COMPUTATIONAL IDENTIFICATION AND EVOLUTIONARY ANALYSIS OF METAZOAN MicroRNAs

A Dissertation

by

JUAN MANUEL ANZOLA LAGOS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2008

Major Subject: Biology

COMPUTATIONAL IDENTIFICATION AND EVOLUTIONARY ANALYSIS OF METAZOAN MicroRNAs

A Dissertation

by

JUAN MANUEL ANZOLA LAGOS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Rodolfo Aramayo
	Christine G. Elsik
Committee Members,	Rodney Honeycutt
	Michael Thon
Head of Department,	Thomas McKnight

December 2008

Major Subject: Biology

ABSTRACT

Computational Identification and Evolutionary Analysis of Metazoan MicroRNAs.

(December 2008)

Juan Manuel Anzola Lagos, B.Sc. Universidad Nacional de Colombia

Co-Chairs of Advisory Committee: Dr. Rodolfo Aramayo
Dr. Christine G. Elsik

MicroRNAs are a large family of 21-26 nucleotide non-coding RNAs with a role in the post-transcriptional regulation of gene expression. In recent years, microRNAs have been proposed to play a significant role in the expansion of organism complexity. MicroRNAs are expressed in a cell or tissue-specific manner during embryonic development, suggesting a role in cellular differentiation. For example, Let-7 is a metazoan microRNA that acts as developmental timer between larval stages in *C. elegans*. We conducted a comparative study that determined the distribution of microRNA families among metazoans, including the identification of new family members for several species. MicroRNA families appear to have evolved in bursts of evolution that correlate with the advent of major metazoan groups such as vertebrates, eutherians, primates and hominids. Most microRNA families identified in these organisms appeared with or after the advent of vertebrates. Only a few of them appear to be shared between vertebrates and invertebrates. The distribution of these microRNA families supports the idea that at least one whole genome duplication event (WGS) predates the advent of vertebrates. Gene ontology analyses of the genes these microRNA families regulate show enrichments for functions related to cell differentiation and morphogenesis.

MicroRNA genes appear to be under great selective constraints. Identification of conserved regions by comparative genomics allows for the computational identification of microRNAs. We have identified and characterized ultraconserved regions between the genomes of the honey bee (*Apis mellifera*) and the parasitic wasp (*Nasonia vitripennis*), and developed a strategy for the identification of microRNAs based on regions of ultraconservation. Ultraconserved regions preferentially localize within introns and intergenic regions, and are enriched in functions related to neural development. Introns harboring ultraconserved elements appear to be under negative selection and under a level of constraint that is higher than in their exonic counterparts. This level of constraint suggests functional roles yet to be discovered and suggests that introns are major players in the regulation of biological processes.

Our computational strategy was able to identify new microRNA genes shared between honey bee and wasp. We recovered 41 of 45 previously validated microRNAs for these organisms, and we identified several new ones. A significant fraction of these microRNA candidates are located in introns and intergenic regions and are organized in genomic clusters. Expression of 13 of these new candidates was verified by 454 sequencing.

DEDICATION

This manuscript is dedicated to the memory of my father, who was my example in life and role model to follow. To my mother for her immense love and care and for always being supportive of my decisions. To my grandma for always smiling at life even the darkest moments (I hope to have inherited a lot of genes from you). To my wife Martha, for her unconditional love, for putting up with me and my mental driftiness, and for showing me the true meaning of the word bliss.

ACKNOWLEDGEMENTS

I would like to thank the co-chairs of my committee, Dr. Christine Elsik and Dr. Rodolfo Aramayo, for giving me the freedom to pursue my own academic interests, for showing me the true value of science and the importance of the discussion of ideas in the scientific process. I owe them a great deal. I thank my committee members, Dr. Michael Thon and Dr. Rodney Honeycutt, for the helpful discussions and for teaching me important concepts in the fields of computational biology and phylogenetics.

I would like to thank my labmates, Michael Dickens and Anand Venkatraman, for their patience helping me solve code-related problems and for being always able to organize lab meetings to discuss the latest developments in the fields of computational biology and bioinformatics.

I also would like to thank Dr. Justin Reese for giving me examples of how to write elegant and understandable code and Danny Weaver for the most interesting discussions on honey bee biology.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

The central dogma of molecular biology was formally postulated by Francis Crick (1970). Information in biological systems is stored in the form of DNA, RNA works as intermediate step between DNA and proteins, and proteins perform most of the biological functions. This vision still remains valid, although there are an increasing number of molecules being discovered that do not fit into the view of the central dogma. Most of these molecules belong to the world of non-coding RNA. These are transcripts that function directly as RNA, and are not translated into protein (Eddy 1999).

MicroRNAs constitute a fraction of the non-coding RNA world. In their mature form, microRNAs usually range between 22 and 25 nucleotides in length (Ambros 2001), and are highly evolutionarily conserved. Some of them, such as let-7, have almost perfect conservation throughout metazoan evolution (Pasquinelli et al 2003).

This dissertation follows the style and format of *Genome Research.*

The discovery of microRNAs goes back to 1993 when Vicror Ambros and colleagues discovered that the gene lin-4, known to control certain steps in the developmental timing of *C. elegans* (Lee et al 1993), does not code for a protein but instead produces a small RNA molecule with the ability to form a fold-back structure, and whose further processing produces a shorter molecule 22nt long.

Researchers noted that lin-4 had numerous antisense complementary sites in the 3'UTR region of the gene lin-14 (Lee et al 1993). This complementarity was localized in a region of the UTR that was previously proposed to be the responsible for the downregulation of lin-14 by lin-4. This finding led to the proposal of a new kind of small RNA molecules known as "small temporal RNAs", indicating the role of these molecules as timing developmental regulators.

The identification of lin-4 was the first in a chain of events that led to the discovery of thousands of these small molecules in plants, animals, and viruses that had in common a fold-back precursor, a small RNA as functional unit, and antisense complementary to 3'UTR regions of target genes. The role of these RNA molecules was no longer restricted to developmental timing, as hundreds of them were found to be involved in tissue differentiation (Aboobaker et al 2002), cell proliferation, morphogenesis (Giraldez et al 2005), apoptosis (Xu et al 2004), aging and life span (Giraldez et al 2005), and chromatin remodeling. To accomode the broader range functions the "small temporal RNAs" became known as microRNAs (Lagos-Quintana et al 2001).

As of June 2007, about 5000 microRNAs had been identified in a variety of plants, animals, and viruses. Data related to microRNAs is systematically stored in databases such as miRBase (Griffiths-Jones et al 2006) with experimentally validated microRNAs; Tarbase with experimentally validated targets for microRNAs (Sethupathy et al 2006), and miRGen with integrated genomics information for microRNAs (Megraw et al 2007).

What is a microRNA?

MicroRNAs are small single stranded RNA molecules of about 22nt long. They are derived from longer precursors (or pre-microRNAs) ranging from about 70nt to 110nt in length that adopt a fold-back structure, also known as hairpin structure. Precursors of microRNAs have been found residing in introns (Lin et al 2006), exons (Ying and Lin 2004) and intergenic regions (Gu et al 2006). Intronic and exonic microRNAs are likely to share regulatory elements and are encoded in the same transcript as their host gene (Bartel 2004) and are processed by RNA polymerase II. Intergenic transcripts are presumably transcribed from their own promoters and are derived from longer transcripts known as pri-microRNAs. It has been shown that at least some intergenic microRNA transcripts are poly-adenylated just like messenger RNAs (Saini et al 2007) and are also processed by RNA polymerase II (Kim 2005).

Isolation of microRNAs

The nature of mature microRNAs makes them very difficult to isolate by traditional cloning techniques. In order to avoid this problem, Ambros and collaborators have developed a system for microRNA annotation that is aimed at the differentiation of microRNAs from other endogenous RNA molecules such as small interfering RNAs. According to Ambros (Ambros et al 2003), microRNAs can be identified using the following criteria:

Expression criteria:

A. Detection of a distinct ~22nt transcript by Northern Blot, RT-PCR or RNase protection methods, or

B. Identification of a ~22nt sequence in a library of cDNAs made from size fractionated RNA.

Biogenesis criteria:

C. Prediction of a ~70 potential fold-back precursor structure that contains the ~22nt mature microRNA within on arm of the hairpin.

D. Phylogenetic conservation of the ~22nt mature sequence and conservation of the precursor secondary structure.

E. Detection of increased precursor accumulation upon deletion of a major microRNA processing enzyme.

In an ideal scenario, small RNA molecules would be classified as microRNAs if they comply with criteria A, D, and E. But in the absence of such data, fulfillment of criterion D is sufficient.

MicroRNA biogenesis in metazoans

MicroRNAs are transcribed by RNA polymerase II. Primary transcripts are usually several kilobases long and contain one or several hairpin-like structures. These pri-microRNAs are then cleaved by the ds-specific RNA endonuclease Drosha, which releases the precursor of the microRNA, called the pre-microRNA. The cut performed by Drosha is done at the base of the pre-microRNA. It leaves a 5' phosphate end and a 3' OH end with a ~2nt overhang. This cut defines one end of the mature microRNA, the other cut is performed by another RNase III endonuclease known as Dicer and occurs in the cytoplasm. Following the Drosha cut, the pre-microRNA is then exported to the cytoplasm through nuclear pore complexes by a process mediated by the nuclear transport receptor Exportin-5 and Ran-GTP. Once in the cytoplasm, pre-microRNAs are subsequently processed into ~22nt RNA duplexes by Dicer. The mechanism by which Dicer recognizes and chops the pre-microRNA is still not well understood, but experimental evidence suggests the 3' overhang at the base of the stem-loop provides the recognition site. Afterwards, Dicer cuts both strands of the duplex about two helical turns away from the base, generating a double stranded RNA molecule that contains 3’ overhangs at each complementary end.

One of the strands from this duplex is processed as mature microRNA. The other one, known as microRNA*, is thought to be degraded. Evidence from massive cloning efforts (Miska et al 2004) indicates a 100 fold difference in cloning frequency between microRNAs and microRNAs*. This major difference is a strong indication that the RNA duplex is short-lived compared to the single mature microRNA strand (Kim 2005; Liu et al 2008).

Experimental evidence suggests that the differential stability of both ends of the RNA duplex determines which strand is to be selected. The strand with relatively unstable pairing at the 5’end usually is the one used as mature microRNA.

The mature strand is eventually incorporated into the RNA Induced Silencing Complex (RISC) (Gregory et al 2005). The microRNA-RISC complex, also known as miRNP or miRISC, identifies target mRNAs based on near perfect complementation between the microRNA and the mRNA. It is thought that this near-complementarity drives translational repression as opposed to mRNA degradation, as it does in plants. The mechanism of translational repression is still not clear. One possibility is that RISC sequesters the target mRNA away from the translational machinery (Bartel 2004; Liu et al 2005). Another possibility is that the miRNP complex may hamper ribosome movement along the mRNA repressing protein translation (Carrington and Ambros 2003).

Molecular characteristics of Drosha and Dicer

Drosha

Drosha is a monomeric ~160kDa RNase that belongs to the Class II of RNA endonucleases. The class is defined by the presence of two RNAse III domains (RIIIDs) and a dsRNA binding domain (dsRBD) that is critical for catalysis. Drosha forms a complex of about ~650kDa in humans, and ~500kDa in *Drosophila.* This complex is also known as the Microprocessor Complex. Drosha interacts with a cofactor known as DGCR8 (Pasha in *Drosophila*) that is a ~120kDa protein with two dsRDB domains. It is believed that DCGR8/Pasha provides the specificity for Drosha cleavage (Han et al 2006; Zeng and Cullen 2003).

Dicer

Dicer is a ~220 kDa protein that also belongs to the Class II of RNA endonucleases with their characteristic RNase III domains (dsRDB). Besides these, Dicer contains an N-terminus DExH/DEAH-box RNA helicase / ATPase domain, a DUF domain (Domain of Unknown Function) and a PAZ domain (Macrae et al 2006; Song, Liu et al 2003).

Biochemical evidence suggests a tight interaction between Dicer and RISC. This dual role of Dicer, as enzyme required for pre-microRNA processing and its interaction with RISC suggests that the two events may be coupled (Lee and Collins 2007).

Identification of microRNA targets

Despite the fact that we know only a fraction of experimentally validated microRNA targets, evidence indicates that microRNAs play a big role in gene regulation. Computational evidence suggests the average microRNA may regulate more than a hundred genes (Bentwich 2005; Enright et al 2003; Grun et al 2005; Rajewsky 2006; Robins et al 2005; Stark et al 2003). These approaches are modeled based on the binding properties of the most studied microRNAs up to this date: lin-4, let-7 and their experimentally validated targets. These models take into account the following properties: 1) location of microRNA complementary sites in 3'UTR regions of mRNA targets; 2) measurement of base paring in the 5' region of the microRNA, also known as the seed region; 3) Phylogenetic conservation of complementary sites in 3' UTRs of orthologous genes (Ambros 2004).

Although the computational procedures used to detect targets are still plagued by high false positive rate, there are new approaches that seem to partially overcome this and make progress towards a reliable identification of single site targets by combining experimental and computational methodologies (Kiriakidou et al 2004). Pure computational approaches require more information about the mechanism of interaction between microRNA and their targets, and more research is necessary to elucidate the details of this interaction.

Functions of microRNAs in metazoans

The most studied animal microRNAs lin-4 and let-7 regulate developmental timing in *C. elegans*. Lin-4 controls the earlier transition between larval stage 1 to larval stage 2. Let-7 controls the transition between larval stage 2 to larval stage 3 (Lee et al 1993). Knock outs of both of these genes result in delayed development (Zhang et al 2007).

Experiments in Zebrafish Dicer mutants indicate the microRNA pathway is absolutely necessary for proper embryo development. In the first days post-fertilization, mutants start accumulating abnormal levels of microRNAs. This accumulation ceases with a complete developmental arrest around day 10. These results indicate that Dicer is essential for vertebrate development. (Wienholds et al 2003)

The microRNA miR-1 is abundant in cardiac muscle. Experiments suggest miR-1 is necessary for the proper shaping of ventricles in the heart. In fly, miR-1 deletion hampers severely the cascade of events that leads to chamber morphogenesis. This microRNA regulates the Notch signaling pathway by targeting the mRNA of the Notch ligand, Delta, indicating its role as regulator of cardiac cell differentiation. (Artavanis-Tsakonas et al 1999; Zhao and Srivastava 2007).

In *C. elegans*, the microRNAs lsy-6 and miR-273 are expressed in chemosensory neurons that are part of the worm's sensory discriminatory system. Two gustatory neurons, ASE left (ASEL) and ASE right (ASER), exhibit asymmetric molecular features that allow the worm to distinguish various chemical stimuli. Data indicates these microRNAs are heavily involved in the asymmetric features between these two neurons. Lsy-6 represses the ASER fate, whereas miR-273 represses the ASEL fate by binding to mRNAs of specific transcription factors for each cell type. These microRNAs reinforce the genetic programs that lead to left-right asymmetry (Johnston and Hobert 2003).

MicroRNAs also play a role in cancer pathogenesis. In humans, miR-15a and miR-16a are located at the chromosomal location 13q14, which is a region frequently absent in patients with B cell chronic lymphocytic leukemia. A study conducted by Calin and collaborators (Calin et al 2004) finds that 98 of 186 (52.5%) of human miR genes are in cancer-associated genomic regions or in fragile sites. There is also differential expression of microRNAs between cancer cells and their normal counterparts (Lu et al 2005).

In lung cancer cells, expression of let-7 is significantly reduced. Let-7 seems to downregulate the oncogene RAS by binding to multiple sites on its 3’ UTR (Takamizawa et al 2004). Further studies confirm the overexpression of let-7 as inhibiting growth in cancerous lung cells (Johnson et al 2005), suggesting a role for let-7 as tumor suppressor gene in humans. MicroRNAs also play a role in apoptosis. The microRNA Bantam inhibits cell death and promotes cell proliferation (Brennecke et al 2003; Xu et al 2004).

In human cells the microRNA microRNA-32 restricts the accumulation of the retrovirus primate foamy virus type 1 (PFV-1). In vivo, this virus is specific to non-human primates, but the fact that the microRNA is able to restrict the accumulation of the retrovirus suggests a role as antiviral agent (Lecellier et al 2005).

Evolution of microRNAs

Very little is known about the forces that give origin to microRNA genes. A major force of genome innovation is gene duplication, but it is unclear as to what extent the emergence of microRNA genes follows local duplication events (tandem duplication), duplication by transposition or whole genome duplications. In plants it has been demonstrated that at least some microRNAs evolve by inverted duplication of target genes (Allen et al 2004). This has not been demonstrated yet in animals, but is likely the same mechanism is present in metazoans. Recent studies suggest a subset of metazoan microRNAs have originated and evolved from transposable elements. They appear to be conserved in related species and are the origin for lineage specific microRNA innovations (Piriyapongsa and Jordan 2007; Piriyapongsa et al 2007).

MicroRNAs have also been used to give insights into the phylogenetic position of problematic groups like the Platyhelminthes. Gene sequences suggest the acoel flatworms are not members of the phylum Platyhelminthes, but instead are the most basal branch of triploblastic bilaterians. Using microRNAs as genetic markers, Sempere and collaborators obtained a picture in which the acoel flatworms are indeed basal triploblastic bilaterians, suggesting our understanding of the group Platyhelmintha is incomplete, because according to this picture Platyhelmintha is paraphyletic (Sempere et al 2006).

Evolutionary analyses of microRNA distributions suggest microRNA innovation as an ongoing process in metazoans. Major microRNA expansions seem to correlate very well with the advent of bilaterians, vertebrates and placental mammals (Hertel et al 2006), however it is unclear what is the contribution of microRNAs to the specific morphological characteristics and phenotypes of these major groups.

A detailed analysis of basal metazoan genomes indicates the presence of a core of 18 microRNAs found only in protostomes and deuterostomes (Coelomata) but not in sponges or cnidarians. Given the fact the microRNAs are known to be expressed in specific tissues and/or organs, the authors propose this core set of microRNAs could have played a role in the development of organ structures such as the brain and heart because these structures are not present in sponges or cnidarians (Sempere et al 2006).

Objectives of this study

MicroRNAs have been recognized as one of the major forces shaping the developmental processes of plants and animals. We are starting to understand the precise mechanism of action of microRNAs and the evolutionary processes by which they arise. This work is aimed at broadening our understanding of microRNA evolution, their distribution in the tree of life, their functionality throughout the evolution of metazoans, and the computational methods used for their identification.

This work explores different aspects of microRNA biology. We start with a phylogenetic perspective of microRNA distribution throughout the evolution of metazoans. We developed a computational strategy for the identification of microRNA families that allowed us to determine different points of microRNA innovations and allowed us to get an idea of the functionality of these microRNA families.

This work is also aimed at the *de novo* identification of microRNA genes by means of comparative genomics. This goal takes advantage of the fact that functional conserved elements in genomes are also conserved at the sequence level. We used ultraconservation as a way to identify these functional elements, we explore the functionality of these ultraconserved sequences based on gene ontologies, and use them to develop an algorithm for the identification of new microRNA genes between two recently sequenced insect genomes, the honey bee, *Apis mellifera*, and the parasitic wasp, *Nasonia vitripennis*.

CHAPTER II

COMPUTATIONAL IDENTIFICATION, CHARACTERIZATION AND FUNCTIONAL EVOLUTIONARY ANALYSIS OF METAZOAN MicroRNAs

Introduction

The first free-living organism for which we determined its genomic sequence was *Haemophilus influenzae* (Fleischmann et al 1995), followed by *Drosophila melanogaster* (Adams et al 2000), and the milestone of the Human genome in 2001 (Lander et al 2001). As of March of 2008 there were more than 4000 genomes sequenced or in the process of being sequenced. The knowledge and information generated by genome projects grows at a relentless pace, and the analyses of those genomes have changed preconceived ideas that were thought to be true. Before the Human genome project it was estimated that the number of genes on the human genome was closer to 100000 (Claverie 2001; Pennisi 2000). Today that estimate is closer to 30000 (Claverie 2001). This finding was a little bit of a shock for the scientific community and the public in general, given the fact that the fruit fly (*Drosophila melanogaster*) has roughly 14000 genes (Adams et al 2000). *Drosophila* is a very complex organism indeed, but as we learn from the information generated by genome projects, complexity does not correlate well with gene number or genome size (Gregory

2001). Evidence points to the fact that complexity arises by the way the set of genes in an organism interact with each other coupled with their activation/inactivation in different cellular conditions (Claverie 2001). This is particularly true for species with sexual dimorphism in which one genome is able to generate two remarkably different phenotypes or for species like honey bee in which one single genome can generate three completely different phenotypes (queen, drone, worker), each one with a particular role in the well being of the hive and the survival of the species. This picture points to gene regulation as an important driving force in the evolution of species and their genomes, probably the most important factor that leads to increasingly complex organisms.

Gene regulation is a concept we intrinsically associate with transcription factors (Chen and Rajewsky 2007; Hobert 2008). These are proteins that bind to promoter regions and allow the RNA polymerase complex to initiate transcription of a target gene. Our current understanding of transcription factors still depicts them as the most important players in the regulation of genes, but in recent years new mechanisms of gene regulation that operate mostly at the posttranscriptional level have been discovered. These include the siRNA pathway and the microRNA pathway, both of which are thought originated as a way to counteract viral infections by the innate immune system of early metazoans (Lu et al 2005). Today the role of these pathways is no longer restricted to immune response, as they are heavily involved in the regulation of development, morphogenesis and the maintenance of cell types such as stem cells (Carrington and Ambros 2003).

Our understanding of the roles microRNAs play in the cell gets better as more genomes are sequenced and as more experiments are conducted; but we understand very little of microRNA biology in the context of a phylogeny. In other words, what are the primordial functions of microRNAs at different stages along the pathway that led to the human species? Here we propose an evaluation of microRNA functions on the basis of gene ontologies. For this objective, we have developed a computational strategy aimed at the identification of potential orthologues in 24 different metazoan species of microRNAs deposited in the microRNA Registry (miRBase). MicroRNAs identified this way were then mapped into the metazoan phylogeny and their functions evaluated on the basis of gene ontologies.

Results

Computational identification of microRNA homologs

Our searches started with a set of 5071 microRNAs precursors from miRBase v. 10.0. This set of sequences contains microRNA genes from animals (3682), plants (1268) and viral genomes (107). In order to identify potential homologs of these microRNAs in the 24 metazoan species evaluated, we conducted a sensitive (but at the same time unspecific, evalue =10) BLAST search against the

24 genome assemblies (Table 1). BLAST parameters were modified in order to minimize false positives. We used a word size of 16 for seeding. The seed dictionary was built only from the mature region of the microRNA. These regions were extracted from microRNA precursors with the help of a perl script. This step guarantees sequence identity for the mature microRNA between the query and any resulting hit. Sequences collected this way were mapped to their respective genome assemblies to reduce the number of sequences mapping to the same position. In order to minimize the possibility of any hit corresponding to a false positive, we evaluated sequence pairs (microRNA – putative homolog) using PRSS – for details please see methods section - with 1000 shuffles (Pearson 1990). Sequences with p-values lower than 1e-05 were considered homologous. This parameter was selected on the basis of PRSS scores for experimentally validated microRNAs. This resulted in 4856 different microRNA homologs that in some cases included genes not reported previously for the species under study and new paralog genes (Table 1).

Construction of microRNA gene families

In order to compute clusters, it was necessary to conduct an all vs. all comparison of all resulting microRNAs from the evaluated species. These comparisons were carried out using PRSS (Pearson 1990) with 1000 shuffles. This allowed us to determine homology among all the sequences.

Since we estimate homology on the basis of the statistical significance of a sequence alignment, it was necessary to determine the best clustering threshold. There is no magic parameter for sequence clustering. That really depends on the type of sequences being used (proteins, DNA sequences, RNA sequences), database size and the model of evolution. For clustering we used a simple but sensitive clustering method that is based on Smith-Waterman local alignment algorithm and a theoretical database of 1000 sequences (1000 shuffles). Clustering parameters were based on the ones used for the let-7 family. This gene family is ubiquitous among metazoan genomes and it comprises members of let-7 and miR-98. We use different p-value thresholds in order to determine what was the one at which all the members of these family were clustered together into a single family (Table 2), and we used that threshold in order to cluster all the microRNAs found on our search.

The p-value 1e-06 was used as a clustering threshold for all the sequences found. There are other more inclusive thresholds at which clustering could be done, but that results in overclustering of some families, and lower p-values may result in oversplitting. 1e-06 turned out to be sensitive enough but also specific enough to keep cluster size on acceptable ranges. Clusters built this way were compared against microRNA families deposited in miRBase, resulting in similar number of clusters, sizes and memberships.

Table 1. Number of microRNAs identified in this study

Species	Assembly	miRBase 10.0	Identified in this study	New
Mammals				
Bos Taurus	btau3.1	95	272	177
Canis familiaris	CanFam2.0	6	297	291
Equus caballus	EquCab1	0	236	236
Homo sapiens	HSA46	475	523	48
Macaca mulatta	Mmul_051212	60	409	349
Monodelphis domestica	MonDom5	106	169	63
Mus musculus	C57BL/6J	426	435	9
Ornithorhynchus anatinus	IOANA5.46	0	44	44
Pan troglodytes	CCY5Cv1	79	485	406
Pediculus humanus	PhumU1	0	20	20
Rattus norvegicus	RG5C3.4	330	549	219
Other vertebrates				
Danio rerio	ZFISH7.46	290	349	59
Gallus Gallus	May2006	140	153	13
Takifugu rubripes	FUGU4.46	133	223	90
Xenopus tropicalis	Ver4.1	177	225	48
Insects				
Aedes aegiptii	AaegL1	0	43	43
Anopheles gambiae	april 2004	41	44	3
Apis mellifera	Amel4.0	50	50	0
Bombyx mori	ge2k	21	31	10
Drosophila melanogaster	dmel4.2.1	76	78	2
Drosophila pseudobscura	dpse_r2.0	72	76	5
Nasonia vitripennis	nvit1.0	0	43	43
Tribolium castaneum	JGI4.1.46	0	30	30

Table 2. Results of different clustering parameters of the microRNAs evaluated

E-value	Clusters	Clusters for let-7 family
1	1	1
0.1	2	1
1E-02	50	1
1E-03	186	1
1E-04	296	1
1E-05	322	1
1E-06	**338**	**1**
1E-07	357	1
1E-08	378	1
1E-09	393	2
1E-10	417	2
1E-11	436	2
1E-12	456	3

Clustering at 1e-06 resulted in 408 different gene families, 270 having 3 or more members, 99 having 2 members, and 39 being singletons. Most of the singletons were unique to *C. elegans*, as this is the most divergent species used in our analyses and was the one we used as outgroup. It is also the only representative we used of the nematode lineage. Had *C. briggsae* being used in this study, most of these singletons would have appeared as shared genes between the two species.

The largest gene family found in our analysis is the miR-343/miR-369/miR-510 family containing 256 members. This family is divergent, their members are not very conserved and some of them are related to transposable elements. It also appears to be specific to eutherian mammals. The second largest is the miR-302/miR-320/miR-524 family, which contains 204 members and appears to be specific to vertebrates. The third largest is the let-7/mir-98 family, which is common to all metazoans and contains 172 members.

The phylogenetic position of each individual family was determined by looking at the distribution of their members along the metazoan phylogeny with the help of a binary matrix in which it was plotted whether or not the family was present (1) or not present (0) in each species. This matrix along with the phylogeny was imported into Macclade (Maddison and Maddison 1989) and the characters (presence/absence of family) were plotted along the phylogeny using the parsimony method.

Reconstructing evolutionary events using phylogenetic methods is not a simple task. If something have we learned from evolutionary analyses is that genomes evolve at different rates; there are gene gains, gene losses, gene duplications and gene expansions that may complicate our interpretation of the results. MicroRNAs are no different than other genes when it comes to this. Figure 1 shows a scenario where the phylogenetic position of the gene family is ambiguous at best. Under parsimony, explaining how this family has evolved requires multiple gains or multiple losses. For example, we can assume the family was present in the common ancestor of coelomates and then it has been lost *independently* in *Pediculus humanus*, *Aedes aegyptii*, *Anopheles gambiae*, *Tribolium castaneum*, *Danio rerio* and *Monodelphis domestica*. A different scenario would be to assume the family wasn't present in the common ancestor of coelomates and it appeared as a result of convergent evolution in amniotes, hymenopterans and dipterans.

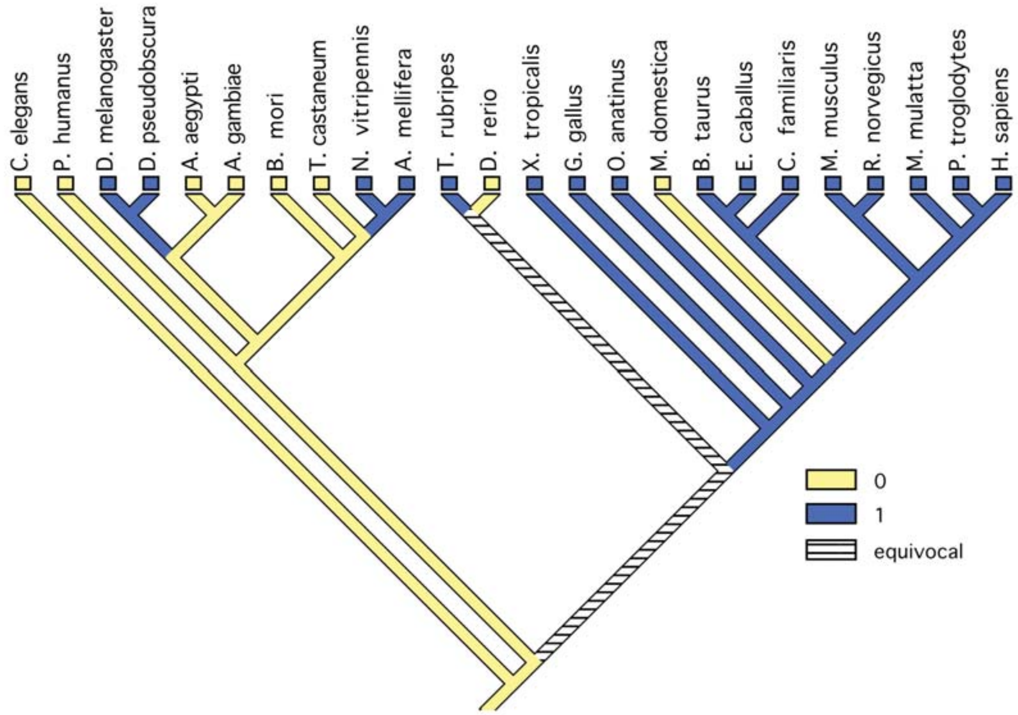


Figure 1. Example of a microRNA gene family with ambiguous distribution. Explaining the evolution of this gene familiy requires multiple gains and multiple losses. It is unclear what was the point at which the family made its appearance in the metazoan phylogeny.

These complicated scenarios can be the result of:

1. The method used to determine homology wasn't sensitive enough to detect a homolog on the genome under scrutiny.
2. Compositional bias of the genome compared to sister species (i.e. *Apis mellifera* is AT rich whereas *Nasonia vitripennis* is not AT rich.)

3. microRNA not being present in the version of the genome used for the analysis. Some of the genomes used are on early stages of assembly (Platypus), some others have been refined over time (*Drosophila*).
4. microRNA being present in euchromatic region in one genome and in heterochromatic region on a different genome, in which case it is not possible to detect it, as most of the genomes sequenced so far correspond to euchromatic regions (with the exception of *Drosophila*).
5. Real gene losses and/or real gene gains.
6. Clustering algorithm used left out members of this family that should have been grouped together.

Analysis of all the gene families under the parsimony method shows that the great majority of families are consistent with simple scenarios of gene gains and gene losses and complicated scenarios like the one just described are the minority. Figures 2 and 3 show simple scenarios with gene family distributions that take only one step (one gain for the vertebrate lineage, Figure 2) or two steps at the most (one gain for the eutherian lineage and one loss for the horse, Figure 3). For further gene functional analyses we narrow our set of microRNA families only to the ones that were perfectly consistent (CI index =1) and the ones with simple scenarios of gains and losses. Complicated scenarios were avoided. Figure 4 shows the distribution of microRNA families along the metazoan phylogeny.

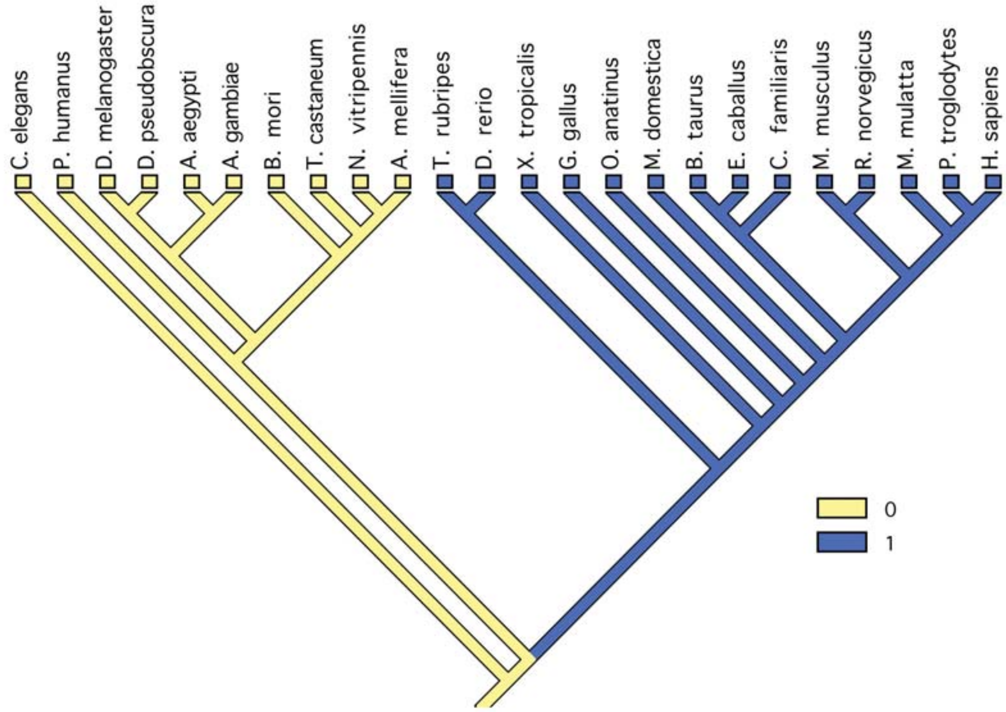


Figure 2. Example of a perfectly consistent microRNA family. Members of this family are shared by all the vertebrate species used in this study. This family was present in the most recent common ancestor between Humans and Fish.

The branch that goes from coelomates to humans is the one that shows most of the microRNA innovations (Figure 4). It appears that there have been bursts of microRNA innovations in the branch that leads to vertebrates (67 microRNA families), therian-eutherian mammals (77 families), primates (41 families) and hominids (58 families). Interestingly the branch that leads to the mammals does not show major innovations. There appears to be only five different families at this position and most of them with consistency indexes of less than 1.

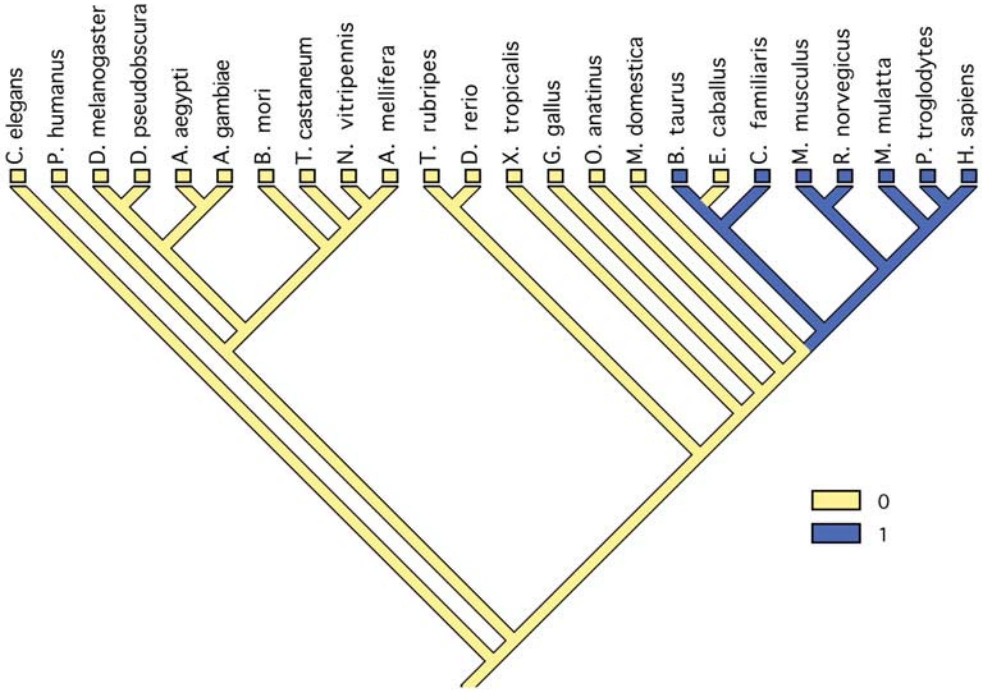


Figure 3. Example of a microRNA family with simple homoplasy. The explanation of how this gene family has evolved only requires one gain for the eutherian lineage and one loss for the horse genome. Most of the microRNA families of this study are consistent with simple scenarios like this one.

The fact that the mammalian clade does not appear to show major innovations is puzzling, since mammalian species certainly show major morphological innovations compared to other groups. It is possible that *some* vertebrate/mammalian microRNA innovations have indeed been lost during the evolution of monotremes, but it is *unlikely* that this phenomenon accounts for the number of families that appear to be lost according to our result (a number that might be close to 30 families). We consider that the most likely explanation for this result is that that mammalian microRNA innovations in Platypus are so divergent that are unrecognizable with current methods of identification. Similar results have been found in other studies (Hertel et al 2006; Warren et al 2008).

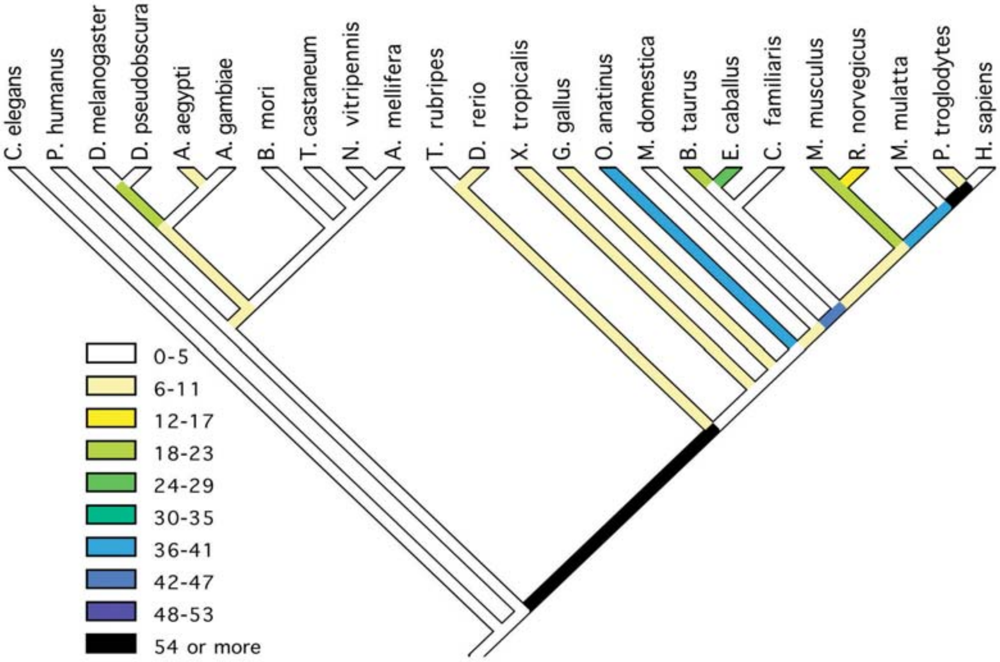


Figure 4. Distribution of microRNA gene families in metazoans. Different colors represent the amount of change throughout the evolution of metazoans.

MicroRNA family size

Our analysis of the distribution of family sizes in the human genome shows that most families contain 3 or fewer members, and only 11 families contain 4 or more members (Figure 5). Two of these 11 families are expanded in the human genome, mir-320c and mir-369 with at least 52 and 33 members respectively. These two families are vertebrate innovations and are also expanded in other vertebrate genomes. Mir-302 from *Mus musculus* contains at least 10 members, and mir-369 at least 27. In humans, at least five of these 11 families have invertebrate counterparts: let-7, mir-34, mir-10, mir-8 and mir92b. Taking all the genomes into consideration, the number of families shared between vertebrates and invertebrates increases to at least 21: let-7, mir-1, mir-7, mir-8, mir-10, mir-31a, mir-33, mir-34, mir-79, mir-92b, mir-124, mir-125, mir-133, mir-137, mir-184, mir-190, mir-210, mir-219, mir-263, mir-375 and mir-739.

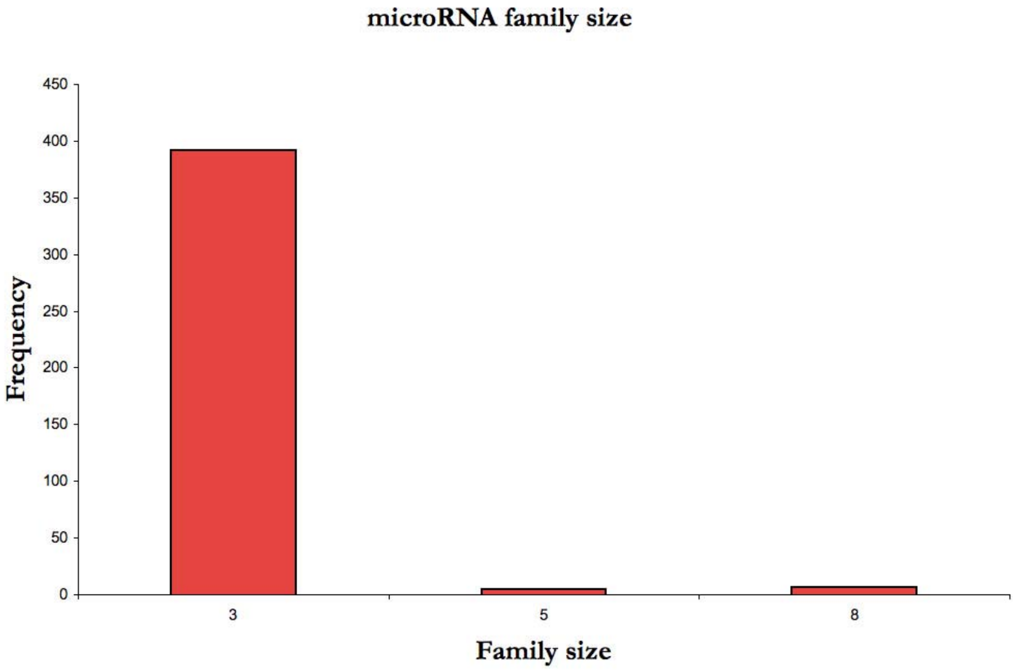


Figure 5. Average microRNA family size in the human genome. Most families contain 3 or fewer members. Similar family sizes are found in other vertebrate and mammalian genomes.

We computed the relationship in number of members for these families between invertebrate and vertebrate genomes (Figure 6). On average, microRNA family sizes have been expanded more for vertebrates than for their invertebrate counterparts. Vertebrate gene families are at least two or four times larger than invertebrate ones.

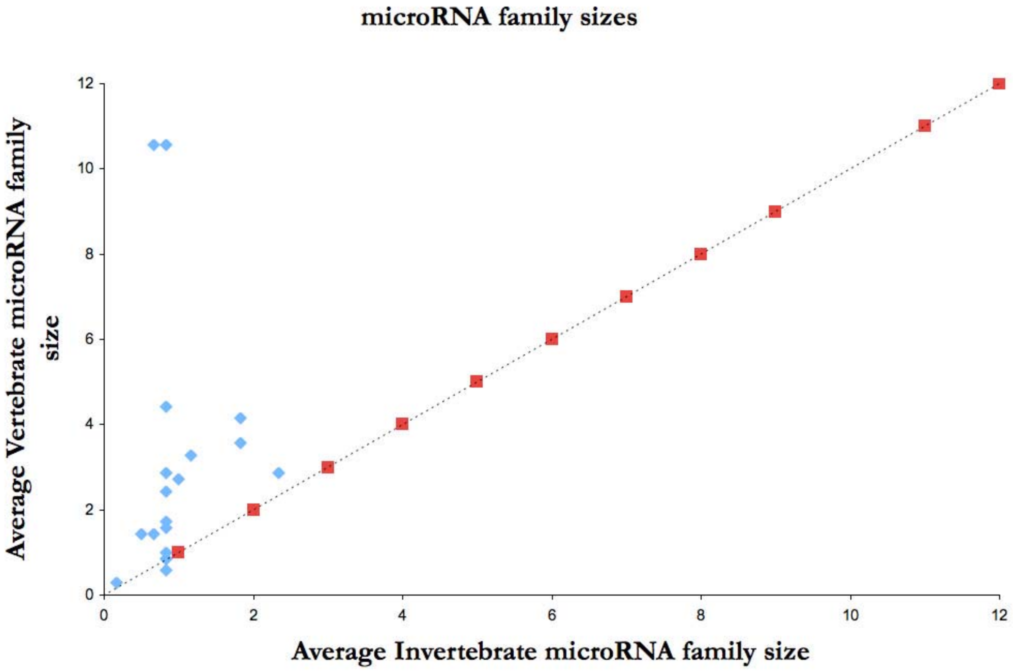


Figure 6. MicroRNA family size between vertebrates and invertebrates. Shared families between invertebrate and vertebrate genomes. The dotted line represents one to one relationship between family sizes. On average, vertebrate family sizes appear to be twice or four times the size of their invertebrate counterparts.

The relationship found between shared invertebrate and vertebrate families does not exist for families unique to the vertebrate lineage. Comparisons in the number of members for families shared between vertebrate-non-mammalian genomes (Zebrafish, Chicken, Xenopus, Fugu) and their eutherian counterparts (Horse, Mouse, Dog, Cow, Human) shows that there is almost a 1:1 relationship in family size. Eutherian microRNA families don't seem to be particularly expanded when compared to their vertebrate-non-mammalian counterparts, exceptions to this behavior come from the microRNA family miR-466, which appears to be derived from transposable elements and is relatively expanded in vertebrate non-mammalian genomes. (Figure 7).

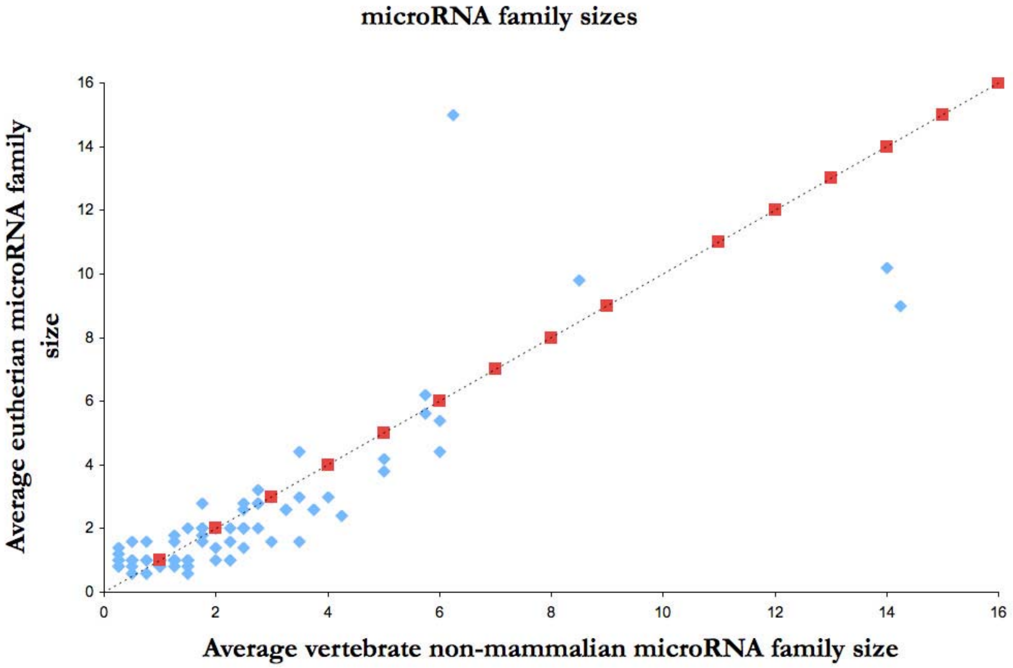


Figure 7. MicroRNA family size between vertebrate-non-mammalian and eutherian genomes. The dotted line represents one to one relationship between family sizes. On average, family sizes don't appear to deviate significantly from a one to one scenario.

Functions of microRNAs

In order to explore the functionality of microRNAs at different points during the evolution of metazoans, we identified target genes of microRNAs that should have been present in the common ancestor of the group under scrutiny. So for example, in order to evaluate the functions of microRNAs of eutherians, we identified orthologous genes between *Bos taurus* and *Homo sapiens*, as Bos and Human shared a common ancestor 148 million years ago and are the most divergent genomes representative of eutherians in our phylogeny. We identified target genes in this ortholog set for microRNAs that made their appearance at the time of the Bos – Human split and evaluated their functionality in the context of gene gntologies. Functionality was determined by evaluating if there were any GO enrichment within the gene set. For details please see the methods section.

We were interested in assessing the contribution of microRNAs with respect to the evolution of the major morphological innovations seen throughout metazoans. In the scientific community there is the consensus that these innovations are the result of new protein coding genes performing new functions. We want to demonstrate, although at a very coarse level, that gene regulation performed by microRNA genes may play a significant role in the appearance of major innovations. For this particular purpose we have assessed whether or not there were any Gene Ontology enrichments for microRNA targets at specific points during the evolution of metazoans. We have found several enrichments that are worth noting. These are summarized in Table 3.

Our results indicate that there is a common theme along the branch that starts with coelomates and leads to humans, as we found the term "Nervous System Development" being enriched three different times, at the basal level (coelomate level), in therian mammals (marsupials and placentals) and in euarchontoglire mammals. The terms "Positive regulation of transcription from RNA polymerase II promoter" and "Multicellular Organismal Development" enriched in vertebrates, eutherians, euarchontoglires, primates and hominids. The term "Regulation of Cell Proliferation" enriched in vertebrates, eutherians, euarchontoglires and hominids. The term "Homoiothermy" in vertebrates, eutherians and hominids; and finally the terms "Transcription", "Cell differentiation" and "Wnt Signaling Pathway" in euarchontoglires.

Table 3. Significant gene ontology enrichments for different metazoan taxa

GO Category	e-score corrected for multiple testing	GO identifier	Description
Basal			
Biological Process	0.012	GO:0007399	Nervous system development
Coelomata			
Molecular Function	2.16E-10	GO:0005515	Protein binding
Vertebrata			
Biological Process	2.78E-06	GO:0045944	Positive regulation of transcription from RNA polymerase II promoter
Biological Process	0.00012	GO:0007275	Multicellular organismal development
Biological Process	0.010	GO:0008284	Positive regulation of cell proliferation
Biological Process	0.037	GO:0042309	Homoiothermy
Cellular Component	0.0035	GO:0005769	Early endosome
Amniota			
Cellular Component	0.012	GO:0048471	Perinuclear region
Cellular Component	0.020	GO:0030014	CCR4-NOT complex
Cellular Component	0.020	GO:0042175	Nuclear envelope-endoplasmic reticulum network
Cellular Component	0.032	GO:0000139	Golgi membrane
Theria			
Biological Process	0.00020	GO:0007399	Nervous system development
Biological Process	0.020	GO:0006355	Regulation of transcription, DNA-dependent
Molecular Function	0.006	GO:0003700	Transcription factor activity
Molecular Function	0.10	GO:0005515	Protein binding
Molecular Function	0.018	GO:0003677	DNA binding
Cellular Component	0.0064	GO:0005634	Nucleus
Eutheria			
Biological Process	3.06E-13	GO:0007275	Multicellular organismal development
Biological Process	0.0007	GO:0045944	Positive regulation of transcription from RNA polymerase II promoter
Biological Process	0.002	GO:0008284	Positive regulation of cell proliferation
	0.0023	GO:0005769	Early endosome
Cellular Component			
Cellular Component	0.019	GO:0005622	Intracellular

Table 3. Continued

Category	e-score corrected for multiple testing	GO identifier	Description
Euarchontoglires			
Biological Process	0.043	GO:0008284	Positive regulation of cell proliferation
Biological Process	4.62E-10	GO:0006355	Regulation of transcription, DNA-dependent
Biological Process	1.47E-06	GO:0007275	Multicellular organismal development
Biological Process	4.37E-05	GO:0006350	Transcription
Biological Process	9.70E-05	GO:0045944	Positive regulation of transcription from RNA polymerase II promoter
Biological Process	0.00038	GO:0030154	Cell differentiation
Biological Process	0.0017	GO:0016055	Wnt receptor signaling pathway
Biological Process	0.0088	GO:0007399	Nervous system development
Molecular Function	5.18E-14	GO:0005515	Protein binding
Molecular Function	4.94E-8	GO:0003700	Transcription factor activity
Molecular Function	2.95E-05	GO:0043565	Sequence specific DNA binding
Molecular Function	0.00020	GO:0046872	Metal ion binding
Molecular Function	0.00047	GO:0003677	DNA binding
Molecular Function	0.0022	GO:0008270	Zinc ion binding
Molecular Function	0.027	GO:0016563	Transcription activator activity
Molecular Function	0.035	GO:0003676	Nucleic acid binding
Molecular Function	0.039	GO:0046875	Ephrin receptor binding
Cellular Component	2.06E-10	GO:0005634	Nucleus
Primates			
Biological Process	3.89E-08	GO:0045944	Positive regulation of transcription from RNA polymerase II promoter
Biological Process	9.16E-08	GO:0007275	Multicellular organismal development
Cellular Component	0.0011	GO:0005769	Early endosome
Hominidae			
Biological Process	1.39E-13	GO:0007275	Multicellular organismal development
Biological Process	5.45E-08	GO:0045944	Positive regulation of transcription from RNA polymerase II promoter
Biological Process	0.000494	GO:0008284	Positive regulation of cell proliferation
Biological Process	0.007276	GO:0030154	Cell differentiation
Biological Process	0.012	GO:0042309	Homoiothermy

Functional bias of microRNAs

In order to get a better understanding of the functions being performed by microRNAs in the context of a phylogeny, we conducted a GO Slim analysis of microRNA targets of the microRNAs identified in this study. Target analysis was carried out in the human genome, as this is the one with the most microRNA target annotation. The data was partitioned according to the phylogenetic position of microRNA families and divided according to the clades that show the most microRNA innovation. This analysis only includes microRNAs shared between humans and other species at particular points during the evolution of metazoans. MicroRNA families that appear to be specific to other clades (i.e. rodents) were excluded from this analysis. For a detailed explanation please see the methods section.

We found a significant bias in microRNA functionality for all the different clades evaluated (Figure 8). In general microRNAs appear to be enriched in Go Slim terms related to "multicellular organism development", "regulation of biological process", "cell differentiation", "anatomical structure morphogenesis", "cellular component organization", "protein modification process", "embryonic development", "cell proliferation", "transcription" and "cell - cell signaling". All these biological processes are related to one another in the sense that all of them are involved in the progression that under-specialized cells undergo in order to create specialized tissues and organs that will constitute multicellular organisms with organized anatomical structures.

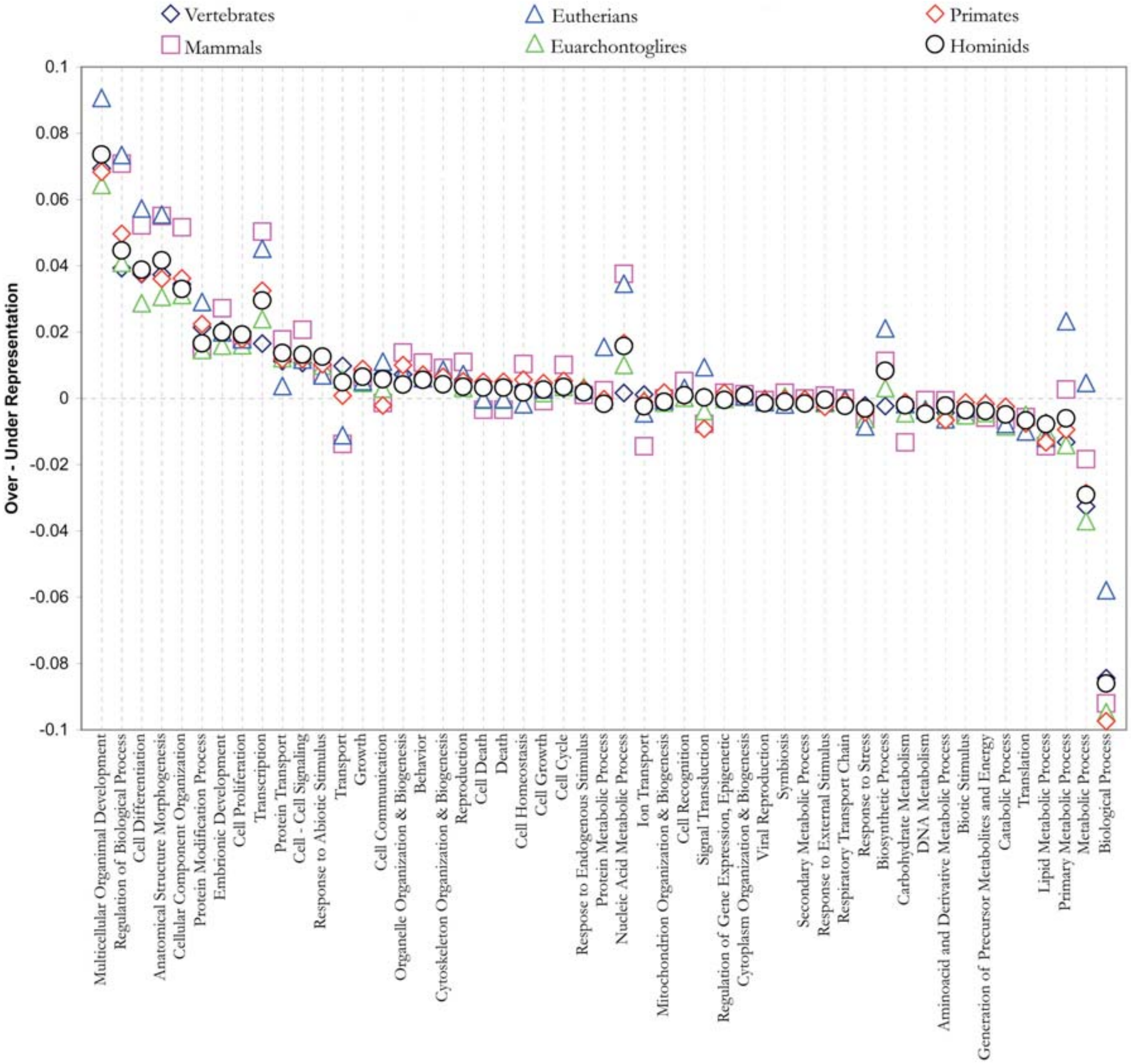


Figure 8. Functional bias for targets of microRNA families. Go Slim Biological Process categories. The magnitude of over or under representation (above or below 0) is represented in the Y axis. MicroRNA families appear to be enriched in functions related to morphogenesis, cell differentiation and multicellular organism development, and appear to be depleted in functions related to energy metabolism.

Interestingly, we also found microRNA targets to be under-represented in the following GO Slim categories: "carbohydrate metabolism", "DNA metabolism", "amino acid metabolism", "biotic stimulus", "energy", "catabolic process", "translation", "lipid metabolism", "primary metabolic process" and "metabolic process". All these biological processes are related to chemical reactions and pathways whose role is in the breakdown of carbon compounds coupled to processes that result in the liberation of energy to be used by the cell or organism. Most of these processes are related to electron transport and metabolic processes, and they are largely localized in the mitochondrion.

Discussion

Our results regarding the distribution of family sizes across metazoan genomes show a clear distinction among microRNA families common to vertebrate and invertebrates and those that are unique to vertebrate genomes. Despite the fact that the number of vertebrate-invertebrate microRNA families is small (~21) it appears that these families have been expanded in vertebrate genomes and supports the idea that there has been one or two rounds of whole genome duplication early in the evolution of vertebrates (Blomme et al 2006; Meyer and Schartl 1999; Panopoulou et al 2003). This phenomenon is not seen in microRNA families that are specific to vertebrate genomes, as most vertebrate-non-mammalian microRNA families are in a 1:1 relationship with their mammalian counterparts. This result, as opposed to protein coding genes, rules-out whole genome duplication as the principal mechanism for microRNA innovation and suggests the mechanism for the creation of new microRNA molecules is related to local duplications, perhaps duplications of inverted repeats as it has been suggested before for plant microRNAs (Allen et al 2004). The fact that only 21 microRNA families can be mapped around the time of whole genome duplication (WGD) events implies that the great majority of the microRNA families found in vertebrate genomes are innovations that occurred after WGD events.

Our results regarding the distribution of microRNA families throughout the evolution of metazoans strongly indicates a correlation (and perhaps causation) of microRNA innovations and organism complexity. In molecular biology, there are two puzzling paradoxes of organism complexity: the first one, the C-value paradox (Mattick 2007), refers to the fact that genome size does not correlate with genome complexity. The second one, the G-value paradox (Taft et al 2007), refers to the fact that the number of protein coding genes found in several organisms also does not correlate with organism complexity. A few studies on this issue have found that the fraction of non-protein DNA increases almost in a linear fashion from lower prokaryotes to higher eukaryotes. The linearity of this model suggest that non-coding DNA elements in genomes scale consistently with developmental complexity, and these elements contain large amounts of regulatory information (Taft et al 2007).

We found most microRNA innovations at key points during the evolution of metazoan species that led to humans. Most of these innovations appear to be specific to vertebrates, eutherians, primates and hominids. Interestingly, we didn't find such innovation for mammalian genomes, but we believe this is due to a sampling problem with the *Platypus* genome, either because genetic divergence didn't allow us to detect the microRNAs that should have been detected or because the genome assembly used for this study wasn't optimal for microRNA finding. We believe that a significant fraction of microRNA families that appear to be therian and eutherian innovations are indeed mammalian innovations. Similar problems have been found in the past by other groups sampling for microRNAs (Heimberg et al 2008). Their results indicate that northern blots and expression techniques are necessary in order to get a complete picture of microRNA distribution.

The bias found in GO Slim categories for microRNA targets goes in accordance to previous findings in microRNA functionality (Zhang et al 2007). MicroRNAs are heavily involved in processes that lead to cell differentiation and morphogenesis and seem to be depleted in functions related to energy metabolism. It is not clear why microRNAs seem to be depleted in the regulation of "energy metabolism" and in other processes like "cell homeostasis", "cell recognition" and "response to endogenous stimulus". A recent study in gene duplicability indicates that protein coding genes related to energy metabolism are less likely to be duplicated than genes in other Go Slim categories like "multicellular organismal development", "transcription" and "protein modification". Gene families related to energy metabolism have remained largely unchanged when it comes to family size since the divergence of vertebrates and invertebrates (Prachumwat and Li 2008). This is an indication that the machinery for energy metabolism was in place and working very well before the divergence of vertebrates and invertebrates, and post – transcriptional regulation of these pathways by microRNAs is not really necessary.

Metazoan microRNAs seem to be particularly underrepresented in functions related to cell adhesion, sensory reception of smell, immune response, defense response and response to stimulus. This result sets them apart from plant microRNAs whose functions tend to relate not only to cell differentiation and morphogenesis, but also are related to stress and defense responses (Chen 2005).

In general, all microRNA families for all clades considered are enriched in terms related to cell differentiation and morphogenesis. We didn't find major differences in GO Slim enrichment when different clades were considered. This finding suggest that *derived* microRNA innovations don't regulate *more* targets than more basal or *primitive* innovations, but suggest that *new* innovations pick up new genes not being regulated by more basal innovations. This mechanism avoids gene regulation overlaps and allows the generation of a fine tuned system able to generate enormous morphological diversity.

Materials and methods

Search strategy

Mature microRNA sequences are short (22nt on average) and their information content is quite low compared with their protein counterparts. DNA is composed by a 4-letter alphaber whereas proteins are composed by a 20-letter alphabet. The information per position in a DNA alphabet is in the order of 2 bits ($\log_2(4)=2$). In proteins information per position is in the order of ~5 bits ($\log_2(20)= 4.3$). A 22 nucleotide word has only 44 bits of information, whereas a 22 amino acid word has about ~110 bits of information. This makes the identification of microRNAs by BLAST searches alone a difficult task, particularly considering the problem of false positives. One might think that using precursors could solve this problem, but precursors preferentially have structural constraints, whereas mature microRNAs have functional constraints. Using BLAST to search with precursors would only get closely related microRNAs, missing the more divergent ones. Filtering the results on the basis of the evolutionary constraints exerted on microRNAs can solve this problem.

Since the functional evolutionary constraint is exerted on the mature sequence, one can take advantage of that and search for homologous microRNAs using mature sequences as queries with very sensitive but unspecific parameters (i.e. high evalue). This would basically guarantee a hit for every microRNA homolog present in the subject sequence, at the price of having lots of false positives (Dezulian et al 2006). Then the results can be filtered for structural constraints given a set of rules common to microRNA stem-loop precursors. For this work, we have developed a pipeline that uses this strategy, avoiding the problem of precursor sequences. The pipeline has been called microRNAscan and its process goes as follows (Figure 9):

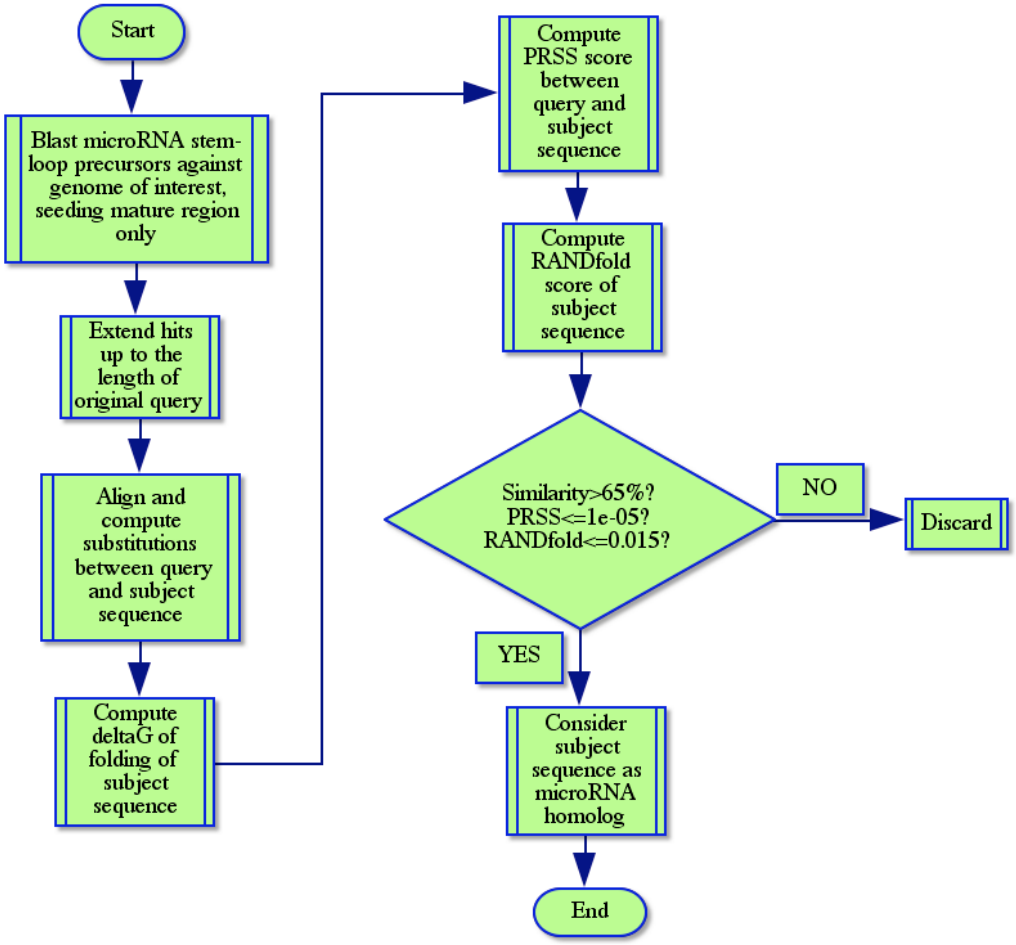

Start
Blast microRNA stem-loop precursors against genome of interest, seeding mature region only
Extend hits up to the length of original query
Align and compute substitutions between query and subject sequence
Compute deltaG of folding of subject sequence
Compute PRSS score between query and subject sequence
Compute RANDfold score of subject sequence
Similarity>65%? PRSS<=1e-05? RANDfold<=0.015?
NO
Discard
YES
Consider subject sequence as microRNA homolog
End

Figure 9. Flowchart of the method used to detect microRNA homologs.

1. Mature microRNAs and stem-loop precursors were downloaded from miRbase v10. Mature microRNAs and their respective precursors were combined into a single sequence with the mature region in lower case format.

2. Each microRNA sequence was aligned against the genome of interest using WUBLAST (Gish, 1996-2004) with an evalue of 10 and automated with the help of a perl script. BLAST jobs were performed by only seeding the mature region of the

microRNA. This step minimizes false positives resulting from unrelated sequences being similar to the stem-loop precursor. Seed extensions were allowed outside the mature region.

3. BLAST output was parsed and a sequence corresponding to each hit was extracted out of the genome being evaluated. This sequence was extended up to approximately the length of the original query if necessary.

4. A PRSS (Pearson 1990) analysis between the two sequences is performed in order to assess the statistical significance of the alignment and determine whether or not the two sequences are homologous. PRSS works by constructing a local alignment between the two sequences and calculating their alignment score. Then the second sequence is shuffled, a new alignment is constructed and its score computed. This process is repeated between 100 - 1000 times. At the end a distribution of scores is computed and the significance of the original alignment is assessed using the distribution of scores. For this particular step, 1000 shuffles were computed.

5. A RANDfold (Bonnet et al 2004) analysis of the subject sequence was performed in order to determine how likely the sequence resembles a microRNA. Most of the microRNAs found to this day are in a structural conformation corresponding to a free energy of folding that is considerable lower than that for shuffled sequences with the same nucleotide composition, indicating a tendency in the sequence towards a stable secondary structure. This conformation tendency is not seen frequently in other non-coding RNA molecules (Figure 10), but is common for microRNAs

(Figure 11). An analysis of the statistical significance of secondary structure conformation for microRNAs deposited in miRBase shows that the great majority of them follow this property, as most of them have a RANDfold scores significant at an alpha level of 0.05 (Figure 12). This very same behavior is expected for microRNA homologs. RANDfold was run with 1000 iterations per sequence, and the results were tabulated. This is the first time this strategy is used for identification of microRNAs.

6. In the filtering step, putative microRNA homologs were kept if the following conditions were met: Similarity score of at least 65% throughout the entire global alignment, free energy of folding of -20 Kcal/mol or lower, PRSS score equal or lower than 1e-05, RANDfold score equal or lower than 0.015. This filtering step was overrun if the percent identity was of 95% or higher.

7. Redundancy was addressed, because in many cases there were more than one microRNA per genomic position. This was particularly true for microRNAs that are known to have several paralogs, or from orthologous genes that are redundant in the database used (miRBase v.10). In order to minimize redundancy, the genomic positions of the putative microRNAs were compared in an all versus all fashion. All overlapping microRNAs were clustered and sequences within the cluster were scored based on similarity against the overlapping queries. Only the best pair was kept and used for further analyses.

8. Putative microRNAs were analyzed through Repeatmasker in order to minimize repetitive and transposable elements.

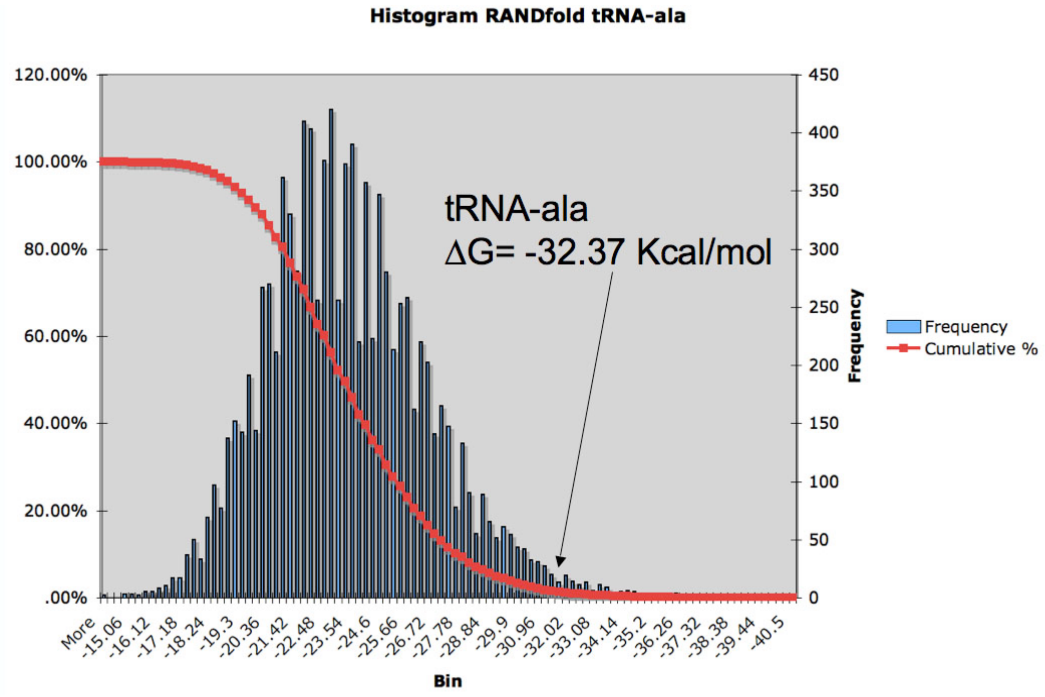


Figure 10. Distribution of RANDfold scores of a tRNA molecule. On average, tRNAs are not in a structural conformation corresponding to a free energy of folding that is considerable lower than that for shuffled sequences with the same nucleotide composition. The free energy of folding for this tRNA is NOT located at the very end of the distribution.

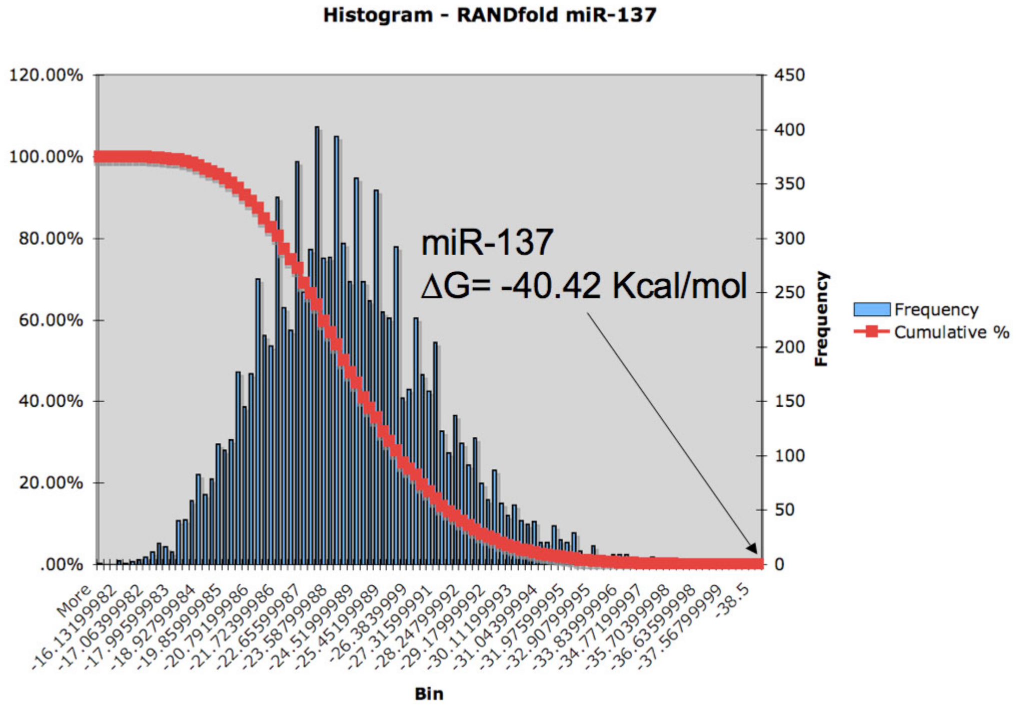


Figure 11. Distribution of RANDfold scores of a microRNA molecule. On average, microRNAs are in a structural conformation corresponding to a free energy of folding that is considerable lower than that for shuffled sequences with the same nucleotide composition. The true free energy of folding for this microRNA is at the very end of the distribution and is highly statistically significant.

Genomes used in this study

Aedes aegyptii (AaegL1), *Anopheles gambiae* (april 2004), *Apis mellifera* (Amel4.0), *Bombix mori* (ge2k), *Bos Taurus* (btau3.1), *Caenorhabditis elegans* (WB170.46), *Canis familiaris* (CanFam2.0), *Danio rerio* (ZFISH7.46), *Drosophila melanogaster* (dmel4.2.1), *Drosophila pseudobscura* (dpse_r2.0), *Equus caballus* (EquCab1), *Homo sapiens* (HSA46), *Macaca mulata* (Mmul_051212), *Monodelphis domestica* (MonDom5), *Mus musculus* (C57BL/6J), *Nasonia vitripennis* (nvit1.0), *Ornithorhynchus anatinus* (IOANA5.46), *Pan* *troglodytes* (CCY5Cv1), *Pediculus humanus* (PhumU1), *Rattus norvegicus* (RG5C3.4), *Takifugu rubripes* (FUGU4.46), *Tribolium castaneum* (JGI4.1.46).

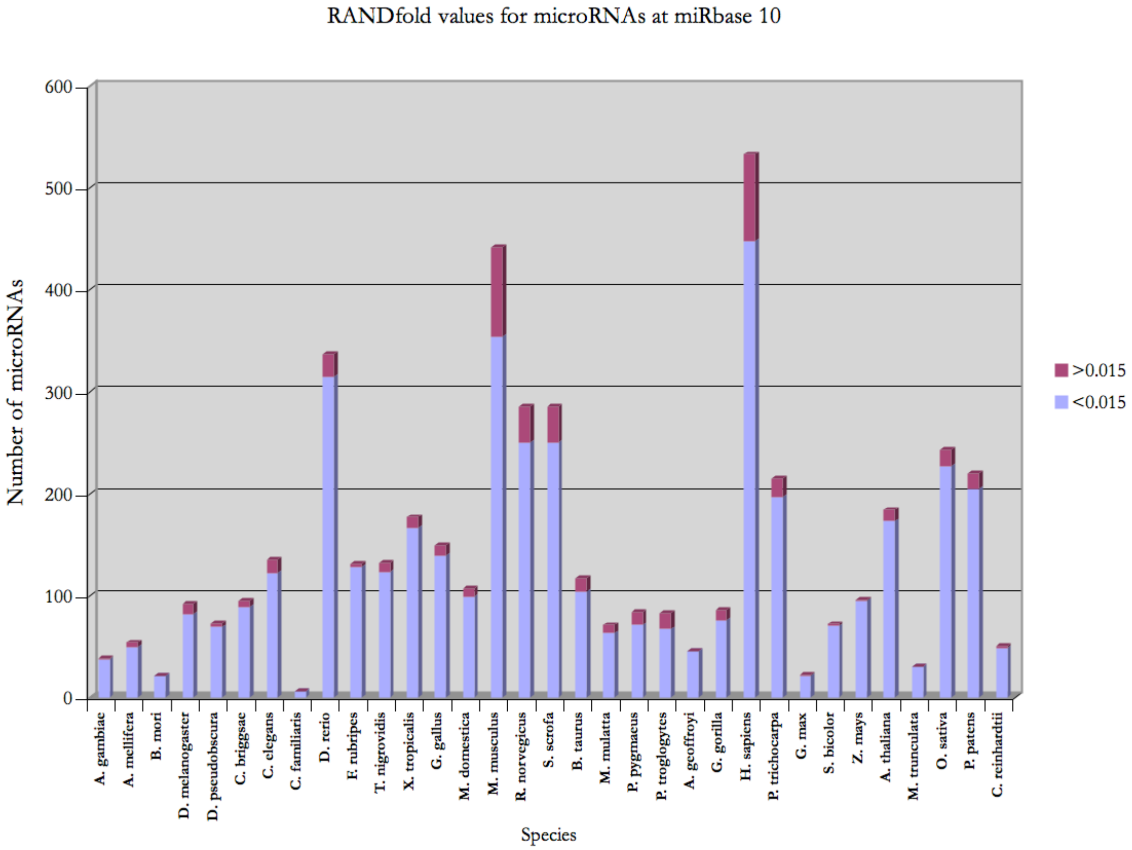


Figure 12. RANDfold values of validated microRNAs from miRBase. Most of the validated microRNAs stored in databases are in a structural conformation corresponding to a free energy of folding that is considerable lower than that for shuffled sequences with the same nucleotide composition, indicating a tendency in the sequence towards a stable secondary structure. More than 90% of the microRNAs from MirBase have a Randfold score equal or less than 0.015, which was the cutoff for the homology search pipeline.

Clustering of microRNA into families

Putative microRNAs from all species were combined into a single FASTA list. Then an all versus all comparison was performed using PRSS with 1000 shuffles. The p-values obtained from each pair wise comparison were used as a basis to cluster the microRNAs. This algorithm is similar to other single-linkage clustering methods used for sequences (i.e. blastclust). In this case the PRSS p-value was used as linkage criterion. In order to determine the best threshold, clustering performance was evaluated on the basis of the let-7 family. This family is known for having several paralogs that should be clustered into the same gene family. The p-value scores that clustered all the members into a single gene family was the criterion used to cluster all the other microRNA families. P-values equal or higher than 1e-06 were appropriate to cluster all the members of let-7 into a single family.

Mapping gene families into metazoan phylogeny

A presence/absence matrix that includes all the resulting microRNA gene families and all the species used was constructed. A "1" was placed in position if the gene family was present in the species under scrutiny, a "0" was placed in the same position if the family was absent from the species under scrutiny. This matrix was imported into MACCLADE (Maddison and Maddison 1989), and all characters (presence/absence of family) were mapped into the phylogeny.

Gene ontologies and functional evaluation of microRNAs throughout the metazoan phylogeny

The functions of microRNAs were assessed on the basis of the genes they regulate (their targets). For this objective microRNA target information was downloaded from miRBase (Griffiths-Jones et al 2008) and the TargetScanS website (Grimson et al 2007). Other target prediction programs were considered, but TargetScanS seems to be the best in terms of both sensitivity and specificity (Martin et al 2007). Both MirBase and TargetScanS contain information from computationally predicted and experimentally validated targets. The information was collected for the following organisms: *Homo sapiens*, *Pan troglodytes*, *Canis familiaris*, *Mus musculus, Rattus norvegicus* and *Drosophila melanogaster*.

In order to determine microRNA functions at different points in the phylogeny a representative species was selected and only the microRNAs that were present in the most recent ancestor of that branch were evaluated for the species under scrutiny. For example, in order to evaluate the microRNA functions common to all eutherian mammals, we first determined the microRNAs that are common to all of them and then evaluated the targets of this set for orthologous genes between Human and Cow as these represents the gene set present in the common ancestor between these two organisms.

All analyses were centered on Human. Similar procedures were followed for other points in throughout the phylogeny as follows:

Orthology sets:

Coelomata:	*Homo sapiens – Drosophila melanogaster.*
Amniota:	*Homo sapiens – Gallus gallus.*
Vertebrata:	*Homo sapiens – Danio rerio.*
Theria:	*Homo sapiens – Monodelphis domestica.*
Eutheria:	*Homo sapiens – Bos taurus.*
Supraprimates:	*Homo sapiens – Mus musculus.*
Primates:	*Homo sapiens – Macaca mulatta.*
Hominidae:	*Homo sapiens – Pan troglodytes.*

Gene ontologies were determined for each orthologous set and enrichments were determined using GeneMerge (Castillo-Davis and Hartl 2003).

Gene ontologies and functional evaluation of microRNAs in the human genome

Functions of microRNAs were assessed on the basis of the genes that they regulate. Targets for human microRNA genes were collected from TargetScanS and the data was partitioned according to the phylogenetic position of each microRNA. Functional bias was is represented as a function of the proportion of genes that are associated with a function with respect to all human genes having gene ontology terms. The statistical significance of the functional bias was assessed by the X2 test with a false discovery rate (FDR) of 0.05. The FDR value was obtained with the QVALUE software library (Storey and Tibshirani 2003).

CHAPTER III

COMPUTATIONAL IDENTIFICATION OF NEW MicroRNAs BY COMPARATIVE GENOMICS

Introduction

Identification of new microRNAs in new species relies heavily on a combination of computational and experimental approaches. The first attempts to develop experimental strategies for the identification of microRNAs relied on size fractionation of total extracts of RNA, cloning and then sequencing. To be considered microRNA, sequenced RNA fragments have to be cloned multiple times and the clone length must be between 20 and 24 nt long. This methodology is cumbersome and time consuming and only allows the identification of abundant microRNAs in abundant tissues. Low copy microRNAs in non-abundant tissues usually escape detection. In other words, cloning screenings are far from being saturated as they are biased towards abundant microRNAs.

To overcome these problems, computational biologists have developed different strategies aimed at the identification of putative microRNA encoding sequences that can be validated using traditional experimental approaches. Most of these strategies rely on the fact that microRNAs are conserved throughout metazoan evolution, and by comparing genomic sequences between different species conserved microRNAs can be detected.

Computational detection of microRNAs

Computational approaches usually rely on the fact that evolutionary constraints are an indication of functional or structural constraints. MicroRNAs are derived from precursor transcripts with an extended stem-loop structure, are usually conserved between genomes of related species and display a pattern characteristic of evolutionary divergence (Thomassen et al 2006). This makes them suitable for computational identification.

MirScan (Limet al 2003) was one of the first tools developed to identify microRNA genes. It was successfully applied to the genomes of *C. elegans* and *C. briggsae*. MirScan scans genome "A" for sequences that can form stem-loop precursors, and then checks whether or not these sequences are conserved in genome "B". The resulting sequences are then compared against a "training set" of experimentally validated microRNAs, and then, based on these comparisons, microRNA candidates are given a score. Best scoring candidates are then labeled as putative microRNAs. Using this methodology, Lim et al found about 35 microRNAs in *C. elegans*, 19 of which have been experimentally validated.

Similar approach is used by miRseeker (Lai et al 2003). A whole genome alignment between *D. melanogaster* and *D. pseudobscura* is used as a basis to detect phylogenetically conserved hairpin precursors. A sliding window of about 100 nt is used to systematically score every piece of the whole genome alignment. Features like conservation percentage, gap percentage and free energy of folding are used to detect potential microRNAs. miRseeker requires good candidates to have a free energy of folding of at least -23 kcal/mol and a minimum arm length of 23nt. miRseeker was able to identify 208 microRNA candidates, 38 of those were subject to experimental verification and 24 turned out to be real.

Using phylogenetic shadowing, Berezikov et al. identified and sequenced 122 new microRNAs in 10 different primate species. Phylogenetic shadowing starts with a multiple whole sequence alignment of a set of related organisms. Then a nucleotide conservation graphic is generated from the alignment. Highly conserved regions are seen as peaks, whereas low conserved regions are seen as valleys. Primate microRNAs have a high conservation peak corresponding to the stem-loop precursor, accompanied by a sudden drop in conservation for the surrounding sequence. This characteristic conservation pattern was used to identify novel microRNAs (Berezikov et al 2005).

Computational methods that do not rely on conservation for microRNA identification also have been developed, although less successfully. In general, it is thought that secondary structure by itself is not a good measure to identify non-coding RNAs (Rivas and Eddy 2000); but when combined with features of known microRNAs like size of loops, helices, CG composition, length, and deltaG of folding, a good identification tool can be constructed. A study aimed at the identification of microRNAs in viral genomes used the concept of Support Vector Machines (SVMs) to identify microRNAs. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. In an SVM, features of the object to be classified are mapped into a vector of n-dimensional space. This transforms the space of the original object from linear mapping to non-linear mapping. The linear model constructed in the new space can represent a nonlinear decision boundary (Witten and Frank 2000). In other words, classification can be fine-tuned given the fact that the data is now in a multidimensional space with as many parameters given by the user. In the viral microRNA study, a SVM was used to classify all genomic regions able to form stem-loop precursors, identifying previously known microRNAs and new ones with a low rate of false positives (Pfeffer et al 2005).

A study conducted in humans identified 45 new microRNAs by homology searching using mature and stem-loop precursors from mouse. The study highlights the importance of signals coming from secondary structures and compensatory mutations to identify microRNAs (Weber 2005).

The computational pipeline used for this work started as an observation made in the past by scientists who identified a particular property of microRNAs in the context of comparative genomics (Glazov et al 2005; Tran et al 2006). This observation suggested that microRNAs are enriched in ultraconserved elements (UCE). That is, at close to moderate phylogenetic distance, microRNAs have runs of absolute conservation with no substitutions, insertions or deletions. In this study we developed a novel method to systematically evaluate and score homologous regions between the genomes of honey bee (*Apis mellifera*) and parasitic wasp (*Nasonia vitrippenis*) with the purpose of identifying new microRNA genes on the basis of ultraconservation. This is the first time ultraconservation is used in a systematic way to identify *new* microRNA genes by comparative genomics. Our strategy uses ultraconservation as first step, but the identification of putative microRNA genes relies on the careful scoring of candidate pairs in order to maximize microRNA finding.

Results

Analysis of previously validated microRNAs shared between Apis *and* Nasonia *(Experimental Set)*

Using a computational approach devoted to the identification of microRNA homologs from databases, we were able to identify a set of 45 microRNAs common between honey bee and wasp (Appendix A). We refer to this group of microRNAs as our "Experimental Set" and it was used as a basis to develop the computational method aimed at the identification of new microRNAs shared between these two genomes.

Our analyses started with a conservation analysis of homologous genes from the Experimental Set (Figure 13, Appendix A). MicroRNAs shared between honey bee and wasp show different degrees of conservation. The microRNA MiR-133 has two long ultraconserved elements (UCEs) that span the entire length of each arm and part of the terminal loop, with only one substitution in the terminal loop. MiR-2-2 has a short (20nt) UCE that overlaps the mature region. This absolute conservation of mature regions in microRNAs of species that have diverged about ~120 mya is an indication that the biological roles performed by these RNA molecules are also conserved.

```
ame-mir-133 & nvi-mir-133
TAATGTTAAGCTTAGCTGGTTGAACACGGGTCAAATATATCGCACGATTGACGCATttggtccccttcaaccagctgtAGTTGACATTA
TAATGTTAAGCTTAGCTGGTTGAACACGGGTCAAATATAACGCACGATTGACGCATttggtccccttcaaccagctgtAGTTGACATTA
*************************************** *************************************************

ame-mir-2-2 & nvi-mir-2-2
TCGACTGTTCCTCCCATCAGAGTGGTTGTGATGTGGTA-ACTTGGACTCGtatcacagccagctttgatgagcGGAACGGTGCGA
TCGGCTGTTTCGCCCGTCAGAGTGGTTGTGATATGGTGCTATTGAACGCAtatcacagccagctttgatgtgcGTAACAGTTCGA
*** ***** * *** **************** ****    *** ** * ******************** *** *** ** ***
```

Figure 13. Ultraconserved elements located within microRNAs. The UCEs span the region that corresponds to the mature region. Runs of ultraconservation are represented in yellow and mature microRNAs in lowercase.

Percent identities of genes within the Experimental Set range between 100% for microRNAs such as mirR-iab-4 and 66% for microRNAs such as miR-279. The absolute conservation of miR-iab-4 is the exception to the rule, as most of the microRNAs have percent identities between 75 – 95%, indicating some extent of divergence within the Hymenoptera. Most of the substitutions found are located along the terminal loop and in the stem regions that do not correspond to the mature microRNA. The mature microRNA is the region with the highest degree of conservation (Figure 14). This is certainly an indication of functional evolutionary constraint exerted upon the mature region. We used this property as part of our methodology for the identification of new microRNAs shared between honey bee and wasp.

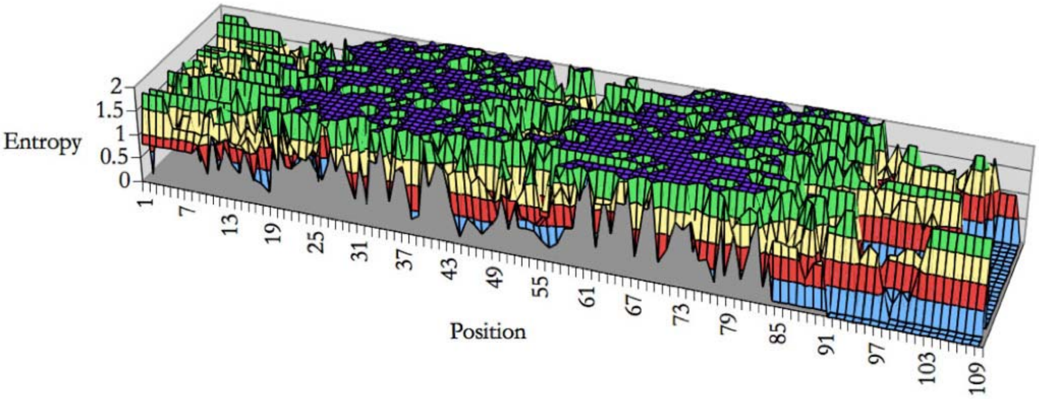


Figure 14. Entropy conservation plot for 25 different microRNA families. The horizontal line represents the nucleotide position along the microRNA precursor. Vertical lines represent the amount of conservation in bits, up to a maximum of 2, which represents complete conservation. Regions of maximal conservation correspond to the mature region of the microRNA precursor and the mature* region. Variable regions are seen as valleys. Variability is higher in the regions corresponding to the terminal loop and in regions of the stem that do not correspond to the mature region. This suggests nucleotide conservation for variable regions is not as important as conservation of secondary structure for microRNA function.

Patterns of nucleotide substitution of microRNAs from the Experimental Set

The mutation patterns of microRNAs were classified according to the rules proposed by (Lai et al 2003). Mutation patterns imply a canonical progression in microRNA evolution (Figure 15). In general microRNAs start accumulating mutations preferentially in the terminal loop, and then mutations accumulate along the stem. According to this model it is unlikely for a microRNA to accumulate mutations along the stem having no mutations in the terminal loop. The pattern of nucleotide substitution for each microRNA class is as follows:

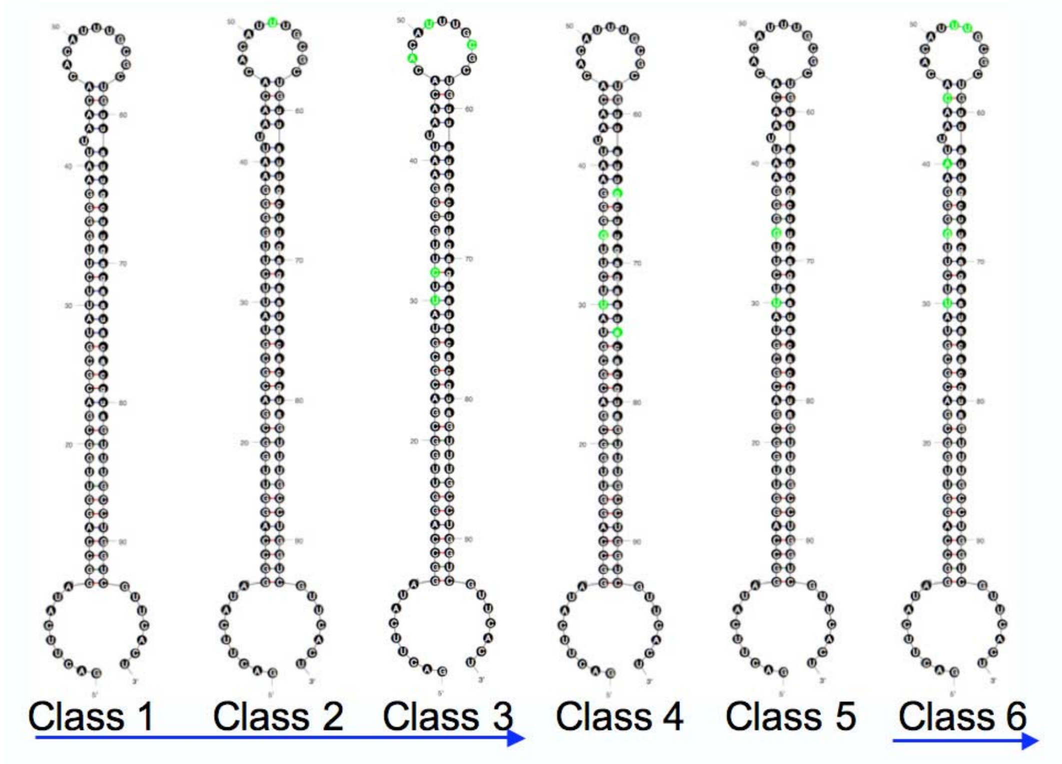


Figure 15. Proposed evolutionary model for conserved microRNAs from Lai et al. (2003). Class 1, 2 and 3 represent the typical progression in the molecular evolution of microRNAs from a state of complete conservation. Class 1, 2, 3, and 6 are considered good microRNA candidates.

Class 1: No substitutions. Complete conservation.

Class 2: One or more substitutions or gaps contained exclusively in the terminal loop.

Class 3: Equal or greater number of mutations within the loop compared to the non-microRNA-encoding arm.

Class 4: Substitutions along both arms and no substitutions in the terminal loop.

Class 5: Substitutions along one arm only and conservation in the terminal loop.

Class 6: Substitutions along one arm greater than substitutions in the terminal loop.

Of the 45 microRNA pairs evaluated, we found one belonging to Class 1, five belonging to Class 2, twenty eight belonging to Class 3, one belonging to Class 4, two belonging to Class 5 and six belonging to class 6 (Figure 16). MicroRNAs from classes 4 and 5 are rare and their frequency of appearance supports the idea that there is a negative selective pressure along both arms of the stem loop, and only mild selective pressure along the terminal loop. For our computational method, we considered sequences from classes 1, 2, 3 and 6 as good microRNA candidates. Classes 4 and 5 were considered poor candidates.

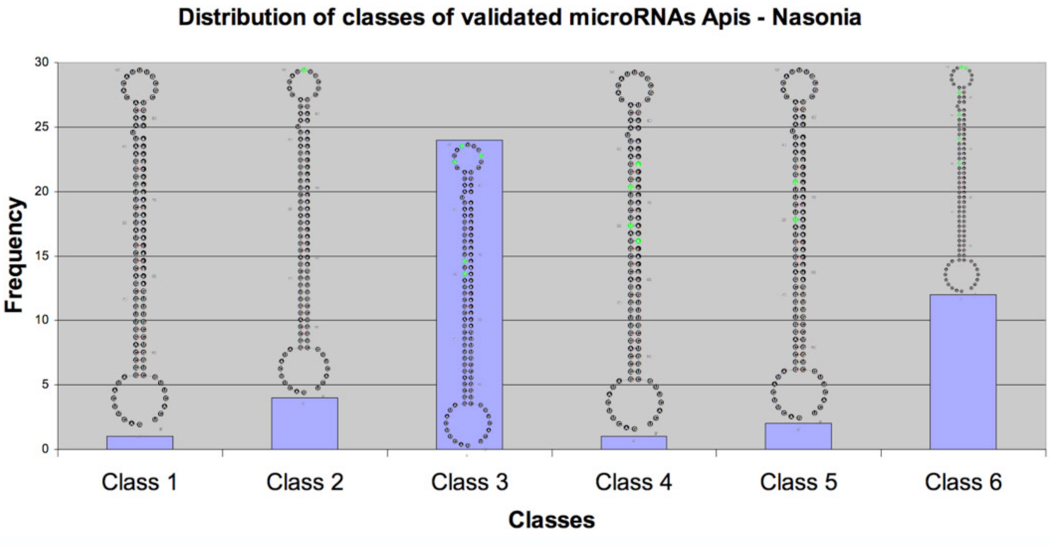


Figure 16. Distribution of mutational pattern classes among validated microRNAs at the honey bee – wasp intersection. Most of the microRNAs belong to classes 3 and 6, which represent the most common mutational pattern of microRNAs.

Computational identification of microRNAs

Our computational strategy started with the identification of ultraconserved DNA elements between the honey bee and wasp genomes. For this purpose we used the WUBLAST (Gish, W. 1996-2004) package with modified parameters suited for the identification of ultraconserved elements.

The first scanning of both genomes identified 294,196 non redundant UCEs. The great majority of these UCEs are short sequences of 20 nt on average and they may come from homologous regions between the two genomes or from non-homologous regions as well. In order to discard the UCEs coming from non homologous regions, the UCEs were extended in honey bee in both directions by 65 nucleotides. These extended sequences were then blasted against the wasp genome using with an evalue of 1, resulting in 9,711 sequences and a 30-fold reduction of the number sequences with of potential of being microRNAs.

Potential candidate sequences were mapped to the honey bee genome in order to minimize redundancy. Redundant sequences mapping to the same chromosomal position were grouped into one single candidate. This step reduced the candidate set to 6,324 sequences.

Filtering for microRNAs

Comparisons of these results against RNA molecules from databases shows that our computational strategy indeed identified microRNAs, but also showed other RNA molecules different to microRNAs. Among these results we found snoRNA genes, particularly U2, U6 and U1. We also found a significant fraction of the results as being related to repeat elements, particularly Non-LTR elements belonging to TAHRE family (Telomere-Associated and HeT-A-Related Element), which was recently described for *Drosophila* (Abad et al 2004; Shpiz et al 2007) and interspersed repeats. All these molecules have in common the capability of adopting secondary structures resembling microRNA stem-loop precursors (Figure 17). These sequences were not considered for further analysis and were discarded.

Candidate sequence pairs were aligned and evaluated for microRNA canonical structures. Any sequence pair not resembling a microRNA was discarded at this step. This filter removed a significant fraction of sequence pairs, resulting in 999 sequence pairs for scoring.

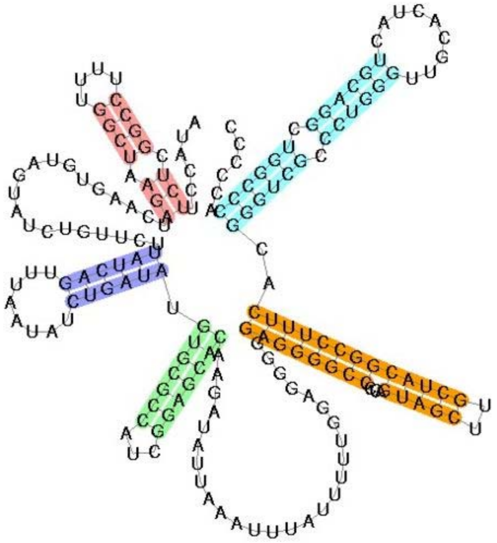


Figure 17. Secondary structure of a U2 spliceosomal RNA. Parts of this RNA are structurally similar to microRNAs (i.e. light blue arm). From RFAM (2008).

Scoring sequence-pairs

The set of 999 candidate sequence-pairs was processed using the using the Smith – Waterman algorithm as implemented in the program Water from the EMBOSS package (Olson 2002). Alignments were scored and normalized based on length. They also were parsed and given as input to RNAalifold (Hofacker 2003), in order to evaluate their secondary structure and free energy of folding. Evaluation of candidate structures took into consideration the physical resemblance to canonical stem – loop precursors and free energy of folding. Statistical significance of folding was assessed by with RANDfold and 1000 randomizations. Candidate pairs were scored on the basis of redundancy, similarity, free energy of folding, and statistical significance of the fold (see methods for details).

Recovering experimentally validated microRNAs

Our computational strategy was able to recover 41 of the 45 microRNAs from the Experimental Set. The distribution of scores of candidate sequence-pairs shows that the great majority of the microRNAs from the Experimental Set score very high in our computational model, as 40 of the 41 recovered are within the top 200 (Figures 18 and 19). This is a very good indication that putative microRNAs scoring high in our computational model may correspond to genuine microRNAs. Only four microRNAs from the experimental set weren't recovered by our computational strategy. These correspond to miR-124, miR-190, miR-279 and miR-307.

Only one of the recovered microRNAs from the Experimental Set wasn't within the top 200 set. Let-7 scored relatively low on our computational strategy with a score 7.41 and a rank of 324. The low score may be related to the fact that this microRNA is AT rich, making the deltaG of folding not as significant as other microRNAs that are not AT rich. The next low scoring microRNA was miR-375 with a score of 8.41 and a rank of 195, scoring high enough to be included within the top 200 set.

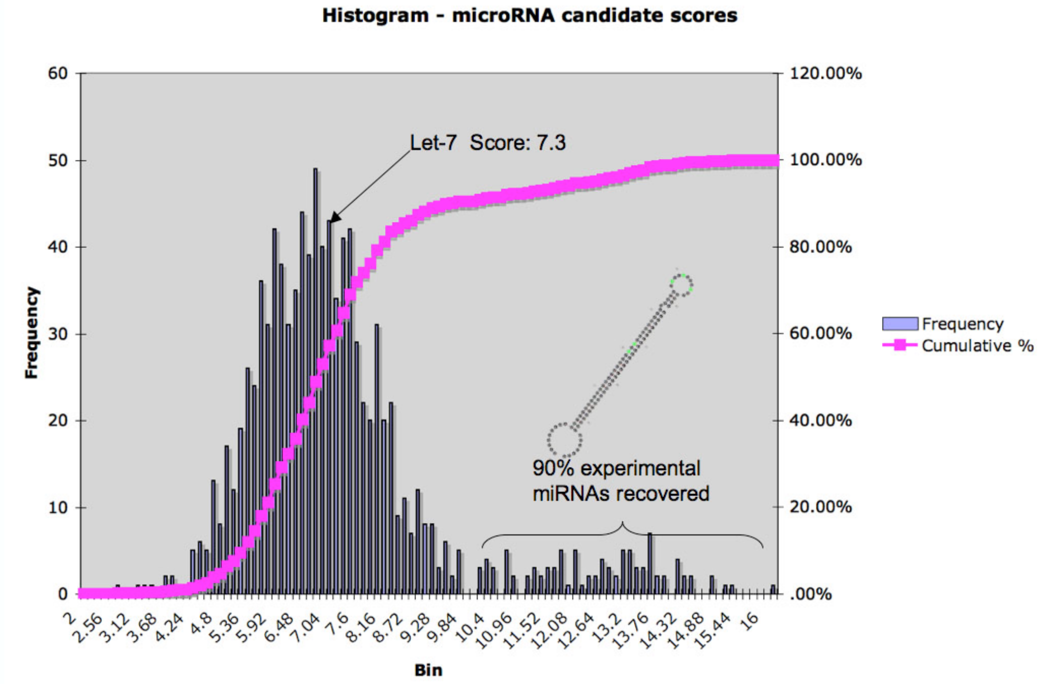


Figure 18. Distribution of scores of microRNA candidates. The great majority of the microRNAs from the Experimental Set score very high on our computational pipeline with scores higher than 10.

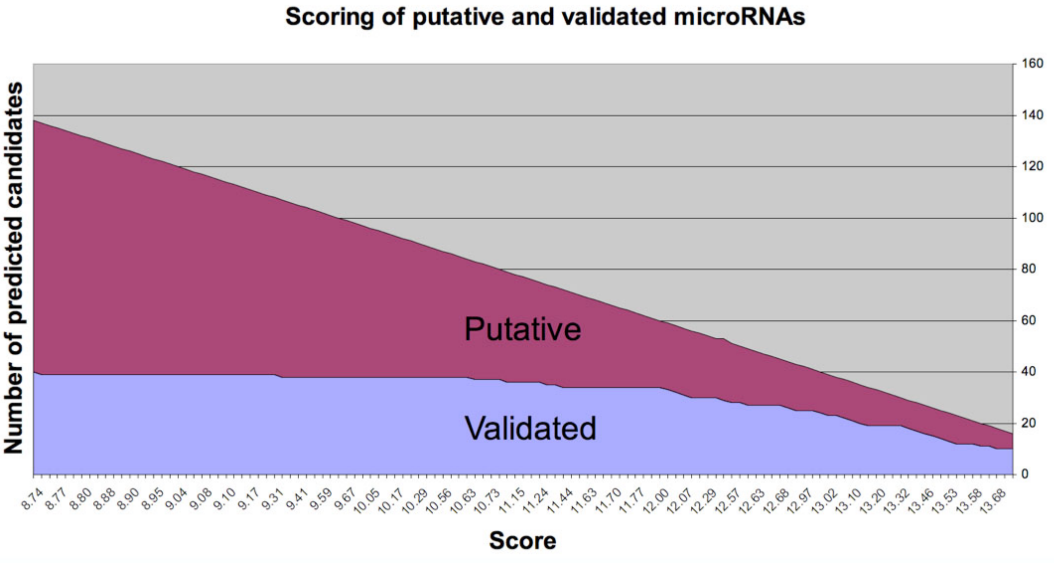
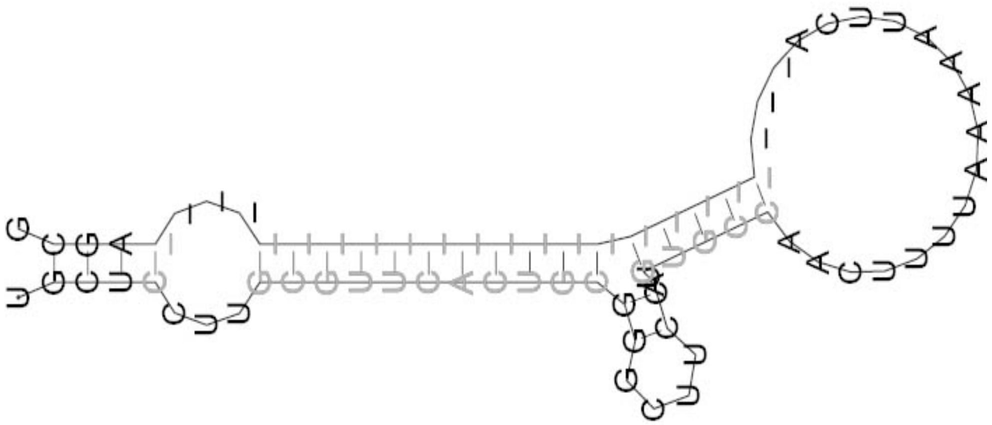


Figure 19. Recovering of experimentally validated microRNAs. Putative and real microRNAs at different scoring cutoffs. At a score of 10, we are able to recover 90% of the known microRNAs shared between wasp and honey bee, and 57 new predictions.

MicroRNAs from Experimental Set that were not recovered

Four of the microRNAs from the Experimental Set weren't identified in our computation. These are mir-124, mir-190, mir-279 and mir-307. Mir-124 is common to invertebrates and vertebrates and it seems to be the only microRNA common between honey bee and wasp that does not adopt a canonical stem-loop structure on its lowest energy form. This result is very interesting, given the fact that there are only two substitutions between the honey bee / wasp version of mir-124, both substitutions located at the base of the terminal loop (Figure 20).



```
TGCTCCTTGCGTTCACTGCGGGCTTCCATGTGCCAACTTTTCAAAATTCAtaaggcacgcggtgaatgccaagAGCG
TGCTCCTTGCGTTCACTGCGGGCTTCCATGTGCCAAGTTTTAAAAATTCAtaaggcacgcggtgaatgccaagAGCG
************************************ **** ***********************************
.((((...((((((((((((....))..(((((.....................)))))))))))))))...)))).
```

Figure 20. Consensus structure of honey bee - wasp mir-124. This microRNA does not adopt the canonical stem-loop precursor structure seen in other microRNAs.

Mir-124 adopts the canonical structure of a stem-loop precursor at a suboptimal free energy of folding, suggesting the microRNA exists in different conformations in the cellular environment. Since our current computational approach requires a canonical secondary structure, any microRNA not adopting a canonical stem-loop precursor would not be identified. Mir-190, mir-279 and mir-307 were identified as part of the ultraconserved set but our current parameters for extension and structural filtering discarded them at an earlier step. Future improvements on our current computational pipeline will address these issues.

Evaluation of microRNA candidates

The mutation pattern of the candidates was evaluated and inspected visually for stem-loop precursor characteristics. We focused on candidates of classes 1,2,3 and 6 as these classes represent the substitution pattern seen in the great majority of known microRNAs. Candidates were also partitioned according to their position with respect to protein coding genes in the genome. This step allowed us to determine whether the microRNA candidate was located in intergenic regions, splice sites, exons or introns.

Our results indicate that there is a tendency of microRNA candidates to localize within introns or intergenic regions, as their score gets incremented (Figure 21). In other words, the higher the score, the more concentrated the candidates seem to be in intronic and intergenic regions, at the expense of exons and splice sites. All of the microRNAs from the Experimental Set overlap either intergenic regions of introns. None of them overlap splice sites or exons. High scoring microRNA candidates also follow this tendency. This result is similar to what has been found in other insect genomes, particularly *Drosophila melanogaster* (Stark et al 2007).

Since known microRNAs in the honey bee genome are strictly localized in introns of protein coding genes, or as independent units, we believe our method successfully identified genuine microRNAs.

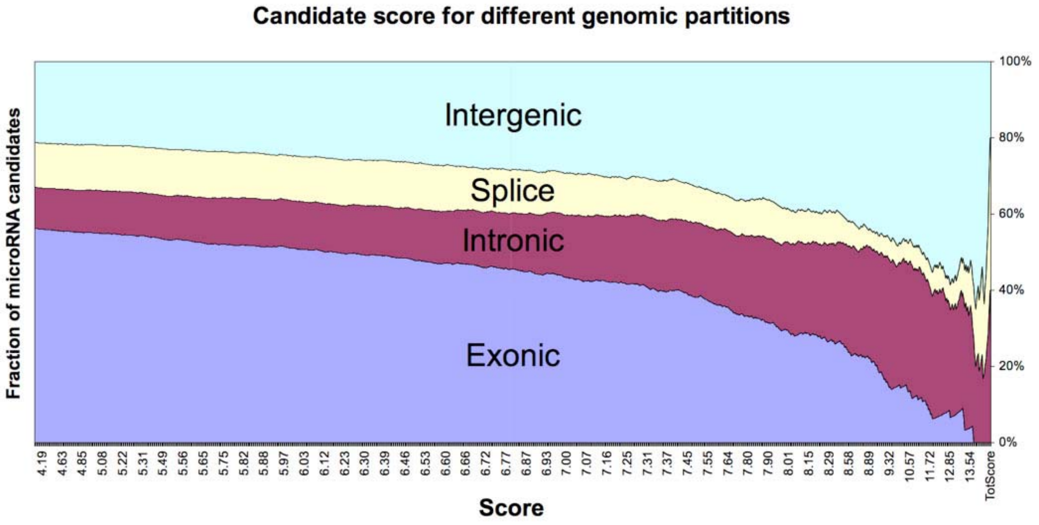


Figure 21. High scoring MicroRNA candidates tend to localize in introns and intergenic regions. High scoring predictions exclude exons and partially exclude splice sites.

MicroRNA predictions are supported by 454 data

Candidate sequence pairs were compared against 454 data coming from pools of RNA extracted independently from both honey bee and wasp. The 454 data was generated from the small RNA fraction of both organisms. Pools were constructed from different developmental conditions (i.e. larva, adult) and from different castes (i.e. queen, worker, drone). These comparisons give support to several of our predictions. These data allowed us to get an idea of the location of the mature region for several of our computational predictions. We found 28 of the 45 microRNAs from the Experimental Set as expressed sequences in the 454 data, and found support for 13 new microRNAs. Interestingly, most of the microRNA candidates supported by 454 data correspond to high scoring microRNAs according to our computational strategy, as 11 of them are within the top 200 set. The lowest scoring candidates supported by 454 data have ranks of 232 and 257 respectively. All the microRNA candidates supported by 454 data are located within introns or intergenic regions, and all the 454 matching regions are located within the stem of the candidate (Appendix B). This result strongly indicates these are real microRNAs

We only found three microRNA candidates that match 454 data in regions NOT corresponding to where mature microRNAs should be located. Matches for these tree candidates occur within the region corresponding to the terminal loop or regions with large bulges. One of the candidates is located at medium rank in our strategy (position 289) and the other two are very low rank candidates (positions 574 and 713 respectively, Appendix D). We didn't consider these candidates as real, given the position of the matches and the fact that they scored low in our strategy.

Functionality of microRNA candidates

At least 17 microRNAs from the Experimental Set are located within introns of protein coding genes. The gene that contains the most intronic microRNAs is GB-15727. This gene corresponds to a serine/threonine phosphatase that appears to have been lost from *Drosophila*, but it is present in both vertebrate and more ancient metazoan species. This gene contains 5 microRNAs from the experimental set (mir-2-1, mir-2-2, mir-2-3, mir-71, mir-13a) and a new candidate from the top 20 set (candidate No. 22). GB16497 is another gene that appears to contain at least three intronic microRNAs, two from the experimental set (mir-12, mir-283) and one from the top 20 set (candidate No. 48). This gene is a homolog of *Drosophila* CG33206-PA (dGMAP). In *Drosophila*, the overexpression of this protein blocks anterograde and retrograde transport between the endoplasmic reticulum and the Golgi apparatus.

We evaluated the functions of the best 20 intronic microRNA candidates belonging to classes 1, 2, 3, and 6. This set of 20 resides within 18 different genes in the honey bee genome. One of these genes, GB16072 is a homolog of the Iron regulatory protein 1B CG6342-PA. The protein it contains an aconitase-like domain with aconitase-hydratase activity. In the TCA cycle, Aconitase catalyzes the conversion of citrate to isocitrate via *cis*-aconitate. The gene contains two microRNAs from the top 20 set (candidate No. 67 and No. 106).

GB10180 corresponds to the *Drosophila* homolog of Scratch. This is a zinc finger protein involved in dendrite development (Roark et al 1995). GB10650 is similar to the Hormone receptor-like in 46 (Hr46 or DHR3). This protein is involved in metamorphosis and mushroom body development. Hr46 is known as a regulator that acts on a negative fashion on the Ecdysone receptor. In *Drosophila*, expression of Hr46 occurs after the expression of ecdysone, but before the expression of late hormones involved in metamorphosis. Expression of Hr46 is required for the transition between prepupal and pupal stages (Ren et al 2005). GB12790 is the homolog of rgk1 in *Drosophila*, a GTP mediated signal transduction protein. GB14007 corresponds to FucTA, which is a phospholipase A2 glycoprotein well known as a bee-venom allergen. Expression of FucTA was recently evaluated in honey bees (Rendic et al 2007). The gene is predominantly expressed in brain tissue and venom glands, with weaker expression in other tissues. GB13919 corresponds to muscleblind, a gene that appears to be involved in alternative splicing and that, in humans, is implicated in myotonic dystrophy, a form of muscular dystrophy that is characterized by muscle weakness and myotonia (slow relaxation of muscle after contraction) (Ho et al 2004). Muscleblind harbors the longest ultraconserved element found between honey bee and wasp (see Chapter IV). GB15055 correspond to VHDL (Very High Density Protein). This is a larval-specific lipoprotein of high density that belongs to the family of Vitellogenins and works as storage protein. GB16274 is homolog to Mmp2 (Matrix metalloproteinase 2). In *Drosophila*, this protein is involved in motor axon fasciculation (Miller et al 2008). GB19979 is a glutamic decarboxylase homolog to *Drosophila*'s *Gad1* gene. Proper expression of this gene is necessary for the correct development of synapse junctions between neurons or synaptogenesis (Featherstone et al 2002).

Table 4. Filtering of candidate microRNA sequences

Step	Number of resulting sequences	454 Matches
UCE scanning	294,196	521
Extend, then blast	9,710	45
Genomic Clustering	6,324	45
Remove snoRNA and repeat elements	6154	45
Evaluation of secondary structure	999	13
Substitution classification (Classes 1, 2, 3, 6)	686	13

Table 5. Genes harboring novel and previously known intronic microRNAs

Gene	Description	Intronic microRNA candidate #
GB10191	Similar to Rpb8, RNA polymerase II core complex	Mir-277, mir-317, mir-34
GB18684	ABC-type phosphate/phosphonate transport system, periplasmic component. Inorganic ion transport and metabolism.	Mir-278
GB12486	LOC724926 similar to DNA polymerase 73kD CG5923-PA, isoform A	Mir-279
GB16086	LOC413964 similar to grapes CG17161-PA, mitotic control.	Mir-79
GB10038	Similar to Histone-lysine N-methyltransferase, (Nuclear receptor binding SET domain containing protein)	Mir-8
GB11212	LOC410951 similar to CG32062-PD, isoform D, DNA binding.	Mir-925
GB15597	Similar to *eag* ether a go-go. Voltage-gated potassium channel activity;	Mir-928
GB17673	Similar to rhea, cytoskeletal protein concentrated at regions of cell–substratum contact.	Mir-190
GB14516	Similar to Distal-less. First genetic signal for limb formation to occur in the developing zygote.	Mir-930
GB10066	Similar to neuroligin. Post-synaptic adhesion molecule.	Mir-932
GB15727	Similar to Serine/threonine-protein phosphatase 4 regulatory subunit 1-like	Mir-13a, mir-2-1, mir-2-2, mir-2-3, mir-71, candidate 22
GB16497	Similar to lethal CG33206-PA. Unknown function.	Mir-12, mir-283, candidate 48
GB15055	Vhdl larval-specific very high density lipoprotein. Storage protein.	Candidate 62
GB16072	Similar to Iron regulatory protein 1B CG6342-PA. Aconitase.	Candidate 67, 106
GB16572	Similar to ANF-receptor	Candidate 77
GB12790	Similar to Rgk1 CG9811-PA. GTPase mediated signal transduction.	Candidate 83
GB14007	FucTA core alpha1,3-fucosyltransferase A. Venom-allergen protein.	Candidate 85
GB13919	Similar to muscleblind CG33197-PD, isoform D. RNA binding factor involved in muscular dystrophy.	Candidate 86
GB19979	Similar to Glutamic acid decarboxylase 1. Larval locomotory behavior, neuromuscular junction development, neurotransmitter receptor.	Candidate 94
GB10650	Similar to Hormone receptor-like in 46 (Hr46). Regulation of development, metamorphosis, mushroom body development.	Candidate 103

Discussion

Identification of microRNAs in hymenopterans, and in honey bee in particular is of great importance given the potential of honey bee as a model organism for brain development, aging, morphological differentiation and social behavior. It is well known the role that microRNA genes play as regulators of these functions (Krichevsky et al 2003). A complete identification of the genetic components of the honey bee genome coupled with the tools of molecular biology and genetic knockouts would enable us to get a systematic understanding of the biology of the honey bee and would provide insights into the areas where the honey bee excels as a model organism. This work is aimed at the identification of some of these core genetic components of the honey bee biology.

In recent years the role of microRNAs in the development and maintenance of neurons has been identified. Neurons that cannot produce microRNAs slowly die in a manner similar to what is seen in human neurodegenerative disorders as Alzheimer and Parkinson's diseases (Hebert et al 2008). MiR-124, a microRNA present in most metazoans is responsible, in humans, for the downregulation of the expression of nonneuronal genes in an in vitro cell system, suggesting an important role for microRNAs as regulators of neuronal differentiation (Makeyev et al 2007). We have predicted two different conformations for honey bee miR-124. We speculate that these two different conformations may lead to at least two different processed microRNA products with different target genes leading to different phenotypes. On the other hand, we did not predict multiple conformations for miR-124 in wasp. Perhaps the additional conformation possibility in honey bee renders higher information-processing capabilities, assuming miR-124 as a brain specific microRNA, like in humans.

Several of the intronic microRNA candidates reported in this study are located in the introns of genes that in some way or another are implicated in cell differentiation and/or neuronal development, maintenance or functionality. Examples of those genes include GB10180 (Scratch), involved in dendrite development; GB16274 (Mmp2), involved in motor axon fasciculation; GB15597 (ether a gogo), involved in glial growth and regulation of heart contraction; GB14007 (FucTA), a phospholipase heavily expressed in the mushroom body; GB19979 (Glutamic Acid Decarboxylase 1), involved in the correct development of synapses; GB10650 (Hormone Receptor-like in 46), involved in mushroom body development and pupation, and the previously reported GB10066 (neurogilin), implicated in neuron signaling (Craig and Kang 2007; Weaver et al 2007). The functional properties of these genes in addition to several other properties seen in real microRNAs and exhibited by our top scoring candidates like genomic clustering and intronic placement (properties that were not used as part of our scoring system), heavily suggest our top predictions correspond to real microRNA genes.

Materials and methods

Entropy conservation plot

Twenty-five microRNA families from Chapter I were used to generate the entropy conservation plot. This plot uses the same principle as the sequences logos (Crooks et al 2004; Vacic et al 2006), but instead of plotting the frequencies of each base, the sum of the contribution of the four bases is taken into account. In this sense, the more conserved the position in the alignment, the higher the peak in the plot. Divergent positions are seen as valleys. For each position in the alignment a scoring function based on the Shannon entropy index (Shannon 1997) was used.

$$H(l) = \sum_{b=a}^{l} f(b,l) \cdot \log_2 f(b,l)$$

Where:
H(l) = Uncertanty at position *l* (in bits).
b= one of the bases (A,C,T,G).
f(*b*,*l*)= frequency of base b at position *l*.

Use of ultraconservation to predict new microRNAs

Ultraconserved regions in honey bee and wasp were extended by 65nt at each end and extracted for processing. Both sets were then reversed complemented and a new file containing the original sequence and its reverse complement was generated. Pairwise alignments of the honey bee – wasp pairs were constructed using the Smith – Waterman algorithm as implemented in the program Water of the EMBOSS package (Olson 2002). Pairwise alignments were then evaluated using RNAalifold, in order to determine the secondary structure and the free energy of folding of each pair of sequences. Only sequence pairs having the canonical stem – loop precursor found in microRNAs were further considered.

Sequence pairs were then evaluated using RANDfold with 1000 iterations per sequence. Sequence pairs were scored on the basis of their alignment score, deltaG of folding and RNAfold score. Alignment scores and deltaG of folding were normalized to sequence length. Each sequence pair was scored using the following formula:

Pair Score = (Alignment length / alignment score) + (1/((RANDfold score in honeybees + RANDfold score in wasp)/2))/200) + (ABS (deltaG of folding for the pair)/alignment length)*10)

The scoring function was developed this way in order to give equal weight to the different parameters evaluated on the candidate sequences.

Results were sorted according to their score and partitioned according to their position with respect to genomic features in the honey bee genome. Data was partitioned in the following way:

Exons: microRNA candidates that overlap exons of protein coding genes.

Introns: microRNA candidates that overlap introns of protein coding genes.

Splice site: microRNA candidates that overlap splice sites of protein coding genes.

Intergenic: microRNA candidates that do not overlap any of the above.

CHAPTER IV

ULTRACONSERVED ELEMENTS IN HYMENOPTERANS

Introduction

The tools of comparative cenomics allow for the identification of numerous conserved coding regions and conserved non-coding regions as well. Efforts aimed at cataloguing all regulatory elements in the human genome have come to the realization that the fraction of non-coding regions conserved between vertebrate species is much higher than what would be expected by chance (Prabhakar et al 2006), and this amount of conservation is indicative of function. Detection of functional elements is based in the fact that conservation over significant evolutionary distances indicates negative (purifying) selection, which in turns indicates function (Simons et al 2006).

Several publications in the past few years have explored the idea of conservation as a way to detect functional elements. These publications differ in their methodologies, definitions, and in the way they identify conserved regions. For example, Gill Bejerano and collaborators have identified more than ~450 segments longer than 200bp that are absolutely conserved (with no substitutions, insertions or deletions) between orthologous regions of Human, Mouse and Rat (Bejerano et al 2004). Most of these regions can be aligned to the dog genome with 99.2% identity, indicating strong purifying selection at least for the past 92 million years. Analysis of these regions against SNP databases and against the chimp genome suggest these elements are changing at a rate that is 20 times slower that the average for the genome. Bejerano and collaborators refer to these regions as "ultraconserved elements" or UCEs. In a similar approach Tran and collaborators identified regions of short "ultraconservation" between honey bee, fruit fly and mosquito (Tran et al 2006). These are DNA segments between 20 – 50 nt long with no substitutions, insertions or deletions, and were identified by an all versus all approach in which regions of similarity were not restricted to the traditional whole genome alignment approach. These "microconserved" elements have a clear overlap with microRNAs; of the 154 microconserved elements longer than 23nt identified in this study, 24 of them overlap with experimentally validated fly microRNAs, which constitutes about 30% of the known *Drosophila* microRNAs. The authors emphasize the point that this methodology is suitable to detect previously unrecognized microRNAs. Comparisons between human and pufferfish have also detected functional non-coding UCEs, although in a much less number than the ones detected between human and mouse/rat. This is of course a reflection of the evolutionary distance between human and pufferfish (~450 MYA), but it also suggests that any conserved non-coding element detected is likely to be functional (Aparicio et al 2002).

Hierarchical clustering of UCEs in the human genome indicates that longer elements (about 5% of the total elements) are unique in the genome, with no obvious paralogs, or genomic cluster organizations. About 4% of the total is organized in clusters that range between two and 1000 copies, in a manner similar to rDNA and transposable elements. This cluster organization increases the chances of function (Bejerano et al 2004).

UCEs also seem to be organized in a very particular way in the human genome. Using a clever yet simple statistical analysis, Derti and collaborators explored the possibility of positional correlations between the location of UCEs and the location of segmental duplications and copy number variants. In this analysis the number of UCEs within duplicated regions was calculated, and the size of segmented duplicated regions was determined. Using these numbers as a basis, the authors generated a set of 1000 random samples from the human genome, each one of the same size as the size of the segmental duplications. For each random sample, the number of UCEs was calculated. This methodology allowed them to generate a distribution of UCEs in the human genome. This distribution was then used to determine how likely (or unlikely) the observed number of UCEs within segmental duplications is. The authors concluded that the number of UCEs observed within segmental duplications is lower that what should be expected by chance (P=1e-06). The fact that UCEs try to "avoid" segmental duplications suggest a mechanism by which UCEs are eliminated. The authors propose different mechanisms for this behavior: lethality, segregation distortion and lowered fitness (Derti et al 2006).

Distribution of ultraconserved elements

In humans UCEs are distributed throughout the entire genome. It has been determined that they overlap basically all human genome features, including: exons, introns, 5' UTRs, 3' UTRs, gene-proximal 5' regions, gene-proximal 3'regions, intergenic regions, and gene deserts, although with different frequencies (Shabalina et al 2001). UCEs seem to be particularly enriched in introns and intergenic regions. The reason for this enrichment is not clear right now, but it emphasizes the idea that strong purifying selection can be also exerted in introns, which sometimes can be higher than their exon counterparts, suggesting essential functional roles (Thomas et al 2003).

For this work we have identified a series of UCEs between the genomes of *Apis mellifera* (honey bee) and *Nasonia vitripennis* (parasitic wasp). We explore the possible functional roles of these elements in the context of gene ontologies and the evolutionary forces that appear to maintain these ultraconserved elements.

Results

Sequence conservation at the protein or DNA level is usually the result of functional or structural evolutionary constraints. We understand how functional constraints at the protein level lead to sequence conservation at the DNA level, but we understand very little about conservation of non-protein coding DNA sequences.

This type of conservation is often associated with non-coding RNAs or transcription factor binding sites, elements that we associate mostly with gene regulation.

For this study we were able to identify UCEs between the genomes of *Apis mellifera*, *Nasonia vitripennis* and *Drosophila melanogaster*. The first two species belong to the order Hymenoptera and have a divergence time of about ~100 mya (Belayeva et al 2002), which is comparable to the divergence time between human and mouse estimated in 90 mya (Ureta-Vidal et al 2003). *Drosophila* belongs to the order Diptera, which diverged from the Hymenoptera around 330 million years ago (Belayeva et al 2002), a time comparable to the divergence between human and birds.

The distribution of results of our UCE analysis is summarized in Table 6. It shows the frequency of elements at each intersection and their distribution with respect the structural properties of protein coding genes in the honey bee genome. We found 869 different UCEs of 50bp or longer between honey bee and wasp. These elements appear to be preferentially located in intergenic regions (57% of them), followed by intronic elements (30%), and exon/splice sites (14%).

This distribution is similar to what has been seen in other insects (Glazov et al 2005) and vertebrates, where the great majority of elements are located in intergenic and intronic regions (Bejerano et al 2004).

Table 6. Ultraconserved elements found among insects

Intersection	Length	Total # Elements	Intergenic	Intronic	Exonic	Splice site
Amel – Nvit	**>=50**	869	495	253	51	70
	>=100	17	10	4	0	3
	>=150	1	0	1	0	0
Amel - Dmel	**>=50**	113	80	29	2	2
	>=100	0	0	0	0	0
	>=150	0	0	0	0	0
Amel – Nvit – Dmel	**>=50**	93	67	24	0	2
	>=100	0	0	0	0	0
	>=150	0	0	0	0	0

UCEs are distributed throughout the entire honey bee genome. They are particularly abundant in chromosomes 1 (being the longest), 13 and 16. The latter two are particularly concentrated with intergenic elements despite the fact that these two chromosomes are among the shortest chromosomes in honey bee and chromosome 16 contains the least number of genes (260 total). This result might be an indication of regulatory elements being abundant in these chromosomes or an indication that these elements form parts of genes. We raise this possibility given the fact that the official gene set appears to be an underestimate of the total number of genes in honey bee. The total number of genes in the official gene set is around ~10.000 (The HoneyBee Genome Sequencing Consortium 2006), this number seems rather low compared to *Drosophila* with an estimated gene count of ~14.000 (Clark et al 2007).

The longest element found at the intersection between honey bee – wasp corresponds to an intronic element found in a gene that corresponds to the *Drosophila* homolog of muscleblind. This gene belongs to the MBNL protein family, whose members have been shown, in humans, to be involved in myotonic dystrophy, a form of muscular dystrophy that is characterized by muscle weakness and myotinoia (slow relaxation of muscle after contraction). Recent reports indicate that Muscleblind in humans may work as a regulator of alternative splicing of two pre-mRNA genes that are misregulated in myotonic dystrophy: cardiac troponin T (cTNT) and insulin receptor (IR). Both of these proteins have homologs in honey bee, as cTNT is homologous to TpnT and IR has a homolog with the same name in honey bee (Herranz et al 2005). This element bears no significant similarity to other sequences in databases and interestingly its lowest free energy conformation corresponds to a stable stem-loop structure suggesting a role as a functional RNA molecule. We don't discard the possibility of this element encoding a microRNA since in a separate analysis (Chapter III) this element showed up as one of the highest scoring microRNA candidates (candidate No. 86). For a complete list see Table 7.

We also found other interesting genes harboring longer intronic UCE elements. The gene LMO4 (GB17398) harbors the second longest UCE (117bp). This gene encodes a cysteine-rich protein that contains two LIM domains but lacks a DNA-binding homeodomain. The encoded protein may play a role as a transcriptional regulator or as an oncogene. Several genes harboring intronic ultraconserved elements include homeotic developmental regulators such as *cut* (GB17945), the protein GB15643, which contains both a homeodomain and the brain-specific homeobox POU, and the protein *mub* (mushroom-body expressed, GB10111), which acts as a regulator of alternative nuclear mRNA splicing via spliceosome and is preferentially expressed in brain tissue (Grams and Korge 1998; Mutsuddi et al 2004). We also found the protein *toutatis* harboring a long intronic UCE. This protein contains a zinc-finger domain and mutational analyses indicate the protein is essential for neural development (Vanolst et al 2005).

Genes harboring long exonic UCEs appear to be involved in ion transport (GB15328, GB15412) and transcription factor activities (GB17328, GB15089). Genes harboring long splice site elements are also related to ion transport (GB12929) and neural circuitry (GB17617).

In order to get a better understanding of the possible functions associated with UCEs, we carried out a gene ontology analysis of genes harboring exonic, intronic, and splice site UCEs. This analysis allowed us to identify any potential overrepresented GO term in out dataset. Our results indicate significant enrichments for each one of these categories (Table 8). At the Molecular Function level, intronic UCE sequences appear to be enriched in "transcription factor activity" (P=6e-08) and "DNA binding" (P=3e-05). At the Biological Process level, they appear to be enriched in "regulation of transcription, DNA - dependent" (P=2e-09), "leg formation" (P=5e-04), and "multicellular organismal development" (P=6e-04).

Table 7. Functions of genes associated with longest UCEs in honey bee and wasp. Numbers in parenthesis correspond to the fraction of the UCE that overlap introns

Gene	Function	UCE size (bp)
Intronic Elements		
GB13919	Similar to muscleblind CG33197-PD, isoform D	170
GB17398	Similar to LIM domain only 4	117
GB17945	Similar to Homeobox protein cut	111,101,91
GB10111	Similar to *meb* (mushroom-body expressed) CG7437-PC, isoform C	97
GB13576	Similar to CG9850-PA, isoform A, Zinc-dependent metalloprotease, salivary_gland	96
GB15643	Similar to CG11641-PA, Homeodomain	95
GB13918	Similar to One cut domain family member 2 (Transcription factor ONECUT-2) (OC-2)	94
GB20131	Similar to CG11323-PA. Tubulin-tyrosine ligase family.	92
GB14851	Similar to Muscle protein 20 CG4696-PA, isoform A	90
GB16601	Similar to toutatis CG10897-PA, isoform A	89
Exonic Elements		
GB19480	Similar to lethal (1) G0196 CG14616-PD, isoform D	86
GB19429	Similar to CG17838-PE, isoform E	84
GB15089	Similar to Sp1 CG1343-PA, isoform A, RNA polymerase II transcription factor activity	80,67,62
GB15328	Similar to Potassium voltage-gated channel protein Shaw (Shaw2)	74
GB15412	Similar to Ca2+-channel protein 1 subunit D CG4894-PA, isoform A	68
GB11661	Similar to jim CG11352-PC, isoform C	65
GB17328	Putative transcription factor mblk-1	64, 60
GB30541	No description found	63
GB13371	Similar to lethal (1) G0269 CG1696-PA	62
GB30126	No description found	60
Splice Sites		
GB12929	Similar to paralytic CG9907-PA	136 (125), 99(22), 84(76)
GB18272	Similar to transportin 1	118(76)
GB16271	No description found	110(92)
GB10313	No description found	98(59)
GB19490	Similar to RluA-1 CG31719-PA, isoform A	94(75)
GB18263	Similar to crooked legs CG14938-PA, isoform A	94(76)
GB10502	No description found	88(40)
GB10395	Similar to drumstick CG10016-PA, isoform A	88(68)
GB17617	Similar to fruitless CG14307-PB, isoform B	86(34)
GB13445	Similar to CG12467-PA	86(15)

Table 8. GO enrichment of genes with UCEs longer than 50bp shared between honey bee and wasp

GO Term	Description	e-score	GO Category
Introns			
GO:0003700	transcription factor activity	6.53E-08	Molecular Function
GO:0043565	sequence-specific DNA binding	1.65E-06	Molecular Function
GO:0003677	DNA binding	3.39E-05	Molecular Function
GO:0006355	regulation of transcription, DNA-dependent	2.89E-09	Biological Process
GO:0006350	Transcription	0.000100932	Biological Process
GO:0007479	leg disc proximal/distal pattern formation	0.000580918	Biological Process
GO:0007275	multicellular organismal development	0.000663494	Biological Process
GO:0010092	specification of organ identity	0.004037661	Biological Process
GO:0007432	Salivary gland boundary specification	0.004037661	Biological Process
GO:0045449	regulation of transcription	0.005139759	Biological Process
Exons			
GO:0003705	RNA polymerase II transcription factor activity, enhancer binding	0.00580904	Molecular Function
GO:0042045	epithelial fluid transport	0.02727066	Biological Process
Splice sites			
GO:0003676	Nucleic acid binding	0.000104693	Molecular Function
GO:0005244	voltage-gated ion channel activity	0.001842488	Molecular Function
GO:0030955	potassium ion binding	0.04769905	Molecular Function
GO:0045433	male courtship behavior, veined wing generated song production	0.008769452	Biological Process

Exonic sequences show enrichment for “RNA pol II transcription factor activity” (P=5e-03) at the Molecular Function level. There is also a slightly significant enrichment for “epithelial transport” (P=0.02) at the Biological Process level.

Splice site sequences show enrichment for “Nucleic Acid Binding” (P=1e-04), “Voltage-gated channel activity” (P=2e-03) and “potassium ion binding” (P=0.04).

A significant fraction of UCEs are located in introns

Our results clearly indicate that the great majority of UCEs are localized in introns and intergenic regions. Conservation in these regions suggests functional roles being performed by these regions in the genome. We know really very little about these molecules and the mechanisms implicated in their conservation. In order to get a better understanding of this phenomenon, we carried out an analysis aimed at determining the strength of selection at which these regions are maintained. For tractability purposes we concentrated our analysis in introns, as they have different statistical signals that allow us to determine their homology and make possible to do comparisons with their exon counterparts.

Orthology between genes is usually computed by using the protein product encoded by the gene of interest. The advantage of using proteins instead of DNA has to do with the fact that proteins have more information content per position than DNA. Determining orthology between introns is a much more difficult task because they don't encode for proteins. Any product encoded by an intron remains (as far as we know) in the form of RNA. To overcome the problem of working with DNA sequences, we developed a procedure that allowed us to determine orthology between introns. We started by first determining orthology between honey bee and wasp proteins and then we determined orthology between introns by using the corresponding DNA sequences (the complete gene structure) of the genes for which we determined orthology.

Using this approach we were able to determine orthology for 2705 intron pairs corresponding to 768 different genes. We refer to this set as the "Gene-Intron Orthology set". We calculated genetic distances based on the Kimura 2-parameter model (K2) for these introns as a whole and in a "by window" approach. Briefly, the Kimura 2-parameter (Kimura 1980) is a model of evolution that takes into account the fact that transitions are more probable than transversions in nucleotide sequences (Kimura 1980). The model is widely used and it has been demonstrated that it is one of the most accurate models of evolution for DNA sequences.

Conservation in introns is variable across their length

Analysis of intron conservation using a window approach, shows that conservation appears to be higher for regions located towards the 5' and 3' ends and lower for regions located towards the center (Figure 22). All centered-located windows show an amount of divergence higher than 0.38 substitutions per site, whereas windows located at the 3' and 5' have less than 0.36 substitutions per site. The higher degree of conservation towards the ends of introns is probably an indication of the presence of control elements necessary for the splicing machinery, as previously suggested (Gazave et al 2007).

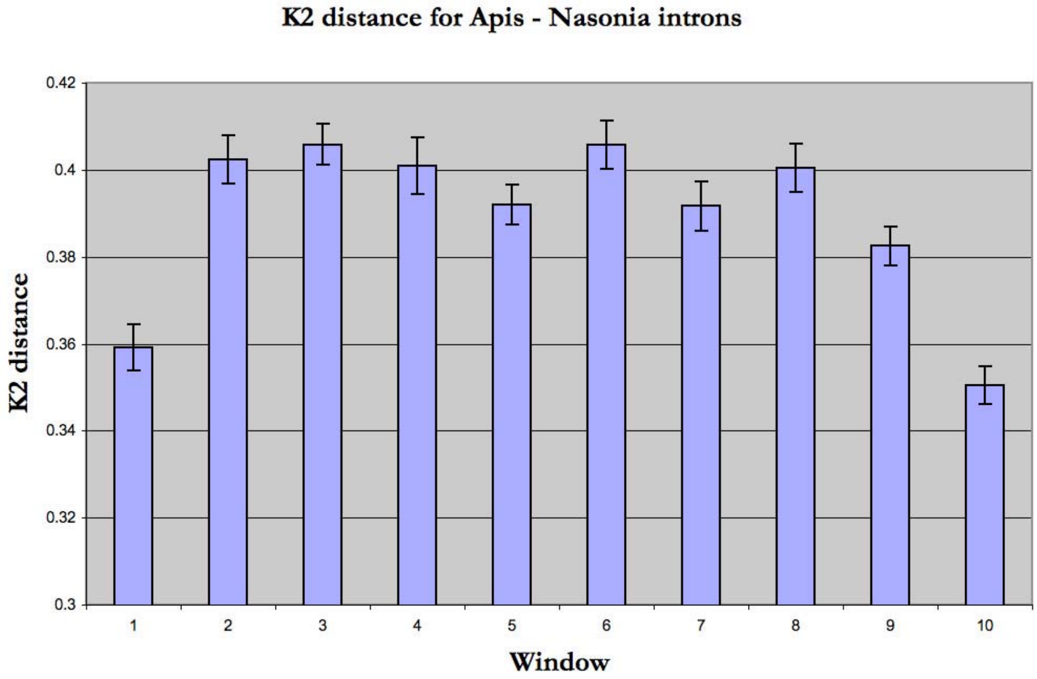


Figure 22. Genetic distance (Kimura 2 parameter) between homologous introns in honey bee - wasp. Conservation tends to be higher for regions close to splice sites (intron / exon) boundaries.

Exon divergence is higher than intron divergence

The same genetic distance analysis carried out for introns was applied to their exon counterparts. Given the gene structure of every gene, we were able to reconstruct their coding parts, assemble them and compute genetic distances among them. Interestingly, genetic distances in coding regions (K2 distances) show a degree of divergence higher than their intron counterparts (Figure 23). Coding regions for these genes appear to be more divergent than introns by at least 60%.

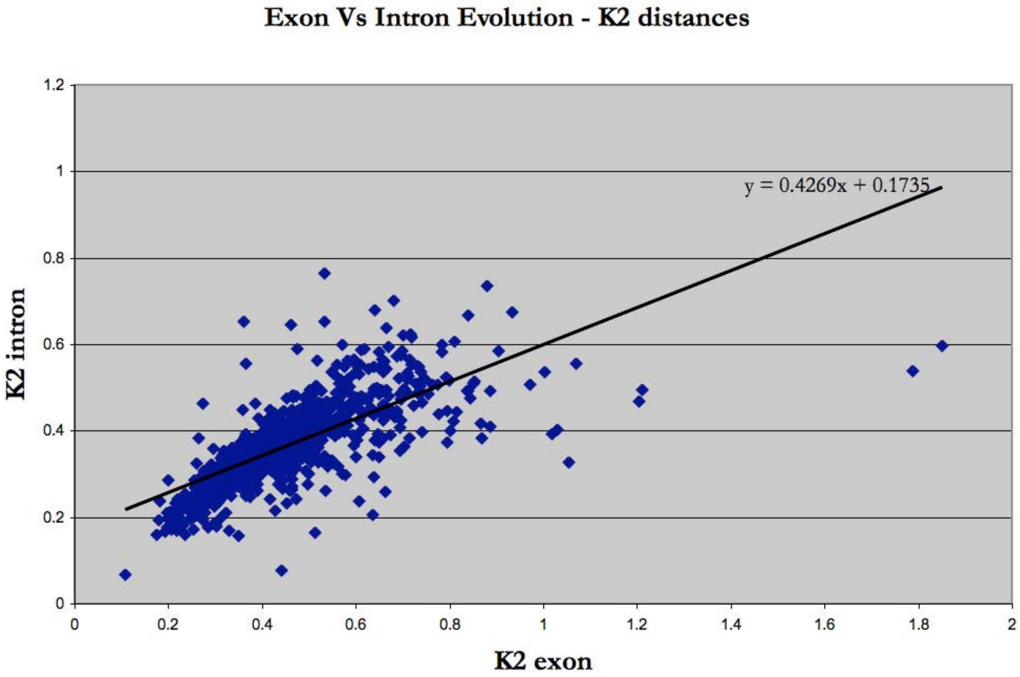


Figure 23. K2 distances of exons and introns within the same gene. On average intron distances appear to be smaller than their intron counterparts by at least 60%.

Genes within the Gene Intron Orthology Set are under purifying selection

We computed nonsynonymous (Ka) Vs synonymous (Ks) substitutions for coding regions of these genes. Substitutions that result in amino acid replacements are said to be nonsynonymous while substitutions that do not cause an amino acid replacement are said to be synonymous. For example, a change from GGG to GGA in a codon results in a synonymous substitution, given the fact that the identity of the amino acid is not changed, both code for Glycine. Genes for which the ratio Ka/Ks is higher than 1 are said to be under negative selection, whereas genes for which the ratio Ka/Ks is lower than 1 are said to be under negative selection (Nei and Gojobori 1986; Nekrutenko et al 2002).

We compared K2 distances of introns vs. Ka/Ks ratios for exons and our result indicates that there is no obvious correlation between intron distance and Ka/Ks ratios. Interestingly, the great majority of the genes analyzed appear to be under negative selection (Figure 24). This result is despite the fact that K2 distances in exons appear to be more divergent than their intron counterparts. This indicates that most of the divergence in exons is the result of synonymous substitutions, and this substitution rate is higher that the substitution rate found in introns. This implies that the great majority of substitutions seen in Figure 23 are the result of synonymous substitutions.

Conservation for introns harboring ultraconserved elements

Genetic distances for introns harboring ultraconserved elements longer than 50bp is significantly lower that for introns not harboring UCEs. It is expected that introns with UCEs to be more conserved than on average (given the ultraconservation), but it appears that high levels of conservation are not only restricted to the UCE harboring region, but it is a general phenomenon of these introns.

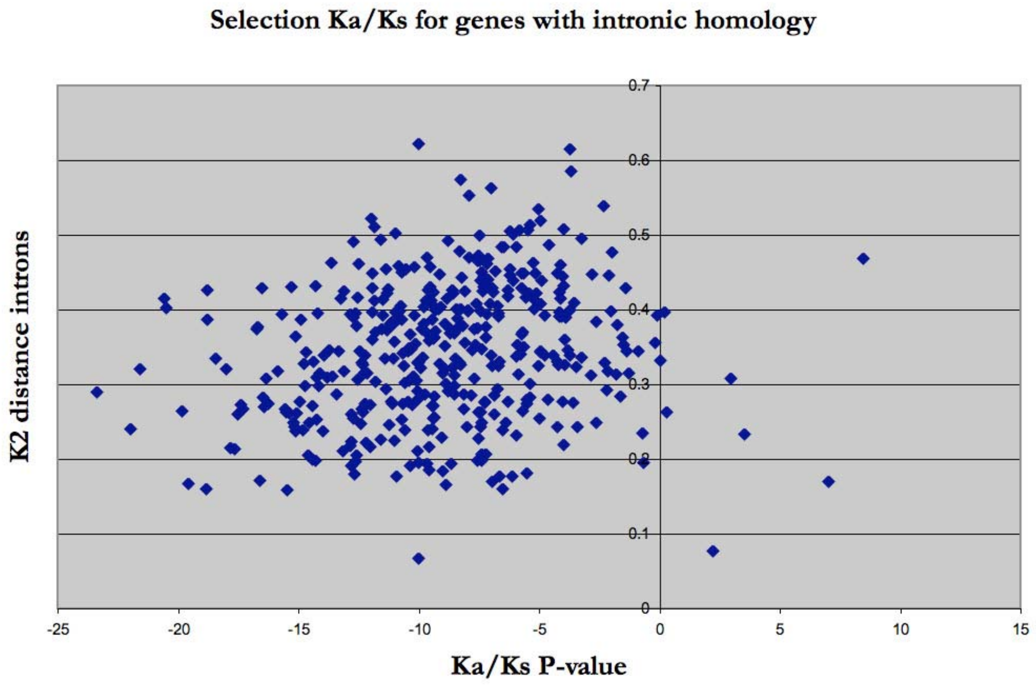


Figure 24. The great majority of orthologous genes with intronic homology are under purifying selection. Ka/Ks values < 0.

Figure 25 shows a comparison of introns harboring UCEs and introns not harboring them. Introns with UCEs show a level of conservation that is significantly higher than on average, and this conservation spans the entire length of these introns. Most likely this is the result of functional conservation.

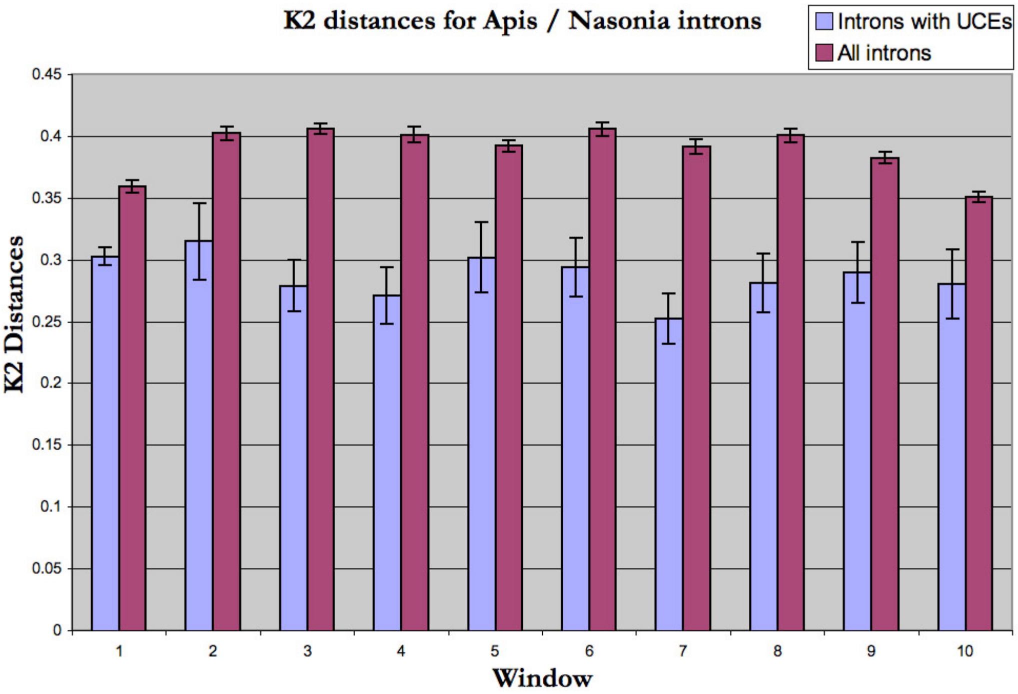


Figure 25. K2 distances for introns harboring ultraconserved elements vs introns not harboring them. Introns with UCEs are remarkably more conserved than not-UCE introns.

Discussion

The results from this work show that there are several non-coding regions between the genomes of honey bee and wasp that are selectively constrained. We were able to identify almost 900 different UCEs longer than 50bp between these two genomes, which seem to be preferentially located in introns and intergenic regions. This finding is very similar to what has been seen in humans, in which among the 327.000 conserved nongenic sequences that were recently found in the human genome, more than 35% were located in introns (Dermitzakis et al 2005). Also, approximately 100 of the 481 UCEs found in the human genome map within introns

(Bejerano et al 2004). This is a recurring theme that shows that introns may have roles that we haven't even thought about before, and despite the fact that we still don't understand the possible roles of these sequences, there is strong evidence pointing out that they play a regulatory role (Rizzolio et al 2008).

We found different kinds of enrichments depending on the genomic features the UCEs overlapped with (introns, exons, splice sites). GO terms for UCEs located in introns are highly enriched with terms related with transcription factor activity. This is not only a feature of the intersection bee / wasp, but is present in other insect genomes, and mammals as well, reinforcing the idea that these elements functions as regulator of the genetic programming of higher metazoans.

We have provided evidence indicating that the strength of selection for intronic regions among a group of genes in honey bee and wasp is greater than their exonic counterparts. Introns and exons of these genes are under negative selection, but interestingly, the strength of selection in introns is significantly higher than in exons. It is unclear why these high levels of conservation are necessary in introns. One possibility is that these introns play a significant role in maintaining essential splicing alternatives for these genes. That would explain why conservation is higher towards splice sites, and higher towards the middle of the introns. Expression experiments using oligonucleotide microarrays designed to capture splicing information across the honey bee genome, would be able to provide information about the prevalence of alternative splicing among these genes, and would be able to provide information as to what sequence motifs are the ones that lead to the expression of particular splicing alternatives. Clustering genes according to their expression profiles can accomplish this goal, as it has been previously done (Sugnet et al 2006).

Gene ontology experiments using the entire set of genes from the Gene-Intron Orthology Set didn't show any statistically significant enrichment. According to this result, functions of these genes are divergent and probably don't constitute a natural population connected by function. Their commonality is to have highly unusually conserved introns.

The identification of ultraconserved regions in the honey bee / wasp genomes, the fact that a significant fraction of them localize within introns or intergenic regions, and the demonstration that introns for a fraction of the genes shared between honey bee and wasp appear to be under negative selection raises important questions related to our understanding of introns and non-coding regions of the genome in general. In the past, introns and intergenic regions were regarded as non-functional and therefore evolved under neutrality (Waterston et al 2002). This view of the non-coding regions of the genome is rapidly changing for a vision in which these regions perform regulatory functions. Most of these functions related to cell differentiation and morphogenesis. These two processes were usually considered to be directed by proteins including transcription factors, homeodomain proteins, and chromatin-modifying proteins. This view is likely to change with the elucidation of the functional roles of non-coding RNA molecules and the discovery of all the players. There are new proposals that point out RNA regulation as the leading mechanism for differentiation and development of higher eukaryotes in a manner that allows a precise interaction between modulator and effector (i.e. microRNA – target gene by base pairing). This level of precise "digital" interaction that can be achieved with RNA molecules is not as precise in the protein world. This new view proposes RNA molecules and not proteins as the most important players in the developmental biology of higher eukaryotes (Mattick 2007).

Materials and methods

Assembled genomes were obtained from online repositories. Analyses were carried out in the following repeat masked genomes: *Apis mellifera* (Amel4.0 scaffolds), *Nasonia vitripennis* (nvit1.0 contigs), and *Drosophila melanogaster* (dmel4.2.1 chromosomes).

Genome intersections

Genome intersections were done in a binary fashion, starting with *Apis mellifera* and *Nasonia vitripennis.* Each intersection was calculated using a combination of perl scripts and the WUBLAST package (Gish, 2004). Given the size of the genomes and the limitations of memory it was necessary to divide the query sequence into different fragments and search them independently. This strategy is known as query chopping and it is mostly used to minimize disk swapping and improve sequence processing (Korf et al 2003).

Parameters used for this search were the following:

- B = 100000 (number of database hits to report).
- V = 100000 (number of online summaries).
- spoutmax=0 (no limit to the number of segment pairs to report).
- nogaps (ungapped alignments).
- Hspsepmax = 100 (max separation between hsps = 100nt).
- filter = seg.
- mformat=2 (Table in tabulated format).
- hspmax=0 (no limit on the number of hsp reported).
- W=20 (word size).
- N=9000 (mismatch penalty).

Individual BLAST reports were parsed using a perl script. Sequences were extracted using xdget (Gish 2004). The extracted sequence and its reverse complement were stored in a different file.

In order to minimize redundancy in the extracted sequences, every individual sequence that was a perfect substring of another sequence of equal of longer length was removed. This step guarantees that the set of UCEs is a set of maximal N-mers (Tran et al 2006). This step was carried out using patdb (Gish 2004).

In order to intersect a third genome (i.e. *Apis* – *Nasonia* - *Drosophila*), the previous binary intersection was used as query for the next database search.

Mapping ultraconserved elements to genome

The non-redundant set of UCEs resulting from the previous step was mapped into the genome of interest in order to determine their relationship with respect to other genomic features, particularly gene features like exons, introns and splice sites. A General Feature Format file (GFF) was generated with these mappings.

Overlaps between UCEs and previously published genomic features for the honey bee genome (The HoneyBee Genome Sequencing Consortium 2006) were determined using a perl script. The resulting data was then partitioned according to the type of feature the UCE overlapped with in the following way:

- Exons: UCEs that were completely contained within exon coordinates.
- Introns: UCEs that were completely contained within intron coordinates.
- Splice sites: UCEs that overlapped splice site junctions.
- Intergenic: UCEs that did not overlapped any genomic feature.

Gene ontologies

Ontologies for honey bee genes were determined by transferring annotation from orthologous genes in fruit fly, mouse, and human. Orthology between proteins in these organisms and honey bee proteins was determined using Reciprocal Smallest Distance (RSD) (Wall and Deluca 2007). The file with orthology relationships between honey bee, human, mouse, and fruit fly was gently provided by Dr. Christine Elsik (personal communication).

Computation of orthologies

The RSD algorithm refines the concept behind the Reciprocal Best Hit (RBH) algorithm (Tatusov et al 1997) by incorporating a model of evolution. The result of this extra step minimizes the problems inherent to the RBH algorithm, which often assigns orthology relationships to genes that are not really orthologs.

Briefly, RSD employs BLASTP as a first step, starting with a subject proteome, J, and a protein query sequence, i, belonging to genome I. A set of hits H, exceeding a predefined significance threshold (e.g., E < 10-20, though this is adjustable) is obtained. Then, using clustalW, each protein sequence in H is aligned separately with the original query sequence i. If the alignable region of the two sequences exceeds a threshold fraction of the alignment's total length, the program PAML (Yang 2007) (specifically, the package codeml) is used to obtain a maximum likelihood estimate of the number of amino acid substitutions separating the two protein sequences, given an empirical amino acid substitution rate matrix (Jones et al 1992). The model under which a maximum likelihood estimate is obtained in RSD may include variation in evolutionary rate among protein sites, by assuming a gamma distribution of rate across sites and by setting the shape parameter of this distribution, α, to a level appropriate for the phylogenetic distance of the species being compared (Nei et al 2001). (This parameter, α, may be altered to accommodate different degrees of phylogenetic distance.) Of all sequences in H for which an evolutionary distance is estimated, only j, the sequence yielding the shortest distance, is retained. This sequence j is then used for a reciprocal BLAST against genome I, retrieving a set of high scoring hits, L. If any hit from L is the original query sequence, i, the distance between i and j is retrieved from the set of smallest distances calculated previously. The remaining hits from L are then separately aligned with j and maximum likelihood distance estimates are calculated for these pairs as described above. If the protein sequence from L producing the shortest distance to j is the original query sequence, i, it is assumed that a true orthologous pair has been found and their evolutionary distance is retained.

Gene ontology overrepresentation

Overrepresentation of gene ontology terms for the different ultraconserved partitions was determined using the hypergeometric distribution as implemented by GeneMerge (Castillo-Davis and Hartl 2003). Briefly, the hypergeometric distribution gives a quantification of the level of over-representation for a particular item in a given sample of size *k* drawn from a larger population, size *n*. In GeneMerge, *k* is always the study set of genes and *n* is the population set, the set from which *k* is drawn. The number of genes with a particular identifier is *r*. *p* is the fraction of genes in the population *n* associated with the particular identifier under investigation. The hypergeometric gives the exact probability of drawing *r* genes with a particular identifier from a sample of size *k* from a population of size *n* given that the identifier exists in fraction p in the population set of genes (Castillo-Davis and Hartl 2003). For this dissertation in particular, the number of UCEs having a particular identifier (i.e. intronic, exonic) is represented by *r*. The fraction of genes in the population with a particular GO term is represented by *p*. The population of genes having a GO term is represented by n.

$$\Pr(r \mid n, p, k) = \frac{\binom{pn}{r}\binom{(1-p)n}{k-r}}{\binom{n}{k}}$$

Identification of orthologous introns and calculation of genetic distances

Gene IDs from genes shown as orthologs by our orthology search were used to grab the information about the genomic structure of the genes from GFF files previously computed for the honey bee project (The Honey Bee Genome Sequencing Consortium 2006) and the wasp project (unpublished). Gene Structures were mapped to their corresponding genomes and different features (intron, exons) were isolated individually.

In order to consider an intron as being orthologous, the gene and the introns had to fulfill the following criteria:

a. Genes evaluated had to come as orthologs by our RSD analysis.
b. Number of introns for both genes had to be equal.
c. Introns with the same relative locations in honey bee / wasp were aligned using the Smith – Waterman algorithm and the statistical significance of their score was evaluated by PRSS. Only intron pairs with a p-value lower than 1e-05 were considered homologous.

Introns fulfilling these criteria were then aligned using t-coffee (Notredame et al 2000), and their patterns of substitutions evaluated using the Kimura 2-parameter method as implemented in DNADIST from the phylip package (Felsestein 1989).

Exons were assembled into one single "in silico" cDNA molecule and aligned with its protein product in order to verify the DNA sequence as being correct. K2 distances were computed with the program DNADIST. Ka/Ks ratios were computed using the module DNASTATISTICS as implemented in BIOPERL (Stajich 2007).

CHAPTER V

SUMMARY AND CONCLUSIONS

Our understanding of the evolution and functionality of species, genomes and genes depends greatly on our ability to identify and catalogue all the functional components that are present in a genome. A "list of parts" that allow us to understand what is happening "under the hood" of our species of interest. Without a complete list of parts, our knowledge of the biology and our ability to make predictions would be severely hampered. Despite the fact that we are getting better at identifying new genes in species with a sequenced genome, our knowledge is still incomplete. The field of comparative genomics offers immense possibilities to learn how to "fine tune" our tools and techniques for this purpose.

This work is aimed at the identification and functional characterization of small genetic components that have been conserved in different species as divergent as 450 mya of evolution. We have developed a new strategy useful for the identification of homologs of microRNAs in different species, with a process that takes advantage of the different statistical signals present in microRNAs in order to maximize the finding of homologs. We have also shown how the microRNA component of metazoan genomes has gone through several rounds of innovation in the branch that leads to humans. Bursts of innovation appeared in vertebrates, eutherians, primates and hominids. These innovations appear to be correlated all the time with processes related to cell specialization, morphogenesis, cell differentiation, and in general all the processes by which expressed genes are able to create multicellular organisms with well organized anatomical structures. According to this result, even the most divergent microRNA families regulate processes related to cell differentiation, suggesting that different morphologies are mostly the result of microRNA regulation and not the result of the expression of species-specific protein families.

We have also shown how the tools of comparative genomics can be used to identify microRNA genes in a systematic way. Our computational strategy clearly identified known microRNAs and novel ones based on a set of “seeds” of ultraconserved elements between two genomes. Several of these candidates have been experimentally validated using 454 sequence data derived from small RNA fractions isolated from honey bee and wasp. These candidates exhibit features such as genomic clustering, which have been seen in real microRNAs, but wasn’t part of our algorithm. This result increases the chances of our predictions correspond to genuine microRNAs. Our computational strategy identified a large number of novel microRNA candidates common to both of them, and identified 42 of the 45 previously validated microRNAs for both organisms. This represents more than 90% of the microRNAs from the experimental set.

It is worth noting that this strategy does not require whole genome alignments as other microRNA search strategies do (Sandmann and Cohen 2007). This represents an advantage for the analysis of sequences from species that lack a genome assembly but contain sequence information in the form of partial reads, BACs or plasmids. It is also useful for genomes on their first stages of assembly, or for genomes for which there is no plan for deep coverage.

Our search strategy relies on the scoring of putative homologous sequences on the basis of sequence conservation and structural conservation. We demonstrated that this strategy works well even for species with moderate divergence such as honey bee and wasp, with a time of divergence of about 120 million years.

Comparative genomics is indeed a powerful tool that allows us to identify functional elements within genomes. Our analyses identified several conserved non-coding regions, a significant fraction of them located within introns that appear to be under negative (purifying) selection. Not very long ago we use to regard introns as non-functional, and assumed they were discarded after splicing and further degraded. Introns were considered as a mechanism to facilitate exon shuffling and nothing more.

We have identified several functional genetic components (microRNAs and ultraconserved elements) within honey bee introns and have demonstrated selective constraints on several of them. Most likely we have only identified just a few of the players of the non-coding RNA component present between honey bee and wasp.

Knockouts of these regions would provide insights of their functionality and ideas for new experiments.

The generation of a sequence genome at half the evolutionary distance existing between honey bee and wasp would be of great help to increase the chances of finding non-coding genetic components in these organisms. Stingless bees of the genus *Mellipona* would be great candidates for such a comparison that would also provide a great deal of information about genes involved in the generation of the venom proteins and apparatus of honey bees.

REFERENCES

Abad, J.P., B. De Pablos, K. Osoegawa, P.J. De Jong, A. Martin-Gallardo, and A. Villasante. 2004. TAHRE, a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. *Mol Biol Evol* **21:** 1620-1624.

Aboobaker, A.A., P. Tomancak, N. Patel, G.M. Rubin, and E.C. Lai. 2005. *Drosophila* microRNAs exhibit diverse spatial expression patterns during embryonic development. *Proc Natl Acad Sci U S A* **102:** 18017-18022.

Adams, M.D. S.E. Celniker R.A. Holt C.A. Evans J.D. Gocayne P.G. Amanatides S.E. Scherer P.W. Li R.A. Hoskins R.F. Galle et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185-2195.

Allen, E., Z. Xie, A.M. Gustafson, G.H. Sung, J.W. Spatafora, and J.C. Carrington. 2004. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana. Nat Genet* **36:** 1282-1290.

Ambros, V. 2001. MicroRNAs: tiny regulators with great potential. *Cell* **107:** 823-826.

Ambros, V. 2004. The functions of animal microRNAs. *Nature* **431:** 350-355.

Ambros, V., B. Bartel, D.P. Bartel, C.B. Burge, J.C. Carrington, X. Chen, G. Dreyfuss, S.R. Eddy, S. Griffiths-Jones, M. Marshall et al. 2003. A uniform system for microRNA annotation. *RNA* **9:** 277-279.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J.M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301-1310.

Artavanis-Tsakonas, S., M.D. Rand, and R.J. Lake. 1999. Notch signaling: cell fate control and signal integration in development. *Science* **284:** 770-776.

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

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, and D. Haussler. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321-1325.

Belayeva, N.V., A.P. Rasnitsyn, and D.L.J. Quicke. 2002. *History of insects*. Kluwer Academic Publishers, Boston.

Bentwich, I. 2005. Prediction and validation of microRNAs and their targets. *FEBS Lett* **579:** 5904-5910.

Berezikov, E., V. Guryev, J. van de Belt, E. Wienholds, R.H. Plasterk, and E. Cuppen. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120:** 21-24.

Blomme, T., K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, and Y. Van de Peer. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* **7:** R43.

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

Brennecke, J., D.R. Hipfner, A. Stark, R.B. Russell, and S.M. Cohen. 2003. *bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*. *Cell* **113:** 25-36.

Calin, G.A., C. Sevignani, C.D. Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini et al. 2004. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc Natl Acad Sci U S A* **101:** 2999-3004.

Carrington, J.C. and V. Ambros. 2003. Role of microRNAs in plant and animal development. *Science* **301:** 336-338.

Castillo-Davis, C.I. and D.L. Hartl. 2003. GeneMerge--post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* **19:** 891-892.

Chen, K. and N. Rajewsky. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8:** 93-103.

Chen, X. 2005. MicroRNA biogenesis and function in plants. *FEBS Lett* **579:** 5923-5931.

Clark, A.G. M.B. Eisen D.R. Smith C.M. Bergman B. Oliver T.A. Markow T.C. Kaufman M. Kellis W. Gelbart V.N. Iyer et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203-218.

Claverie, J.M. 2001. Gene number. What if there are only 30,000 human genes? *Science* **291:** 1255-1257.

Craig, A.M. and Y. Kang. 2007. Neurexin-neuroligin signaling in synapse development. *Curr Opin Neurobiol* **17:** 43-52.

Crick, F. 1970. Central dogma of molecular biology. *Nature* **227:** 561-563.

Crooks, G.E., G. Hon, J.M. Chandonia, and S.E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188-1190.

Dermitzakis, E.T., A. Reymond, and S.E. Antonarakis. 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet* **6:** 151-157.

Derti, A., F.P. Roth, G.M. Church, and C.T. Wu. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* **38:** 1216-1220.

Dezulian, T., M. Remmert, J.F. Palatnik, D. Weigel, and D.H. Huson. 2006. Identification of plant microRNA homologs. *Bioinformatics* **22:** 359-360.

Eddy, S.R. 1999. Noncoding RNA genes. *Curr Opin Genet Dev* **9:** 695-699.

Enright, A.J., B. John, U. Gaul, T. Tuschl, C. Sander, and D.S. Marks. 2003. MicroRNA targets in *Drosophila*. *Genome Biol* **5:** R1.

Featherstone, D.E., E. Rushton, and K. Broadie. 2002. Developmental regulation of glutamate receptor field size by nonvesicular glutamate release. *Nat Neurosci* **5:** 141-146.

Felsestein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164 - 166.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269:** 496-512.

Gazave, E., T. Marques-Bonet, O. Fernando, B. Charlesworth, and A. Navarro. 2007. Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol* **8:** R21.

Giraldez, A.J., R.M. Cinalli, M.E. Glasner, A.J. Enright, J.M. Thomson, S. Baskerville, S.M. Hammond, D.P. Bartel, and A.F. Schier. 2005. MicroRNAs regulate brain morphogenesis in zebrafish. *Science* **308:** 833-838.

Glazov, E.A., M. Pheasant, E.A. McGraw, G. Bejerano, and J.S. Mattick. 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res* **15:** 800-808.

Grams, R. and G. Korge. 1998. The mub gene encodes a protein containing three KH domains and is expressed in the mushroom bodies of *Drosophila melanogaster*. *Gene* **215:** 191-201.

Gregory, R.I., T.P. Chendrimada, N. Cooch, and R. Shiekhattar. 2005. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell* **123:** 631-640.

Gregory, T.R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc* **76:** 65-101.

Griffiths-Jones, S., R.J. Grocock, S. van Dongen, A. Bateman, and A.J. Enright. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34:** D140-144.

Griffiths-Jones, S., H.K. Saini, S. van Dongen, and A.J. Enright. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36:** D154-158.

Grimson, A., K.K. Farh, W.K. Johnston, P. Garrett-Engele, L.P. Lim, and D.P. Bartel. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27:** 91-105.

Grun, D., Y.L. Wang, D. Langenberger, K.C. Gunsalus, and N. Rajewsky. 2005. microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput Biol* **1:** e13.

Gu, J., T. He, Y. Pei, F. Li, X. Wang, J. Zhang, X. Zhang, and Y. Li. 2006. Primary transcripts and expressions of mammal intergenic microRNAs detected by mapping ESTs to their flanking sequences. *Mamm Genome* **17:** 1033-1041.

Han, J., Y. Lee, K.H. Yeom, J.W. Nam, I. Heo, J.K. Rhee, S.Y. Sohn, Y. Cho, B.T. Zhang, and V.N. Kim. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125:** 887-901.

Hebert, S.S., K. Horre, L. Nicolai, A.S. Papadopoulou, W. Mandemakers, A.N. Silahtaroglu, S. Kauppinen, A. Delacourte, and B. De Strooper. 2008. Loss of microRNA cluster *miR*-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression. *Proc Natl Acad Sci U S A* **105:** 6415-6420.

Heimberg, A.M., L.F. Sempere, V.N. Moy, P.C. Donoghue, and K.J. Peterson. 2008. MicroRNAs and the advent of vertebrate morphological complexity. *Proc Natl Acad Sci U S A* **105:** 2946-2950.

Herranz, R., J. Mateos, J.A. Mas, E. Garcia-Zaragoza, M. Cervera, and R. Marco. 2005. The coevolution of insect muscle *Tpn*T and *Tpn*I gene isoforms. *Mol Biol Evol* **22:** 2231-2242.

Hertel, J., M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I.L. Hofacker, and P.F. Stadler. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7:** 25.

Ho, T.H., B.N. Charlet, M.G. Poulos, G. Singh, M.S. Swanson, and T.A. Cooper. 2004. Muscleblind proteins regulate alternative splicing. *EMBO J* **23:** 3103-3112.

Hobert, O. 2008. Gene regulation by transcription factors and microRNAs. *Science* **319:** 1785-1786.

Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res* **31:** 3429-3431.

Johnson, S.M., H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier, K.L. Reinert, D. Brown, and F.J. Slack. 2005. RAS is regulated by the *let*-7 microRNA family. *Cell* **120:** 635-647.

Johnston, R.J. and O. Hobert. 2003. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature* **426:** 845-849.

Jones, D.T., W.R. Taylor, and J.M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8:** 275-282.

Kim, V.N. 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* **6:** 376-385.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16:** 111-120.

Kiriakidou, M., P.T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou. 2004. A combined computational-experimental approach predicts human microRNA targets. *Genes Dev* **18:** 1165-1178.

Korf, I., M. Yandell, and J. Bedell. 2003. *BLAST*. O 'Reilly, Cambridge.

Krichevsky, A.M., K.S. King, C.P. Donahue, K. Khrapko, and K.S. Kosik. 2003. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA* **9:** 1274-1281.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853-858.

Lagos-Quintana, M., R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl. 2002. Identification of tissue-specific microRNAs from mouse. *Curr Biol* **12:** 735-739.

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

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lecellier, C.H., P. Dunoyer, K. Arar, J. Lehmann-Che, S. Eyquem, C. Himber, A. Saib, and O. Voinnet. 2005. A cellular microRNA mediates antiviral defense in human cells. *Science* **308:** 557-560.

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

Lee, S.R. and K. Collins. 2007. Physical and functional coupling of RNA-dependent RNA polymerase and Dicer in the biogenesis of endogenous siRNAs. *Nat Struct Mol Biol* **14:** 604-610.

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

Lin, S.L., J.D. Miller, and S.Y. Ying. 2006. Intronic microRNA (miRNA). *J Biomed Biotechnol* **2006:** 26818.

Liu, J., M.A. Valencia-Sanchez, G.J. Hannon, and R. Parker. 2005. MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nat Cell Biol* **7:** 719-723.

Liu, N., K. Okamura, D.M. Tyler, M.D. Phillips, W.J. Chung, and E.C. Lai. 2008. The evolution and functional diversification of animal microRNA genes. *Cell Res*.

Lu, J., G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando et al. 2005. MicroRNA expression profiles classify human cancers. *Nature* **435:** 834-838.

Lu, R., M. Maduro, F. Li, H.W. Li, G. Broitman-Maduro, W.X. Li, and S.W. Ding. 2005. Animal virus replication and RNAi-mediated antiviral silencing in Caenorhabditis elegans. *Nature* **436:** 1040-1043.

Macrae, I.J., K. Zhou, F. Li, A. Repic, A.N. Brooks, W.Z. Cande, P.D. Adams, and J.A. Doudna. 2006. Structural basis for double-stranded RNA processing by Dicer. *Science* **311:** 195-198.

Maddison, W.P. and D.R. Maddison. 1989. Interactive analysis of phylogeny and character evolution using the computer program MacClade. *Folia Primatol (Basel)* **53:** 190-202.

Makeyev, E.V., J. Zhang, M.A. Carrasco, and T. Maniatis. 2007. The MicroRNA *miR*-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell* **27:** 435-448.

Martin, G., K. Schouest, P. Kovvuru, and C. Spillane. 2007. Prediction and validation of microRNA targets in animal genomes. *J Biosci* **32:** 1049-1052.

Mattick, J.S. 2007. A new paradigm for developmental biology. *J Exp Biol* **210:** 1526-1547.

Megraw, M., P. Sethupathy, B. Corda, and A.G. Hatzigeorgiou. 2007. miRGen: a database for the study of animal microRNA genomic organization and function. *Nucleic Acids Res* **35:** D149-155.

Meyer, A. and M. Schartl. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr Opin Cell Biol* **11:** 699-704.

Miller, C.M., A. Page-McCaw, and H.T. Broihier. 2008. Matrix metalloproteinases promote motor axon fasciculation in the *Drosophila* embryo. *Development* **135:** 95-109.

Miska, E.A., E. Alvarez-Saavedra, M. Townsend, A. Yoshii, N. Sestan, P. Rakic, M. Constantine-Paton, and H.R. Horvitz. 2004. Microarray analysis of microRNA expression in the developing mammalian brain. *Genome Biol* **5:** R68.

Mutsuddi, M., C.M. Marshall, K.A. Benzow, M.D. Koob, and I. Rebay. 2004. The spinocerebellar ataxia 8 noncoding RNA causes neurodegeneration and associates with staufen in *Drosophila*. *Curr Biol* **14:** 302-308.

Nei, M. and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3:** 418-426.

Nei, M., P. Xu, and G. Glazko. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc Natl Acad Sci U S A* **98:** 2497-2502.

Nekrutenko, A., K.D. Makova, and W.H. Li. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res* **12:** 198-202.

Notredame, C., D.G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302:** 205-217.

Olson, S.A. 2002. EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief Bioinform* **3:** 87-91.

Panopoulou, G., S. Hennig, D. Groth, A. Krause, A.J. Poustka, R. Herwig, M. Vingron, and H. Lehrach. 2003. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res* **13:** 1056-1066.

Pasquinelli, A.E., A. McCoy, E. Jimenez, E. Salo, G. Ruvkun, M.Q. Martindale, and J. Baguna. 2003. Expression of the 22 nucleotide *let*-7 heterochronic RNA throughout the Metazoa: a role in life history evolution? *Evol Dev* **5:** 372-378.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* **183:** 63-98.

Pennisi, E. 2000. Human Genome Project. And the gene number is...? *Science* **288:** 1146-1147.

Pfeffer, S., A. Sewer, M. Lagos-Quintana, R. Sheridan, C. Sander, F.A. Grasser, L.F. van Dyk, C.K. Ho, S. Shuman, M. Chien et al. 2005. Identification of microRNAs of the herpesvirus family. *Nat Methods* **2:** 269-276.

Piriyapongsa, J. and I.K. Jordan. 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE* **2:** e203.

Piriyapongsa, J., L. Marino-Ramirez, and I.K. Jordan. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176:** 1323-1337.

Prabhakar, S., J.P. Noonan, S. Paabo, and E.M. Rubin. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314:** 786.

Prachumwat, A. and W.H. Li. 2008. Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes. *Genome Res* **18:** 221-232.

Rajewsky, N. 2006. microRNA target predictions in animals. *Nat Genet* **38 Suppl 1:** S8-S13.

Ren, N., C. Zhu, H. Lee, and P.N. Adler. 2005. Gene expression during *Drosophila* wing morphogenesis and differentiation. *Genetics* **171:** 625-638.

Rendic, D., J. Klaudiny, U. Stemmer, J. Schmidt, K. Paschinger, and I.B. Wilson. 2007. Towards abolition of immunogenic structures in insect cells: characterization of a honey-bee (*Apis mellifera*) multi-gene family reveals both an allergy-related core alpha1,3-fucosyltransferase and the first insect Lewis-histo-blood-group-related antigen-synthesizing enzyme. *Biochem J* **402:** 105-115.

Rivas, E. and S.R. Eddy. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16:** 583-605.

Rizzolio, F., S. Bione, C. Sala, C. Tribioli, R. Ciccone, O. Zuffardi, N. di Iorgi, M. Maghnie, and D. Toniolo. 2008. Highly conserved non-coding sequences and the 18q critical region for short stature: a common mechanism of disease? *PLoS ONE* **3:** e1460.

Roark, M., M.A. Sturtevant, J. Emery, H. Vaessin, E. Grell, and E. Bier. 1995. *scratch*, a pan-neural gene encoding a zinc finger protein related to snail, promotes neuronal development. *Genes Dev* **9:** 2384-2398.

Robins, H., Y. Li, and R.W. Padgett. 2005. Incorporating structure to predict microRNA targets. *Proc Natl Acad Sci U S A* **102:** 4006-4009.

Saini, H.K., S. Griffiths-Jones, and A.J. Enright. 2007. Genomic analysis of human microRNA transcripts. *Proc Natl Acad Sci U S A* **104:** 17719-17724.

Sandmann, T. and S.M. Cohen. 2007. Identification of novel *Drosophila* melanogaster microRNAs. *PLoS ONE* **2:** e1265.

Sempere, L.F., C.N. Cole, M.A. McPeek, and K.J. Peterson. 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zoolog B Mol Dev Evol* **306:** 575-588.

Sethupathy, P., B. Corda, and A.G. Hatzigeorgiou. 2006. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA* **12:** 192-197.

Shabalina, S.A., A.Y. Ogurtsov, V.A. Kondrashov, and A.S. Kondrashov. 2001. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet* **17:** 373-376.

Shannon, C.E. 1997. The mathematical theory of communication. 1963. *MD Comput* **14:** 306-317.

Shpiz, S., D. Kwon, A. Uneva, M. Kim, M. Klenov, Y. Rozovsky, P. Georgiev, M. Savitsky, and A. Kalmykova. 2007. Characterization of *Drosophila* telomeric retroelement TAHRE: transcription, transpositions, and RNAi-based regulation of expression. *Mol Biol Evol* **24:** 2535-2545.

Simons, C., M. Pheasant, I.V. Makunin, and J.S. Mattick. 2006. Transposon-free regions in mammalian genomes. *Genome Res* **16:** 164-172.

Song, J.J., J. Liu, N.H. Tolia, J. Schneiderman, S.K. Smith, R.A. Martienssen, G.J. Hannon, and L. Joshua-Tor. 2003. The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nat Struct Biol* **10:** 1026-1032.

Stajich, J.E. 2007. An introduction to BioPerl. *Methods Mol Biol* **406:** 535-548.

Stark, A., J. Brennecke, R.B. Russell, and S.M. Cohen. 2003. Identification of *Drosophila* MicroRNA targets. *PLoS Biol* **1:** E60.

Stark, A., P. Kheradpour, L. Parts, J. Brennecke, E. Hodges, G.J. Hannon, and M. Kellis. 2007. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Res* **17:** 1865-1879.

Storey, J.D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100:** 9440-9445.

Sugnet, C.W., K. Srinivasan, T.A. Clark, G. O'Brien, M.S. Cline, H. Wang, A. Williams, D. Kulp, J.E. Blume, D. Haussler et al. 2006. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput Biol* **2:** e4.

Taft, R.J., M. Pheasant, and J.S. Mattick. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29:** 288-299.

Takamizawa, J., H. Konishi, K. Yanagisawa, S. Tomida, H. Osada, H. Endoh, T. Harano, Y. Yatabe, M. Nagino, Y. Nimura et al. 2004. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res* **64:** 3753-3756.

Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. *Science* **278:** 631-637.

Thomas, J.W., J.W. Touchman, R.W. Blakesley, G.G. Bouffard, S.M. Beckstrom-Sternberg, E.H. Margulies, M. Blanchette, A.C. Siepel, P.J. Thomas, J.C. McDowell et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788-793.

Thomassen, G.O., O. Rosok, and T. Rognes. 2006. Computational prediction of microRNAs encoded in viral and other genomes. *J Biomed Biotechnol* **2006:** 95270.

Tran, T., P. Havlak, and J. Miller. 2006. MicroRNA enrichment among short 'ultraconserved' sequences in insects. *Nucleic Acids Res* **34:** e65.

Ureta-Vidal, A., L. Ettwiller, and E. Birney. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* **4:** 251-262.

Vacic, V., L.M. Iakoucheva, and P. Radivojac. 2006. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **22:** 1536-1537.

Vanolst, L., C. Fromental-Ramain, and P. Ramain. 2005. Toutatis, a TIP5-related protein, positively regulates Pannier function during *Drosophila* neural development. *Development* **132:** 4327-4338.

Wall, D.P. and T. Deluca. 2007. Ortholog detection using the reciprocal smallest distance algorithm. *Methods Mol Biol* **396:** 95-110.

Warren, W.C. L.W. Hillier J.A. Marshall Graves E. Birney C.P. Ponting F. Grutzner K. Belov W. Miller L. Clarke A.T. Chinwalla et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453:** 175-183.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Weaver, D.B., J.M. Anzola, J.D. Evans, J.G. Reid, J.T. Reese, K.L. Childs, E.M. Zdobnov, M.P. Samanta, J. Miller, and C.G. Elsik. 2007. Computational and transcriptional evidence for microRNAs in the honey bee genome. *Genome Biol* **8:** R97.

Weber, M.J. 2005. New human and mouse microRNA genes found by homology search. *FEBS J* **272:** 59-73.

Wienholds, E., M.J. Koudijs, F.J. van Eeden, E. Cuppen, and R.H. Plasterk. 2003. The microRNA-producing enzyme Dicer1 is essential for zebrafish development. *Nat Genet* **35:** 217-218.

Witten, I.H. and E. Frank. 2000. *Data mining : practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, CA.

Xu, P., M. Guo, and B.A. Hay. 2004. MicroRNAs and the regulation of cell death. *Trends Genet* **20:** 617-624.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586-1591.

Ying, S.Y. and S.L. Lin. 2004. Intron-derived microRNAs--fine tuning of gene functions. *Gene* **342:** 25-28.

Zeng, Y. and B.R. Cullen. 2003. Sequence requirements for micro RNA processing and function in human cells. *RNA* **9:** 112-123.

Zhang, B., Q. Wang, and X. Pan. 2007. MicroRNAs and their regulatory roles in animals and plants. *J Cell Physiol* **210:** 279-289.

Zhao, Y. and D. Srivastava. 2007. A developmental view of microRNA function. *Trends Biochem Sci* **32:** 189-197.

APPENDIX A

Position of ultraconserved elements and subsitution pattern of known microRNAs from honey bee – wasp

MicroRNA	Class	Percent identity
Bantam	2	84
ame-bantam AAACGAAACTGGTTTTCACAATGATTTGACAGATAGATTCGATTCtgagatcattgtgaaagctgattTTGTTGAAAAG nvi-bantam AGACGAAACTGGTTTTCACAATGATTTGACAGATGTATTTGATTCtgagatcattgtgaaagctgattTTGTTCCGAAT Conservation * ******************************** *** ********************************* **		
Let-7	3	88
ame-let-7 ACGATGCCTGGGtgaggtagtaggttgtatagtAGGGAATGGAAATTCGCGATATACGAAGTCCACTGTACAACTTGCTAACTTTCCCGGTCGTCGACGC nvi-let-7 GCGACGCCTGGGtgaggtagtaggttgtatagtGGGGAATGAAAATTCGTCAT-TATGGAATCCACTGTACAACTTGCTAACTTTCCCGGTTGTCGGCGC Conservation *** **************************** ******* ******* ** ** * * ****************************** **** ***		
Mir-1	2	87
ame-mir-1 CCGGGCGATGCTGTTCCGTGCTTCCTTACTTCCCATAGTGGATGCGACGTAtggaatgtaaagaagtatggagCTGCGCCCGG nvi-mir-1 CCGAGCGT--CTGTTCCGTGCTTCCTTACTTCCCATAGTGCACGTGACGTAtggaatgtaaagaagtatggagCCGCGCTCTG Conservation *** *** ****************************** * * ***************************** **** * *		
Mir-10	3	80
ame-mir-10 AATGCTCTACATCTaccctgtagatccgaatttgtTTGATAAGAGGCGACAAATTCGGTTCTAGAGAGGTTTGTGTGGTGCATA nvi-mir-10 AATGCTCTACATCTaccctgtagatccgaatttgtTTGAAGTCAGGCGACAAATTCGGTTCTAGAGAGGTTTGTGTGGTGCATA Conservation *************************************** ***		
Mir-100	3	92
ame-mir-100 --GCTATACTTGATGATACTaacccgtagatccgaacttgtgGGCTTTTTTATATGTATACCGCAAGCTCCTATCTACCGGTACATGTAGTCAGGCCAGCAT nvi-mir-100 CAGC--TACTTGATGACACTaacccgtagatccgaacttgtgGGCTTTTTTATATGCATACCGCAAGCTCCTATCTACCGGTACATGTAGTCAGGCCGGCGT Conservation ** ********** *************************************** ** ** *		
Mir-12	3	82
ame-mir-12 AAGACATGGGTGtgagtattacatcaggtactggtGTG—-ATATTCAGACAACCAGTACTTGTGTTATACTTACGCTCATGTCTT nvi-mir-12 AAGAGGAGGGTGtgagtattacatcaggtactggtGTCTTTTTTTCAG-CGACCAGTACTTGTGTTATACTTGCGTTCGCTTCTT Conservation **** ****************************** * ***** * ********************* ** ** ****		
Mir-124	3	97
ame-mir-124 TGCTCCTTGCGTTCACTGCGGGCTTCCATGTGCCAACTTTTCAAAATTCAtaaggcacgcggtgaatgccaagAGCG nvi-mir-124 TGCTCCTTGCGTTCACTGCGGGCTTCCATGTGCCAAGTTTTAAAAATTCAtaaggcacgcggtgaatgccaagAGCG Conservation ************************************ **** ***********************************		
Mir-125	3	92
ame-mir-125 GTAAAGCCT---GCCGCGTCGCCGGTcccctgagaccctaacttgtgaCGTCGCGACCGATATCTCACAGGCTAGATTCTCTGGTATTGGCGATGAGTGCTGCCTTTTGC nvi-mir-125 ATAAAGCCTGCCGCCGCGTCGCCGGTtccctgagaccctaacttgtgaCGTCGCGTACGATATCTCACAGGCTAGATTCTCTGGTATTGGCGATGAGTGCTGCCTTTTGA Conservation ******** ************** **************************** **		
Mir-133	5	98
ame-mir-133 TAATGTTAAGCTTAGCTGGTTGAACACGGGTCAAATATATCGCACGATTGACGCATttggtccccttcaaccagctgtAGTTGACATTA nvi-mir-133 TAATGTTAAGCTTAGCTGGTTGAACACGGGTCAAATATAACGCACGATTGACGCATttggtccccttcaaccagctgtAGTTGACATTA Conservation *************************************** ***		

```
                                        Mir-137                                                  3      91
ame-mir-137    GACTTCATAGGCCAGGTTGGCGACGCGTATTCTTGGGGAATTAACACACATTTGCGCTGttattgcttgagaatacacgtaGTTTGCCTGGTCGTTCACT
nvi-mir-137    TACATCGGAGACCAGGTTGGCGACGCGTATTCTTGGGGAATTAACACACATACGCGCTGttattgcttgagaatacacgtaGTTTGCCTGGTCGATCTCT
Conservation    ** **  ** ****************************************  ***************************************** ** **
                                        Mir-13a                                                  3      80
ame-mir-13a    ----ACCGAAATGAAAATACCTTTTGCGGTCCGATACATCAAATTGGTTGTGGAATGTTTCGAGT---CAtatcacagccattttgatgagCTTGGCCCGCAGAATC
nvi-mir-13a    GAGCAGAGACATGGAA----CTTTTGCGGTCCGATACATCAAATTGGTTGTGGAATGTCTCGTCTTTCCAtatcacagccattttgatgagCTTGGCCCGTAGAACT
Conservation       *  ** *** **    ************************************** ***  *   ******************************** ****
                                        Mir-14                                                   3      76
ame-mir-14     CTTTTTCTCGGTCGCTAGGTCAGTGGGGGTGAGAAACTGGCTTGGCTCTCTGTGCTAC--------GATAGtcagtcttttctctcctaTCGGCTTTGCGACATA
nvi-mir-14     CTTTTTAT---TCGCTCGGTCAGTGGGGGTGAGAAACTGGCTTGGCT--ATGTACTCCCCATTGTGGACAGtcagtcttttctctcctaTTGGCCCGGCGAGAAT
Conservation   ****** *   ***** ******************************   *** ** *        ** ********************* ***   **** *
                                        Mir-184                                                  3      85
ame-mir-184    TTCGTGCCCAAAGCCCCTTATCATTCTCCTGTCCGGTGTAGAATTGTTAGACGACtggacggagaactgataagggcCCGAGGGTCACAGAA
nvi-mir-184    AACGTGCCTAAAGCCCCTTATCATTCTCCTGTCGAGTGTTAAAATTCTCTACGACtggacggagaactgataagggcCCGAGGGTTACAGAT
Conservation     ****** ************************  ****  ** *  *  *********************************** *****
                                        Mir-190                                                  6      72
ame-mir-190    CAAACAAGTCGTCTGGTTTCCGTAagatatgtttgatattcttggttgttTTTTAAAGAATCGACCAGGAATCAAACATATTATTATGGTGGTCAGAAAA
nvi-mir-190    GAATCCATTCC-CAGTATACTGCTagatatgtttgatattcttggttgtaTAATAATAA--CGACCAGGAATCAAACATATTATTACAGTG-TCTGGTTT
Conservation    ** * * **  * *  * * *  ************************* *  ***  *  *************************  *** ** *
                                        Mir-2-1                                                  3      70
ame-mir-2-1    GGCGCGTGTGCACCGCTCACAAAGTGGTTGTGATATG-CTGAT-ACGAGCGTTCAtatcacagccagctttgatgagc-GTG-GCGTCGCGTC
nvi-mir-2-1    GTGGAGATCGCGTCGCTCACAAAGTGGCTGTTGTATGTCGGATTTCTTTGGCTCAtatcacagccagctttgttgagcGGTGAGCGTCTTCCT
Conservation   *  * *   **  ************** ***  **** * ***  *    * ******************** ***** *** *****
                                        Mir-2-2                                                  6      81
ame-mir-2-2    TCGACTGTTCCTCCCATCAGAGTGGTTGTGATGTGGTA-ACTTGGACTCGtatcacagccagctttgatgagcGGAACGGTGCGA
nvi-mir-2-2    TCGGCTGTTTCGCCCGTCAGAGTGGTTGTGATATGGTGCTATTGAACGCAtatcacagccagctttgatgtgcGTAACAGTTCGA
Conservation   *** ***** * *** **************** ****    *** ** * ******************** *** *** ** ***
                                        Mir-2-3                                                  3      80
ame-mir-2-3    AAATATCCCCGGACAAG-GACATGCTTTTACCATCAAAGTTGGTTTGTCATAGAGA-TCGACtatcacagccagctttgatgagcAAAATTGTGTCCGTCTA
nvi-mir-2-3    GGAGACAAGAGAGCGAGTGGCATGCTTTTACCATCAAAGCTGGTTTGTCATAGGGCTTCGACtatcacagccagctttgatgagcAAAATTGTGTCCGGCAT
Conservation     * *     *  * ** * ******************* ************* *  ***************************************** *
                                        Mir-219                                                  3      69
ame-mir-219    AATTGAATGTCTCAGGCAAtgattgtccaaacgcaattcttgTCTAAACGGTACGAAATCAAGAATTGTGTGGGGACATCAGCGCTCGAGGTGCGATTCAAC
nvi-mir-219    ---------TCTCGGGCTAtgattgtccaaacgcaattcttgTCTGTGCCTTGAGATACCAAGAATTGTGTGGGGACATCAGCGCTCG------GAG-----
Conservation            **** *** ***************************   *  *  ** * *****************************      **
                                        Mir-263                                                  3      90
ame-mir-263    AGCTTGGACTCTgtaaatggcactggaagaattcacGGGGGATTTAAGAAACGGGCCCGTGGAGCTCCCGTGTCATACACAGCGTCCGGCT
nvi-mir-263    AGCTTGGGTTCTgtcaatggcactggaagaattcacGGGGGAATTTAGCAACAGTCCCGTGGAACTCCCGTGTCATACACAGCGTCCGGCT
Conservation   *******  ***** *************************** ** ** *** * ******** ***************************
                                        Mir-275                                                  3      73
ame-mir-275    AAACGTTACTTGTCGTGCGCAACGCGCGTTACTCGGGTACTTTAGGCTGTGCCAATTTCGAATCAGtcaggtacctgaagtagcgcgcgCTGCGGCGAAA
nvi-mir-275    TTACGT-ACAGCCAGTTTGCAACTCGCGCTACTCCGGTACTTACGACTGTGC---ATTCGAT--AGtcaggtacctgaagtagcgcgcgCTGCGACTGGA
Conservation     **** **     **  ***** **** ***** *******  * ******    *****   ****************************** *   *
```

```
Mir-276                                                                                    2      98
ame-mir-276     TGGTAGAGATCCAGCAGCGAGGTATAGAGTTCCTACGTAGTGTTCAGAAAGtaggaacttcataccgtgctcttGGACTTGCCG
nvi-mir-276     TGGTAGAGATCCAGCAGCGAGGTATAGAGTTCCTACGTAGATTTCAGAAAGtaggaacttcataccgtgctcttGGACTTGCCG
Conservation    ****************************************  ******************************************
Mir-277                                                                                    3      93
ame-mir-277     GGCAG-TTGGGGCTCGTGCCAGATGCGCGTTTACAC--GGGCCCTGAATACTGtaaatgcactatctggtacgacaTCTCTCCTGTC
nvi-mir-277     GGCAGGCTGGGGCTCGTGCCAGATGCGCGTTTACACGAGAGCCCTGAATACTGtaaatgcactatctggtacgacaTCTCTCCTGCC
Conservation    *****  *****************************  * ********************************************* *
Mir-279                                                                                    6      66
ame-mir-279     -----AGAAAATGAAAAAATTTCCTGAATTTGCCAAATGAGTGAAGGTCTAGTGCACAGAAAATGAAATTGtgactagatccacactcattaaGTACGTTCAGGT
nvi-mir-279     CCATCGGACAACGACCAG-----CCGATTGTACTGAGTGAGTGATGGTCTGGTGCACGGTTTATCGATCTGtgactagatccacactcattaaGTACGTTCGGCT
Conservation          ** ** **  *      * ** * * *  * ******* ***** ****** *   **  *  ******************************** * *
Mir-282                                                                                    2      84
ame-mir-282     GGACAGAGTAACTTgatttagcctctcctaggctttgtctgtATATAAAGAACGGAGACATAGCCTAGAATAGGTTAGGTCAGGGCTCGTTC
nvi-mir-282     ATATCGAGCGACGTgatttagcctctcctaggctttgtctgtCAGTGAAAAAACGAGACATAGCCTAGAATAGGTTAGGTCGGGGCTCGTTC
Conservation      *  ***  ** *****************************   * ** **  *************************** **********
Mir-283                                                                                    3      80
ame-mir-283     AATAATCTGGTGATGTAGTCaaatatcagctggtaattctGGGATTTTGACAAT--AACCCAGGATTCTTGCTGGTATCCGGCTACGAACTGGACGATCGCC
nvi-mir-283     TGCTGTCCAGTCACGTAGTCaaatatcagctggtaattctGGGATACTTATTATGCAGCCCAGGATTCTTGCTGGTATCCGGTTACGAACTGGCTCGTCGCC
Conservation         **  ** * *******************************  * *  **  * ************************ **********    *****
Mir-29b                                                                                    3      80
ame-mir-29b     ATTTAAAGACAATAAGAAGATAGAGGTACTGACTTCTATGCGTGCTGGGGTTTGTGCTAAATCTCCtagcaccatttgaaatcagtACTACTCTTCTTAG
nvi-mir-29b     ACACGAAGA-ATTAAGATAGCAGAGGTACTGACTTCGGTGCGTGCTGGGGTTACCGATCAAGCGCCtagcaccatttgaaatcagtACTACTCGTCTTAG
Conservation    *    **** * *****    ***************  **************   * * ** * ***************************** ******
Mir-305                                                                                    6      92
ame-mir-305     GGAGGCTGCATGTTAattgtacttcatcaggtgctctgGTGAACTCGATACCCGGCACCTGTTGGAGCGCAATTCATATGACTGTGCCCT
nvi-mir-305     GGAGGCTGCATGTTAattgtacttcatcaggtgctctgGTGATTTTGATACCCGGTGCCTGTTGGAGCGCAATTCATATGATTGTGCCTT
Conservation    ******************************************  * *********  ************************ ****** *
Mir-315                                                                                    6      89
ame-mir-315     GCTCTTTATGCttttgattgttgctcagaaagcCTTGATTATGATATTGGCTTTCGGGCAATAATCATAATCACGAAAGGGT
nvi-mir-315     GCTCTCTATGCttttgattgttgctcagaaagcCTCGATATCGATACTGGCTTTCGGGCAATAATCATAATCACGGGAGAGT
Conservation    ***** ***************************** ***   **** ****************************  ** **
Mir-317                                                                                    2      95
ame-mir-317     GCTCTCGGAGAACAGGGAGCCACTCTGCGTTCACTCGGTGGGTAATGAAGCGGGtgaacacagctggtggtatctcagtTTTCTGAGGGC
nvi-mir-317     GTGCTCGGAGAACAGGGAGCCACTCTGCGTTCACTCGGTAGGTAAAGAAGCGGGtgaacacagctggtggtatctcagtTTTCTGAGGGC
Conservation    *  ************************************ ***** ********************************************
Mir-31a                                                                                    3      78
ame-mir-31a     ATCACGATTCTAACTGGGCGCCTCGAAggcaagatgtcggcatagctgaTGCGATTTTAAAATTCGGCTGTGTCACATCCAGCCAACCGAACGCTCAGAC
nvi-mir-31a     TGCAGCGTAAAATATGAGCGTGTAGAAggcaagatgtcggcatagctgaTGC-TTTTGAAATTTCGGCTGTGTCACATCCAGCCAGCCAAACGCTCAAAA
Conservation      **   *   *  ** ***  * ****************************  *** *** *********************** ** ******** *
Mir-33                                                                                     3      70
ame-mir-33      TATTTATTTGATTGCTTACCTGTTACAACTgtgcattgtagttgcattgCATGAA-ATATAACTATGCAATACTTCTACAGTGCAACTCTTGTGGCAGATT
nvi-mir-33      TATGCAACGAAAATATTGTCTGCTACGAGAgtgcattgtagttgcattgCATGAATATTTGAAAGTGCAGTACTTCTGCAGTGCAACCCTTGTGGTTGGCG
Conservation    ***  *    *    **  *** *** *  ************************* ** * *   **** ******* ********* *******  *
```

```
                                       Mir-375                                                   3         67
ame-mir-375   ---ATCGATTGAATTATCAGTTTGGTGCATCGATCCTAACGATCAACAAACTTTTCACTTGA--AAGtttgttcgttcggctcgagttaTCAAACTGAATGG-ATG
nvi-mir-375   CTCAGCAATAGGCTT--CAC-CTGGTCCATCGATCCCAACGATCAATAAACTGTGTTATTAAACAAGtttgttcgttcggctcgagttaT-TAGGTGAGCCTTGTC
Conservation     * * ** *  **  **   **** ********* ********* ***** *    ** *  **************************  *  ***      *
                                       Mir-7                                                     3         92
ame-mir-7     CGAGCGCCGTTGCAtggaagactagtgattttgttgtTCTACTTTCGATATAACAAGGAATCACTAATCATCCTACAAAGGCGCTCG
nvi-mir-7     TGAGCGTCGTTGTAtggaagactagtgattttgttgtTCTACTTAAGATGTAACAAGGAATCACTAATCATCCTACAAAGACGCTCG
Conservation   ***** ***** *******************************  *** ****************************** ******
                                       Mir-71                                                    3         74
ame-mir-71    G-TCCTCCTTCGGGCGGATTCCGTCtgaaagacatgggtagtgaGATG-TTCTCACGCTATCGCGTCTCACTATCTTGTCTTTCATCCGGCGTTCGTTCTGC--
nvi-mir-71    GAAAATCAATTGGCGAGATTCCGCTtgaaagacatgggtagtgaGATGCTTACCTCGGATTCGCGTCTCACTACCTTGTCTTTCATGCGGCGCTCG--CAGCAA
Conservation  *    **  * **   *******  *********************** **  * **   ************* ************ ***** ***  * **
                                       Mir-8                                                     3         92
ame-mir-8     GGAGTATCTGTTCACATCTTACCGGGCAGCATTAGATTGAAGTTGA-CCTTCtaatactgtcaggtaaagatgtcGTCAGGATTCC
nvi-mir-8     GGAGTATCTGTTTACATCTTACCGGGCAGCATTAGATTACATTTGAATTTTCtaatactgtcaggtaaagatgtcGTCAGGATTCC
Conservation  ************ *************************  * ****   *************************************
                                       Mir-927                                                   3         78
ame-mir-927   AGATAAAAGCGTGGTATTTGttttagaattcctacgctttacc-GATGTTCGAAGTGGCAAAGCGTTTGAAATCTGAAACGAATGCGCATAA-ACCTTCATC
nvi-mir-927   AGACAATAAGGTGATATTTGttttagaattcctacgctttaccGGTTTTTAAAAATGGCAAAGCGTTTGAAGTCTAAAACAAAAGCACATAACAACTCGCTG
Conservation  *** ** *  *** ***************************** * * **  ** **************** *** **** ** ** ***** * **   *
                                       Mir-929                                                   5         95
ame-mir-929   ACTTAACTGGGGTCAAattgactctagtagggagtccCTGCATTCAATATGGCGACTTCCTAATAGAGTCAGGCTGACTCCTTTTAAGACGTTCAACGGA
nvi-mir-929   ACTTAACTGGGGTCAAattgactctagtagggagtccCTGCATTCAATATGGCGACTTCCTAACAGAGTCAGGCTGACTCCTTTTAAGACGCTCACCGAC
Conservation  *************************************************************** *************************** *** **
                                       Mir-92a                                                   6         83
ame-mir-92a   TTATTTTGCATAGAAGATAGGCCGAGATTTGTGACAATGTTTCGTGATGATGTAATCTTCAATattgcacttgtcccggcctatCGGAATGCATTATATT
nvi-mir-92a   GTTTGTTGCATAAAAGATAGGTCGAGATCGGTGGCAATGTTTCGTAATGGTTTT-CCCTCAATattgcacttgtcccggcctatCGGAATGCACTA-ATT
Conservation   * * ******* ******** ******  *** *********** *** * *   * *********************************** ** ***
                                       Mir-9a                                                    3         92
ame-mir-9a    TGGCGCGGACATTTtctttggttatctagctgtatgaGTATTATTCGACATCATAAAGCTAGGTTACCGGAGTTAAGCTCCTCGCCA
nvi-mir-9a    TGGCGCGGACATTTtctttggttatctagctgtatgaGTTTGGTTGTACATCATAAAGCTAGGTTACCGGAGTTGAGATCCTCGCCA
Conservation  *************************************** *  **  *************************** ** *********
                                       Mir-iab-4                                                 1         100
ame-mir-iab-4  GTGAAACCCCCTGTacgtatactgaatgtatcctgaGTGTATTTCTGTCCGGTATACCTTCAGTATACGTAACAGGAGGCTACAC
nvi-mir-iab-4  GTGAAACCCCCTGTacgtatactgaatgtatcctgaGTGTATTTCTGTCCGGTATACCTTCAGTATACGTAACAGGAGGCTACAC
Conservation   *************************************************************************************
                                       Mir-210                                                   4         95
ame-mir-210   TGGACCCTAATGCAGCTGCTGGCCACTGCACAAGATTAGACATAAGACTCttgtgcgtgtgacagcggctaTGATGGGGTTTCCA
nvi-mir-210   TGGACCCCAGTGCAGCTGCTGGCCACTGCACAAGATTAGACATAAGACTCttgtgcgtgtgacagcggctaTGATGGGGCTTCCG
Conservation  ******* * ********************************************************************* ****
                                       Mir-281                                                   3         91
ame-mir-281   GCGCGCGCTATAAAGAGAGCTATCCATCGACAGTATGGTTATAATAGACACtgtcatggagttgctctctttgtAGACACTGCT
nvi-mir-281   GCGGACGCTATAAAGAGAGCTATCCATCGACAGTATGGTGATAGATGACACtgtcatggagttgctctctttgtGGACGTCGCT
Conservation  ***  ********************************** ***   **************************** ***   ***
```

Mir-307		6	78
ame-mir-307	AAATGGCGGTCACGTGGACTCACTCAACCTGGGTGTGATGCTTGCCTGTGTA-----TCA-GGCCCTAGCGGTCAtcacaacctttttgagtgagCGAACG----CGACTG		
nvi-mir-307	AGATGGCGACCGCGTGAACTCACTCAACCTGGGTGTGATGCTTGCCCGTGAAATTGAACAGGGCCCTTGCGGTCAtcacaacctttttgagtgagCGAACGCGCTCGTCAG		
Conservation	* ****** * **** ***************************** *** * ********* ********************************* ** * *		
Mir-932		3	83
ame-mir-932	CGCGTTGCCTCTtcaattccgtagtgcattgcaGATGATTGTTCGAATTGACGAGAAAGAACCTGCAAGCACCGCGGGAGTGAGGTGGCCTCGCG		
nvi-mir-932	CGCGTTGCCTCTtcaattccgtagtgcattgcaGTAGA---------TTTAAAAGAAAGATTCTGCAAGCACCGCGGGAGTGAGGCGGCTTCGTG		
Conservation	********************************* ** ** * ****************************** *** *** * *		
Mir-9b/79			
ame-mir-9b	TGGCGCGGACATTTTCTTTGGTTATCTAGCTGTATGAGTTTGGTTGTACATCATAAAGCTAGGTTACCGGAGTTGAGATCCTCGCC		
nvi-mir-9b	TGGCGCGGACATTTTCTTTGGTTATCTAGCTGTATGAGTATTATTCGACATCATAAAGCTAGGTTACCGGAGTTAAGCTCCTCGCC		
Conservation	*************************************** * ** *************************** ** ********		

APPENDIX B

MicroRNA candidates supported by 454 data

<table>
<tr><th>RANK/ID</th><th>Score</th><th>Contig</th><th>Pos1</th><th>Pos2</th><th>Overlap</th><th>Strand</th><th>Worm</th><th>Anopheles</th><th>Drosophila</th><th>Human</th><th>Class</th></tr>
<tr><td>18</td><td>13.61611111</td><td>Group4.13</td><td>369199</td><td>369276</td><td>Intergenic</td><td>+</td><td>0</td><td>0</td><td>1</td><td>0</td><td>Class 3</td></tr>
<tr><td colspan="12">CTGGCGGGTCTTGGCACTGGAAGAATTCACAGATGTGCAG---TGTATAATCGTGGATCTTGCAATGCCATCACTTGCTGG
CTGGCGGGTCTTGGCACTGGAAGAATTCACAGATGTGCGGTTGTGTGTAATCGTGGATCTTGCAATGCCATTACCCGCTGG
************************************ *   *** ************************* **   *****
(..((((((..(((((.((.((((.(((((.((((((((....))))).))))))))))))).)).)))))..))))))..)</td></tr>
<tr><td>22</td><td>13.53487013</td><td>Group1.1</td><td>136120</td><td>136195</td><td>Intronic:
GB15727</td><td>-</td><td>0</td><td>1</td><td>1</td><td>0</td><td>Class 3</td></tr>
<tr><td colspan="12">GATCCAATCGTCAAATTGGTTGTGGCGTGTTGCTTTTCTAGAT-TTCATATCACAGCCATTTTTGACGATTTGGATC
GATCCGATCGTCAAATTGGTTGTGGCGTGTTG----TCTAGATTTTCATATCACAGCCATTTTTGACGAGTTGGATC
***** ***************************   ******* ************************** *******
((((((((((((((.((((((((..(((.................)))..))))))))..)))))))).)))))))</td></tr>
<tr><td>34</td><td>13.10424242</td><td>Group16.11</td><td>38051</td><td>38181</td><td>Intergenic</td><td>+</td><td>0</td><td>1</td><td>1</td><td>0</td><td>Class 6</td></tr>
<tr><td colspan="12">GTGGTTGCATAAGATAGGCACATTCGTGATCTACCCTGTAGATCCGGGCTTTTGTAGAATTGTAAATATCAGAAGCTCGTCTCTACAGGTATCTTACGGATGACATGCCACGCGACTCTAGATTGCA-AT
GTGTTTGCGAAAGATAGGCACATTCGTGATCTACCCTGTAGATCCGGGCTTTTGTAGAATTGTAGATATCAGAAGCTCGTCTCTACAGGTATCTTACGGATGACATGCCACGCGACTCTTTATCGCAGAT
*** **** ** ** ** *** **
((((((((((((((..((((((((((((((..((((.((((((..((((((((((...............))))))))))..))))))))))..))))))))))...))))).......)))))..)))))))</td></tr>
<tr><td>42</td><td>12.84833333</td><td>Group2.34</td><td>226591</td><td>226677</td><td>Intergenic</td><td>+</td><td>0</td><td>0</td><td>0</td><td>0</td><td>Class 3</td></tr>
<tr><td colspan="12">TGGATGCAAACGTCTGGGTTTCGTGACAGGCGAGCCGTT--CTTTACGACTTGGTTCGTTGTCAACGAAACCTGCACGATCTGCAACCA
TGGACGCAGACGTACGGGTTTCGTGACAGGCGAGCCGTTTGAATTA--ACGTGGTTCGTTGTCGACGAAACTTGCACGATTTGCAACCA
**** *** **** ************************   ***   ** ************ ******* ******** ********
.((.(((((((((.((((((((((((((((.(((((((..............))))))))))))).))))))))).))).))))))).)).</td></tr>
<tr><td>48</td><td>12.57744898</td><td>Group2.17</td><td>22929</td><td>23021</td><td>Intronic:
GB16497</td><td>-</td><td>0</td><td>0</td><td>0</td><td>0</td><td>Class 3</td></tr>
<tr><td colspan="12">TGGTAGCGTTGGAGTGGGTAATCTCATGCGGTAACTGTGAGTGTGTGAATTGTAAAAGCTCATATTACCTCGTGGGGTTTCCCACCCGTTACC
TGGTAGTGTCTGTGTAGGTAATCTCATGCGGTAACTGTGAGCGTGTGAAAAACTATGGCTCATATTACCTCGTGGGATTTCCCACCCACTACC
****** **   * ** ************************* *******   *   ******************* ********** ****
.(((((((...(.(((((.((((((((((.(((((.(((((((...............))))))))))))).))))))))).)))))))))))))</td></tr>
<tr><td>51</td><td>12.44766667</td><td>Group9.15</td><td>253723</td><td>253795</td><td>Intergenic</td><td>+</td><td>0</td><td>0</td><td>0</td><td>0</td><td>Class 3</td></tr>
<tr><td colspan="12">CCTGATGGAGCTCTGGCTGTGACTTGTGTGTTAACTGAATTAAAATCACATCACAGGCAGAGTTCTAGTTAGG
CCTGGTGGGACTCTGTCTGTGACTTGTGTAGTC-TTGAAT---GGTCACATCACAGGCAGAGTTCTAGTTAGG
**** *** ***** *************   *   *****   ****************************
((((((((((((((((((((..((((...................)))))))))))))))))))))))).)))))</td></tr>
</table>

<table>
<tr><td>62</td><td>11.73688406</td><td>Group13.6</td><td>502702</td><td>502765</td><td>Intronic:
GB15055</td><td>-</td><td>0</td><td>0</td><td>0</td><td>0</td><td>Class 3</td></tr>
<tr><td colspan="12">GGTAGGTAACGACTGATGGGAACACTCTGAGATT--TTCTTTAATGTTCTCTTTGGTTGTTACCAC
GGTAGGTAACGACTGATGGGAACACATAGATAATCAAACTAACATGTTCTCTTTGGTTGTTACCAC
*************************   ** * *   **   **********************
....(((((((((..(.(((((((..(((((.....)))))...))))))).)..)))))))))..</td></tr>
<tr><td>69</td><td>11.49905941</td><td>Group10.18</td><td>8982</td><td>9080</td><td>Intronic:
GB13125-RA</td><td>-</td><td>0</td><td>0</td><td>0</td><td>0</td><td>Class 6</td></tr>
<tr><td colspan="12">GAGGATGCACCGGGGCCGGTTCACTTTAAGTTCGAATACCAAGCCTGTCTATA-AACAGTTTGGTGTTCTACCTTACAGTGAGTCGACCGTGGTATCGTC
GAGGAGAAACCGTAGGCGGGTCACTCTAAGTTCGAATACCAAGCCCTGCTGGAGAATAGCTTGGTGTTCTACCTTACAGTGATTTGATTGCGGTCACGGC
*****   **** * *** ***** ******************   ** * ** ** ********************** * **   * ***   ** *
........(((((((.(((((((((.((((...(((((((((((.............)))))))))))...)))).))))))))).)))))))......</td></tr>
<tr><td>78</td><td>11.03595238</td><td>Group15.13</td><td>54829</td><td>54909</td><td>Intronic:
GB16072-RA</td><td>-</td><td>0</td><td>0</td><td>0</td><td>0</td><td>Class 6</td></tr>
<tr><td colspan="12">AGACTGCTGTCGGATGAAATCTCGTCCGGTGTGGTT-GGAAA--AAAAAAACCTCACCGGGTAGGATTCATCCAATGTCCGCCT
AGGCTGCTGTCGGGTGGAATC----CCGCT-TGGTTAGGAAATTATAAATACCTCACCGGGTAGGATTCATCCAATATCAGCCT
** ********** ** ****   *** * ***** ***** * *** ************************** ** ****
((((((.(((.((((((((.(((..((((((((.(((...............))).))))))))..)))))))))).))).))))))</td></tr>
<tr><td>85</td><td>10.55820513</td><td>Group4.16</td><td>137099</td><td>137176</td><td>Intronic:
GB14007</td><td>+</td><td>0</td><td>1</td><td>0</td><td>0</td><td>Class 3</td></tr>
<tr><td colspan="12">GCGATAAGGTTAGGGGTTTCTATCGGCCTCCAGCAGCAACGATGGTTGTAGGCCGGCGGAAACTACTTGCTCTTGTCG
GCAACAAGGTTAGGGGTTTCTATCGGCCTCCAGCAGCTAAGACAGCTGTAGGCCGGCGGAAACTACTTGCTCTTGTTG
** * ******************************** * ** * ****************************** *
.(((((((((.((..((((((..((((((...(((((.......)))))))))))..))))))..)).)).))))))))</td></tr>
<tr><td>98</td><td>9.636020619</td><td>GroupUn.2585</td><td>1</td><td>97</td><td>Intergenic</td><td>+</td><td>0</td><td>0</td><td>1</td><td>0</td><td>Class 6</td></tr>
<tr><td colspan="12">CCTGTCTTGTTCATAAGTACTAGTGCCGCAGGAGTGACTAGGTTGGGTTAGAAATTACCATATCTCCTGCTGCTCAAGTGCTTATCAATGGGTCCGG
CCCGTCCTGTTCCTAAGTACTAGTGCCGCAGGAGTGACTAGATTGGGCTAGAAATTACCGCGTCTCCTGCTGCTCAAGTGCTTATCAATGGTTCGGG
** *** ***** **************************** ***** *********** ****************************** ** **
((((.((..((.(((((((((.(.((.(((((((((....((((........))))....)).))))))).)).).))))))))).))..)).))))</td></tr>
<tr><td>232</td><td>7.902565603</td><td>Group15.28</td><td>185995</td><td>186088</td><td>Intronic:
GB14851-RA</td><td>-</td><td>1</td><td>1</td><td>1</td><td>1</td><td>Class 5</td></tr>
<tr><td colspan="12">CGTAGCCGATGGTATTTCACATCGTCATGGCGGGGTATTGGTAAAAGTTTTCAACTAGCAATAATCGCACCTCGGTAGAACCTCATTGGTTACG
CGTAGCCGATGGTATTTCACATCGTCTTGGCGGGGTATTGGTAAAAGTTTTCAACTAGCAATAATCGCACCTCGGTAGAACCTCATTGGTTACG
************************** **
((((((((((((...(((..((((.....((((..(((((.((.............)).))))).))))....)))).)))..))))))))))))</td></tr>
<tr><td>257</td><td>7.763298193</td><td>GroupUn.8772</td><td>1963</td><td>2041</td><td>Intergenic</td><td>+</td><td>0</td><td>0</td><td>0</td><td>0</td><td>Class 3</td></tr>
<tr><td colspan="12">CTTGCAGTTGGAAGTGAGGATCTAGGCAGTGAGCAGCA-GCGCTCTGA-TAACTTGCCAGATCTAACTCTTCCAGCTCAAG
CTTG-AGTTGGAAGTGCGGATCTAGACAGTGAGCGCCTAAGACTCTGAGGAATTTGCCAGATCTAACTCTTCCAGCTCGAG
**** *********** ******** ********   *   ******   ** ************************ **
((((.(((((((((...((((((.(((((.((((.......))))........))))))))))....)))))))))))))</td></tr>
</table>

APPENDIX C

454 matches. Low scoring candidates overlap with non putative mature regions

289	7.566941176	Group12.18	206149	206228	Intergenic	+	1	1	1	1	Class 1
AAAATGTGGAACGCTTCACGATTTTGCGTGTCATCCTTGCGCAGGGGCCATGCTAATCTTCTCTGTATCGTTCCAATTTT AAAATGTGGAACGCTTCACGATTTTGCGTGTCATCCTTGCGCAGGGGCCATGCTAATCTTCTCTGTATCGTTCCAATTTT ** ......(((((((...((.((((..(((((...(((((....))))).))))).)))).....))...))))))).....											
574	6.572575481	Group16.19	234267	234350	Intergenic	+	1	1	1	1	Class 6
GCAATGGGAGAGCCTCACCCTGGAGGAACCGCCTTGATGATCACGGTAACCTCTACGCCAGGTAAGTATGCTTTTATCCGGTGC GCAATGGAAGAGCCTCGCCCTGGAGGAACCGCCTTGTTGATCACGGTAGCCTCTGCGCCAGGTAAGTATGCTTTCAGCCGTTCC ******* ******** ******************* *********** ***** ******************* * *** * * (((((((.(((((..(..(((((.(((...(((.((.....)).)))....)))...)))))...)...)))))...)))))))											
713	6.030044893	Group3.1	131	211	Intergenic	+	1	1	1	1	Class 6
CCAAGTACTAACCGTGCCCGACGTAGCTTGACTTTGGTGATCGAACGAGAACCAGTAGTTCCTACGTGGTATGGTCGTTGG CCAAGTACTAACCGCGCCTAACGCTGCTTAACTTCGGTGATCGGACGAGAACCGGTGCATTCAGCGTAGTATGGCTGTTGG ************** *** *** **** **** ******** ********* ** ** *** ****** ***** ((((....((((((((((..((((((.......(((((..((....))..)))))......))))))))))))))))))))											

APPENDIX D

Top 200 scoring sequences – honey bee – Amel4.0.

RANK	Class	TotalScore	Alifold deltaG	RANDfold Apis	RANDfold Nasonia	Aln length	Identity	Gaps	Aln _score	microRNA	Group	Coord 1	Coord 2	Strand	Overlap
1	Class 6	16.51552632	-67.25	0.000999001	0.000999001	95	93.70%	0.00%	421	Putative	Group8.5	10732	10826	-	Overlap splice: GB30239-RB
2	Class 3	15.16253425	-47.25	0.000999001	0.000999001	73	86.30%	1.40%	269	Putative	Group7.6	284963	285034	+	Intergenic
3	Class 5	15.02473684	-39.05	0.000999001	0.000999001	76	98.70%	0.00%	371	Putative	Group2.40	27078	27153	-	Overlap splice: GB14719-RA
4	Class 5	14.86132184	-43.15	0.000999001	0.000999001	87	98.90%	0.00%	426	ame-mir-929	Group15.29	779992	780078	-	Intronic: GB12095-RB
5	Class 2	14.7847619	-41.95	0.000999001	0.000999001	84	97.60%	0.00%	402	ame-mir-317	Group5.13	300667	300750	+	Intronic: GB10191-RA
6	Class 4	14.36452381	-39.32	0.000999001	0.000999001	84	96.40%	0.00%	393	ame-mir-210	Group2.38	687279	687362	+	Intergenic
7	Class 2	14.355	-38.34	0.000999001	0.000999001	84	97.60%	0.00%	402	ame-mir-276	Group7.35	693906	693989	+	Intergenic
8	Class 3	14.31452381	-38.9	0.000999001	0.000999001	84	96.40%	0.00%	393	ame-mir-137	Group14.17	281279	281362	+	Intergenic
9	Class 2	14.24084906	-23.95	0.000999001	0.000999001	53	98.10%	1.90%	250	Putative	Group10.36	100810	100861	+	Intergenic
10	Class 3	14.10626582	-35.1	0.000999001	0.000999001	79	96.20%	0.00%	368	ame-mir-125	Group8.27	75742	75820	+	Intergenic
11	Class 3	14.09323529	-40.15	0.000999001	0.000999001	85	92.90%	0.00%	371	ame-mir-7	Group9.22	487151	487235	+	Intergenic
12	Class 4	14.08174419	-36.86	0.000999001	0.000999001	86	97.70%	0.00%	412	Putative	Group13.3	22647	22732	+	Overlap splice: GB11776-RA
13	Class 4	14.05045455	-26.1	0.000999001	0.000999001	55	92.70%	3.60%	236.5	Putative	Group2.35	149863	149915	-	Intronic: GB11268-RA
14	Class 1	13.88412088	-35.3	0.000999001	0.000999001	91	100.00%	0.00%	455	ame-mir-iab-4	Group16.9	878072	878162	+	Intergenic
15	Class 5	13.82522472	-34.9	0.000999001	0.000999001	89	98.90%	0.00%	436	ame-mir-133	Group16.18	199372	199460	+	Intergenic
16	Class 2	13.67642857	-28.4	0.000999001	0.000999001	70	95.70%	0.00%	323	ame-bantam	Group14.24	1302342	1302411	+	Intergenic
17	Class 4	13.66170103	-43.22	0.000999001	0.000999001	97	89.70%	9.30%	407.5	Putative	Group12.1	545771	545858	+	Intronic: GB17926-RA
18	Class 3	13.61611111	-36.35	0.000999001	0.000999001	81	90.10%	3.70%	334	Putative	Group4.13	369199	369276	+	Intergenic
19	Class 3	13.57791667	-40.6	0.000999001	0.000999001	96	92.70%	0.00%	417	ame-mir-9a	Group1.64	474355	474450	+	Intergenic
20	Class 4	13.55584746	-29.05	0.000999001	0.000999001	59	84.70%	0.00%	214	Putative	Group3.33	39123	39181	+	Intergenic
21	Class 3	13.53625	-26.8	0.000999001	0.000999001	64	93.80%	1.60%	278	ame-mir-2-3	Group1.1	136757	136819	-	Intronic: GB15727-RA
22	Class 3	13.53487013	-33.63	0.000999001	0.000999001	77	90.90%	6.50%	320.5	Putative	Group1.1	136120	136195	-	Intronic: GB15727-RA
23	Class 6	13.53330189	-23.2	0.000999001	0.000999001	53	90.60%	0.00%	220	Putative	GroupUn.189	32050	32102	+	Overlap exon: GB16323-RA
24	Class 3	13.505	-34.9	0.000999001	0.000999001	82	91.50%	3.70%	348	ame-mir-13a	Group1.1	136463	136541	-	Intronic: GB15727-RA
25	Class 6	13.46477011	-45.55	0.000999001	0.000999001	87	80.50%	5.70%	280.5	ame-mir-2-2	Group1.1	134955	135038	-	Intronic: GB15727-RA
26	Class 3	13.42264706	-39.25	0.000999001	0.000999001	85	88.20%	2.40%	323	ame-mir-281	Group1.65	143838	143921	+	Intergenic

27	Class 6	13.39825843	-36.5	0.000999001	0.000999001	89	92.10%	0.00%	382	ame-mir-305	Group4.22	624610	624698	+	Intergenic
28	Class 3	13.32	-37.92	0.000999001	0.000999001	80	85.00%	1.20%	286	ame-mir-71	Group1.1	137134	137212	-	Intronic: GB15727-RA
29	Class 6	13.23993976	-35.25	0.000999001	0.000999001	83	90.40%	2.40%	331	ame-mir-315	Group2.33	181912	181993	+	Intergenic
30	Class 3	13.21558824	-41.64	0.000999001	0.000999001	85	82.40%	3.50%	281.5	ame-mir-12	Group2.17	22430	22512	-	Intronic: GB16497-RA
31	Class 6	13.195	-37.47	0.000999001	0.000999001	80	83.80%	2.50%	280.5	Putative	Group9.16	38513	38590	+	Intergenic
32	Class 6	13.14893939	-27.05	0.000999001	0.000999001	66	89.40%	0.00%	267	Putative	Group1.40	45413	45478	-	Overlap exon: GB12475-RA
33	Class 6	13.10794118	-31.6	0.000999001	0.000999001	68	83.80%	1.50%	235	Putative	Group7.37	249435	249501	+	Overlap exon: GB18988-RA
34	Class 6	13.10424242	-48.71	0.000999001	0.000999001	132	93.90%	0.80%	582	Putative	Group16.11	38051	38181	+	Intergenic
35	Class 3	13.08454545	-34	0.000999001	0.000999001	88	92.00%	1.10%	371	ame-mir-8	Group11.33	642192	642278	-	Intronic: GB10038-RA
36	Class 3	13.05142857	-35.44	0.000999001	0.000999001	84	88.10%	3.60%	321.5	ame-mir-31a	Group8.21	34659	34741	+	Intergenic
37	Class 3	13.02366667	-34.49	0.000999001	0.000999001	75	82.70%	6.70%	256.5	ame-mir-275	Group4.22	624394	624468	+	Intergenic
38	Class 2	13.01369565	-27.36	0.000999001	0.000999001	69	91.30%	2.90%	279	ame-mir-282	Group14.23	30961	31028	+	Intergenic
39	Class 4	12.99280488	-32.55	0.000999001	0.000999001	82	90.20%	3.70%	329.5	Putative	Group11.31	1263918	1263998	+	Intergenic
40	Class 3	12.96517699	-41.75	0.000999001	0.000999001	113	92.90%	4.40%	482	ame-mir-10	Group16.10	82747	82857	+	Intergenic
41	Class 3	12.88308219	-29.36	0.000999001	0.000999001	73	87.70%	2.70%	281.5	ame-mir-283	Group2.17	23710	23780	-	Intronic: GB16497-RA
42	Class 3	12.84833333	-38.89	0.000999001	0.000999001	90	83.30%	6.70%	317	Putative	Group2.34	226591	226677	+	Intergenic
43	Class 6	12.68357143	-27.85	0.000999001	0.000999001	84	94.00%	3.60%	366.5	Putative	Group10.24	4082	4162	+	Intergenic
44	Class 3	12.68289474	-40.34	0.000999001	0.000999001	95	80.00%	9.50%	326	ame-mir-932	Group1.75	243971	244065	-	Intronic: GB10066-RA
45	Class 3	12.65442529	-33.5	0.000999001	0.000999001	87	88.50%	4.60%	330.5	ame-mir-184	Group13.5	467380	467464	+	Intergenic
46	Class 3	12.63357143	-32.34	0.000999001	0.000999001	77	83.10%	9.10%	264	Putative	GroupUn.139	31096	31169	+	Intergenic
47	Class 6	12.58335821	-51.65	0.000999001	0.000999001	134	85.80%	0.00%	499	Putative	Group7.31	108471	108604	-	Overlap exon: GB14642-RA
48	Class 3	12.57744898	-40.71	0.000999001	0.000999001	98	80.60%	10.20%	335	Putative	Group2.17	22929	23021	-	Intronic: GB16497-RA
49	Class 6	12.57230769	-23.25	0.000999001	0.000999001	52	78.80%	0.00%	161	Putative	Group13.17	186815	186866	+	Intergenic
50	Class 3	12.47067164	-25.22	0.000999001	0.000999001	67	86.60%	1.50%	248	ame-mir-927	Group8.29	233074	233139	+	Intronic: GB16304-RA
51	Class 3	12.44766667	-31.92	0.000999001	0.000999001	75	81.30%	10.70%	239	Putative	Group9.15	253723	253795	+	Intergenic
52	Class 6	12.28570175	-24.7	0.000999001	0.000999001	57	77.20%	0.00%	168	ame-mir-9b	Group15.33	20342	20398	+	Intronic: GB16086-RA
53	Class 6	12.28570175	-24.7	0.000999001	0.000999001	57	77.20%	0.00%	168	ame-mir-79	Group15.33	20342	20398	+	Intronic: GB16086-RA
54	Class 6	12.15743902	-31.45	0.000999001	0.000999001	82	82.90%	2.40%	272	Putative	Group15.13	55468	55548	-	Intronic: GB16072-RA
55	Class 4	12.07155914	-36.2	0.003996004	0.001998002	62	95.20%	0.00%	283	Putative	Group3.31	172863	172924	+	Intergenic
56	Class 5	12.0544086	-26.9	0.001998002	0.000999001	62	93.50%	3.20%	271.5	Putative	Group11.33	352527	352588	-	Overlap splice: GB30128-RA
57	Class 3	12.05361111	-26.15	0.000999001	0.000999001	72	83.30%	1.40%	246	ame-mir-33	GroupUn.1070	1108	1178	-	Intronic: GB30339-RA
58	Class 3	11.99918605	-28.75	0.000999001	0.000999001	86	84.90%	3.50%	314	ame-mir-29b	Group8.25	59469	59551	+	Intergenic

59	Class 3	11.97201031	-35.48	0.000999001	0.000999001	97	79.40%	10.30%	321	ame-mir-219	Group11.1	42866	42957	+	Intergenic
60	Class 3	11.85275862	-36.84	0.000999001	0.001998002	87	93.10%	3.40%	372.5	ame-mir-277	Group5.13	322869	322952	+	Intronic: GB10191-RA
61	Class 6	11.76792135	-21.99	0.000999001	0.000999001	89	92.10%	0.00%	382	Putative	Group2.38	681487	681575	+	Intergenic
62	Class 3	11.73688406	-23.7	0.000999001	0.000999001	69	81.20%	11.60%	227.5	Putative	Group13.6	502702	502765	-	Intronic: GB15055-RA
63	Class 6	11.71809524	-30.54	0.000999001	0.000999001	84	79.80%	3.60%	258.5	Putative	GroupUn.1659	1290	1372	+	Intergenic
64	Class 6	11.70352941	-22.95	0.000999001	0.000999001	68	82.40%	1.50%	226	Putative	Group8.24	608177	608243	+	Intergenic
65	Class 4	11.67491379	-27.15	0.001998002	0.001998002	58	94.80%	3.40%	260.5	Putative	Group14.23	541010	541065	+	Intergenic
66	Class 4	11.65909657	-43.65	0.001998002	0.000999001	107	91.60%	0.00%	454	Putative	Group16.20	328261	328367	-	Overlap exon: GB30062-RA
67	Class 3	11.63202703	-26.44	0.000999001	0.000999001	74	77.00%	10.80%	226	Putative	Group15.13	54961	55029	-	Intronic: GB16072-RA
68	Class 4	11.5352968	-30.55	0.000999001	0.001998002	73	89.00%	0.00%	293	Putative	Group1.72	355261	355333	+	Overlap exon: GB12139-RA
69	Class 6	11.49905941	-37.59	0.000999001	0.000999001	101	77.20%	3.00%	280	Putative	Group10.18	8982	9080	-	Intronic: GB13125-RA
70	Class 5	11.44351351	-38	0.001998002	0.004995005	74	98.60%	0.00%	361	Putative	Group12.11	38382	38455	+	Overlap exon: GB10111-RA
71	Class 5	11.4025	-26.25	0.000999001	0.002997003	65	98.50%	0.00%	316	Putative	Group2.35	113160	113224	+	Intergenic
72	Class 5	11.34002817	-31.7	0.000999001	0.003996004	71	98.60%	0.00%	346	Putative	Group9.16	90241	90311	+	Overlap exon: GB12929-RA
73	Class 3	11.24302817	-24.94	0.000999001	0.000999001	71	70.40%	22.50%	193.5	ame-mir-2-1	Group1.1	135912	135973	-	Intronic: GB15727-RA
74	Class 6	11.2087037	-51.9	0.000999001	0.000999001	162	77.80%	0.00%	486	Putative	Group8.24	315499	315651	-	Intronic: GB14709-RA
75	Class 2	11.19604839	-25.6	0.001998002	0.001998002	62	95.20%	0.00%	283	ame-mir-1	Group16.18	5866	5927	+	Intergenic
76	Class 6	11.14916667	-27.45	0.001998002	0.000999001	72	88.90%	0.00%	288	Putative	Group2.16	121189	121260	-	Overlap exon: GB15177-RA
77	Class 3	11.04582022	-42.89	0.000999001	0.003996004	89	92.10%	1.10%	376	Putative	Group4.7	1559823	1559911	-	Intronic: GB16572-RA
78	Class 6	11.03595238	-24.06	0.000999001	0.000999001	84	81.00%	9.50%	266	Putative	Group15.13	54829	54909	-	Intronic: GB16072-RA
79	Class 5	10.72940972	-30.45	0.007992008	0.000999001	64	98.40%	0.00%	311	Putative	Group9.10	515685	515748	-	Overlap splice: GB17617-RA
80	Class 6	10.70598485	-32.5	0.000999001	0.001998002	88	86.40%	3.40%	323.5	ame-mir-92a	Group15.34	28349	28435	+	Intergenic
81	Class 4	10.67653005	-28.95	0.002997003	0.002997003	61	91.80%	0.00%	260	Putative	Group1.25	234267	234327	-	Overlap exon: GB15780-RA
82	Class 5	10.63037838	-27.75	0.001998002	0.002997003	74	98.60%	0.00%	361	Putative	Group16.9	1406492	1406565	+	Intergenic
83	Class 3	10.62219298	-28.47	0.001998002	0.000999001	76	81.60%	9.20%	269	Putative	GroupUn.69	22994	23069	-	Intronic: GB12790-RA
84	Class 3	10.56929825	-37.16	0.000999001	0.001998002	95	78.90%	14.70%	315.5	ame-mir-14	Group11.14	243981	244065	+	Intergenic
85	Class 3	10.55820513	-39.4	0.005994006	0.000999001	78	89.70%	0.00%	318	Putative	Group4.16	137099	137176	+	Intronic: GB14007-RA
86	Class 1	10.37374384	-35.2	0.013986014	0.014985015	70	100.00%	0.00%	350	Putative	Group1.7	129414	129483	+	Intronic: GB13919-RA
87	Class 4	10.33261905	-30.15	0.004995005	0.000999001	70	92.90%	0.00%	305	Putative	Group7.33	466907	466976	+	Overlap exon: GB30231-RA
88	Class 6	10.28953704	-21.35	0.001998002	0.001998002	54	87.00%	0.00%	207	Putative	Group1.54	681258	681311	-	Overlap exon: GB11223-RA
89	Class 5	10.2425	-27.35	0.002997003	0.008991009	60	98.30%	0.00%	291	Putative	Group14.23	657560	657619	+	Intergenic
90	Class 4	10.20083333	-39.1	0.014985015	0.028971029	75	97.30%	0.00%	357	Putative	Group6.29	204528	204602	+	Intergenic

91	Class 6	10.17189394	-27.85	0.000999001	0.001998002	88	85.20%	0.00%	323	Putative	GroupUn.549	28613	28700	+	Overlap exon: GB19591-RA
92	Class 1	10.12384615	-28.3	0.005994006	0.006993007	65	100.00%	0.00%	325	Putative	Group13.18	75309	75373	-	Overlap exon: GB15328-RA
93	Class 2	10.1072807	-35.62	0.002997003	0.002997003	95	96.80%	2.10%	445.5	Putative	Group4.23	268120	268212	+	Intergenic
94	Class 3	10.052	-28.9	0.001998002	0.002997003	80	93.80%	0.00%	355	Putative	Group13.10	877163	877242	-	Intronic: GB19979-RA
95	Class 4	10.00179012	-41.11	0.011988012	0.005994006	81	93.80%	1.20%	354	Putative	Group8.5	22573	22652	-	Intronic: GB30239-RB
96	Class 5	9.670705128	-21.75	0.010989011	0.000999001	52	96.20%	0.00%	242	Putative	Group13.4	190326	190377	+	Intergenic
97	Class 5	9.669206407	-31.95	0.028971029	0.257742258	67	98.50%	0.00%	326	Putative	Group6.15	248126	248192	+	Intergenic
98	Class 6	9.636020619	-34.55	0.003996004	0.000999001	97	89.70%	0.00%	395	Putative	GroupUn.2585	1	97	+	Intergenic
99	Class 6	9.598627451	-35.95	0.000999001	0.040959041	68	89.70%	0.00%	277	Putative	Group2.43	449675	449742	+	Overlap exon: GB10264-RA
100	Class 3	9.591538462	-22.35	0.005994006	0.000999001	65	96.90%	0.00%	307	Putative	Group5.22	397929	397993	+	Intergenic
101	Class 5	9.500181818	-36.98	0.001998002	0.002997003	110	90.90%	3.60%	455	Putative	Group11.32	309771	309876	+	Intergenic
102	Class 6	9.430571429	-37.4	0.003996004	0.000999001	91	81.30%	0.00%	302	Putative	Group3.33	199637	199727	-	Overlap splice: GB15382-RA
103	Class 3	9.406124031	-34	0.024975025	0.017982018	75	96.00%	0.00%	348	Putative	Group1.55	98080	98154	+	Intronic: GB10650-RC
104	Class 5	9.32389372	-31	0.02997003	0.014985015	69	95.70%	0.00%	318	Putative	Group3.7	171972	172040	+	Intronic: GB18730-RA
105	Class 6	9.315183291	-34.25	0.114885115	0.022977023	68	91.20%	0.00%	286	Putative	Group2.2	89797	89864	-	Intronic: GB10324-RA
106	Class 3	9.306666667	-26.51	0.001998002	0.000999001	80	71.20%	13.80%	212.5	Putative	Group15.13	55307	55381	-	Intronic: GB16072-RA
107	Class 4	9.301350575	-25	0.001998002	0.001998002	87	87.40%	8.00%	341.5	Putative	GroupUn.19	63647	63726	+	Intergenic
108	Class 3	9.286947497	-36.99	0.001998002	0.006993007	91	90.10%	0.00%	374	ame-mir-263	Group10.29	368549	368639	+	Intergenic
109	Class 6	9.170642857	-23.15	0.007992008	0.001998002	56	89.30%	0.00%	226	Putative	Group12.30	1279544	1279599	+	Overlap exon: GB10560-RA
110	Class 6	9.168365079	-31.1	0.033966034	0.055944056	70	95.70%	0.00%	323	Putative	Group8.28	239333	239402	-	Intronic: GB11761-RA
111	Class 4	9.146553846	-27.95	0.022977023	0.001998002	65	93.80%	0.00%	289	Putative	Group15.20	175075	175139	-	Overlap exon: GB11254-RA
112	Class 3	9.098270677	-29.7	0.00999001	0.008991009	63	87.30%	0.00%	243	Putative	Group4.27	305392	305454	-	Overlap exon: GB11807-RA
113	Class 6	9.086125356	-30.45	0.004995005	0.021978022	65	89.20%	0.00%	262	Putative	Group9.12	168115	168179	+	Intergenic
114	Class 6	9.079954233	-37	0.00999001	0.012987013	76	88.20%	2.60%	287	Putative	Group9.25	425313	425387	-	Overlap exon: GB14416-RA
115	Class 2	9.07803767	-54.8	0.201798202	0.191808192	133	99.20%	0.00%	656	Putative	Group13.9	541542	541674	+	Intergenic
116	Class 6	9.074307692	-30.7	0.032967033	0.021978022	65	90.80%	0.00%	271	Putative	Group7.33	520753	520817	+	Overlap exon: GB13516-RA
117	Class 4	9.051048387	-40.65	0.007992008	0.01998002	93	92.50%	0.00%	402	Putative	Group5.13	247556	247648	+	Overlap exon: GB20055-RE
118	Class 6	9.042294776	-26.7	0.002997003	0.004995005	67	86.60%	4.50%	255	Putative	Group10.32	161890	161953	-	Overlap exon: GB13430-RA
119	Class 4	9.006265893	-27.4	0.004995005	0.048951049	67	97.00%	0.00%	317	Putative	Group16.9	1021525	1021591	+	Intergenic
120	Class 3	8.957304176	-31.35	0.052947053	0.003996004	71	93.00%	0.00%	310	Putative	Group14.24	544736	544806	+	Intergenic
121	Class 2	8.946865942	-26.9	0.02997003	0.025974026	69	98.60%	0.00%	336	Putative	Group4.7	68470	68538	-	Overlap exon: GB30205-RB
122	Class 6	8.920452586	-48.12	0.001998002	0.013986014	116	90.50%	0.00%	481	Putative	Group6.29	143431	143546	-	Overlap exon: GB15419-RA

123	Class 6	8.90911211	-39.3	0.243756244	0.001998002	76	85.50%	0.00%	281	Putative	Group2.38	427015	427090	+	Overlap exon: GB15698-RA
124	Class 5	8.89865942	-27.05	0.056943057	0.034965035	69	98.60%	0.00%	336	Putative	Group1.82	250087	250155	+	Intergenic
125	Class 6	8.897898551	-28.55	0.24975025	0.002997003	60	90.00%	0.00%	246	Putative	Group13.5	172260	172319	-	Overlap exon: GB30150-RC
126	Class 6	8.889507246	-26.2	0.688311688	0.000999001	56	91.10%	0.00%	235	Putative	Group8.36	134675	134730	+	Overlap exon: GB30298-RA
127	Class 6	8.878391304	-39.35	0.01998002	0.014985015	92	92.40%	0.00%	397	Putative	Group6.40	148631	148722	+	Intergenic
128	Class 3	8.842647849	-34.7	0.005994006	0.001998002	93	87.10%	6.50%	359	Putative	Group1.18	41267	41353	-	Intronic: GB16274-RA
129	Class 6	8.81641327	-26.7	0.175824176	0.162837163	61	93.40%	0.00%	269	Putative	Group10.30	162642	162702	-	Overlap exon: GB10640-RA
130	Class 4	8.800432099	-24.47	0.002997003	0.002997003	81	90.10%	0.00%	333	Putative	Group1.55	212100	212180	-	Overlap exon: GB12769-RA
131	Class 4	8.799565217	-28.6	0.002997003	0.003996004	92	92.40%	4.30%	392	Putative	Group11.29	46559	46646	+	Intergenic
132	Class 4	8.79694747	-35.35	0.106893107	0.017982018	83	94.00%	0.00%	370	Putative	Group13.5	160657	160739	-	Overlap exon: GB30150-RB
133	Class 2	8.766766667	-26.2	0.020979021	0.078921079	69	98.60%	0.00%	336	Putative	Group1.82	688445	688513	-	Intronic: GB14440-RA
134	Class 1	8.764097938	-29.7	0.102897103	0.090909091	80	100.00%	0.00%	400	Putative	Group11.32	246408	246487	+	Overlap splice: GB15792-RA
135	Class 4	8.745438437	-26.1	0.011988012	0.018981019	71	97.20%	0.00%	337	Putative	Group11.11	166869	166939	-	Overlap splice: GB10578-RA
136	Class 6	8.743333333	-25.9	0.000999001	0.00999001	66	87.90%	0.00%	258	Putative	Group8.18	144921	144986	-	Overlap exon: GB18762-RA
137	Class 4	8.740862069	-28.35	0.012987013	0.000999001	58	79.30%	0.00%	182	Putative	GroupUn.3059	5342	5392	+	Intergenic
138	Class 3	8.732340426	-29.23	0.007992008	0.002997003	94	96.80%	0.00%	443	ame-mir-100	Group8.27	74565	74658	+	Intergenic
139	Class 6	8.725623729	-26.2	0.046953047	0.002997003	59	89.80%	0.00%	241	Putative	Group12.30	913153	913211	+	Overlap exon: GB19579-RA
140	Class 5	8.71619403	-22.75	0.013986014	0.007992008	67	98.50%	0.00%	326	Putative	Group16.9	1281056	1281122	+	Intergenic
141	Class 4	8.704808207	-26.5	0.016983017	0.001998002	59	88.10%	6.80%	217.5	Putative	Group2.36	260320	260375	+	Intergenic
142	Class 6	8.654905149	-31.65	0.002997003	0.005994006	82	85.40%	0.00%	302	Putative	Group8.3	9235	9316	-	Overlap splice: GB18118-RA
143	Class 3	8.624754642	-32.1	0.074925075	0.040959041	78	93.60%	0.00%	345	Putative	Group10.30	165629	165706	-	Overlap exon: GB10640-RA
144	Class 4	8.618225806	-40.2	0.004995005	0.008991009	124	96.80%	0.80%	578	Putative	GroupUn.2532	4613	4735	+	Intronic: GB15283-RA
145	Class 6	8.606092153	-32.05	0.24975025	0.086913087	72	90.30%	0.00%	297	Putative	Group11.9	113656	113727	+	Overlap exon: GB18111-RA
146	Class 6	8.605126103	-55.7	0.015984016	0.105894106	130	91.50%	0.00%	551	Putative	Group1.40	330985	331114	+	Intronic: GB10180-RA
147	Class 6	8.57803114	-30.15	0.176823177	0.003996004	66	89.40%	1.50%	261	Putative	Group2.24	337294	337358	+	Intronic: GB16625-RA
148	Class 4	8.564714729	-23.8	0.021978022	0.048951049	59	93.20%	0.00%	259	Putative	Group6.55	61363	61421	-	Overlap splice: GB17945-RA
149	Class 4	8.55953271	-32.55	0.000999001	0.00999001	107	96.30%	0.90%	493	Putative	GroupUn.7907	820	925	+	Intergenic
150	Class 4	8.551116139	-32.8	0.316683317	0.03996004	65	83.10%	0.00%	226	Putative	Group5.12	1560842	1560906	+	Overlap exon: GB11380-RA
151	Class 6	8.550350877	-24.05	0.001998002	0.008991009	57	82.50%	0.00%	195	Putative	Group14.24	965293	965349	-	Overlap exon: GB14793-RA
152	Class 1	8.540598592	-22.6	0.013986014	0.013986014	71	100.00%	0.00%	355	Putative	Group5.7	154830	154900	+	Intergenic
153	Class 6	8.526528266	-30.25	0.002997003	0.097902098	62	83.90%	0.00%	220	Putative	Group4.15	28558	28619	-	Overlap exon: GB15295-RA
154	Class 6	8.505316108	-30.4	0.001998002	0.104895105	68	88.20%	0.00%	268	Putative	Group4.23	24615	24682	+	Overlap exon: GB30210-RA

155	Class 6	8.5025	-25.2	0.000999001	0.002997003	63	63.50%	25.40%	126	Putative	GroupUn.1226	4261	4315	+	Overlap exon: GB17195-RA
156	Class 4	8.475010935	-25.85	0.05994006	0.001998002	59	88.10%	0.00%	232	Putative	Group4.27	305518	305576	-	Overlap splice: GB11807-RA
157	Class 4	8.471461353	-32.5	0.057942058	0.010989011	72	86.10%	5.60%	274.5	Putative	Group12.23	46078	46149	+	Overlap exon: GB11571-RA
158	Class 6	8.434423077	-23.6	0.015984016	0.011988012	65	93.80%	0.00%	289	Putative	Group9.17	3477	3541	+	Overlap exon: GB11915-RA
159	Class 5	8.419589314	-24.4	0.072927073	0.002997003	66	95.50%	0.00%	303	Putative	Group3.39	48046	48111	-	Overlap splice: GB15719-RA
160	Class 6	8.410589744	-23.6	0.022977023	0.006993007	65	93.80%	0.00%	289	Putative	GroupUn.452	59336	59400	+	Overlap exon: GB12507-RA
161	Class 6	8.406430678	-24.85	0.675324675	0.001998002	60	91.70%	0.00%	255	Putative	Group4.21	53347	53406	-	Overlap exon: GB13109-RA
162	Class 4	8.398277844	-23.7	0.03996004	0.056943057	61	93.40%	0.00%	269	Putative	Group8.14	141601	141661	-	Overlap splice: GB11896-RA
163	Class 5	8.393768187	-35.1	0.573426573	0.737262737	101	99.00%	0.00%	496	Putative	Group1.7	129484	129584	+	Intronic: GB13919-RA
164	Class 4	8.373466855	-42.6	0.285714286	0.684315684	95	88.40%	6.30%	368.5	Putative	Group3.19	164468	164561	-	Overlap splice: GB12179-RA
165	Class 5	8.362769282	-32.3	0.245754246	0.201798202	94	98.90%	0.00%	461	Putative	GroupUn.1348	3702	3795	+	Intergenic
166	Class 6	8.301280488	-30	0.002997003	0.000999001	82	68.30%	9.80%	175.5	Putative	Group8.46	214707	214784	+	Intergenic
167	Class 6	8.293932109	-40.65	0.717282717	0.011988012	91	86.80%	0.00%	347	Putative	Group5.22	594213	594303	-	Overlap exon: GB18706-RA
168	Class 6	8.293442116	-28.85	0.00999001	0.030969031	71	88.70%	0.00%	283	Putative	Group10.36	1092328	1092398	-	Overlap exon: GB16085-RA
169	Class 4	8.290107513	-25.3	0.671328671	0.861138861	67	95.50%	1.50%	302	Putative	Group11.21	368378	368444	+	Intergenic
170	Class 3	8.282442529	-31.2	0.001998002	0.00999001	87	87.40%	0.00%	336	Putative	Group4.18	272373	272459	+	Overlap exon: GB10337-RA
171	Class 4	8.281099081	-25.25	0.260739261	0.509490509	69	95.70%	0.00%	318	Putative	GroupUn.646	17761	17829	+	Overlap splice: GB30491-RA
172	Class 6	8.28080303	-24	0.016983017	0.002997003	66	90.90%	3.00%	273.5	Putative	Group2.38	785134	785197	-	Intronic: GB18541-RA
173	Class 4	8.279294069	-21.7	0.015984016	0.021978022	63	95.20%	0.00%	288	Putative	Group13.4	192878	192940	+	Intergenic
174	Class 4	8.277558287	-26.75	0.006993007	0.581418581	71	94.40%	0.00%	319	Putative	Group2.38	537558	537628	+	Intergenic
175	Class 4	8.276908602	-25.95	0.177822178	0.193806194	66	92.40%	0.00%	285	Putative	Group14.21	224590	224655	+	Overlap exon: GB15446-RA
176	Class 4	8.264205247	-22.8	0.048951049	0.031968032	64	95.30%	0.00%	293	Putative	Group4.15	369252	369315	+	Overlap exon: GB11420-RA
177	Class 6	8.250491551	-29.2	0.017982018	0.074925075	70	88.60%	0.00%	278	Putative	Group5.31	100836	100905	-	Overlap exon: GB12614-RA
178	Class 4	8.245647356	-28.31	0.184815185	0.344655345	71	91.50%	0.00%	301	Putative	Group13.11	97800	97870	+	Intergenic
179	Class 6	8.242761645	-22.9	0.001998002	0.055944056	57	89.50%	0.00%	231	Putative	Group2.23	107483	107539	+	Intergenic
180	Class 5	8.24243956	-30.55	0.01998002	0.034965035	91	96.70%	0.00%	428	Putative	Group16.9	1439592	1439682	+	Intergenic
181	Class 6	8.240717239	-24.7	0.080919081	0.000999001	59	88.10%	0.00%	232	Putative	Group5.11	474899	474957	-	Overlap exon: GB14300-RA
182	Class 6	8.224299517	-31.45	0.066933067	0.031968032	69	84.10%	0.00%	246	Putative	Group1.62	404712	404780	-	Overlap exon: GB17390-RA
183	Class 4	8.213488045	-21.1	0.005994006	0.002997003	79	93.70%	0.00%	350	Putative	Group16.19	740682	740760	+	Intergenic
184	Class 4	8.211678744	-26.8	0.100899101	0.312687313	72	94.40%	2.80%	321.5	Putative	Group10.33	125284	125353	-	Intronic: GB18613-RA
185	Class 4	8.209635712	-33.15	0.006993007	0.18981019	85	91.80%	0.00%	362	Putative	Group7.11	49452	49536	-	Overlap exon: GB17867-RA
186	Class 6	8.208366228	-23.5	0.301698302	0.001998002	57	89.50%	0.00%	231	Putative	Group1.1	659855	659911	-	Overlap exon: GB19470-RA

187	Class 6	8.170507991	-30.3	0.022977023	0.000999001	73	86.30%	2.70%	263	Putative	Group1.17	120362	120433	+	Intergenic
188	Class 6	8.169074447	-26.45	0.012987013	0.035964036	71	91.50%	0.00%	301	Putative	GroupUn.281	28777	28847	+	Intronic: GB17956-RA
189	Class 4	8.154231037	-40.1	0.145854146	0.332667333	105	92.40%	0.00%	453	Putative	Group6.28	183735	183839	-	Overlap exon: GB13823-RA
190	Class 4	8.148972146	-31.99	0.12987013	0.021978022	77	88.30%	6.50%	302.5	Putative	Group14.24	362350	362421	+	Intergenic
191	Class 4	8.126242938	-21.15	0.015984016	0.04995005	59	93.20%	0.00%	259	Putative	Group8.29	278765	278823	+	Overlap splice: GB16304-RA
192	Class 6	8.117377173	-38.32	0.033966034	0.262737263	98	90.80%	0.00%	409	Putative	Group8.11	78656	78753	+	Overlap splice: GB15150-RA
193	Class 6	8.116469528	-37	0.007992008	0.050949051	94	90.40%	2.10%	377	Putative	Group5.33	445585	445677	-	Overlap exon: GB18386-RA
194	Class 4	8.114517635	-37.8	0.064935065	0.416583417	96	90.60%	0.00%	399	Putative	Group14.21	224492	224587	+	Overlap exon: GB15446-RA
195	Class 3	8.110333333	-17.65	0.000999001	0.003996004	60	76.70%	16.70%	190	ame-mir-375	Group3.18	653811	653864	+	Intergenic
196	Class 6	8.107739446	-30.25	0.281718282	0.000999001	76	90.80%	1.30%	311	Putative	Group4.15	144372	144446	+	Intergenic
197	Class 4	8.105166898	-30.56	0.093906094	0.197802198	79	91.10%	0.00%	332	Putative	GroupUn.9	90106	90184	-	Overlap exon: GB15740-RA
198	Class 4	8.101449222	-30	0.036963037	0.000999001	71	85.90%	4.20%	256.5	Putative	Group13.11	139026	139093	+	Intergenic
199	Class 4	8.099351266	-21.8	0.004995005	0.002997003	79	89.90%	0.00%	323	Putative	Group16.20	2935	3013	+	Intronic: GB30066-RA
200	Class 6	8.094166667	-23.4	0.041958042	0.041958042	60	90.00%	8.30%	244.5	Putative	Group14.24	1273957	1274014	+	Intronic: GB12714-RA

APPENDIX E

Major computational steps of the methods used for this work.

Chapter II - MicroRNA homologs, Evolution

These are the major steps of the methodology followed in order to identify microRNA homologs in different genome assemblies. The main part of the method is contained in the second script: miRNA_homolog_finder_V7.1.pl

COMPUTING HOMOLOGS:

1. Generating microRNA precursors in a format suited for the pipeline.
The pipeline requires microRNA precursor sequences in FASTA format and with the mature sequence indicated in lower case. MicroRNA databases usually don't provide sequences with this peculiarity, so it is neccesary to generate one with the help of a perl script.

Download stem loop precursors and mature sequences from miRBase. Combine the two files into a single file by running the following perl script:

```
Perl miRNAs_in_proper_format.pl input1 input2 > output
```

Where:
Input1: mature_microRNAs_10.0.fa
Input2: stem-loop_precursor_microRNAs_10.0.fa
Output: hairpin_miRNAs.10.0.PRC.TAB

2. Blasting precursors against genome of interest.
The script miRNA_homolog_finder_V7.1.pl is a wrapper around different programs used to compute homologous microRNAs in different genomes. This script uses WUBLAST, ViennaRNA package and t-coffee. Make sure you have installed these programs before you run the script and indicate the paths to the programs in the section of the script where is indicated.

```
perl miRNA_homolog_finder_V7.1.pl --miRNA_file input1 –-genome input2 --species input3 –-outfile output.file
```

where:
Input1: hairpin_miRNAs.10.0.PRC.TAB
Input2: genome.fasta
Input3: ‘hsa’ if human, ‘ptr’ if chimp, ‘ame’ if honey bee, etc.
You can check abbreviations for different species in the stem-loop file.es

This script will generate two files:
file 1: output.file
file 2: output.file.representative

The second file is the result of comparing genomic positions between the different results of the blast search. Removing the ones that overlap in the same genomic position and selecting the one that represents the best homolog of the sequence query. It also removes redundancy of sequences that are repeated in the database i.e.: same miRNA from two different species. This scripts also runs RANDfold on the subject sequence.

COMPUTING CLUSTERS

In order to determine microRNA family affiliation the resulting miRNA candidates from the species evaluated need to be processed in the following way:

Resulting files from the species evaluated need to be concatenated and redirected to a new file. Then the following programs can be run in order to determine microRNA clusters.

1. PRSS

This step establish the statistical significance of an alignment score between two sequences. This script runs an all vs all PRSS evaluation among the sequences present in the file.

```
perl prss_miRNA_families.pl -miRNA_TAB_file input1 –unique_ID input2 –
step_loop_precursor input3 –outfile out.file
```

Where:
Input1: concatenated files in TAB format. i.e.: file_with_miRNAs.TAB
Input2: Column representing the UNIQUE identifier of the microRNA.
Input3: Column containing the stem-loop precursor.

This step produces a file called out.file in TAB format with PRSS scores between sequences. When the sequences are too divergent, no score is calculated and a dummy p-value of 100 is reported.

2. Create adjacency list

```
perl prss2adjLists.pl out.file 1 >adjacency.list
```

3. Create clusters

```
perl adja2cluster.pl adjacency.list >clusters.file
```

4. Create clusters of FASTA files

```
perl creating_fasta_from_clusters_and_fasta.pl -clusters_file input1
clusters.file -fasta_file file_with_miRNAs.TAB -outfile
sequence_clusters.fas
```

Where:
Input1: clusters.file
Input2: file_with_miRNAs.TAB

This script produces and output with the actual clusters in FASTA format.

Chapter III - Computing genome intersections and microRNAs

These are the major steps of the methodology followed in order to identify ultraconserved elements and microRNAs by comparative genomics.

Ultraconservation is calculated with the first script: Binary_intersection_WU-BLAST_V2.pl. Subsequent scripts perform different analysis on the data generated by the first script.

A) Compute genome intersection to identify ultraconserved elements

This script is a wrapper around WUBLAST with modified parameters to compute ultraconserved elements. You will notice that is neccesary to redirect standard error to an empty file given that WUBLAST complains quite a bit when asked to compute short ultraconserved sequences. BLASTN parameters are harcoded within the script and are optimized for ultraconserved sequences. In order to change parameters it is necessary to change one line the script, the line where blastn is called.

```
perl Binary_intersection_WU-BLAST_V2.pl -query input1 -database
genome2.fasta -min_word_size input3 -filter seg &>/dev/null &
```

Input1: genome1.fasta
Input2: genome2.fasta
Input3: word size (minimum length of UCE) (i.e.: 20)
Input4: seg, dust, none.

Input1 and input2 must be in FASTA format.
min_word_size refers to the minimun length of the ultraconserved elements to be found between two genomes

Three different files will be generated out of this step:
File1: genome1.fasta_Vs_genome2.fasta.W20
-> If you used a different word size your W will be different. This file is TAB delimited and contains matches between the two genomes.

File2: genome1.fasta_Vs_genome2.fasta.W20.fas
-> This file contains the same information as the previous one but is in FASTA format. Sequences reported in this file are from genome1

File3: genome1.fasta_Vs_genome2.fasta.W20.extended_regions
-> Every hit between the two genomes gets extended upstream and downstream by 75 nt. Sequences are reported in TAB format.

Actual files:
File1: Amel_4.0_scaffolds.masked_Vs_Nvit_1.0.linear.fa.masked.W20
File2: Amel_4.0_scaffolds.masked_Vs_Nvit_1.0.linear.fa.masked.W20.fas
File3: Amel_4.0_scaffolds.masked_Vs_Nvit_1.0.linear.fa.masked.W20.extended_regions

B) Removing substrings - optional step

The file generated in (1) is redundant, and having so much redundancy is problematic in further steps. This step is recommended in order to remove redundancy. Any sequence that is a perfect substring of another sequence of the same length or longer is removed. This is a wrapper around PATDB, a program in the WUBLAST suite that does the trick.

```
perl redundancy reductor.pl input
```

Where:
Input: File in FASTA format to be slim down by removing perfect substrings. (i.e.: genome1.fasta_Vs_genome2.fasta.W20.fas).

This step will generate a file called genome1.fasta_Vs_genome2.fasta.W20.fas.nr

Actual file:
Amel_4.0_scaffolds.masked_Vs_Nvit_1.0.linear.fa.masked.W20.nr

C) Blasting non redundant regions (nr file) from genome1 against genome2.

```
blastn genome2.fasta genome1.fasta_Vs_genome2.fasta.fas.nr -e1e-10 -mformat=2 >output.file
```

This step minimizes the chances of comparing non-homologous regions between two genomes. The evalue was determined empirically to be good for a bee-wasp comparison. Other genomes may require different evalue.

This step generates a file called output.file

Actual file:
1.apis_extended_regions.fas.nr_Vs_Nasonia.BLAST_1e-10.TAB

(This file was further modified so that the TAB format produced by BLAST includes the actual sequence)

D) Removing repeats

Repeats were removed by running CENSOR and by removing sequences for which it was not possible to calculate KARLIN-ALTSCHUL STATISTICS. CENSOR can

```
censor input_file_name -bprm '-filter=none' -s -nofilter -mode {sens}
-lib hum -lib mam -lib vrt -lib rod -lib ang -lib ath -lib cbr -lib
cel -lib chl -lib cin -lib dia -lib dro -lib fng -lib fug -lib inv
-lib ory -lib pln -lib pri -lib smp -lib spu &
```

```
perl KA_parameters_masking_TAB.pl input_file
```

Actual files:
2.1.apis_extended_regions.fas.nr_Vs_Nasonia.BLAST.NR.CENSOR_DATA
2.2.apis_extended_regions.fas.nr_Vs_Nasonia.BLAST.NR.KA_masking_DATA
2.apis_extended_regions.fas.nr_Vs_Nasonia.BLAST_1e-10.NR.TAB
3.Candidates_after_CENSOR_KA_masking.TAB

E) Since microRNAs could be generated from either strand, generate a file with both strands

File:
4.Candidates_BOTH_STRANDS.TAB

F) Generate t_coffee alignments of sequence pairs and evaluate with ALIFOLD.

ALIFOLD is a program that comes with ViennaRNA. It computes the deltaG of folding of a pairwise or multiple sequence alignment.

```
perl pipeline.pl 4.Candidates_BOTH_STRANDS.TAB
```

Actual files:

```
4.Candidates_BOTH_STRANDS.T_COFFEE.ALIFOLD.RAW
4.Candidates_BOTH_STRANDS.T_COFFEE.ALIFOLD.RAW.sub_final_candidates
```

G) Local alignments of sub_final_candidates

Sequence pairs are evaluated aligned with the program WATER of the EMBOSS package. This step computes alignment score based in the score of a local alignment, this is not possible with t-coffee.

```
perl water_alignments.pl input
```

Where:

Input:

```
4.Candidates_BOTH_STRANDS.T_COFFEE.ALIFOLD.RAW.sub_final_candidates
```

Output:

```
4.Candidates_BOTH_STRANDS.T_COFFEE.ALIFOLD.RAW.sub_final_candidates.WATER
```

Actual file generated:

```
5.Candidates_BOTH_STRANDS.WATER_SCORING.RAW
```

H) Parsing file generated from step (G).

```
Perl water_accumulative_parser.pl
5.Candidates_BOTH_STRANDS.WATER_SCORING.RAW
```

Actual file:

```
5.Candidates_BOTH_STRANDS.WATER_SCORING.TAB.WATER
```

I) Evaluating candidates with RANDfold

```
perl RANDfold_TAB_both.pl
5.Candidates_BOTH_STRANDS.WATER_SCORING.TAB.WATER
```

Actual file:

```
5.Candidates_BOTH_STRANDS.WATER_SCORING.TAB.RANDfold
```

Other files related to this step:

File from this step in FASTA format:

```
5.Candidates_BOTH_STRANDS.WATER_SCORING.TAB.RANDfold.FAS
```

Blast of fasta file against RFAM and Stem Loop Precursors of MiRBase - TAB format:

```
5.Candidates_BOTH_STRANDS.WATER_SCORING.TAB.RANDfold.STLP_RFAM.TAB
```

Blast of fasta file against RFAM and Stem Loop Precursors of MirBase - FASTA format

```
5.Candidates_BOTH_STRANDS.WATER_SCORING.WITH_MIRBASE
```

J) Generate ALIFOLD of previous file:

```
perl TAB_alifold.pl
5.Candidates_BOTH_STRANDS.WATER_SCORING.WITH_MIRBASE
```

Actual file:

```
5.Candidates_BOTH_STRANDS.WATER_SCORING.WITH_MIRBASE_ALIFOLD
```

K) Import file into excel

Remove candidates with deltaG higher than -20Kcal/mol

Score candidates based on the following formula

Pair Score = (Alignment length / alignment score) + (1/((RANDfold score in Apis + RANDfold score in Nasonia)/2))/200) + (ABS (deltaG of folding for the pair)/alignment length)*10)

Chapter IV – Ultraconserved Elements

Follow steps **(A)** and **(B)** from chapter III.

C) Mapping UCEs to genome.

```
perl mapping_to_genome.pl -query UCE_file.nr -database Amel4.0 -
min_word_size 20 -filter seq -var UCE_vs_Amel_4.0
```

where:
query: non redundant file result of the previous step
database: genome of interest
var: description for GFF track file.
This step generates a GFF file called UCE_file.nr_Amel4.0.mapping.GFF

D) Intersecting UCEs with genomic features in genome

```
perl UCE_genes_intersector_honeybee.pl -GFF_query input1 –GFF_ref
input2
```

Where:
Input1: UCE_file.nr_Amel4.0.mapping.GFF
Input2: Amel_release2_combined_ncbi_apollo.gff
This step will generate 4 files:
Out_intronic: with UCEs that locate within introns
Out_exonic: with UCEs that locate within exons
Out_splice: with UCEs that overlap splice sites by at least 1 nt
Out_intergenic: UCEs that diddn't overlap with any of the previous

VERSIONS OF PROGRAMS USED FOR THIS WORK

CENSOR: censor-4.2.12.tar.gz
WUBLAST: blast2.macosx-x86.tar.Z
Vienna RNA package: ViennaRNA-1.6.1.tar.gz
t-coffee: T-COFFEE_distribution_Version_5.05.tar.gz
clustalw: Version 1.83

```
RANDfold: Ony one version – perl based
PRSS: prss34 – version 3 of FASTA package
WATER (EMBOSS SUITE): EMBOSS-6.0.1.tar.gz
```

VITA

Name:	Juan Manuel Anzola Lagos
Address:	Calle 146 No. 99a – 80 Int. 5 Santafe de Bogota, Colombia.
Email Address:	biojmal@gmail.com
Education:	Ph.D., Biology, Texas A&M University, 2008 B.Sc., Biology, Universidad Nacional de Colombia, 2000