MICRORNA DISCOVERY IN *BELGICA ANTARCTICA*: MICRORNA LOCI RELOCATION BY DUPLICATION ACROSS TAXA

By

KARL SWANSON

A thesis submitted to the

Graduate School-Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of Master of Science

Graduate Program in Biology

Written under the direction of

Andrey Grigoriev

And approved by:

_____

Dr. Andrey Grigoriev

_____

Dr. Eric Klein

_____

Dr. Nir Yakoby

Camden, New Jersey

May 2015

THESIS ABSTRACT

MICRORNA DISCOVERY IN *BELGICA ANTARCTICA*:

MICRORNA LOCI RELOCATION BY DUPLICATION ACROSS TAXA

by Karl Swanson

Thesis Director:
Dr. Andrey Grigoriev

Small non-coding RNAs are a diverse class of molecules with wide biological importance, including regulatory roles, implications for evolution and possible medical therapeutics. The advent of next generation sequencing technology and various computational tools has aided in increasing the throughput and methods of discovery for these molecules. In this thesis we utilize and expand upon the most current methodologies of computational discovery, sequencing analysis and visualization for non-coding RNA, particularly microRNA (miRNA), in the Antarctic midge, *Belgica antarctica* and *Drosophila melanogaster.* These methods and the unique properties of *B. antarctica's* genome lead to discoveries of evolutionary and functional importance, especially for a class of miRNA called mirtrons. We show that mirtrons within the *B. antarctica* can relocate to an alternative gene loci, or are lost from their host gene. This relocation and loss of mirtrons is based on computational discovery and predictions, but is supported and validated by other examples in literature covering a wide range of taxa. The data and results suggest a re-examination of the mechanisms that birth miRNA,

specifically in terms of evolutionary duplication events. Additionally we describe and expand upon a tool for the *in silico* visualization of small non-coding RNA sequencing data, Genome Navigator. This tool can be used interactively to visualize concepts generated from high-throughput DNA and RNA sequencing data. We applied the new functionalities of Genome Navigator to elucidate biogenesis properties of another class of small non-coding RNA, called tRNA-derived fragments (tRFs). These properties strikingly resemble the canonical biogenesis cleavage patterns of miRNA.

Acknowledgements

I would first and foremost like to thank my advisor, Dr. Grigoriev. In my undergraduate years, I was aware of the work his lab was performing and I aspired to one day be a part of that team. Needless to say, principal investigators have a difficult task finding the right approach and balance of micro-management and macro-management for each student. Thank you, Dr. Grigoriev's for always being available to guide my project, yet never overbearing and constantly understanding.

Dr. Grigoriev's lab not only guided my thesis, but also introduced me to my first group of friends within The Center for Computational and Integrative Biology. I am a student of The Biology Department, but this group also mentored me throughout my program. I would like to thank my lab members. Thank you, Ammar Naqvi, soon to be Dr. Naqvi, for sharing your enthusiasm of non-coding RNA and allowing me to work on your projects. Not only are you an inspiring lab mate, but you are also one of my greatest friends that I've met through my academic studies. I wish you the best for the future and hope you find success for the remainder of your research endeavors.

I would also like to thank Spyros Karaiskos. He is the newest member of the lab, but quite possibly one of the most capable. Spyros was an instant "classic" amongst the CCIB students and in a short time, also became one of my closest friends. I look forward to moving into the new apartment with you. I know you will have success with your studies, as you are one of the brightest people I've met.

I thank Sulbha Choudari for always keeping the lab in line. I would also like to congratulate you on your defense, Dr. Choudari. Thank you Joe Kawash and Dr. Sean Smith, a.k.a. The Dream Team, your work is truly impressive and your input for all of the

lab members' projects was invaluable. Overall, I couldn't have asked for better lab members.

I'd like to thank Dr. Nam for the opportunity to work in his lab. You taught me a lot, if not the most and molded me from the beginning of my graduate career. It was in your lab that I learned the intricacies and difficulties of experimental molecular biology. You also supplied me with the foundation of my skills for computational biology. I wish I had the capacity to repay you for showing me such priceless skill sets.

I'd like to thank the members of other labs as well, especially the lunch club, Steve, Harish, Nastassia, Aylin, Sean and Kyle. Thank you, Ruchi, Sruthi and Min. All of these friends mentioned were a great pleasure to hangout with in and out of school. Thank you members of the fly lab, Rob, Nicole, Matt and Vikrant, you always provided great ideas for my experimental work. A special thanks to the leader of the fly lab, Dr. Yakoby, who took on being a committee member for my thesis on such a short notice. You were also a great inspiration in my undergraduate years and your genetics class helped influence my decision to study molecular and computational biology.

Thank you, Dr. Klein, for also being a part of my committee from the beginning. Your class on molecular carcinogenesis was one of the most interesting courses I've ever taken. You have a unique ability to teach and discuss those complex concepts. Thank you Dr. Shain and Dr. Saidel, who were essential to me in joining the program. I will never forget your words of encouragement. Thank you for believing in me. Lastly, thank you, Kevin Abbey, Janet Caruso, Peter Fazzino and Karen Taylor. Without your diligent work, none of our work would be possible.

Table of Contents

Introduction

MicroRNA (miRNA) are molecules that have helped pioneer the world of non-coding RNA (ncRNA) and its functional context, especially in regards to small ncRNAs. The study of small ncRNAs biogenesis, role in cellular function and possible medical application have opened intriguing areas of research for ncRNA (Bartel, 2004). With the advent of high throughput sequencing, discovering these small molecules and abstracting the biological concepts has become both facilitated and increasingly complex. Depending on the data available for an organism, different pipelines of experimental and computational software become applicable.

There are several classes of small ncRNAs and each have their own characteristics that make identifying them and classifying them possible. Here we define miRNAs as ~22 nucleotide long RNA molecules, which are derived from DNA-encoded RNA precursors with a hairpin, or stem-loop, secondary structure (Bartel, 2004). Dicer processes the mature miRNA from the stem loop of the precursor (**Figure 1**). The mature miRNA recognize their targets, usually protein coding mRNA, via hybridization by base pairing. After pairing, the targets expression is down regulated by cleavage, or by inhibiting the translational machinery's ability to translate the message (**Figure 1**) (Bartel, 2004). These properties, as noted earlier, allow for the prediction and validation of miRNA.

Figure 1 – miRNA Biogenesis Figure
This is a cartoon representation of miRNA biogenesis as described in the text. Figure credit: "MiRNA" by Kelvin Song - Own work. Licensed under CC BY 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:MiRNA.svg#/media/File:MiRNA.svg

History of miRNA Discovery

Forward genetics played an important role in miRNA discovery. Many of the first miRNAs discovered were due to noticeable phenotypic differences in organisms, which could be attributed to a mutation in a miRNA (Berezikov et al., 2006). One such discovery was that of *lin-4* in a *Caenorhabditis elegans* mutant, which failed to repress the expression and accumulation of *lin-14,* resulting in a heterochronic mutation and the subsequent loss of vulva and cuticle structures in the nematode (Lee et al., 1993). Following this discovery another *Caenorhabditis elegans* miRNA mutant was found, *let-7.* This was an even more progressive discovery as the let-7 miRNA is conserved among several taxa.

While forward genetics proved a thorough method of discovery, not all mutant miRNAs cause a noticeable phenotypic change and the throughput was too low to keep up with sequencing technology trends and demands; for these reasons researchers sought to develop new methods of miRNA discovery, with higher throughput.

Foundation of Current MicroRNA Discovery Techniques

Discovery methods can be defined in two categories, experimental and computational. Experimental techniques often rely on size fractionated RNA libraries and northern blotting. There is now an impressive overlap between experimental and computational approaches using high throughput sequencing data and often times both

are used. For example, computational predictions may be made and then a particular candidate RNA of interest can be validated to exist via northern blotting. Another example and probably the most utilized method for experimental discovery is sequencing of size fractionated RNA libraries (Bartel, 2004; Berezikov et al., 2006). While several methods for size fractionated RNA libraries have been developed, they all follow the same general principle: *I. First, size fractionate an RNA sample with a gel and select for 20 – 25 nt length RNAs. II. Attach a 3' and 5' adapter to all selected RNAs. III. Reverse transcribe amplify the RNAs via PCR. IV. Clone the cDNA into a vector to create a library. V. Sequence the library* (Berezikov et al., 2006). After this library is sequenced, or alternative experimental processes are used, further computational pipelines can be run to support the RNA's status as a miRNA vs. some other form of small ncRNA, or degradation product.

After features of miRNA were recognized via cloning and with the explosion of genomic sequencing data, computational approaches to miRNA discovery became increasingly popular. The most common features of miRNA dictated how to approach discovery with genomic, or transcriptomic data. A predictor used for all computational approaches is a secondary structure called an inverted reverse compliment palindromic repeat, or concisely, fold-back structures and hairpins; based on the concept that miRNA precursor usually have this structure. The next most common approach is phylogenetic analysis for sequence and structure conservation; this separates random fold-back structures from potential meaningful ones. Other approaches include thermodynamic stability of secondary structures sequence complexity, patterns specific to miRNA and potential miRNA targets (Berezikov et al., 2006).

The pioneering computational discovery approaches are conservation-based algorithms; such as, MirScan, Snarloop, MIRcheck and miRSeeker. MirScan and Snarloop predicted several hundreds of miRNA in *C. elegans* through pure conservation (Grad et al., 2003; Lim et al., 2003a; Lim et al., 2003b). MiRSeeker differed from these algorithms in that it used conservation of patterns specific to miRNA and identified 48 miRNA in *Drosophila melanogaster* (Lai et al., 2003). MIRCheck is another conservation based approach, but miRNA pattern specific algorithm that was originally designed for plant miRNA (Jones-Rhoades and Bartel, 2004). MIRcheck has proved effective for animals as well (Wu et al., 2013).

Several other groups sought alternative methods than conservation to detect miRNA. It was shown that certain motifs in the sequences of 3' UTRs of mRNA had higher frequencies than expected. Many of these motifs corresponded to the seed regions of miRNA (Xie et al., 2005). The seed region is the portion of the mature miRNA largely responsible for target expression inhibition. The seed region binds via, reverse complimentarity to the target mRNA. Additionally, miRNA hairpins were also shown to have a lower thermodynamic folding energy than other RNA (Bonnet et al., 2004). Therefore, one can distinguish a miRNA hairpin from other hairpins with folding energy. RNAz is notable secondary structure conservation and thermodynamic approach to discovering non-coding RNAs, including miRNA (Washietl et al., 2005)

With all of this in mind the MIRCheck pipeline developed by Jones-Rhoades et al., is strikingly clever. It takes into account all of the previously discussed features of miRNA. While the MIRCheck algorithm is mainly based on sequence complexity and conserved patterns in miRNA, the whole pipeline used by Bartel's lab included the

canonical hairpin search, using einverted (Rice et al., 2000); as well as thermodynamic stability by RNAfold (Lorenz et al., 2011); later in the analysis it also uses pure conservation via patscan; and finally they introduced the method of miRNA target prediction to help determine the validity of candidate miRNAs. Some of these methods, such as thermodynamics and target prediction, were designed prior to the robust analyses of other researchers, like Xie et al. and Bonnet et al, respectively (Bonnet et al., 2004; Xie et al., 2005).

Despite the success that MIRcheck and other fixed descriptor models showed for predictions in organisms that the model was based on, the algorithms often suffer from extremely low sensitivity. For example, MIRcheck can have as low of sensitivity 0.82% in other organisms. Granted, these algorithms showed high specificity. MIRcheck's redeeming quality is that it has a low false positive discovery, or specificity, as low as 0.03% (Ballén-Taborda et al., 2013). Varying secondary structure and differences in sequence homology across taxa create the main challenges in detecting miRNA. These variations are generated from mutations and evolutionary events such as nascent duplications followed by single-nucleotide polymorphisms, or entire deletions of genomic loci (Cuperus et al., 2011; Li et al., 2009; Lu et al., 2008). The fixed descriptor algorithms could not find all of the miRNA they were searching for, however, they maintained high confidence about the ones they did find. To solve the issue of sensitivity from the predecessor fixed algorithms, we employed algorithms that use support vector machine learning (SVM). A general overview of the SVM approach is described in the **Appendix**. The specific SVM tool we employed is described in the **Materials and Methods**.

Shortened Genome and Intronic Sequences of *Belgica Antarctica*

Belgica antarctica (midge), commonly called the Antarctic midge (**Figure 2)** is a wingless midge dipteran of the family Chironimidae and order Diptera (Convey and Block, 1996; SUGG et al., 1983). It was first discovered off of the Antarctica Peninsula (Peckham, 1971). This midge can survive extreme environments including temperature, freezing, dessication, ultraviolet radiation, high velocity winds, osmotic pressure of both high and low salinity and high nitrogen environments due to penguin and elephant seal breeding grounds (Elnitsky et al., 2009; Lopez-Martinez et al., 2008; Teets and Denlinger, 2014). It has adaptations for this extreme environment such as winglessness, freeze tolerance, desiccation survival and high expression of heat shock proteins (Rinehart et al., 2006).

Figure 2 –Belgica antarctica, the wingless Antarctic midge.
Photo credit: Richard Lee.

The midge has had some molecular studies on the expression of a handful of genes including heat-shock proteins, antioxidant enzymes catalase and superoxide dismutase (Lopez-Martinez et al., 2008), some genes that respond to changes in hydration states (Lopez-Martinez et al., 2009; Teets et al., 2012) and an aquaporin (Goto et al., 2011). The ability to survive an extreme environment and the constant expression of normally stress-induced genes are not the only unusual property about the midge. The genome was sequenced and found to be the most compact insect genome to date and is

the first dipteran of Chironomidae to have a sequenced genome at 99 megabase pairs (Kelley et al., 2014).

In general, genome size is correlated to intron size and the amount of non-coding regions within the genome; logically it is also correlated to the amount of non-coding RNA transcripts (Vinogradov, 1999). The group who sequenced the genome of the midge has also shown there is a significant reduction of intron size and transposable elements as compared with other insects with larger genome sizes. The midge only has about 0.12% of the genome as transposable elements (TEs), which is little when compared with other insects like *Aedes aegypti* (47%), *Anopheles gambiae* (16%), *Culex quinquefasciatus* (29%) and *Drosophila melanogaster* (20%) (**Figure 3)** (Kelley et al., 2014). It has about also has reduced intron length, when compared with same insects (**Figure 3)** (Kelley et al., 2014). The reduction in genome size, TEs and intron length do not affect the protein coding gene content, however, as approximately 19.4% (just under 19 Mbp) of the genome is protein coding in *B. antarctica* and contains 97% of the core eukaryotic genes. The midge has a larger portion of the genomes coding for proteins, when compared with the other insects (**Figure 3**) (Kelley et al., 2014).

Figure 3.
a.) Shows the lengths of Genome Size, Intron, CDS and Transposable Elements in 5 species, *Ae. aegypti*, *C. quinquefasciatus, An. Gambiae, D. melanogaster and B. antarcica.* b.) Shows the median, mean and maximum intron lengths in the same species; both figures were produced by Kelley et al. (Kelley et al., 2014).

We hypothesize that the evolution and radically different genomic landscape of the midge, due to the extreme selective pressures, when compared with other known insects, should directly affect the type of miRNA one can predict within this organism.

This in turn may shed light directly into the interactive effects of evolution and the creation and loss of small ncRNAs. Herein, we utilize a pipeline to detect these miRNA. This pipeline uses some homology-based approaches, so we also needed a genome with a relatively close phylogenetic distance and a well-defined reference set of miRNA. *Drosophila melanogaster* displays a close phylogenetic distance with the midge (**Figure 4**) and has greater annotation of miRNAs than any other insect species in miRbase.org. For this reason we selected the *D. melanogaster* (fly) genome, release 5, from flybase.org (St Pierre et al.).

Figure 4 – Phylogenetic tree of 7 insect species
This tree shows the lineage of 7 insect species, two flies from Drosophila, *D. melanogaster* and *D. willistoni;* 4 mosquitos from Aedes and Anopheles, *A. aegypti, A. albocephalus, A. stephensi and A. gambiae;* and the midge *B. Antarctica*. It was constructed by phyloT, using information from the NCBI taxonomic database. This tree displays the close phylogeny between *B. Antarctica* and *D. melanogaster.* Despite the close phylogenetic relationship of the *B. antarctica* and the *D. melanogaster*, the midge has a severely reduced genome size.

Results

We ran the pipeline (**Materials and Methods**), on both the fly and the midge genomes (**Table 1)**. The pipeline identified miRNA candidates that wanted to investigate. The SVM RNAmicro filtering step lowers the sensitivity of detecting miRNA slightly by ~14%, but increases confidence of true positives. Therefore, we did the initial downstream analysis on the 1,133 midge hairpin candidates, prior to filtering with RNAmicro, which had homology to 154 fly reference miRNA. (**Table 1, rows 4 and 5)**.

Table 1 – Pipeline Results:
The categories on the left define the number of items detected within the pipeline. The column on the furthest right indicates the pipelines sensitivity to detect known miRNA in the fly. In order, for both the fly and midge we list for the fly and the midge respectively.
Row 1: the total number of hairpins detected with einverted 447,218 (fly) and 421,318 (midge)
Row 2: the number of secondary structures predicted with RNAfold 894,436 (fly) and 842,636 (midge)
Row 3: the hairpin loci which contain a miRNA sequence that have homology to a fly reference miRNA, found by BLAST, 1,928 (fly) 1,164 (midge)
Row 4: the unique hairpin loci from BLAST, which contain a miRNA candidate, that is homologous to a fly reference miRNA 1,590 and 1,133
Row 5: the number of fly reference miRNA found in hairpin candidate loci from BLAST 230/238 (fly candidate/fly reference) and 154/238 (midge candidate/fly reference)
Row 6: the number of reference mirtrons found in hairpin candidate loci from BLAST 114/125 (fly candidate/fly reference) and 78/125 (midge candidate/fly reference)
Row 7: total hairpins scored as precursor by RNAmicro 614 (fly) and 379 (midge)
Row 8: total unique hairpin loci scored as precursor by RNAmicro 563 (fly) and 350 (midge)
Row 9: total miRNA references found in candidates scored as precursor by RNAmicro 193/238 (fly candidate/fly reference) and 105/238 (midge candidate/fly reference)
Row 10: total mirtron references scored as precursor by RNAmicro 94/125 (fly candidate/fly reference) and 51/125 (midge candidate/fly reference).

The 1,133 midge and 1,590 hairpin candidates were categorized and grouped to their respective 154 fly reference miRNA as either a mirton, or intergenic/antisense exonic miRNA (**Table 2**). This illustrates that there is retention of homologous miRNA of all classes in the midge; however, we wanted to know if the miRNA in the midge

retained their classification according to their genomic context. That is, do midge candidate miRNA, which are homologous to fly mirtrons, still reside within an intron in the midge and do the midge miRNA candidates that are homologous to fly intergenic/antisense miRNA still reside in intergenic, or antisense exonic regions in the midge.

Table 2 – miRNA Candidate Classification
This table contains the final candidates from the pipelines divided into their categories. There is a common category, for mirtrons and intergenic/antisense miRNA. Common miRNAs are defined as those candidates found with the pipeline that have homology to a reference fly miRNA that are found in both *D. melanogaster* and *B. Antarctica.* Uncommon miRNAs are defined as those candidates found with the pipeline that have homology to a reference fly miRNA that are found in *D. melanogaster*, but not in *B. Antarctica*. Mirtrons and intergenic/antisense categories are defined as those miRNA, which have homology to the *D. mleanogaster* references and are based on the references genomic location in *D. melanogaster.*

We tested whether the midge candidate miRNA retain their class according to their genomic context by mapping all of the predicted hairpins to the introns of their respective species. We then selected hairpins that had candidates from the pipeline that passed RNAmicro to increase confidence in the candidates being true-positives. In the midge we found that there are 51 candidates in the hairpin dataset that have homology to fly reference mirtrons, only 21 of those 51 map to actual midge introns, while 30 of the 51 do not map to introns (**Table 3**), rather they map to intergenic loci.

Table 3 – Hairpin and Mirtron Mapping to Intron Analysis
This table lists results from the analysis in which we mapped hairpin candidates to the intronic sequences for both the fly and midge. In order we list: total hairpins; total hairpins mapping to introns; unique candidate hairpins mapping to introns; total reference mirtrons hosted in candidate hairpins; reference mirtrons hosted in candidate hairpins, which map to introns; reference mirtrons hosted in candidate hairpins, which do not map to introns.

We began investigating the cause of the relocated and the putatively missing mirtrons, especially the possibility of the mechanism being due to intron reduction. We did this by obtaining all of the genes in the fly which host a mirtron, within its introns and performing a reciprocal BLAST against all of the midge genes. We divided the reciprocal gene pairs and plotted their introns by the following four categories: *all-set*, all 114 genes hosting mirtrons in the fly detected by the pipeline, prior to filtering with RNAmicro and their reciprocal hits in the midge; *uncommon-set*, the homologous reciprocal gene sets corresponding to the 36 uncommon mirtrons not found in the midge from **Table 2**; *common-retained-set*, the homologous reciprocal gene sets corresponding to the 21 candidates with homology to a fly mirtron, which also exist in a midge intron from **Table 3**; *common-relocated-set*, the homologous reciprocal gene sets corresponding to the 30 candidates with homology to a fly mirtron, but do not exist in a midge intron. The average intron length reduction remains the same for all four datasets with a 4-fold decrease (**Figure 5**). If there was no midge homolog for the corresponding fly gene detected by reciprocal BLAST, we took the mirtron sequences from the fly, plus 500bp upstream and downstream and searched against the midge genome using tblastx. No missing gene homologs were detected with this approach.

Table 4 – Statistics for Intron Length Shortening Across Four Datasets
The first three rows show percentages of intron number fluctuation with regards to homologous pair of genes in the fly and the midge. The last row shows the global decrease of intron lengths for the four datasets of homologous genes. In each percentage there is a decrease in the midge for the number (row 1-3), or length of introns (row 4), for homologous pairs for the fly and midge. This table shows a consistent decrease in the midge of introns and intron length for all pairs of homologs.

Figure 5 – Intron length distributions for all candidate/reference subsets.
We choose the four datasets as described in the text, *all-set* (fly introns are black bars; midge introns are white hatched bars), *common-retained-set* (fly introns are orange bars; midge introns are blue bars), *common-relocated-set* (fly introns are yellow bars; midge introns are purple bars), and *uncommon-set* (fly introns are green bars; midge red bars). These sets illustrate the point that the reduction in intron length in the midge genome, when compared with the fly genome, is not a likely mechanism for retained, relocated, or missing miRNA candidates that have reference to a fly mirtron. The sets for both species, are plotted as intron length on the x-axis against a percentage of introns at that length on the y-axis. The bars corresponding to lengths of 500 or less are increased in the midge, while bars corresponding to lengths 500 or greater are increased in the fly. Also observed in **Table 4, row 4**.

Interestingly, after analyzing the fly and midge reciprocal BLAST hit genes of this set we found only 9 of the 21 candidates retained their intronic position within the homologous gene pairs in both species. Moreover, 8 of the 21 relocated to an intron of another gene in the midge, which do have homologs for the fly. Finally, 4 of the 21 did not have an identifiable reciprocal homolog in the fly. We categorized these as new classes called "*retained mirtrons*" and "*relocated mirtrons*," these relocated miRNA are different from the original *common-relocated-set* because they relocate to an intron of another gene, rather than an intergenic space; therefore "*relocated mirtrons"* retain their class (**Figure 6**).

*****Figure 6 – Examples of "retained *mirtons*" and "*relocated mirtrons*"**
Figure 6a shows an alignment of mir-11 in all Arthropoda species in miRbase. The first sequence in the alignment is the midge mirtron. This mirtron maps to introns of homologous fly and midge genes, the *common-retained-set*. In this alignment there are several examples of insertions into the hairpin loop region, as well as the more conserved region of the mature and star. The mature strand is boxed in blue, where there are no SNPs. The seed region is boxed in red, within the blue mature box, there are also no SNPs within the seed region, exhibiting high homology. **Figure 6b** shows an alignment of the midge mir-966 putative mature region against the fly mir-966, it has no other known arthropoda homologs, so a more basic alignment is shown. It maps to a different gene's intron in the midge than the fly homolog. In **Figure 6c** the red ad green boxes represent fly and midge homologous gene exons that surround homologous mirtrons, respectively. **Example 1** shows the mirtron is retained in both genes. In **Example 2** the mirtron is relocated to the intron of an alternative gene in the midge, which may, or may not have a homolog in the fly. mir-11 would

be categorized in Example 1 where the mirtron is retained in a homologous intron, whereas mir-966 would be categorized in Example 2, as it relocates to a different intron.

We examined whether a reduction of intron numbers in the midge influenced relocation of mirtrons. We did this by plotting the number of introns of reciprocal homolog genes in the fly and midge, which all host introns in the fly, for three different dataset (**Figure 7a-c and Table 4; rows 1 - 3)**. The datasets we used are three of the four previously mentioned, *common-retained-set*, *common-relocated-set* and *uncommon-set*. In all of the datasets the number of introns within the midge genes are reduced when compared to their *Drosophila* homolog regardless if a mirtron is retained, relocated, or lost.

Figure 7a – Intron Count for Homologous Genes Set 1
This figure shows a bar chart counting the introns of genes for the *common-retained-set*. The homologs in the midge may, or may not host the same mirtron. For the first 9 mirtrons, left to right, the mirtrons are conserved within the introns of homologous pairs for both species, which we dubbed "*retained mirtrons"*. The remaining mirtrons past the first nine, left to right, are genes in which the mirtron relocates, or "*relocated-mirtrons*.*"* The red (midge) and blue (fly) bars signify the homologs where the mirtron exists in the midge and not in the fly and the yellow and green are the homologs where the mirtron exists in the fly, but not the midge. For the genes that host mir-274 in the midge there is no homolog detected for the fly, which is why there is no blue bar. For the gene that hosts mir-987 in the midge there is no homolog detected for the fly which is why there is no blue bar. For the gene that hosts mir-987 in the fly, the homolog in the midge, where the mirtron is missing, has no introns, which is why there is no green bar. The homolog's intron count for both pairs of genes is plotted as blue and yellow for fly and red and green for midge. There is a reduction in intron count in the midge when compared with the fly, as seen by the horizontal mean lines in blue, red, yellow and green.

Figure 7b – Intron Count for Homologous Genes Set 2
This figure shows a bar chart counting the introns of homologous genes in the *common-relocated-set*. The blue bars, fly, and the red bars, midge, show the intron counts. There is a reduction in intron count in midge when compared with the fly, as seen by the mean lines in blue and red.

Figure 7c – Intron Count for Homologous Genes Set 3
This figure shows a bar chart counting the introns of homologous genes for the *uncommon-set*. The blue bars, fly, and the red bars, midge, show the intron counts. There is a reduction in intron count in midge when compared with the fly, as seen by the mean lines in blue and red.

Discussion

Discussion of Results

The results show a likely retention of all miRNA classes in the midge (**Table 1**). Some of these miRNA from the fly to the midge may change class based on genomic context and location, or they appear to be missing. This is likely due to evolutionary mutation events, such as duplication (**Table 3**). Upon comparison to the fly, it seems plausible that the drastic change in genomic makeup of the midge is responsible for the miRNA that are missing, or relocated, in the midge genome. A reduction of mirtrons due to loss of host genes seems unlikely, since despite the shortened genome, the midge retains a majority of protein coding genes (Kelley et al., 2014) and in general once a miRNA is gained, it is rarely lost (Peterson et al., 2009).

To examine if the reduction in intron length effects whether or not a mirtron exists in the midge, we plotted the distributions of midge vs fly intron length for four datasets (**Figure 5**). All four dataset had a reduction of introns relative from the fly to the midge. This suggests that intron lengths of homologous genes are reduced in the midge regardless if a fly homolog mirtron in the midge is retained, relocated, or missing in the midge. Additionally, we were confident that some midge mirtrons are truly relocated, from the tblastx of mirtron sequences plus the 500 upstream and downstream nucleotides. No missing midge homologs were detected with this alternative approach, suggesting that these genes are truly missing, or at least highly diverged and therefore the mirtrons are truly relocated from a gene to an intergenic space, relative from the fly to the midge.

The 30 out of the 51 candidates from the *common-relocated-set* in the midge that have homology to fly mirtrons and are relocated to intergenic regions, exemplify that miRNA may change from class to class across different species, however, we wanted to see if there were any further discrepancies even within miRNA that maintain their class across these two species. We expect the 30 candidates from the *common-relocated-set* would change their class from species to species, but we wanted to see if the 21 candidates from the *common-retained-set* have other changes between the fly and the midge. We shifted focus from all midge candidates with homology to fly mirtrons to just the *common-retained set.* Only 9 of the 21 mirtrons are retained within homolog pairs for the fly and the midge and 8 of the 21 mirtrons are found in an alternative gene in the midge, which may or may not have a homolog in the fly. This suggests mirtrons may also relocate from gene to gene (**Figure 6**).

Since, the midge also has a reduction in the number of introns that reside within exons of protein coding genes that host mirtrons, it is tempting to hypothesize that the reduction of intron numbers plays a role in the loss of mirtrons in one set of homologous genes and the appearance, or relocation to others. In all of the datasets the number of introns within the midge genes are reduced when compared to their *Drosophila* homolog regardless if a mirtron is retained, relocated, or lost. This suggests that neither intron loss, nor length reduction, is the primary mechanism of mirtron loss or relocation. An alternative hypothesis is that of nascent gene duplication, followed by loss of sequence homology in the original, or duplicated gene, via mutations, and/or entire gene deletions, over the course of evolution. We assume this hypothesis because we are “missing”

several reference miRNA that likely diverged homologs and undetected due to stringent filtering criteria.

Since these discoveries and assumptions hold implications for the evolution and function of miRNAs, we wanted to confirm that similar trends could occur across different species. One group also recently demonstrated that in five platyhelminth species, mirtrons are able to relocate their loci (Jin et al., 2013). They noted that approximately ~41% of 22 known miRNA loci have transcriptional direction conversion and differing genomic loci, relative to other known protein coding genes, across the five species. One specific example of these relocations across the platyhelminth species, mir-2a, was also found in the dataset *common-relocated-set.* It was found within an intergenic space in the midge, but as a mirtron within the fly. This supports the hypothesis that mirtrons are often relocated to different genes, or different intergenic loci, however, it does not shed light into the possible mechanisms.

In another study of *Aedes aegypti* and *Anophele stephensi* the authors found a similar example of what they thought to be missing mirtrons mir-304 and mir-306. In the Anopheles small RNA-seq data they found mir-304 clustered with mir-283 and mir-12, and mir-306 clustered with mir-9b and mir-79, as they are in *Drosophila*; however, when they compared this cluster to Aedes clusters, mir-304 and mir-306 were missing from their respective clusters in Aedes (Mead and Tu, 2008).

The authors performed a follow-up study on Aedes, which showed a novel miRNA in both Aedes and Anopheles is actually a diverged homolog to mir-304 in Drosophila. This homolog is similar to Drosophila mir-304's reverse complement strand, but so different that it had to be classified as a new miRNA, mir-1889 in Anopheles and

Aedes(Li et al., 2009). They also investigated what they previously thought was the missing mir-306 miRNA from its cluster in Aedes. With a closer analysis of the cluster assembly they found mir-306 in Aedes in the same order of its cluster in Drosophila and Anopheles, however, its sequence had two mismatches to the other insect mir-306 **(Figure 8)**.

In the midge common mirtron dataset we also detect these two hairpin miRNA clusters. In the midge, 1 out of the 3 miRNA, for each cluster, is completely missing. Different miRNA are missing in the midge than were missing in *Aedes* and *Anopheles*. In the mir-306 cluster, mir-79 is conserved and mir-9b is missing. In the midge at the position where mir-9b is located for the fly and mosquitoes, there is a hit that is more similar to mir-9c, or mir-9a. Like the mir-306 cluster from the literature on Aedes, we are missing mir-306 completely as well **(Figure 8)**. In the mir-304 clusters from the midge datasets, mir-304 is not missing, however, mir-12 is missing. It is conserved, but with two mismatches to Drosophila and mir-283 is present with 100 percent identity to Drosophila (**Figure 8)**. The mirtrons that are retained in these clusters also have a conserved order when compared with the 3 other insect species. In both clusters there are hairpins, which hit in the region where the "missing" mirtrons should be. These are perhaps the diverged mirtron precursor hairpins.

They hypothesize that these differences are due to the then 17 known cases where pre-miRNA are duplicated, so one hairpin may produce the same or similar sequences as observed in the cases of mir-304 and mir-306 clustered hairpins. These duplications may have also been inverted, however, they acknowledge they have no direct evidence to support this hypothesis(Li et al., 2009). The paralog miRNA, which are likely products of

duplication, are usually annotated as miR-X-1 and miR-X-2, or miR-Xa and miR-Xb. There are several examples of these in Drosophila.

Figure 8 – "Missing" mirtrons from 4 insect species
This figure provides a cartoon representation of the mir-306 (**Figure 8a)** and mir-304 (**Figure 8b**) clusters of mirtrons across 4 different insect species, which likely result as a duplication and the divergence of sequences by mutation: *A. gambiae, A. aegypti, D. melanogaster and B. antarctica.*In the cluster that harbors mir-306 in *A. aegypti,* mir-306 was originally believed to be missing. It was later discovered to exist, but with 2 mismatches to the mature miRNA of *D. mel* and *A. gambiae*(Li et al., 2009; Mead and Tu, 2008). In the midge datasets mir-306 is also not detected, which means it is likely mutated to an undetectable degree, or missing altogether. In the second cluster that harbors mir-304, a similar scenario happened to mir-304, which was believed to be missing from *A. gambiae* and *A. aegypti*, but was later found as a distant homolog which was renamed to mir-1889(Li et al., 2009; Mead and Tu, 2008). In the midge dataset mir-304 exists in its cluster, but with 2 mutations, we are, however, missing mir-12, which may be a similar circumstance as described with mir-306.

We also speculate that duplication has a much larger role for the generation of miRNA than originally expected. It is a common mechanism known in plants, but originally it was thought, according to one notable study that only 1.7% of the miRNA in Drosophila have a paralog within the Drosophila genome(Lu et al., 2008). The authors concluded that most miRNAs are birthed from non-miRNA sequences, which accumulate enough mutations to create a novel miRNA, rather than duplication and subsequent divergence. We found contradictory results to this claim. Out of the 614 candidates at unique loci that passed all filtering in the pipeline, there were 193 unique miRNA from the Drosophila reference set. Out of those 186 reference miRNA 103 had alternative hairpin loci, on different chromosomes that passed the pipeline as well. This suggests that ~50% of candidates have a paralog on a different chromosome, which is much higher than the 1.7% estimate from 2008. This birth of new miRNA would be more consistent

with the idea of duplication. These duplicate miRNA either have redundant function as their paralog, or they are mutated resulting in a new alternative function.

The discoveries and data supporting the hypotheses herein and from the discussed literature suggest several properties of miRNA and its various classes: *I.) Mirtrons are often preferentially retained as homologs across taxa, despite severe differences in genomic context and these mirtron homologs may retain their location/classification as a mirtron, or change location across taxa to become intergenic, or antisense miRNA. II.) if the homolog of a mirtron still exists as a mirtron in the corresponding species, it may actually exist in a different gene's intron entirely. III.) if a miRNA's, especially a mirtron's, sequence and genomic location are conserved, then the pathways this miRNA plays a part in are of great importance.*

These properties support the idea that the various genomic locations of homologous miRNAs across taxa are effected by the evolution of organisms and vice versa. The transcriptional regulatory elements that control miRNA may also be removed in *"relocated-mirtrons,"* since the mirtron is transcribed with its host gene. This also implies that these miRNA may, or may not, have a different role in different gene regulatory networks of another organism, despite having very similar sequence composition. The sequences of the retained and relocated mirtrons remain highly conserved over large evolutionary distances. For example, within a retained mirtron, mir-11, there appears to be some insertions in various species (**Figure 6a**). It is commonly know that the most conserved portion of the stem-loop structure, across evolutionary distant species, is the stem composed of the seed, mature strand and star strand region.

Due to random mutation and evolutionary events these regions are freckled with single-nucleotide-polymorphisms, but are mostly conserved

In the datasets we also did a gene ontology (GO) analysis using GO terms from flybase (**Supplementary Table 1**). We found that there are about 13 genes that host mirtrons in the fly which play a role in ATP and GTP binding pathway. This was interesting because in cold-adapted species this pathway is of great importance and ATP is generally up-regulated (Parry and Shain, 2011). Interestingly the most instances of ATP binding and regulatory GO terms were associated with the *uncommon-set* group, which could indicate a need for less ATP pathway repression. Another interesting finding from the GO analysis was the association of host genes with several other pathways, which are known to be involved with cold tolerance, such as ion binding (Koštál et al., 2007), diacylglycerol (Moellering et al., 2010), heme binding (Yang and Brill, 1991) and fatty acid chain synthesis (Finegold, 1986).

Future Directions

The results hold many implications for the evolution and function of miRNA. Validating and examining them further would be an advisable next step. The other examples in literature of miRNA evolution, duplication and divergence, while supportive, do not fully validate our findings, so we propose the following further analyses. First and foremost next generation RNA-sequencing data could be obtained to validate the expression of the predicted midge miRNA. These could be mapped back to the genome to show the true locations of these miRNA. If the small RNA-sequencing reads map to

locations of our predicted hairpins, this will validate their expression. As noted, we detected several possible miRNA paralogs in the midge and the fly. If these reads map to multiple locations, it would support the hypothesis that duplication and paralogous miRNA have a much higher importance than their previously reported.

After validating the expression of miRNA it would be interesting to see the outcomes of miRNA knockdown experiments. In the midge and fly the validated miRNAs can be knocked down via antisense morpholinos. We can see if there is a difference in the functionality of miRNA across both species for homologous miRNA by observing any resulting phenotypic changes. Additionally, the miRNA-mRNA target interactions for each species can be predicted and validated and then analysis of the expression of mRNA can be examined with qPCR, and/or full transcriptomic RNA-sequencing data. In conjunction with miRNA expression data, it would be interesting to see the temporal, or spatial differences in miRNA expression for homologs, of the *common-relocated-set*. Since these are relocated across species, there is a chance that the regulatory elements that control the homologous miRNAs expression have also changed. We could predict and validate miRNA regulatory elements, potential transcription factor, and their binding sites. Then knockdowns of transcription factors and mutating cis-regulatory binding sites for transcription factors can be generated and analyzed by via knockdown.

Finally, if the formerly discussed future experiments validate the findings of "relocated" miRNA, this pipeline and the future experiments can be done across different species from a wide-variety of taxa. First, we would compare with several other insects. We could predict and validate miRNA; find miRNAs that "relocate" loci; and finally

examine their temporal and spatial expression, as well as functional context, within the different species. This would effectively show that miRNA relocate across various species, due to duplication and other possible mechanisms, which in turn directly affects their expression and therefore their biological functionality.

Materials and Methods

Generating Candidate Precursors

We sought to utilize, develop and apply existing computational miRNA discovery techniques to discover miRNA in *Belgica antarctica.* An overview of the pipeline can be found in **Figure 9**.

Figure 9 – This is a flowchart for the pipeline overview we used to detect miRNA in *Belgica antarctica* and *Drosophila melanogaster.*
From top down the information flowed as follows: genomes of two species are masked. One species genome and annotated miRNAs are used as a reference for predicted miRNA and computing the sensitivity of the pipeline. These masked genomes are then run through a program called Einverted to detect imperfect inverted palindromic sequences, or hairpins for brevity. The hairpins are "folded" with RNAfold, which calculates the mininum free energy (mfe) for all possible structures in the hairpin. After the mfe is calculated for each hairpin's forward and reverse strand these hairpins are checked against the reference set. This is done finding alignments between reference miRNAs and the candidate hairpin set via BLAST with parameters as specified in the text. The hairpins that pass BLAST and filtering by the said parameters are entered into ClustalW to be aligned with all known Arthropoda miRNA homologs of the same miRNA family, from miRBase. Those with alignments of at least one homolog are then analyzed by RNAmicro to score its probability as a miRNA.

We started by acquiring the genomes of two organisms one of interest, *Belgica antarctica* (midge)(Kelley et al., 2014) and the *Drosophila melanogaster* (fly) genome from the flybase.org as for a homology driven reference. Since miRNA are not often found in repeat rich regions of a genome, we mask both of the genomes for repeat regions. We used RepeatMasker (Smit, A.F.A. and Green, P., http://www.repeatmasker.org/) with default parameters. RepeatMasker masks DNA input sequences by replacing simple repeats, tandem repeats, segmental duplications and interspersed repeat sequences with a string of the letter N.

After masking the genomes, the second step requires a program from the EMBOSS suite called einverted to find imperfect palindromes in the genomes using a 1000 bp sliding window. Einverted searches for a reverse complement sequence upstream of the query sequence with the following parameters: threshold = 30, match score = 3, mismatch score = -3, gap penalty = 40, and maximum repeat length = 240. This effectively finds all potential hairpin structures in the masked genome. After identifying loci with an inverted reverse complement palindromic repeat, the loci of the inverted repeats have 10 nucleotides added to both the 5' and 3' end. As noted earlier, miRNA precursors have a lower folding energy than most other non-coding RNAs, so in the third step, we use RNAfold to predict minimum folding energy secondary structure for each palindrome data set (Hofacker et al., 1994).

Homology of miRNA Approach

After generating all of the possible candidate precursors, with the parameters set in Einverted and RNAfold, as described above, we sought to filter out hairpins that displayed low homology to the known fly precursors. We did this by using the NCBI BLAST+ command line utility(Camacho et al., 2009). We searched against all of the generated hairpins for the mature miRNA sequences as a query, since the mature miRNA have the higher sequence conservation across taxa, than do their precursor hairpins (Warthmann et al., 2008). We used the blastn functionality, with parameters of 90 percent identity and a word size of 7. After BLAST we parsed and filtered the data for a minimum alignment of 20 nucleotides and a maximum mismatch of 2 nucleotides,

between the hairpin and the queries. The word size parameter, generally set at 11, is used to increase the sensitivity of BLAST's algorithm, at the sacrifice of increasing the false positive discovery rate. We decreased it to 7 because miRNA detection requires a very delicate balance between sensitivity and specificity. The 90 percent identity parameter allows us to filter out sequences with an alignment of 20 or greater, with two, or more, mismatches.

miRNA Candidate Scoring by RNAmicro a Support Vector Machine Approach for Sequence Descriptors

After generating the initial hairpin candidates we align them with ClustalW (Thomopson et al., 1994) to phylogenetically similar species, specifically, all Arthropoda miRNA hairpins from miRbase and run the aligned hairpins through an SVM that ranks, scores and classifies hairpins as various non-coding RNAs. This yields the final dataset in the pipeline for predicitions. We selected RNAmicro(Hertel and Stadler, 2006) as the SVM algorithm to score the hairpins from the homology search. Briefly, this algorithm was trained on all know metazoan miRNAs at the time. It has sensitivity as high as 90% and specificity of 99%. It uses a window with adjustable incremental steps and analyzes 12 descriptors in total from each of the following properties miRNA, *lengths of the stem and hairpin loop regions* (2 descriptors), *sequence composition* (1 descriptor G+C content), *sequence conservation* (4 entropy descriptors), *thermodynamic stability* (4 descriptors, the average z of the energy z-scores, the folding energy, adjusted minimum free energy and minimum free energy index) and *structural conservation* (1 descriptor).

Each candidate hairpin precursor is scored with a probability according to the SVM's training for all 12 descriptors(Hertel and Stadler, 2006).

Appendix: Support Vector Machine Learning for Descriptors

There is a delicate balance between sensitivity and specificity for the fixed model algorithms, which needed to be solved. Ideally, an approach to miRNA discovery should have high sensitivity and low false positive discovery, or high specificity. With the advent of machine learning and support vector machine learning (SVM), the balance for high sensitivity and high specificity was achieved. Machine learning methods first started gaining popularity in miRNA discovery in 2005 with such algorithms as ProMiR(Nam et al., 2005), a Hidden Markov Model (HHM) machine learning approach, and SVM approach by Pfeffer et al. 2005(Pfeffer et al., 2005). Most current machine learning methods have above an 80% sensitivity and specificity(Yousef et al., 2009).

Vladimir N. Vapnik and Alexey Ya Chervonenkis originally developed SVM in 1963 and the algorithm approach was modified from 1993 through 1995 to its current flavor, by Vapnik, Boser, Guyon and Cortes (Boser et al., 1992; Cortes and Vapnik, 1995). SVM, a specific machine learning approach, and other machine learning methods essentially allow a computer to find the best statistical fit for a selected number of descriptors. These descriptors attempt to describe the ideal properties for a particular subject, which has the properties of the descriptors. The machine learns the best fit by using positive and negative training sets. More explicitly, all of the numbers for each feature, or descriptor of a miRNA, are combined into a single vector in an n-dimensional space. The algorithm compares all positive to all negative vectors for each class and finds a ‘hyperplane’ that separates the classes. The vectors which lie closest together, but on

opposite sides of the hyperplane, 'support' the hyperplane, hence support vector machine learning (Yousef et al., 2009). A cartoon representation can be seen in **Figure 13.** This differs from the non-dynamic programs like MIRcheck, in that MIRcheck's statistical fit for descriptors is fixed and based on a fixed set of observations. That is, if a new dataset, or subject has the same descriptors, but the best statistical fit for the descriptors changes, the algorithm will not work.

Figure 10 – Support Vector Machine Learning
This is a cartoon representation of a dataset that would be analyzed using SVM. The arrow on top points to the "hyperplane" which separates the data point vectors into categories. The vectors closest to the hyperplane fit, or "support," the hyper plane.

References

Ballén-Taborda, C., Plata, G., Ayling, S., Rodríguez-Zapata, F., Becerra Lopez-Lavalle, L.A., Duitama, J., Tohme, J., 2013. Identification of cassava MicroRNAs under abiotic stress. International journal of genomics 2013.

Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281-297.

Berezikov, E., Cuppen, E., Plasterk, R.H.A., 2006. Approaches to microRNA discovery. Nature Genetics 38, S2-S7.

Bonnet, E., Wuyts, J., Van de Peer, Y., Rouzé, P., 2004. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 20, 2911-2917.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, Proceedings of the fifth annual workshop on Computational learning theory. ACM, pp. 144-152.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC bioinformatics 10, 421.
Convey, P., Block, W., 1996. Antarctic Diptera: ecology, physiology and distribution. European Journal of Entomology 93, 1-14.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273-297.
Cuperus, J.T., Fahlgren, N., Carrington, J.C., 2011. Evolution and functional diversification of MIRNA genes. The Plant Cell Online 23, 431-442.

Elnitsky, M.A., Benoit, J.B., Lopez-Martinez, G., Denlinger, D.L., Lee, R.E., 2009. Osmoregulation and salinity tolerance in the Antarctic midge, Belgica antarctica: seawater exposure confers enhanced tolerance to freezing and dehydration. Journal of Experimental Biology 212, 2864-2871.

Finegold, L., 1986. Molecular aspects of adaptation to extreme cold environments. Advances in space research 6, 257-264.

Goto, S.G., Philip, B.N., Teets, N.M., Kawarasaki, Y., Lee, R.E., Denlinger, D.L., 2011. Functional characterization of an aquaporin in the Antarctic midge Belgica antarctica. Journal of insect physiology 57, 1106-1114.

Grad, Y., Aach, J., Hayes, G.D., Reinhart, B.J., Church, G.M., Ruvkun, G., Kim, J., 2003. Computational and experimental identification of C. elegans microRNAs. Mol Cell 11, 1253-1263.

Hertel, J., Stadler, P.F., 2006. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics 22, e197-e202.
Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie/Chemical Monthly 125, 167-188.

Jin, X., Lu, L., Su, H., Lou, Z., Wang, F., Zheng, Y., Xu, G.T., 2013. Comparative analysis of known miRNAs across platyhelminths. FEBS Journal 280, 3944-3951.
Jones-Rhoades, M.W., Bartel, D.P., 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell 14, 787-799.

Kelley, J.L., Peyton, J.T., Fiston-Lavier, A.S., Teets, N.M., Yee, M.C., Johnston, J.S., Bustamante, C.D., Lee, R.E., Denlinger, D.L., 2014. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. Nature communications 5, 4611.

Koštál, V., Renault, D., Mehrabianova, A., Bastl, J., 2007. Insect cold tolerance and repair of chill-injury at fluctuating thermal regimes: role of ion homeostasis. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology 147, 231-238.

Lai, E.C., Tomancak, P., Williams, R.W., Rubin, G.M., 2003. Computational identification of Drosophila microRNA genes. Genome biology 4, R42.

Lee, R.C., Feinbaum, R.L., Ambros, V., 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843-854.

Li, S., Mead, E.A., Liang, S., Tu, Z., 2009. Direct sequencing and expression analysis of a large number of miRNAs in Aedes aegypti and a multi-species survey of novel mosquito miRNAs. BMC genomics 10, 581.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., Bartel, D.P., 2003a. Vertebrate microRNA genes. Science (New York, N.Y.) 299, 1540.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., Bartel, D.P., 2003b. The microRNAs of Caenorhabditis elegans. Genes & development 17, 991-1008.

Lopez-Martinez, G., Benoit, J.B., Rinehart, J.P., Elnitsky, M.A., Lee Jr, R.E., Denlinger, D.L., 2009. Dehydration, rehydration, and overhydration alter patterns of gene expression in the Antarctic midge, Belgica antarctica. Journal of Comparative Physiology B 179, 481-491.

Lopez-Martinez, G., Elnitsky, M.A., Benoit, J.B., Lee, R.E., Denlinger, D.L., 2008. High resistance to oxidative damage in the Antarctic midge Belgica antarctica, and

developmentally linked expression of genes encoding superoxide dismutase, catalase and heat shock proteins. Insect biochemistry and molecular biology 38, 796-804.

Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L., 2011. ViennaRNA Package 2.0. Algorithms for Molecular Biology : AMB 6, 26-26.

Lu, J., Shen, Y., Wu, Q., Kumar, S., He, B., Shi, S., Carthew, R.W., Wang, S.M., Wu, C.-I., 2008. The birth and death of microRNA genes in Drosophila. Nature genetics 40, 351-355.

Mead, E.A., Tu, Z., 2008. Cloning, characterization, and expression of microRNAs from the Asian malaria mosquito, Anopheles stephensi. BMC genomics 9, 244.

Moellering, E.R., Muthan, B., Benning, C., 2010. Freezing tolerance in plants requires lipid remodeling at the outer chloroplast membrane. Science (New York, N.Y.) 330, 226-228.

Nam, J.-W., Shin, K.-R., Han, J., Lee, Y., Kim, V.N., Zhang, B.-T., 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic acids research 33, 3570-3581.

Parry, B.R., Shain, D.H., 2011. Manipulations of AMP metabolic genes increase growth rate and cold tolerance in Escherichia coli: implications for psychrophilic evolution. Molecular biology and evolution 28, 2139-2145.

Peckham, V., 1971. Notes on the chironomid midge Belgica antarctica Jacobs at Anvers Island in the maritime Antarctic. Pac Insects Monogr 25, 145-166.

Peterson, K.J., Dietrich, M.R., McPeek, M.A., 2009. MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. Bioessays 31, 736-747.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grässer, F.A., van Dyk, L.F., Ho, C.K., Shuman, S., Chien, M., 2005. Identification of microRNAs of the herpesvirus family. Nature methods 2, 269-276.

Rice, P., Longden, L., Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16, 276-277.

Rinehart, J.P., Hayward, S.A., Elnitsky, M.A., Sandro, L.H., Lee, R.E., Denlinger, D.L., 2006. Continuous up-regulation of heat shock proteins in larvae, but not adults, of a polar insect. Proceedings of the National Academy of Sciences 103, 14223-14227.
St Pierre, S.E., Ponting, L., Stefancsik, R., McQuilton, P., FlyBase 102-advanced approaches to interrogating FlyBase.

SUGG, P., EDWARDS, J.S., BAUST, J., 1983. Phenology and life history of Belgica antarctica, an Antarctic midge (Diptera: Chironomidae). Ecological Entomology 8, 105-113.

Teets, N.M., Denlinger, D.L., 2014. Surviving in a frozen desert: environmental stress physiology of terrestrial Antarctic arthropods. The Journal of experimental biology 217, 84-93.

Teets, N.M., Peyton, J.T., Colinet, H., Renault, D., Kelley, J.L., Kawarasaki, Y., Lee, R.E., Denlinger, D.L., 2012. Gene expression changes governing extreme dehydration tolerance in an Antarctic insect. Proceedings of the National Academy of Sciences 109, 20744-20749.

Thomopson, J., Higgins, D.G., Gibson, T., 1994. ClustalW. Nucleic acids research 22, 4673-4680.

Vinogradov, A.E., 1999. Intron-genome size relationship on a large evolutionary scale. Journal of molecular evolution 49, 376-384.

Warthmann, N., Das, S., Lanz, C., Weigel, D., 2008. Comparative analysis of the MIR319a microRNA locus in Arabidopsis and related Brassicaceae. Molecular biology and evolution 25, 892-902.

Washietl, S., Hofacker, I.L., Stadler, P.F., Frauenfelder, H., 2005. Fast and Reliable Prediction of Noncoding RNAs. National Academy of Sciences, p. 2454.

Wu, P., Han, S., Chen, T., Qin, G., Li, L., Guo, X., 2013. Involvement of microRNAs in infection of silkworm with bombyx mori cytoplasmic polyhedrosis virus (BmCPV). PloS one 8, e68209.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., Kellis, M., 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434, 338-345.

Yang, A., Brill, A., 1991. Influence of the freezing process upon fluoride binding to hemeproteins. Biophysical journal 59, 1050.

Yousef, M., Showe, L., Showe, M., 2009. A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification. FEBS journal 276, 2150-2156.