsn R1 TCTTTGGTTATCTAGCTGTATGA
miR-10	Pmi AACCCTGTAGATCCGAATTTGT**G** Hsn R1 AACCCTGTAGATCCGAATTTGT-
miR-22, 745, 980	Pmi TCAGCTGCCCGGTGAAGTGTAG Hsn TCAGCTGCCCGGTGAAGTGTAG
miR-29, 83, 285	Pmi AAGCACCAGTTGAAATCAGAGC Hsn R1 AAGCACCAGTTGAAATCAGAGC

miR-31	Pmi 31a AGGCAAGATGTTGGCATAGCTG 31b AGGCAAGATG**C**TGGCATAGCTG Hsn R1 AGGCAAGATGTTGGCATAGCT-
miR-33	Pmi GTGCATTGTAGTTGCATTGCAT Hsn GTGCATTGTAGTTGCATTGCAT
miR-34	Pmi TGGCAGTGTGGTTAGCTGGTTG Hsn TGGCAGTGTGGTTAGCTGGTTG
miR-71	Pmi TGAAAGACATGGGTAGTGAGAT Hsn R1 TGAAAGACATGGGTAGTGAGAT
miR-79	Pmi ATAAAGCTAGGTTACCAAAGATA Hsn –TAAAGCTAGGTTACCAAAGAT-
miR-92	Pmi 92a TATTGCACTTGTC**C**CGGCC**A**GC 92b TATTGCACTTGTCTCGGCCTGC 92c TATTGCACT**C**GTC**C**GGCCTGC 92d TATTGCACT**C**GTC**C**GGCCT**AG** Hsn TATTGCACTTGTCTCGGCCTGC
miR-96	Pmi TTTGGCACTAGCACATTTTGC- Hsn TTTGGCACTAGCACATTTTGCT
miR-100	Pmi AACCCGTAGATCCGAACTTGT Hsn AACCCGTAGATCCGAATTTGT
miR-124	Hsn TAAGGCACGCGGTGAATGCCA
miR-125	Pmi TCCCTGAGACCCTAACTTGTGA Hsn R1 TCCCTGAGACCCTAACTTGTGA
miR-133	Hsn TTTGGTCCCCTTCAACCAGCCGT
miR-137	Pmi TATTGCTTGAGAATACACGTAG Hsn TATTGCTTGAGAATACACGTAG
miR-153	Hsn TTGCATAGTCACAAAAGTGATT
miR-1692	Pmi TGTAGCTCAGTTGGTAGAG
miR-182, 263b	Pmi TTTGGCAATAGATAGAATTCACA Hsn TTTGGCAATAGATAGAATTCAC-
miR-183	Pmi TATGGCACTGTAGAATTCACT

APPENDIX E

ALIGNMENT OF MATURE MIRNA SEQUENCES IN TWO ECHINODERMS AND A HEMICHORDATE OUTGROUP SPECIES.

spu - S. purpuratus; pmi - P. miniata; sko - S. kowalevskii.

```
spu-let-7         TGAGGTAGTAGGTTATATAGTT 22
sko-let-7         TGAGGTAGTAGGTTGTATAGTT 22
                  ********************** *******

spu-miR-1         TGGAATGTAAAGAAGTATGTAT 22
pmi-miR-1         TGGAATGTAAAGAAGTATGTAT 22
sko-miR-1         TGGAATGTAATGAAGTATGTAT 22
                  ********** ***********

spu-miR-1b        TGGAATGTAAAGAAGTATGTAC 22
sko-miR-1b        TGGAATGTAATGAAGTATGTAT 22
                  ********** **********

spu-miR-10        AACCCTGTAGATCCGAATTTGTG 23
pmi-miR-10        AACCCTGTAGATCCGAATTTGTG 23
sko-miR-10        TACCCTGTAGATCCGAATTTGTG 23
                   **********************

pmi-miR-100       AACCCGTAGATCCGAACTTGT- 21
sko-miR-100       AACCCGTAGATCCGAACTTGTG 22
                  ********************
```

```
spu-miR-124       TAAGGCACGCGGTGAATGCCA- 21
sko-miR-124       TAAGGCACGCGGTGAATGCCAA 22
                  **********************

spu-miR-125       TCCCTGAGACCCTAACTTGTGA 22
pmi-miR-125       TCCCTGAGACCCTAACTTGTGA 22
sko-miR-125       TCCCTGAGACCCTAACTTGTGA 22
                  **********************

spu-miR-133       TTTGGTCCCCTTCAACCAGCCGT 23
sko-miR-133       -TTGGTCCCCTTCAACCAGCTGT 22
                   ******************* **

spu-miR-137       TTATTGCTTGAGAATACACGT—- 21
pmi-miR-137       -TATTGCTTGAGAATACACGTAG 22
sko-miR-137       -TATTGCTTGAGAATACACGTAG 22
                   ********************

spu-miR-153       TTGCATAGTCACAAAAGTGATT 22
sko-miR-153       TTGCATAGTCACAAAAGTGATT 22
                  **********************

pmi-miR-1692      TGTAGCTCAGTTGGTAGAG 19

spu-miR-182       TTTGGCAATTGATAGAATTCACACT 25
pmi-miR-182       TTTGGCAATAGATAGAATTCACA-- 23
sko-miR-182       TTTGGCAATAGATAGAATTCACA-- 23
                  ********* *************

spu-miR-183       TATGGCACTATA-GAATTCACTG 22
pmi-miR-183       TATGGCACTGTA-GAATTCACT- 21
sko-miR-183       AATGGCACTGTATGAATTCACTG 23
                   ******** ** **********


spu-miR-184       TGGACGGAGAACTGATAAGGGC 22
pmi-miR-184       TGGACGGAGAACTGATAAGGGC 22
sko-miR-184       TGGACGGAGAACTGATAAGGGC 22
                  **********************

spu-miR-200       TAATACTGTCTGGTGATGATGTT 23
pmi-miR-200       TAATACTGTCTGGTAATGATGTT 23
sko-miR-200       TAATACTGTCTGGTAATGATGTT 23
                  ************** ********
```

```
spu-miR-2001      ATGTGACCGATATAATGGGCAT 22
pmi-miR-2001      ATGTGACCGTTACAATGGGCAT 22
sko-miR-2001      TTGTGACCGTTATAATGGGCAT 22
                   ******** ** *********

spu-miR-2002      TGAATACATCTGCTGGTTTTTAT 23

spu-miR-2003      AACCCGTAAGGTCTTAACTTGTG 23

spu-miR-2004      TCACACACAACCACAGGAAGTT 22
pmi-miR-2004      TCACACACAACCACAGGAAGTT 22
                  **********************

spu-miR-2005      AGTCCAATAGGGAGGGCATTGCA 23

spu-miR-2006      GAGCACACTTGGTAGCGGTGCC 22
pmi-miR-2006      GAGCACACTTGGTAGCGGTGCC 22
                  **********************

spu-miR-2007      TATTTCAGGCAG-TATACTGGTAA  23
pmi-miR-2007      TATTTCAGGCGG-TATACTGGTAA  23
sko-miR-2007      TATTTCAGGCGTTTATACTGGTGA  24
                  **********   ********* *

spu-miR-2008      ATCAGCCTCGCTGTCAATACGA 22
sko-miR-2008      ATCAGCCTCGCTGTCAATACGG 22
                  **********************

spu-miR-2009      TGAGTTGTCCCACAAAGAACAC 22
pmi-miR-2009      TGAGTTGTCCCACAAAGAACAC 22
                  **********************

spu-miR-2010      TTACTGTTGATGTCAGCCCCTT 22
pmi-miR-2010      TTACTGTTGATGTCAGCCCCTC 22
                  **********************

spu-miR-2011      ACCAAGGTGTGCTAGTGATGAC 22
pmi-miR-2011      ACCAAGGTGTGTTAGTGATGAC 22
sko-miR-2011      ACCAAGGTGTGTTAGTGATGAC 22
                  ********** **********

spu-miR-2012      TAGTACTGGCATATGGACATTG 22
pmi-miR-2012      TAGTACTGGCATATGGACATT- 21
sko-miR-2012      TAGTACTGGCATATGGACATTG 22
                  **********************
```

```
spu-miR-2013      TGCAGCATGATGTAGTGGTGTA 21
pmi-miR-2013      TGCAGCATGATGTAGTGGTG-A 22
sko-miR-2013      TGCAGCATGATGTAGTGGTGTG 22
                  ********************

spu-miR-210       TTGTGCGTGCGACAGCGACTGA 22
sko-miR-210       TTGTGCGTGCGACAGCGACTTC 22
                  ********************

spu-miR-219       TGATTGTCCGAACGCAATTCTTG 23

spu-miR-22        TCAGCTGCCCGGTGAAGTGTATA 23
pmi-miR-22        TCAGCTGCCCGGTGAAGTGTAG- 22
                  **********************

spu-miR-242       TTGCGTAGGCGTTGTGCACAGT- 22
pmi-miR-242       TTGCGTAGGCGTTGTGCACAGT- 22
sko-miR-242       -TGCGTAGGCGTTGTGCACAGTG 22
                   **********************

spu-miR-252a      CTAAGTACTAGTGCCGTAGGTT- 22
pmi-miR-252a      CTAAGTACTAGTGCCGCAGGTTG 23
sko-miR-252a      CTAAGTACTAGTGCCGCAGGAGT 23
                  *************** ***

spu-miR-252b      CTAAGTAGTAGTGCCGCAGGTA- 22
pmi-miR-252b      CTAAGTAGTAGTGCCGCAGGTA- 23
sko-miR-252b      CTAAGTAGTAGTGCCGCAGGTAA 23
                  ***********************

spu-miR-278       TCGGTGGGACTTTCGTTCGATT 22
pmi-miR-278       TCGGTGGGACTTTCGTTCGATT 22
sko-miR-278       TCGGTGGGACTTTCGTTCGTTT 22
                  ****************** **

spu-miR-29        AAGCACCAGTTGAAATCAGAGC 22
pmi-miR-29        AAGCACCAGTTGAAATCAGAGC 22
sko-miR-29b       TAGCACCATTTGAAATCAGTGT 22
                   ******* ********** *

spu-miR-29b       TAGCACCATGAGAAAGCAGTAT 22
sko-miR-29        TAGCACCATATGAAATCAGTTT 22
sko-miR-29b       TAGCACCATTTGAAATCAGTGT 22
                  *********  ********* *
```

```
spu-miR-31a       AGGCAAGATGTTGGCATAGCTG 22
pmi-miR-31a       AGGCAAGATGTTGGCATAGCTG 22
sko-miR-31a       AGGCAAGATGTTGGCATAGCTG 22
                  **********************

spu-miR-31b       AGGCAAGATGCTGGCATAGCT 21
pmi-miR-31b       AGGCAAGATGCTGGCATAGCT 21
                  ********************

spu-miR-33        GTGCATTGTCGTTGCATTGCAT 22
pmi-miR-33        GTGCATTGTAGTTGCATTGCAT 22
                  ********* ************

spu-miR-34        CGGCAGTGTAGTTAGCTGGTTG 22
pmi-miR-34        TGGCAGTGTGGTTAGCTGGTTG 22
sko-miR-34        TGGCAGTGTGGTTAGCTGGTTG 22
                   ******** ************

spu-miR-375       -TTGTTCGTTCGGCTCGCGTCAA 22
sko-miR-375       TTTGTTCGTTCGGCTCGCGCGA- 22
                   ******************  *

pmi-miR-4171      TGACTCTCTTAAGGTAGCC 19

spu-miR-4847      TAATGATGGCGCGGTGCGGTGC 22

spu-miR-4848a     TGGGTTGAGGCTTTTGGGCAGGA 23

spu-miR-4848b     TGGGTTGAGGCTTTGGGGCAGGA 23

spu-miR-4849      TAATGATGGCGCGGTGCGGTGC 22

spu-miR-4850      TTATCATGACTGTAAACAGGAGG 23

spu-miR-4851      TGATTACTTGCTTTGGAGTTCTT 23

spu-miR-4852      AATTCTATCATTTTGGCTGCAT 22

spu-miR-4853      TAGCTCCGTTGTTGCGTCTTGGTA 24

spu-miR-4854      TGTTGCAGTGACGACTTCGCGC 22

spu-miR-4855      TGTGTAACATCTCATTCAGTGGGT 24
```

```
spu-miR-7         TGGAAGACTAGTGATTTTGTTGT 23
pmi-miR-7         TGGAAGACTAGTGATTTTGTTGT 23
sko-miR-7         TGGAAGACTAGTGATTTTGTTGT 23
                  ***********************

spu-miR-71        TGAAAGACATGGGTAGTGAGATT 23
pmi-miR-71        TGAAAGACATGGGTAGTGAGAT- 22
sko-miR-71        TGAAAGACACAGGTAGTGAGAT- 22
                  *********  ***********

spu-miR-79        ATAAAGCTAGGTTACCAAAGATA 23
sko-miR-79        ATAAAGCTAGGTTACCAAAGACA 23
                  ********************** *

spu-miR-9         TCTTTGGTTATCTAGCTGTATG- 22
pmi-miR-9         TCTTTGGTTATCTAGCTGTATGA 23
sko-miR-9         TCTTTGGTTATCTAGCTGTAT-- 21
                  **********************

spu-miR-92a       TATTGCACTTGTCCCGGCCTAC 22
pmi-miR-92a       TATTGCACTTGTCCCGGCCAGC 22
sko-miR-92a       TATTGCACTTGTCCCGGCCTAA 22
                  ******************

spu-miR-92b       TATTGCACTTGTCCCGGCCTGC 22
pmi-miR-92b       TATTGCACTTGTCTCGGCCAGC 22
sko-miR-92b       TATTGCACTTGTCCCGGCCTGC 22
                  ****************** **

spu-miR-92c       TATTGCACTCGTCCCGGCCTGC 22
pmi-miR-92c       TATTGCACTCGTCCCGGCCTGC 22
sko-miR-92c       TATTGCACTCGTCCCGGCCTGT 22
                  **********************

pmi-miR-92d       TATTGCACTCGTCCCGGCCTAG 22

spu-miR-96        TTTGGCACTAGCACATTTTGC 21
pmi-miR-96        TTTGGCACTAGCACATTTTGC 21
sko-miR-96        TTTGGCACTAGCACATTTTGC 21
                  **********************

spu-miR-981       TTCGTTGTCAACGAAACCTGC 21
```

APPENDIX F

PRIMERS USED FOR VARIOUS EXPERIMENTS

F.1 NORTHERN BLOT

miRNA	Sequence	Signal found in Northern blot
miR-124	AGGCAAGAUGUUGGCAUAGCUGA	
miR-125	UUGCAUAGUCACAAAAGUGAUC	
miR-10	AGGCAAGAUGUUGGCAUAGCUGA	
miR-1	UUGCAUAGUCACAAAAGUGAUC	
miR-7	CACGCUCAUGCACACACCCACA	No
miR-9	UGAAAGACAUGGGUAGUGA	
miR-31/72	AGGCAAGAUGUUGGCAUAGCUGA	
miR-153	UUGCAUAGUCACAAAAGUGAUC	
miR-574	CACGCUCAUGCACACACCCACA	No
miR-71	UGAAAGACAUGGGUAGUGA	

F.2 MORPHOLINO SEQUENCES

Splice junction morpholinos for *SpDicer*

SpDicer-1 (Exon-intron boundary) Exon3-Intron3:

ATAGATCTCATAAGGACATT[AAAAAgtgagtggctcttcactattc]tgtt

Morpholino sequence: GAATAGTGAAGAGCCACTCACTTTT

SpDicer-2 (Exon-intron boundary): Intron14-Exon15

gtctgt[actctcctattctacacagGTTAGT]AATTTCAACCTGTACTGCT

Morpholino sequence: ACTAACCTGTGTAGAATAGGAGAGT

SpAgo1

CTGACATGCTGAAATAAACGTTCAACTACGGCCTATTTTGCCATTATTGAAAT

TTATATTGTATAATTTTTGAA[AGAGGAAAG(ATG)TATCAACCACCCT]TTCCG

Morpholino sequence: AGGGTGGTTGATACATCTTTCCTCT

SpAgo2

CAAATACAACTCAATGTCAATATACTATG[GCCCTAATAGTAACGACAAACTG

AA]ATAACATC(ATG)TACCAGCCACCACCGCATCCGA

Morpholino sequence: TTCAGTTTGTCGTTACTATTAGGGC

F.3 3'UTR PRIMERS FOR REPORTER CONSTRUCTS

Gene Name	Forward Primer	Reverse Primer	Length of amplicon
Alx1	GGGAGTCCTGAAGCTTAGTG	ATCCATGCTCTTTCCACCGA	2.3kb
Blimp1	GTGTGCTCGCTTTGGCTAGT	AGGCACAATCCTGTTGGAAG	1.6kb
Ets1	TGTGCACTGCGCAAGAATAC	TCTCGACATTCTGCTGATCC	1.3kb
FoxJ1	AATGTTGTGTGAGGACCAGG	GAACGTACGCTATGTTTCGC	1.1kb
Gsc	GACTGTTTTGATGTGCTTCT	AGGAAGGGAACATCTCGTTG	1.7kb
Hox11/13b	CTCTTCTGTTGTAGGCACGC	TTGGACAAGAAAGCGATCGG	2.1kb
Nk2.1	TGATAAGCCTCCTAAGGCCG	ACACCTTCTTCCGTCATAGC	1.9kb
SoxB1	AATAGTATGCGACGAAACGG	AAACACACACGCTAACATCC	1.3kb
Tgif	GCAGATTGAAACGGTACAGC	CATACATTGCACAGGGACGG	1.8kb
GataE	CCAGAGGAAATCACCAGAGA	CAGCATATCCCTTCTGGTCAG	1.6kb

F.4 RESULTS OF LIBRARY SCREEN IN *P. MINIATA*

Gene Name	Forward Primer	Reverse Primer	Clone number
PmDicer	TCCAACTACCAGCAGCCTCT	TTGAATCACCTTGGCTTTCC	233N24
PmDGCR8	AGGAGAGCCATCAACAATGG	GTATGCTTGCCAAAGGTGGT	

F.5 PRIMERS FOR DICER GENE IN *P. MINIATA*

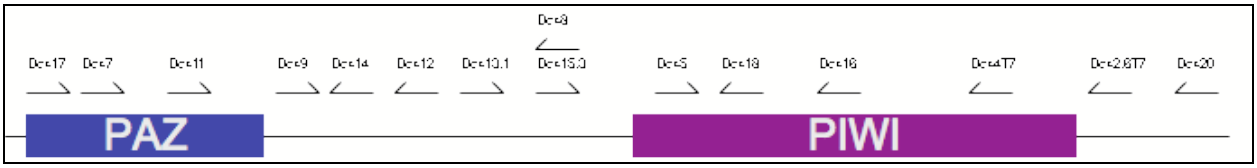


Figure F.5. 1: Primers for cloning, WMISH and RT-PCR *PmDicer*: The direction of the arrows represent the sense and antisense primers (→: Forward; ←:Reverse).

F.6 RACE PRIMERS

Gene Name	Primer sequence	Type of RACE Primer
AmDGCR8-R1	AGCCAATGCAGACGGGGCACCAGAC	3’ RACE
AmDGCR8-R3	CCCCGGGGACAAAAGTACGGGCAAAGG	
SpAgo-R-4	GGAAACCAAACCACACCTCTC	5’ RACE
SpAgo2-R-6	CGCTGGATAGGTGCCTGGGGGA	5’ RACE

APPENDIX G

SECONDARY RNA STRUCTURES OF CONSERVED MIRNAS FOUND IN SEA URCHIN EMBRYONIC LIBRARY

R: miRNA

: miRNA

>spu-miR-31

AGGCAAGATGTTGGCATAGCT

>Scaffold56613_109 strand:+ excise_beg:252614 excise_end:252723

TACAAACCAGTGGAGAAAAGGCAAGATGTTGGCATAGCTGTGATTTAAATATTAACCCAGCTGTGTCTTCATACTGCCATTTATTCACTTGGTTGT

((((.(((((((.(((.((((..(((..(((((((((((((.....)))).)))..)))))))))..)))..)))).))).)))))).)))))))

RRRRRRRRRRRRRRRRRRRR **********

>spu-miR-184

TGGACGGAGAACTGATAAGGGC

>Scaffold74947_52 strand:+ excise_beg:134114 excise_end:134232

TGTCTTTTATCTTTCCTTATCATTCTCTTGCCCGGCCGAATACTATGTTATTGAAATTGCGCTGGACGGAGAACTGATAAGGGCTTATATAAGACA

((((((.(((...(((((((((((((((.(.(((((((.................))).)))))).)))))))).))))))))).)...))).))))))

********** RRRRRRRRRRRRRRRRRRRRR

>spu-miR-2012

TAGTACTGGCATATGGACATTG

GGTTCTTGATAGTACTGGCATATGGACATTGTTGTAATAATCTTACTCTCAATGTTCATCTGTCCGTACTGCCAAGTGCC

(((.((((.(((((((.((((.((((((((((..((((.....))))...)))))))))).))))).))))))).))))).)))

RRRRRRRRRRRRRRRRRRRRR **********

>spu-miR-2009

TGAGTTGTCCCACAAAGAACACA

>Scaffold57328_15 strand:+ excise_beg:19762 excise_end:19872

TCCTCAGTAATCTTTGGTTTTTGTGGAACAGCTCCTGTGAATTATATAAGACCATGAGTTGTCCCACAAAGAACACAGATCGCTGTGA

((..(((..((((.((.(((((((((.(((((((.((.(..((....))..)))).)))))))).)))))))).)).)))))..)))))

********** RRRRRRRRRRRRRRRRRRRRR

```
>spu-miR-183
TATGGCACTATAGAATTCACTG
>Scaffold58778_183 strand:+ excise_beg:323022 excise_end:323135
TCTGCTCCTCGTCTATGGCACTATAGAATTCACTGTGTGTTGTGTTGCCATGGTAACAGCTGGTTCTACAGTGCCATTGCATGATGGGTAGA
(((((((.(((((.(((((((((.(((((((..((((.((..(((....)))..)))))))..)))))))).)))))))).).)))).)))))))
             RRRRRRRRRRRRRRRRRRRRR                        **********************

>spu-miR-1
TGGAATGTAAAGAAGTATGTAT
>Scaffold4906_62 strand:+ excise_beg:262910 excise_end:263018
TCCCATGCCATAACATACTTCTTTAGAATTCCATACTGAATCTCCTCAACTCTATGGAATGTAAAGAAGTATGTATTGAAACTGGGA
(((((...((..(((((((((((((..(((((((((.(((......))).....)))))))).)))))))))))))..)).....)))))
             **********************                   RRRRRRRRRRRRRRRRRRRRR

>spu-miR-182
TTTGGCAATTGATAGAATTCACACT
>Scaffold58778_188 strand:+ excise_beg:328065 excise_end:328176
TCCGCCTCTATTTTTGGCAATTGATAGAATTCACACTGTAACCTCATTATGTAATCAGTGGTTCTCTCATTGCCAAGCCTGGTGGCGTGATT
((((((.(((..((((((((((.((.(((((...(((((..((........))...)))))))))).)))))))))))..))).)))).))..
            RRRRRRRRRRRRRRRRRRRRRRRRR                     ************************

>spu-miR-2001
ATGTGACCGATATAATGGGCAT
>Scaffold41617_52 strand:+ excise_beg:105259 excise_end:105365
ATCACGTGATCTTGTCAATGTGACCGATATAATGGGCATGAATAATCATGATCAATGTCCAGTATAGCGGTGACAATGAGATGGTCATGTGGT
((((((((((.(.(((.(((.((((.((((.(((((((((.((....)).)).)))))))).)))).)))).)))).)))).)))))))))
                 RRRRRRRRRRRRRRRRRRRRR                  **********************

>spu-miR-92c
TATTGCACTCGTCCCGGCCTGC
>Scaffold40011_170 strand:+ excise_beg:295409 excise_end:295520
GAGGGCAGCAAGCTGGTCGTGAGGAGTTGCAATTTGTCCACATGATAATAATCATCATATTGCACTCGTCCCGGCCTGCCTGCTTGCCCTC
((((((((((.((.(((((.((.((((.(((((.((.....(((((....)))))))).)))))))))).)).)))))).)).)))).)))))))
              **********************                   RRRRRRRRRRRRRRRRRRRRR

>spu-miR-96
TTTGGCACTAGCACATTTTGC
>Scaffold58778_184 strand:+ excise_beg:324584 excise_end:324701
ATACTAGTCTGTTTGGCACTAGCACATTTTGCTGTACCTCGAAAGAGGCTCATTGTCGTGTGCAAGATATGTTTGTGCCAAGCCAGATTGG
...((((((((((((((((.((((.((((((((((..(((((....)))))..))).........)))))))).)))).)))))))).))))))))
            RRRRRRRRRRRRRRRRRRRRR                                **********************

>spu-miR-92b
TATTGCACTTGTCCCGGCCTGC
>Scaffold40011_162 strand:+ excise_beg:293726 excise_end:293839
TGGTGATTGATGGTCAAGTCGGACCGAGCGCAATGTTGTTCCTCTATTGAGGTTTTTCGAATATTGCACTTGTCCCGGCCTGCTCATCAACACCA
(((((.(((((((..((.((((((..((((.((((((((((..(((((....)))))).....).)))))))).)))).))))))..)).)))))))))))
               **********************                      RRRRRRRRRRRRRRRRRRRRR
```

```
>spu-miR-2007
TATTTCAGGCAGTATACTGGTAA
>Scaffold20011_6 strand:+ excise_beg:3791 excise_end:3903
CGGCATCATGCATTATTTCAGGCAGTATACTGGTAAAGGGTTTTATTTGCACCATTCTTACCTGTTGCTACCTGAAATTAATGCAAGATGT
..(((((.(((((((((((((((.((((.((.((((((..(((.........)))....)))))).)))))).)))))))).)))))))).)))))
              RRRRRRRRRRRRRRRRRRRRR                        **********************

>spu-miR-2008
ATCAGCCTCGCTGTCAATACG
CACTTTTTACCCCGTATTAACAATGTGGCTGATGAGGAATCTTATGAACCATCAGCCTCGCTGTCAATACGAGGTAAGGAGTG
((((((((((.((((((.(((.((.(((((((((.(...((....)).)))))))))).)).)))).)))))).))))))))))
         **********************                  RRRRRRRRRRRRRRRRRRRRR

>spu-miR-219
TGATTGTCCGAACGCAATTCTTG
>Scaffold4943_18 strand:- excise_beg:24301 excise_end:24406
TTTCCATGCGAGTGTTCTCCGTTGATTGTCCGAACGCAATTCTTGTTGAATTTCCTAGTCAAGAACTGTGTACGGACATCAGTGGTTGATACCGC
.......(((.((((((..(((((((((.((((((.((((((.((((((((((.......))).)))))).)))))).))))))))))))..)))))))))
                      RRRRRRRRRRRRRRRRRRRRR                **********************

>spu-miR-7
TGGAAGACTAGTGATTTTGTTGT
>Scaffold15487_10 strand:- excise_beg:9657 excise_end:9766
CACATACCTCTGGCCCTTTGGAAGACTAGTGATTTTGTTGTTGTGAATGTCAAAGAACTAACAAAAAATCACTGTCGGCCACTGGGGACAGAGGT
.....(((((((.((((.(((..(((.(((((((((.(((((((.............)).)))))))))))))))))..))))..)))).)))))))
                 RRRRRRRRRRRRRRRRRRRRR                      **********************

>spu-miR-242
TGCGTAGGCGTTGTGCACAGT
>Scaffold65146_255 strand:- excise_beg:99185 excise_end:99289
GTAACACCATGTCGGGCATTGCGTAGGCGTTGTGCACAGTGTTCTTTAAATTACAATTGTGCATCTCATCCACGTAAATCACGGCACTGGT
.....(((((((((....((((((.(..(..((((((((((.........)).)))))))))..)..).)))))).....)))))).))))
                   RRRRRRRRRRRRRRRRRRRR                  *********************

>spu-miR-2002
TGAATACATCTGCTGGTTTTTAT
>Scaffold34693_338 strand:+ excise_beg:947283 excise_end:947392
TTATTCATCAGACTGATAAGACCAACAGGTGTTATTCTCATGTGTGATGTACAGATCTGTGAATACATCTGCTGGTTTTTATTAGGCTCGATGA
....(((((((.((((((((((((.(((((((.((((.((.((.((.....)).)).)).)))))))))))).)))).)))))))))).)))))
                 **********************                   RRRRRRRRRRRRRRRRRRRRR

>spu-miR-9
TCTTTGGTTATCTAGCTGTATG
>Scaffold32321_166 strand:- excise_beg:474453 excise_end:474559
ACTGACTCTGCTTTTCTCTTTGGTTATCTAGCTGTATGAGTGATGTCAATGCATCATAAAGCTAGGTTACCAAAGATAAGTGCCTGTTGTCAC
..((((.(.((.(((.(((((((((.(((((((((......(((((((....)))))))..)))))))).)))))))).)))).))..)..)))))
                RRRRRRRRRRRRRRRRRRRRR                     **********************
```

APPENDIX H

WHOLE MOUNT IN SITU HYBRIDIZATION OF PKS IN EMBRYOS INJECTED WITH CONTROL MASO AND MASO TARGETING AGO2

BIBLIOGRAPHY

Alon, U., 2007. *An introduction to systems biology:design principles of biological circuits*, Chapman & Hall/CRC.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403-410.

Alvarez-Garcia, I. & Miska, E.A., 2005. MicroRNA functions in animal development and human disease. *Development (Cambridge, England)*, 132(21), pp.4653-4662.

Ambros, V., 2003. MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell*, 113(6), pp.673-676.

Ambros, V., 2004. The functions of animal microRNAs. *Nature*, 431(7006), pp.350-355.

Angerer, L M et al., 2001. Sea urchin goosecoid function links fate specification along the animal-vegetal and oral-aboral embryonic axes. *Development (Cambridge, England)*, 128(22), pp.4393-4404.

Anon, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696), pp.636-640.

Arnold, K. et al., 2005. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2), pp.195-201.

Ason, B. et al., 2006. Differences in vertebrate microRNA expression. *Proceedings of the National Academy of Sciences of the United States of America*, 103(39), pp.14385-14389.

van Bakel, H. et al., 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biology*, 8(5), p.e1000371.

Bannister, S.C. et al., 2009. Sexually Dimorphic MicroRNA Expression During Chicken Embryonic Gonadal Development. *Biology of Reproduction*, 81(1), pp.165-176.

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

Bartel, D.P., 2009. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2), pp.215-233.

Bartel, D.P. & Chen, C.-Z., 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature Reviews. Genetics*, 5(5), pp.396-400.

Baskerville, S. & Bartel, D.P., 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA (New York, N.Y.)*, 11(3), pp.241-247.

Benson, D.A. et al., 2008. GenBank. *Nucleic Acids Research*, 36(Database issue), pp.D25-30.

Benson, S., 1987. A lineage-specific gene encoding a major matrix protein of the sea urchin embryo spicule *1I. Authentication of the cloned gene and its developmental expression. *Developmental Biology*, 120(2), pp.499-506.

Benson, S.C., Benson, N.C. & Wilt, F., 1986. The organic matrix of the skeletal spicule of sea urchin embryos. *The Journal of Cell Biology*, 102(5), pp.1878-1886.

Bentwich, I. et al., 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics*, 37, pp.766-770.

Berezikov, E. et al., 2010. Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. *Nature Genetics*, 42(1), pp.6-9.

Bernstein, Emily et al., 2003. Dicer is essential for mouse development. *Nature Genetics*, 35(3), pp.215-217.

Blakaj, A. & Lin, H., 2008. Piecing Together the Mosaic of Early Mammalian Development through MicroRNAs. *Journal of Biological Chemistry*, 283(15), pp.9505-9508.

Borchert, G.M., Lanier, W. & Davidson, B.L., 2006. RNA polymerase III transcribes human microRNAs. *Nature Structural & Molecular Biology*, 13(12), pp.1097-1101.

Brennecke, J. et al., 2005. Principles of MicroRNA–Target Recognition. *PLoS Biology*, 3(3), p.e85.

Bushati, N. & Cohen, S.M., 2007. microRNA functions. *Annual Review of Cell and Developmental Biology*, 23, pp.175-205.

Calestani, C., Rast, J.P. & Davidson, Eric H, 2003. Isolation of pigment cell specific genes in the sea urchin embryo by differential macroarray screening. *Development (Cambridge, England)*, 130(19), pp.4587-4596.

Campo-Paysaa, F. et al., 2011. microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evolution & Development*, 13(1), pp.15-27.

Catlett, C. et al., 2007. TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications. In *HPC and Grids in Action*. Advances in Parallel Computing. Amsterdam: IOS Press, pp. 225-249.

Cerutti, H. & Casas-Mollano, J.A., 2006. On the origin and functions of RNA-mediated silencing: from protists to man. *Current Genetics*, 50(2), pp.81-99.

Chen, K. & Rajewsky, N., 2006. Natural selection on human microRNA binding sites inferred from SNP data. *Nature Genetics*, 38(12), pp.1452-1456.

Chendrimada, T.P. et al., 2005. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, 436(7051), pp.740-744.

Chi, S.W. et al., 2009. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254), pp.479-486.

Christodoulou, F. et al., 2010. Ancient animal microRNAs and the evolution of tissue identity. *Nature*, 463(7284), pp.1084-1088.

Cline, M.S. et al., 2007. Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, 2(10), pp.2366-2382.

Corcoran, D.L. et al., 2009. Features of Mammalian microRNA Promoters Emerge from Polymerase II Chromatin Immunoprecipitation Data C. K. Patil, ed. *PLoS ONE*, 4(4), p.e5279.

Davidson, Eric H, 2009. Network design principles from the sea urchin embryo. *Current Opinion in Genetics & Development*, 19(6), pp.535-540.

Davidson, Eric H et al., 2002. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Developmental Biology*, 246(1), pp.162-190.

Dunn, E.F. et al., 2007. Molecular paleoecology: using gene regulatory analysis to address the origins of complex life cycles in the late Precambrian. *Evolution & Development*, 9(1), pp.10-24.

Enright, A.J. et al., 2003. MicroRNA targets in Drosophila. *Genome Biology*, 5(1), p.R1.

Ettensohn, C.A., Wray, G.A. & Wessel, G.M., 2004. *Development of Sea Urchins, Ascidians, and Other Invertebrate Deuterostomes: Experimental Approaches.*, San Diego: Elsevier Academic Press.

Fine, S., Singer, Y. & Tishby, N., 1998. The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, 32(1), pp.41-62.

Friedländer, M.R. et al., 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology*, 26(4), pp.407-415.

Gajewski, M. et al., 2006. Comparative analysis of her genes during fish somitogenesis suggests a mouse/chick-like mode of oscillation in medaka. *Development Genes and Evolution*, 216(6), pp.315-332.

Gilbert, S., 2000. *Developmental Biology. 6th edition.*, Sunderland (MA): Sinauer Associates.

Giraldez, A.J. et al., 2005. MicroRNAs regulate brain morphogenesis in zebrafish. *Science (New York, N.Y.)*, 308(5723), pp.833-838.

Giraldez, A.J. et al., 2006. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science (New York, N.Y.)*, 312(5770), pp.75-79.

Goujon, M. et al., 2010. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research*, 38(Web Server issue), pp.W695-699.

Grad, Y. et al., 2003. Computational and experimental identification of C. elegans microRNAs. *Molecular Cell*, 11(5), pp.1253-1263.

Griffiths-Jones, S., 2006. miRBase: the microRNA sequence database. *Methods in Molecular Biology (Clifton, N.J.)*, 342, pp.129-138.

Grün, D. et al., 2005. microRNA target predictions across seven Drosophila species and comparison to mammalian targets. *PLoS Computational Biology*, 1(1), p.e13.

Hinman, V.F. & Davidson, Eric H, 2007. Evolutionary plasticity of developmental gene regulatory network architecture. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49), pp.19404-19409.

Hinman, V.F., Nguyen, A. & Davidson, Eric H, 2007. Caught in the evolutionary act: precise cis-regulatory basis of difference in the organization of gene networks of sea stars and sea urchins. *Developmental Biology*, 312(2), pp.584-595.

Hinman, V.F., Nguyen, A.T. & Davidson, Eric H, 2003. Expression and function of a starfish Otx ortholog, AmOtx: a conserved role for Otx proteins in endoderm development that predates divergence of the eleutherozoa. *Mechanisms of Development*, 120(10), pp.1165-1176.

Hinman, V.F. et al., 2003. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23), pp.13356-13361.

Hofacker, I.L., 2003. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13), pp.3429-3431.

Howardashby, M. et al., 2006. Gene families encoding transcription factors expressed in early development of Strongylocentrotus purpuratus. *Developmental Biology*, 300(1), pp.90-107.

Humphrey, W., Dalke, A. & Schulten, K., 1996. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1), pp.33-38, 27-28.

Hutvagner, G., 2001. A Cellular Function for the RNA-Interference Enzyme Dicer in the Maturation of the let-7 Small Temporal RNA. *Science*, 293(5531), pp.834-838.

Ivey, K.N. & Srivastava, D., 2010. MicroRNAs as Regulators of Differentiation and Cell Fate Decisions. *Cell Stem Cell*, 7(1), pp.36-41.

Kadri, S., Hinman, V. & Benos, P.V., 2009. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics*, 10(Suppl 1), p.S35.

Kadri, S., Hinman, V.F. & Benos, P.V., 2011. RNA Deep Sequencing Reveals Differential MicroRNA Expression during Development of Sea Urchin and Sea Star. *PloS One*, 6(12), p.e29217.

Kent, W.J. et al., 2002. The human genome browser at UCSC. *Genome Research*, 12(6), pp.996-1006.

Ketting, R F et al., 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes & Development*, 15(20), pp.2654-2659.

Khvorova, A., Reynolds, A. & Jayasena, S.D., 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2), pp.209-216.

Kiefer, F. et al., 2009. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*, 37(Database), p.D387-D392.

Kim, V.N., 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nature Reviews. Molecular Cell Biology*, 6(5), pp.376-385.

Kim, V.N., Han, J. & Siomi, M.C., 2009. Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology*, 10(2), pp.126-139.

Kim, Y.-K. & Kim, V.N., 2007. Processing of intronic microRNAs. *The EMBO Journal*, 26(3), pp.775-783.

Kloosterman, W.P. & Plasterk, R.H.A., 2006. The diverse functions of microRNAs in animal development and disease. *Developmental Cell*, 11(4), pp.441-450.

Kloosterman, W.P. et al., 2007. Targeted Inhibition of miRNA Maturation with Morpholinos Reveals a Role for miR-375 in Pancreatic Islet Development. *PLoS Biology*, 5(8), p.e203.

Kloosterman, W.P. et al., 2006. Cloning and expression of new microRNAs from zebrafish. *Nucleic Acids Research*, 34(9), pp.2558-2569.

Krek, A. et al., 2005. Combinatorial microRNA target predictions. *Nature Genetics*, 37(5), pp.495-500.

Lai, Eric C et al., 2003. Computational identification of Drosophila microRNA genes. *Genome Biology*, 4(7), p.R42.

Larkin, M.A. et al., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21), pp.2947-2948.

Lee, I. et al., 2009. New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. *Genome Research*, 19(7), pp.1175-1183.

Lee, R., 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5), pp.843-854.

Legendre, M., Lambert, A. & Gautheret, D., 2005. Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics (Oxford, England)*, 21(7), pp.841-845.

Lewis, B.P., Burge, C.B. & Bartel, D.P., 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1), pp.15-20.

Licatalosi, D.D. et al., 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221), pp.464-469.

Lim, L.P. et al., 2003. The microRNAs of Caenorhabditis elegans. *Genes & Development*, 17(8), pp.991-1008.

Lytle, J.R., Yario, T.A. & Steitz, J.A., 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proceedings of the National Academy of Sciences of the United States of America*, 104(23), pp.9667-9672.

Marson, A. et al., 2008. Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells. *Cell*, 134(3), pp.521-533.

Martello, G. et al., 2007. MicroRNA control of Nodal signalling. *Nature*, 449(7159), pp.183-188.

McClay, D. R., 2011. Evolutionary crossroads in developmental biology: sea urchins. *Development*, 138(13), pp.2639-2648.

Millar, A.A. & Waterhouse, P.M., 2005. Plant and animal microRNAs: similarities and differences. *Functional & Integrative Genomics*, 5(3), pp.129-135.

Miska, E.A. et al., 2007. Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. *PLoS Genetics*, 3(12), p.e215.

Murchison, E.P. et al., 2007. Critical roles for Dicer in the female germline. *Genes & Development*, 21(6), pp.682-693.

Nam, J.-W. et al., 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Research*, 33(11), pp.3570-3581.

Nelson, P.T. et al., 2007. A novel monoclonal antibody against human Argonaute proteins reveals unexpected characteristics of miRNAs in human blood cells. *RNA (New York, N.Y.)*, 13(10), pp.1787-1792.

Ohler, U. et al., 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA (New York, N.Y.)*, 10(9), pp.1309-1322.

Oliveri, P., Carrick, D.M. & Davidson, Eric H, 2002. A regulatory gene network that directs micromere specification in the sea urchin embryo. *Developmental Biology*, 246(1), pp.209-228.

Pasquinelli, A.E. et al., 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808), pp.86-89.

Peitsch, M.C., 1995. Protein Modeling by E-mail. *Bio/Technology*, 13(7), pp.658-660.

Peterson, R.E. & McClay, David R, 2003. Primary mesenchyme cell patterning during the early stages following ingression. *Developmental Biology*, 254(1), pp.68-78.

Pfeffer, S., Sewer, A., et al., 2005. Identification of microRNAs of the herpesvirus family. *Nature Methods*, 2(4), pp.269-276.

Prather, R.S. et al., 2009. Transcriptional, post-transcriptional and epigenetic control of porcine oocyte maturation and embryogenesis. *Society of Reproduction and Fertility Supplement*, 66, pp.165-176.

Preker, P. et al., 2008. RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. *Science*, 322(5909), pp.1851-1854.

Prochnik, S.E., Rokhsar, D.S. & Aboobaker, A.A., 2007. Evidence for a microRNA expansion in the bilaterian ancestor. *Development Genes and Evolution*, 217(1), pp.73-77.

Rajewsky, N., 2006. microRNA target predictions in animals. *Nature Genetics*, 38 Suppl, pp.S8-13.

Rehmsmeier, M., 2004. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10), pp.1507-1517.

Reinhart, B J et al., 2000. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772), pp.901-906.

Reinhart, Brenda J et al., 2002. MicroRNAs in plants. *Genes & Development*, 16(13), pp.1616-1626.

Romano, L.A. & Wray, G.A., 2003. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Development (Cambridge, England)*, 130(17), pp.4187-4199.

Ruby, J.G., Jan, C.H. & Bartel, D.P., 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149), pp.83-86.

Saetrom, P. et al., 2006. Conserved microRNA characteristics in mammals. *Oligonucleotides*, 16(2), pp.115-144.

Samanta, M.P. et al., 2006. The Transcriptome of the Sea Urchin Embryo. *Science*, 314(5801), pp.960-962.

Samollow, P.B., 2008. The opossum genome: insights and opportunities from an alternative mammal. *Genome Research*, 18(8), pp.1199-1215.

Sea Urchin Genome Sequencing Consortium et al., 2006. The Genome of the Sea Urchin Strongylocentrotus purpuratus. *Science*, 314(5801), pp.941-952.

Selbach, M. et al., 2008. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209), pp.58-63.

Sempere, L.F. et al., 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 306B(6), pp.575-588.

Sewer, A. et al., 2005. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, 6, p.267.

Shannon, P., 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11), pp.2498-2504.

Sherwood, D.R. & McClay, D R, 1997. Identification and localization of a sea urchin Notch homologue: insights into vegetal plate regionalization and Notch receptor regulation. *Development (Cambridge, England)*, 124(17), pp.3363-3374.

Song, J.L. & Wessel, G.M., 2007. Genes involved in the RNA interference pathway are differentially expressed during sea urchin development. *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, 236(11), pp.3180-3190.

Song, J.L. et al., 2011. Select microRNAs are essential for early development in the sea urchin. *Developmental Biology*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22155525 [Accessed January 3, 2012].

Su, Y.-H., 2009. Gene regulatory networks for ectoderm specification in sea urchin embryos. *Biochimica Et Biophysica Acta*, 1789(4), pp.261-267.

Suh, N. & Blelloch, R., 2011. Small RNAs in early mammalian development: from gametes to gastrulation. *Development (Cambridge, England)*, 138(9), pp.1653-1661.

Taft, R.J. et al., 2009. Tiny RNAs associated with transcription start sites in animals. *Nature Genetics*, 41(5), pp.572-578.

Wada, H. & Satoh, N., 1994. Phylogenetic relationships among extant classes of echinoderms, as inferred from sequences of 18S rDNA, coincide with relationships deduced from the fossil record. *Journal of Molecular Evolution*, 38(1), pp.41-49.

Wheeler, B.M. et al., 2009. The deep evolution of metazoan microRNAs. *Evolution & Development*, 11(1), pp.50-68.

Wienholds, E. & Plasterk, R.H.A., 2005. MicroRNA function in animal development. *FEBS Letters*, 579(26), pp.5911-5922.

Wienholds, E. et al., 2005. MicroRNA expression in zebrafish embryonic development. *Science (New York, N.Y.)*, 309(5732), pp.310-311.

Wienholds, E. et al., 2003. The microRNA-producing enzyme Dicer1 is essential for zebrafish development. *Nature Genetics*, 35(3), pp.217-218.

Wightman, B., Ha, I. & Ruvkun, G, 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5), pp.855-862.

Willmann, M.R. et al., 2011. MicroRNAs Regulate the Timing of Embryo Maturation in Arabidopsis. *PLANT PHYSIOLOGY*, 155(4), pp.1871-1884.

Wu, S.-Y., Ferkowicz, M. & McClay, David R, 2007. Ingression of primary mesenchyme cells of the sea urchin embryo: a precisely timed epithelial mesenchymal transition. *Birth Defects Research. Part C, Embryo Today: Reviews*, 81(4), pp.241-252.

Xu, P. et al., 2003. The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Current Biology: CB*, 13(9), pp.790-795.

Xue, C. et al., 2005. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6, p.310.

Yang, J.-S. et al., 2010. Widespread regulatory activity of vertebrate microRNA* species. *RNA*, 17(2), pp.312-326.

Yang, W.J., 2004. Dicer Is Required for Embryonic Angiogenesis during Mouse Development. *Journal of Biological Chemistry*, 280(10), pp.9330-9335.

Yu, Z. et al., 2007. Global analysis of microRNA target gene expression reveals that miRNA targets are lower expressed in mature mouse and Drosophila tissues than in the embryos. *Nucleic Acids Research*, 35(1), pp.152-164.

Zhang, L. et al., 2007. Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Molecular Cell*, 28(4), pp.598-613.

Zhao, X.-F. et al., 2008. Treatment with small interfering RNA affects the microRNA pathway and causes unspecific defects in zebrafish embryos. *FEBS Journal*, 275(9), pp.2177-2184.