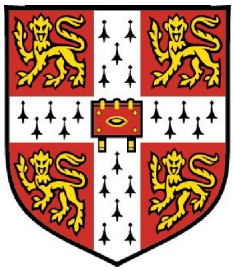
Evolutionary analysis of animal microRNAs



José Afonso Guerra Martins dos Santos Assunção

Clare Hall

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

October 2012

To my Family

In particular to my parents, who always did everything they could so I could follow my dreams, even when they didn't believe I could achieve them. Thanks to them, I've been very privileged to always have the best education available, and a supportive environment to foster my curiosity and love for science, music and art. From buying me mechanical clocks to disassemble since I was 5 to giving me my own computer at an early age while putting up with my utopic programming projects later on during high school all contributed in no small part to the development of my curiosity and skills, for which I'm very grateful.

Finally, allowing me to read Biology as an undergraduate gave me the opportunity to fulfill my ambitions and avoid the frustrations that beset my uncle in his own time.

As hard as I might try, I will never be able to put into words how grateful I am for my perfect childhood and endless support throughit the years. Thank you!

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This thesis does not exceed the specified length limit of 60.000 words as defined by the Biology Degree Committee.

This thesis has been typeset in 12pt font using LaTeX according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Evolutionary analysis of animal microRNAs

José Afonso Guerra Martins dos Santos Assunção

Summary

In recent years, microRNAs (miRNAs) have been recognised as important genetic regulators of gene expression in Animals and Plants. They can potentially target a large fraction of the cellular transcriptome, having been shown to be important for diverse biological processes such as development, cell differentiation, proliferation and metabolism. The publication of the Human genome in 2001 marked the start of a great community effort to sequence a variety of other species. These data have great potential for comparative genomics, that can lead to better biological understanding.

Some miRNA families are known to be highly conserved, across long evolutionary distances, many found in co-transcribed clusters across the genome. While these phenomena have been previously reported, a large-scale analysis of evolutionary patterns was still lacking. Furthermore, the rate at which new relevant data is being made available makes it challenging to keep up and many of the evolutionary studies performed before are now significantly out of date.

This thesis describes a number of approaches taken to analyse miRNA datasets, harnessing the full potential of currently available data for comparative genomics. These were used, not only to revisit many of the notions in the field with a larger and updated dataset, but also to develop novel strategies that enable a coherent view of miRNA evolution at different evolutionary time-scales.

A new tool, described within this thesis, was developed for large-scale, species independent miRNA mapping. An assessment of the evolution of the miRNA repertoire across species was performed, together with detailed sequence conservation analysis and miRNA family clustering. Phylogenetic profile analysis uncovered interesting co-evolution between miRNAs and protein coding genes. The genomic organisation of miRNAs and their conservation across species was also studied, providing detailed conserved synteny maps for miRNAs and proteins across more than 80 species. Finally, at the intra-specific level, I analysed the occurrence of single nucleotide polymorphisms affecting miRNA loci or their predicted target sites.

All the tools built and integrated in this research were made available to the community and designed to be easily updated, making it easier to keep up with the data that is constantly being made available. Many aspects of miRNA biology are still being uncovered, and the ability to easily put these findings into an evolutionary context will potentially be useful for the community.

Acknowledgements

Since finishing my undergraduate degree in 2007, I began the curious journey of discovery, leading to this thesis. My acceptance into the Ph.D. Program in Computational Biology (PDBC) at Instituto Gulbenkian de Ciência (IGC) was crucial in providing me with a more formal computational background and a wide overview of the field of computational biology.

I wish to thank all the members of the PDBC community, who shared this journey with me, and provided support and useful comments, especially during our yearly meetings.

I am very grateful to Jorge Carneiro. First, for taking me on as a "short term apprentice" in his group at IGC, which was a fantastic first hands-on experience as a researcher in computational biology. Secondly and more importantly, for his friendship, encouragement and mentoring.

I am indebted to my Ph.D. supervisor, Anton Enright, for taking me as a student and allowing me to develop my own ideas. His infinite patience and the guidance were pivotal to this work. Most of all, I thank him for being great fun to be around and for frequently sharing with the lab his passion for science, photography, gadgets and German beer.

The environment in which this work was carried out was extremely nice, informal and at times completely bonkers, thanks to the great colleagues within the Enright Lab. A big thank you to Stijn van Dongen, Cei Abreu-Goodger, Harpreet Kaur Saini, Sergei Manakov, Nenad Bartonicek, Mat Davis, Iain Wallace and Leonor Quintais. I learned a lot from all of you, from being a better programmer, to being able to throw a frisbee!

I am thankful for the guidance provided by Nick Goldman, Dónal O'Carroll and Duncan Odom as part of my Thesis Advisory Committee. I am sure that my work greatly benefited from the insightful comments received during the formal and informal meetings we had during the past four years.

I would also like to mention my collaborators, with whom I had many fruitful discussions during the projects we shared, Min Hu, Qasim Ayub, Daniel Jeffares, Leopold Parts, David Adams, Kerstin Howe, Derek Stemple, Vania Pobre, Cecilia Arraiano and Andreia Amaral.

I also wish to thank the Fundação para a Ciência e Tecnologia (FCT) for funding the first four years of my Ph.D. studies, and to the European Molecular Biology Laboratory for funding my university fees and the last year of my studies.

The predoc community in the Genome Campus is quite dynamic and many people contributed to my well being while here. I would like to mention Pablo Moreno, Tamara Steijger, Rita Santos, Mar Gonzales-Porta and Ana Rita Gomes for cheering me up and general support.

I have found that the world of science is not devoid of bureaucracy, a skill I still have to master. I am indebted to Manuela Cordeiro at IGC and Kathryn Hardwick and Tracey Andrew at EBI for all the assistance provided with all my paperwork, and for always making sure I did not forget anything important. I also wish to thank Peggy Nunn and Sally Wedlock for always greeting me with a smile in the mornings, and handling all my mailing needs at EBI.

I wish to thank Clare Hall, at the University of Cambridge, for fostering such a nice environment to be in. Sharing the dinner table with brilliant scholars from other fields of study was always an enlightening experience that I will no doubt miss. I also wish to thank the Clare Hall Music Committee for all I learned during our meetings, and for organising such an interesting musical program throughout the year.

Cambridge is the perfect place to improve oneself and to do things one would not expect to be able to do. Joining the Cambridge Samulnori Society greatly helped me maintain my sanity, and made sure I would take time every week to be with friends, socialising and drumming! A special thank you to Nami Morris, Justin Charity, Andrew Segar and all the other members who accompanied me in this endeavour.

I am indebted to my friends that, despite the physical distance, kept sending me postcards and words of encouragement throughout my time at EBI, Catarina Bourgard, Violeta Barradas and Sara Borges.

Finally, I wish to thank my family, and in particular my parents, José Afonso and Susana, for raising me to be what I am today, with all my characteristics and peculiarities, and for fully supporting and guiding me through the ongoing process of growing-up.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The discovery of the first microRNA (miRNA) in *Caenorhabditis elegans* opened up new horizons for biologists by showing that there could be eukaryotic regulation of gene expression by small, non-coding RNAs (Lee *et al.*, 1993; Wightman *et al.*, 1993). The field of miRNA research is rapidly expanding, and is seen by many as the "tip of the iceberg" of the non-coding regulation that is present in eukaryotic cells. Soon after the discovery of a second miRNA, in the year 2000, it became apparent that this form of regulation was not specific to nematodes and that it had implications in many important biological processes. The small length and high degree of similarity between miRNAs in different species meant that the same molecular probes could be used to detect homologs in different organisms, with good sensitivity.

As more miRNAs were described, a plethora of alternative methods started being devised, to find novel miRNA candidate loci and their targets computationally, thus avoiding the expensive and time-consuming process of using purely experimental techniques.

The origin of small interfering RNAs appears to pre-date the emergence of eukaryotes (Shabalina & Koonin, 2008). The miRNA repertoires seem to have arisen independently in animals and plants, being totally absent in fungi. Fungi possess elements of the processing machinery but not functional miRNAs (Shabalina & Koonin, 2008).

Expansions in morphological complexity in metazoans have previously been shown to correlate with expansions in miRNA repertoire (Heimberg *et al.*, 2008). This seems to indicate that miRNAs are particularly advantageous for defining cell and tissue types.

With the advent of new sequencing technologies, it is now much faster and affordable to sequence the genomes of new organisms. For the same reason, the amount of data concerning messenger RNA and miRNA transcriptomes is also rapidly expanding. Although several facets of the evolution of miRNAs and other small regulatory non-coding RNAs have been reported in the literature (Hertel *et al.*, 2006; Murphy *et al.*, 2008; Tanzer & Stadler, 2004, 2006), these studies tend to be small scale attempts to understand the evolution of a few miRNA families, within a set of closely related species.

The main motivation for this work was to make sense of the vast amount of unexplored data currently available. I perform phylogenetic analysis, comparative genomics and use post-genomic techniques to explore the evolution of animal miRNAs at different evolutionary time-scales, under a common framework.

1.1 Eukaryotic Non-coding RNAs

Upon the publication of the Human genome (Lander *et al.*, 2001; Venter *et al.*, 2001), many scientists were bemused by the apparent lack of correlation between the perceived complexity of the organism and the number of protein coding genes in its genome. While some of the transcript diversity can be explained by alternative splicing, it still would not justify the vast amounts of non protein-coding DNA observed. This led to the establishment of the concept of "genomic dark matter". Work to better understand these data was soon started, headed by the Encyclopaedia of DNA Elements (ENCODE) project, and the Functional Annotation of the Mammalian Genome (FANTOM) project (Carninci *et al.*, 2005; ENCODE Project Consortium, 2004). Both projects accumulated evidence supporting the conclusion that the majority ($> 70\%$) of the Human and Mouse genomes are actively transcribed.

It has been known for some time that not all transcripts give rise to proteins, with the latest data indicating there are more non-coding transcripts than protein-coding transcripts (ENCODE Project Consortium, 2004). Besides miRNAs, many non-coding RNA classes have been described based on classical molecular biology techniques and forward genetics studies. They are transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small-nuclear RNAs (snRNAs) and small-nucleolar RNAs (snoRNAs). The advent of high-throughput sequencing technologies, led to the discovery of many novel classes of non-coding RNAs. Even though they are still not

fully characterised, there is evidence to suggest they play important biological roles (Mattick, 2009). These are PIWI interacting RNAs (piRNAs), endogenous small-interfering RNAs (endo-siRNAs), and long non-coding RNAs (lncRNAs). It seems that we are still just beginning to glimpse the immense complexity of the transcriptional landscape and its regulation within mammalian cells (Saxena & Carninci, 2010).

Even though the work in this thesis is focused exclusively on animal miRNAs (Figure 1.1), I hope it will provide insights and methodologies that can be applied to other classes of ncRNAs.

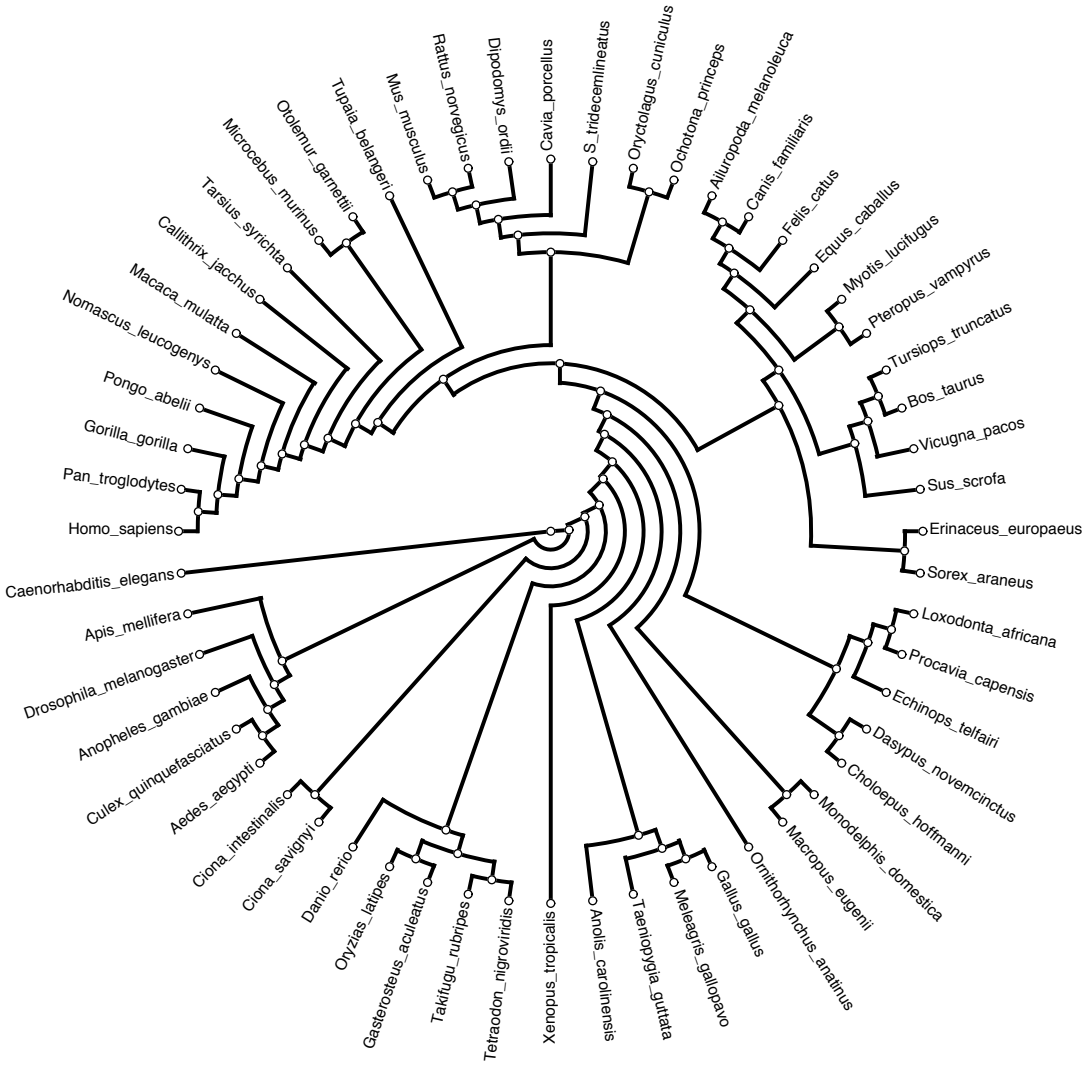


Figure 1.1: Evolutionary tree representing the phylogenetic relationship between the main species analysed within this work. It was computed from molecular data and retrieved from Ensembl (see Section 1.9.2).

1.2 microRNAs

1.2.1 The Discovery of miRNA Regulation

In 1993, the first miRNA, lin-4, was described as a regulator of lin-41 in *C. elegans* (Lee *et al.*, 1993; Wightman *et al.*, 1993). Although the mutation in lin-4 and its effects were known before (Ambros & Horvitz, 1987), this was the first time that it was demonstrated that lin-4 was a non-coding RNA, that directly regulates lin-41. This regulation is essential to define a development stage in at least 4 species of the *Caenorhabditis* genus. They also demonstrated that its mode of action was through the anti-sense binding to the 3'UTR of the lin-41 transcript, and that this regulation occurred post-transcriptionally.

This was first interpreted as an unusual mechanism of regulation, specific to the nematode lineage, and it remained that way until the identification of let-7 in 2000 (Reinhart *et al.*, 2000). Unlike lin-4, let-7 was found in a wide range of other species, spanning more than 400 million years of evolution (Pasquinelli *et al.*, 2000). Curiously, let-7 is also involved in the separation of developmental phases in *C. elegans*. Given their role in the setting of developmental timing, they were initially referred to as small temporal RNAs (stRNA) (Lee & Ambros, 2001).

The importance of this class of regulators was recognised soon after these studies, leading to a significant amount of research, focusing on the identification of novel loci (Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee & Ambros, 2001), their target sites (Enright *et al.*, 2003; Lewis *et al.*, 2005; Stark *et al.*, 2003) and likely cellular function (Giraldez *et al.*, 2006; Vigorito *et al.*, 2007).

1.2.2 miRNA Biogenesis

In parallel with the search for novel miRNA loci, there was a large community effort to identify the components of the miRNA maturation and processing machinery.

Primary miRNA transcripts (pri-miRNAs) are transcribed by RNA Pol II, possess a 5' cap and are 3' poly-adenylated (Cai *et al.*, 2004). Pri-miRNAs can encode one or more stem-loop secondary structures that will give rise to precursor miRNAs (pre-miRNAs), which are approximately 70 nucleotides long (see Figure 1.2).

Precursor miRNA loci are approximately 70bp long. In the canonical miRNA processing pathway (Figure 1.3) are formed by the recognition and cleavage of the pri-miRNA stem-loop structures by the RNAse III like enzyme Drosha (Lee *et al.*,

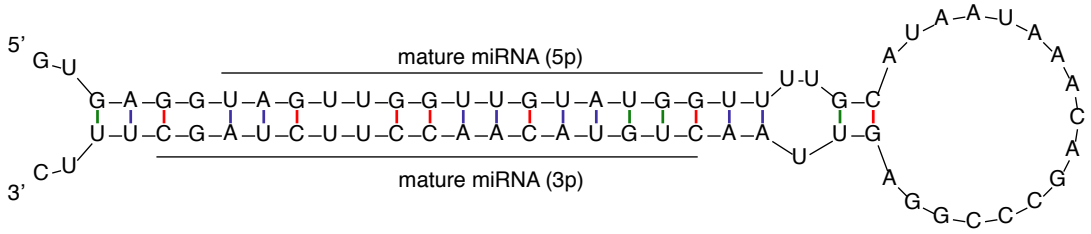


Figure 1.2: Schematic representation of a primary miRNA hairpin encoding one miRNA (let-7d). The interactions within the hairpin are coloured according to the base pair involved. Red corresponds to cytosine/guanine pairs, blue to adenine/uracil pairs, and green represents guanine/uracil "wobble" pairs.

2002). Drosha acts in a complex with Pasha (Partner of Drosha)/DGCR8 (DiGeorge Syndrome Critical Region 8), which is a double-stranded RNA binding protein (Gregory *et al.*, 2004).

This precursor hairpin is exported from the nucleus by Exportin-5, in a RanGTP dependent manner (Bohnsack *et al.*, 2004). In the cytoplasm, the pre-miRNA is cleaved by Dicer, another RNAse III like enzyme, that acts in a complex with other proteins including TRBP (the human immunodeficiency virus Transactivating Response RNA-Binding Protein). The result of this cleavage is a double stranded duplex of approximately 22bp in length, containing the mature miRNA and the miRNA* molecule (also known as the guide strand and passenger strand respectively). This duplex will frequently have 2bp 3' overhangs containing a 5' hydroxyl group, as is characteristic of this family of RNAses (Grishok *et al.*, 2001; Hutvagner *et al.*, 2001).

After cleavage by the Dicer/TRBP complex, the mature miRNA is loaded into a protein of the Argonaute (Ago) family, usually Argonaute 2 (Ago2), that will in turn recruit the other elements of the RNA-induced silencing complex (RISC) (Sontheimer, 2005). Upon loading, the passenger strand of the miRNA duplex will in most cases be degraded, while the guide strand will stay tethered to Argonaute and mediate target recognition. The loading process and RISC formation is slightly different between species (Yoda *et al.*, 2009). When loaded, this complex is responsible for target-recognition and for inactivating the target transcript or promoting its degradation.

It is believed that from each precursor, only one of the duplex strands will be functionally incorporated into Ago2 (Matranga *et al.*, 2005). This leads to the distinction between the mature miRNA, which is incorporated, and the miRNA* that

is degraded. By using the high coverage available in current sequencing methods, it has been found that the strand that gets incorporated can change depending on cellular conditions (Li *et al.*, 2012; Marco *et al.*, 2010). As either strand can be functional, a new naming scheme was devised, indicating from which arm of the pre-miRNA hairpin the mature sequence is being produced, deprecating the previous miRNA* annotation. As of release 19 of miRBase (Kozomara & Griffiths-Jones, 2011), all mature forms are now annotated as 5p for the mature form on the 5' arm of the hairpin and 3p for the mature form on the 3' arm of the hairpin, regardless of their relative expression level in the conditions profiled (Figure 1.2).

As our knowledge of the process expanded, some exceptions to these rules have been reported, where certain miRNA families are processed in a non-canonical fashion. For instance, the pre-miRNA can be formed in a Drosha independent fashion, by using the splicing machinery, if the miRNA forms an intron by itself (Okamura *et al.*, 2007; Ruby *et al.*, 2007). There have also been reports that the Dicer slicing step can be performed by Ago2, for instance, in the case of miR-451 (Cheloufi *et al.*, 2010; Cifuentes *et al.*, 2010).

1.2.3 miRNAs in Animals and Plants

Although they are functionally similar, the processing and mode of action of miRNAs in plants and animals show several differences (Axtell *et al.*, 2011; Voinnet, 2009).

In plants, the length of each miRNA hairpin is more heterogeneous, and can range from 70 to hundreds of nucleotides. Interestingly, while in animals each miRNA loci tends to produce only one miRNA/miRNA* duplex, some plant miRNAs are able to produce multiple duplexes from the same hairpin.

Arabidopsis thaliana, for example, does not have an homolog of Drosha. Instead, the main miRNA maturation steps occur in the nucleus and are performed by DCL1 (Kurihara & Watanabe, 2004). The miRNA/miRNA* duplex is exported to the cytoplasm by HASTY, an Exportin-5 homolog. In the cytoplasm, the mature miRNA gets loaded into AGO1. As opposed to animals, in plants target cleavage is usually performed by the slicer activity of Argonaute itself. Furthermore, while the target sites in animals are mostly restricted to the 3'UTR of the target transcripts, plant miRNAs frequently target the coding region directly.

The prevalence of miRNA regulation also seems to be different between the two kingdoms. It is predicted that 30% of human transcripts are targeted by miRNAs.

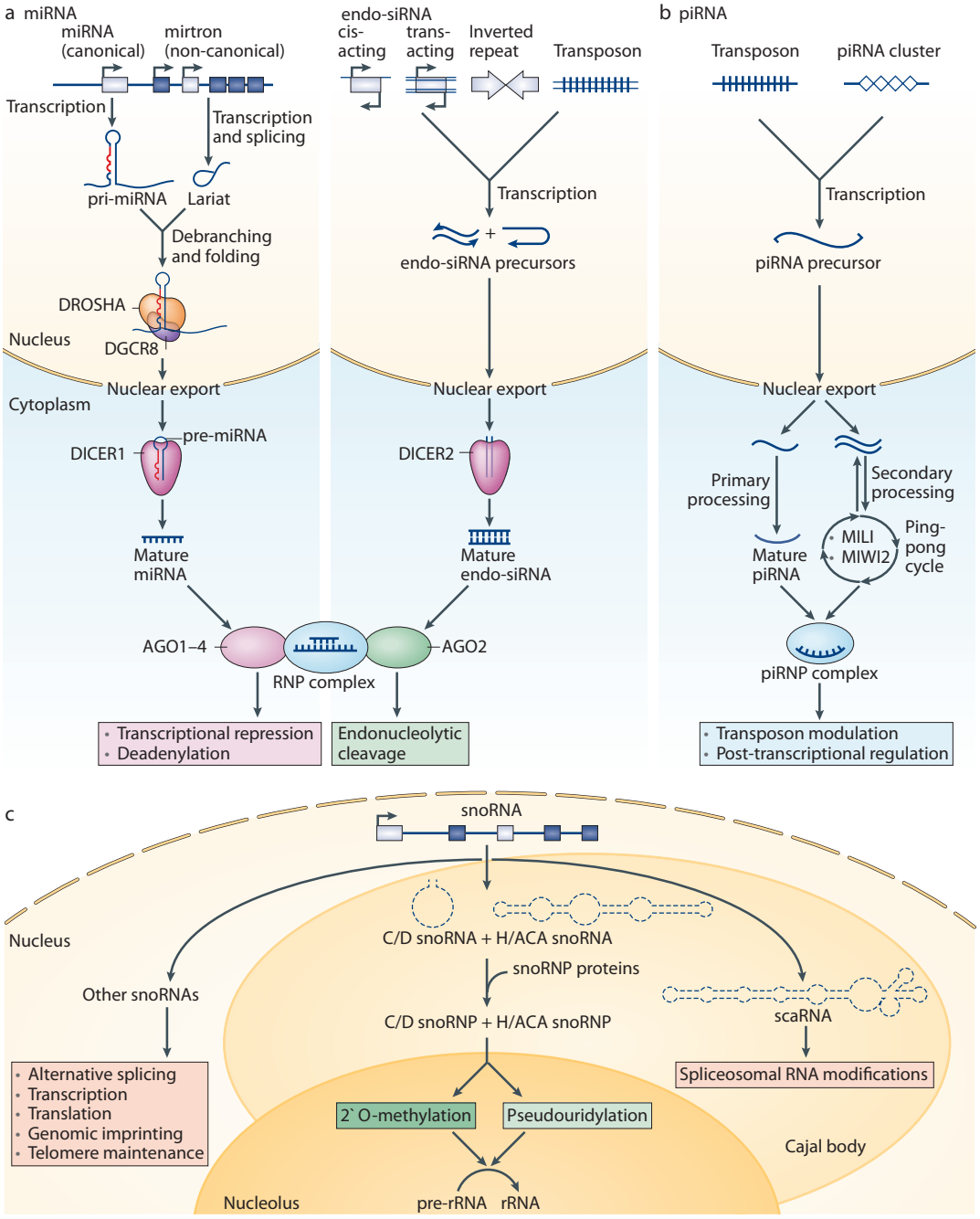


Figure 1.3: Schematic representation of the biogenesis and functions of animal ncRNAs. (Figure adapted from Qureshi & Mehler (2012))

In contrast, less than 1% of *A. thaliana* transcripts appear to be targeted by miRNAs (Fahlgren *et al.*, 2007).

While plant and animal miRNAs are though to have evolved independently, there have also been reports of miRNAs that appear to be shared by plants and animals (Arteaga-Vázquez *et al.*, 2006), and a recent report of rice miRNAs that are taken up and act on humans upon ingesting them (Zhang *et al.*, 2012).

1.2.4 Evolution of miRNA Biogenesis

Many of the catalytic domains contained within the miRNA processing machinery are already present in prokaryotes, albeit in proteins unrelated to small regulatory RNA processing. The different elements of the miRNA processing pathway tell different evolutionary stories. It is generally agreed that the active domains of the proteins that are part of the RNAi processing machinery were already present in Bacteria and Archaea (Shabalina & Koonin, 2008).

The phylogenetic distribution of these proteins is scattered in many of the more simple eukaryotes, with many species unable to use interference RNA effectively, but still retaining parts of the processing machinery. Interestingly, canonical miRNAs have not been found in Fungi (Drinnenberg *et al.*, 2009). Thus, work has also been done to understand the role of the RNAi machinery in the fission yeast *Schizosaccharomyces pombe*. Whilst Dicer is not essential for the viability of *S. pombe*, the deletion of Dicer causes slow growth, lagging chromosomes during anaphase and lack of silencing of centromeric repeats (Provost *et al.*, 2002). The known role of Dicer within RNAi processing pathways of slicing double stranded RNA into approximately 22bp fragments, appears to be conserved in budding yeast (Dang *et al.*, 2011). Furthermore, it has also been shown that the insertion of human Dicer partially rescues the endogenous Dicer deletion, supporting an evolutionarily conserved function (Provost *et al.*, 2002). In fission yeast, this protein seems to play an essential role in the formation and maintenance of heterochromatin at the centromeres and mating type loci, and its loss correlates with the loss of cohesin at centromeres (Hall *et al.*, 2003). It was also shown that the slicer activity of the Argonaute protein plays a key role in the process (Zoffal & Grewal, 2006).

Species that express Dicer usually possess a single dicer encoding gene. Curiously, in arthropods, there seems to have been a duplication, with species possessing two Dicer homologs. Nematodes possess a single Dicer copy, which supports the existence of a specific adaptation in arthropods. Studies in *Drosophila melanogaster* have

shown that Dcr1 seems to be responsible for miRNA processing, but not essential for dsRNA processing, whilst Dcr2 shows the opposite phenotype (Lee *et al.*, 2004).

The Pasha/DGCR8 protein seem to be one of the few components of the miRNA processing pathway that have a direct one-to-one orthologous relationship, without any known clade specific expansions. Different paralogues arose from large-scale protein duplications within vertebrates (e.g. in fish). There are also some lineage specific changes, which seem to occur in short evolutionary time-spans, such as the expansion of Argonaute proteins in plants (Mallory *et al.*, 2009).

Exportins have an unusual phyletic distribution, likely due to loss and re-adaptation of the available paralogues. Usually, Exportin-5 is responsible for the export of miRNAs out of the nucleus, with Exportin-1 exporting snRNAs and Exportin-T exporting tRNAs. However, in organisms that lack one of the exportins, the other seems to relax its specificity allowing the export of other ncRNA families (Murphy *et al.*, 2008).

The phylogenies of miRNA processing enzymes in general support the notion that plants and animals all evolved specific adaptations in this context. *Ciona intestinallis*, a deuterostome, has a single copy of each of the miRNA processing enzymes. It is interesting to note the case of Argonaute, that is usually present in multiple genomic copies in other species. The sequence of *C. intestinallis* Argonaute suggests it is ancestral to the Argonaute orthologues in vertebrates (Murphy *et al.*, 2008). The divergence and specialisation of the different Argonaute paralogues found in other organisms, seem to indicate that there might be other classes of small non-coding RNAs with specific functions that we might not be fully aware of (Ender & Meister, 2010). The full characterisation of the molecular functions of all these enzymes is ongoing, and is likely to provide some interesting insights into their evolution and functions outside of the miRNA/siRNA pathway.

1.2.4.1 miRNA Strand Selection

After Dicer cleavage, a duplex of approximately 22bp is formed. Unfortunately, it is not trivial to predict which strand of the miRNA hairpin will be incorporated into the RISC complex (mature sequence) and which strand will be targeted for degradation (star sequence). There is some evidence that the relative stability of the 5' end of the sequence will play a role (Jazdzewski *et al.*, 2008; Sun *et al.*, 2009). Nevertheless, these rules do not provide sufficient accuracy in predicting which strand is incorporated.

More recently, there have been reports of a correlation between strand selection and target availability (Chatterjee *et al.*, 2011). In this model, it is argued that both strands have comparable probabilities of being incorporated, nevertheless, the presence of a potential target sites causes the miRNA to be protected from degradation, causing it to be detected more frequently in sequencing runs. As it stands, the exact rules of miRNA recognition by Dicer and the RISC complex and subsequent incorporation still seem to be open to debate, precluding the development of accurate prediction tools.

1.2.5 Genomic Organisation

MiRNAs are not randomly distributed throughout the genome. It was found early on that miRNAs can form polycistronic transcripts consisting of clearly defined clusters within the genome (Lagos-Quintana *et al.*, 2001). It is often found that clusters were formed by local duplication of an existing miRNA locus. Nevertheless, there are also many cases of miRNA families with paralogues at different genomic locations, and also miRNA clusters containing a wide variety of miRNA families (Olena & Patton, 2009).

These loci can be found in several different patterns of genomic organisation (see Figure 1.4). MiRNA loci can be intergenic, encoded in monocystronic or polycistronic transcripts. They are also frequently found in the introns of protein-coding genes. In rare circumstances, miRNAs can also be found in the exons of protein-coding genes (Rodriguez *et al.*, 2004), or be derived from other classes of non-coding RNAs. It is important to note that what we consider to be exonic miRNAs is dependent on our knowledge of precise gene splicing patterns. It has also been found that miRNAs can form a whole intron by themselves, thus bypassing the requirement of Drosha for their processing (see Section 1.2.2).

Genomic miRNA clusters tend to be relatively small, rarely containing more than five or six distinct loci. Nevertheless there are exceptions. Human chromosome 14 contains the largest known cluster of miRNA loci that is conserved among many species, containing 37 miRNA loci, belonging to 6 distinct miRNA families. Other large clusters have been described, namely the cluster that is present on Human chromosome 19, and is conserved in most other primates that have been sequenced to date (see Figure 4.8). Repeat derived miRNAs can be located in locally duplicated clusters along the genome (e.g. miR-427 and miR-430) or be spread in an almost random fashion throughout the genome (e.g. miR-548).

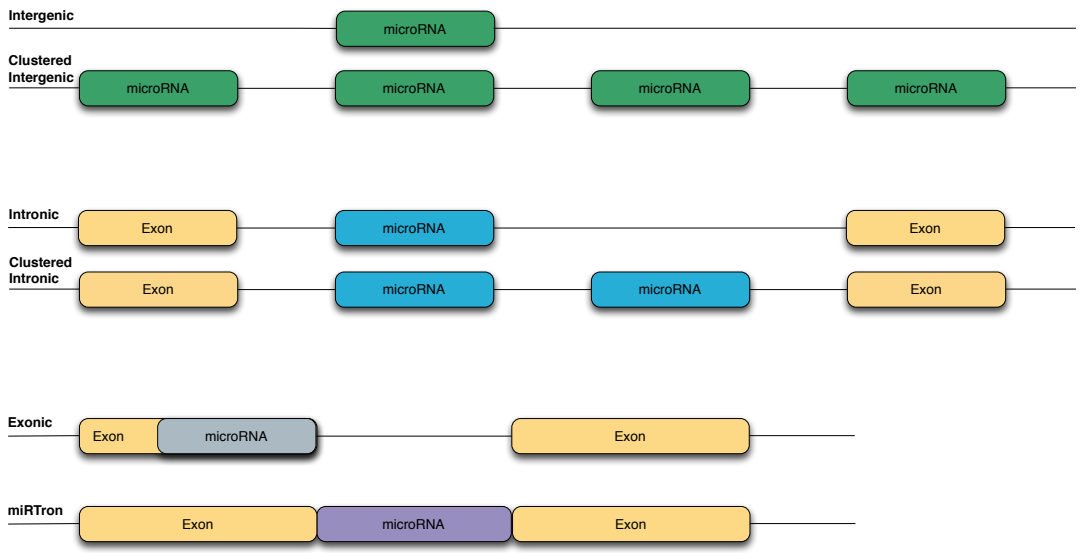


Figure 1.4: Possible genomic organisation of miRNA loci. Intergenic miRNA loci are illustrated in green, intronic miRNAs in blue. In rarer cases, miRNA loci can be contained inside an exon (grey), or be the exclusive feature within an intron (purple), which allows its maturation using the splicing machinery instead of requiring Drosha (miRTron).

1.3 miRNA Loci Profiling

1.3.1 Experimental Methods for miRNA Loci Discovery

Profiling of miRNAs can be defined as the assessment of miRNA expression in a given cell type and condition (Pritchard *et al.*, 2012). Several methods are available to do this, and are preferentially used depending on a wide range of factors. The most important considerations tend to be related to the amount of biological material available, the experimental design and final objectives. Initial miRNA profiling studies relied on capillary sequencing, frequently followed by northern blot analysis for validation of miRNA loci (Lagos-Quintana *et al.*, 2001, 2002, 2003). Despite the lower throughput compared with sequencing methods now available, these methods allowed an accurate, high-specificity profile of highly expressed miRNAs in several model organisms.

Currently, three main methods are commonly used for profiling miRNA sequences: qPCR is specific and sensitive, allowing for a wide dynamic range and is appropriate for absolute quantification of miRNA levels. It requires the smallest amount of biological material of the three methods presented, but its low-throughput

makes this approach impractical for large scale profiling. It is also not suitable for finding novel miRNAs. An alternative are miRNA microarrays, allowing a much higher throughput, albeit with the same limitation, whereby only known miRNAs can be profiled. There is also a loss of sensitivity and difficulties with quantification (Wang & Yang, 2010).

The final approach is small RNA high-throughput sequencing, which enables the search for novel miRNAs, provides a higher throughput and dynamic range than is possible with microarray technology. It also allows single base-pair resolution, making it possible to distinguish between iso-miRs (miRNAs that share the same set of targets, but which can have small differences in their mature sequence).

Its drawbacks are the higher cost, potential biases introduced during library preparation (e.g. amplification biases), and the significant computational resources that are required for the processing and analysis of the data produced. The quality of the data itself is highly dependent on the protocol used for library preparation, and can be prone to biases introduced at different steps of the protocol (Nekrutenko & Taylor, 2012).

1.3.2 Computational Methods for miRNA Loci Discovery

The initial challenges in experimental profiling of miRNA loci led to the development of computational methods for miRNA loci prediction (Lim *et al.*, 2003). By assessing the properties of previously annotated *C. elegans* miRNAs, the authors developed an algorithm to computationally detect novel miRNA. Their results were then validated using sequencing data, greatly expanding the number of *C. elegans* miRNA loci known at the time.

Since then, many other methods were developed to predict miRNA loci, based on conservation, sequence and structural properties of the candidate loci. MiRNA precursors form characteristic hairpin structures that can be assessed based on existing secondary structure prediction methods (Hofacker *et al.*, 1994; Jacobson & Zuker, 1993). Many methods exploit this information to compute metrics for miRNA candidate loci classification. The miRNA classifier methods then use different features to evaluate the structural stability and sequence properties of the candidate loci, to produce a final classification. The number of methods and implementations currently available make it challenging to compare and evaluate the performance of all existing methods.

In general, most methods follow a common logical flow: A candidate hairpin is provided to a secondary structure predictor; A diverse set of metrics, like thermodynamic stability, the number of unpaired nucleotides in the stem, the number of loops and loop length are then computed from the predicted secondary structure. The result of these computations is passed to a classifier function that will integrate these data and produce a final score. This function can range from a set of hard thresholds on each metric, a linear equation that combines these into a final score that is then filtered, or more complex machine learning techniques that provide a binary classification or probability for each candidate. Some of these methods will be described in more detail later (Chapter 2).

One of the major issues that affects the description of novel miRNA loci arises from the fact that purely computational methods require some sort of training and/or validation procedure. While this is not a bad thing in itself, researchers are then faced with the difficult choice of specifying a negative dataset for the analysis. Our understanding of the characteristics of genomic hairpins that are recognised and processed as miRNAs by Drosha and Dicer is still limited (Chiang *et al.*, 2010). Nevertheless, the community is now making efforts to address this problem, using sequencing data to identify which miRNAs are processed (Ritchie *et al.*, 2012).

The problem of using a non-optimal negative dataset for method development is particularly evident with machine learning methods. The strong statistical model and cross-validation procedures used in these cases can lead to over-fitting, hence obtaining good accuracy for recognising known miRNAs used for the training procedure, but producing less accurate scores for other miRNAs that were not included in the original training dataset.

Another common issue concerning the use of purely computational methods for *de novo* miRNA loci prediction, is that it is difficult to predict, without sequencing data, when a potential genomic hairpin will be expressed as RNA, and thus be available to be processed.

Many of these issues have started to be addressed with the use of small RNA sequencing data. This enables the definition of the exact mature sequence for a candidate miRNA loci and assessment of the expression level. Furthermore, processing enzymes of the RNAse III class leave characteristic 3' overhangs that can be detected if the sequencing depth and miRNA expression are high enough. Different methods were developed to explore these data to their full potential, for example miREna (Mathelier & Carbone, 2010), miRDeep (Friedlander *et al.*, 2008) and miRnouveau

which is described in detail later in this thesis (see Section 2.5). These methods are thus expected to deliver better, species independent, predictions of miRNA loci.

1.4 miRNA Targeting and Specificity

In animals, the mature miRNA guides the RNA induced silencing complex (RISC) to the binding site that is normally located in the 3' UTR of the target transcripts, with the binding specificity provided by the sequence complementarity of the seed region (nucleotides 2 to 8) of the mature miRNA to the target UTR (Lewis *et al.*, 2005). It is also reported that imperfect complementarity of the seed region can be compensated for by further complementarity between the 3' end of the mature miRNA and the target UTR (Bartel & Chen, 2004).

While the majority of miRNA::target occur through the binding to the 3' UTR, examples have been found of target sites located within the exons of protein coding genes (Lewis *et al.*, 2005; Tay *et al.*, 2008), and 5' UTRs (Lee *et al.*, 2009). Nevertheless, these are rarer and it has been postulated that ribosomes acting on these regions will compete with the RISC complex, reducing the effect of the miRNA mediated regulation (Bartel, 2009).

The binding of a RISC complex loaded with a miRNA to the target transcript can have a range of effects (Figure 1.5). Typically the translation of the target transcript can be inhibited by promoting ribosomal drop-off and degradation of the nascent peptide or blocking ribosome assembly and the initiation process itself. The target mRNA can also be de-adenylated and de-capped and thus marked for degradation (Fabian *et al.*, 2010; Giraldez *et al.*, 2006).

Whilst the common effect of miRNA regulation is target repression, there have been reports of transcription activation by miRNAs in a Human cell-culture system (Vasudevan *et al.*, 2007). Even so, it is still unclear how general and reproducible this phenomenon is.

Even though it is difficult to be sure of the biological relevance of miRNA targets predicted by current algorithms, there are a few rules that all of them take into consideration. The miRNA target recognition is mediated by the seed region (nucleotides 2-8) of the miRNA, that form Watson-Crick pairs with the target site, that is normally found in the 3' UTR of the target transcript. The first nucleotide of the miRNA seems to mediate tethering to Ago2, but not be necessarily complementary to the target site. It would appear that miRNAs which have an uracil at

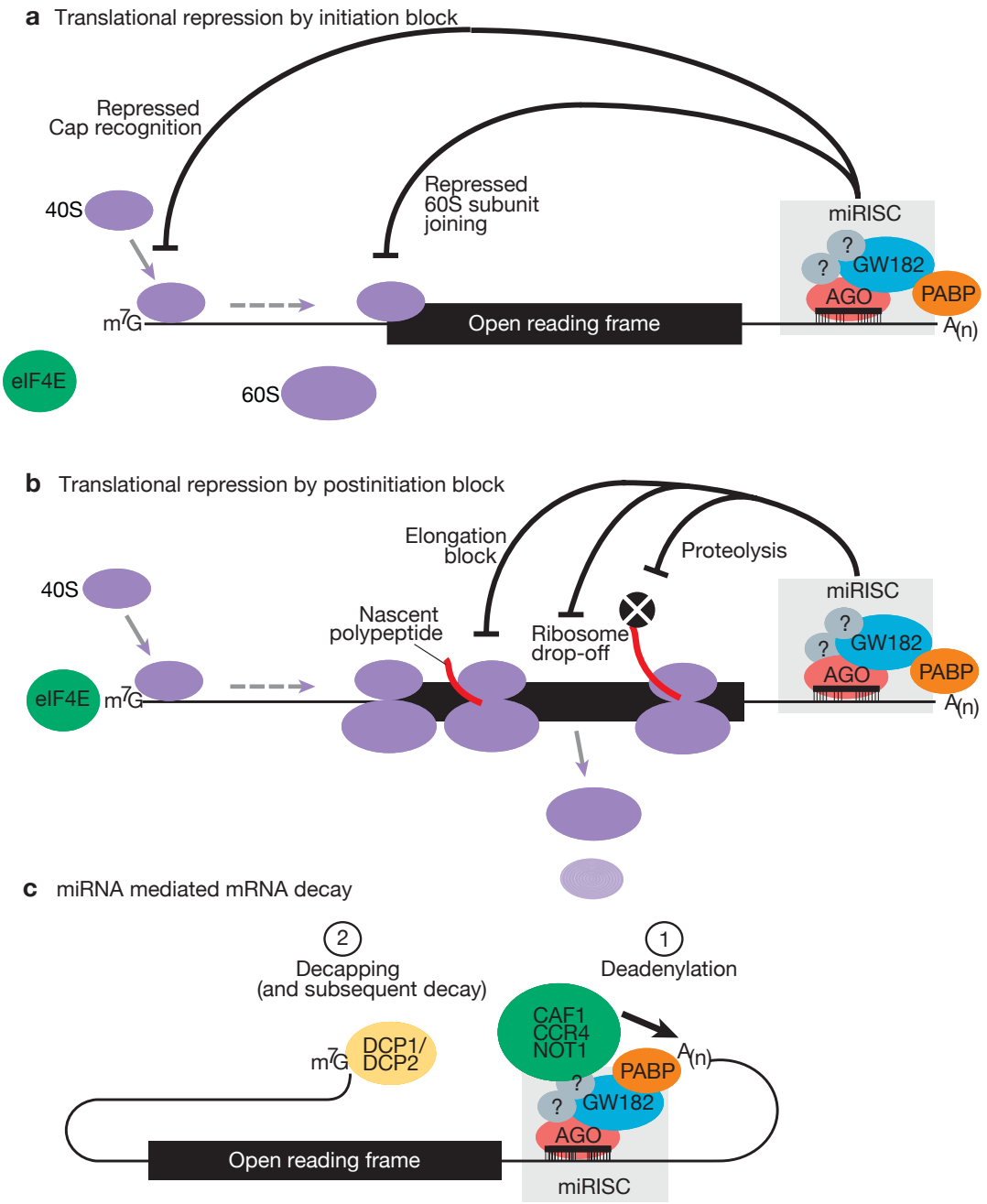


Figure 1.5: Illustration of the different modes of action of miRNAs. a) the RISC complex can act by preventing the assembly of the ribosome, thus blocking the initiation of translation. b) the RISC complex can block elongation, degradation of nascent peptide and ultimately drop-off of the ribosome. c) Alternatively, the RISC complex can induce the degradation of the target mRNAs by promoting its decapping and deadenylation (Figure adapted from Fabian *et al.* (2010))

this position are more efficient at repressing targets than other nucleotides that do not form Watson-Crick complementarity at the first nucleotide (Baek *et al.*, 2008).

The functional role of the 3' end of the mature miRNA is still open to debate. It has been postulated that in certain situations it can compensate for weaker seed matches (Bartel & Chen, 2004). To reduce false-positives in target prediction algorithms, it is recommended that putative target sites are filtered with high stringency criteria, requiring longer seed matches and excluding G:U wobble pairs in the seed region (Friedman *et al.*, 2009). However, it has been shown that some target interactions filtered out by these criteria are functional (e.g. let-7/lin-41 regulation) (Brennecke *et al.*, 2005).

In extreme cases where the miRNA is perfectly complementary to its target, it will act like an siRNA causing target cleavage instead of temporary repression. This repression mechanism seems to be common in plant miRNAs, sometimes targeting the coding region directly. Whilst direct target cleavage was thought to be rare in animal miRNAs, there has been evidence suggesting it is more prevalent than initially thought (Pillai *et al.*, 2007).

Using site conservation across species for target finding algorithms helps reduce the number of false positives. It is assumed that loci that are biologically relevant are more likely to be under purifying selection and thus be a biologically active target site, when compared to non-conserved putative target sites. On the other hand, it has also been shown that non-conserved predicted target sites can be functional (Ellwanger *et al.*, 2011; Farh *et al.*, 2005), illustrating that work still needs to be done before we fully understand miRNA target recognition mechanisms, and are able to achieve accurate computational predictions. Another aspect of miRNA-based regulation is that the target sites are not always independent (Doench & Sharp, 2004). If two target sites are located between 8 and 40 nucleotides apart, then they will act cooperatively and the repressive effects will be more significant than if the two sites were acting independently (Grimson *et al.*, 2007).

Whilst we now have a general picture of how miRNA regulation and targeting works, there is still work to be done to improve our understanding of miRNA targeting. Combining expression information for both target transcripts and the miRNAs themselves is essential to define many aspects of miRNA targeting mechanisms. So far, existing assays have been limited to highly expressed genes and miRNAs, where the effect can unambiguously be detected. As new technologies improve, allowing the expression profiling with greater sensitivity and dynamic range, this will lead to

a refinement of our current knowledge, by allowing the assessment of the effect of regulation by miRNAs in cases where their regulatory effects are subtle.

1.4.1 miRNA Target Prediction

1.4.1.1 Computational Methods

A series of computational methods were devised to predict and assess the potential for miRNA regulation, with many of the methods taking into account the sequence complementarity observed between the mature miRNA and 3' UTRs. The majority of algorithms use a similar set of features to classify each candidate target site, albeit with different weights given to each factor. These are usually seed region complementarity between the miRNA and the target site, free energy of the RNA duplex formed between the two, with some of the methods also taking into account regions surrounding the target site. To increase specificity, target site conservation is also often taken into account.

Two of the most used algorithms, and among the first to be proposed in the field are miRanda (Enright *et al.*, 2003; John *et al.*, 2004), and TargetScan (Friedman *et al.*, 2009; Garcia *et al.*, 2011; Grimson *et al.*, 2007; Lewis *et al.*, 2005).

The miRanda algorithm, was first described and applied for *D. melanogaster* miRNA targets (Enright *et al.*, 2003), and was later applied to *Homo sapiens* (John *et al.*, 2004) and incorporated into miRBase::Targets for a series of other species (Griffiths-Jones *et al.*, 2006). Its core algorithm is based on the local alignment of the miRNA sequence to the 3'UTR, giving different weights to different positions of the miRNA, favouring matches in the seed region, but not requiring full complementarity. Each match is then assessed for the thermodynamic stability of the RNA-RNA duplex, after which a conservation filter is applied.

TargetScan, on the other hand requires perfect seed complementarity between at least 6 nucleotides of the seed region and the 3' UTR, giving more significance to the perfect complementarity of the full seed sequence. The rules used in this algorithm were derived by a maximisation of the signal-to-noise ratio when comparing TargetScan predictions with validated miRNA targets and the background level of conserved heptamers in 3' UTRs (Lewis *et al.*, 2005). The algorithm has successively been updated to take more information into account, aiming at increasing its accuracy. It now includes information about the context of the target site within

the UTR (Garcia *et al.*, 2011; Grimson *et al.*, 2007), and conservation based metrics (Friedman *et al.*, 2009).

A wide range of other algorithms and approaches were proposed, many based on machine learning techniques. Nevertheless, as for miRNA loci finding algorithms, target finding algorithms that are not trained are also less prone to over-fitting, resulting in improved prediction accuracy. This can also make them easier to interpret in biological terms, something that is quite difficult to do with some of the machine learning based approaches that have been proposed.

It has also been found, when comparing target prediction methods, that the 3'UTR dataset used plays an important role in the results obtained. This shows, not only that it is not trivial to define the exact UTR sequence, but also that it is important to pay close attention to the dataset being used when comparing methods and results (Ritchie *et al.*, 2009).

As our knowledge about miRNA targeting rules increases, the available target prediction methods have also been updated to take more information into account. Nevertheless, we are still missing crucial biological insights into the targeting process, and the accuracy of these purely computational methods is still below what would be desired.

1.4.1.2 Expression-based Target Prediction

An alternative to the purely computational, *ab initio* target prediction methods described above, are methods that take into account the combined effects of miRNAs in target transcripts. One of these approaches, Sylamer (Van Dongen *et al.*, 2008) uses a gene-list, sorted from the most up-regulated to the most down-regulated transcripts in two contrasting conditions, and the 3'UTR sequences for those transcripts. These conditions can be a miRNA knock-out or knock-in experiment, contrasts between different time points of a time-course or the difference in transcript expression between disease states. Sylamer searches for significantly over-represented and under-represented k-mers at each end of the sorted list. As expected, the seed sequence for the miRNA that was perturbed appears over-represented in the transcripts that are being depleted when the miRNA is over-expressed. This allows the identification of miRNA-like effects and produces a characteristic plot. It is then possible to identify candidate transcripts that contain a miRNA seed in their 3' UTR sequences, thus confirming them as potentially direct targets of the miRNA.

A different approach, GenMIR++ (Huang *et al.*, 2007), uses a Bayesian data analysis algorithm to integrate expression data from miRNAs and mRNAs simultaneously, between different conditions, to infer the miRNA target network.

These approaches enable the easy computational detection of miRNA-like effects, at a large scale, between biological conditions. This allows a broad overview of the full regulatory network of the miRNA under study. Nevertheless, these approaches have two main drawbacks. They are only applicable to miRNA families whose expression gives rise to significant changes in gene expression and cannot unambiguously distinguish between direct and indirect targets of the miRNAs.

With the increase of the number of experiments profiling the changes in transcript level upon miRNA perturbation being made available, it is expected that these approaches will become more popular and useful.

1.4.1.3 Experimental Target Validation

Classically, miRNA targets have been validated *in vitro* by first creating a construct that fuses the 3'UTR of a candidate transcript to a reporter gene (e.g. Luciferase), and then measuring the reporter intensity in the presence and absence of candidate miRNAs.

The binding at a certain target site within the UTR can be further assessed by the insertion of point mutations within that target site do be probed. Luciferase activity when coupled with the mutant UTR is then compared with the wild-type in the presence of a miRNA mimic to assess the extent of the disruption caused by the mutation.

The main drawbacks of this approach are that it is a labour intensive process, and it is difficult to distinguish negative results from experimental failures. Additionally, the concentrations of miRNA that are present *in vitro* are much higher than the biological concentrations, raising concerns that the conclusions are not necessarily biologically relevant (Thomas *et al.*, 2010).

1.4.1.4 Experimental Target Determination

Technological advances now enable the direct assessment of miRNAs and their bound target sites *in vivo*. A method denominated *Cross-linking and Immuno-precipitation* (CLIP) (Ule *et al.*, 2003), initially developed to study alternative splicing in mouse brain, has been optimised for use in miRNA research.

These CLIP protocols use ultra violet (UV) radiation to induce a stable cross-linking of the protein Ago2 and the bound RNA, that can be either a miRNA, its target, or both. An antibody specific to Ago2 is then used to immunoprecipitate the protein-RNA complex and, because of the irreversible nature of the covalent bond, stringent purification conditions can be applied to remove remaining unbound RNA. Before the protein is depleted from the complex by a proteinase digestion, the RNA is partially digested in order to obtain short RNA tags containing the binding site. These tags are then sequenced and mapped to the corresponding genome. After further computational analysis, they can be used to infer the active miRNA/target duplexes.

There are already several different CLIP protocols, with the most frequently used being *High-Throughput Sequencing of RNA isolated by Cross-linking Immuno-precipitation* (HITS-CLIP) (Licatalosi *et al.*, 2008) and *Photoactivatable-ribonucleoside-enhanced Cross-linking and Immuno-precipitation* (PAR-CLIP) (Hafner *et al.*, 2010).

These approaches are changing our view of miRNAs, not only by providing miRNA targets but also by providing evidence of new modes of target recognition that can in turn be fed back into the computational analyses to improve predictions (Chi *et al.*, 2012). Ultimately, this creates an even bigger challenge, as it provides evidence for more models of how miRNAs recognise their targets, running the risk of increasing the false-discovery rate of the current computational methods. Furthermore, CLIP protocols are still very much in active development, to improve the cross-linking efficiency, antibody affinity and analysis of results.

1.5 Regulatory Function

Although originally found to regulate developmental timing in *C. elegans*, it soon became apparent that the potential range of activity of miRNAs was more far reaching. It has been predicted that over 30% of human protein-coding genes have targets sites for miRNAs (Lewis *et al.*, 2005), spanning most classes of biological processes (Filipowicz *et al.*, 2008). The transfection of particular tissue specific miRNAs (miR-1 and miR-124) in HeLa cells, that are then analysed using microarray technologies, showed that hundreds of genes change their expression profiles upon the over-expression of these miRNAs (Lim *et al.*, 2005).

The biological function of each miRNA family depends on its targets. Initially, miRNAs were identified based on the phenotypic consequences upon mutation, using

classic mutation based studies. It has been shown that, at least in *Caenorhabditis elegans*, many of the miRNAs identified, by sequencing and computational methods, are not essential, and few have detectable phenotypes upon mutation (Miska *et al.*, 2007). This can be explained by assuming that most miRNAs are fine-tuners of gene expression, that in most cases, would need a specific stress, in addition to the deletion of the miRNA, for the phenotype to manifest itself. It is also likely that the scope of action of the miRNA might be specific to a small number of cells within the organism, making its detection difficult.

Basing functional prediction purely on computational target prediction has several potential problems. The noise in target prediction makes it difficult to get statistically significant results for particular functional classes. Furthermore, these classes are frequently defined based on gene ontology (GO). This can present some challenges, as some miRNAs can have diverse cellular functions that do not necessarily fit the classes defined within the ontology in a statistically significant way.

A model of regulation by miRNAs has been summarised in (Bartel & Chen, 2004). Three main modes of action were proposed: Switch-like interaction, Tuning interaction and Neutral interaction. An example of a miRNA that acts in a switch like fashion is the development regulator, let-7 that represses lin-4. A tuning interaction can be characterised as one where the target needs to be kept at a reasonable level within the cell, but not eliminated. The miRNA acts to dampen protein output, but there is still an active pool of protein in the cell. Finally, neutral interactions are interactions that are not predicted to play an important biological role. These are normally not as conserved as the target sites that participate in the other types of miRNA/target interaction, as they are not under selective pressure.

Selective pressures affecting miRNA target sites can be varied. Besides neutrally evolving target sites, there are also known cases of purifying and positive selection in miRNA target sites. Purifying selection is expected to act on conserved target sites that play an essential role in cell regulation, ensuring that the pairing between the miRNA and the target is conserved. Conversely, it has also been shown that some transcripts show strong selection against the formation of potential target sites for miRNAs (Farh *et al.*, 2005). These so called anti-targets are particularly evident for miRNAs that show high expression and tissue specificity.

1.6 The Evolution of the miRNA Repertoire

1.6.1 On the Use of Gene Presence or Absence for Evolutionary Analysis

One of the first applications of a computer in evolutionary studies, was performed in the 1950s by Sneath. He used a computer system to classify bacterial strains based on a series of binary characters, determined by biochemical tests, which were used to compute a numeric value of similarity between strains (Sneath, 1957). This initial work was later developed into the seminal book on numerical taxonomy that details many of the methods still in use today (Sokal & Sneath, 1963). Despite their apparent simplicity, these algorithms would lead to many important evolutionary insights. However, their initial goal was the categorisation of species and the inference of phylogenetic trees, based on discretely coded characters, as large-scale sequence analysis was not a feasible option at that time.

It is generally believed that the path that requires the least state changes (e.g. gene gains or losses), and therefore the most parsimonious evolutionary scenario, is more likely than an explanation that requires many character state changes to justify the observed data. Various methods have been proposed based on this principle. They differ in the weight given to each transition, or restrictions applied to certain transitions. A choice must then be made based on the biological knowledge available (Felsenstein, 1983).

One such method, based on Dollo's principles (Dollo, 1893), was first suggested by Walter Le Quesne (Le Quesne, 1974), and further specified by Farris (Farris, 1977). This particular parsimony variant specifies that a character is only allowed to appear once (0 to 1 transition), with no restriction imposed on the number of times the gene can be lost. This is particularly useful for characters that are thought to appear rarely, and where no convergent evolution is to be expected. This approach can also be used, when a valid species phylogeny is available, to infer the ancestral state of each character under analysis on the internal nodes in the provided phylogenetic tree. When applied to the presence and absence of miRNA families throughout the metazoan phylogenetic tree, this provides a more detailed overview of the evolution of the miRNA repertoire within sequenced animal species.

1.6.2 On Exploring the Evolution of miRNA Gene Family Sizes

While some miRNA families are usually present in a single locus per genome, others have expanded, having large number of paralogues per species. These expansions can happen as a result of whole genome duplication events, transposable element activity for repeat element derived miRNAs, or other local duplications. Certain instances of local duplications of miRNA families that act in a switch-like fashion have already been described (e.g. dre-miR-430 (Giraldez *et al.*, 2006)).

Given the significant wealth of miRNA information currently available, I sought to identify and characterise other miRNA families that show unexpected loci expansions or deletions. This analysis builds upon the results obtained with Dollo parsimony analysis, by integrating loci count data. It is thus possible to explore changes in the number of paralogues per family per species.

To perform this analysis I focused on the CAFE tool (De Bie *et al.*, 2006) which implements a stochastic model of the birth and death of miRNA loci, to estimate the birth and death rate characteristic of miRNA families for a certain species phylogeny. In turn, this allows the detection of miRNA families that diverge significantly from what is expected. This can be either the sudden disappearance of a miRNA family that was normally present in multiple paralogues, or more commonly, miRNA families that sudden expand in a certain clade.

It is important to point out that due to potential issues with the available genome assemblies, it is difficult to distinguish a technical inability to detect a certain loci in a genome from actual gene loss. For this reason, my analyses focused primarily on gene gains, as these are likely to be more reliable.

1.6.3 Detection of Functional Associations Based on Correlated Evolution of Gene Families

In recent years, sequencing of new species is becoming commonplace, greatly expanding the amount of information available for genomic research. The vast amounts of sequencing data being produced need to be accompanied by genomic annotation, so that the sequence differences between organisms or experimental conditions can be interpreted in a biological context. One of the ways to perform functional annotation of genes, within biological pathways, is the use of correlations within phylogenetic

profiles. It seems reasonable to assume that genes that are part of the same biological pathways have a higher tendency to co-evolve, being gained and lost together more frequently than unrelated genes. For example, if the pathway is essential, there is a tendency for its elements to be maintained, while if the pathway is disrupted, it is likely that this will also affect the conservation of the other elements of the pathway. This phenomenon can be assessed through the analysis of phylogenetic profiles (Pellegrini *et al.*, 1999).

Phylogenetic profiles are matrices containing the presence and absence of the genes in sets of species (Figure 3.2 on page 66). This approach was shown to work well with simple correlation metrics in Prokaryotic genomes. Various metrics to correlate between the presence and absence profiles can be used (Kensche *et al.*, 2008). The application of these metrics to more divergent species, is likely to suffer from spurious correlations due to the phylogenetic distribution of the species under analysis. This can be addressed by taking the species phylogeny into consideration (Pellegrini, 2012).

Determining the function of poorly expressed or less studied miRNA families is still a significant challenge in the field. Therefore I sought to bridge this gap, integrating protein coding genes and miRNA loci information in a coherent dataset across species, and applying phylogenetic profile analysis.

1.7 The Evolution of miRNA Genomic Organisation

Even with the rapid expansion of sequencing data available, we still lack a global overview of the genomic organisation of miRNAs across a broad range of species, and an overview of their evolutionary relationships. Most previous studies, focused on specific clusters in a small set of species (Olena & Patton, 2009). A miRNA cluster is often transcribed as a single pri-miRNA, hence all its members are co-transcribed and are likely to participate in similar biological functions (Ooi *et al.*, 2011). Thus, if new miRNAs appear in already existing clusters, they will have a pre-defined expression pattern, and will more easily integrate in the cellular regulatory network.

There are very few miRNA families for which their evolutionary history has been inferred. Even for these families, the process was mostly inferred by manual curation (Hertel *et al.*, 2006; Tanzer & Stadler, 2004, 2006; Tanzer *et al.*, 2005). This does not provide a representative overview of miRNA evolution, and makes it difficult

to expand the analysis as more species are sequenced, or better genomic assemblies become available. Taking this new information into consideration will likely lead to the closing of gaps in our knowledge, especially in relation to miRNA families of more divergent species, or miRNA families that appear to be clade specific. For this reason it is important that an easy way to maintain and expand these analysis is provided.

As illustrated before, the evolution of the miRNA repertoire is far from being a static process. Many miRNAs show signs of having recently arisen, while others seem to have been lost in particular species. Local duplications within existing miRNA clusters are a frequent mode of evolution for new miRNA paralogues. It is essential that unambiguous homology links exist between the members of certain miRNA families for an evolutionary analysis to succeed. The evolutionary changes can appear in several ways, from local duplications within the same cluster, to whole cluster duplications.

There exists a wide range of conservation patterns within miRNA clusters. While some show perfect conservation across a wide range of species, others show minor changes in particular lineages, very rarely showing major rearrangements. Non-local duplications are almost exclusively associated with genome-wide duplications.

Some miRNA families are known to be derived from repetitive elements. These are spread across the genome in a variety of ways, making gene order analysis somewhat challenging. In some cases, all repeats seem to be clustered together, apparently deriving from a series of local duplications (e.g. dre-miR-430). Due to the difficulty in inferring the exact evolutionary scenario for repeat-element derived miRNAs, they are commonly excluded from evolutionary analysis.

A careful exploration of these phenomena, and of the conservation of miRNA clusters in general, can provide insights into the the evolutionary conservation of miRNA genomic organisation and the ways new loci integrate into the existing miRNA regulatory network. Furthermore, the availability of the resources developed in this thesis enable researchers that are interested in a particular miRNA family to take advantage of these analyses and to quickly access a list of related miRNAs that are co-localised with their miRNA family of interest, in a broader set of species.

1.8 Intra-specific miRNA Evolution

When proposing his theory of natural selection, Charles Darwin highlighted the importance of variability within a population (Darwin, 1859). Naturally, he was referring to the general phenotypic differences observed between individuals, and not changes at a molecular level.

At the molecular level, mutations arise when there is imperfect copying of the genetic information from one cell to the next during cell replication. Mutation heritability depends on the type of cell where mutations occur. Somatic mutations occur outside of the germ-line and thus are not passed on to the next generation. This type of mutation is of particular interest within cancer studies.Natural selection, on the other hand, is detectable at longer time-scales and is evident on heritable mutations, such as those occurring in the germ-line or during gamete production.

Mutations can be of different categories, depending on which process caused them and on their effects on the genome. These can be large genome rearrangements, insertions or deletions commonly called indels, or point mutations that are also called single nucleotide polymorphism (SNP). Insertions and deletions are rarer than SNPs and can have large effects on the sequences they affect. It is more difficult to computationally analyse their biological effects, in particular within non-coding sequences.

The consequences of these changes on the regions they affect can be detrimental to the original function, have no noticeable functional effect or improve the original sequence. These consequences are reflected by corresponding selection forces. Purifying selection, also known as negative selection corresponds to the selective removal of deleterious mutations from the population. By contrast, positive selection describes the fixation in the population of advantageous mutations.

With the advent of new sequencing technologies, there has been a significant increase in the amount of variation data available in public datasets. The study of variation of outbred populations, allows an overview of how natural selection is acting on certain genomic elements. This can in turn be used to make inferences regarding their biological importance within the organism. The different forces affecting a particular genomic region can be computationally detected by looking at the frequency at which mutations are occurring, and by analysing the allele frequencies of a particular SNP to determine the rate at which fixation is occurring within the population.

Some of these SNPs and indels will occur in miRNA loci or in the miRNA target sites. Depending on its location, a SNP in a miRNA loci can have a range of effects. The most extreme and disruptive, are mutations of the seed region of a mature miRNA, as these will dramatically change the target set the miRNA regulates. SNPs in other regions of the mature sequence will likely change the binding affinity of the miRNA to its target sites, but will be potentially less deleterious. Changes in the remaining parts of the precursor can change the structural stability of the hairpin, which will potentially affect miRNA processing.

Interestingly, despite their importance in gene regulation and homeostatic maintenance, only a small proportion of miRNAs in *C. elegans* are essential for survival under lab conditions (Miska *et al.*, 2007). Although not essential, many mutations disrupting miRNA loci or specific target sites have now been associated with disease phenotypes (Brest *et al.*, 2011; Esteller, 2011; Lewis *et al.*, 2009). It is thus natural to assume that functionally important regions will show evidence of negative selection, exhibiting a lower SNP frequency than adjacent regions.

Despite the fact that most mutations affecting miRNA loci are likely to cause disruption (Jazdzewski *et al.*, 2008), some mutations within miRNA loci seem to be advantageous, and show signs of recent positive selection (Hu *et al.*, 2012; Quach *et al.*, 2009). It is likely that with recent re-sequencing and genotyping efforts, other examples of positive selection within miRNAs will be reported, and a better understanding of the patterns of positive selection will be achieved.

While it is theoretically possible for miRNA loci and respective target sites to co-evolve, thus accumulating mutations while maintaining sequence complementarity between the most relevant miRNA target sets within its regulatory network, this is expected to be a rare event and there is no evidence at the moment that this process actually occurs. This is likely due to technical challenges posed by the relatively small number of SNPs found to affect miRNA loci. On the other hand, it may also be related to the requirement for variation affecting multiple independent target sites simultaneously, to maintain complementarity in order to prevent the disruption of the miRNA regulatory network, which would likely have a deleterious effect.

1.9 Data Resources

This research, being purely computational in nature, relies on publicly available datasets from several sources. As much as possible the work was carried out us-

ing miRNA information from miRBase (Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006, 2008; Kozomara & Griffiths-Jones, 2011) and associated links, and extra genomes and annotations obtained from the Ensembl (Flicek *et al.*, 2011a; Kersey *et al.*, 2011) family of resources.

1.9.1 miRBase

The microRNA Registry (Griffiths-Jones, 2004) was created as part of RFAM (Griffiths-Jones *et al.*, 2003), to address the need of a coherent resource for nomenclature, storage and annotation of miRNA sequences being discovered by a fast-growing community of researchers focused on miRNA biology. This resource was also entrusted with providing unique names, as agreed by the community (Ambros, 2003). With the continued growth of the data on miRNAs the resource was separated from RFAM and renamed miRBase, which to this day continues to be the primary repository of miRNA sequence. It is now hosted at the University of Manchester[1] (Kozomara & Griffiths-Jones, 2011).

1.9.1.1 Nomenclature

Recognising the importance of having a coherent naming scheme, especially due to the high degree of conservation across species of some miRNA families, an agreement was reached on suitable miRNA classification guidelines and a naming scheme (Ambros, 2003). The classification guidelines attempt to define a set of rules that distinguish miRNAs from siRNAs. The latter are also processed by Dicer, and thus have some similarities to miRNAs.

Animal miRNA names consist of "mir-" followed by a unique sequential numeric identifier. Slightly different variants of the same miRNA, present in multiple copies in the same genome, can have a letter appended to the name (e.g. mir-1a), whilst paralogues have a dash followed by the paralogue number (e.g. mir-2-1). When referring to the mature miRNA, an uppercase 'R' in miR- should be used. Names are also preceded by a three letter species identifier, where the first letter corresponds to the genus, while the two others correspond to the species (e.g. hsa for *H. sapiens*).

When new miRNAs are discovered, a submission should be made to miRBase. The unique miRNA identifiers will be attributed by miRBase upon acceptance of a new manuscript for publication.

[1] http://www.mirbase.org

1.9.1.2 Genomic Context

Besides ensuring the correct nomenclature of miRNAs, miRBase also provides resources to the community that enable an easy exploration of the context of each miRNA. Since most miRNAs are co-transcribed either with other miRNAs or with protein-coding genes, knowing which genomic features surround the loci is important, particularly if genetic manipulation is being performed. This information is provided in each miRNA page. Further details can be found on the miRBase::Genomics section of the website, providing in depth information of pri-miRNA boundaries, transcription start site (TSS), expressed sequence tag (EST) evidence, information on CpG islands, Poly-A sites and promoter elements (Saini *et al.*, 2008).

1.9.1.3 miRNA Validity

As the primary repository for miRNA sequences, miRBase is often used as a gold-standard for miRNA analyses. To support miRNA validity, miRBase provides a simple classification indicating the source of evidence (e.g. Sequencing, Northern Blotting, Cloning, Homology). Furthermore, it lists the original literature sources that describe the miRNA discovery. More recently (Kozomara & Griffiths-Jones, 2011), data from small RNA sequencing experiments is being incorporated in miRBase, further providing evidence for miRNA expression in certain tissues and conditions. It is still difficult to define the validity of some miRNAs where there is very little evidence present, as it is impossible to distinguish between profiling biases and lack of sufficient evidence (Chiang *et al.*, 2010).

1.9.1.4 miRNA Targets

Another crucial aspect of miRNA biology is the identification of miRNA targets. Since no definitive computational method has been found so far to accurately predict miRNA targets, and experimentally validating targets is still a low-throughput process, miRBase links to several resources that contain miRNA target predictions. This allows the user to decide which sources to use depending on the research questions being asked. Besides linking to tarBase (Vergoulis *et al.*, 2011) which contains experimentally validated miRNA targets, miRBase links to several computational resources, including TargetScan (Lewis *et al.*, 2005), microCosm (previously known as miRBase::Targets (Griffiths-Jones *et al.*, 2006)), DIANA-microT (Maragkakis *et al.*,

2011) and miRDB (Wang, 2008). Some of these methods were briefly described in Section 1.4.1.1.

1.9.2 Ensembl

Initially developed to allow the easy analysis and data mining of the data produced by the Human genome project (Lander *et al.*, 2001; Venter *et al.*, 2001), Ensembl has since greatly expanded encompassing more than 100 species and microbial strains across all kingdoms of life (Flicek *et al.*, 2011a; Kersey *et al.*, 2011). Ensembl is now divided into different sub-projects, each focused on providing data and resources for the study of different facets of genomics. I will focus on the resources used for the analyses within this thesis.

1.9.2.1 Ensembl and Ensembl Genomes

The main Ensembl web-resource focuses on providing an easy to access graphical web interface for the exploration of genomic regions of interest, for a wide variety of vertebrate genomes. Behind each new release, there are complex and robust data integration and analysis pipelines for genome annotation that use the data that is publicly available in other resources, as well as computational prediction methods. The results are a set of annotations across species. The use of the same pipeline across species makes these data ideal for cross species analyses. Furthermore, the dataset can be easily downloaded, queried through the web interface, or integrated in scripts using the API provided.

The same pipelines and interfaces are now applied to other organisms beyond vertebrates, available through Ensembl Genomes (Kersey *et al.*, 2011), covering invertebrate species as well as the other kingdoms of life. The availability of a standardised procedure for the annotation of different species is essential when the objective is comparative analysis. Furthermore, the resources are updated following a predictable schedule, which makes it easier to manage resources that depend on it.

1.9.2.2 Ensembl Compara

Ensembl Compara (Flicek *et al.*, 2011b; Vilella *et al.*, 2009), builds on the main Ensembl dataset and provides resources for comparative genomics analysis. Protein-coding genes are organised into families, based on the clustering with Uniprot, en-

abling coherent annotation between species. Ensembl Compara also provides other resources useful for comparative genomics: multiple sequence alignments between pairs of species and within certain groups of taxonomic units; Phylogenetic trees for protein and some ncRNA families; inter-species conservation tracks.

1.9.2.3 Ensembl Variation

As more and more studies are focused on detecting and quantifying intra-species variation in different populations, it is increasingly important to have resources that integrate the resulting data from these studies.

Collecting data from dbSNP (Sherry *et al.*, 2001) and other publicly available sources, Ensembl Variation provides intra-specific variation data, and its annotation depending on the genomic features it affects. This information is then combined with phenotypic information available. This resource provides a common source for variation information across a wide variety of species.

Chapter 2

Defining microRNA Loci Based on Homology and RNA Sequencing

2.1 Aim

Although the data available for microRNA (miRNA) research are expanding exponentially, not all species are being assessed at the same depth. To produce a consistent dataset enabling comparative genomics and building on existing resources, I developed a novel resource for species-independent miRNA mapping. Its use within this project greatly expands the evolutionary space that can be confidently explored, contributing to the main aim of finding evolutionary patterns affecting miRNA loci.

2.2 Introduction

The miRBase database is the primary repository for miRNA data (Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006, 2008; Kozomara & Griffiths-Jones, 2011). It focuses on both nomenclature and recording of precursor and mature sequences and their probable genomic loci. Currently, a large proportion of deposited miRNAs are derived from model organisms (e.g. *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*), with many other species lacking proper miRNA profiling and annotation (e.g. *Gorilla gorilla*).

In other cases even in one species there may be multiple genomic loci which could encode for a particular miRNA family and not all of these may be annotated in miRBase. This implicit bias towards model organisms hampers miRNA research

in other organisms and makes evolutionary analysis of miRNA families across species extremely difficult.

Given that many miRNAs are highly conserved between species (Pasquinelli *et al.*, 2000) it is likely, for example, that a miRNA discovered in *C. elegans* will also be present in *Caenorhabditis briggsae* or other nematodes. With a characterised mature miRNA sequence in one species it is possible to detect the likely location of its orthologue in another species, or further paralogues in the original species, by combining sequence analysis and RNA secondary structure prediction (Berezikov *et al.*, 2006).

Although many methods have been developed to map miRNAs across species, very few were made available to the research community. This restricts their use outside the lab they were developed at. The tools that are freely available to the community are frequently clade specific or exhibited significant bugs when local analyses were attempted. This hampered my efforts to use pre-existing methods to create a high-confidence expanded dataset of miRNA loci mappings across a wide range of animal genomes. For these reasons, I decided to create MapMi, a novel method that builds on the knowledge acquired in the field and that would be released in an open-source model so that others can modify and improve it. The method aims to be accurate, species-independent and fast enough to be useful when processing large datasets (Guerra-Assunção & Enright, 2010).

The assumption is that an orthologous miRNA will possess both a high degree of similarity to the miRNA mature sequence and that identified orthologous loci should have the capability to form the stem-loop structure typical of miRNA precursors. Some groups use *ad hoc* methods for miRNA mapping analysis, however such approaches are generally either not available to the community, have not been validated or are too specific for general use.

For example, miROrtho (Gerlach *et al.*, 2009) provides web-access but not software or raw data, while CoGemiR (Maselli *et al.*, 2008) provides raw data but does not allow sequence searches. Another tool, miRNAminer (Artzi *et al.*, 2008) requires the user to provide both the mature sequence and the precursor sequence and runs on a limited set of species. For these reasons, it is very difficult to directly compare the existing methods to MapMi in terms of performance. However, when possible, MapMi results were compared against predictions from CoGemiR, miRNAminer and miROrtho (see Section 2.4.3). The most complete comparison is with miROrtho where there is a high degree of overlap between the methods, for

the species where data from miROrtho is available. When human miRBase (v14) miRNAs are used as a reference set, both methods predict a shared set of 478 loci, while miROrtho predicts 49 loci that MapMi does not and MapMi detects 139 loci that were not identified by miROrtho.

Many methods have focused exclusively on the classification and prediction of novel miRNAs from genomic hairpins (Mendes *et al.*, 2009) which is a non-trivial problem, when addressed in a purely computational way. This approach focuses, in the first instance on the simpler task of mapping an identified miRNA in one species to others using both sequence similarity and RNA secondary structure. To confidently map novel miRNA loci, sequencing data and extra filtering steps are used to improve the prediction accuracy (see Section 2.5).

The MapMi pipeline is freely available as both software and a web interface[1]. For convenience, a full dataset of pre-computed mappings can also be downloaded or browsed through the available web interface.

This method was developed based on 46 Ensembl genomes (Hubbard *et al.*, 2009) and 21 Ensembl Metazoa genomes (Kersey *et al.*, 2009). During the course of the project, this has subsequently been updated to incorporate suggestions and feature requests from users, as well as the addition of extra species and updates to the genome assemblies to match the latest versions of Ensembl and Ensembl Metazoa.

2.3 Implementation

2.3.1 Pipeline

The MapMi pipeline works as follows (Figure 2.1). The system is supplied with a set of input sequences corresponding to mature miRNA sequences. The user then decides which species to map these sequences against. The stand-alone version of MapMi allows the user to supply their own candidate mature miRNA sequences, as well as genomic sequences. The provided input sequences are scanned against selected genomes using the Bowtie algorithm (Langmead *et al.*, 2009), which is designed for efficient short sequence matching. The system allows no gaps but up to three mismatches, allowing one mismatch by default. Each match is extended to produce a pair of potential miRNA precursors through extension of 110nt (e.g. 70nt

[1] http://www.ebi.ac.uk/enright-srv/MapMi/

5' and 40nt 3' and *vice versa*). Each of these potential precursors is then folded using RNAfold (v1.8.5) from the *ViennaRNA* package (Hofacker *et al.*, 1994).

A scoring function is used to evaluate each candidate. The scoring function (see Equation 2.1) takes into account both the quality of the sequence match and the structure of any predicted hairpin. The best candidate is selected based on the score (either 5' or 3'). Candidates are further filtered according to a score-threshold.

This is defined by the user, however a number of suggested thresholds are provided. These thresholds have been selected according to an empirical analysis of known miRNA and di-nucleotide shuffled miRNA sequences (see Table 2.4). All miRNA precursor loci above threshold are reported to the user with their associated scores and other relevant information. As an alternative, the user can query a database of pre-computed results, using a miRNA name as a query, and selecting the desired species and threshold.

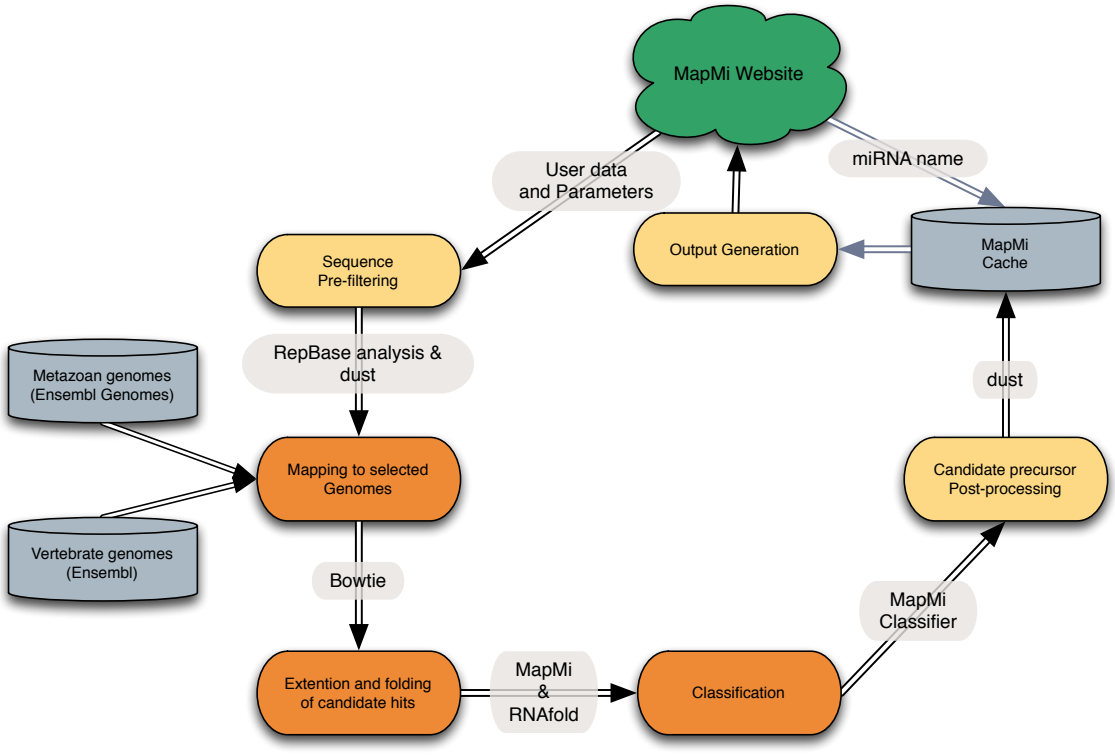


Figure 2.1: Workflow of the MapMi webserver and pipeline. The user can use the service by either providing a set of potential mature sequences to map against one of the available genomes, or by querying the results database. The results can be queried either using a miRNA name or a job ID from a previous run.

2.3.2 Repeat Element Derived microRNAs

Several miRNA families have been shown to be derived from repeat elements, in particular in mammalian species (Borchert *et al.*, 2011; Piriyapongsa *et al.*, 2007). Repeat elements present a challenge for miRNA searches. Their faster rate of evolution and the fact that they have a higher number of loci in the genome, makes it difficult to determine with confidence which of the candidate loci are actively producing miRNAs and which are miRNA pseudo-genes. Therefore, removing repeat elements from miRNA mapping analyses is likely to be the most reliable option, in order to reduce the number of potential pseudo-genes in the dataset.

A repeat masking procedure applied to the genomes prior to the analysis is useful to avoid the detection of repeat elements that contain sequences similar to known miRNAs. For this reason, the genomes under analysis were processed using RepeatMasker (Smit *et al.*, 2004) to remove repetitive elements. (see Table 2.3.2). Nevertheless, as a consequence of this procedure some miRBase annotated miRNAs may be masked and therefore reduce the sensitivity of the MapMi method (see also Tables 2.1 and 2.2).

This filtering can be disabled if the pipeline is to be used to study a particular family of miRNAs that is repeat associated (Hu *et al.*, 2012). Furthermore, a detailed analysis of which miRNA families were excluded during this filtering step can also be performed, to get a better insight in which families are derived from repeat elements (Tables 2.1, 2.2 and 2.3.2).

Repeat Element Type	Frequency
Type II Transposons	28.78%
Type I Transposons/SINE	21.70%
Type I Transposons/LINE	14.29%
Tandem repeats	10.75%
Unknown	9.79%
LTRs	7.73%
RNA repeats	3.28%
Low complexity regions	2.46%
Simple repeats	0.96%
Satellite repeats	0.18%
Other repeats	0.10%

Table 2.1: Summary of repeat elements overlapping MapMi predicted miRNAs, for the species under analysis if no repeat filtering is used. The Ensembl API was used to assess the overlap of MapMi predicted loci with annotated repeat elements. The parameters used for this run were the default.

Repeat Element Type	Frequency
Type I Transposons/SINE	31.14%
Type II Transposons	28.45%
Type I Transposons/LINE	17.18%
Tandem repeats	11.57%
LTRs	5.65%
Unknown	2.13%
Simple repeats	1.74%
Low complexity regions	1.70%
RNA repeats	0.30%
Satellite repeats	0.11%
Other repeats	0.03%

Table 2.2: Summary of the overlap between repeat elements and miRBase deposited miRNA loci for the species that are present in Ensembl and have miRBase coordinates available. The same procedure as for Table 2.1 was used.

bmo-miR-2728	eca-miR-1302d	eca-miR-1302d	hsa-miR-548h	mmu-miR-2138	mmu-miR-709
bmo-miR-2743	eca-miR-1302e	eca-miR-1302e	hsa-miR-720	mmu-miR-2140	mmu-miR-720
bmo-miR-2747	gga-miR-1810	gga-miR-1810	mdo-miR-151	mmu-miR-2141	oan-miR-1386
bmo-miR-2749	hsa-miR-1246	hsa-miR-1246	mdo-miR-739	mmu-miR-2142	ptr-miR-1227
bmo-miR-2750	hsa-miR-1255b	hsa-miR-1255b	mml-miR-616	mmu-miR-2144	ptr-miR-1246
bmo-miR-2753	hsa-miR-1260	hsa-miR-1260	mmu-miR-1937a	mmu-miR-2146	ptr-miR-1274b
bta-miR-1814a	hsa-miR-1274a	hsa-miR-1274a	mmu-miR-1937b	mmu-miR-466f	ptr-miR-1302
bta-miR-544b	hsa-miR-1274b	hsa-miR-1274b	mmu-miR-2132	mmu-miR-467g	ptr-miR-548f
cfa-miR-1271	hsa-miR-1975	hsa-miR-1975	mmu-miR-2133	mmu-miR-690	ptr-miR-720
eca-miR-1302	hsa-miR-548f	hsa-miR-548f	mmu-miR-2135	mmu-miR-706	

Table 2.3: List of input miRNAs, retrieved from miRBase version 13, that were found to be associated with repeat elements, either because they were overrepresented in the analysis performed without repeat masking and/or matched without mismatches to one or more sequences in Repbase Update (Volume 14, Issue 8).

2.3.3 Phylogenetic Analysis of microRNAs

The high degree of sequence conservation between miRNAs across a wide range of species makes them ideally suited as phylogenetic markers in large scale phylogenetic studies, in particular in conjunction with morphological markers (Rota-Stabelli *et al.*, 2011). Nevertheless, their small length, high sequence similarity across the mature sequence and higher divergence within the loop region, pose some challenges for the analysis of the phylogenetic signal exclusively from the miRNA sequences.

Instead, classic phylogenetic methods can be used to detect particular patterns within each miRNA family, such as conservation profiles and rapid sequence divergence within specific clades. To explore these facets of miRNA evolution, the dataset generated by MapMi was subdivided by miRNA families. Subsequently, a multiple sequence alignment, phylogenetic tree and consensus sequence and structure were calculated for each family. These results are available in interactive viewers within the pre-computed results section of the MapMi website.

Multiple sequence alignments were performed using the MUSCLE program (Edgar, 2004), and can be interactively explored on the website in Jalview (Waterhouse *et al.*, 2009). Maximum-Likelihood phylogenetic trees were computed using PhyML (Guindon & Gascuel, 2003) and are displayed on the website using the PhyloWidget interface (Jordan & Piel, 2008).

Finally, to display the patterns of conservation in the context of the predicted secondary structure assumed by the miRNA hairpin, RNA structural logos were generated using RNALogo (Chang *et al.*, 2008), enabling an easy visualisation of these properties for each miRNA family. The RNA structural logos combine the properties of the common sequence logos, where for each position, the relative size of each nucleotide is proportional to the frequency with which the nucleotide appears in said position in a multiple sequence alignment, with a consensus secondary structure of the RNA being analysed.

In most miRNA families, this also enables the easy identification of the limits of the pre-miRNA hairpin, due to their higher conservation across species in comparison with adjacent base pairs.

2.3.4 Scoring Function

MapMi takes into account several properties of known miRNAs in its scoring function (Equation 2.1). In this context, *Mismatches*, *Matches* and *PerfectMatches*

correspond to the number of nucleotides that are part of the predicted structure between the two arms of the stem loop. *Mismatches* correspond to the number of structurally unpaired bases, *Matches* correspond to the number of structurally paired bases and *PerfectMatches* to actual base-pairing.

Mature Mismatches are obtained by parsing the output of Bowtie. The $Hairpin\Delta G$ is the value of minimum free energy (MFE) returned by RNAfold, which corresponds to the estimated energy required to sever the bonds that form the secondary structure of the RNA. *MismatchPenalty* is a parameter specified by the user. The *MismatchPenalty* parameter is important to distinguish sequences with mismatches from sequences with no mismatches, that can match to the same loci. This is frequent for miRNA families that possess many subfamilies, with few differences at the mature miRNA level. The parameter can be set to a value that is large enough to enable this distinction but at the same time does not hamper the accuracy of the method by penalising mismatches too much (i.e. excluding sequences that have less than the maximum allowed number of mismatches, because the penalty is too high). A warning is displayed if this is likely to be the case.

The scoring function is composed of three parts. The first part scores the structural pairings between the two arms of the candidate hairpin. The second component integrates the minimum free energy of the hairpin. The third and last part scores the mapping of the candidate mature sequence against the genome.

$$\begin{aligned} Score &= (\tfrac{PerfectMatch*Match}{Match+Mismatch}) \\ &+ abs(\tfrac{Hairpin\Delta G}{2}) \\ &+ (1 - MatureMismatches) * MismatchPenalty \end{aligned} \tag{2.1}$$

2.3.5 Imposing Constraints on Hairpin Properties

Besides the score, a certain number of constraints can be placed on the predicted hairpins to increase specificity or simply to tailor the pipeline to more specific searches. It is possible to specify the minimum precursor length. The user can also define a minimum value for the ratio between paired and unpaired bases within the hairpin stem, and the absolute minimum number of paired bases within the stem. Even though there are cases described where the mature miRNA is part of the loop (Cheloufi *et al.*, 2010), this is uncommon. For this reason an option enables the user to restrict the number of base pairs that are allowed to overlap the loop, if any.

Concerning the actual mapping of the candidate mature sequences to the genome, it is possible to specify the number of mismatches that are allowed for the mapping. By default, no mismatches in the seed region are allowed, as this would change the miRNA target set, and thus the family the sequence belongs to.

Finally, it is possible to exclude candidate mature sequences that are present in the genome more than a specified number of times. This complements the repeat element analysis and filtering while excluding candidates that can potentially be generated from degradation fragments around the genome, and not really from a potential miRNA loci.

2.4 Results

I applied the MapMi pipeline to the repeat masked genome of the 67 species in Ensembl (release 55) and Ensembl Metazoa (release 2). This was done using all 7,844 miRBase (v14) metazoan miRNA sequences, corresponding to 4,237 unique mature sequences. In total, 16,025 loci were identified in all genomes under analysis using the default threshold of 35 (see Table 2.4), including 10,944 loci not previously reported in miRBase (Table 2.7). The phylogenetic profiles of miRNAs in each species are shown (see Chapter 3, Figures 2.2 and 2.3). For short evolutionary distances, the dendrogram obtained from the clustering of these phylogenetic profiles broadly agrees with known phylogenetic relationships between organisms (Figure 2.4). Genomes were masked for repetitive elements before further analysis (see Section 2.3.2).

2.4.1 Validation Datasets

The negative dataset was generated by using *ushuffle* (Jiang *et al.*, 2008) to generate 10 and 100 shuffles per initial nucleotide sequence. Due to their nucleotide composition, some of the 4,237 initial sequences could not be shuffled the required number of times. The resulting datasets contained 42,366 and 423,343 random shuffled sequences respectively. These datasets were mapped against all 67 genomes under analysis.

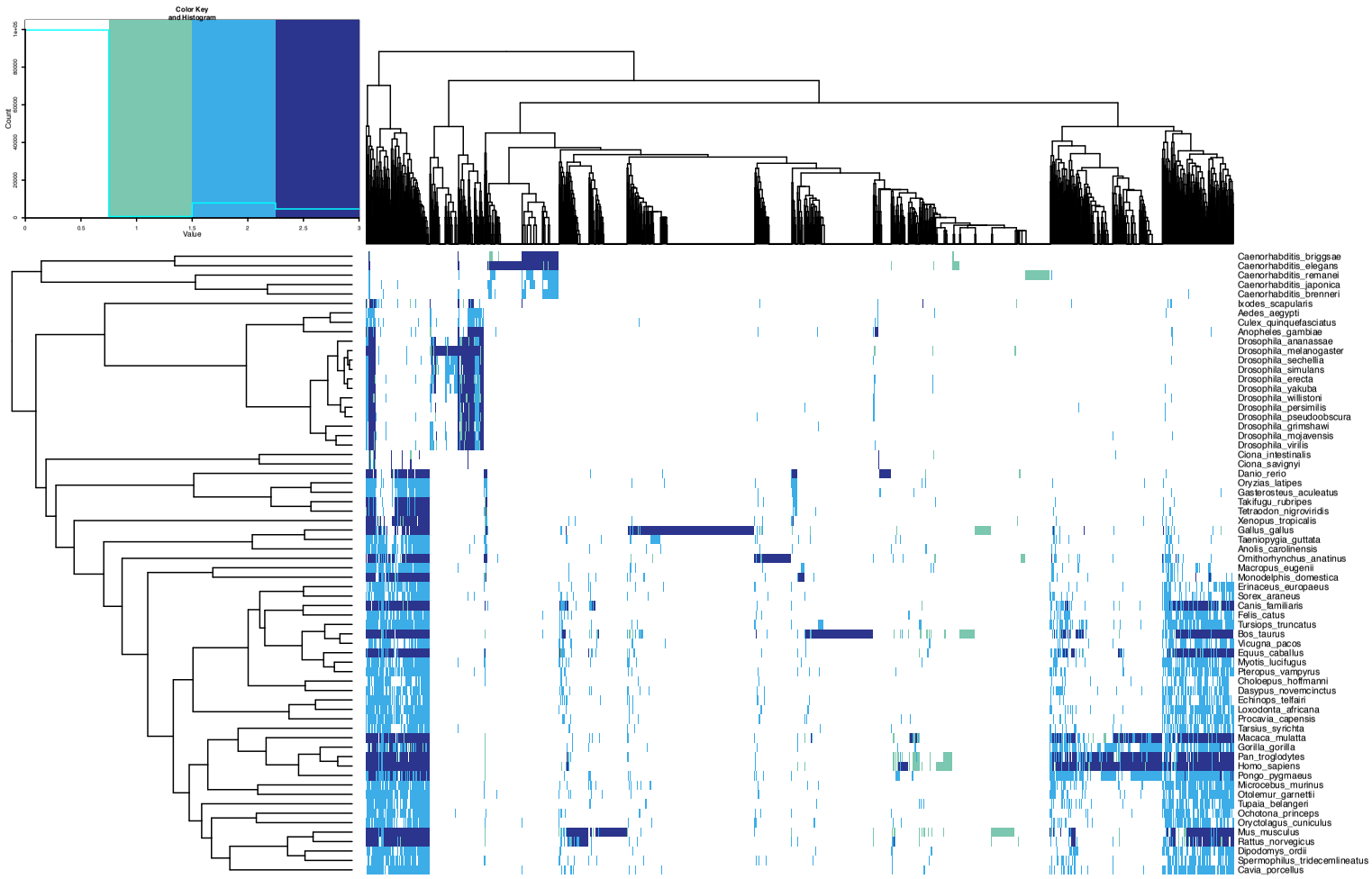


Figure 2.2: This heatmap shows an overview of Metazoan miRNAs, represented as a presence/absence matrix. It is colour coded to illustrate the effect of mapping using MapMi in the overall view of miRNAs in the species under analysis. Dark purple corresponds to an overlap between MapMi predictions and miRBase annotation. Blue indicates miRNAs that are only present in MapMi, while green indicates miRNAs that are on miRBase but are missing from the MapMi predictions. Bias towards model organisms is readily apparent in this view. It is also clear from the image that MapMi is complementing miRBase in a way that is broadly coherent with the expected evolution of miRNAs across the metazoan lineage. The different species are ordered respecting their phylogenetic relationships, as present in the NCBI taxonomy.

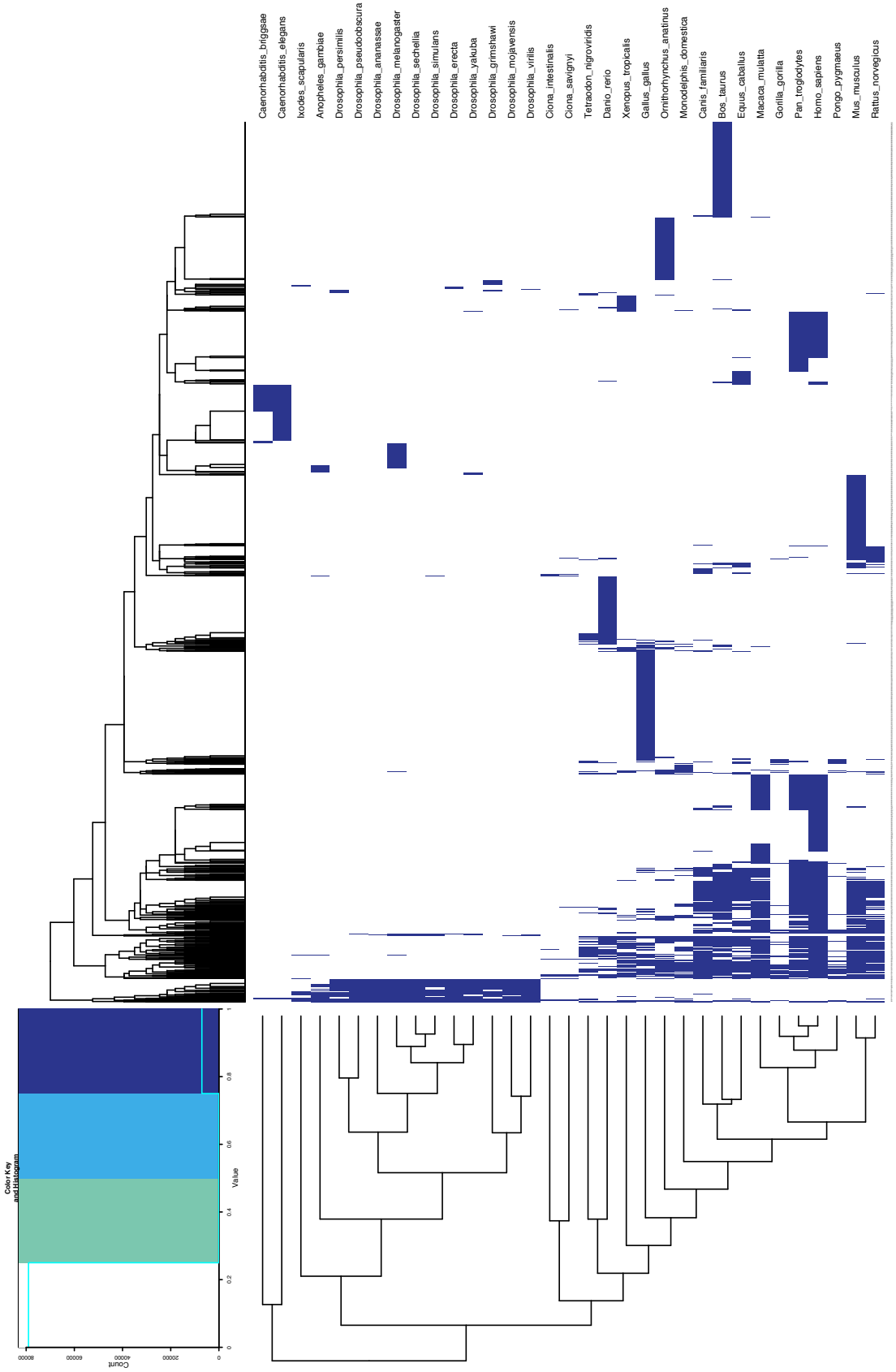


Figure 2.3: Heatmap containing information regarding miRNAs present in miRBase for the set of species under analysis. It was generated from a binary presence/absence matrix with the same parameters as Figure 2.2. It is easy to see that some miRNAs are missing in miRBase. This is particularly evident for *Pan troglodytes* and *Pongo pygmaeus*.

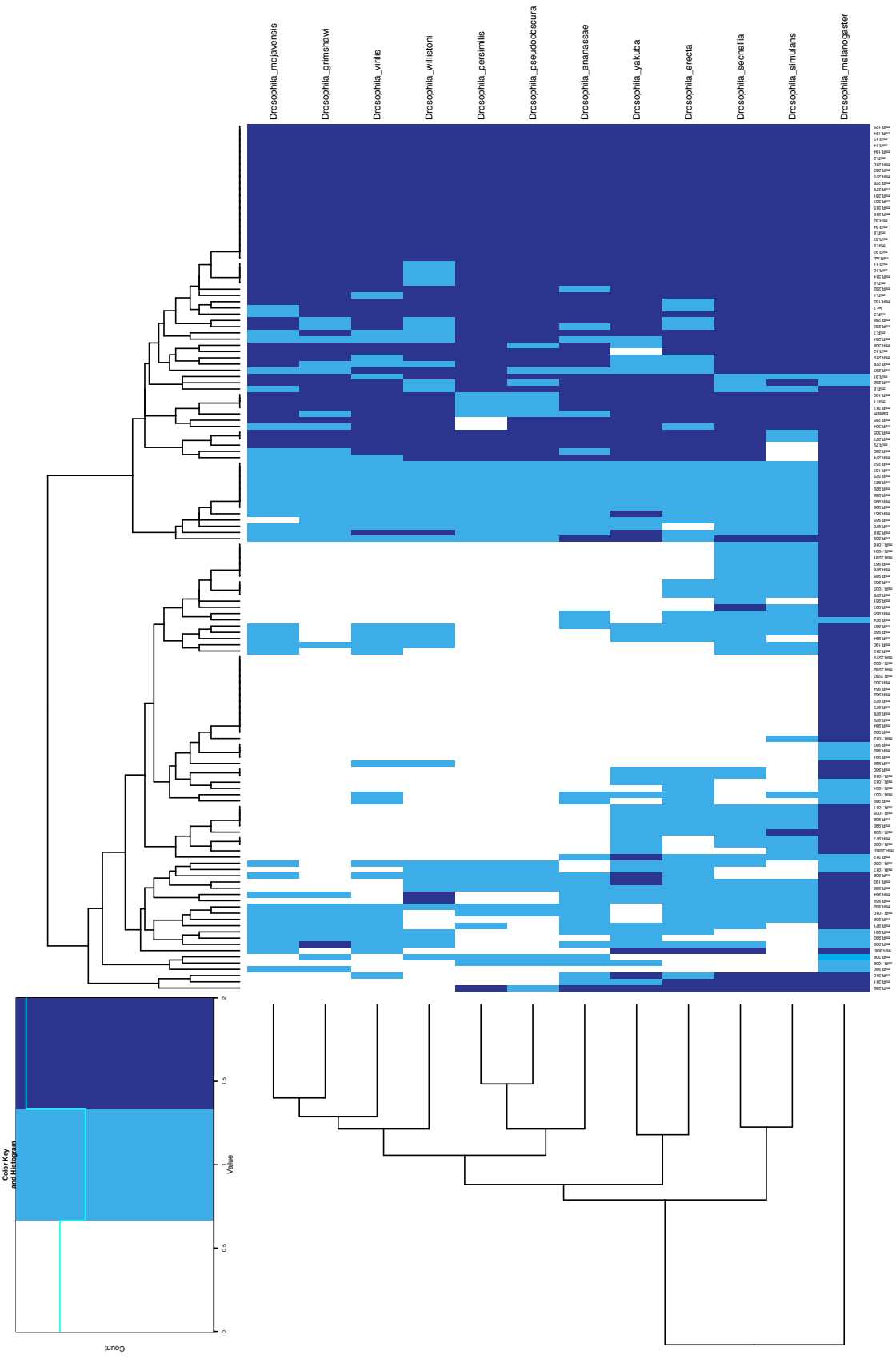


Figure 2.4: Heatmap of *drosophilid* miRNAs and hierarchical clustering of miRNAs present in the 12 drosophilid genomes, as predicted by MapMi using miRBase deposited *D. melanogaster* miRNA sequences as query. Dendrograms were produced by clustering of the data matrix. Dark blue indicates a miRNA present both in MapMi and in miRBase, light blue indicates a miRNA present in only one of the sets.

2.4.2 Validation Procedure

The performance of the scoring function (Equation 2.1) was evaluated by comparing score distributions from a positive dataset containing 4,237 miRBase (v 14) deposited unique sequences from Metazoan species, to a negative dataset composed of di-nucleotide shuffled versions of the sequences in the positive control (see Section 2.4.1). MiRBase deposited miRNAs have MapMi scores that are significantly higher than shuffled miRNAs (see Figures 2.5 and 2.6). This validation procedure was also used to derive thresholds for large-scale mapping projects in a way that balances sensitivity and specificity (Table 2.4).

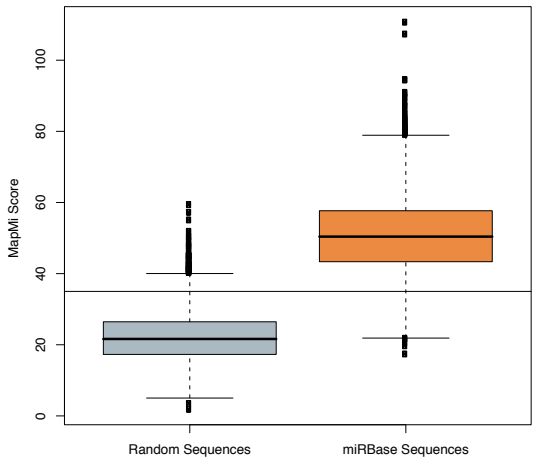


Figure 2.5: Boxplot illustrating the MapMi score distribution (y-axis) for 12 Drosophilids, queried with randomly di-nucleotide shuffled sequences (grey box) and miRBase deposited mature sequences (orange box). The horizontal line represents the default threshold (35).

To assess the performance of this pipeline when predicting miRNA orthologues in a different analysis scenario, MapMi predictions of horse miRNAs were analysed. Horse was chosen because it was recently introduced at the time of the analysis, in version 14 of the miRBase database. For this analysis, miRBase v13 deposited Metazoan miRNAs, that do not include any horse sequences, were used to predict miRNA loci in the horse genome, that are homologous to other previously known miRNA families. The overlap of MapMi predictions and miRBase v14 deposited

Threshold	Specificity: 10 Shuffles	Specificity: 100 Shuffles	Sensitivity
25	89.08%	88.04%	99.23%
26	90.68%	89.74%	98.97%
27	92.07%	91.22%	98.61%
28	93.28%	92.53%	98.23%
29	94.32%	93.65%	97.76%
30	95.22%	94.62%	97.30%
31	95.99%	95.46%	96.56%
32	96.64%	96.17%	95.66%
33	97.19%	96.78%	94.68%
34	97.66%	97.30%	93.61%
35	**98.05%**	**97.73%**	**92.20%**
36	98.38%	98.11%	90.67%
37	98.66%	98.42%	88.98%
38	98.89%	98.69%	86.98%
39	99.08%	98.92%	84.78%
40	99.24%	99.10%	82.64%
41	99.37%	99.25%	80.63%
42	99.48%	99.38%	78.56%
43	99.57%	99.49%	76.08%
44	99.64%	99.57%	72.64%
45	99.71%	99.65%	69.34%

Table 2.4: Summary of values of specificity and sensitivity of the method for each threshold. Default threshold (35) is in bold-face. The negative dataset was composed of random di-nucleotide shuffled versions of each of the 4,237 unique miRBase deposited metazoan miRNA sequences. The positive dataset consisted of miRBase deposited precursors.

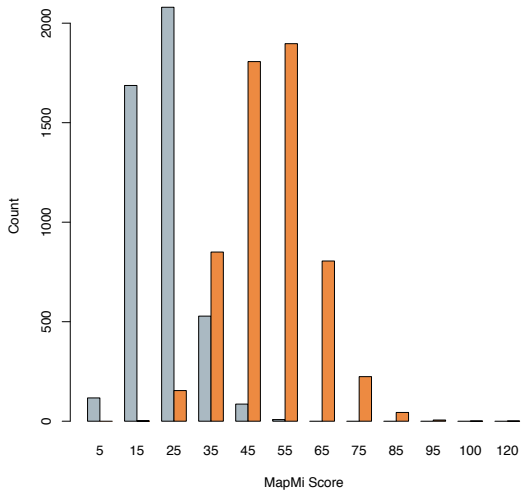


Figure 2.6: Histogram of MapMi scores (y-axis) for the same run shown in Figure 2.5. Grey bars correspond to the random di-nucleotide shuffled sequences, while the orange bars represent miRBase deposited miRNAs.

horse miRNAs was 82.99% (Table 2.5). This indicates not only that the method is sound in itself, but also that this approach works as intended, as the majority of known horse miRNA families are homologous to a previously known miRNA family.

The ability of the classifier function to distinguish miRNA hairpins from other genomic hairpins was verified by classifying a dataset containing 8,494 non-miRNA hairpins as reported in (Batuwita & Palade, 2009). MapMi obtained a correctly classified ratio of 93.14%.

Further verification was done for the genomes for which miRBase genomic coordinates are available, to assess how many MapMi predictions overlap with miRBase annotated miRNA loci and how many of those are correctly named. It was found that 87.04% of the predictions overlap with miRBase with 99.13% of those being assigned the same name as miRBase (Table 2.7).

Nine miRNAs appear to be highly conserved across the majority of species (Table 2.6). These miRNAs include the well-known *let-7* family, previously known to be highly conserved (Pasquinelli *et al.*, 2000). Conversely, a total of 636 miRNAs were shown to be species-specific, mostly in Chicken, *C. elegans*, Cow, Platypus, Human and Mouse. This result may arise due to some organisms being more heavily profiled. Additionally, some species have few related species available for comparison (e.g.

Query miRBase version	Allowed Mismatches	MapMi detected miRNAs	Percent Overlap with miRBase 14
13	0	271	78.09%
13	1	288	82.99%
13	2	291	83.86%
14	0	314	90.48%
14	1	314	90.48%

Table 2.5: The set of detected miRNA sequences in horse (*Equus caballus*) was recently introduced in miRBase in release 14. This table illustrates the predictive power of MapMi for finding horse miRNAs, and the importance of allowing mismatches to find orthologues when no annotation is present.

X. tropicalis) and as a result appear to have an excess of species-specific miRNAs. *Saccharomyces cerevisiae* is not believed to possess machinery for miRNA processing (Drinnenberg *et al.*, 2009), however it is present in Ensembl and was retained as a negative control. As expected, no miRNAs were found in *S. cerevisiae*.

Taken together, these results indicate that while miRBase has excellent coverage of many species, many others remain to be accurately profiled for miRNAs. Even though methodologies based on homology, like MapMi, cannot recover unknown species specific miRNA families, I believe that these results can complement miRBase in a useful way.

2.4.3 Comparison with Other Methods

Several methods are described in the literature with similar aims to MapMi. Many of these methods lack the openness and flexibility of MapMi regarding data sharing and availability of an implementation that can easily be run by other researchers in the field. I decided to compare the performance of MapMi against three of these methods, that have an available dataset. These three methods, CoGemiR (Maselli *et al.*, 2008), miRNAminer (Artzi *et al.*, 2008) and miROrtho (Gerlach *et al.*, 2009) were all designed with the aim to complement miRBase, filling in the gaps by using specialised homology searches.

	let-7	miR-1	miR-124	miR-125	miR-133	miR-219	miR-34	miR-7	miR-92
Present in miRBase only	3	3	1	0	1	4	2	3	0
Predicted by MapMi only	34	30	29	27	26	22	31	28	28
Present in MapMi and miRBase	28	28	34	29	32	29	29	29	32
Total times the miRNA is present	**65**	**61**	**64**	**56**	**59**	**55**	**62**	**60**	**60**

Table 2.6: This table presents the total number of miRNAs that are present in the majority of species from those under analysis (present at least in 55 out of 67). In here, the counts refer to the presence of at least one orthologue of the specific miRNA family in a species, not taking into account conservation of the number of paralogous loci in each species.

Species	Loci in miRBase	Overlapping Loci (1)	New Loci (1)	Overlapping Loci (2)	New Loci (2)
A. gambiae	67	59	27	59	12
B. taurus	626	517	1002	515	187
C. elegans	174	150	96	150	3
C. familiaris	325	310	251	309	89
C. intestinalis	25	21	5	21	1
C. savignyi	27	23	4	23	3
D. melanogaster	157	129	4	129	2
D. pseudoobscura	73	59	33	59	24
D. simulans	70	55	50	55	47
E. caballus	347	311	332	310	99
G. gallus	476	410	172	410	71
H. sapiens	750	620	874	619	138
M. mulatta	483	442	730	440	166
M. domestica	161	146	162	145	58
M. musculus	600	428	133	427	51
O. anatinus	348	289	238	289	58
P. troglodytes	604	514	751	512	149
R. norvegicus	320	297	152	297	60
T. rubripes	133	123	124	122	95
X. tropicalis	208	190	58	190	24
Total Loci in miRBase: 5974		5093 overlapping loci and 5232 new loci		5081 overlapping loci and 1365 new loci	
Correctly named:		5046		5035	
Overlap ratio:		(5093/5974): 85.25%		(5081/5974): 85.05%	
Correct Name Ratio:		(5046/5093): 99.07%		(5035/5081): 99.09%	

Table 2.7: Summary of the number of loci per species that overlap miRBase annotated loci, and the number of times the overlapping loci is correctly named by MapMi. This analysis could not be performed for all species, as miRBase loci coordinates were not readily available. Results are presented for two different parameter sets. (1) MapMi default parameters with no repeat element post-filtering. (2) MapMi allowing only perfect matches, post-filtering for sequences that are associated with repeat elements and map to multiple places in the genome (details of filtered sequences in Tables 2.1 and 2.3.2).

2.4.3.1 CoGemiR

It is challenging to directly compare MapMi with the CoGemiR database because loci location data is not readily available. While MySQL dumps are available for download, there is no documentation that allows the conversion of the database tables back to a simple loci location table. In the supplementary information for their manuscript it is possible to see a list of miRNAs they predicted. This list was compared to the list of miRNAs predicted by MapMi (regardless of loci location) and from the 188 predictions provided, MapMi only misses 6 miRNAs (96.8% of overlap, see also Table 2.8).

ete-mir-107	laf-mir-363	dno-mir-454
dno-mir-140	oan-mir-363	oan-mir-490

Table 2.8: List of miRNA families only found in "CoGemiR".

2.4.3.2 miRNAminer

The comparison with miRNAminer was performed using the latest version of the predictions available from their website. Nevertheless, these predictions are based on an old version of Ensembl (v48) and thus it is not expected that the genomic coordinates are coherent between genome assemblies. This likely accounts for a large proportion of loci that appear here to be miRNAminer specific. Since miRNAminer predictions exclude miRBase deposited miRNAs by design, it is not possible to properly compare miRBase overlap between these two methods.

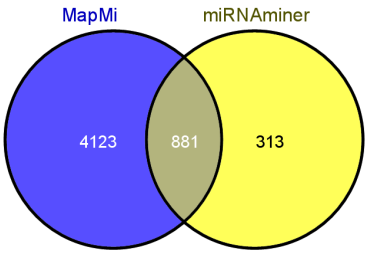


Figure 2.7: Venn diagram of loci overlap between MapMi and miRNAminer predicted miRNA loci. MapMi data is shown in blue, while miRNAminer is shown in yellow.

2.4.3.3 miROrtho

It is currently impossible to do a full direct comparison with miROrtho, as neither their method nor their full dataset is available for download. They provide the users with annotation tracks for the UCSC genome browser, but only for three of the species in their dataset. To perform the comparison, these files were downloaded and used to compute the genomic overlap between their predictions, MapMi predictions and miRBase deposited miRNAs (see Figure 2.8). Both methods agree that, even for these highly profiled species, there are still some miRNA loci belonging to known miRNA families that are missing from miRBase (v14).

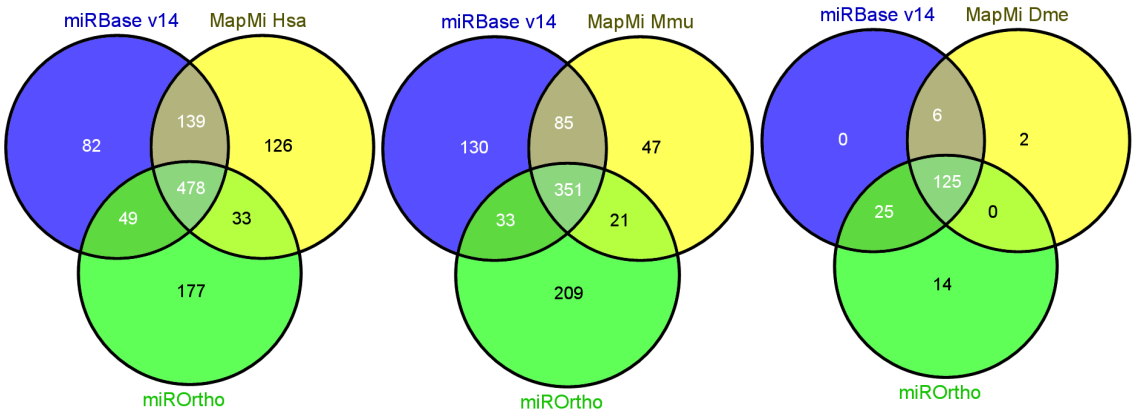


Figure 2.8: Venn diagrams summarising the loci overlap between MapMi and miROrtho predictions with miRBase v14 for the three species where miROrtho data is available for download. MapMi data is shown in yellow, miROrtho in green and miRBase in blue.

2.5 Predicting Novel microRNA Loci Using Small RNA Sequencing Data

With the recent advances in sequencing technology, it is now more affordable and widespread to use sequencing to profile miRNAs in a wide set of species and biological conditions. This enables not only the quantification of known miRNAs in more biological conditions, but can also be used to identify novel miRNAs that are expressed in particular circumstances that were previously unknown.

2.5.1 Existing Approaches

There are many precursor classifiers that aim to distinguish *bona fide* miRNA precursors from non-miRNA genomic hairpins (Batuwita & Palade, 2009; Jiang *et al.*, 2007; Lee & Kim, 2008). Others were specifically created to harness the power of the vast amounts of high-throughput sequencing data being generated, for novel miRNA discovery (Friedlander *et al.*, 2008; Friedländer *et al.*, 2012; Mathelier & Carbone, 2010). Nevertheless, these methods do not allow much control over selection criteria, and have implementation requirements that are sometimes difficult to meet. Furthermore, some methods are species specific and based on machine learning approaches depending on a training step. Whilst there are clear rules to define what an animal miRNA is, it is harder to unambiguously define a negative dataset of RNAs forming a stem-loop structure, for which one can be confident that they do not act as miRNAs. This may lead to model over-fitting and lower accuracy, when the classifier is used outside the training dataset.

2.5.2 The miRNouveau Approach

The MapMi pipeline has been expanded to meet the needs of researchers trying to identify novel miRNA loci in a species independent way. It builds on the knowledge that was acquired while developing MapMi, adding the necessary steps to ensure that the criteria for defining novel miRNAs are met. This approach is species independent, and works for any animal species where a genome has been sequenced. It is particularly suited for species that are present within the Ensembl resource (Flicek *et al.*, 2011a), as it is easy to access genomic sequences and annotations under a common framework, but can be used with other data sources. Since it does not require any training step, it is also not affected by over-fitting the particular data it was developed on.

The MapMi classifier was not developed for *de novo* miRNA discovery, and does not have enough information in its scoring function to properly assess all properties of miRNAs and thus detect novel miRNAs by itself. Instead, it relies on the fact that the candidates given as input are experimentally validated miRNA sequences. If such a sequence maps unambiguously to a different genome, and the locus has the general properties of a miRNA, then there is enough evidence it is a good locus. Moreover, most classifiers that were developed for *de novo* miRNA loci discovery, require a candidate hairpin to classify. Thus MapMi can be used a low-pass filter

in this situation, providing candidate hairpins that can then be scored and filtered by other more specialised loci classifiers. With this in mind, I searched for a loci classifier to complement MapMi.

2.5.3 Comparison of Classifiers for *de novo* microRNA Prediction

Over the years, a wide variety of methods were developed to distinguish between miRNA hairpins and other genomic hairpins. Most approaches score the structural properties of the RNA hairpin, nucleotide composition and structural stability. Some methods compute a probability based on a randomised trial, others employ more sophisticated machine learning methods such as Support Vector Machines (SVM), Hidden Markov Models (HMM) or Random Forests.

The goal of this comparison is to find an accurate, species independent classifier, that can complement MapMi and enable *de novo* miRNA loci finding. For this reason I restricted the comparison to three methods that were likely to produce good results, and had compatible underlying assumptions.

2.5.3.1 Randfold

The Randfold approach (Bonnet *et al.*, 2004) is based on the observation that miRNA hairpins consistently have lower minimum free energy (MFE) values than randomised hairpins, thus suggesting that real miRNA hairpins evolved to be stabler.

The algorithm consists of randomising the nucleotide sequence of the hairpin, computing the MFE, and using a modified Z-score test to assess if the candidates MFE increases significantly upon randomisation (i.e. the candidate hairpin is optimised to have a low MFE). Three randomisation algorithms can be chosen within Randfold. The results presented used di-nucleotide shuffling. The other randomisation methods produced results with lower or similar accuracy.

The MFE of each randomised sequence is used to compute a distribution of MFE scores for randomised sequences. The MFE of the original candidate sequence is then compared to this empirical distribution using a Z-score test to produce a p-value, that is then reported back to the user.

2.5.3.2 Self-containment

Another randomisation based method is based on Self-containment (Lee & Kim, 2008). It has been found that the structure of miRNA hairpins tends to be more stable to changes in contiguous sequence. The Self-containment method takes the candidate hairpin and inserts it into randomised genomic contexts with the same di-nucleotide composition of the candidate. It then assesses how many times did the candidate hairpin change its predicted secondary structure. It is expected that a miRNA will not change its structure due to its surrounding sequences as frequently as non-miRNA hairpins.

2.5.3.3 microPred

Even though our focus is on species independence, it is undeniable that there is significant biomedical potential for finding novel miRNAs in Human. To this end several methods have been developed with *Homo sapiens* in mind. One of the latest is microPred (Batuwita & Palade, 2009). This SVM based method relies on an array of 42 features, assessing both structure and sequence patterns within each candidate. Curiously, the authors claim in their manuscript that the method produced reasonable results when applied to other species. Unfortunately, the manuscript is too vague in this regard, namely, regarding which conditions and species it was tested.

2.5.3.4 Testing Procedures

To test the accuracy of these methods a set of five well profiled species was chosen (Human, Mouse, Rat, Fly and Worm). For each of the species, the complete genome was folded using RNALfold, part of the Vienna RNA package. RNALfold aims to find locally stable secondary structures across larger sequences. Each of the resulting hairpins was evaluated by the MapMi classifier, with the default threshold (35). The highest scoring non-overlapping hairpins were then filtered to avoid overlaps with known genomic annotations (all non-coding RNAs and protein-coding genes from Ensembl). After filtering, the dataset was still several orders of magnitude larger than the positive control dataset composed of miRBase deposited miRNAs for each of the species. This imbalance was resolved by randomly selecting hairpins from the dataset. Each method was then used to evaluate a negative dataset containing the unannotated hairpins and a positive dataset containing miRBase sequences.

2.5.3.5 Comparison Results

All the methods assessed require significant computational resources to perform large analyses. This is mainly due to the procedures used to determine hairpin stability, involving permutations of the sequence or adjacent sequences, and reassessment of secondary structure. On average[1], microPred takes 1 minute 30 seconds per candidate sequence, Randfold and Self-containment with 1000 randomisations take approximately 60 seconds, while Self-containment using 100 randomisations takes approximately 8 seconds per candidate.

Each of the methods produces different output: microPred reports a binary (-1/1) classification for each candidate; Randfold returns a p-value that was thresholded at 0.05 to provide a binary classification; Self-containment reports a score between 0 and 1, that was thresholded at 0.75, as suggested by the authors and in agreement with my own testing. These results were used to compute sensitivity and specificity for each species and in each test condition (see Table 2.5.3.5). As the positive and negative dataset are not exactly the same size, the geometric-mean of sensitivity and specificity was used as a proxy for method accuracy (Batuwita & Palade, 2009).

It was found that microPred has the lowest accuracy in the set, with acceptable sensitivity at the expense of a relatively low specificity. Interestingly, the results seem to be consistent between species, supporting the author's claim that the method can be used in a species-independent context. The specificity, lower than originally claimed, might be related to the machine-learning nature of the method. While SVM based methods are known to work well for classification problems involving large datasets, they are also prone to over-fitting if the datasets used for training do not contain enough information to fully train the classifier. To ameliorate this problem, there are ongoing efforts to establish better training datasets for miRNA loci prediction (Ritchie *et al.*, 2012).

Randfold performed better than microPred (see Table 2.5.3.5), albeit still biased towards sensitivity at the expense of specificity. It can be noted that this method is used by miRDeep (Friedlander *et al.*, 2008), to complement its scoring scheme for predicting novel miRNAs. The Self-containment method was assessed in two configurations, to assess the loss of accuracy that occurs when reducing the number of randomisations, in an attempt to reduce run-time. In its default configuration of

[1]These programs were run on a 64-bit Linux machine using a single 2.93GHz CPU

1000 randomisations, it achieves the best result of the methods compared, with a good balance between sensitivity and specificity. The reduction of the number of randomisations reduces the accuracy by approximately 5%. Although this loss affects the specificity of the method the most, this configuration is much less demanding computationally and still performs better than the other two methods under analysis. Overall, I found that Self-containment is, among these three methods, the best suited to complement MapMi for *de novo* miRNA genome-wide searches.

2.5.4 Criteria for Novel microRNA Identification: Revisited

Based on the data available in the literature concerning miRNA characteristics, a list of criteria has been set for the definition of novel miRNAs based on sequencing data. The pipeline is set up in a way that allows easy and fast determination of all these criteria in a user-friendly way, by automatically gathering relevant information from a wide variety of sources (Figure 2.9). This enables the researcher to make informed decisions taking into account the goal of the analysis and specific biological information about the project at hand. The pipeline also allows easy customisation (e.g. adding a different classifier). This system is more flexible than other approaches, allowing the user to make the decisions based on the project at hand, instead of leaving those decisions to a black-box automatic classifier. It is possible to argue that this approach is less objective and reproducible than a fully automated method, as it depends on the decisions made by a particular user. However, those decisions are made based on biological evidence that should be clear independently of who analyses the data and that conforms to the criteria set down to define animal miRNAs (Ambros, 2003).

Here I suggest a set of criteria that can be applied sequentially and are required for the identification of novel miRNA loci:

1. The read maps to the stem of a genomic hairpin (MapMi Pipeline).
2. The candidate hairpin is stable and robust to the surrounding genomic context (Self-containment Classifier).
3. An acceptable number of reads (depending on total read depth) map to the whole length of the hairpin (Automatic threshold based filtering).

	microPred			Randfold 1000 randomisations			Self-containment 100 randomisations			Self-containment 1000 randomisations		
	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
Human	52.05%	89.42%	68.22%	62.43%	85.69%	73.14%	69.17%	84.64%	74.67%	78.88%	80.61%	79.74%
Mouse	50.13%	89.58%	67.02%	61.33%	87.80%	73.38%	70.21%	83.78%	77.58%	79.63%	85.71%	82.62%
Rat	51.41%	93.38%	69.29%	62.28%	92.65%	75.96%	70.19%	87.25%	78.26%	79.97%	89.22%	84.47%
Fly	50.56%	90.34%	67.59%	57.89%	93.18%	73.45%	72.11%	89.77%	80.46%	81.83%	94.74%	88.05%
Worm	50.93%	95.43%	69.72%	55.58%	94.29%	72.39%	68.47%	90.29%	78.63%	77.13%	92.00%	84.24%

Table 2.9: Comparison of the accuracy between species-independent precursor classifiers for *de novo* miRNA prediction.

4. Mapped reads pile up in a way that is consistent with Drosha and Dicer processing (Visual analysis of the read pile-ups within the hairpin).

5. The locus does not overlap any other annotated genomic element, either protein coding or other classes of ncRNAs (Automatic filtering and visual filtering based on an Ensembl GFF file).

6. The locus does not overlap significantly with known intra-species variation in the mature sequence (Filtering based on Ensembl Variation data).

7. The locus is potentially conserved in other species (Based on multiple sequence alignments retrievable from Ensembl Compara).

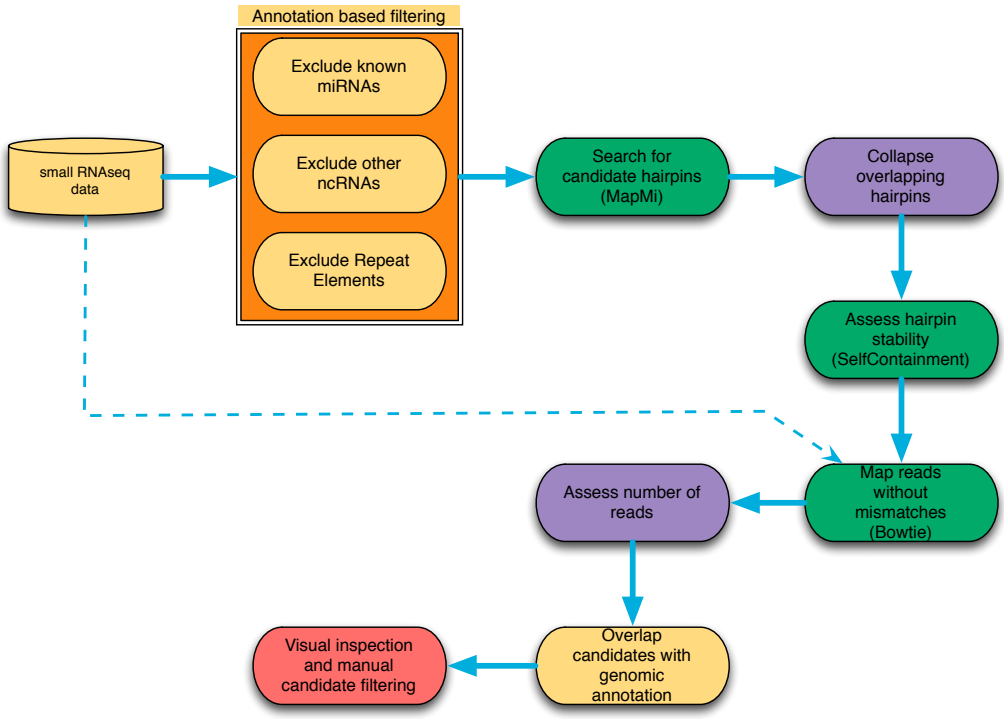


Figure 2.9: Workflow of the miRNouveau pipeline. The pipeline combines annotation based steps (salmon) and computational scoring of candidates (green) with automated dataset reduction (purple) and manual candidate assessment and selection (red).

2.6 Conclusion

In this chapter, I presented a computational pipeline designed to expand the information that is present in miRBase to a broader set of species that have had their genome sequenced, but have yet to be experimentally profiled for miRNAs, as well as expanding the known repertoire on some other less profiled species. I have also demonstrated the suitability of my miRNA mapping pipeline for the discovery of novel miRNAs based on small RNA sequencing data, when combined with different classifiers.

The field of miRNA research is relatively recent and great efforts are still underway to understand the functional importance of a large fraction of the miRNA repertoire. The number of miRNAs in miRBase is still increasing at a fast pace with each new release. It is essential for comparative genomics work to have a dataset that is as complete as possible, spanning a large range of species in a way that is as unbiased as current methods allow.

Many others have recognised this necessity and developed their own methods and approaches to enable their studies. Nevertheless, no usable implementation of such a method was found when this project started. To bridge this gap, I developed MapMi and made it available to the community under the GPL licence (Guerra-Assunção & Enright, 2010). Releasing the source code under an open-source license and making the methodology easily accessible through a web service will hopefully address the needs of a wide range of researchers, from the bench biologist that wants to quickly analyse a few sequences, to the pure bioinformatician that wants to adapt MapMi to their own needs and have a local installation for larger analyses.

The method was well received by the research community. At the time of writing, the MapMi manuscript (Guerra-Assunção & Enright, 2010) has 9 citations in peer-reviewed publications. The MapMi web server receives a monthly average of 40 custom runs as well as 150 accesses to the pre-computed MapMi dataset. The stand-alone version of MapMi was downloaded more than 500 times during the last 9 months.

For the method to be useful, it is also necessary that it produces good results. I demonstrate that this method has good sensitivity and specificity in a wide variety of datasets and research scenarios. I then used it to build a coherent dataset that enables the evolutionary studies presented in later chapters. This dataset is also easily accessible through the MapMi website, together with conservation and

phylogenetic analysis for each miRNAs family. In collaboration with other research groups, I also applied this method to search for previously unknown miRNA paralogues in the human genome (Hu *et al.*, 2012), and to the search for novel miRNA loci in human adipose tissue, based on sequencing data (Parts *et al.*, 2012).

As new species are being sequenced and small RNA sequencing data is being produced at an unprecedented pace, it is important to have tools that can easily take advantage of this new data. As it is based on the Ensembl genome browser and the miRBase database, MapMi can easily be updated to provide a dataset with the latest miRNA data and genome assemblies. This allows the research community to take advantage of the new data regularly being integrated in these resources.

Chapter 3

Evolutionary Analysis Based on microRNA Family Presence and Absence Across Evolutionary Time

3.1 Aim

This chapter deals with the evolution of the microRNA (miRNA) repertoire. Determining the miRNA families that are present in 80 metazoan species allows the analysis of miRNA evolution at an unprecedented scale. Besides looking at the turnover of miRNA families at each node of the species phylogenetic tree, these data also allow the detection of patterns of simultaneous appearance and loss of miRNA families and the detection of miRNA families that are expanding significantly faster in particular clades when compared to the general trend.

3.2 Introduction

Each species has its own miRNA repertoire, that is, a set of miRNA families that have at least one locus encoded in the genome. Having accurate information regarding the miRNA repertoire of each species, combined with existing phylogenetic information for the species under analysis enables a detailed analysis of the evolutionary history of each miRNA family, detection of associations between patterns or unexpected expansions of miRNA loci (Figure 3.1 and Guerra-Assunção & Enright (2012)).

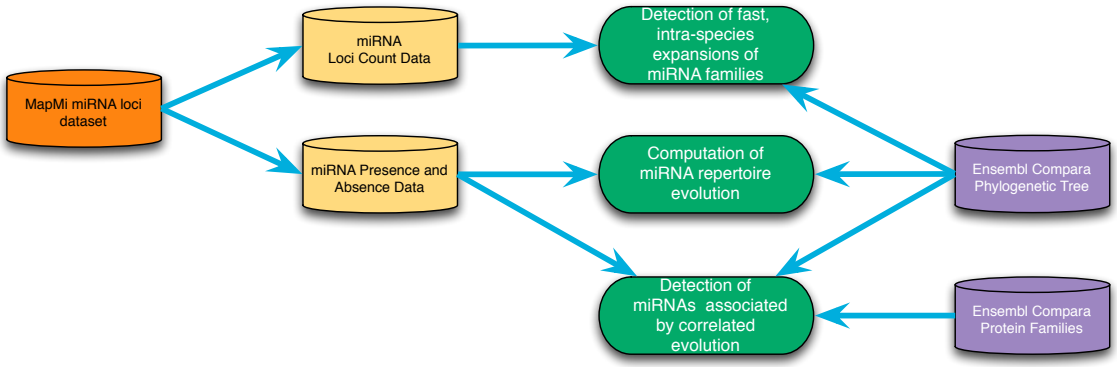


Figure 3.1: Flowchart of the analyses and datasets used within this chapter.

Application of Parsimony Approaches to Infer microRNA Family Gains and Losses

While many miRNAs are present in multiple species and are highly conserved, there are a growing number of miRNAs restricted to specific lineages. These data enable the exploration of the evolution of miRNA families across the metazoan lineage. The objective is to look at the history of each miRNA family and look for properties of miRNA families that depend on evolutionary age, while at the same time revisiting the concept of the correlation between the number of distinct miRNA families in a species and its morphological complexity (Heimberg *et al.*, 2008).

Furthermore, this view highlights biases present in the miRNA dataset. A lower than expected diversity within the miRNA repertoire is potentially caused by the quality of the genome assembly of some of the less studied species. Alternatively, a higher than expected number of miRNA families in some species is likely due to the comprehensive profiling of model species of biomedical importance. Whilst this problem poses a difficulty to large scale analysis in general, I sought to address this by selecting methods that are less affected by biases in the dataset.

Maximum Parsimony Methods

Maximum parsimony can be defined as a search for the minimum number of evolutionary changes that explain the patterns of evolution in a particular dataset. Many methods based on maximum parsimony have been described over the years. The different types of maximum parsimony methods were summarised by Felsenstein (Felsenstein, 1983). I describe three alternative approaches applicable to binary characters. In this context each character corresponds to a miRNA family, and the

two possible states represent the presence or absence of the miRNA family in a certain species.

The first discrete character parsimony method was described by Camin and Sokal (Camin & Sokal, 1965). In this model, it is assumed that the ancestral state was the absence of the character, and that characters could not revert from state 1 (miRNA presence) to state 0 (miRNA absence). The goal is to minimise the number of 0 to 1 transitions. This is not ideal for the analysis of this dataset, as convergent evolution of miRNAs is thought to be rare or non-existing (Wheeler *et al.*, 2009). Furthermore, the results from this set of rules would be significantly affected by assembly errors that could show as a spurious loss of miRNA families in certain species.

A more relaxed approach comes with Wagner parsimony (Eck & Dayhoff, 1966; Kluge & Farris, 1969). In this approach no assumptions are made regarding the state at the hypothetical common ancestor of all species under analysis (root node). The transitions between character states are weighted equally. While this method is a more generic approach that could yield good results, it is potentially affected by biases in the dataset. It is known that not all species in this dataset have been profiled to the same depth, or have high quality genome assemblies. Within the framework of Wagner parsimony, this could originate scenarios where a miRNA family would appear to have arisen more than once at different points of the phylogenetic three, which is thought to be highly unlikely (Wheeler *et al.*, 2009).

Farris presented an approach to maximum parsimony based on Dollo's Law (Dollo, 1893; Farris, 1977), where new gene families are restricted to arising only once throughout evolution, which is in concordance with what is thought to occur with miRNA families. Under these conditions, the most parsimonious scenario is the one that minimises the number of miRNA family losses. This seems ideal as it correctly accounts for the biases in the data, producing results that are unlikely to be affected by spurious gene losses.

Species that have potentially been over-sampled in regard to the rest of the tree will exhibit a large number of species specific miRNA families, without disrupting the overall results. At the same time, species that due to poor sampling and assembly issues have less miRNA families than would be expected would also not affect the general results obtained with this method.

For these reasons, Dollo parsimony was chosen for this study to detect instances of miRNA family gains and losses throughout evolution. The original objective of

Dollo parsimony is the inference of phylogenetic relationships based on the characters given as input. However, the same algorithm can be used, provided there is a good phylogenetic tree, to infer the evolutionary history of miRNA families (see Figure 3.2).

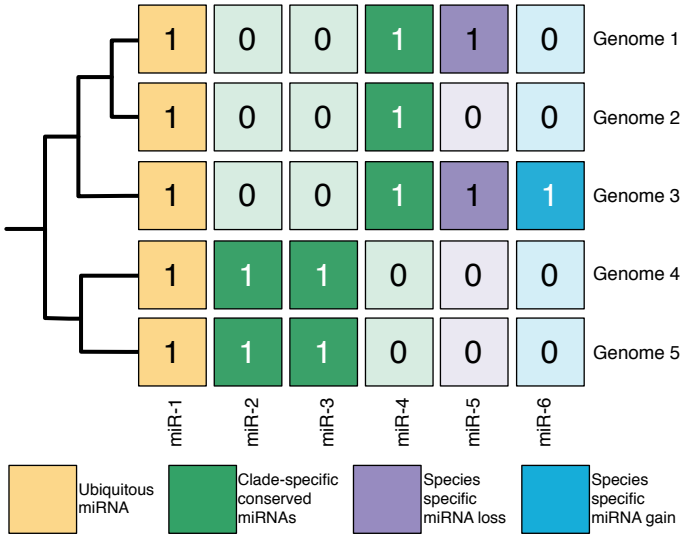


Figure 3.2: Illustration of Dollo parsimony applied to miRNA presence and absence data, to infer the evolutionary history of miRNA families. The phylogenetic tree represented is specified by the user. From the presence [1] or absence [0] of a particular miRNA family in each of the genomes, the evolutionary events represented in different colours can be inferred.

Exploring Correlated Evolution Between microRNA Families and Protein Coding Gene Families

Phylogenetic profile analysis was developed to tackle the annotation of protein-coding genes in bacterial genomes, based on the detection of correlated gains and losses of genes, between different species (Pellegrini *et al.*, 1999).

It is an assumption of this approach that genes gained and lost in a correlated manner, are likely to be involved in the same biological processes. This phenomenon is easy to understand in prokaryotes, as full biochemical pathways are frequently encoded as operons (Jacob *et al.*, 1960), that is, a set of genes encoding the different components of the pathway are transcribed from a single region of the genome under a common regulator.

It becomes clear in this scenario that if the two genes are present in the same pathway, they are likely to be found together in every species where the pathway is present. On the other hand, if it is not a selective advantage for the organism, the pathway will become inactive, with the consequent loss of the whole operon in a correlated fashion.

Over the years, different methods have been developed to extend this methodology to eukaryotic genomes and to explore the information in more complex evolutionary scenarios (Kensche *et al.*, 2008). These methods are globally known for having good specificity, albeit at the cost of having low sensitivity. Taking into account the difficulties faced by miRNA target prediction algorithms, it is an interesting to use this approach to infer potentially functional associations between miRNA genes and protein-coding genes within the dataset.

Detection of Rapidly Expanding microRNA Loci

The appearance of a new miRNA family is likely to be a major event in the evolution of a species, leading to a different regulatory regime for the genes it targets. In many cases, miRNAs act as fine-tuners and not as switch-like regulators (Bartel & Chen, 2004). There have also been reports of miRNA families that act as essential switches that mark the change between different regulatory regimes (e.g. dre-mir-430 in the maternal to zygotic transition in *Danio rerio*). To this effect, dre-mir-430 has different sub-families and is present in a high number of loci in the genome (Giraldez *et al.*, 2006). A similar pattern is also found in *Xenopus tropicalis* with respect to xtr-mir-427 (Lund *et al.*, 2009).

Expansion of the number of paralogues is likely to cause little disruption in the normal functioning of the cell, as it is just a question of dosage of a miRNA that is already present (Akita *et al.*, 2012; Mclysaght *et al.*, 2002). Furthermore, the increase of the number of paralogues, if expressed in a correlated fashion, will cause a stronger repression of the target genes, as the concentration of miR in the cell will be higher.

To detect unusually rapid expansions of other miRNA families and species, besides mir-430 and mir-427, I used the CAFE approach (De Bie *et al.*, 2006). This enables the detection of loci expansion in other cases, suggesting this is not a singular event or restricted to the maternal to zygotic transition. Some of these expansions have already been reported in the literature. Curiously, where a function has been described before, the miRNAs are often involved in development and pluripotency.

3.3 Results

3.3.1 Dataset Definition

Large-scale analysis of miRNA repertoire evolution depends on accurate information about the occurrence of miRNA loci across many species. I addressed this by expanding the miRBase loci annotation using the MapMi approach (see Chapter 2 and (Guerra-Assunção & Enright, 2010)). The expanded dataset contains 80 species (see Table 7.1). One factor hampering this analysis can arise from low-coverage genomes (Milinkovitch *et al.*, 2010; Vilella *et al.*, 2011) which makes mapping and identification of miRNAs difficult. Even though the methods used for the analyses described herein are robust to gene loss, I assessed all available genomes for completeness, specifying where results are likely due to a genome being low-coverage (see Table 7.1 in the Appendix).

This dataset is based on Ensembl (Flicek *et al.*, 2011b), Ensembl Metazoa (Kersey *et al.*, 2009) genomic sequences and protein family annotations (Ensembl Families). Annotations for miRNAs were obtained by mapping all metazoan sequences in miRBase (Griffiths-Jones *et al.*, 2008) using MapMi (see Chapter 2 and Section 3.5.1). The dataset contains 52 species containing both protein coding annotation and miRNA annotation, and 28 species where just miRNA annotation is present. This corresponds to 774,002 protein coding loci and 31,237 miRNA coding loci across all species under analysis.

Defining microRNA Families

Different miRNAs usually belong to the same family if they share the same seed sequence (i.e. nucleotides 2-8 of the mature miRNA (Lewis *et al.*, 2005)). It is believed that these miRNAs have similar targets and thus similar cellular functions although they may have very different spatial and temporal expression profiles. Given that many miRNAs are present in multiple related copies it is essential that I can accurately place them into families.

The miRBase database provides the grouping of miRBase deposited miRNAs into distinct families, based on seed matches and some hand curation. Nevertheless, the dataset is comprised of an expanded set of miRNA loci, making it difficult to use the miRBase provided grouping. Hence, I define 3,053 miRNA families based on all the miRNA loci under analysis (see Section 3.5.2).

3.3.2 Exploration of the Evolution of the microRNA Repertoire

While the miRNA repertoire is highly dynamic, significant changes in the miRNA family repertoire are thought to be rare throughout metazoan evolution (Heimberg *et al.*, 2008; Hertel *et al.*, 2006; Wheeler, 2008). The miRNA repertoire in each species can be simplified into presence or absence vectors, that can then be analysed using maximum parsimony approaches to determine their likely evolution.

This kind of analysis has been hampered in the past due to poor coverage of miRNAs in non-model organisms. I sought to address this by using an expanded dataset (see Section 3.5.1). As such, I explored when, in evolutionary time, miRNA families were generated and lost across a dataset of eighty species. By using Dollo parsimony analysis (see Section 3.5.3) I was able to infer the most likely ancestral nodes in the species phylogenetic tree where each miRNA family appeared (Figure 3.3). This analysis also highlights significant expansions in the number of miRNA families in certain branching points of the phylogenetic tree, as well as the species that are likely to have profiling biases (see Figure 3.3).

One drawback of this approach is that while I seek to detect miRNA orthologues across species, I cannot detect novel miRNAs present in species that have been poorly characterised at the miRNA level. This creates issues for analysis of gains and losses due to these sampling biases. Some species are well profiled for small RNAs, while for others there exists little or no validated data. However for those sets of species which are well profiled, such analyses provide useful information about the evolutionary dynamics of miRNA families, and benefit from the extra information obtained from the low-coverage species (Milinkovitch *et al.*, 2010; Vilella *et al.*, 2011). Furthermore, the rules of Dollo parsimony make it less likely to be affected by missing data within the dataset.

The results of this analysis are striking and show a large number of miRNA expansions across the phylogenetic tree (see Figure 3.3). As previously reported (Heimberg *et al.*, 2008), I observe a significant increase in miRNA number as morphological complexity increases with significant growth starting for metazoans and in particular across vertebrates (Heimberg *et al.*, 2008). The largest growth is observed for rodents and primates with a significant gain observed for great apes (see Figure 3.3). Globally the tree highlights sampling biases between clades. Some clades (e.g. Mammals) are well profiled while others (e.g. Insectivora, Bilateria) are poorly profiled. Individual species (e.g. *Tarsius syrichta*), although they are

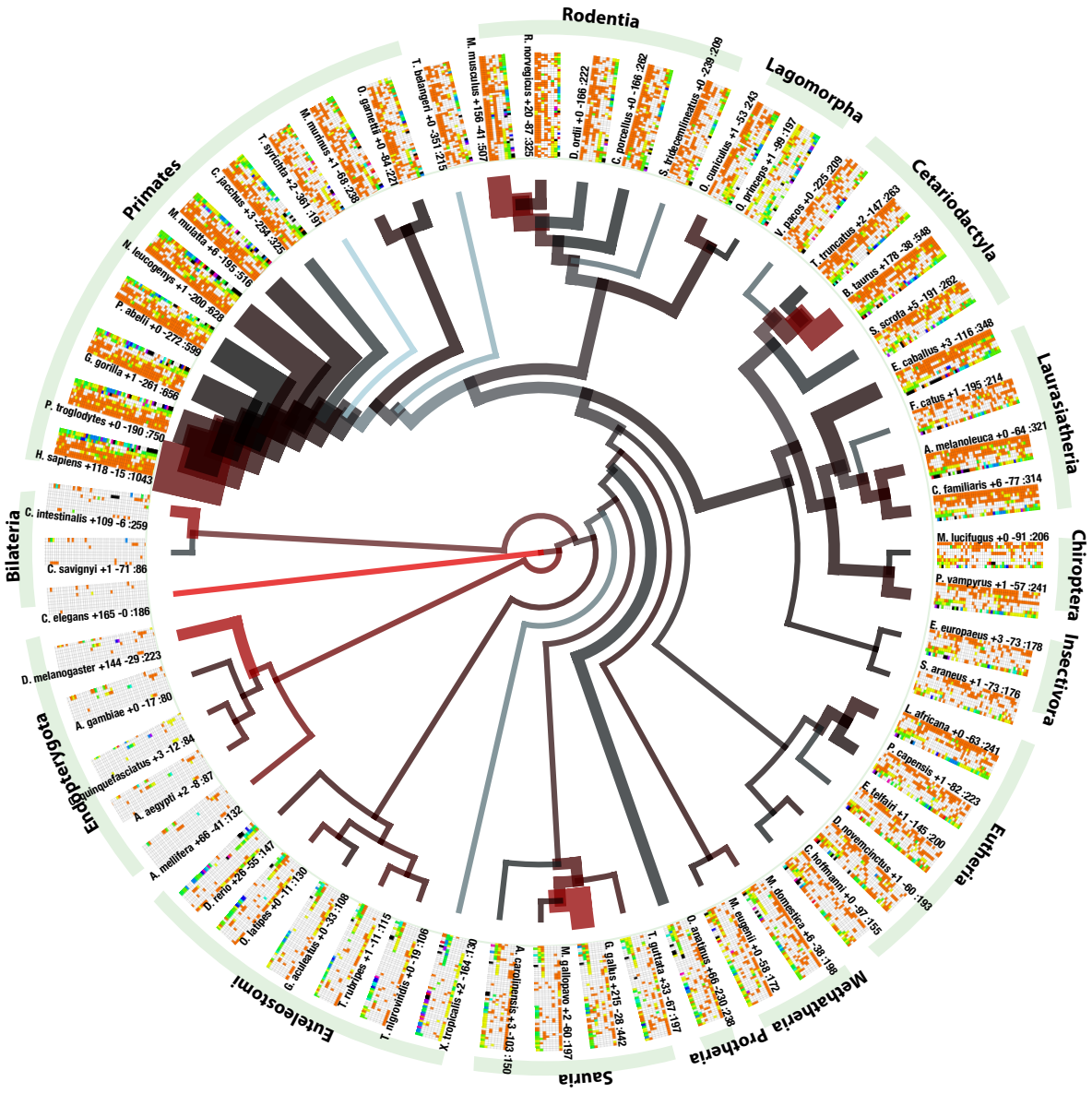


Figure 3.3: Phylogenetic tree with information regarding miRNA family gains and losses. Branch width represents the number of miRNA families present among leaves of the branch, while the colour represents significant miRNA family loss (blue) or gain (red). For each of 408 miRNA families present at multiple loci on at least two species, I also built a graphical "glyph" for each species. This glyph can be used to quickly assess presence, absence or expansion of families between clades. Each square represents a specific miRNA family. Squares are coloured as follows: white, indicates that this species does not contain a particular family, black indicates that this species contains at least 10 copies of miRNAs within that family. Copies between 1 and 10 are indicated as a rainbow gradient (red through violet). Groups of species are labelled according to the name of the evolutionary branch preceding them.

in a well-profiled clade, may have poor assemblies that hamper miRNA identification. Hence care must be taken in the interpretation of miRNA repertoire and the prediction of large gains and losses.

Additionally, I observe gains within Insects and Nematodes; this is particularly striking due to the absence of many species in these groups in the phylogenetic tree. A small number of clades exhibit significant losses, such as frog, marsupials, squirrel and hedgehog. Some of these perceived losses are most likely due to poor miRNA characterisation within these species that, possibly due to assembly problems, cannot be recovered by the MapMi pipeline.

3.3.3 Detection of Associations Based on Phylogenetic Profiles

A number of approaches have been successfully used to predict functional associations between protein-coding genes based on both their sequence and their genomic context (Dandekar *et al.*, 1998; Enright *et al.*, 1999; Kensche *et al.*, 2008; Marcotte *et al.*, 1999). In the context of protein-coding genes, these approaches have usually been applied to detect possible protein-protein interactions. I applied phylogenetic profile analysis to miRNA data for the first time, retrieving several meaningful associations between miRNA families and protein-coding genes.

An exploration of the most appropriate method to be used for miRNA application was performed. I attempted to use simple correlation metrics (Hamming distance and Pearson correlation), followed by clustering using the Markov Clustering algorithm (Van Dongen, 2000). However, it was found that this approach is heavily influenced by the species sampling biases within the dataset. This is probably due to the large evolutionary distances between the species under analysis, and the distinct oversampling of some of the clades (e.g. Mammals). Faced with this scenario, I focused on phylogeny aware approaches, settling for BayesTraits (Barker & Pagel, 2005) for its flexibility and suitability for the task at hand.

BayesTraits is a further development of the original BayesMultiState program from the same research group. Both approaches were designed for the comparison of pairs of phylogenetic profiles within a bayesian statistical framework. It attempts to fit, using maximum likelihood, two models to the data: one assuming the independence of the underlying gene families and one assuming correlated evolution. The two models are then assessed via a Likelihood-Ratio test. The pairwise comparisons

are automatically and sequentially called by the *BMS_runner* script developed for this purpose (Barker *et al.*, 2007).

Associations Between microRNAs and Protein-coding Genes

A small number of proteins appear to exhibit significant associations with distal miRNAs (>10kb) based on phylogenetic profile analysis (see Table 3.1). The associations detected are for three independent miRNA families (miR-876, miR-1251 and miR-1788). The associations for miR-876 are particularly interesting as all four protein-coding families involved appear to play a role in immune response. Two of the proteins, IL1A and CD86 have well established roles in immune response (Cytokine signalling and T-cell receptor signalling). The ASGR1 protein is involved in endocytosis of glycoproteins and is a target of the Hepatitis virus. MGL2 is a C-type lectin active in Macrophages. Finally MEFV is a protein producing Pyrin in white blood cells (eosinophils and monocytes) and plays a role in inflammation. Mutations in the MEFV gene cause the Mediterranean fever, an inflammatory disease (Karadag *et al.*, 2012).

While miR-876 associations appear to have strong connections to immune response, little is known about the expression or activity of miR-876. The only experimentally validated target so far for this miRNA in human is MCL1 (Induced myeloid leukaemia cell differentiation) (Hsu *et al.*, 2011), which is important for immune response. Predicted targets of this miRNA from both MicroCosm and TargetScan (Friedman *et al.*, 2009; Griffiths-Jones *et al.*, 2008) indicate a preference for receptor proteins.

Similarly, the miR-1251 family is poorly characterised but shows an interesting association with PRAME, a protein that normally is found exclusively in testis, but that is also highly expressed in melanoma. Finally, I detected a strong association between the fish specific miRNA miR-1788 and the TLCD2 protein family. Again in this instance little is known about the miRNA and the co-evolving protein. These associations represent interesting cases for further analysis both computationally and experimentally.

I also searched for significant phylogenetic associations between different miRNA families. Nevertheless, after filtering of associations found based on small numbers of species, I found no significant miRNA:miRNA associations.

miRNA Family	Protein Family Identifier	Protein Family Description	Likelihood Ratio
SF00154 miR-876	ENSFM00250000004087	IL1A interleukin 1 alpha	54.311
SF00154 miR-876	ENSFM00250000003359	CD86 antigen	54.311
SF00154 miR-876	ENSFM00440000236904	MGL2 Macrophage galactose N-acetyl-galactosamine specific lectin 2 ASGR1 Asialoglycoprotein receptor 1	49.285
SF00154 miR-876	ENSFM00500000270948	MEFV Mediterranean fever	49.285
SF01004 miR-1788	ENSFM00500000279147	TLCD2 TLC domain containing 2	49.614
SF01198 miR-1251	ENSFM00250000000393	PRAME Preferentially Expressed Antigen in Melanoma	53.283

Table 3.1: This table contains the most significant results from the phylogenetic profile analysis performed using the BayesTraits method. The table was sorted by miRNA family name (see Section 3.5.2). The protein families were obtained from Ensembl Compara.

3.3.4 Detection of Rapid microRNA Family Expansions

The CAFE (Computational analysis of gene family evolution) method (De Bie *et al.*, 2006) was developed to estimate the birth and death rate of gene families, inference of the number of gene loci in the internal nodes of the phylogenetic tree and more importantly, the detection of gene families and species with accelerated rates of gene gain and loss. With some interesting cases described in the literature, of miRNAs that have vastly expanded in specific species (e.g. dre-miR-430 (Giraldez *et al.*, 2006)), I aimed to obtain a global view of rapidly expanded miRNA families, followed by a literature search to determine the biological processes those families are involved in.

The CAFE approach implements a probabilistic model (first described in Hahn *et al.* (2005)). It models a random birth and death process, governed by parameter λ, that is estimated from the data by maximum likelihood. The model then uses it to estimate the most likely gene family size for all internal nodes, as well as the likely direction of change (e.g. gains or losses). Finally, when violations to the model are found, the program assesses each branch of the tree to determine where is the likely cause of the model violation occurring. For each family, the results are displayed in tabular and graphical forms (see Figure 3.4).

One of the assumptions of the CAFE algorithm is that gene families under analysis are present on the root of the phylogenetic tree provided as input. Due to the way miRNAs evolved, the majority of miRNA families are specific to a subset of species under analysis (see Figure 3.3). To address this issue, I limited the search to three distinct sub-trees of the main species phylogeny (see Table 3.4). These were chosen because they cover the majority of fully assembled and well profiled species, and correspond to major miRNA family expansion events, as highlighted by the parsimony analysis.

The CAFE algorithm (De Bie *et al.*, 2006) was used to detect rapidly expanding families within specific clades (see Section 3.5.3). In particular, I focused on three clades: primates (Table 3.2), fish and insects (Table 3.3). A large number of expansions were detected in primates, most significantly for embryonic stem (ES) cell expressed and repeat-associated miRNA families.

Two large families of miRNAs appear to have expanded rapidly in primates. The first large family (see Table 3.2) contains miR-130 and miR-301, miRNAs which have been previously reported as ancient miRNAs arising from tandem repeat duplications and which have been remodelled in animals (Hertel *et al.*, 2006). Members

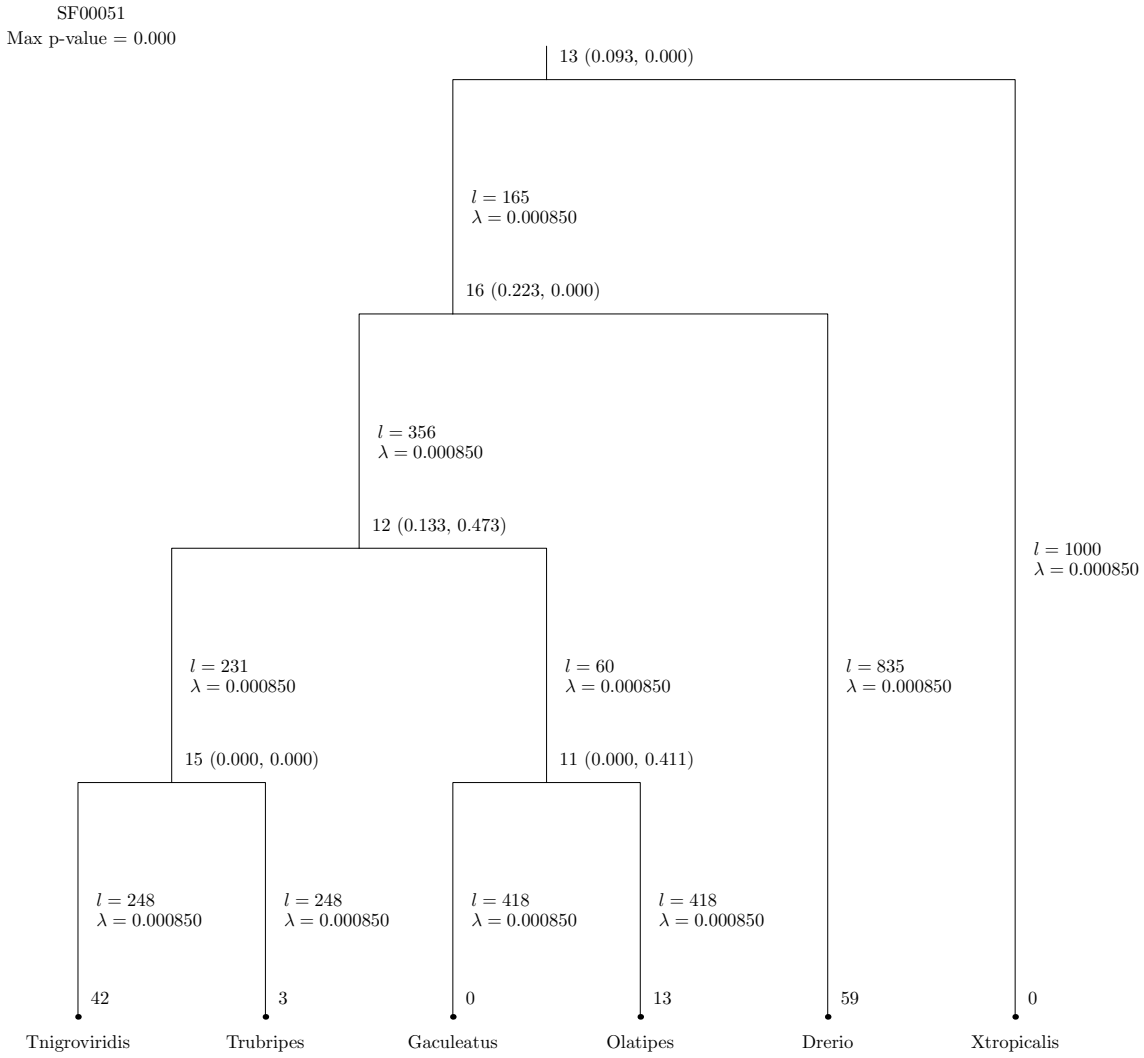


Figure 3.4: CAFE results for miR-430 family of miRNAs, and the detected fast expansion in fish species, in particular zebrafish. In each branch of the cladogram, the value of the scaled branch length (l) and the value of the birth and death parameter estimated by CAFE (λ) are shown. At each branching point, the estimated number of paralogues in the node and the p-values corresponding to each of the sub-branches are shown.

of this primate expanded family have been shown to have embryonic stem cell (ES cell) expression (Houbaviy *et al.*, 2003; Landgraf *et al.*, 2007). The second cluster is also linked to ES cell expression and contains members such as miR-290 - miR-294. Interestingly, not only is the miR-290-294 set of miRNAs expressed in ES cells, but it has also been postulated to be a putative maternal-zygotic switching mechanism in mouse oocytes (Tang *et al.*, 2007).

It is intriguing that such families of miRNAs involved in pluripotency and early embryonic development have expanded in primates. Interestingly these expansions mirror what is observed for other maternal-zygotic switches, described below for Insects and Fish. The increase of both morphological complexity and longevity in primates possibly requires increasingly complex control of gene-expression in stem cells. These results suggest that miRNAs are expanding in unison.

Aside from these two groups of ES cell related miRNAs I observe significant expansion of two large families of repeat-associated miRNAs. It has previously been shown that Alu elements were expanded in the ancestor of Old and New World monkeys and that this facilitated expansion of segmental duplications (Enard & Pääbo, 2004). Other studies have shown that such Alu expansion might also support frequent duplication of short units such as miRNAs (Zhang *et al.*, 2008).

The first cluster contains a number of miRNAs derived from simple repeats, (LINE and LTR elements), which have previously been shown to have expanded in primates, again likely through segmental duplication. The second family contains miRNAs likely derived from MADE1 elements (Piriyapongsa & Jordan, 2007), while the third family contains MER63 derived miRNAs (Yuan *et al.*, 2011).

These data support the hypothesis that many primate expanded miRNA families are derived from repetitive elements and may form genomic clusters through successive rounds of local duplication. The relevance and function of such miRNAs are difficult to establish. One possibility, that has been suggested before, is that such repeats may act as generators of novel miRNA sequences which have yet to find functional relevance.

Another interesting expansion involves a family of X chromosome miRNAs, including miR-465 and miR-509. A large number of expansions are also listed for miRNAs whose function and expression are not well characterised yet (Tables 3.2 and 3.3). A number of other expansions are observed for other miRNA families, however in many cases little is known about the family members involved.

For fish, amphibians and insects, few expansions are detected (see Table 3.3). However, two out of the four detected expansions involve miRNA families implicated in the maternal-zygotic transition, a process in early development that is regulated by miRNAs (Giraldez *et al.*, 2006). In particular mir-430 has been reported to have rapidly expanded in *Danio rerio*. I also detect a similar expansion in mir-427, an equivalent MZ-switch miRNA in *Xenopus tropicalis* (Lund *et al.*, 2009). An expansion is also detected for mir-2185 in *D. rerio*, however this miRNA has been poorly characterised with the limited expression information pointing to a possible role in heart development. For insects a single expansion is detected within *Aedes aegypti* for mir-2951, however this miRNA is also uncharacterised.

Table 3.2: List of genomic loci expansions as found by CAFE for the primate species under analysis (p-value < 0.01). The description for each miRNA family was obtained by manually assessing the literature (continued overleaf).

Family	Family Members	Description
SF00001	mir-1186, mir-1186b, mir-130, mir-1303, mir-130a, mir-130b, mir-130c, mir-1972, mir-301, mir-301a, mir-301b, mir-301c, mir-3090, mir-3590, mir-4452, mir-5095, mir-5096, mir-544, mir-544a, mir-544b, mir-619	ES Cell Expressed
SF00003	mir-1283, mir-1283a, mir-1283b, mir-290, mir-291a, mir-291b, mir-292, mir-293, mir-294, mir-371, mir-371b, mir-373, mir-512, mir-515, mir-516, mir-516a, mir-516b, mir-517, mir-517a, mir-517b, mir-517c, mir-518a, mir-518b, mir-518c, mir-518d, mir-518e, mir-518f, mir-519a, mir-519b, mir-519c, mir-519d, mir-519e, mir-519f, mir-520a, mir-520b, mir-520c, mir-520d, mir-520e, mir-520f, mir-520g, mir-520h, mir-521, mir-522, mir-523, mir-523a, mir-523b, mir-524, mir-525, mir-526a, mir-526b, mir-527	ES Cell Expressed (Maternal-Zygotic transition)
SF00022	mir-1254, mir-1268, mir-1273, mir-1273c, mir-1273d, mir-1273e, mir-1273f, mir-1273g, mir-1304, mir-297, mir-297a, mir-297b, mir-297c, mir-4419b, mir-4459, mir-4478, mir-466, mir-466a, mir-466b, mir-466c, mir-466d, mir-466e, mir-466f, mir-466g, mir-466h, mir-466i, mir-466j, mir-466k, mir-466l, mir-466m, mir-466n, mir-466o, mir-466p, mir-467a, mir-467b, mir-467c, mir-467d, mir-467e, mir-467g, mir-467h, mir-566, mir-669a, mir-669b, mir-669c, mir-669d, mir-669e, mir-669f, mir-669g, mir-669h, mir-669i, mir-669j, mir-669k, mir-669l, mir-669m, mir-669o, mir-669p	Repeat-associated miRNAs (simple repeats, SINE, LTR)

Continued on next page

SF00030	mir-2284a, mir-2284b, mir-2284c, mir-2284d, mir-2284e, mir-2284f, mir-2284g, mir-2284h, mir-2284i, mir-2284k, mir-2284l, mir-2284m, mir-2284n, mir-2284o, mir-2284p, mir-2284q, mir-2284r, mir-2284s, mir-2284t, mir-2284v, mir-2284w, mir-2284x, mir-2285a, mir-2285b, mir-2285c, mir-2285d, mir-2312, mir-2435, mir-548a, mir-548ab, mir-548ac, mir-548ad, mir-548ae, mir-548ag, mir-548ah, mir-548ai, mir-548aj, mir-548ak, mir-548al, mir-548am, mir-548an, mir-548b, mir-548c, mir-548d, mir-548e, mir-548f, mir-548g, mir-548h, mir-548i, mir-548j, mir-548k, mir-548l, mir-548m, mir-548n, mir-548o, mir-548p, mir-548q, mir-548t, mir-548u, mir-548v, mir-548w, mir-548x, mir-548y, mir-570, mir-603	Repeat-associated miRNAs (MADE1 Elements)
SF00037	mir-3586	Uncharacterised
SF00069	mir-1261, mir-1302, mir-1302b, mir-1302c, mir-1302d, mir-1302e	MER 53 derived, repeat-associated miRNAs
SF00090	mir-1587	Uncharacterised
SF00099	mir-3585, mir-463, mir-465, mir-465a, mir-465b, mir-465c, mir-470, mir-506, mir-507, mir-508, mir-509, mir-509a, mir-509b, mir-510, mir-513a, mir-513b, mir-513c, mir-514, mir-514b, mir-547, mir-652, mir-742, mir-743a, mir-743b, mir-871, mir-878, mir-880, mir-881, mir-883, mir-883a, mir-883b, mir-888, mir-890, mir-892, mir-892a, mir-892b	X-linked miRNA cluster
SF00160	mir-378b, mir-378d, mir-378f, mir-378g	Uncharacterised
SF00227	mir-4426	Uncharacterised
SF00280	mir-703	Uncharacterised
SF00332	mir-1233	Uncharacterised
SF00335	mir-4310	Uncharacterised
SF00379	mir-1244	Uncharacterised
SF00386	mir-4646	Uncharacterised
SF00447	mir-1236	Uncharacterised
SF00481	mir-1973, mir-4485	Uncharacterised
SF00485	mir-4640	Uncharacterised
SF00731	mir-3118	Uncharacterised
SF00807	mir-4509	Uncharacterised
SF00912	mir-663, mir-663a, mir-663b	Tumor Suppressor
SF00954	mir-3689a, mir-3689c, mir-3689d, mir-3689e, mir-3689f	Uncharacterised
SF01055	mir-877	Uncharacterised miRtron
SF01979	mir-3675	Uncharacterised
SF01987	mir-3180	Uncharacterised

Clade	Family	Family Members	Description
Amphibian	SF00050	mir-427	Maternal-Zygotic Switch
Fish	SF00051	mir-430a, mir-430b, mir-430c, mir-430i	Maternal-Zygotic Switch (see Figure 3.4)
Fish	SF01291	mir-2185	Uncharacterised
Insects	SF01286	mir-2951	Uncharacterised expansion in *Culex quinquefasciatus*

Table 3.3: miRNA family expansions in Amphibians, Fish and Insects.

3.4 Conclusion

This chapter explores the dynamics of miRNA repertoire evolution across the metazoan lineage. I have used Dollo Parsimony to perform a fully automated, large-scale analysis of metazoan miRNA families. My findings support the previous reports that the number of miRNA families seem to correlate with the morphological complexity of the organism. These results also identify the genomes whose low-coverage or poor genomic assembly makes them difficult to work with, and whose results should be interpreted more carefully. I have performed, for the first time, a large-scale automated phylogenetic profile analysis of miRNA and proteins, discovering a number of novel associations between miRNAs and protein coding genes with implications for the roles of miRNAs in immune response. Finally, I identified miRNA families that are fast expanding in certain clades. Within this analysis I found that the largest miRNA expansions detected frequently involve miRNA families connected with pluripotency and switching from maternal to zygotic gene expression in the early embryo.

Many of the results within this chapter would benefit from further functional validation. It would be advantageous to integrate a high-quality target sets to complement the results on the evolution of the miRNA loci themselves. Unfortunately, there are many reasons why this analysis is still challenging with currently available data.

There are still many unknowns in regard to miRNA target specificity, that hamper purely computational methods for miRNA target prediction. While this can be better addressed using experimental techniques for target prediction, there are still too few datasets available for this to be applied on a large-scale.

Furthermore, the search for miRNA targets is strongly dependent on the correct definition of the 3' UTRs of each potential target transcript. These are usu-

ally poorly defined in non-model organisms, making the determination of miRNA targets even more challenging. The low sequence coverage and unfinished assembly of the genomes of certain species and research biases towards model species, pose a challenge for computational genomics. Further sequencing and validation of miRNA families will be useful to remove erroneously predicted miRNA families and to mitigate biases. I believe that, as more data is being produced, these automated approaches will become more useful to handle the larger datasets, while at the same time producing better results, as more information is integrated in the analysis.

3.5 Materials and Methods

3.5.1 Dataset

I retrieved genomic sequences from all species in Ensembl (Flicek *et al.*, 2011b) (version 62) and Ensembl Metazoa (Kersey *et al.*, 2009) (version 9). I used MapMi (see Chapter 2 and (Guerra-Assunção & Enright, 2010)) (version 1.0.4) to map all the metazoan miRNAs in miRBase (Griffiths-Jones *et al.*, 2008; Kozomara & Griffiths-Jones, 2011) (release 17) against all genomes, using the default MapMi score threshold of 35. This dataset was merged with miRBase annotations, to retain the full miRNA annotation and increase sensitivity. The protein coding data was obtained using the Ensembl API to retrieve coordinates, ID and family information for all proteins. Proteins with no family information or with ambiguous family attribution were removed from the dataset to ensure coherence of the homology attributions across species.

Phylogenetic Tree

The phylogenetic trees shown are based on the tree provided by Ensembl[1]. This is a rooted, binary branching phylogram built from molecular data. All format conversions and node sorting necessary for compatibility with the programs used in this research were performed using the Mesquite framework for phylogenetic analysis (Maddison & Maddison, 2008).

[1] http://tinyurl.com/ensembltree

3.5.2 microRNA Family Attribution

Different miRNA families within the dataset can have different divergence within themselves. While the majority of miRNA families are present in a single loci within each genome, some others are highly divergent, containing several sub-families that can be present in several loci within each genome. It is advantageous for these analyses to group all related sequences in a coherent fashion, as a miRNA family can be represented by different sub-families in different clades.

For this purpose, miRNA loci were grouped, based on seed-weighed sequence homology. All miRNA loci sequences were compared using the Needleman-Wunsch algorithm for global-global alignment (Needleman & Wunsch, 1970), as implemented in ggsearch (FASTA package version 36.3.5a) (Pearson & Lipman, 1988). A scoring matrix that gives double weight to in-seed matching was used. This differentiation was performed by modifying the default ggsearch scoring matrix and using an expanded set of nucleotide codes to define the seed region within each loci sequence.

Families are then defined by single-linkage clustering of the bit scores of each pairwise alignment. Single-linkage clustering was chosen for its computational simplicity, and ease of interpretation of the results. The appropriate threshold was determined by minimising the split-join distance (Van Dongen, 2000) between my clustering and miRBase::Families. The full list of miRNA family attributions used within these analyses are available in the Appendix (Table 7.2).

3.5.3 Birth and Death of microRNA Families

There are several models to infer the most parsimonious evolutionary scenario (Felsenstein, 1983). The major difference between them concerns the assumptions of the model in regard to the relative birth and death rate for each gene family.

In the case of miRNA families, current data indicates a low probability of convergent evolution. Based on this, I have selected Dollo parsimony, an approach that allows each gene family to be gained once, with no restrictions on the number of times it suffers secondary loss. It is thus robust to losses due to genome assembly issues. I used this approach as implemented in the PHYLIP package (Felsenstein, 1993) (version 3.69).

Binary presence/absence data for each of the miRNA families were used allowing us to obtain an estimate of the evolutionary time of birth for each of the miRNA

	Primates	Fish	Insects
Species	*Homo sapiens*	*Tetraodon nigroviridis*	*Apis mellifera*
	Pongo pygmaeus	*Takifugu rubripes*	*Aedes aegypti*
	Pan troglodytes	*Gasterosteus aculeatus*	*Culex quinquefasciatus*
	Macaca mulatta	*Oryzias latipes*	*Anopheles gambiae*
	Callithrix jacchus	*Danio rerio*	*Drosophila melanogaster*
Outgroup	*Mus musculus*	*Xenopus tropicalis*	

Table 3.4: List of species present in each of the sub-trees used for the CAFE analysis.

families in the dataset. This was used to explore miRNA evolution from different perspectives (see Figures 3.3 and 4.3).

3.5.4 Association Analysis

I studied correlated miRNA gene gains and losses by using the BayesTraits package (Barker & Pagel, 2005) in a sequential fashion as implemented in the bms_runner script (Barker *et al.*, 2007) (version 1.4). This approach performs a Maximum Likelihood based analysis taking into account the phylogenetic distribution of the species under analysis, removing potential biases caused by uneven sampling of the phylogenetic space.

3.5.5 Rapid Loci Expansions and Deletions

While some miRNA families are present in a single copy in each genome, some families have rapidly expanded in some clades. To assess these fast expansions or unexpectedly fast deletions I used CAFE (De Bie *et al.*, 2006) (Version 2.2). This approach uses quantitative data for the number of elements of each family at each species, and requires that the gene families being studied are present at the root node of the provided phylogenetic tree. To accommodate this requirement, I performed this analysis in a selected set of sub-trees (see Table 3.4).

Chapter 4

Analysis of the Genomic Organisation and Evolution of microRNA Loci

4.1 Aim

In the previous chapters I described a pipeline for accurate mapping of microRNA (miRNA) loci across species, and explored the evolution of the miRNA repertoire.

In this chapter, I present a large-scale analysis of conserved synteny of miRNA loci. MiRNA loci are not randomly located throughout the genome. Several loci can be co-transcribed as part of the same primary miRNA, appearing in the genome as clusters of miRNA loci. This genomic organisation has been found to be conserved between species.

The conserved synteny blocks were analysed to detect patterns of miRNA cluster organisation throughout metazoan evolution, as well as highlighting specific cases of clade specific cluster expansion. The conservation of miRNA loci genomic organisation was also compared with conserved synteny blocks containing protein-coding genes.

4.2 Introduction

Synteny can be defined as the co-location of two or more genes along the same chromosome. Consequently, synteny conservation refers to the maintenance of the gene order across the genomes of two or more species (Sankoff *et al.*, 1997).

The conservation of gene order across species is a well researched topic in regard to protein-coding genes, therefore I attempted to revisit established approaches, methods and implementations and adapt them as necessary for the analysis of miRNA loci.

It is computationally hard to compute synteny breakpoints and conserved gene order simultaneously, which is needed for the reconstruction of ancestral synteny maps (Dasgupta *et al.*, 1998). Nevertheless, it is feasible to infer either the breakpoints or conserved gene order from a set of homologous anchors across species independently. There are many approaches that use genomic location information to infer genomic context and its evolutionary conservation. It can be meaningful to infer either the breakpoint history between species or the actual gene order conservation.

In this dataset, miRNA genes are spread non-randomly across the genome, and may not provide enough resolution for accurate breakpoint determination, even with the inclusion of protein-coding genes. To avoid this problem, I have chosen to focus on gene order conservation across species, combining miRNA loci and protein-coding genes. Each miRNA is potentially capable of regulating hundreds (or even thousands) of mRNA targets simultaneously (Lewis *et al.*, 2005). It is therefore important that their regulation be tightly controlled. Moreover, it has been postulated that intronic miRNAs may regulate the same biological pathway as their host genes. Several examples of this have been found, namely in the regulation of Myosin expression (van Rooij *et al.*, 2009) and cholesterol biosynthesis (Rayner *et al.*, 2011). This suggests that miRNAs that are consistently co-localised with proteins might be involved in the same biological processes.

4.2.1 Methods for the Identification of Conserved Syntenic Blocks

The problem of determining blocks of synteny conservation across species has been tackled in many different ways. The main challenge of this research was to determine which approach would be more suitable to automatically find conserved synteny blocks involving miRNAs, which have different genomic distribution when compared to protein-coding genes (Altuvia *et al.*, 2005). The first issue to be addressed was the choice between pairwise methods and methods that handle data from multiple species simultaneously.

Several methods and implementations were applied to the dataset presented here. The vast majority were unsuccessful in recovering conserved syntenic blocks for miRNA data when taking into account multiple species simultaneously. Pairwise comparisons are easier to compute, but more difficult to integrate in the context of a multi-species comparison, and showed errors when comparisons between evolutionarily distant species were compared. The use of well established pairwise synteny methods was thus not possible as the generalisation back to multiple species is not trivial.

A more advanced approach is implemented in Cyntenator (Rödelsperger & Dieterich, 2010), based on a *progressive alignment* approach. In a similar way to progressive sequence alignment methods, this method uses known phylogenetic relationships between species, adding one species at a time, starting with more related species and further traversing through the phylogenetic tree, comparing each new species to the conserved blocks detected in previous comparisons. While this approach was developed and demonstrated to work for protein coding genes, between a set of closely related species, it cannot be applied to miRNA data on a large-scale, as the miRNA repertoire is dynamic and there are very few miRNAs that are present in all species of the evolutionary tree.

Such an algorithm, when applied to an example case (illustrated in Figure 4.1), would start with the comparison between genomes 1 and 2, and immediately exclude miR-3 and miR-4 from further analysis, as their synteny does not appear to be conserved. Looking at the broad picture it possible to see that this was not the best approach. Progressive synteny methods can not resolve scenarios where a miRNA family is missing from a certain clade that shares a common ancestor with other clades that possess the miRNA in question within a conserved cluster. Since very few miRNA families are conserved in all species under analysis, this method does not appear to be applicable to this dataset.

To address these issues in a way that is compatible with this dataset, I sought different methods for inferring conserved synteny. I focused on the Enredo approach (Paten *et al.*, 2008), a method that handles all species in parallel and is not too demanding computationally. The simplicity of the algorithm enables easy parameterisation based on previous knowledge of the genomic organisation of miRNA loci. Even so, Enredo is designed to find syntenic blocks that span the maximum number of species, as opposed to finding the largest syntenic block, even if it is only present in a small number of species. To address this issue I designed a post-processing

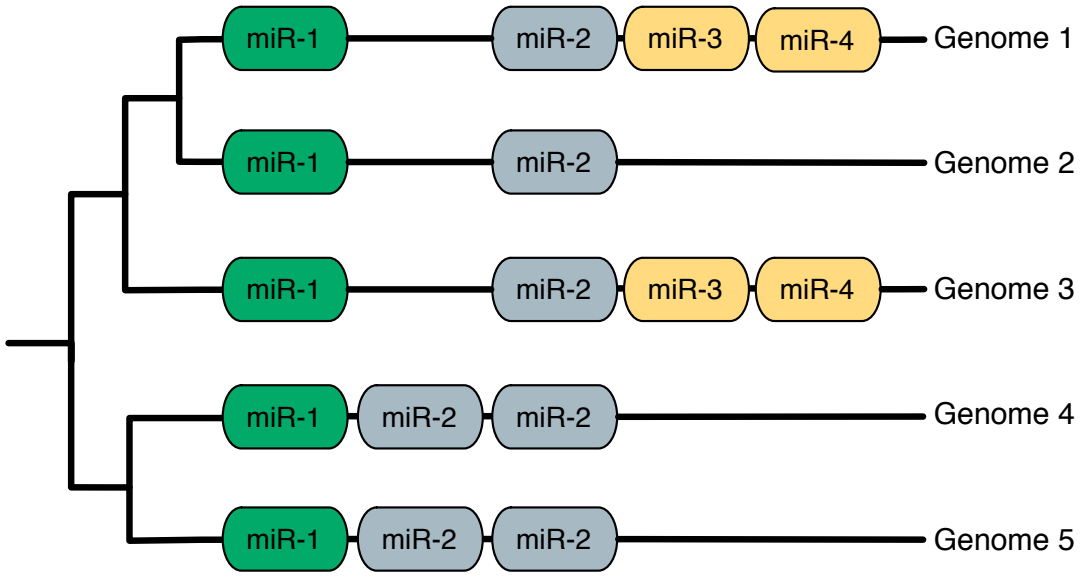


Figure 4.1: Common evolutionary scenarios for a widely conserved miRNA cluster. In this example, miR-1 and miR-2 are widely conserved. miR-2 shows a local duplication in genomes 4 and 5. MiRNAs 3 and 4 are clade specific, arising in the ancestor of genomes 1, 2 and 3, but were lost in genome 2.

step that extends the blocks found by Enredo, chaining them together. This also addresses some of the issues with unassembled genomes, by recovering large syntenic blocks that span more than one *contig*, as long as there are species where the block is intact.

Enredo will analyse these data in a different way than the progressive methods explained previously (see Figure 4.1). The first difference is that Enredo is not dependent on the phylogenetic relationship between the species, or a particular order for the genome analysis. Instead, it considers all the genomes under analysis simultaneously, building a directed graph with homologous anchors as vertices, and edges representing links or adjacency (see Figure 4.2).

The maximum distance for the link between adjacent anchors is specified as a parameter. Based on previous work in the field and the analyses within this thesis, it was defined as 10 kb. This threshold was shown to be the ideal compromise between capturing the majority of known miRNA clusters without including spurious clusters (Saini *et al.*, 2008). This also happens to be the smallest distance recommended for use with Enredo.

After the initial graph is built, Enredo proceeds in an iterative fashion to merge

the edges that are not conflicting and splitting edges that are in conflict, until a stable non-conflicting set is found. This corresponds to the set of co-linear segments that are present in the highest possible number of species. Each of these segments will then be reported as a conserved synteny block.

Due to the nature of the algorithm, some anchors may be present in several syntenic blocks (see Figure 4.2).

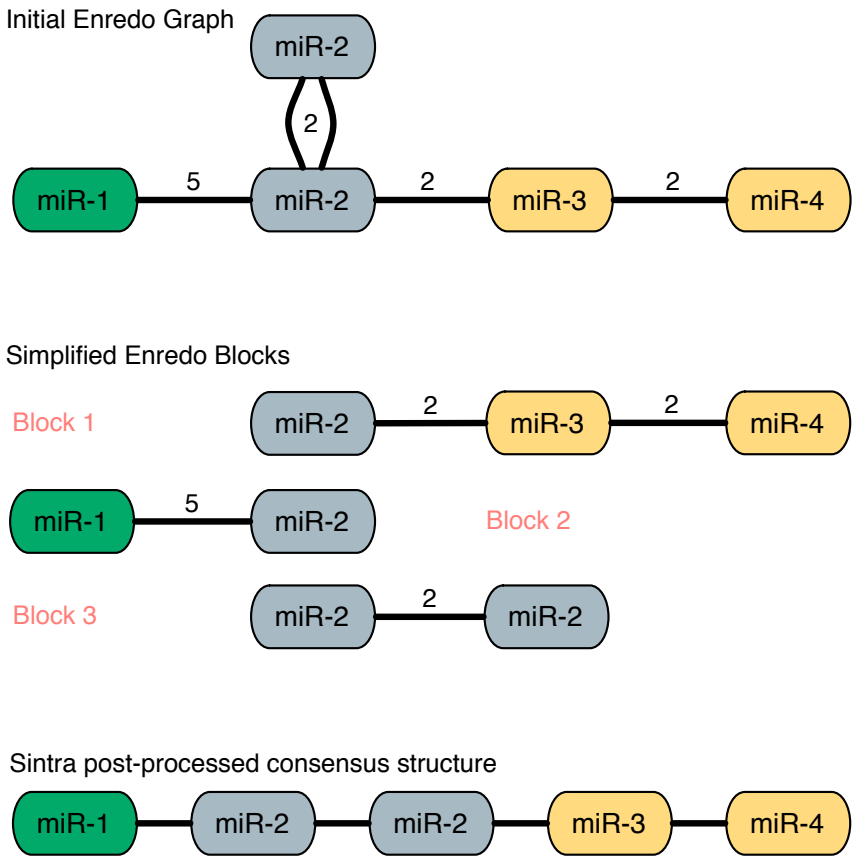


Figure 4.2: Simplified illustration of the Enredo algorithm and post-processing steps, based on the example in Figure 4.1. In a first step, a graph is generated where miRNAs are represented by edges linked by weighted vertices. The weights encode the number of regions where the miRNAs are found together. In a second step, the graph is simplified to remove cycles in the graph. In the final step, the linear blocks produced by Enredo are chained together to form a consensus structure that will then be used to visualise the synteny plots (e.g. Figure 4.1).

To illustrate the conserved synteny blocks, it is convenient to join these in a methodical fashion. For this, I compute a consensus string of elements in the block, and use the Needleman-Wunsch (Needleman & Wunsch, 1970) algorithm to align

the loci in each genome under analysis to this consensus structure (see Figure 4.2). This helps not only with visualisation of results, but also in the detection of synteny breakages across this dataset. Additionally, it does not affect the composition of the conserved synteny blocks or any of the metrics applied throughout this analysis.

4.2.2 Length Distribution of Conserved Syntenic Blocks

Before the availability of complete genome sequences, molecular markers were used to find comparable anchor points between genomes. These could then be used for genome organisation analysis. The first study that attempted to compare and quantify the synteny conservation between Human and Mouse was reported in 1984 (Nadeau & Taylor, 1984).

In this pivotal study, Nadeau and Taylor used the molecular marker data available at the time to propose a mathematical model to explain the patterns of genomic rearrangements between these two species. Their results supported the random breakage model of genome evolution, as proposed by Susumo Ohno (Susumo, 1973). The random breakage model postulates that the distribution of the length of conserved genomic segments follow an exponential distribution. This model was further validated and expanded by David Sankoff and colleagues (Ehrlich *et al.*, 1997; Nadeau & Sankoff, 1998; Sankoff *et al.*, 1997), and others (Bader *et al.*, 2001).

In the analysis presented in this chapter, the search for conserved synteny was extended to the detection of conserved synteny blocks across multiple species simultaneously. Nevertheless, the patterns found within this study (Figure 4.5) approximate an exponential distribution, in a similar way to what was originally described for Human and Mouse (Nadeau & Taylor, 1984).

To explore the contribution of miRNA genes to the patterns of synteny conservation, each conserved synteny block was classified depending on the coding-nature of its elements. Consequently, protein-coding blocks represent conserved synteny blocks containing only protein coding genes, labeled by their Ensembl Families ID. Mixed blocks contain miRNA loci and protein-coding genes, frequently corresponding to the maintenance of an intronic miRNA within a conserved protein. Finally, miRNA blocks contain conserved polycistronic miRNA clusters.

The genomes under analysis vary significantly in terms of their length, number of repeats and gene density. Block lengths are affected by these properties, acting as a confounding factor for biologically relevant results (Figure 4.4). To address

this issue, the data were normalised according to genome length, providing a more consistent view of block length distribution among species (Figure 4.5).

An interesting outlier is *Ciona savignyi*, whose normalised block lengths seem to be significantly longer than expected (see Figure 4.5). This can be explained by taking into account the fact that the genome assembly is still incomplete, and was probably generated by overlapping sequencing data to other species like *Ciona intestinalis*, which is then used as a scaffold. For this reason the available genomic sequences are extremely gene rich by comparison with other fully assembled organisms, and hence appear different in this analysis.

4.2.3 Conserved Synteny Analysis

Analysis of linkage and synteny is a useful tool for establishing both orthology relationships and functional linkages between genes. The application of synteny analysis to miRNA genes (both intronic and intergenic) has not been previously applied on a large scale.

To enable this analysis, a dataset was built based on a comprehensive, multi-species miRNA dataset derived from the MapMi pipeline for miRNA loci across species (see Chapter 2). This dataset was then combined with protein-coding gene information derived from Ensembl. The homology between proteins was determined by the Ensembl Compara pipeline (Flicek *et al.*, 2011b), with the coherent naming between species assured by using Ensembl Families.

I then set out to explore the question of whether synteny blocks containing miRNAs exhibit differences compared to those blocks that contain solely protein-coding genes. Moreover, I assessed whether particular species illustrated unexpected arrangements for miRNA genes when compared to other species (Guerra-Assunção & Enright, 2012).

4.3 Results

In this dataset, many miRNAs (48%) are encoded as independent non-coding transcripts while the rest (52%) are encoded within the introns of protein-coding genes. Some miRNAs exist as individual molecules encoded by a single locus while others occur in transcripts encoding multiple copies of the same miRNA or multiple transcripts at different genomic loci (Olena & Patton, 2009). It has been postulated

that in some cases multiple loci are required to increase the copy-number of specific miRNA molecules in certain circumstances (e.g. miR-430 in early development of the Zebrafish embryo (Giraldez *et al.*, 2006)). The expression of a large number of similar miRNA loci simultaneously causes a rapid increase of miRNA level in the cell, resulting in a switch-like repression mechanism (Bartel & Chen, 2004).

4.3.1 Implementation Notes

The automated detection of conserved genomic segments within the dataset were performed using the Enredo algorithm (Paten *et al.*, 2008). For this algorithm, each genomic element with a defined position within the genome is considered as an "anchor". One of the requirements of this algorithm is that no anchors overlap. This is particularly important due to the prevalence of intronic miRNAs. Moreover, the algorithm allows anchors to be part of multiple syntenic blocks. This is important as it allows for a certain anchor to exhibit specific syntenic patterns depending on the clade. To take these characteristics of the algorithm into account, anchors were reduced to 2bp, in the centre of the original coordinates, before running Enredo. This fully resolves overlaps, as only two elements with exactly the same coordinates, or exactly in the centre of the larger element will have the same 2bp coordinates. Furthermore, having a fixed length allows straightforward chaining of different conserved synteny blocks whose last element has a 2bp shift in coordinates from the first element of a different block. This was used to build the final syntenic blocks that are plotted and analysed.

4.3.2 Evolutionary Comparison of miRNA Genomic Context

One of the advantages of having a detailed and coherent dataset across species for evolutionary analysis, is that the information can be combined in different ways to look at the data from different perspectives.

It is believed that more ancient miRNAs, appearing early in metazoan development, tend to be more conserved and be more highly expressed than recent, species specific miRNAs (Ason *et al.*, 2006). I wanted to explore if there was any trend concerning miRNA age and the time of the largest expansions of diversity in the metazoan miRNA repertoire and genomic context. To achieve this, the information presented in the previous chapter (see Section 3.3.2), regarding repertoire evolution

using Dollo parsimony (see Section 3.5.3), was combined with the analysis of conserved synteny and genomic context. The phylogenetic distance (branch length) between the ancestral node of the phylogenetic tree and other nodes was taken as a proxy for node age.

In agreement with previous reports (Hertel *et al.*, 2006), I observe major miRNA expansions in the bilaterian and vertebrate splits (Figure 4.3a). While it is difficult to infer clearcut conclusions from the results, there is a tendency for more recent miRNA families to be intronic rather than intergenic (Figure 4.3b). There is also a tendency for ancestral miRNA families to be found in polycistronic clusters more often than more recent families (Figure 4.3c).

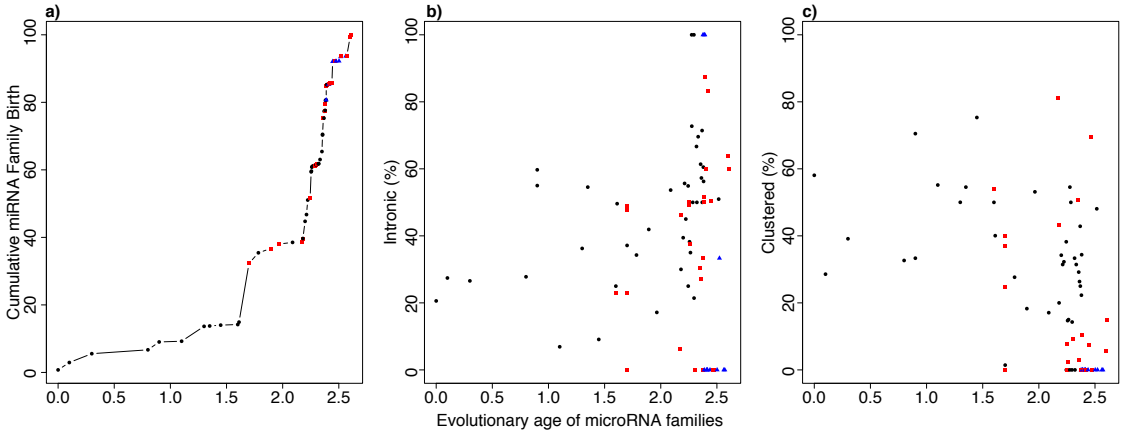


Figure 4.3: Evolution of the genomic organisation of miRNA families. The evolutionary age is represented by the distance between each node of the phylogenetic tree in Figure 3.3 and the root node of the tree (x-axis). Internal nodes are represented by black dots. Terminal nodes corresponding to high-coverage genomes are represented by red squares, while low-coverage genomes are represented by blue triangles. The panels represent on their y-axis: a) Cumulative number of miRNA families appearing at each node. b) Percentage of appearing miRNA families that are intronic per node. c) Percentage of appearing miRNA families that are part of miRNA clusters per node.

4.3.3 Length Distribution of Conserved Syntenic Blocks Containing microRNAs

In a similar fashion to previously published work (Nadeau & Taylor, 1984), I computed block-length distributions (Figure 4.5) in all genomes for three distinct classes of synteny blocks (I) Protein-coding only blocks (II) Mixed blocks (encoding both miRNA and protein coding genes) and (III) miRNA only blocks.

For protein-coding blocks, the data approximates an exponential distribution of conserved block lengths, as previously described by Nadeau and Taylor (Nadeau & Taylor, 1984). This is characterised by a high fraction of short conserved segments, with extremely long blocks being found rarely. Blocks that encode only miRNAs have a different distribution where long blocks occur at a higher frequency, giving a bimodal distribution where both short and long blocks are favoured. Mixed blocks predominantly follow the observed patterns seen for protein-coding only blocks but again have more long blocks than expected.

Genome compaction among fish is readily observable (Figure 4.4) for both protein-coding and mixed blocks, hence I normalise for total genome length (Figure 4.5 and Materials and Methods). For mixed blocks the only outlier is *Ciona savignyi*, which exhibits longer than expected blocks, however this may in fact be due to poor genome assembly. Interestingly, for miRNA-only blocks, most species exhibit similar block length distributions, except for *C. elegans*, *C. intestinalis*, *C. savignyi*, *D. melanogaster* and *D. rerio*, *T. rubripes* and *O. latipes*. These species have the smallest genomes in the dataset yet would seem to have longer miRNA encoding blocks than expected.

This finding suggests that miRNA encoded blocks may not have been subject to genome compaction and appear to be relatively stable in terms of length across species and independent of genome length. One possibility is that miRNA syntenic blocks are already at a maximal compaction state and hence do not appear to be affected by genome compaction.

4.3.4 Conserved Synteny Blocks Among microRNA Clusters

The majority (59%) of the miRNA loci in this dataset are found to be encoded on the genome by transcripts containing several miRNA loci. It was also found that a large part of the miRNA clusters analysed (63%) are found in conserved synteny

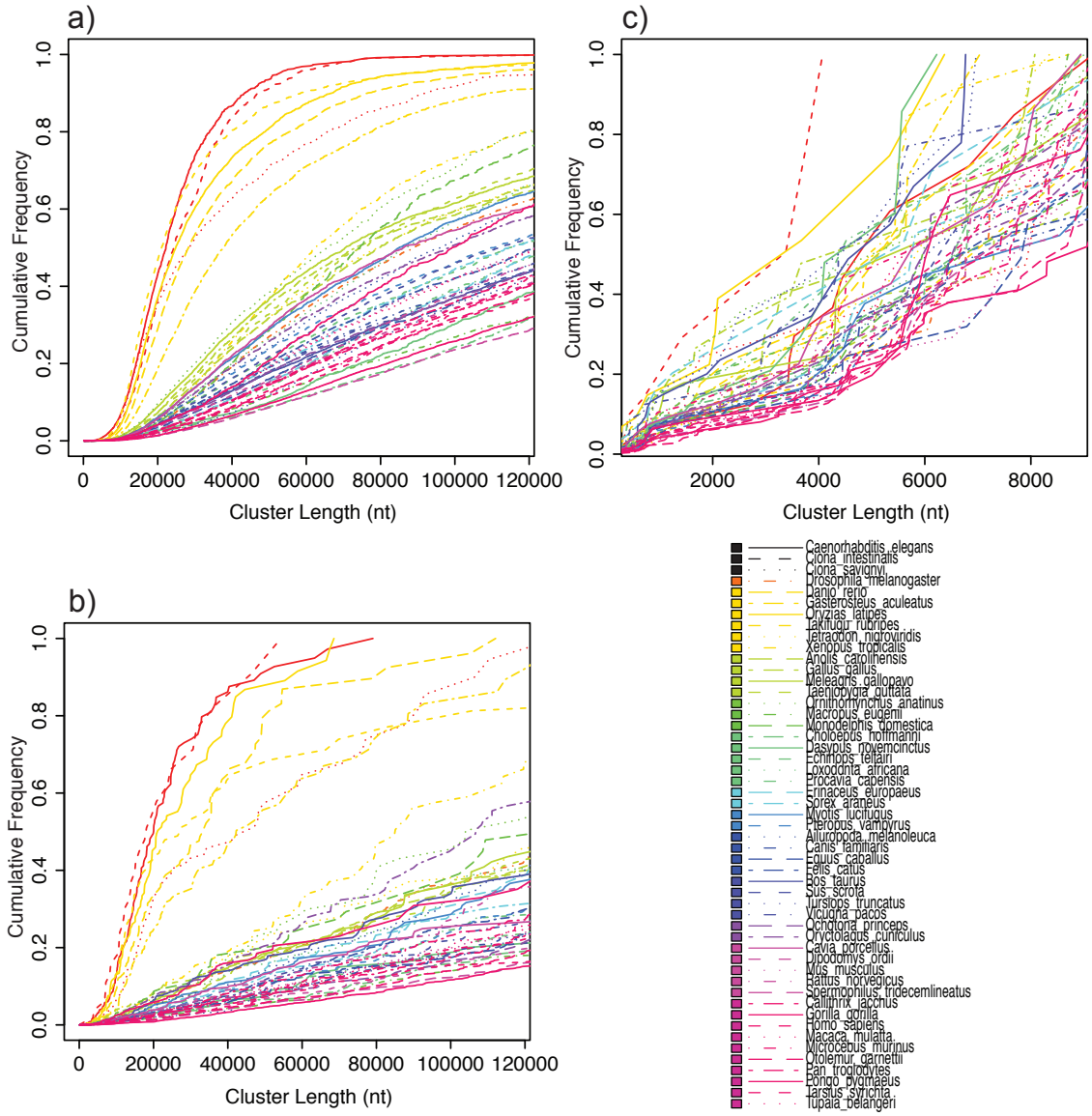


Figure 4.4: Non-normalised genomic cluster length per species. The x-axis represents the number of nucleotides spanned by each genomic cluster. The y-axis represents the cumulative frequency of cluster length in each species. Species are coloured according to the grouping shown in Figure 3.3. Different classes of genomic clusters are shown in each panel: a) clusters that contain only protein-coding genes; b) mixed clusters containing both protein-coding and miRNA genes; c) clusters containing only miRNA loci.

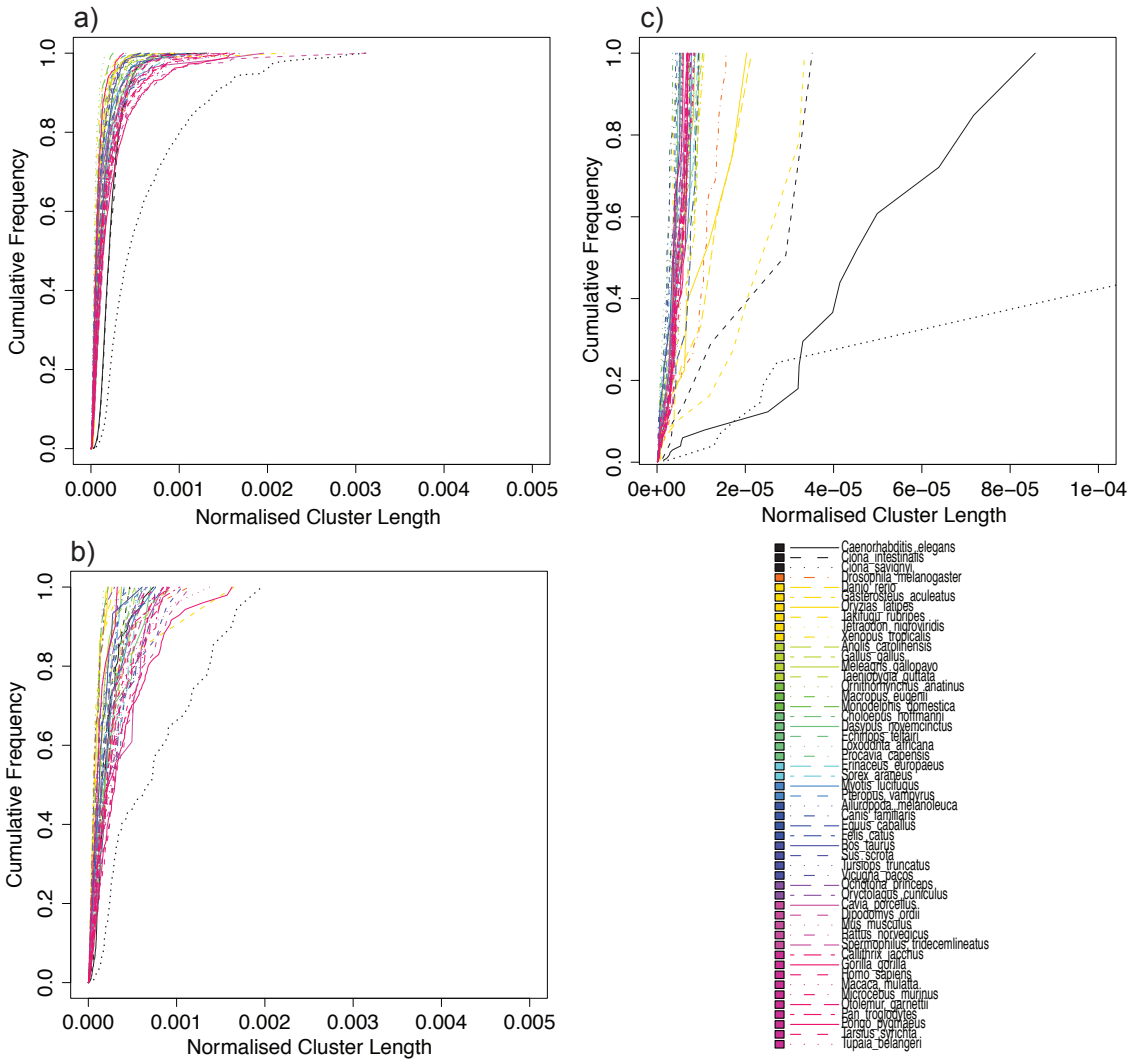


Figure 4.5: Normalised genomic cluster length per species. In these plots, the cluster length (x-axis) was normalised (cluster length/total genome length). The y-axis represents the cumulative frequency of cluster length in each species. Species are coloured according to the grouping shown in Figure 3.3. Different classes of genomic clusters are shown in each panel: a) clusters that contain only protein-coding genes; b) mixed clusters containing both protein-coding and miRNA genes; c) clusters containing only miRNA loci.

blocks across two or more species. A small fraction (3%) of non-clustered miRNA loci are found to be in conserved synteny with protein coding genes.

A number of syntenic blocks illustrating different evolutionary scenarios are shown (Figures 4.6 to 4.9). These striking cases were chosen to illustrate the variety of the different evolutionary contexts I observe within synteny blocks. In some situations new miRNA families can appear integrated in already existing, conserved syntenic clusters, albeit in a subset of species, such as mir-434, mir-540 and mir-3070 in mouse and rat (see Figure 4.6). One of the elements of this cluster, miR-127 has previously been shown to be involved in fetal lung development (Bhaskaran *et al.*, 2009). In other cases, part of a cluster duplicates locally, such as miR-302 (Figure 4.7). This cluster has been widely studied and is important in the definition of human embryonic stem cells (Barroso-delJesus *et al.*, 2011) and germ-cell tumours (Murray *et al.*, 2010). In extreme cases, a miRNA family, containing multiple miRNAs, has significantly expanded in primates and rodents (Figure 4.8). These miRNAs have also been shown to be important in ES cells and are likely involved in maternal zygotic switching in animals (Tang *et al.*, 2007). I also found clusters that duplicated within the genome, but to different chromosomes. One example of this phenomenon is the cluster shown in Figure 4.9. It contains a member of the let-7 family, one of the most conserved miRNAs known, and has been implicated in many fundamental biological processes, namely: development timing, ageing and malignancy (Thornton & Gregory, 2012). As a cluster, it has also been implicated in the regulation of primitive hematopoietic cells in mouse (Gerrits *et al.*, 2012).

The organisation of miRNAs between species seems to be more constrained than that of the nearby protein coding genes. Due to the diversity of possible scenarios, it is challenging to accurately reconstruct the series of events that led to the current organisation of genes (Nadeau & Sankoff, 1998). In general, the results are coherent with the hypothesis that miRNA genomic organisation is more conserved than expected compared to both random models and protein-coding genes (Altuvia *et al.*, 2005).

4.4 Conclusion

I have constructed a global synteny map and phylogenetic analysis for miRNAs across 80 animal species (Guerra-Assunção & Enright, 2012). These data not only form the basis of the analysis presented in this chapter, but are also an useful

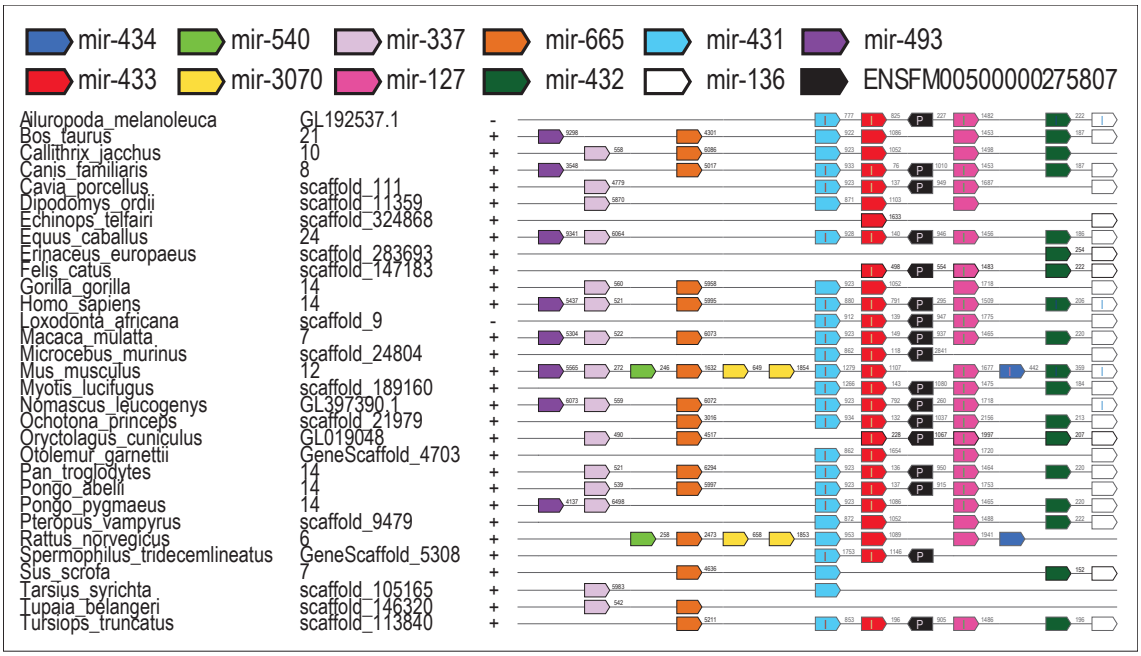


Figure 4.6: An example of clade specific evolution. In these synteny plots the species are sorted alphabetically, with the names shown in the first column. The other columns provide chromosome and strand information for the collinear block shown. Different colours represent different families. Each coloured arrow represents a locus and the direction it is encoded within the conserved block. In this conserved miRNA cluster, it is possible to detect an insertion of mir-540 and two copies of mir-3070, as well as miR-434, specifically in rodents (*Mus musculus* and *Rattus norvegicus*). In these plots, "I" indicated an intronic miRNA, "P" indicates a protein coding gene.

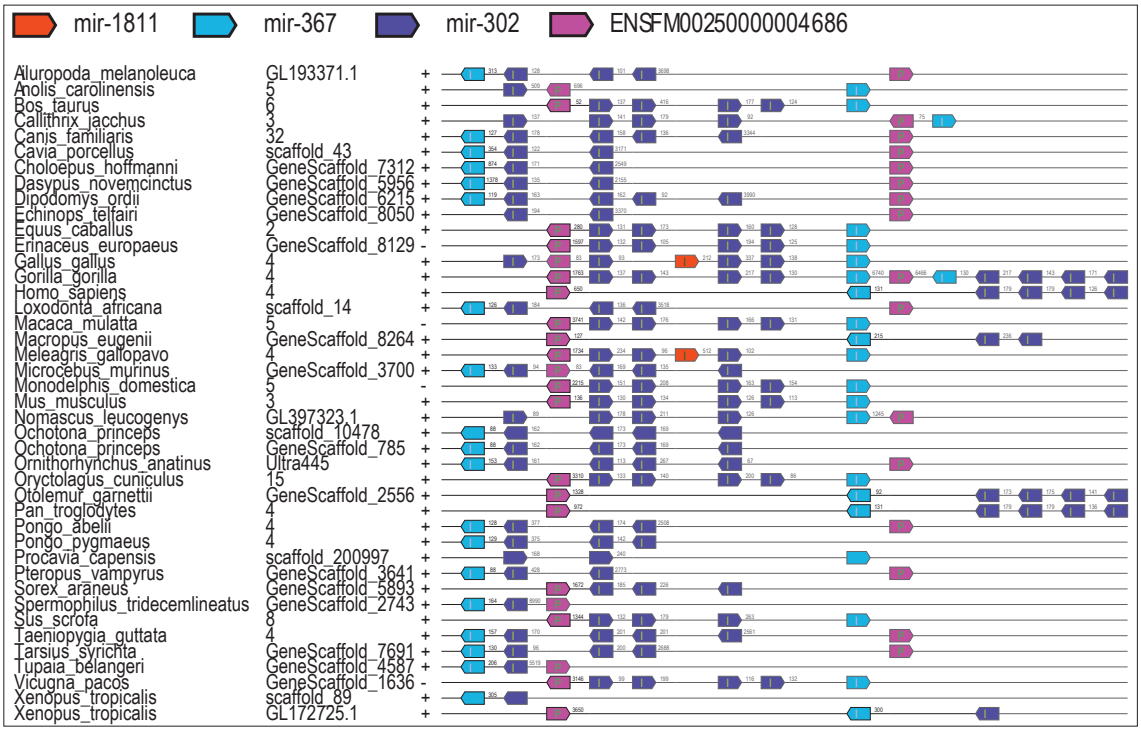


Figure 4.7: This example illustrates the evolution of miR-302, an intronic miRNA that is conserved across a broad range of species. The whole protein is duplicated in Gorilla, leading to the duplication of the whole miRNA cluster as well. It is also possible to see the insertion of mir-1811 between different copies of mir-302 in avian species (*Gallus gallus* and *Meleagris gallopavo*).

Figure 4.8: Fast local expansion of a mammalian miRNA family. This miRNA cluster is known to be highly expressed in embryonic stem cells, and was detected as being significantly expanding in primates (see Table 3.2).

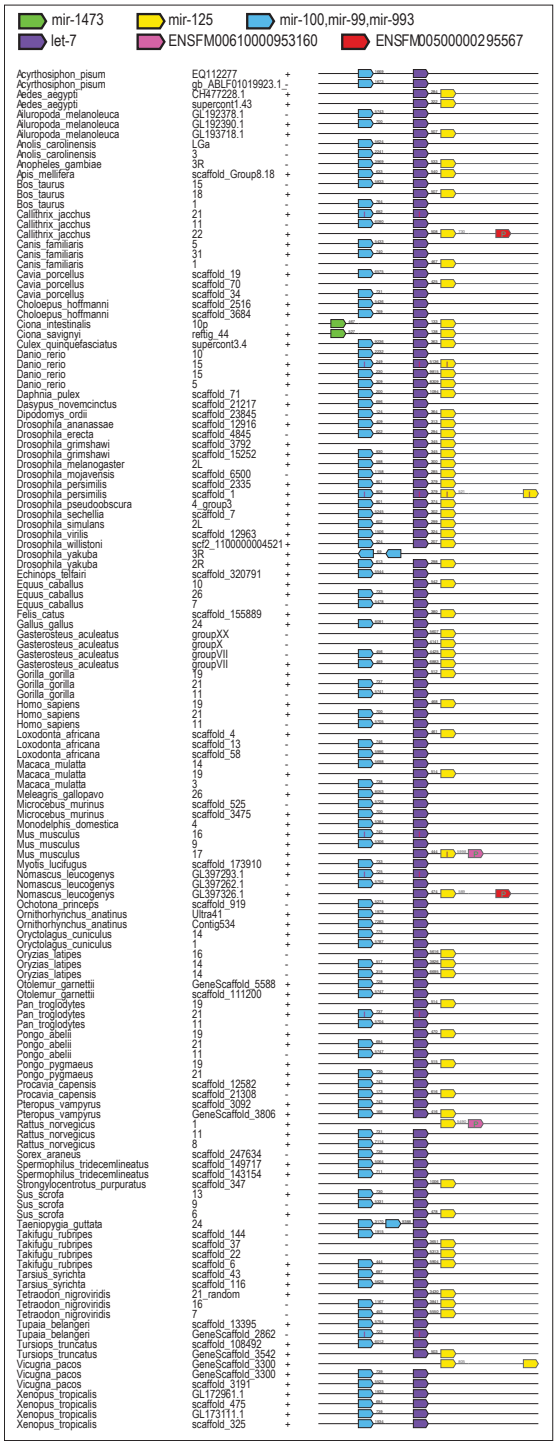


Figure 4.9: The miRNA cluster mir-99/let-7/mir-125, shown here, is the most widely conserved miRNA cluster in the dataset. It can be seen that the whole cluster underwent several rounds of duplication across the genome, with paralogues in different chromosomes in most species. It is also interesting to note the case of mir-1473 in cionidae, that is found in the place of the mir-99 family.

resource for the community. The complete dataset is available for query or download at the Sintra webserver[1]. This resource will be kept up-to-date, as new genomic and miRNA data are made available.

Using these data, I have undertaken a large-scale analysis of miRNA synteny, genomic organisation and evolution. These results recapitulate a number of earlier findings (Hertel *et al.*, 2006), in a fully automated fashion with many more genomes and miRNAs. This work revisits previous studies on the evolution of the miRNA repertoire and its correlation with morphological complexity (Heimberg *et al.*, 2008), whilst also highlighting the fact that few miRNA families are shared between different clades. I show that miRNAs have atypical patterns of synteny with preferences for longer clustered regions, which do not appear to be affected by genome compaction.

In general, four different patterns of miRNA cluster evolution can be considered. Clusters can potentially arise by the aggregation under a common promoter of independently evolved miRNA loci, even though this is thought to be rare. More common scenarios include simple local duplication of a particular miRNA family. Finally, in line with the proposal by Ohno for protein coding genes (Susumo, 1970), there is also evidence for functional divergence following gene duplication. These clusters can then further evolve by further local duplications of the whole cluster (see Figure 4.8), or duplicate non-locally, forming a paralogous cluster somewhere else in the genome (see Figure 4.9).

Even though the analysis presented in this thesis does not include a detailed quantification of each of this scenarios, due to the inherent difficulty of automatically reconstructing ancestral cluster structure, it provides a sound basis for this type of assessment on a per-cluster basis by manual assessment.

4.5 Materials and Methods

4.5.1 Synteny Block Detection

The syntenic anchor dataset was built by combining MapMi miRNA data and from protein-coding datasets retrieved from Ensembl. These were identified by their family name, making the homology between anchors coherent between species.

[1] http://www.ebi.ac.uk/enright-srv/Sintra

The file was sorted by its genomic coordinates and duplicates were eliminated according to the Enredo documentation, using the provided tool. I detected conserved collinear segments using Enredo (Paten *et al.*, 2008) (version 0.5) using the following options: max-gap-length=10000, max-path-dissimilarity=10, min-regions=2, min-anchors=2, simplify-graph=7. Blocks sharing a terminal anchor were chained together, according to standard operating procedures (Javier Herrero, personal communication).

4.5.2 Synteny Block Visualisation

To visualise miRNA containing synteny blocks, I developed a set of Perl scripts to align the conserved synteny blocks by miRNA family using a Perl implementation of the Needleman-Wunsch algorithm producing plots using PostScript. While this does not change any of the properties of the blocks found by Enredo, it makes the synteny plots easier to assess visually. Furthermore, each anchor is coloured based on its family (e.g. Figures 4.6 to 4.9).

Different visualisation modes are available. Using the web resource, it is possible to download a text version of the conserved synteny plots, that can be visually assessed using a text editor or parsed automatically for further analysis. For the graphical visualisations in PostScript and PDF format, the user can choose to see the alignment view, which is the default, or plot the cluster alignments in the context of the species phylogenetic tree, allowing an easier assessment of evolutionary events shaping miRNA clusters.

4.5.3 Analysis of Block Length Distribution

The length of each conserved synteny block in each species was used. Following Enredo analysis, the original anchor coordinates were recovered and the start and end coordinates of the first and last element of each cluster were used to compute block size. For the normalisation procedure, genome lengths were inferred from the available genome FASTA files, retrieved from Ensembl (v 62). For ease of analysis, each species was coloured based on the phylogenetic clade it belongs to, according to the same classification that was employed in Figure 3.3, with different dashing to distinguish each species. The R statistical analysis framework was used to produce the final figures and perform the normalisation.

4.5.4 Integration of Repertoire Evolution and Genome Context

This analysis integrates many data sources found within this thesis as a whole. For each miRNA family, Dollo parsimony analysis (see Section 3.5.3) was used to identify the node within the provided phylogenetic tree was the most likely for its origin. As a proxy for evolutionary age, the species phylogeny was used to compute the distance between this node and the root of the phylogenetic tree under analysis. An assessment was made to determine the genomic context (intronic and clustered states) for each family in each of the species that contains it. The data were then combined for all families that appear at each of the nodes of the phylogenetic tree and the percentage of loci with each of the properties was plotted using the R statistical analysis framework. Each node was coloured to distinguish internal nodes, terminal nodes corresponding to species with low-coverage genomes, and terminal nodes corresponding to species with high-coverage genome sequences.

Chapter 5

Intra-species Variation of microRNA Loci and Their Targets

5.1 Aim

In previous chapters, I examined different facets of the evolution of microRNAs (miRNAs) across a broad set of species, representing all the main metazoan clades.

In this chapter I seek to explore the selective pressures affecting miRNA loci and their predicted target sites, at an intra-specific level, with particular interest in the characteristics of miRNA families derived from repetitive elements.

By dividing each miRNA locus and its corresponding target sites into relevant classes, I searched for patterns indicating strong negative selection. Using this approach it is possible to identify critical nucleotides and features that are likely to be important for the correct function of this class of regulatory molecules within the mouse genome.

5.2 Introduction

When cells replicate, DNA is usually faithfully copied from one cell to the next. However, there can be errors in the replication process. These mutations occur in a seemingly random fashion during the process of DNA replication. While most mutations occur in somatic cells, affecting only the individual they occur in, some mutations occur in germ line cells that will produce gametes, thus spreading the mutation to the next generation.

There are several types of mutations, which can affect the message encoded by the DNA in different ways. Some mutations will affect several nucleotides, either consisting of either a small insertion, or a small deletion, frequently called *indels*. In this chapter, I focus on mutations that occur in a single nucleotide, and are usually termed point mutation or single nucleotide polymorphism (SNP). It is still challenging to computationally predict the effect of an indel affecting a miRNA loci. Since indels are less frequent than SNPs, their removal from these analyses does not significantly influence the results.

Normally such mutations are mostly harmless, and have little functional effect (neutral mutation) (Kimura, 1968). This can occur when the mutation occurs in a region of the genome that does not affect any regulatory element, or does not affect the amino acid sequence of a protein (synonymous mutation). This is possible due to the redundancy of the genetic code (usually on the 3^{rd} position of the codon), whereby different codons will code for the same amino acid. Mutations within protein-coding genes that do not change its amino acid sequence, are called synonymous, while mutations that cause a change of amino acid are called non-synonymous.

This natural source of intra-specific variation is then affected by natural selection acting at the population level. As a consequence, at an evolutionary time-scale, mutations that are beneficial will tend to increase in frequency within the population (positive selection), while mutations that disrupt the correct functioning of the affected element, having a negative impact on the organism, have a tendency to be eliminated from the population (negative or purifying selection).

It is challenging to directly trace back all evolutionary changes in wild populations, as it would require information from the original ancestors of modern day species, that are no longer available. Nevertheless, it is possible to compare the rate of mutations and allele frequencies in different regions of the genome, offering a way to infer where and how elements are being affected by natural selection. This allows the identification of elements and regions that are under stronger purifying selection, potentially indicating essential roles within the cellular environment.

While the analysis of SNPs affecting protein-coding genes is common practice, the study of SNPs affecting non-coding elements in the genome has only recently become a focus of the scientific community (Georges *et al.*, 2007). In this chapter, I will discuss the aspects of animal miRNA biology that should be taken into account

when analysing the effect of SNPs, as well as general intra-specific evolutionary analysis within these elements and what can be inferred from it.

5.2.1 Single Nucleotide Polymorphisms Affecting microRNA Loci in Mouse Strains

For a more complete understanding of the way miRNAs evolve, it is important to look at intra-specific variation affecting miRNA loci and their target sites, as it provides a view at a shorter time-scale on the selective pressures affecting these genomic sites. The study of conservation profiles between species (Chapter 2, Section 2.3.3) reveals many complexities and constraints associated with miRNAs over long evolutionary distances. Nevertheless, not all miRNAs are widely conserved between species, hampering this kind of analysis. Furthermore, different species evolve at different rates which makes detailed analysis difficult (Graur & Li, 2000).

The high degree of similarity between the mature sequence across species, and simultaneous high divergence of the loop region can hamper phylogenetic approaches, as it is challenging to estimate the actual number of substitutions between sequences. The sequence of the mature sequence is more constrained, as it is important for target recognition. Conversely, the loop region is mainly a structural feature (Cullen, 2004), so a change of sequence or the presence of an insertion or deletion is though to be of little consequence. This creates difficulties to model the evolutionary changes that affect the precursor sequence of miRNA genes in large multi-species datasets.

Current sequencing technologies enable the detection of single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels). When high sequencing depth is available, it is also possible to detect large insertions and deletions, routinely referred to as copy number variants (CNVs). Given high biomedical interest, there is a particular focus on human and mouse and there are many variation datasets available for these species (Sherry *et al.*, 2001).

Due to the nature of the genetic code, synonymous substitutions are generally believed to be under little or no selection. For this reason, the rate of non-synonymous to synonymous substitutions is frequently used as a measure of the strength of selection affecting a certain protein-coding gene (Nei & Gojobori, 1986). This is commonly represented by the ratio between rate of non-synonymous substitutions and the rate of synonymous substitutions (dN/dS). It has so far been impossible, due to their small size and non-coding properties, to find a similar metric that can

be applied to non-coding RNAs. While regions like the stem are under stronger purifying selection than the loop, both are still far less variable than the rest of the genome, so it can not be assumed that any region is under neutral selection (Mimouni *et al.* (2009) and also Figures 5.2 and 5.3).

The biological characteristics of miRNAs (Section 1.2) are likely to be affected by selective pressures in particular ways. For instance, variation affecting processing enzyme recognition can have an effect on the efficiency of miRNA production, cellular levels of miRNA and even strand selection, which can lead to potential deleterious effects (Jazdzewski *et al.*, 2008).

The importance of correct base pairing between the miRNA seed region and target sites for correct miRNA target recognition is likely to be responsible for the strong selective pressures that have been reported within this region (Mu *et al.*, 2011).

Although miRNA loci can potentially remain functional upon mutation outside the mature region without any known consequences, a change in the seed region can drastically change the regulatory network of miRNA targets, exhibiting serious phenotypes. One of the first examples of a mutation in a miRNA locus being responsible for a disease phenotype, was reported for mmu-miR-96 and hereditary deafness (Lewis *et al.*, 2009). While the study was performed in mice, it was also reported that a similar mutation in the same miRNA has the same effect in Human (Mencía *et al.*, 2009). Another interesting aspect, is that a mutation affecting the seed region will not only affect the regulation of existing miRNA targets, but also generate a new set of targets, causing a spurious set of genes to be down-regulated at a wrong moment, if they were being expressed (Lewis *et al.*, 2009).

5.2.2 Single Nucleotide Polymorphisms Affecting Predicted microRNA Target Sites

There is much interest in studying natural intra-specific variation that affects predicted target sites. The effect of a mutation disrupting a miRNA target site is potentially less deleterious than a change in the seed region of the miRNA itself, as it changes only one target and not the full regulatory network. Nevertheless the loss of a target site can have important biological implications.

Several studies have described phenotypes of medical relevance caused by mutations in 3' UTRs affecting miRNA target sites (Esteller, 2011; Gong *et al.*, 2012;

Zorc *et al.*, 2012). In the work described in this chapter, I focus not so much on the identification of particular disease associated variation or their effects, but rather on the general patterns of variation and the study of selective pressures acting on predicted target sites.

This study builds upon research previously reported for Human (Chen & Rajewsky, 2006), using the most complete dataset of Mouse genomic variation to date, and focusing on the differences between repeat-associated and non repeat-associated miRNAs. This larger dataset and the existence of many more target sites due to better UTR annotation and target prediction, provides added resolution enabling a more detailed study.

5.2.3 A Reflection on the Role of Repeat Derived microRNAs for the Evolution of the miRNA Repertoire

Throughout the work described within this thesis, I have found that repeat-derived miRNAs behave in a different fashion and have different evolutionary patterns when compared to non-repeat associated miRNAs. In this chapter, I aim to explore this further by using the available information to compare the SNP frequencies of repeat-associated miRNAs and their predicted target sites with those of non repeat-associated miRNA families, as well as a set of non-miRNA neutrally evolving controls.

The number of paralogous loci per genome is much higher for repeat-derived miRNAs. Their genomic organisation is quite diverse, not obeying the same patterns found in other, non repeat-associated, miRNAs. The results within this chapter aim to elucidate the importance, from an evolutionary perspective, of these higher copy-number miRNA families within the genome.

The main open questions regarding these elements are still their biological roles. There have been reports that repeat-associated miRNAs bind to Argonaute (Goff *et al.*, 2009) and some are expressed at detectable levels and were found to be similar to miRNAs in a mouse validation study (Chiang *et al.*, 2010). The fact that they are present in multiple copies in the genome make them less suitable for standard knock-down or knock-out studies.

Three main scenarios can be hypothesised for the function of repeat-derived miRNAs within the cell. Some repeat-derived loci might be non-functional, miRNA

pseudo-genes that retain a hairpin shape. Some might be functional canonical miRNAs that are under less selective pressure because they are present in multiple copies in the genome. Hence, if one of the loci is affected, the function can be compensated by other loci of the same miRNA family. It might also be the case that some repeat-derived miRNA families are rapidly evolving as they are being integrated into the already established miRNA regulatory network.

Computational analysis of repeat-derived miRNAs is not free from challenges. Frequently, these miRNAs are recent additions to the miRNA repertoire of the species, making it difficult to use inter-species conservation patterns. This diminishes the power of synteny based approaches. Therefore, a general analysis of the patterns of intra-specific variation affecting repeat-associated miRNAs can potentially provide the best overview of the importance of these elements within the cell.

5.3 Results

To study selective pressures affecting miRNA loci and predicted target sites, I inferred the rate of variation across different biologically relevant regions, and compared it with overall rates of variation across different regions of the genome. It is important to highlight that the rate of variation that is detected, is the composite effect of the variation within the population combined with selective pressures affecting each genomic region.

Regions where less SNPs are present than expected, are believed to be under purifying selection, that is, attempting to preserve the original sequence without accumulating mutations that will likely have a deleterious effect. The challenge is then to identify, with as much detail as possible, the regions that are actively being selected, and try to infer their biological function.

Of particular interest for this study is the comparison of repeat-derived miRNAs to other miRNA loci. For this purpose, each miRNA family under analysis was classified as either repeat-associated or non repeat-associated. This dataset is composed of 736 miRNA precursors that have coordinates defined in miRBase (v18). Of these, 528 correspond to non repeat-associated miRNA families, while 208 are classified as repeat-associated. This was established by assessing for each hairpin that overlaps Ensembl repeat element annotations, and combined by miRNA families. This approach is consistent with previous analyses presented in this thesis where the repeat status of the miRNA is taken into account.

This dataset was matched as closely as possible to non-miRNA, unannotated genomic hairpins (see Materials and Methods) to provide a solid basis for the comparison of selective pressures affecting miRNA loci.

5.3.1 Single Nucleotide Polymorphisms Affecting microRNA Loci

From the overlap of the SNP dataset with the 736 miRNA loci under analysis, 720 SNPs were identified as being within the boundaries of miRNA hairpins. When this was performed for my background dataset, 4,081 SNPs were detected within the boundaries of 2,224 genomic hairpins that are part of the background dataset used.

The SNPs were then grouped into classes as the SNP density per nucleotide was found to be too low to accurately distinguish differences at the base-pair level from random noise. Therefore, these polymorphisms were grouped according to two independent classification systems.

A functional, sequence-based classification was devised to take into account the perceived biological properties of each sequence fragment within the hairpin. The three classes are the "seed" region, the "mature" sequence (excluding the seed region) and the remaining SNPs within the hairpin being classified as "precursor" (see Materials and Methods Section 5.5.2).

The second classification groups SNPs depending on which predicted secondary structural element they are likely to affect (see Figure 5.1). The structural classes were defined based on inter-species conservation results analysed within Chapter 2, and by previous analysis of variation affecting ncRNAs (Mimouni *et al.*, 2009).

As expected, I found that the seed region and mature sequence significantly avoid the accumulation of mutations when compared to the rest of the precursor molecule ($P < 6.05 \times 10^{-11}$, 6.04×10^{-5} respectively, two-sided Mann-Whitney-Wilcoxon (MWW) test). Surprisingly, the difference between seed region and the remaining mature sequence is not as striking, not reaching significance ($P < 0.14$ two-sided MWW test). A contributing factor is that very few of the SNPs (59) fall in the seed region of known miRNAs. Significant results were obtained for all classes when the respective functional classes were compared with the background dataset (see Figure 5.2).

When the results are sub-divided between SNPs affecting repeat-associated and non repeat-associated miRNAs, it becomes clear that the SNP frequency is higher for

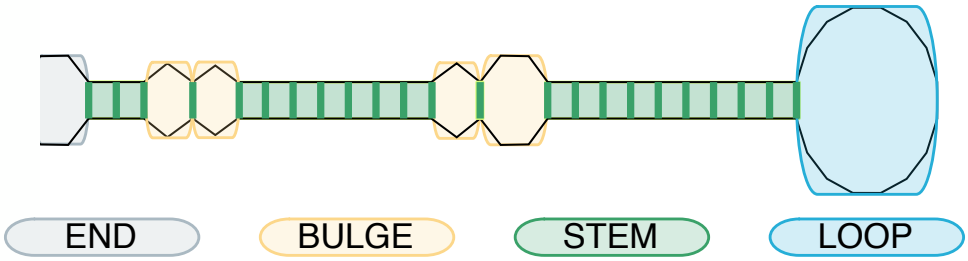


Figure 5.1: Schematic representation of the division of a miRNA hairpin into structure based classes.

repeat-associated miRNAs (see Figures 5.2 and 5.3). Interestingly, I found that background hairpins that overlap repeat elements seem to have a higher SNP frequency than non repeat-associated background hairpins consistent with the hypothesis that repeat elements evolve faster than other genomic elements. For both repeat and non repeat-associated miRNAs, it was found that SNP frequency was lower than for background genomic hairpins, suggesting that at least some of these elements are under purifying selection.

This structure based classification yielded interesting results (Figure 5.3). As foreseen, the paired nucleotides of the stem region have significantly less variation than the loop region ($P < 0.028$ two-sided MWW test). Surprisingly though, bulges within the structure seem to accumulate a similar number of mutations when compared to the loop region ($P < 0.68$ two-sided MWW test). A more detailed analysis of this class of SNPs indicates that the majority (51%) of these mutations do not change hairpin stability at all, as assessed by comparing the minimum free energy (MFE) of the hairpin with and without the mutation. The MFE is an estimation of the energy required to dissociate the bonds between nucleotides in a secondary structure and is commonly used as a proxy for hairpin stability. Even when the global hairpin stability changes, the mutations do not indicate a pressure to reduce the number of bulges within the structure. Since these bulges often overlap mature sequences, it is possible that the effect of these mutations is more relevant to the sequence context rather than structural context.

When compared to the background hairpin dataset, I found that all structural classes were also significantly different, indicating that even the loop region is under

some level of purifying selection.

In general, the comparison between repeat and non repeat-associated loci shows, as with the functional classes, a qualitative difference between these two types of miRNA families, showing no significant qualitative differences.

5.3.2 Single Nucleotide Polymorphisms Affecting microRNA Predicted Target Sites

To detect variation affecting miRNA target sites, I computed a high sensitivity target set using TargetScan v6.2 (see Section 5.5.4). These were separated based on the association of the miRNA family this site corresponds with repeat element annotations present in Ensembl. The SNP frequency of these regions were also compared with the set of unannotated, 21bp background regions, randomly selected from 3' UTRs as control regions.

The TargetScan method includes a series of progressive filtering steps that aim to increase specificity, based on genomic context, phylogenetic properties and conservation. In this particular study, my interest focused on the general patterns, rather than specific target sites for particular miRNA families. To maximise sensitivity, this dataset contains both conserved and non-conserved target sites. Recently developed experimental methods to assess miRNA targets have shown that many non-conserved target sites are functional (Ellwanger *et al.*, 2011; Giraldez *et al.*, 2006). Furthermore, given my interest in repeat-associated miRNAs in particular, the loci of which are not widely conserved, I have chosen not to enforce a conservation constraint on predicted target sites. It is thus reasonable to assume that even with genomic context-based scoring and filtering, there are still likely to be false positive predicted target sites within the dataset, that might dilute observed signals of negative selection.

The target site dataset is composed of 1,022,782 TargetScan predicted miRNA target sites for miRNAs that are not repeat-associated, 337,182 predicted target sites for miRNAs that are repeat-associated and 24,709 unannotated background regions. The SNP frequencies for each of the positions in each sequence were analysed (Figure 5.4).

The variability between positions, in particular in the background set of regions, is likely due to their random nature, and to the lower number of regions in this class.

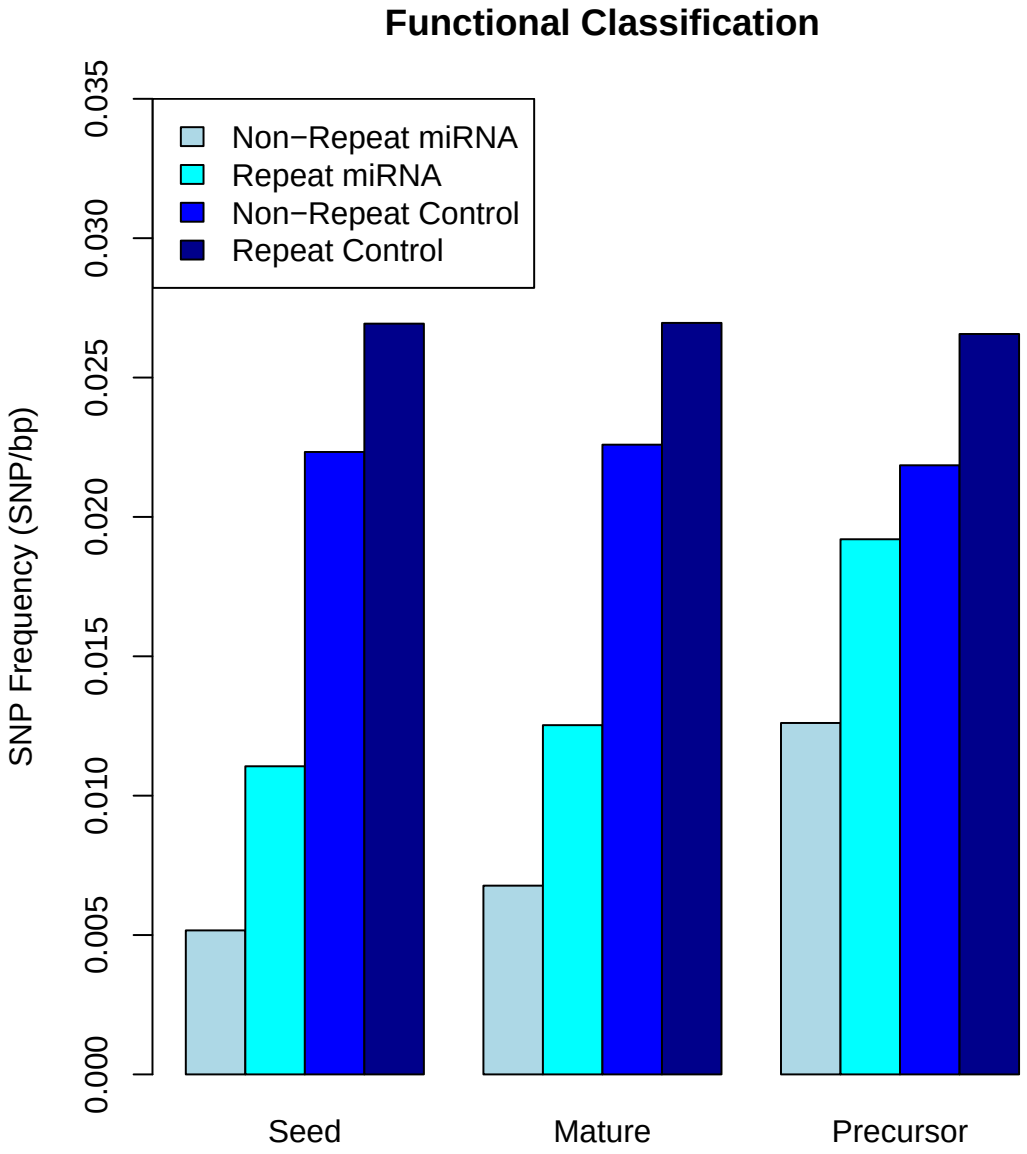


Figure 5.2: Comparison of SNP frequency between functional regions within miRNA loci. The number of polymorphisms per nucleotide is represented on the y-axis. The different regions are represented on the x-axis and are colour coded by class: non repeat-associated miRNA loci, repeat-associated miRNA loci, non repeat-associated background regions and repeat-associated background regions respectively.

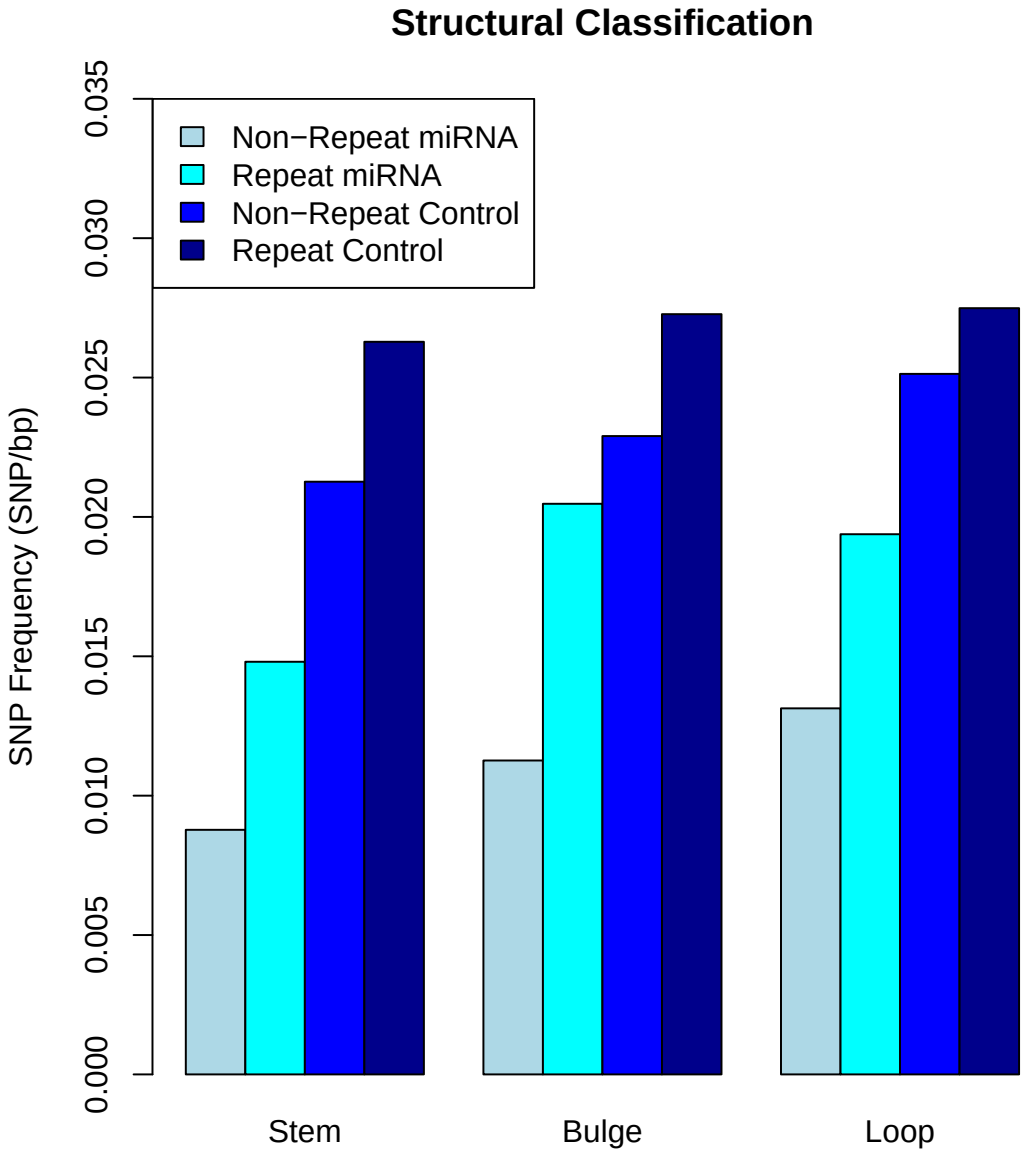


Figure 5.3: Comparison of SNP frequency between structural classes within miRNA loci. The number of polymorphisms per nucleotide is represented on the y-axis. The different regions are represented on the x-axis and are colour coded by class: non repeat-associated miRNA loci, repeat-associated miRNA loci, non repeat-associated background regions and repeat-associated background regions respectively.

Predicted targets for non repeat-associated miRNAs show a clear pattern highlighting the important role of the seed sequence, that is accumulating less mutations than the remaining positions of the target sites. It is interesting to note that even less constrained positions seem to be less variable than the background set. This indicates, as previously suggested (Bartel & Chen, 2004) that these positions, although not essential, may also play a role in miRNA targeting. Furthermore, even though TargetScan favours larger seed sequences (7mers and 8mers), the profile seems to suggest that non repeat-associated miRNAs are being mostly constrained in 6mers or slightly less so in 7mers. Interestingly, in contrast with previously reported results (Chen & Rajewsky, 2006), this analysis seems to differentiate the first position of the seed region, in both classes of predicted target sites, as being less constrained than the remaining seed region. This is likely due to the more complete dataset, that has better resolution than the previous analyses.

Intriguingly, predicted target sites for repeat-associated miRNA loci show several unexpected quantitative and qualitative differences. The SNP frequency affecting predicted target sites for miRNA families associated with repeat elements is higher than for those of the non repeat-associated families. Nevertheless, a seed region is still clearly observable, indicating that at least some of these target sites and respective miRNA loci appear to be acting as canonical miRNAs. Curiously, the hallmarks of purifying selection in the seed region seem to suggest that the seed region is longer (7mer or 8mer) for the repeat-associated miRNAs, than for the other miRNAs (6mer or 7mer). It is also interesting to take into account the fact that the remaining positions of the predicted target sites have a SNP frequency similar to that of the background regions, suggesting that these positions are not under strong purifying selection.

Finally, it is noticeable that position 6 of the predicted miRNA target sites for both classes of miRNA families seem to have a lower SNP frequency than adjacent base pairs. This is particularly evident for the repeat-associated families. Nevertheless, a careful analysis of the dataset did not produce any clear explanation to justify the observed pattern. There are also no other reports of this phenomenon in the literature. It thus remains to be seen if this is a somewhat unusual characteristic of this dataset, or if it can be found in other datasets and potentially other species.

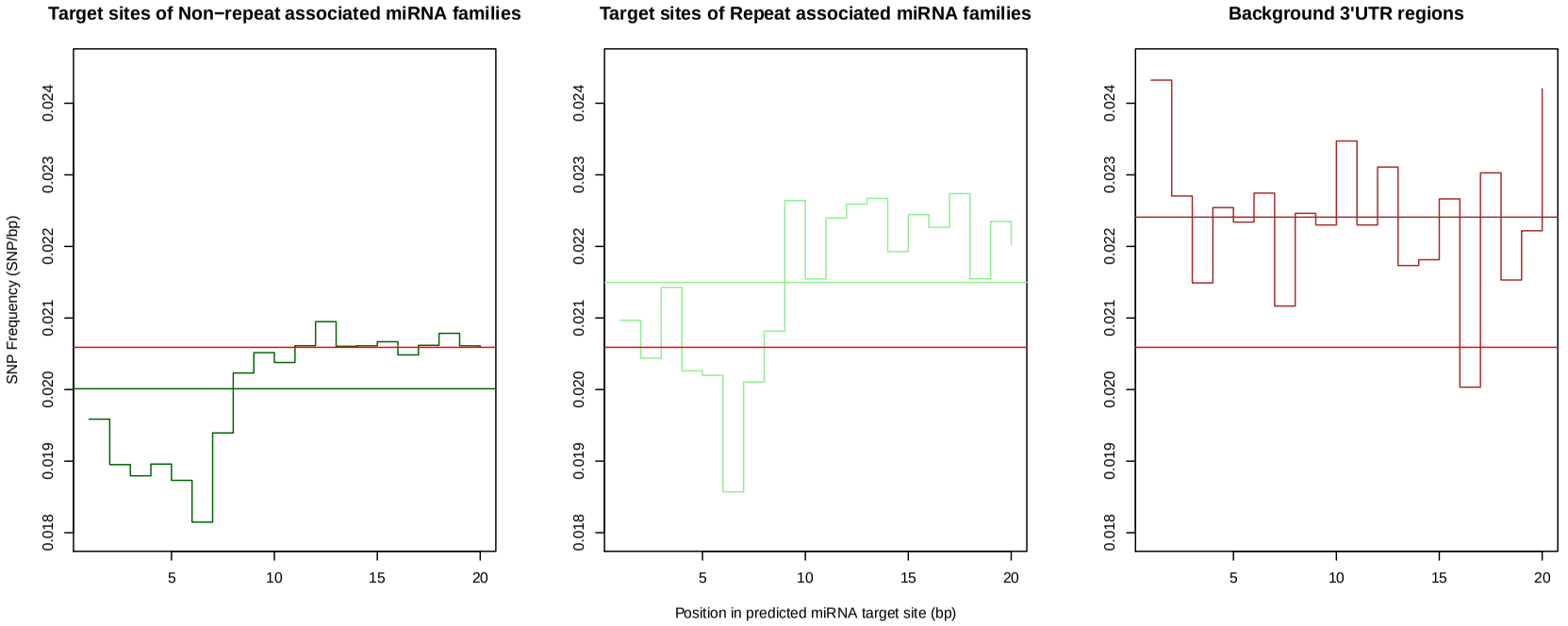


Figure 5.4: SNP frequency within different 21bp regions of 3' UTRs. The x-axis represents the position within the predicted target sites for each of the classes. A) SNP frequency for TargetScan predicted target sites for non repeat-associated miRNA families. B) SNP frequency for TargetScan predicted target sites for repeat-associated miRNA families. C) SNP frequency for non-miRNA related background windows within 3' UTR. The red line represents the average SNP frequency for all 3' UTRs in the dataset. Other coloured lines represent the median of all the base pairs for the whole target site in each of the situations.

5.3.3 Comparison of Evolution Rates Between microRNAs and Other Genomic Elements

It is still a challenge to create a suitable self-contained metric to evaluate the selective pressures affecting ncRNA elements within the genome. It is frequently assumed that synonymous substitutions within protein-coding genes are evolving neutrally, serving as comparison for non-synonymous substitutions affecting the same genes, allowing a wide range of approaches to detect selective pressures. Since no such region has ever been reported for small ncRNAs and in particular miRNAs, the rates of mutation per base pair were compared with other comparable regions of the genome. These results were combined, providing a quick overview of which elements are varying the most within the mouse genome, for the 17 strains under analysis (Keane *et al.*, 2011) (see Figure 5.5).

In this dataset the highest SNP frequency is found for synonymous SNPs within protein-coding genes. This is coherent with the assumption that these positions evolve neutrally. Both sets of background hairpins accumulate less substitutions than synonymous SNPs. This can be somewhat justified by the way they were selected, as they are present in the vicinity of known miRNA genes, which are under strong purifying selection. To the best of our knowledge, these background regions are unannotated and are not expected to have any function within the cell. Nevertheless, it is also possible that a fraction of these hairpins have a yet to be discovered function within the cell.

It is curious to note that repeat sequences within the genome appear to be under weaker constraints than the corresponding non repeat-associated regions. This is particularly interesting for the predicted target sites of repeat-associated miRNAs, as the target sites themselves are not repeat-associated.As previously mentioned, predicted target sites for both repeat and non repeat-associated miRNA families have a lower SNP frequency compared to the control unannotated regions within 3'UTRs, even though the repeat-associated sites seem to accumulate more mutations than the average 3' UTR region. As expected, the miRNA loci appear to be under the strongest purifying selection of all the elements analysed, in line with their important regulatory functions within the cellular environment, and the deleterious effect most mutations seem to have on these loci (Jazdzewski *et al.*, 2008).

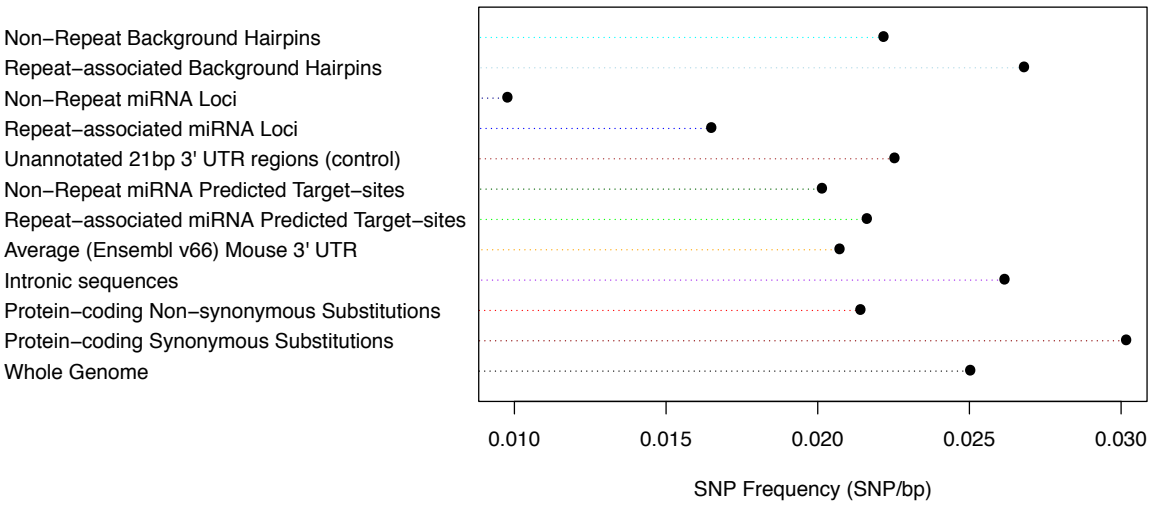


Figure 5.5: Illustration of the different rates of SNP accumulation within different regions of the mouse genome relevant to this study.

5.3.4 An Overview of the Mouse microRNA Repertoire and Their Accumulated Variation

The miRNA repertoire within the mouse genome contains families with a wide range of inter-specific conservation patterns, ranging from families like let-7, that are present in all sequenced metazoans to date, to other families that are only found in the mouse genome. Extending the analysis previously presented (Chapter 4, Section 4.3.2), I used Dollo parsimony to determine the location within the phylogenetic tree where each of the miRNA families arose in evolutionary time. This information was used to calculate the SNP frequency of the families in each node of the tree leading to mouse (see Figure 5.6).

As expected, it appears that widely conserved miRNA families accumulate less SNPs than more recent miRNA families. In particular, miRNA families only found in the mouse genome seem to have the highest SNP frequency. This result is coherent with the general idea that recent miRNA families have a higher turn-over rate as they are adapting to the cellular regulatory network. As new information is produced, and new species sequenced, it will be possible to reconstruct the evolutionary history of the current miRNA regulatory network with greater precision, improving on these findings.

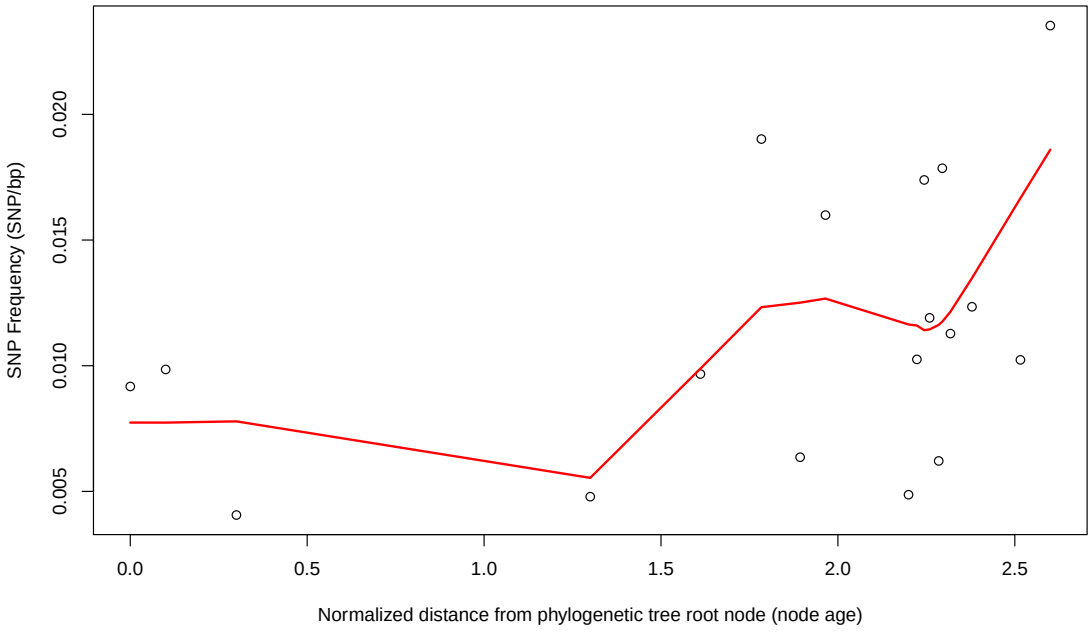


Figure 5.6: SNP frequency per miRNA family evolutionary age. The x-axis represents the evolutionary distance between the root of the species phylogeny to the node under analysis. The y-axis represents the average SNP frequency for the miRNA families that arising on each of the nodes of the tree leading to mouse. The red line is the LOWESS (locally weighted scatterplot smoothing) regression line.

5.4 Conclusion

This study explores the patterns of intra-specific variation within 17 mouse strains, in respect to miRNA loci and predicted target sites. This research not only confirms previous findings, but also brings in a new level of detail due to the higher density of variation features within the dataset. I also present new findings, in particular regarding the differences in selective pressures between repeat-associated miRNA loci and their respective targets, when compared with those of miRNA families that are not associated with repeat elements.

This study has shown that although repeat-associated miRNAs show different properties and patterns of selection when compared with the other miRNA families, they still retain some features that suggest that at least some of them act as canonical miRNAs.

At the beginning of this study, three main hypotheses were set forward, regarding the function and characteristics of repeat-derived miRNAs within the mouse genome. It can be that repeat-derived miRNAs arise randomly and do not have a relevant regulatory role within the cell. A different view suggests that these loci, due to the higher number of paralogues per genome, can withstand the presence of more SNPs without loss-of-function within the cell. It is also possible that the higher rate of mutations observed within these loci are hallmarks of the adaptation of the recent addition of these miRNA families within the miRNA regulatory network. Depending on the miRNA family and the loci itself, it is likely that the three scenarios are present within this class of miRNAs. Although a specific answer cannot be obtained for each miRNA family, this study provides evidence for negative selection affecting both repeat-associated and non repeat-associated miRNA loci and target sites, indicating that at least some of these miRNA families are acting as miRNAs within the cell.

The analysis of repeat-associated elements in comparative genomics still raises several technical challenges. At the experimental level, the multitude of loci spread over the genome pose new challenges for directed studies, turning them into a challenge for functional analysis. Nevertheless, it seems that until the research community is able to overcome these issues, there is a fraction of miRNA biology that will remain to be fully explored. Many of these repeat-derived miRNA families are quite recent on an evolutionary time-scale, with some of them being specific to the mouse

genome. It is thus reasonable to assume that adaptation between new miRNA loci and the cellular transcriptome is still occurring.

To further test this hypothesis, it would be necessary to detect hallmarks of recent positive selection, and to have a more detailed view on the selection forces acting upon each SNP in particular. Unfortunately, these analyses are not possible with the current dataset. While this dataset was chosen for its completeness regarding divergence time and number of variation features it provides, it lacks good estimates for allele frequencies, which will be needed for further research.

Further directions of research might include the extension of this research to other datasets, and even other species (e.g. Human) where more data is available. Although there are previous studies that characterise polymorphisms affecting miRNAs in Human, no distinction between repeat-associated and non repeat-associated miRNA families was made. The use of different datasets will enable further analyses based on derived allele frequency (DAF). Another approach is to restrict the current dataset to the variants for which allele frequency has already been estimated.

There is a large effort underway to explore experimental techniques for miRNA target determination (see Section 1.4.1.4). A closer inspection at these datasets might be useful to support the findings related to different seed lengths for repeat-associated miRNAs when compared to non repeat-associated miRNA families.

It is my hope that these analyses serve as a stepping stone for further research that will lead to a better understanding of the evolution of repeat-associated miRNAs and their biological functions.

5.5 Materials and Methods

5.5.1 Dataset

5.5.1.1 Genomic Data

The genomic sequence for *Mus musculus* assembly NCBIM37 for both the repeat-masked and non-repeat-masked versions was retrieved from the Ensembl FTP site. This site was also used to retrieve the genomic annotation (Ensembl 66) in GTF format. Custom Perl scripts based on the Ensembl API were used to create tables containing coordinates for 3'UTRs, protein-coding exons and introns. In case of several 3'UTR versions being available for the same gene, the longest UTR was used.

5.5.1.2 Variation Data

This analysis was performed based on SNP data from 17 mouse strains, as described in (Keane *et al.*, 2011). While the dataset also includes small insertion and deletions, due to the difficulty of determining the biological effects of these on miRNAs, they were excluded from this analysis.

5.5.2 microRNA Loci

The miRBase database (release 18) was used to retrieve the 736 *Mus musculus* miRNA loci for which coordinates were available, and their respective 5' and 3' mature sequences.

For each locus, a comparison was made between the repeat masked and non repeat masked versions of the genome. Loci that were found to overlap repeat elements were classified as *repeat-associated*. These data was integrated in a coherent way within miRNA families that have multiple paralogues within the mouse genome.

The loci were then placed into classes for further analysis. Sequence classes are "Seed", "Mature" and "Precursor". The "Seed" region was defined as nucleotides 2 to 7 of the mature sequence and "Mature" as the remaining nucleotides in the mature miRNA sequence. The remaining regions of the pre-miRNA were classified as "Precursor".

The secondary structure for each locus was predicted using RNAfold (v. 1.8.5). The hairpin structure was then classified into three structure classes: Stem, Bulge and Loop. Within the hairpin, the base pairs that are part of the stem and form Watson-Crick base pairs are classified as "Stem", base pairs that are part of the stem but are unpaired with the opposite side of the hairpin, are classified as "Bulge" while the remaining base pairs that form part of the main hairpin loop are classified as "Loop".

5.5.3 Dataset of Background non-microRNA Genomic Hairpins

The background dataset was built to be as comparable as possible to *bona fide* miRNA hairpins, while ensuring they do not overlap any known element of the mouse genome.

The full genomic sequence of *Mus musculus* (NCBIM37) was retrieved from the Ensembl FTP website. RNALfold (Vienna RNA package v1.8.5) was used to find locally stable RNA secondary structures across the whole genome. These secondary structures were filtered according to a series of criteria designed to match as closely as possible those of known miRNAs. Hairpins were selected to be between 50 and 150 nucleotides in length, with a maximum of one main hairpin loop, and a maximum minimum free energy, as provided by RNALfold of -20 Kcal/mol. If several hairpins matching these criteria overlap, the longest one was retained. The background hairpins were then filtered, excluding those that overlap existing genomic annotation, retrieved in GTF format from the Ensembl FTP site. The final filtering step removed hairpins responding to sequences of low complexity, using the dust algorithm (Morgulis *et al.*, 2006).

The hairpins and sequences that are part of this final dataset were classified based on their genomic context (e.g. intronic or intergenic) and assessed for overlap with known repetitive elements. For each background hairpin, one potential mature sequence was selected, based on the average distance between the end of the hairpin and the start of the mature sequence in known miRNAs. The arm of the hairpin yielding the mature sequence was randomly selected.

As expected, this dataset contains a much larger number of hairpins than the dataset of known miRNAs. To address this issue and to remove potential biases due to different rates of evolution at different locations within the genome, only background hairpins within 1Kb of a known miRNA loci were kept for further analysis. The remaining analysis on these regions was performed with the same scripts used for miRNAs.

5.5.4 Target Prediction

The prediction of putative miRNA target sites was done using the Perl implementation of the TargetScan algorithm version 6.2 (Friedman *et al.*, 2009; Garcia *et al.*, 2011; Grimson *et al.*, 2007; Lewis *et al.*, 2005). For coherency between annotations, the sequences and coordinates for 3' UTRs were retrieved using the Ensembl Perl API. MiRNA families, defined by their seed sequences, were grouped together according to the instructions provided on the TargetScan website.

This dataset was used to compute predicted target sites. These were then scored using TargetScan's *ContextScore+* metric and finally filtered by the context score

percentile (50%). The coordinates of each target site were used to overlap this dataset with the existing SNP data.

Even though predicted target sites that share the same seed sequence are collapsed (e.g. targets of different members of the same miRNA family), predicted targets for different miRNA families are still allowed to overlap. While it is not guaranteed that a SNP would not be counted multiple times, the position within the target site will be different, and the normalisation by number of base pairs addresses the biases that are potentially caused by this.

5.5.5 Control Dataset for Target Analysis

The 3' UTRs in the dataset containing at least one predicted miRNA target site were used to build the background dataset. Each UTR was assessed using a sliding window approach. To be coherent with TargetScan predictions, the first 15 bp of each UTR were excluded. After that, a randomised start position was chosen for the first window. This iterative approach then proceeds to find 21 bp windows within the UTR that are not overlapping each other, known annotations or any other predicted target site. The distance between these windows was also randomised, according to a uniform distribution, with values between 1 and 21.

The final dataset is comprised of 24,709 non-overlapping un-annotated 21bp regions, within the 10,330 3' UTR sequences that are part of the dataset. A purpose-built Perl script was used to overlap the SNP data with these regions. These where then analysed using the same procedures as the predicted target sites.

5.5.6 Estimation of SNP Frequencies for Protein-coding Genes

The genome annotations provided by Ensembl were used to establish the coordinates of all mouse protein-coding genes. The Ensembl Perl API was then used to assess the transcripts generated from each gene, determining whether intronic miRNA loci are present. For genes with multiple transcripts, the longest protein-coding transcript was taken as representative.

The SNP were overlapped with this data to create the necessary input files for the analysis. The SNAP Perl implementation (Rodrigo & Learn, 2001) of Nei and Gojobori's algorithm (Nei & Gojobori, 1986), was used to compute rates of synonymous and non-synonymous mutations per base pair.

5.5.7 Analysis of microRNA Variation Throughout Evolutionary Time

The Dollo parsimony approach, was applied to each miRNA family, as previously detailed (Chapter 4). The results were then purged of miRNA families that are not thought to be present in mouse. This provides information regarding in which ancestor of mouse did a particular miRNA family arise. This information was then combined with SNP frequencies and the R framework was used to perform the LOWESS regression.

Chapter 6

Conclusions

The latest technological advances have enabled the acquisition of vast amounts of genomic data. Since the discovery of miRNAs, large efforts have been put in the profiling of the small RNA transcriptome in as many species and biological conditions as possible.

In this thesis, the first aim was to develop methods that would allow the exploration of this wealth of information for evolutionary analysis of metazoan miRNA families. The method development was complemented by a large-scale analysis exploring different types of data to investigate the evolution of miRNAs at multiple evolutionary time-scales, in a fully automated fashion. To maximise their usefulness, the developed methods and respective results were integrated into easy to use online resources that are freely available to the research community in general.

I have presented the MapMi resource, designed to perform automated mapping of miRNA loci across animal genomes, in a species independent way. This research tool has been well accepted by the community. To fully take advantage of its flexibility, it was further developed and integrated into the miRNouveau approach for discovery of novel miRNA loci based on small RNA sequencing.

The dataset produced using the MapMi approach represents a marked improvement over the data available in the miRNA repository, in particular for comparative genomics research.

This enabled a detailed analysis of the evolution of the miRNA repertoire, as well as the detection of patterns of co-evolution between miRNA families and protein coding genes, as well as the detection of fast evolving miRNA families.

The dataset also supported a fully automated search for conservation patterns in the genomic organisation of miRNA-containing synteny blocks, across metazoan evolution. In this analysis I was able to identify different evolutionarily conserved

patterns of miRNA genomic organisation, as well as evidence for the integration of novel, clade specific miRNAs within existing primary miRNA transcripts.

Finally, I analysed patterns of intra-specific genomic variation affecting miRNA loci and their predicted target sites. In this analysis I took particular interest in the selective pressures affecting repeat-associated miRNA families when compared to non repeat-associated miRNA families, discovering several novel evolutionary patterns.

The current investigation was mainly limited by accuracy of the underlying data. The challenges in producing genome assemblies still hampers genomic analysis for some of the species used within this thesis. This is a problem especially for the analysis of repeat-associated miRNAs, as it is difficult to determine the correct position within the genome and the number of copies of these elements. These issues are likely to be mitigated as sequencing technologies advance and new algorithms are developed.

The lack of accurate genome assemblies also hinder a large-scale, multi-species analysis of miRNA targets and function. Most computational target prediction approaches still have an accuracy that is far from optimal, and are dependent on the accurate definition of 3' UTR sequences. Experimental approaches to target prediction will likely help address these issues, but the number of publicly available dataset is still not sufficient to replace computational target prediction.

I believe the integration of the methods and results presented within this thesis into web resources can be useful to the scientific community, allowing the integration of evolutionary information into more specific miRNA projects. Examples are the assessment of homologs for novel miRNA families and the exploration of a particular miRNA family or cluster across species.

The research community is still making sense of the non-coding transcriptome within our cells, with several novel classes of ncRNA having recently been described. Although these analyses focus on miRNAs, there are many parallels with classes of non-coding genes. I hope these approaches will prove useful, once enough information is available to enable large-scale evolutionary studies on other non-coding RNA elements. I have very much enjoyed working in this field and hope that the research described here will prove useful to the community.

Chapter 7

Additional Tables

The following two tables contain extra information about the dataset used in Chapter 3 and Chapter 4 of this thesis. Whilst not essential for their understanding, they are provided as a reference, as they may be needed to provide context about the exact results obtained.

Table 7.1: List of genomes analysed in this study, including assembly name, assembly release date, coverage depth and assembly status. This information was retrieved from the Ensembl public MySQL server.

	Assembly Name	Assembly Date	Coverage Depth	Full Assembly
Acyrthosiphon_pisum	Acyr2	2008-06	high	Unassembled
Aedes_aegypti	AaegL1	2005-10	low	Unassembled
Ailuropoda_melanoleuca	ailMel1	2009-07	high	Unassembled
Anolis_carolinensis	AnoCar2.0	2010-05	high	Unassembled
Anopheles_gambiae	AgamP3	2006-02	high	Unassembled
Apis_mellifera	Amel_2.0	2005-01	low	Unassembled
Bos_taurus	Btau_4.0	2007-10	high	Assembled
Caenorhabditis_brenneri	CB601	2007-07	high	Unassembled
Caenorhabditis_briggsae	CB3	2007-07	high	Assembled
Caenorhabditis_elegans	WS220	2010-10	high	Assembled
Caenorhabditis_japonica	CJ302	2007-07	high	Unassembled
Caenorhabditis_remanei	CR2	2007-07	high	Unassembled
Callithrix_jacchus	C_jacchus3.2.1	2010-01	high	Assembled
Canis_familiaris	CanFam 2.0	2006-05	high	Unassembled

Continued on next page

Cavia_porcellus	cavPor3	2008-03	high	Unassembled
Choloepus_hoffmanni	choHof1	2008-09	low	Unassembled
Ciona_intestinalis	JGI 2	2005-03	high	Assembled
Ciona_savignyi	CSAV 2.0	2005-10	high	Unassembled
Culex_quinquefasciatus	CpipJ1	2007-01	high	Unassembled
Danio_rerio	Zv9	2010-04	high	Unassembled
Daphnia_pulex	Dappu1	2009-05	low	Unassembled
Dasypus_novemcinctus	dasNov2	2008-07	low	Unassembled
Dipodomys_ordii	dipOrd1	2008-07	low	Unassembled
Drosophila_ananassae	dana_r1.3_FB2008_07	2005-08	high	Unassembled
Drosophila_erecta	dere_r1.3_FB2008_07	2005-08	high	Unassembled
Drosophila_grimshawi	dgri_r1.3_FB2008_07	2005-08	high	Unassembled
Drosophila_melanogaster	BDGP 5	2006-04	high	Assembled
Drosophila_mojavensis	dmoj_r1.3_FB2008_07	2005-08	high	Unassembled
Drosophila_persimilis	dper_r1.3_FB2008_07	2005-08	high	Unassembled
Drosophila_pseudoobscura	BCM-HGSC 2.8	2004-11	high	Unassembled
Drosophila_sechellia	dsec_r1.3_FB2008_07	2005-08	high	Unassembled
Drosophila_simulans	dsim_r1.3_FB2008_07	2005-04	high	Unassembled
Drosophila_virilis	dvir_r1.2_FB2008_07	2005-08	high	Unassembled
Drosophila_willistoni	dwil_r1.3_FB2008_07	2005-07	high	Unassembled
Drosophila_yakuba	dyak_r1.3_FB2008_07	2005-11	high	Unassembled
Echinops_telfairi	TENREC	2005-07	low	Unassembled
Equus_caballus	Equ Cab 2	2007-09	high	Unassembled
Erinaceus_europaeus	eriEur1	2006-06	low	Unassembled
Felis_catus	CAT	2006-03	low	Unassembled
Gallus_gallus	WASHUC2	2006-05	high	Assembled
Gasterosteus_aculeatus	BROAD S1	2006-02	high	Unassembled
Gorilla_gorilla	gorGor3	2009-12	low	Unassembled
Homo_sapiens	GRCh37.p3	2009-02	high	Assembled
Ixodes_scapularis	IscaW1	2007-08	high	Unassembled
Loxodonta_africana	Loxafr3.0	2009-07	high	Unassembled
Macaca_mulatta	MMUL 1.0	2006-02	high	Unassembled
Macropus_eugenii	Meug_1.0	2008-12	low	Unassembled
Meleagris_gallopavo	Turkey_2.01	2010-09	high	Unassembled
Microcebus_murinus	micMur1	2007-06	low	Unassembled
Monodelphis_domestica	monDom5	2006-10	high	Unassembled
Mus_musculus	NCBIM37	2007-04	high	Assembled
Myotis_lucifugus	myoLuc1	2006-03	low	Unassembled
Nematostella_vectensis	Nemve1	2007-07	low	Unassembled
Nomascus_leucogenys	Nleu1.0	2010-01	high	Unassembled

Continued on next page

Ochotona_princeps	OchPri2.0	2007-06	low	Unassembled
Ornithorhynchus_anatinus	Ornithorhynchus_anatinus-5.0	2005-12	high	Assembled
Oryctolagus_cuniculus	oryCun2	2009-11	high	Unassembled
Oryzias_latipes	HdrR	2005-10	high	Assembled
Otolemur_garnettii	otoGar1	2006-05	low	Unassembled
Pan_troglodytes	CHIMP2.1	2006-03	high	Assembled
Pediculus_humanus	PhumU1	2008-11	high	Unassembled
Pongo_abelii	PPYG2	2007-09	high	Unassembled
Pristionchus_pacificus	pp1	NA	high	Unassembled
Procavia_capensis	proCap1	2008-07	low	Unassembled
Pteropus_vampyrus	pteVam1	2008-07	low	Unassembled
Rattus_norvegicus	RGSC 3.4	2004-12	high	Assembled
Saccharomyces_cerevisiae	EF 2	2010-02	high	Assembled
Schistosoma_mansoni	sma_v3.1	2008-08	low	Unassembled
Sorex_araneus	sorAra1	2005-10	low	Unassembled
Spermophilus_tridecemlineatus	speTri1	2006-06	low	Unassembled
Strongylocentrotus_purpuratus	Spur2.5	2006-11	low	Unassembled
Sus_scrofa	Sscrofa9	2009-04	high	Assembled
Taeniopygia_guttata	Taeniopygia_guttata-3.2.4	2008-08	high	Assembled
Takifugu_rubripes	FUGU 4.0	2005-06	high	Unassembled
Tarsius_syrichta	tarSyr1	2008-07	low	Unassembled
Tetraodon_nigroviridis	TETRAODON 8.0	2007-03	high	Unassembled
Trichoplax_adhaerens	TRIAD1	2006-08	low	Unassembled
Tupaia_belangeri	tupBel1	2006-06	low	Unassembled
Tursiops_truncatus	turTru1	2008-07	low	Unassembled
Vicugna_pacos	vicPac1	2008-07	low	Unassembled
Xenopus_tropicalis	JGI 4.2	2009-11	high	Unassembled

Table 7.2: Table containing all miRBase miRNA subfamilies under analysis and their corresponding family based on the family attribution procedure presented in Chapter 3.

miRNA Names	miRNA Family	miRNA Names	miRNA Family	miRNA Names	miRNA Family	miRNA Names	miRNA Family
bantam	SF00832	let-7	SF00057	let-7a	SF00057	let-7b	SF00057
let-7c	SF00057	let-7d	SF00057	let-7e	SF00057	let-7f	SF00057
let-7g	SF00057	let-7h	SF00057	let-7i	SF00057	let-7j	SF00057
let-7k	SF00057	lin-4	SF01193	lsy-6	SF01153	mir-1	SF00040
mir-10	SF00096	mir-100	SF00096	mir-1000	SF00217	mir-1001	SF00873
mir-1002	SF00853	mir-1003	SF01331	mir-1004	SF02310	mir-1005	SF01619
mir-1006	SF00665	mir-1007	SF00734	mir-1008	SF00945	mir-1009	SF02104
mir-101	SF00348	mir-1010	SF00196	mir-1011	SF02103	mir-1012	SF01670
mir-1013	SF01315	mir-1014	SF00351	mir-1015	SF00231	mir-1016	SF02298
mir-1017	SF02039	mir-1018	SF00811	mir-1019	SF00577	mir-101a	SF00348
mir-101b	SF00348	mir-101c	SF02227	mir-1020	SF02292	mir-1021	SF02915
mir-1022	SF03037	mir-103	SF00282	mir-103a	SF00282	mir-105	SF00101
mir-105a	SF00101	mir-105b	SF00101	mir-106	SF00038	mir-106a	SF00038
mir-106b	SF00038	mir-107	SF00282	mir-107a	SF00282	mir-107b	SF00282
mir-10a	SF00096	mir-10b	SF00096	mir-10c	SF00096	mir-10d	SF00096
mir-11	SF01300	mir-1174	SF02446	mir-1175	SF01403	mir-1178	SF00756
mir-1179	SF01250	mir-1180	SF01632	mir-1181	SF02555	mir-1182	SF00995
mir-1183	SF01680	mir-1184	SF01406	mir-1185	SF00031	mir-1186	SF00001
mir-1186b	SF00001	mir-1187	SF01502	mir-1188	SF00476	mir-1190	SF02149
mir-1191	SF01509	mir-1192	SF00919	mir-1193	SF00031	mir-1195	SF01170
mir-1196	SF02245	mir-1197	SF00031	mir-1198	SF01908	mir-1199	SF02442
mir-12	SF01761	mir-1200	SF02774	mir-1202	SF02317	mir-1203	SF02443
mir-1204	SF01584	mir-1205	SF01760	mir-1206	SF02153	mir-1207	SF00297
mir-1208	SF01225	mir-122	SF00703	mir-1224	SF00170	mir-1225	SF00024
mir-1226	SF00658	mir-1227	SF00270	mir-1228	SF00795	mir-1229	SF00049
mir-122a	SF00703	mir-122b	SF00699	mir-1230	SF02004	mir-1231	SF00889
mir-1232	SF00640	mir-1233	SF00332	mir-1234	SF01144	mir-1235	SF00601
mir-1236	SF00447	mir-1237	SF00255	mir-1238	SF00333	mir-1239	SF01135
mir-124	SF00120	mir-1240	SF00261	mir-1241	SF00590	mir-1243	SF01722
mir-1244	SF00379	mir-1245	SF01898	mir-1245b	SF01407	mir-1246	SF00289
mir-1247	SF00619	mir-1248	SF01183	mir-1249	SF00449	mir-124a	SF00120
mir-124b	SF00120	mir-124e	SF01656	mir-125	SF00398	mir-1250	SF02548
mir-1251	SF01198	mir-1252	SF02262	mir-1253	SF01798	mir-1254	SF00022
mir-1255a	SF01494	mir-1255b	SF00538	mir-1256	SF00229	mir-1257	SF01422
mir-1258	SF02357	mir-125a	SF00398	mir-125b	SF00398	mir-125c	SF00398
mir-126	SF00574	mir-1260	SF00394	mir-1260b	SF00583	mir-1261	SF00069

Continued on next page...

mir-1262	SF02718
mir-1266	SF02189
mir-1269	SF01147
mir-127	SF00420
mir-1273	SF00022
mir-1273f	SF00022
mir-1277	SF00510
mir-1280	SF00583
mir-1283a	SF00003
mir-1285a	SF00635
mir-1288	SF00218
mir-129	SF00448
mir-1291b	SF01461
mir-1295	SF00174
mir-1299	SF02573
mir-130	SF00001
mir-1302c	SF00069
mir-1304	SF00022
mir-130a	SF00001
mir-1321	SF00179
mir-1325	SF02736
mir-1329	SF01519
mir-1332	SF01306
mir-1336	SF01812
mir-133a	SF00466
mir-134	SF00967
mir-1343	SF00384
mir-1347	SF02597
mir-1350	SF00996
mir-1354	SF02933
mir-1358	SF01263
mir-135c	SF00371
mir-1362	SF00612
mir-1366	SF02383
mir-137	SF00091
mir-1373	SF02691
mir-1377	SF01860
mir-137b	SF00091
mir-1382	SF01872
mir-1386	SF00039
mir-138a	SF00070
mir-1391	SF01975

mir-1263	SF02429
mir-1267	SF02916
mir-1269b	SF01147
mir-1270	SF00475
mir-1273c	SF00022
mir-1273g	SF00022
mir-1278	SF02418
mir-1281	SF00696
mir-1283b	SF00003
mir-1285b	SF00635
mir-1289	SF01203
mir-1290	SF00406
mir-1292	SF00023
mir-1296	SF00579
mir-129a	SF00448
mir-1301	SF01246
mir-1302d	SF00069
mir-1305	SF02689
mir-130b	SF00001
mir-1322	SF00155
mir-1326	SF02716
mir-133	SF00466
mir-1333	SF02796
mir-1337	SF01819
mir-133b	SF00466
mir-1340	SF01095
mir-1344	SF02382
mir-1348	SF00788
mir-1351	SF01069
mir-1355	SF01578
mir-1359	SF02123
mir-136	SF00262
mir-1363	SF01662
mir-1367	SF02188
mir-1370	SF02437
mir-1374	SF01881
mir-1378	SF02856
mir-138	SF00070
mir-1383	SF02216
mir-1387	SF01684
mir-138b	SF00070
mir-1392	SF02196

mir-1264	SF00350
mir-1268	SF00022
mir-126a	SF00574
mir-1271	SF00213
mir-1273d	SF00022
mir-1275	SF00243
mir-1279	SF01532
mir-1282	SF01124
mir-1284	SF01199
mir-1286	SF01042
mir-128a	SF00029
mir-1291	SF01461
mir-1293	SF00793
mir-1297	SF00543
mir-129b	SF00448
mir-1302	SF00069
mir-1302e	SF00069
mir-1306	SF01243
mir-130c	SF00001
mir-1323	SF00324
mir-1327	SF02857
mir-1330	SF02377
mir-1334	SF02380
mir-1338	SF01758
mir-133c	SF00466
mir-1341	SF01175
mir-1345	SF01235
mir-1349	SF02529
mir-1352	SF02595
mir-1356	SF00514
mir-135a	SF00371
mir-1360	SF02204
mir-1364	SF03029
mir-1368	SF02346
mir-1371	SF00410
mir-1375	SF01799
mir-1379	SF01265
mir-1380	SF02672
mir-1384	SF01844
mir-1388	SF00586
mir-139	SF00691
mir-1393	SF01505

mir-1265	SF02095
mir-1268b	SF01579
mir-126b	SF00574
mir-1272	SF01567
mir-1273e	SF00022
mir-1276	SF01336
mir-128	SF00029
mir-1283	SF00003
mir-1285	SF00635
mir-1287	SF00518
mir-128b	SF00029
mir-1291a	SF01461
mir-1294	SF00392
mir-1298	SF01283
mir-13	SF00010
mir-1302b	SF00069
mir-1303	SF00001
mir-1307	SF00817
mir-132	SF00498
mir-1324	SF01029
mir-1328	SF01370
mir-1331	SF02652
mir-1335	SF00396
mir-1339	SF02453
mir-133d	SF00466
mir-1342	SF02913
mir-1346	SF00071
mir-135	SF00371
mir-1353	SF00020
mir-1357	SF00971
mir-135b	SF00371
mir-1361	SF01435
mir-1365	SF00515
mir-1369	SF00066
mir-1372	SF00754
mir-1376	SF01650
mir-137a	SF00091
mir-1381	SF02615
mir-1385	SF00105
mir-1389	SF01988
mir-1390	SF01190
mir-1394	SF01865

Continued on next page...

mir-1395	SF01664
mir-1399	SF00126
mir-140	SF00465
mir-1403	SF01497
mir-1407	SF01595
mir-1410	SF02912
mir-1414	SF00741
mir-1418	SF02919
mir-1419d	SF00708
mir-142	SF00377
mir-1420d	SF00286
mir-1421a	SF00045
mir-1421ad	SF00045
mir-1421ah	SF00045
mir-1421al	SF00045
mir-1421d	SF00045
mir-1421h	SF00045
mir-1421l	SF00045
mir-1421p	SF00045
mir-1421t	SF00045
mir-1421x	SF00045
mir-1422b	SF00109
mir-1422f	SF00109
mir-1422j	SF00109
mir-1422n	SF00109
mir-142a	SF00377
mir-144	SF00672
mir-1453	SF00128
mir-1458	SF02948
mir-1461	SF01034
mir-1465	SF00136
mir-1469	SF02618
mir-147	SF00338
mir-147a	SF00338
mir-148a	SF00167
mir-15	SF02635
mir-1502c	SF01452
mir-151b	SF00006
mir-1538	SF01515
mir-153c	SF00337
mir-1542	SF01467
mir-1546	SF01357

mir-1396	SF02452
mir-13a	SF00010
mir-1400	SF02202
mir-1404	SF02599
mir-1408	SF00061
mir-1411	SF00718
mir-1415	SF01484
mir-1419a	SF00708
mir-1419e	SF00708
mir-1420a	SF00286
mir-1420e	SF00286
mir-1421aa	SF00045
mir-1421ae	SF00045
mir-1421ai	SF00045
mir-1421am	SF00045
mir-1421e	SF00045
mir-1421i	SF00045
mir-1421m	SF00045
mir-1421q	SF00045
mir-1421u	SF00045
mir-1421y	SF00045
mir-1422c	SF00109
mir-1422g	SF00109
mir-1422k	SF00109
mir-1422o	SF00109
mir-142b	SF00377
mir-145	SF00178
mir-1454	SF00930
mir-1459	SF02158
mir-1462	SF01104
mir-1466	SF02401
mir-146a	SF00221
mir-1470	SF01321
mir-147b	SF00338
mir-148b	SF00167
mir-150	SF01109
mir-1502d	SF01232
mir-152	SF00167
mir-1539	SF02388
mir-154	SF00031
mir-1543	SF00419
mir-1547	SF00997

mir-1397	SF00436
mir-13b	SF00010
mir-1401	SF03049
mir-1405	SF02288
mir-1409	SF01945
mir-1412	SF00198
mir-1416	SF02601
mir-1419b	SF00708
mir-1419f	SF00708
mir-1420b	SF00286
mir-1420f	SF00286
mir-1421ab	SF00045
mir-1421af	SF00045
mir-1421aj	SF00045
mir-1421b	SF00045
mir-1421f	SF00045
mir-1421j	SF00045
mir-1421n	SF00045
mir-1421r	SF00045
mir-1421v	SF00045
mir-1421z	SF00045
mir-1422d	SF00109
mir-1422h	SF00108
mir-1422l	SF00109
mir-1422p	SF00109
mir-143	SF00336
mir-1451	SF01825
mir-1456	SF02042
mir-146	SF00221
mir-1463	SF00462
mir-1467	SF01830
mir-146b	SF00221
mir-1471	SF00517
mir-148	SF00167
mir-149	SF00041
mir-1502a	SF02327
mir-1504	SF01141
mir-153	SF00337
mir-153a	SF00337
mir-1540	SF01816
mir-1544	SF01010
mir-1548	SF03009

mir-1398	SF02894
mir-14	SF00032
mir-1402	SF02018
mir-1406	SF01552
mir-141	SF00226
mir-1413	SF02740
mir-1417	SF02712
mir-1419c	SF00708
mir-1419g	SF00708
mir-1420c	SF00286
mir-1420g	SF00286
mir-1421ac	SF00045
mir-1421ag	SF00045
mir-1421ak	SF00045
mir-1421c	SF00045
mir-1421g	SF00045
mir-1421k	SF00045
mir-1421o	SF00045
mir-1421s	SF00045
mir-1421w	SF00045
mir-1422a	SF00109
mir-1422e	SF00108
mir-1422i	SF00109
mir-1422m	SF00109
mir-1422q	SF00109
mir-1434	SF01262
mir-1452	SF02922
mir-1457	SF01472
mir-1460	SF02566
mir-1464	SF02517
mir-1468	SF00736
mir-146c	SF00221
mir-1473	SF02331
mir-1487	SF02372
mir-1497	SF00425
mir-1502b	SF01452
mir-151	SF00006
mir-1537	SF00948
mir-153b	SF00337
mir-1541	SF01966
mir-1545	SF02501
mir-1549	SF03014

Continued on next page...

mir-154a	SF00031
mir-1551	SF01415
mir-1555	SF02629
mir-1559	SF02013
mir-1563	SF03043
mir-1567	SF02239
mir-1571	SF00993
mir-1575	SF02134
mir-1579	SF02425
mir-1583	SF01035
mir-1587	SF00090
mir-1591	SF01935
mir-1595	SF01893
mir-1599	SF02964
mir-16	SF00150
mir-1603	SF00018
mir-1607	SF02148
mir-1611	SF02208
mir-1615	SF01719
mir-1619	SF01849
mir-1623	SF00861
mir-1627	SF02523
mir-1631	SF01922
mir-1635	SF00149
mir-1639	SF00152
mir-1643	SF01733
mir-1647	SF02128
mir-1651	SF02169
mir-1655	SF01353
mir-1659	SF02980
mir-1663	SF01688
mir-1667	SF02237
mir-1671	SF02396
mir-1675	SF02535
mir-1679	SF02318
mir-1683	SF01333
mir-1687	SF01044
mir-1691	SF02858
mir-1695	SF02874
mir-1699	SF03032
mir-17	SF00038
mir-1703	SF02592

mir-154b	SF00031
mir-1552	SF01057
mir-1556	SF02334
mir-1560	SF02826
mir-1564	SF02402
mir-1568	SF02817
mir-1572	SF02099
mir-1576	SF02353
mir-1580	SF01925
mir-1584	SF00158
mir-1588	SF00587
mir-1592	SF01562
mir-1596	SF00495
mir-15a	SF00084
mir-1600	SF01479
mir-1604	SF01969
mir-1608	SF02893
mir-1612	SF00113
mir-1616	SF03026
mir-1620	SF00692
mir-1624	SF01679
mir-1628	SF01295
mir-1632	SF03034
mir-1636	SF01329
mir-1640	SF01820
mir-1644	SF02813
mir-1648	SF02697
mir-1652	SF01215
mir-1656	SF02907
mir-1660	SF02743
mir-1664	SF01933
mir-1668	SF01111
mir-1672	SF02052
mir-1676	SF00803
mir-1680	SF01266
mir-1684	SF02074
mir-1688	SF01245
mir-1692	SF01196
mir-1696	SF01735
mir-16a	SF00150
mir-1700	SF02551
mir-1704	SF02634

mir-155	SF00102
mir-1553	SF00104
mir-1557	SF02859
mir-1561	SF00715
mir-1565	SF02361
mir-1569	SF01606
mir-1573	SF01339
mir-1577	SF02490
mir-1581	SF02161
mir-1585	SF01061
mir-1589	SF01989
mir-1593	SF02393
mir-1597	SF00748
mir-15b	SF00084
mir-1601	SF02935
mir-1605	SF02887
mir-1609	SF02121
mir-1613	SF02687
mir-1617	SF02900
mir-1621	SF01046
mir-1625	SF02834
mir-1629	SF01570
mir-1633	SF00720
mir-1637	SF02646
mir-1641	SF00886
mir-1645	SF02313
mir-1649	SF02145
mir-1653	SF00834
mir-1657	SF02336
mir-1661	SF01742
mir-1665	SF02014
mir-1669	SF00620
mir-1673	SF02696
mir-1677	SF00692
mir-1681	SF01715
mir-1685	SF01012
mir-1689	SF01748
mir-1693	SF01050
mir-1697	SF02051
mir-16b	SF00150
mir-1701	SF02534
mir-1705	SF01951

mir-1550	SF02970
mir-1554	SF00553
mir-1558	SF00653
mir-1562	SF00711
mir-1566	SF00124
mir-1570	SF01706
mir-1574	SF00172
mir-1578	SF01942
mir-1582	SF02560
mir-1586	SF01589
mir-1590	SF00938
mir-1594	SF01292
mir-1598	SF02454
mir-15c	SF00084
mir-1602	SF01167
mir-1606	SF02275
mir-1610	SF02498
mir-1614	SF02254
mir-1618	SF01829
mir-1622	SF01807
mir-1626	SF02240
mir-1630	SF00480
mir-1634	SF01648
mir-1638	SF01957
mir-1642	SF00558
mir-1646	SF01683
mir-1650	SF02124
mir-1654	SF02983
mir-1658	SF02781
mir-1662	SF00841
mir-1666	SF02430
mir-1670	SF02612
mir-1674	SF00728
mir-1678	SF01786
mir-1682	SF02538
mir-1686	SF02156
mir-1690	SF00358
mir-1694	SF00804
mir-1698	SF01354
mir-16c	SF00150
mir-1702	SF01997
mir-1706	SF01927

Continued on next page...

mir-1707	SF02582
mir-1711	SF02085
mir-1715	SF02029
mir-1719	SF02943
mir-1723	SF02662
mir-1727	SF02475
mir-1731	SF01074
mir-1735	SF02304
mir-1739	SF02642
mir-1743	SF01817
mir-1747	SF01771
mir-1751	SF01834
mir-1755	SF01006
mir-1758	SF02464
mir-1762	SF02082
mir-1766	SF03041
mir-1770	SF00907
mir-1774	SF02664
mir-1777a	SF00849
mir-1780	SF00827
mir-1784	SF00693
mir-1788	SF01004
mir-1792	SF02283
mir-1796	SF02889
mir-17a	SF00038
mir-1802	SF00900
mir-1806	SF02669
mir-181	SF02753
mir-1814	SF00237
mir-1815	SF00796
mir-1819	SF02122
mir-181d	SF00374
mir-1822	SF01631
mir-1827	SF01200
mir-1829c	SF00145
mir-1832b	SF02571
mir-1836	SF02764
mir-184	SF00618
mir-1843	SF00891
mir-184a	SF00618
mir-187	SF01218
mir-1890	SF00281

mir-1708	SF02488
mir-1712	SF02788
mir-1716	SF01634
mir-1720	SF02850
mir-1724	SF02143
mir-1728	SF02093
mir-1732	SF01700
mir-1736	SF01361
mir-1740	SF00607
mir-1744	SF01737
mir-1748	SF00078
mir-1752	SF01430
mir-1756a	SF01114
mir-1759	SF02503
mir-1763	SF01823
mir-1767	SF00479
mir-1771	SF01398
mir-1775	SF03035
mir-1777b	SF00677
mir-1781	SF01947
mir-1785	SF02135
mir-1789	SF02944
mir-1793	SF01864
mir-1797	SF02364
mir-18	SF00112
mir-1803	SF00692
mir-1807	SF02399
mir-1811	SF03050
mir-1814a	SF00122
mir-1816	SF02033
mir-181a	SF00374
mir-182	SF00107
mir-1823	SF02241
mir-1828	SF02663
mir-183	SF00107
mir-1833	SF01036
mir-1837	SF00927
mir-1840	SF02727
mir-1843b	SF00891
mir-184b	SF00618
mir-188	SF00082
mir-1891	SF02604

mir-1709	SF02048
mir-1713	SF02436
mir-1717	SF01707
mir-1721	SF01096
mir-1725	SF02924
mir-1729	SF01088
mir-1733	SF01681
mir-1737	SF02960
mir-1741	SF01294
mir-1745	SF01995
mir-1749	SF01779
mir-1753	SF02067
mir-1756b	SF01114
mir-1760	SF02287
mir-1764	SF00765
mir-1768	SF00923
mir-1772	SF02777
mir-1776	SF01520
mir-1778	SF01861
mir-1782	SF01911
mir-1786	SF02502
mir-1790	SF01290
mir-1794	SF02833
mir-1798	SF03006
mir-1800	SF02854
mir-1804	SF01544
mir-1808	SF02839
mir-1812	SF02819
mir-1814b	SF01738
mir-1817	SF01125
mir-181b	SF00374
mir-1820	SF02459
mir-1824	SF01646
mir-1829a	SF00145
mir-1830	SF02759
mir-1834	SF01875
mir-1838	SF02088
mir-1841	SF03031
mir-1844	SF03044
mir-185	SF00418
mir-1889	SF01626
mir-1892	SF00545

mir-1710	SF01576
mir-1714	SF01360
mir-1718	SF02337
mir-1722	SF02990
mir-1726	SF01853
mir-1730	SF01636
mir-1734	SF02190
mir-1738	SF02276
mir-1742	SF01238
mir-1746	SF02409
mir-1750	SF02609
mir-1754	SF02221
mir-1757	SF00378
mir-1761	SF02117
mir-1765	SF01309
mir-1769	SF00478
mir-1773	SF01496
mir-1777	SF02265
mir-1779	SF01939
mir-1783	SF01762
mir-1787	SF02491
mir-1791	SF01588
mir-1795	SF01068
mir-1799	SF02969
mir-1801	SF01189
mir-1805	SF00572
mir-1809	SF01404
mir-1813	SF00140
mir-1814c	SF00211
mir-1818	SF02890
mir-181c	SF00374
mir-1821	SF01961
mir-1825	SF00771
mir-1829b	SF00145
mir-1832	SF02572
mir-1835	SF02440
mir-1839	SF00369
mir-1842	SF00242
mir-1845	SF00156
mir-186	SF01054
mir-189	SF00222
mir-1893	SF01674

Continued on next page...

mir-1894	SF01668	mir-1895	SF02800	mir-1896	SF01431	mir-1897	SF01527
mir-1898	SF01621	mir-1899	SF00215	mir-18a	SF00112	mir-18b	SF00112
mir-18c	SF00112	mir-190	SF00073	mir-1900	SF02312	mir-1901	SF01973
mir-1902	SF00187	mir-1903	SF02676	mir-1904	SF03027	mir-1905	SF00125
mir-1905a	SF00125	mir-1905b	SF00125	mir-1905c	SF00125	mir-1906	SF01158
mir-1907	SF01132	mir-1908	SF00729	mir-1909	SF02414	mir-190a	SF00073
mir-190b	SF00073	mir-191	SF00542	mir-1910	SF02951	mir-1911	SF00568
mir-1912	SF01217	mir-1913	SF02616	mir-1914	SF00397	mir-1915	SF00244
mir-192	SF00434	mir-1923	SF03048	mir-1927	SF01985	mir-1928	SF00307
mir-1929	SF02069	mir-193	SF00162	mir-1930	SF00885	mir-1931	SF01525
mir-1932	SF00656	mir-1933	SF01914	mir-1934	SF02186	mir-1935	SF00961
mir-1936	SF02750	mir-1938	SF00389	mir-193a	SF00162	mir-193b	SF00162
mir-194	SF00254	mir-1940	SF00088	mir-1941	SF01434	mir-1942	SF02965
mir-1943	SF02938	mir-1945	SF00119	mir-1946a	SF01930	mir-1946b	SF02905
mir-1947	SF01425	mir-1948	SF01880	mir-1949	SF00916	mir-194a	SF00254
mir-194b	SF00254	mir-195	SF00150	mir-1950	SF02719	mir-1951	SF02748
mir-1952	SF00563	mir-1953	SF02340	mir-1954	SF00913	mir-1955	SF00359
mir-1956	SF02201	mir-1957	SF02333	mir-1958	SF02481	mir-196	SF00064
mir-1960	SF00661	mir-1961	SF00186	mir-1962	SF02828	mir-1963	SF02378
mir-1964	SF02024	mir-1965	SF02330	mir-1966	SF01454	mir-1967	SF01206
mir-1968	SF02820	mir-1969	SF02908	mir-196a	SF00064	mir-196b	SF00064
mir-196c	SF00064	mir-196d	SF00064	mir-197	SF01001	mir-1970	SF00893
mir-1971	SF00970	mir-1972	SF00001	mir-1973	SF00481	mir-1976	SF00634
mir-198	SF01753	mir-1981	SF02829	mir-1982	SF00602	mir-1983	SF01916
mir-199	SF00310	mir-1993	SF01251	mir-199a	SF00310	mir-199b	SF00310
mir-199c	SF00310	mir-19a	SF00036	mir-19b	SF00036	mir-19c	SF00036
mir-19d	SF00036	mir-1a	SF00040	mir-1b	SF00040	mir-1c	SF01399
mir-2	SF00010	mir-20	SF00038	mir-200	SF00226	mir-2001	SF01874
mir-2002	SF00405	mir-2003	SF01594	mir-2004	SF01244	mir-2005	SF01842
mir-2006	SF00456	mir-2007	SF02271	mir-2008	SF01744	mir-2009	SF00881
mir-200a	SF00226	mir-200b	SF00226	mir-200c	SF00226	mir-201	SF00503
mir-2010	SF02243	mir-2011	SF02279	mir-2012	SF01091	mir-2013	SF01877
mir-202	SF00604	mir-2022	SF01788	mir-2023	SF02868	mir-2024a	SF00428
mir-2024b	SF00428	mir-2024c	SF00428	mir-2024e	SF00428	mir-2024f	SF00428
mir-2024g	SF00428	mir-2025	SF02415	mir-2026	SF02407	mir-2027	SF02172
mir-2028	SF02584	mir-2029	SF00738	mir-203	SF00326	mir-2030	SF02360
mir-2031	SF01593	mir-2032a	SF01892	mir-2032b	SF01892	mir-2033	SF01169
mir-2034	SF02473	mir-2035	SF02524	mir-2036	SF02607	mir-2037	SF02673
mir-2038	SF01850	mir-2039	SF02653	mir-203a	SF00326	mir-203b	SF00326
mir-204	SF00416	mir-2040a	SF00644	mir-2040b	SF00173	mir-2041	SF00920
mir-2042	SF02985	mir-2043a	SF00687	mir-2043b	SF00687	mir-2044	SF02115
mir-2045	SF01481	mir-2046	SF02694	mir-2047	SF00301	mir-2048	SF01894

Continued on next page...

mir-2049	SF01572
mir-2050	SF00268
mir-2054	SF00901
mir-207	SF00828
mir-208b	SF00342
mir-210	SF00598
mir-2113	SF00194
mir-2117	SF00139
mir-2128	SF02458
mir-2136	SF02706
mir-2147b	SF02505
mir-2162	SF01458
mir-216c	SF01034
mir-2183	SF00582
mir-2187	SF00034
mir-218b	SF00589
mir-2192	SF02537
mir-2196	SF02214
mir-2207	SF02527
mir-2209b	SF00576
mir-2211	SF02675
mir-2215	SF01558
mir-2218b	SF02991
mir-2221	SF02598
mir-2225	SF02918
mir-2229	SF01098
mir-2230	SF01854
mir-2234a	SF02568
mir-2236b	SF02545
mir-2238a	SF00700
mir-2238e	SF00700
mir-2240b	SF00313
mir-2241c	SF00313
mir-2245	SF03018
mir-2249	SF02479
mir-2253a	SF00799
mir-2256	SF00346
mir-2261	SF01188
mir-2265	SF02166
mir-2269	SF02928
mir-2272	SF02344
mir-2277	SF01093

mir-204a	SF00416
mir-2051	SF01159
mir-205a	SF00791
mir-2073	SF01944
mir-20a	SF00038
mir-210b	SF02450
mir-2114	SF01727
mir-212	SF00472
mir-2129	SF02637
mir-2137	SF03047
mir-215	SF00434
mir-2169	SF00733
mir-217	SF00315
mir-2184	SF00239
mir-2188	SF00195
mir-219	SF00042
mir-2193	SF00709
mir-2197	SF01438
mir-2208a	SF01059
mir-2209c	SF00576
mir-2212	SF02758
mir-2216	SF02955
mir-2219	SF02982
mir-2222	SF01334
mir-2226	SF01299
mir-222a	SF00522
mir-2231	SF01693
mir-2234b	SF02400
mir-2237a	SF00564
mir-2238b	SF00700
mir-2239	SF02888
mir-2240c	SF00313
mir-2242	SF02489
mir-2246	SF02293
mir-2250	SF02782
mir-2253b	SF01665
mir-2258	SF01653
mir-2262	SF02422
mir-2266	SF02455
mir-227	SF00950
mir-2273	SF01097
mir-2278	SF01501

mir-204b	SF00416
mir-2052	SF00594
mir-205b	SF00791
mir-208	SF00342
mir-20b	SF00038
mir-211	SF00416
mir-2115	SF00998
mir-2126	SF01078
mir-2130	SF02947
mir-2139	SF01065
mir-2159	SF01873
mir-216a	SF00181
mir-2176	SF02398
mir-2185	SF01291
mir-2189	SF02496
mir-2190	SF02745
mir-2194	SF02030
mir-2198	SF01712
mir-2208b	SF01059
mir-221	SF00308
mir-2213	SF02961
mir-2217	SF02590
mir-222	SF00522
mir-2223	SF02474
mir-2227	SF01858
mir-222b	SF00522
mir-2232	SF00454
mir-2235	SF01298
mir-2237b	SF00564
mir-2238c	SF00700
mir-224	SF00318
mir-2241a	SF00313
mir-2243	SF00894
mir-2247	SF00875
mir-2251	SF00313
mir-2254	SF03038
mir-2259	SF01627
mir-2263	SF00313
mir-2267	SF02715
mir-2270	SF01081
mir-2274	SF01417
mir-2279	SF01442

mir-205	SF00791
mir-2053	SF01960
mir-206	SF00040
mir-208a	SF00342
mir-21	SF01015
mir-2110	SF02884
mir-2116	SF02910
mir-2127	SF01146
mir-2131	SF00973
mir-214	SF00116
mir-216	SF00181
mir-216b	SF01034
mir-218	SF00589
mir-2186	SF00956
mir-218a	SF00589
mir-2191	SF02392
mir-2195	SF02280
mir-22	SF00701
mir-2209a	SF00576
mir-2210	SF03010
mir-2214	SF01269
mir-2218a	SF02807
mir-2220	SF01379
mir-2224	SF00431
mir-2228	SF02305
mir-223	SF00974
mir-2233	SF00028
mir-2236a	SF02545
mir-2237c	SF00564
mir-2238d	SF00700
mir-2240a	SF00313
mir-2241b	SF00313
mir-2244	SF03039
mir-2248	SF02660
mir-2252	SF02771
mir-2255	SF02613
mir-2260	SF02765
mir-2264	SF01923
mir-2268	SF01052
mir-2271	SF00177
mir-2276	SF01752
mir-228	SF01592

Continued on next page...

mir-2280	SF00714
mir-2284a	SF00030
mir-2284e	SF00030
mir-2284i	SF00030
mir-2284n	SF00030
mir-2284r	SF00030
mir-2284v	SF00030
mir-2285b	SF00030
mir-2287	SF01536
mir-2290	SF01446
mir-2294	SF02741
mir-2298	SF01166
mir-23	SF00205
mir-2301	SF02547
mir-2305	SF00464
mir-2309	SF01754
mir-2312	SF00030
mir-2316	SF02036
mir-232	SF02339
mir-2323	SF00595
mir-2325c	SF01179
mir-2329	SF02941
mir-2332	SF00636
mir-2336	SF02699
mir-234	SF02066
mir-2343	SF02080
mir-2347	SF02034
mir-2350	SF02968
mir-2354	SF01040
mir-2358	SF02767
mir-2360	SF02408
mir-2364	SF01840
mir-2368	SF01628
mir-2371	SF00531
mir-2375	SF02665
mir-2379	SF02521
mir-2382	SF02862
mir-2386	SF01934
mir-239	SF02623
mir-2393	SF00725
mir-2397	SF02165
mir-239b	SF00609

mir-2281	SF01780
mir-2284b	SF00030
mir-2284f	SF00030
mir-2284k	SF00030
mir-2284o	SF00030
mir-2284s	SF00030
mir-2284w	SF00030
mir-2285c	SF00030
mir-2288	SF00717
mir-2291	SF02805
mir-2295	SF00646
mir-2299	SF01642
mir-230	SF01899
mir-2302	SF00501
mir-2306	SF00164
mir-231	SF01599
mir-2313	SF01763
mir-2317	SF00241
mir-2320	SF01896
mir-2324	SF02531
mir-2326	SF01905
mir-233	SF01022
mir-2333	SF01815
mir-2337	SF01981
mir-2340	SF02860
mir-2344	SF02844
mir-2348	SF02155
mir-2351	SF02285
mir-2355	SF00716
mir-2359	SF00757
mir-2361	SF01833
mir-2365	SF01390
mir-2369	SF02806
mir-2372	SF02667
mir-2376	SF01161
mir-238	SF01062
mir-2383	SF02513
mir-2387	SF00760
mir-2390	SF01343
mir-2394	SF00921
mir-2398	SF02574
mir-23a	SF00205

mir-2282	SF01337
mir-2284c	SF00030
mir-2284g	SF00030
mir-2284l	SF00030
mir-2284p	SF00030
mir-2284t	SF00030
mir-2284x	SF00030
mir-2285d	SF00030
mir-2289	SF01967
mir-2292	SF03017
mir-2296	SF02405
mir-22a	SF00701
mir-2300a	SF02877
mir-2303	SF01258
mir-2307	SF02926
mir-2310	SF01565
mir-2314	SF02441
mir-2318	SF01764
mir-2321	SF01257
mir-2325a	SF00191
mir-2327	SF02873
mir-2330	SF01598
mir-2334	SF02832
mir-2338	SF02881
mir-2341	SF02633
mir-2345	SF02223
mir-2349	SF02785
mir-2352	SF02139
mir-2356	SF02460
mir-235b	SF00683
mir-2362	SF02366
mir-2366	SF02150
mir-237	SF00882
mir-2373	SF01573
mir-2377	SF01314
mir-2380	SF01685
mir-2384	SF01247
mir-2388	SF01932
mir-2391	SF01952
mir-2395	SF01087
mir-2399	SF02603
mir-23b	SF00205

mir-2283	SF02703
mir-2284d	SF00030
mir-2284h	SF00030
mir-2284m	SF00030
mir-2284q	SF00030
mir-2284u	SF01366
mir-2285a	SF00030
mir-2286	SF02852
mir-229	SF02068
mir-2293	SF00561
mir-2297	SF02620
mir-22b	SF00701
mir-2300b	SF03052
mir-2304	SF00190
mir-2308	SF02878
mir-2311	SF02037
mir-2315	SF01768
mir-2319b	SF02875
mir-2322	SF02448
mir-2325b	SF00797
mir-2328	SF02270
mir-2331	SF02821
mir-2335	SF02929
mir-2339	SF01660
mir-2342	SF02880
mir-2346	SF01801
mir-235	SF00683
mir-2353	SF01659
mir-2357	SF02348
mir-236	SF00404
mir-2363	SF02668
mir-2367	SF02825
mir-2370	SF00951
mir-2374	SF00768
mir-2378	SF00663
mir-2381	SF02320
mir-2385	SF02899
mir-2389	SF02851
mir-2392	SF01781
mir-2396	SF02512
mir-239a	SF00189
mir-23c	SF01906

Continued on next page...

mir-24	SF00222
mir-2402	SF02325
mir-2406	SF02717
mir-241	SF00085
mir-2413	SF02563
mir-2417	SF02776
mir-2420	SF00230
mir-2424	SF00922
mir-2428	SF01639
mir-2431	SF02552
mir-2435	SF00030
mir-2439	SF02974
mir-2442	SF00489
mir-2446	SF01534
mir-245	SF01661
mir-2451	SF02044
mir-2455	SF02747
mir-2459	SF03028
mir-2462	SF02946
mir-2466	SF02230
mir-247	SF01172
mir-2473	SF01401
mir-2477	SF01453
mir-2480	SF02690
mir-2484	SF02923
mir-2488	SF02779
mir-2491	SF02019
mir-2495	SF00520
mir-2499	SF02244
mir-250	SF01413
mir-2503	SF02211
mir-2507a	SF00414
mir-251	SF01475
mir-2513b	SF01429
mir-2517a	SF01859
mir-252	SF00014
mir-2522b	SF01213
mir-2526	SF00740
mir-252a	SF00014
mir-2531	SF00637
mir-2535	SF02059
mir-2538	SF01831

mir-240	SF00952
mir-2403	SF00548
mir-2407	SF01368
mir-2410	SF00596
mir-2414	SF01284
mir-2418	SF02798
mir-2421	SF00490
mir-2425	SF01669
mir-2429	SF02222
mir-2432	SF00896
mir-2436	SF02397
mir-244	SF00353
mir-2443	SF02972
mir-2447	SF02218
mir-2450a	SF00937
mir-2452	SF01832
mir-2456	SF03016
mir-246	SF02225
mir-2463	SF02570
mir-2467	SF01654
mir-2470	SF00203
mir-2474	SF02999
mir-2478	SF00360
mir-2481	SF02870
mir-2485	SF01470
mir-2489	SF02608
mir-2492	SF02546
mir-2496	SF00621
mir-24a	SF00222
mir-2500	SF01751
mir-2504	SF02090
mir-2507b	SF00983
mir-2510	SF02931
mir-2514	SF02040
mir-2517b	SF01859
mir-2520	SF02802
mir-2523	SF02714
mir-2527	SF01550
mir-252b	SF01556
mir-2532	SF01610
mir-2535b	SF02059
mir-2539	SF01039

mir-2400	SF00761
mir-2404	SF01089
mir-2408	SF02754
mir-2411	SF00899
mir-2415	SF02945
mir-2419	SF00884
mir-2422	SF02532
mir-2426	SF00940
mir-243	SF01041
mir-2433	SF00269
mir-2437	SF00685
mir-2440	SF02307
mir-2444	SF00355
mir-2448	SF02711
mir-2450b	SF00937
mir-2453	SF02061
mir-2457	SF01557
mir-2460	SF02049
mir-2464	SF02212
mir-2468	SF02865
mir-2471	SF02803
mir-2475	SF02958
mir-2479	SF01607
mir-2482	SF01530
mir-2486	SF02530
mir-249	SF01878
mir-2493	SF01630
mir-2497	SF02761
mir-24b	SF00222
mir-2501	SF02724
mir-2505	SF01242
mir-2508	SF01449
mir-2511	SF00695
mir-2515	SF00654
mir-2518	SF00826
mir-2521	SF02688
mir-2524	SF01301
mir-2528	SF01734
mir-253	SF00688
mir-2533	SF00910
mir-2536	SF01837
mir-254	SF02192

mir-2401	SF01273
mir-2405	SF02142
mir-2409	SF02866
mir-2412	SF01332
mir-2416	SF00445
mir-242	SF01671
mir-2423	SF02057
mir-2427	SF02522
mir-2430	SF02260
mir-2434	SF01101
mir-2438	SF02238
mir-2441	SF03013
mir-2445	SF00684
mir-2449	SF00928
mir-2450c	SF00937
mir-2454	SF01092
mir-2458	SF02914
mir-2461	SF03000
mir-2465	SF02259
mir-2469	SF03030
mir-2472	SF01254
mir-2476	SF00135
mir-248	SF02194
mir-2483	SF00762
mir-2487	SF00815
mir-2490	SF02977
mir-2494	SF01223
mir-2498	SF01740
mir-25	SF00421
mir-2502	SF01569
mir-2506	SF01915
mir-2509	SF00749
mir-2513a	SF01429
mir-2516	SF00340
mir-2519	SF01677
mir-2522a	SF01213
mir-2525	SF02273
mir-2529	SF01568
mir-2530	SF02641
mir-2534	SF02352
mir-2537	SF02463
mir-2540	SF01749

Continued on next page...

mir-2541	SF01389
mir-2544	SF01392
mir-2547	SF02413
mir-2550	SF01540
mir-2554	SF02209
mir-2558	SF01116
mir-2561	SF00737
mir-2565	SF01673
mir-2567b	SF00872
mir-2569	SF01289
mir-2572	SF02264
mir-2575	SF02027
mir-2579	SF01613
mir-2582a	SF02419
mir-259	SF01585
mir-262	SF02515
mir-264	SF01615
mir-268	SF01476
mir-26a	SF00357
mir-270	SF02297
mir-273	SF01018
mir-276	SF00570
mir-276c	SF00402
mir-2778b	SF02836
mir-2790	SF00365
mir-279c	SF01032
mir-27d	SF00079
mir-281	SF00606
mir-285	SF00076
mir-286b	SF00946
mir-2882	SF00056
mir-2887	SF01239
mir-2890	SF01940
mir-2894	SF02917
mir-2898	SF01416
mir-29	SF00076
mir-2902	SF02959
mir-2917	SF00106
mir-293	SF00003
mir-2942	SF00839
mir-2944b	SF00840
mir-2951	SF01286

mir-2542	SF02349
mir-2545a	SF01432
mir-2548	SF02315
mir-2551	SF02083
mir-2555	SF01308
mir-2559	SF02309
mir-2562	SF02078
mir-2566a	SF01511
mir-2567c	SF00871
mir-257	SF02733
mir-2573	SF00223
mir-2576	SF00132
mir-258	SF02812
mir-2582b	SF02906
mir-26	SF00357
mir-263	SF01013
mir-265	SF02569
mir-2681	SF01690
mir-26b	SF00357
mir-271	SF02579
mir-274	SF01367
mir-2765	SF00752
mir-277	SF00236
mir-278	SF00844
mir-2796	SF01805
mir-27a	SF00079
mir-27e	SF00079
mir-282	SF01388
mir-286	SF00946
mir-287	SF02219
mir-2883	SF00131
mir-2888	SF00059
mir-2891	SF01697
mir-2895	SF02901
mir-2899	SF01512
mir-290	SF00003
mir-2903	SF00486
mir-291a	SF00003
mir-294	SF00003
mir-2943	SF02007
mir-2945	SF01649
mir-2952	SF02426

mir-2543a	SF01302
mir-2545b	SF01432
mir-2549	SF02102
mir-2552	SF01322
mir-2556	SF02026
mir-256	SF01583
mir-2563	SF02381
mir-2566b	SF01511
mir-2568a	SF01048
mir-2570	SF01726
mir-2574a	SF02587
mir-2577	SF01698
mir-2580	SF00427
mir-2583	SF00250
mir-260	SF02925
mir-263a	SF01013
mir-266	SF00669
mir-2682	SF01547
mir-26c	SF00357
mir-272	SF00585
mir-2742	SF00622
mir-276a	SF00570
mir-2777	SF01293
mir-2788	SF01523
mir-279a	SF00676
mir-27b	SF00079
mir-28	SF00006
mir-283	SF01827
mir-2861	SF01555
mir-288	SF00197
mir-2885	SF01554
mir-2889	SF01204
mir-2892	SF02217
mir-2896	SF02882
mir-28b	SF00865
mir-2900	SF02449
mir-2904	SF01882
mir-291b	SF00003
mir-2940	SF01456
mir-2944	SF00840
mir-2946	SF01318
mir-2953	SF02020

mir-2543b	SF01302
mir-2546	SF02585
mir-255	SF01184
mir-2553	SF01624
mir-2557	SF02861
mir-2560	SF01226
mir-2564	SF01460
mir-2567a	SF00872
mir-2568b	SF01048
mir-2571	SF02055
mir-2574b	SF02587
mir-2578	SF02424
mir-2581	SF00537
mir-2584	SF02827
mir-261	SF01928
mir-263b	SF00107
mir-267	SF01186
mir-269	SF00669
mir-27	SF00079
mir-2723	SF02164
mir-275	SF00460
mir-276b	SF00570
mir-2778a	SF00897
mir-279	SF00782
mir-279b	SF00676
mir-27c	SF00079
mir-280	SF01045
mir-284	SF00925
mir-286a	SF00946
mir-2881	SF00499
mir-2886	SF00127
mir-289	SF00892
mir-2893	SF01790
mir-2897	SF02168
mir-28c	SF01996
mir-2901	SF02486
mir-2909	SF02395
mir-292	SF00003
mir-2941	SF00513
mir-2944a	SF00129
mir-295	SF00206
mir-2954	SF00253

Continued on next page...

mir-2955	SF02011
mir-296	SF00565
mir-2963	SF02723
mir-2967	SF02614
mir-2970	SF02485
mir-2974	SF00549
mir-2978	SF02482
mir-297c	SF00022
mir-2982	SF00312
mir-2986	SF02558
mir-299	SF00031
mir-2994	SF02658
mir-29a	SF00076
mir-29e	SF00076
mir-2d	SF00043
mir-3015a	SF00774
mir-3017a	SF02062
mir-301a	SF00001
mir-3020	SF02823
mir-3024	SF01408
mir-3028	SF02476
mir-302c	SF00144
mir-303	SF03033
mir-3033	SF00597
mir-3037	SF01824
mir-3040	SF02347
mir-3044	SF00033
mir-3047	SF01216
mir-3050	SF00352
mir-3054	SF02228
mir-3058	SF02830
mir-3061	SF02151
mir-3065	SF00664
mir-3069	SF02404
mir-3072	SF02302
mir-3076	SF03045
mir-307b	SF01948
mir-3082	SF00816
mir-3086	SF01784
mir-309	SF00148
mir-3093	SF02993
mir-3096b	SF00292

mir-2956	SF02369
mir-2960	SF01134
mir-2964	SF02795
mir-2968	SF02286
mir-2971	SF02060
mir-2975	SF02593
mir-2979	SF01730
mir-298	SF00869
mir-2983	SF02076
mir-2987	SF02640
mir-2991	SF00277
mir-2995	SF02277
mir-29b	SF00076
mir-2a	SF00010
mir-3	SF00148
mir-3015b	SF00670
mir-3017b	SF02062
mir-301b	SF00001
mir-3021	SF00556
mir-3025	SF01900
mir-3029	SF00837
mir-302d	SF00144
mir-3030	SF01741
mir-3034	SF02132
mir-3038	SF02045
mir-3041	SF00933
mir-3045a	SF02773
mir-3048	SF02108
mir-3051	SF02229
mir-3055	SF02385
mir-3059	SF00220
mir-3062	SF00299
mir-3066	SF03007
mir-307	SF00298
mir-3073	SF01897
mir-3077	SF02611
mir-308	SF01437
mir-3083	SF03051
mir-3087	SF02203
mir-3090	SF00001
mir-3094	SF01439
mir-3097	SF02125

mir-2958	SF01839
mir-2961	SF02098
mir-2965	SF01611
mir-2969	SF01984
mir-2972	SF02248
mir-2976	SF01121
mir-297a	SF00022
mir-2980	SF02713
mir-2984	SF02466
mir-2988	SF02268
mir-2992	SF02987
mir-2996	SF02246
mir-29c	SF00076
mir-2b	SF00010
mir-300	SF00031
mir-3015c	SF00670
mir-3018	SF02576
mir-301c	SF00001
mir-3022	SF02311
mir-3026	SF01160
mir-302a	SF00144
mir-302e	SF00444
mir-3031	SF02362
mir-3035	SF01176
mir-3039	SF00947
mir-3042	SF00964
mir-3045b	SF02773
mir-3049	SF01457
mir-3052	SF02130
mir-3056	SF00986
mir-306	SF00847
mir-3063	SF01084
mir-3067	SF00650
mir-3070a	SF00497
mir-3074	SF00222
mir-3078	SF02818
mir-3080	SF02451
mir-3084	SF01868
mir-3088	SF01714
mir-3091	SF01641
mir-3095	SF01907
mir-3098	SF02685

mir-2959	SF01970
mir-2962	SF01655
mir-2966	SF01828
mir-297	SF00022
mir-2973	SF02760
mir-2977	SF01507
mir-297b	SF00022
mir-2981	SF00327
mir-2985	SF00399
mir-2989	SF02063
mir-2993	SF01491
mir-2997	SF01564
mir-29d	SF00076
mir-2c	SF00010
mir-301	SF00001
mir-3016	SF00836
mir-3019	SF02730
mir-302	SF00144
mir-3023	SF01759
mir-3027	SF00146
mir-302b	SF00144
mir-302f	SF00422
mir-3032	SF01580
mir-3036	SF00012
mir-304	SF00856
mir-3043	SF01436
mir-3046	SF02433
mir-305	SF00188
mir-3053	SF02693
mir-3057	SF00322
mir-3060	SF01618
mir-3064	SF01386
mir-3068	SF01026
mir-3070b	SF00497
mir-3075	SF01982
mir-3079	SF02365
mir-3081	SF02234
mir-3085	SF00163
mir-3089	SF00755
mir-3092	SF00911
mir-3096	SF00292
mir-3099	SF02097

Continued on next page...

mir-309a	SF00148
mir-309b	SF00148
mir-30a	SF00081
mir-30b	SF00081
mir-30c	SF00081
mir-30d	SF00081
mir-30e	SF00081
mir-30f	SF00081
mir-31	SF00719
mir-310	SF01575
mir-3100	SF02814
mir-3101	SF01312
mir-3102	SF01380
mir-3103	SF02494
mir-3104	SF02514
mir-3105	SF01424
mir-3106	SF00829
mir-3108	SF01806
mir-3109	SF01330
mir-310a	SF01575
mir-310b	SF01575
mir-311	SF00344
mir-3110	SF02940
mir-3112	SF02461
mir-3113	SF00532
mir-3115	SF01024
mir-3116	SF01209
mir-3117	SF01311
mir-3118	SF00731
mir-3119	SF02863
mir-311a	SF00344
mir-311b	SF00343
mir-311c	SF00344
mir-312	SF00343
mir-3120	SF00116
mir-3121	SF00403
mir-3122	SF01818
mir-3123	SF02296
mir-3124	SF00467
mir-3125	SF00955
mir-3126	SF01910
mir-3127	SF00645
mir-3128	SF02005
mir-3129	SF00183
mir-313	SF00343
mir-3130	SF02808
mir-3131	SF00775
mir-3132	SF01264
mir-3133	SF01791
mir-3134	SF01347
mir-3135	SF02170
mir-3135b	SF01122
mir-3136	SF01219
mir-3137	SF01275
mir-3138	SF00883
mir-3139	SF00864
mir-314	SF00838
mir-3140	SF00978
mir-3141	SF00316
mir-3142	SF01709
mir-3143	SF02793
mir-3144	SF01992
mir-3145	SF02224
mir-3146	SF01307
mir-3147	SF02477
mir-3148	SF02942
mir-3149	SF00898
mir-315	SF01604
mir-3150a	SF00655
mir-3150b	SF00655
mir-3151	SF00459
mir-3152	SF02389
mir-3153	SF02146
mir-3154	SF01260
mir-3155	SF00713
mir-3155b	SF00264
mir-3156	SF01212
mir-3157	SF01115
mir-3158	SF01148
mir-3159	SF01862
mir-315a	SF01604
mir-315b	SF02079
mir-316	SF00989
mir-3160	SF00876
mir-3161	SF02636
mir-3162	SF00279
mir-3163	SF02580
mir-3164	SF01279
mir-3165	SF01623
mir-3166	SF01017
mir-3167	SF01031
mir-3168	SF00962
mir-3169	SF01267
mir-317	SF01127
mir-3170	SF02116
mir-3171	SF01775
mir-3173	SF02358
mir-3174	SF02891
mir-3175	SF00831
mir-3176	SF02589
mir-3177	SF02257
mir-3178	SF01633
mir-3179	SF01038
mir-318	SF00148
mir-3180	SF01987
mir-3181	SF00328
mir-3182	SF00863
mir-3183	SF02902
mir-3185	SF01720
mir-3186	SF00166
mir-3187	SF01769
mir-3188	SF01745
mir-3189	SF00953
mir-3191	SF01157
mir-3192	SF02299
mir-3193	SF01326
mir-3194	SF01704
mir-3195	SF01374
mir-3196	SF00362
mir-3197	SF00980
mir-3198	SF01387
mir-3199	SF01136
mir-31a	SF00246
mir-31b	SF00246
mir-32	SF00208
mir-320	SF00058
mir-3200	SF01119
mir-3201	SF00905
mir-3202	SF01601
mir-320a	SF00058
mir-320b	SF00058
mir-320c	SF00058
mir-320d	SF00058
mir-320e	SF00058
mir-322	SF00732
mir-323	SF00031
mir-3236	SF01395
mir-323b	SF00031
mir-323c	SF00031
mir-324	SF00138
mir-325	SF00429
mir-326	SF00016
mir-327	SF03003
mir-328	SF00387
mir-328a	SF00387
mir-328b	SF01500
mir-329	SF00031
mir-329a	SF00031
mir-329b	SF00031
mir-33	SF00153
mir-330	SF00184
mir-331	SF00705
mir-335	SF00458
mir-336	SF00252
mir-337	SF01785
mir-338	SF00664
mir-339	SF00528
mir-339b	SF02992
mir-33a	SF00153
mir-33b	SF00153
mir-34	SF00007
mir-340	SF00902
mir-341	SF01028
mir-342	SF00305
mir-3422	SF00855
mir-343	SF00395
mir-3431	SF02274
mir-3432	SF02119

Continued on next page...

mir-344	SF00288
mir-344a	SF00288
mir-344b	SF00288
mir-344c	SF00288
mir-344d	SF00288
mir-344e	SF00288
mir-344f	SF00288
mir-344g	SF00288
mir-345	SF00207
mir-346	SF00540
mir-347	SF02323
mir-3470a	SF02236
mir-3470b	SF02518
mir-3471	SF01689
mir-3472	SF02686
mir-3473	SF00097
mir-3473b	SF02510
mir-3473c	SF02182
mir-3473d	SF01482
mir-3474	SF02300
mir-3475	SF00430
mir-3477	SF00182
mir-3482	SF02191
mir-3483	SF01082
mir-3485	SF01462
mir-3486	SF00786
mir-3488	SF00546
mir-349	SF00450
mir-3490	SF00552
mir-3492	SF00842
mir-3499	SF00842
mir-34a	SF00007
mir-34b	SF00007
mir-34c	SF00007
mir-35	SF00147
mir-350	SF00580
mir-3503	SF02328
mir-351	SF01810
mir-3523	SF02976
mir-3524	SF01310
mir-3526	SF01220
mir-3527	SF01814
mir-3528	SF02176
mir-353	SF01099
mir-3530	SF03012
mir-3531	SF02207
mir-3532	SF01349
mir-3533	SF00544
mir-3534	SF02541
mir-3535	SF00303
mir-3536	SF02235
mir-3537	SF02127
mir-3538	SF00833
mir-3539	SF03025
mir-354	SF02499
mir-3540	SF00370
mir-3541	SF01133
mir-3542	SF02187
mir-3544	SF01919
mir-3545	SF00326
mir-3546	SF00342
mir-3547	SF01448
mir-3548	SF00781
mir-355	SF01094
mir-3550	SF00934
mir-3551	SF00787
mir-3552	SF01197
mir-3555	SF02375
mir-3558	SF01126
mir-3559	SF02849
mir-356	SF01378
mir-3561	SF01980
mir-3562	SF02864
mir-3564	SF01883
mir-3566	SF00541
mir-3568	SF00944
mir-3569	SF02468
mir-357	SF02258
mir-3572	SF01793
mir-3573	SF02799
mir-3575	SF01616
mir-3577	SF01171
mir-3578	SF00626
mir-3579	SF02092
mir-358	SF01876
mir-3580	SF00175
mir-3583	SF02783
mir-3584	SF00364
mir-3585	SF00099
mir-3586	SF00037
mir-359	SF00483
mir-3590	SF00001
mir-3591	SF00703
mir-3593	SF00851
mir-3594	SF02316
mir-3596	SF01359
mir-3596c	SF02294
mir-3597	SF00267
mir-3598	SF02162
mir-3599	SF02617
mir-35a	SF00147
mir-35b	SF00147
mir-35c	SF00147
mir-35d	SF00147
mir-35e	SF00147
mir-35f	SF00147
mir-35g	SF00147
mir-36	SF00147
mir-360	SF01106
mir-3601	SF00309
mir-3604	SF02171
mir-3605	SF00086
mir-3606	SF00005
mir-3607	SF01107
mir-3609	SF00373
mir-361	SF01020
mir-3610	SF00169
mir-3611	SF02872
mir-3612	SF00680
mir-3613	SF01746
mir-3614	SF01129
mir-3615	SF02179
mir-3616	SF01652
mir-3617	SF01027
mir-3618	SF01464
mir-3619	SF01561
mir-362	SF00082
mir-3620	SF00048
mir-3621	SF01426
mir-3622a	SF02656
mir-3622b	SF02656
mir-363	SF01075
mir-3641	SF02766
mir-3642	SF01731
mir-3643	SF01539
mir-3644	SF01549
mir-3645	SF01185
mir-3646	SF01743
mir-3647	SF00675
mir-3648	SF00625
mir-3649	SF01976
mir-365	SF00285
mir-3650	SF01846
mir-3651	SF01352
mir-3652	SF00477
mir-3653	SF00453
mir-3654	SF00698
mir-3655	SF02930
mir-3656	SF01514
mir-3657	SF02106
mir-3658	SF00904
mir-3659	SF01750
mir-3660	SF02536
mir-3661	SF01418
mir-3662	SF01488
mir-3663	SF01319
mir-3664	SF02705
mir-3665	SF00999
mir-3666	SF01682
mir-3667	SF01965
mir-3668	SF02661
mir-3669	SF00072
mir-367	SF00426
mir-3670	SF02131
mir-3671	SF01694
mir-3672	SF02205
mir-3673	SF00536
mir-3674	SF01105

Continued on next page...

mir-3675	SF01979	mir-3676	SF02973	mir-3677	SF02173	mir-3678	SF00083
mir-3679	SF01889	mir-3680	SF02303	mir-3681	SF00550	mir-3682	SF02242
mir-3683	SF02755	mir-3684	SF02903	mir-3685	SF01546	mir-3686	SF01773
mir-3687	SF02137	mir-3688	SF00704	mir-3689a	SF00954	mir-3689b	SF00924
mir-3689c	SF00954	mir-3689d	SF00954	mir-3689e	SF00954	mir-3689f	SF00954
mir-369	SF00031	mir-3690	SF01011	mir-3691	SF00880	mir-3692	SF00225
mir-37	SF00147	mir-370	SF00251	mir-371	SF00003	mir-3713	SF02975
mir-3714	SF01879	mir-3715	SF00245	mir-3716a	SF02770	mir-3716b	SF01797
mir-3717	SF02702	mir-3718a	SF01151	mir-3719	SF00452	mir-371b	SF00003
mir-372	SF00137	mir-3722	SF00413	mir-3726	SF01852	mir-3727	SF02178
mir-3728	SF01647	mir-3729	SF02266	mir-373	SF00003	mir-3730	SF03046
mir-3733	SF02390	mir-3734	SF00311	mir-3736	SF00942	mir-3737	SF02904
mir-3738	SF01274	mir-3739	SF02427	mir-374	SF01382	mir-3741	SF02263
mir-3742	SF01574	mir-3743	SF02423	mir-3745	SF02622	mir-3746	SF01369
mir-3747a	SF02757	mir-3747b	SF02470	mir-3748	SF02110	mir-3749	SF02370
mir-374a	SF01382	mir-374b	SF01382	mir-375	SF00320	mir-3750	SF01443
mir-3751	SF00931	mir-3752	SF03002	mir-3754	SF02762	mir-3764	SF02416
mir-3766	SF02737	mir-3767	SF01803	mir-3768	SF01808	mir-376a	SF00031
mir-376b	SF00031	mir-376c	SF00031	mir-376d	SF00031	mir-376e	SF00031
mir-377	SF00031	mir-3770	SF02368	mir-3771	SF02016	mir-3773	SF01695
mir-3775	SF01077	mir-3776	SF01473	mir-3779	SF02557	mir-378	SF00519
mir-3780	SF02842	mir-3781	SF01076	mir-3782	SF01931	mir-3783	SF02681
mir-3784	SF02886	mir-3785	SF02553	mir-3786	SF01835	mir-3787	SF01596
mir-3788	SF00994	mir-3789	SF01030	mir-378b	SF00160	mir-378c	SF00519
mir-378d	SF00160	mir-378e	SF01551	mir-378f	SF00160	mir-378g	SF00160
mir-378h	SF01421	mir-378i	SF01455	mir-379	SF00031	mir-3790	SF01409
mir-3792	SF02772	mir-3794	SF02025	mir-3795	SF02768	mir-3796	SF03023
mir-3797	SF02655	mir-3798	SF02811	mir-3799	SF02278	mir-38	SF00147
mir-380	SF00031	mir-3800	SF02979	mir-3801	SF00383	mir-3802	SF02492
mir-381	SF00031	mir-3811c	SF01486	mir-3817	SF00778	mir-382	SF00031
mir-3828	SF00200	mir-383	SF00649	mir-3837	SF01102	mir-384	SF00457
mir-3842	SF02562	mir-3852	SF02112	mir-3854	SF00985	mir-3856	SF01150
mir-3859	SF02199	mir-3861	SF01234	mir-3868	SF00643	mir-3873	SF01924
mir-3876	SF03011	mir-3885	SF00052	mir-3888	SF02506	mir-3897	SF00103
mir-39	SF00147	mir-3901	SF02549	mir-3904	SF01493	mir-3906	SF02072
mir-3907	SF01236	mir-3908	SF01651	mir-3909	SF00060	mir-3910	SF02355
mir-3911	SF00629	mir-3912	SF02465	mir-3913	SF00329	mir-3914	SF01341
mir-3915	SF02511	mir-3916	SF00228	mir-3917	SF00053	mir-3918	SF02950
mir-3919	SF02650	mir-392	SF02639	mir-3920	SF02896	mir-3921	SF01433
mir-3922	SF02157	mir-3923	SF02643	mir-3924	SF02126	mir-3925	SF02420
mir-3926	SF01344	mir-3927	SF02053	mir-3928	SF02291	mir-3929	SF01955
mir-3931	SF02610	mir-3934	SF00263	mir-3935	SF02484	mir-3936	SF02578

Continued on next page...

mir-3937	SF01391
mir-3941	SF00935
mir-3945	SF02160
mir-3958	SF00031
mir-3962	SF02105
mir-3966	SF02744
mir-3970	SF02267
mir-3974	SF03020
mir-3978	SF01600
mir-4000b	SF00068
mir-4000f	SF01230
mir-4001a	SF00599
mir-4001e	SF00917
mir-4001i	SF01112
mir-4003c	SF01156
mir-4005b	SF00614
mir-4006c	SF00062
mir-4006g	SF00535
mir-4009a	SF01397
mir-4011a	SF02438
mir-4013b	SF02256
mir-4017	SF00668
mir-4020b	SF00527
mir-4025	SF01384
mir-4029	SF00800
mir-4034	SF02495
mir-4038	SF01423
mir-4043	SF01410
mir-4047	SF02032
mir-4051	SF02838
mir-4055	SF01272
mir-4059	SF01841
mir-4063	SF02709
mir-4067	SF02775
mir-4071	SF02022
mir-4075	SF02586
mir-4077c	SF00388
mir-4081	SF02528
mir-4086	SF00932
mir-409	SF00031
mir-4093	SF02996
mir-4098	SF01822

mir-3938	SF01645
mir-3942	SF01954
mir-3955	SF01248
mir-3959	SF00031
mir-3963	SF00360
mir-3967	SF02971
mir-3971	SF02154
mir-3975	SF01138
mir-4	SF00857
mir-4000c	SF00067
mir-4000g	SF02003
mir-4001b	SF01590
mir-4001f	SF02006
mir-4002	SF00746
mir-4003d	SF01155
mir-4005c	SF00614
mir-4006d	SF00506
mir-4008a	SF01373
mir-4009b	SF02628
mir-4011b	SF02439
mir-4014	SF01211
mir-4018b	SF02526
mir-4021	SF01508
mir-4026	SF01522
mir-4030	SF00100
mir-4035	SF02937
mir-4039	SF01402
mir-4044	SF02539
mir-4048	SF02932
mir-4052	SF00401
mir-4056	SF02198
mir-4060	SF02869
mir-4064	SF03040
mir-4068	SF01517
mir-4072	SF01528
mir-4076	SF01451
mir-4077d	SF00388
mir-4083	SF01918
mir-4087	SF02871
mir-4090	SF02997
mir-4094	SF02594
mir-4099	SF02784

mir-3939	SF01776
mir-3943	SF00991
mir-3956	SF00943
mir-3960	SF00908
mir-3964	SF01110
mir-3968	SF02520
mir-3972	SF02994
mir-3976	SF01060
mir-40	SF00147
mir-4000d	SF00068
mir-4000h	SF00512
mir-4001c	SF01867
mir-4001g	SF02478
mir-4003a	SF01155
mir-4004	SF02627
mir-4006a	SF00506
mir-4006e	SF00468
mir-4008b	SF01373
mir-4009c	SF00823
mir-4012	SF01016
mir-4015	SF01974
mir-4019	SF01699
mir-4022	SF00423
mir-4027	SF02556
mir-4031	SF02649
mir-4036	SF00439
mir-4040	SF02746
mir-4045	SF01447
mir-4049	SF01317
mir-4053	SF01377
mir-4057	SF01870
mir-4061	SF03022
mir-4065	SF01325
mir-4069	SF03036
mir-4073	SF02542
mir-4077a	SF00388
mir-4078	SF02141
mir-4084	SF02701
mir-4088	SF02295
mir-4091	SF02403
mir-4095	SF02129
mir-409a	SF00031

mir-3940	SF01796
mir-3944	SF02031
mir-3957	SF02250
mir-3961	SF01459
mir-3965	SF02493
mir-3969	SF01787
mir-3973	SF02136
mir-3977	SF02588
mir-4000a	SF00860
mir-4000e	SF00075
mir-4000i	SF01480
mir-4001d	SF02210
mir-4001h	SF00599
mir-4003b	SF01155
mir-4005a	SF00614
mir-4006b	SF00063
mir-4006f	SF00506
mir-4008c	SF01373
mir-4010	SF02282
mir-4013a	SF02256
mir-4016	SF00975
mir-4020a	SF00526
mir-4024	SF02742
mir-4028	SF01950
mir-4033	SF02247
mir-4037	SF01885
mir-4042	SF00611
mir-4046	SF02329
mir-4050	SF02631
mir-4054	SF01736
mir-4058	SF02670
mir-4062	SF02519
mir-4066	SF02725
mir-4070	SF01856
mir-4074	SF00812
mir-4077b	SF00388
mir-4079	SF00435
mir-4085	SF02554
mir-4089	SF02751
mir-4092	SF02152
mir-4097	SF02843
mir-41	SF00147

Continued on next page...

mir-410	SF00031
mir-4104	SF01962
mir-4109	SF02232
mir-4112	SF02261
mir-4116	SF00906
mir-412	SF00031
mir-4123	SF01537
mir-4127	SF01728
mir-4131	SF02624
mir-4135	SF02909
mir-4139	SF01936
mir-4143	SF00745
mir-4147	SF02957
mir-4151	SF01717
mir-4155	SF02435
mir-4159	SF02284
mir-4164	SF00686
mir-4169	SF02504
mir-4174	SF01921
mir-4178b	SF02752
mir-4182	SF01066
mir-4186	SF02692
mir-4191	SF02174
mir-4195	SF02835
mir-4199	SF02810
mir-4202	SF01261
mir-4206	SF01521
mir-421	SF00350
mir-4214	SF01071
mir-4218	SF01869
mir-423	SF00651
mir-4251	SF00859
mir-4255	SF00319
mir-4259	SF01543
mir-4263	SF00825
mir-4267	SF00521
mir-4270	SF01716
mir-4274	SF01559
mir-4278	SF02682
mir-4281	SF00151
mir-4285	SF02002
mir-4289	SF01195

mir-4100	SF01845
mir-4105	SF00259
mir-411	SF00031
mir-4113	SF01723
mir-4117	SF00247
mir-4120	SF00743
mir-4124	SF02507
mir-4128	SF02892
mir-4132	SF00879
mir-4136	SF02412
mir-4140	SF02185
mir-4144	SF02897
mir-4148	SF01383
mir-4152	SF02114
mir-4156	SF02065
mir-4160	SF02544
mir-4165	SF01777
mir-4171	SF02012
mir-4176	SF01324
mir-4179	SF00848
mir-4183	SF02674
mir-4187	SF02883
mir-4192	SF00753
mir-4196	SF02824
mir-41b	SF00147
mir-4203	SF01348
mir-4207	SF02096
mir-4211	SF01609
mir-4215	SF01718
mir-4219	SF01766
mir-423a	SF00651
mir-4252	SF01770
mir-4256	SF00121
mir-4260	SF00027
mir-4264	SF00168
mir-4268	SF02290
mir-4271	SF00141
mir-4275	SF00505
mir-4279	SF00331
mir-4282	SF01687
mir-4286	SF01474
mir-429	SF00226

mir-4101	SF01800
mir-4106	SF01214
mir-4110	SF01571
mir-4114	SF02949
mir-4118	SF02081
mir-4121	SF01772
mir-4125	SF02921
mir-4129	SF02120
mir-4133	SF01991
mir-4137	SF02885
mir-4141	SF01705
mir-4145	SF01131
mir-4149	SF00982
mir-4153	SF03019
mir-4157	SF02379
mir-4162	SF01450
mir-4166	SF01721
mir-4172	SF01903
mir-4177	SF00011
mir-4180	SF02434
mir-4184	SF00123
mir-4189	SF01477
mir-4193	SF02421
mir-4197	SF02677
mir-42	SF00147
mir-4204	SF02600
mir-4208	SF01328
mir-4212	SF00432
mir-4216	SF02678
mir-4220	SF01612
mir-424	SF00732
mir-4253	SF00272
mir-4257	SF00077
mir-4261	SF00135
mir-4265	SF01128
mir-4269	SF02831
mir-4272	SF01227
mir-4276	SF00870
mir-428	SF00095
mir-4283	SF01137
mir-4287	SF00662
mir-4290	SF01253

mir-4103	SF01495
mir-4108	SF00293
mir-4111	SF00569
mir-4115	SF01725
mir-411b	SF00031
mir-4122	SF00764
mir-4126	SF02778
mir-4130	SF02575
mir-4134	SF03008
mir-4138	SF01708
mir-4142	SF00133
mir-4146	SF01067
mir-4150	SF01993
mir-4154	SF01597
mir-4158	SF00437
mir-4163	SF01617
mir-4168	SF03024
mir-4173	SF01605
mir-4178a	SF02752
mir-4181	SF02952
mir-4185	SF01304
mir-4190	SF01365
mir-4194	SF00098
mir-4198	SF00415
mir-4201	SF00273
mir-4205	SF01794
mir-4209	SF02816
mir-4213	SF00724
mir-4217	SF02978
mir-422a	SF00482
mir-425	SF00726
mir-4254	SF01890
mir-4258	SF00185
mir-4262	SF01713
mir-4266	SF00087
mir-427	SF00050
mir-4273	SF01510
mir-4277	SF01338
mir-4280	SF01804
mir-4284	SF01471
mir-4288	SF00306
mir-4291	SF00240

Continued on next page...

mir-4292	SF01231	mir-4293	SF00224	mir-4294	SF00494	mir-4295	SF00411
mir-4296	SF00547	mir-4297	SF00114	mir-4298	SF02953	mir-4299	SF01541
mir-429b	SF00226	mir-42a	SF00496	mir-42b	SF01381	mir-43	SF00055
mir-4300	SF01978	mir-4301	SF01533	mir-4302	SF00093	mir-4303	SF00110
mir-4304	SF02301	mir-4305	SF01070	mir-4306	SF00843	mir-4307	SF01162
mir-4308	SF01021	mir-4309	SF00707	mir-430a	SF00051	mir-430b	SF00051
mir-430c	SF00051	mir-430i	SF00051	mir-431	SF00814	mir-4310	SF00335
mir-4311	SF01139	mir-4312	SF02722	mir-4313	SF00571	mir-4314	SF00180
mir-4315	SF02708	mir-4316	SF00557	mir-4317	SF00491	mir-4318	SF00780
mir-4319	SF00287	mir-432	SF00390	mir-4320	SF01747	mir-4321	SF02376
mir-4322	SF01553	mir-4323	SF00209	mir-4324	SF02101	mir-4325	SF01516
mir-4326	SF01696	mir-4327	SF00813	mir-4328	SF00468	mir-4329	SF00647
mir-433	SF00502	mir-4330	SF02540	mir-4331	SF00210	mir-4333	SF01591
mir-4334	SF02023	mir-4335	SF02483	mir-4336	SF01340	mir-4337	SF02428
mir-434	SF02720	mir-44	SF00216	mir-4417	SF00463	mir-4418	SF00089
mir-4419a	SF00330	mir-4419b	SF00022	mir-4420	SF02252	mir-4421	SF00516
mir-4422	SF00356	mir-4423	SF01358	mir-4424	SF02954	mir-4425	SF02107
mir-4426	SF00227	mir-4427	SF01445	mir-4428	SF01692	mir-4429	SF01767
mir-4430	SF02432	mir-4431	SF02175	mir-4432	SF01994	mir-4433	SF03001
mir-4434	SF00487	mir-4435	SF01281	mir-4436a	SF01103	mir-4436b	SF01103
mir-4437	SF00710	mir-4438	SF01350	mir-4439	SF02801	mir-4440	SF02472
mir-4441	SF01941	mir-4442	SF00025	mir-4443	SF00488	mir-4444	SF00960
mir-4445	SF02797	mir-4446	SF01297	mir-4447	SF00304	mir-4448	SF00219
mir-4449	SF01757	mir-4450	SF01123	mir-4451	SF00080	mir-4452	SF00001
mir-4453	SF01303	mir-4454	SF00199	mir-4455	SF00117	mir-4456	SF00354
mir-4457	SF02837	mir-4458	SF01058	mir-4459	SF00022	mir-4460	SF00968
mir-4461	SF00382	mir-4462	SF01165	mir-4463	SF00529	mir-4464	SF01222
mir-4465	SF02193	mir-4466	SF00616	mir-4467	SF01342	mir-4468	SF00523
mir-4469	SF01710	mir-4470	SF01756	mir-4471	SF02780	mir-4472	SF00015
mir-4473	SF02058	mir-4474	SF01411	mir-4475	SF02322	mir-4476	SF02967
mir-4477a	SF01191	mir-4477b	SF01191	mir-4478	SF00022	mir-4479	SF01009
mir-448	SF00671	mir-4480	SF02739	mir-4481	SF00539	mir-4482	SF00660
mir-4483	SF00391	mir-4484	SF01774	mir-4485	SF00481	mir-4486	SF01037
mir-4487	SF00271	mir-4488	SF01240	mir-4489	SF02804	mir-449	SF00008
mir-4490	SF00966	mir-4491	SF02732	mir-4492	SF02728	mir-4493	SF01108
mir-4494	SF01943	mir-4495	SF02363	mir-4496	SF00890	mir-4497	SF00742
mir-4498	SF00835	mir-4499	SF02822	mir-449a	SF00008	mir-449b	SF00008
mir-449c	SF00008	mir-449d	SF01005	mir-45	SF00216	mir-450	SF00642
mir-4500	SF00433	mir-4501	SF01327	mir-4502	SF01821	mir-4503	SF02497
mir-4504	SF02841	mir-4505	SF00809	mir-4506	SF00822	mir-4507	SF00159
mir-4508	SF02815	mir-4509	SF00807	mir-450a	SF00642	mir-450b	SF00642
mir-450c	SF00642	mir-451	SF00300	mir-4510	SF02091	mir-4511	SF02721

Continued on next page...

mir-4512	SF00914
mir-4516	SF02043
mir-451a	SF00642
mir-4522	SF01977
mir-4526	SF01208
mir-453	SF00031
mir-4533	SF02986
mir-4537	SF01629
mir-4540	SF01938
mir-455b	SF00712
mir-458	SF00094
mir-4606	SF02845
mir-462	SF01278
mir-4634	SF00624
mir-4638	SF01811
mir-4642	SF01371
mir-4646	SF00386
mir-465	SF00099
mir-4653	SF01007
mir-4657	SF00819
mir-465a	SF00099
mir-4660	SF01838
mir-4664	SF02017
mir-4668	SF02462
mir-466c	SF00022
mir-466g	SF00022
mir-466k	SF00022
mir-466o	SF00022
mir-4671	SF01643
mir-4675	SF00283
mir-4679	SF01990
mir-467d	SF00022
mir-467h	SF00022
mir-4682	SF00977
mir-4686	SF01051
mir-4690	SF02001
mir-4694	SF01194
mir-4698	SF00628
mir-4700	SF01489
mir-4705	SF03021
mir-4709	SF00588
mir-4712	SF01233

mir-4513	SF01252
mir-4517	SF01201
mir-452	SF00874
mir-4523	SF00862
mir-4527	SF03042
mir-4530	SF02371
mir-4534	SF00026
mir-4538	SF01629
mir-454a	SF00232
mir-456	SF01351
mir-459	SF00238
mir-460a	SF01217
mir-463	SF00099
mir-4635	SF00915
mir-4639	SF02010
mir-4643	SF00929
mir-4647	SF00939
mir-4650	SF02350
mir-4654	SF01412
mir-4658	SF02769
mir-465b	SF00099
mir-4661	SF01958
mir-4665	SF00772
mir-4669	SF01686
mir-466d	SF00022
mir-466h	SF00022
mir-466l	SF00022
mir-466p	SF00022
mir-4672	SF00559
mir-4676	SF01241
mir-467a	SF00022
mir-467e	SF00022
mir-468	SF01587
mir-4683	SF01851
mir-4687	SF00990
mir-4691	SF01149
mir-4695	SF00806
mir-4699	SF02386
mir-4701	SF00193
mir-4706	SF02543
mir-471	SF01003
mir-4713	SF02073

mir-4514	SF00551
mir-4518	SF02789
mir-4520b	SF02853
mir-4524	SF00630
mir-4528	SF02314
mir-4531	SF00363
mir-4535	SF00161
mir-4539	SF01782
mir-454b	SF00232
mir-457a	SF00779
mir-46	SF00606
mir-460b	SF01217
mir-4632	SF01711
mir-4636	SF01271
mir-4640	SF00485
mir-4644	SF00248
mir-4648	SF02289
mir-4651	SF00409
mir-4655	SF01638
mir-4659a	SF01142
mir-465c	SF00099
mir-4662a	SF01983
mir-4666	SF01968
mir-466a	SF00022
mir-466e	SF00022
mir-466i	SF00022
mir-466m	SF00022
mir-466q	SF00446
mir-4673	SF02035
mir-4677	SF00830
mir-467b	SF00022
mir-467f	SF00202
mir-4680	SF00744
mir-4684	SF01427
mir-4688	SF01701
mir-4692	SF00443
mir-4696	SF02086
mir-47	SF00606
mir-4703	SF01008
mir-4707	SF01277
mir-4710	SF01296
mir-4714	SF00903

mir-4515	SF00959
mir-4519	SF00380
mir-4521	SF02335
mir-4525	SF02683
mir-4529	SF00375
mir-4532	SF02417
mir-4536	SF01802
mir-454	SF00232
mir-455	SF00712
mir-457b	SF00150
mir-460	SF01217
mir-461	SF00895
mir-4633	SF00638
mir-4637	SF02367
mir-4641	SF00639
mir-4645	SF00926
mir-4649	SF01739
mir-4652	SF02680
mir-4656	SF01120
mir-4659b	SF01142
mir-466	SF00022
mir-4663	SF01419
mir-4667	SF00509
mir-466b	SF00022
mir-466f	SF00022
mir-466j	SF00022
mir-466n	SF00022
mir-4670	SF00613
mir-4674	SF00909
mir-4678	SF03053
mir-467c	SF00022
mir-467g	SF00022
mir-4681	SF02936
mir-4685	SF00321
mir-4689	SF01895
mir-4693	SF01025
mir-4697	SF01778
mir-470	SF00099
mir-4704	SF01836
mir-4708	SF01202
mir-4711	SF02373
mir-4715	SF00266

Continued on next page...

mir-4716	SF01014	mir-4717	SF02487	mir-4718	SF00784	mir-4719	SF01964
mir-4720	SF00727	mir-4721	SF01524	mir-4722	SF01622	mir-4723	SF00846
mir-4724	SF01640	mir-4725	SF00988	mir-4726	SF01963	mir-4727	SF02015
mir-4728	SF01702	mir-4729	SF00878	mir-4730	SF01180	mir-4731	SF01440
mir-4732	SF01396	mir-4733	SF01542	mir-4734	SF01492	mir-4735	SF01986
mir-4736	SF00442	mir-4737	SF02659	mir-4738	SF01256	mir-4739	SF02920
mir-4740	SF00798	mir-4741	SF01891	mir-4742	SF01672	mir-4743	SF01586
mir-4744	SF02710	mir-4745	SF02109	mir-4746	SF01703	mir-4747	SF01000
mir-4748	SF02749	mir-4749	SF00592	mir-4750	SF01174	mir-4751	SF00850
mir-4752	SF00789	mir-4753	SF01073	mir-4754	SF02341	mir-4755	SF01513
mir-4756	SF01033	mir-4757	SF01313	mir-4758	SF01143	mir-4759	SF02215
mir-4760	SF00735	mir-4761	SF02564	mir-4762	SF02879	mir-4763	SF00690
mir-4764	SF01691	mir-4765	SF00673	mir-4766	SF00074	mir-4767	SF00017
mir-4768	SF01224	mir-4769	SF02133	mir-4770	SF00766	mir-4771	SF02077
mir-4772	SF00852	mir-4773	SF00877	mir-4774	SF01485	mir-4775	SF00578
mir-4776	SF02206	mir-4777	SF02981	mir-4778	SF01428	mir-4779	SF02249
mir-4780	SF01056	mir-4781	SF02410	mir-4782	SF01468	mir-4783	SF00633
mir-4784	SF01288	mir-4785	SF01581	mir-4786	SF00965	mir-4787	SF00408
mir-4788	SF01178	mir-4789	SF00530	mir-4790	SF01783	mir-4791	SF00657
mir-4792	SF00581	mir-4793	SF00801	mir-4794	SF01635	mir-4795	SF00591
mir-4796	SF02272	mir-4797	SF01506	mir-4798	SF00981	mir-4799	SF01884
mir-48	SF01117	mir-4800	SF00678	mir-4801	SF02111	mir-4802	SF02075
mir-4803	SF00824	mir-4804	SF02684	mir-4805	SF02988	mir-4806	SF02847
mir-4807	SF00257	mir-4808	SF02995	mir-4809	SF02324	mir-4810	SF00256
mir-4811	SF02855	mir-4812	SF01498	mir-4813	SF00777	mir-4814	SF02898
mir-4815	SF00776	mir-4816	SF02183	mir-4825	SF01080	mir-483	SF01466
mir-484	SF00941	mir-4847	SF02846	mir-4848a	SF02054	mir-4848b	SF02306
mir-4849	SF01789	mir-485	SF00031	mir-4850	SF01920	mir-4851	SF02567
mir-4852	SF02734	mir-4853	SF00275	mir-4854	SF02255	mir-486	SF01444
mir-4863	SF02911	mir-487a	SF00031	mir-487b	SF00031	mir-488	SF00689
mir-489	SF00759	mir-49	SF00722	mir-490	SF00204	mir-4908	SF01356
mir-4909	SF01956	mir-491	SF00492	mir-4910	SF02666	mir-4911	SF03005
mir-4912	SF00936	mir-4913	SF02989	mir-4914	SF02927	mir-4915	SF02630
mir-4916	SF01917	mir-4917	SF00469	mir-4918	SF02565	mir-4919	SF00750
mir-492	SF01210	mir-4920	SF00887	mir-4921	SF00810	mir-4922	SF01316
mir-4923a	SF01345	mir-4923b	SF01346	mir-4924	SF02326	mir-4925	SF02444
mir-4926	SF00888	mir-4927	SF02159	mir-4929	SF02998	mir-493	SF00004
mir-4930	SF00992	mir-4931	SF02763	mir-4932	SF02809	mir-4933	SF02087
mir-4934	SF02226	mir-4935	SF02963	mir-4936	SF02445	mir-4937	SF01563
mir-4938	SF01843	mir-4939	SF02895	mir-493a	SF00615	mir-493b	SF00004
mir-494	SF00031	mir-4940	SF01478	mir-4941	SF02167	mir-4942	SF02281
mir-4943	SF02731	mir-4944	SF02391	mir-4945	SF01608	mir-4946	SF02648

Continued on next page...

mir-4947	SF01959
mir-4950	SF02962
mir-4954	SF02934
mir-4958	SF02577
mir-4961	SF02704
mir-4965	SF02342
mir-4969	SF00142
mir-4972	SF02356
mir-4976	SF01582
mir-498	SF01221
mir-4983	SF02431
mir-4987	SF02411
mir-50	SF01441
mir-501	SF00082
mir-503	SF00438
mir-505	SF00470
mir-509	SF00099
mir-5098	SF00913
mir-51	SF00666
mir-5102	SF02447
mir-5106	SF00867
mir-511	SF00504
mir-5113	SF02118
mir-5117	SF02966
mir-5120	SF02645
mir-5124	SF00134
mir-5128	SF01483
mir-5132	SF01173
mir-5136	SF02374
mir-514	SF00099
mir-516a	SF00003
mir-517b	SF00003
mir-518c	SF00003
mir-519a	SF00003
mir-519e	SF00003
mir-520b	SF00003
mir-520f	SF00003
mir-522	SF00003
mir-524	SF00003
mir-527	SF00003
mir-54	SF01163
mir-543	SF00031

mir-4948	SF00385
mir-4951	SF01577
mir-4955	SF02791
mir-4959	SF02200
mir-4962	SF02046
mir-4966	SF02606
mir-497	SF01152
mir-4973	SF00002
mir-4977	SF01280
mir-4980	SF01499
mir-4984	SF01887
mir-499	SF00603
mir-500	SF00082
mir-502	SF00082
mir-504	SF00317
mir-506	SF00099
mir-5095	SF00001
mir-5099	SF02056
mir-510	SF00099
mir-5103	SF02406
mir-5107	SF03004
mir-5110	SF01663
mir-5114	SF01064
mir-5118	SF00961
mir-5121	SF02644
mir-5125	SF02698
mir-5129	SF02738
mir-5133	SF02794
mir-513a	SF00099
mir-514b	SF00099
mir-516b	SF00003
mir-517c	SF00003
mir-518d	SF00003
mir-519b	SF00003
mir-519f	SF00003
mir-520c	SF00003
mir-520g	SF00003
mir-523	SF00003
mir-525	SF00003
mir-53	SF00667
mir-540	SF02550
mir-544	SF00001

mir-4949	SF01083
mir-4952	SF01400
mir-4956	SF00455
mir-496	SF00031
mir-4963	SF01795
mir-4967	SF01168
mir-4970	SF00573
mir-4974	SF02251
mir-4978	SF01953
mir-4981	SF00284
mir-4985	SF01913
mir-499a	SF02786
mir-500a	SF00082
mir-502a	SF00082
mir-5046	SF02457
mir-507	SF00099
mir-5096	SF00001
mir-509a	SF00099
mir-5100	SF01405
mir-5104	SF01937
mir-5108	SF02047
mir-5111	SF01531
mir-5115	SF00560
mir-5119	SF00201
mir-5122	SF02467
mir-5126	SF01620
mir-5130	SF02591
mir-5134	SF01043
mir-513b	SF00099
mir-515	SF00003
mir-517	SF00003
mir-518a	SF00003
mir-518e	SF00003
mir-519c	SF00003
mir-52	SF00667
mir-520d	SF00003
mir-520h	SF00003
mir-523a	SF00003
mir-526a	SF00003
mir-532	SF00082
mir-541	SF00341
mir-544a	SF00001

mir-495	SF00031
mir-4953	SF01792
mir-4957	SF03015
mir-4960	SF01857
mir-4964	SF00424
mir-4968	SF00234
mir-4971	SF00143
mir-4975	SF00987
mir-4979	SF01047
mir-4982	SF00783
mir-4986	SF02008
mir-5	SF00130
mir-500b	SF00082
mir-502b	SF00082
mir-5047	SF00641
mir-508	SF00099
mir-5097	SF00969
mir-509b	SF00099
mir-5101	SF00751
mir-5105	SF02792
mir-5109	SF01676
mir-5112	SF00511
mir-5116	SF01053
mir-512	SF00003
mir-5123	SF02596
mir-5127	SF00260
mir-5131	SF02671
mir-5135	SF02269
mir-513c	SF00099
mir-516	SF00003
mir-517a	SF00003
mir-518b	SF00003
mir-518f	SF00003
mir-519d	SF00003
mir-520a	SF00003
mir-520e	SF00003
mir-521	SF00003
mir-523b	SF00003
mir-526b	SF00003
mir-539	SF00031
mir-542	SF00309
mir-544b	SF00001

Continued on next page...

mir-545	SF00350
mir-548aa	SF00368
mir-548ae	SF00030
mir-548aj	SF00030
mir-548an	SF00030
mir-548e	SF00030
mir-548i	SF00030
mir-548m	SF00030
mir-548q	SF00030
mir-548v	SF00030
mir-549	SF02840
mir-54d	SF01163
mir-550b	SF00554
mir-552	SF02471
mir-556	SF01285
mir-55a	SF01855
mir-562	SF00296
mir-567	SF01385
mir-570	SF00030
mir-574	SF00367
mir-578	SF00773
mir-581	SF00979
mir-585	SF02561
mir-589	SF01538
mir-590	SF01848
mir-595	SF01079
mir-599	SF00249
mir-601	SF01602
mir-605	SF00652
mir-609	SF02021
mir-612	SF02070
mir-616	SF00361
mir-61a	SF02113
mir-622	SF01675
mir-626	SF00047
mir-629	SF01181
mir-632	SF00176
mir-636	SF01282
mir-63a	SF00866
mir-63e	SF00349
mir-64	SF01130
mir-642b	SF01320

mir-546	SF00706
mir-548ab	SF00030
mir-548ag	SF00030
mir-548ak	SF00030
mir-548b	SF00030
mir-548f	SF00030
mir-548j	SF00030
mir-548n	SF00030
mir-548s	SF00821
mir-548w	SF00030
mir-54a	SF01163
mir-55	SF01268
mir-551	SF00412
mir-553	SF02308
mir-557	SF00648
mir-55b	SF02144
mir-563	SF00679
mir-568	SF00291
mir-571	SF00808
mir-575	SF01946
mir-579	SF00674
mir-582	SF00325
mir-586	SF00632
mir-58a	SF00854
mir-591	SF00400
mir-596	SF01888
mir-6	SF00019
mir-602	SF01866
mir-606	SF02319
mir-61	SF02197
mir-613	SF00610
mir-617	SF02848
mir-62	SF00958
mir-623	SF01929
mir-627	SF01904
mir-63	SF01154
mir-633	SF02700
mir-637	SF00584
mir-63b	SF00454
mir-63f	SF00454
mir-640	SF01187
mir-643	SF00802

mir-547	SF00099
mir-548ac	SF00030
mir-548ah	SF00030
mir-548al	SF00030
mir-548c	SF00030
mir-548g	SF00030
mir-548k	SF00030
mir-548o	SF00030
mir-548t	SF00030
mir-548x	SF00030
mir-54b	SF01163
mir-550	SF00339
mir-551a	SF00412
mir-554	SF01207
mir-558	SF01518
mir-56	SF01504
mir-564	SF00274
mir-569	SF02533
mir-572	SF01971
mir-576	SF00723
mir-58	SF00854
mir-583	SF02321
mir-587	SF02071
mir-58b	SF00854
mir-592	SF00440
mir-597	SF00984
mir-60	SF01644
mir-603	SF00030
mir-607	SF01545
mir-610	SF02195
mir-614	SF00393
mir-618	SF02147
mir-620	SF01998
mir-624	SF01376
mir-628	SF00372
mir-630	SF02651
mir-634	SF00334
mir-638	SF00739
mir-63c	SF00454
mir-63g	SF00349
mir-641	SF02332
mir-644	SF01901

mir-548a	SF00030
mir-548ad	SF00030
mir-548ai	SF00030
mir-548am	SF00030
mir-548d	SF00030
mir-548h	SF00030
mir-548l	SF00030
mir-548p	SF00030
mir-548u	SF00030
mir-548y	SF00030
mir-54c	SF01163
mir-550a	SF00339
mir-551b	SF00412
mir-555	SF01503
mir-559	SF02480
mir-561	SF00054
mir-566	SF00022
mir-57	SF01625
mir-573	SF00792
mir-577	SF00631
mir-580	SF01362
mir-584	SF00471
mir-588	SF02253
mir-59	SF02500
mir-593	SF00858
mir-598	SF02181
mir-600	SF02516
mir-604	SF00366
mir-608	SF01666
mir-611	SF01667
mir-615	SF00021
mir-619	SF00001
mir-621	SF00593
mir-625	SF02657
mir-628a	SF00372
mir-631	SF00214
mir-635	SF00508
mir-639	SF00451
mir-63d	SF00349
mir-63h	SF00349
mir-642	SF01320
mir-645	SF02619

Continued on next page...

mir-646	SF00407
mir-64a	SF01909
mir-64e	SF00473
mir-650b	SF00417
mir-652	SF00099
mir-656	SF00031
mir-66	SF01393
mir-663	SF00912
mir-664b	SF00567
mir-668	SF00235
mir-669c	SF00022
mir-669g	SF00022
mir-669k	SF00022
mir-669o	SF00022
mir-671	SF00118
mir-675	SF01323
mir-677	SF02525
mir-681	SF01394
mir-686	SF01724
mir-691	SF01205
mir-695	SF02679
mir-70	SF02140
mir-703	SF00280
mir-707	SF02581
mir-710	SF01871
mir-718	SF01657
mir-720	SF00044
mir-724	SF00314
mir-728	SF00972
mir-731	SF00726
mir-735	SF01270
mir-739	SF00659
mir-741	SF01826
mir-744	SF01902
mir-75	SF02231
mir-753	SF02213
mir-759	SF01372
mir-762	SF00730
mir-766	SF00562
mir-77	SF02050
mir-785	SF01228
mir-789	SF01560

mir-647	SF00721
mir-64b	SF00295
mir-65	SF01130
mir-650c	SF00417
mir-653	SF00794
mir-657	SF02009
mir-660	SF00082
mir-663a	SF00912
mir-665	SF01614
mir-669	SF00347
mir-669d	SF00022
mir-669h	SF00022
mir-669l	SF00022
mir-669p	SF00022
mir-672	SF00035
mir-675a	SF01323
mir-678	SF02707
mir-682	SF00608
mir-687	SF01999
mir-692	SF01487
mir-697	SF02632
mir-700	SF02394
mir-704	SF01972
mir-708	SF00006
mir-711	SF00767
mir-719	SF02387
mir-721	SF01809
mir-725	SF02084
mir-729	SF01063
mir-732	SF02695
mir-736	SF00342
mir-73b	SF01490
mir-742	SF00099
mir-745	SF01249
mir-750	SF02094
mir-754d	SF00617
mir-76	SF00192
mir-763	SF00886
mir-767	SF00600
mir-770	SF00092
mir-786	SF00758
mir-79	SF00267

mir-648	SF00525
mir-64c	SF00294
mir-650	SF00417
mir-650d	SF00417
mir-654	SF00820
mir-658	SF00818
mir-661	SF01469
mir-663b	SF00912
mir-666	SF02726
mir-669a	SF00022
mir-669e	SF00022
mir-669i	SF00022
mir-669m	SF00022
mir-67	SF01566
mir-673	SF02638
mir-675b	SF01323
mir-679	SF02343
mir-683	SF00697
mir-688	SF01603
mir-693	SF02956
mir-698	SF00681
mir-701	SF01335
mir-705	SF00278
mir-709	SF01463
mir-713	SF00115
mir-71b	SF00770
mir-722	SF01118
mir-726	SF00345
mir-73	SF01490
mir-733	SF02220
mir-737	SF01529
mir-74	SF00627
mir-743a	SF00099
mir-74a	SF00507
mir-751	SF00769
mir-757	SF02867
mir-760	SF00790
mir-764	SF00763
mir-769	SF00111
mir-78	SF01287
mir-787	SF02384
mir-790	SF00212

mir-649	SF01049
mir-64d	SF00474
mir-650a	SF00417
mir-651	SF02163
mir-655	SF00031
mir-659	SF01305
mir-662	SF00323
mir-664	SF00567
mir-667	SF02654
mir-669b	SF00022
mir-669f	SF00022
mir-669j	SF00022
mir-669n	SF00376
mir-670	SF00976
mir-674	SF02233
mir-676	SF00157
mir-680	SF02621
mir-684	SF00957
mir-690	SF01090
mir-694	SF00555
mir-7	SF00534
mir-702	SF02180
mir-706	SF01237
mir-71	SF00171
mir-717	SF02583
mir-72	SF00719
mir-723	SF01140
mir-727	SF02041
mir-730	SF00623
mir-734	SF02000
mir-738	SF01145
mir-740	SF02602
mir-743b	SF00099
mir-74b	SF00507
mir-752	SF01019
mir-758	SF00031
mir-761	SF01229
mir-765	SF00918
mir-769b	SF00111
mir-784	SF01420
mir-788	SF01072
mir-791	SF01023

Continued on next page...

mir-792	SF02469
mir-796	SF02177
mir-7a	SF00534
mir-800	SF00805
mir-82	SF00493
mir-84b	SF01949
mir-871	SF00099
mir-875	SF01086
mir-879	SF02509
mir-881	SF00099
mir-883b	SF00099
mir-889	SF00031
mir-891b	SF01375
mir-9	SF00267
mir-920	SF02028
mir-927	SF00461
mir-92a	SF00009
mir-92e	SF02605
mir-934	SF01182
mir-938	SF01847
mir-940	SF01637
mir-944	SF02100
mir-956	SF00265
mir-96	SF00747
mir-963	SF01192
mir-967	SF02756
mir-971	SF00845
mir-975	SF01355
mir-978a	SF00682
mir-980	SF00702
mir-982b	SF00484
mir-985	SF00963
mir-989	SF02064
mir-992	SF02876
mir-995	SF00605
mir-999	SF00694

mir-793	SF01961
mir-797	SF02787
mir-7b	SF00534
mir-802	SF00238
mir-83	SF00076
mir-85	SF02338
mir-872	SF01100
mir-876	SF00154
mir-87a	SF01364
mir-882	SF01113
mir-885	SF02038
mir-890	SF00099
mir-892	SF00099
mir-90	SF01863
mir-921	SF02735
mir-927b	SF02345
mir-92b	SF00009
mir-93	SF00038
mir-935	SF02351
mir-939	SF00575
mir-941	SF02508
mir-95	SF00350
mir-957	SF00524
mir-960	SF01548
mir-964	SF02089
mir-968	SF01755
mir-972	SF02184
mir-976	SF02984
mir-978b	SF02354
mir-981	SF00046
mir-982c	SF00484
mir-986	SF01259
mir-99	SF00096
mir-993	SF00096
mir-996	SF00676
mir-99a	SF00096

mir-794	SF01732
mir-798	SF01813
mir-8	SF00226
mir-804	SF02456
mir-84	SF00441
mir-86	SF00233
mir-873	SF00566
mir-877	SF01055
mir-87b	SF01363
mir-883	SF00099
mir-887	SF00500
mir-891	SF01375
mir-892a	SF00099
mir-90b	SF01863
mir-922	SF01276
mir-928	SF01926
mir-92c	SF00009
mir-932	SF00302
mir-936	SF00868
mir-93a	SF00038
mir-942	SF02939
mir-954	SF01886
mir-958	SF01414
mir-961	SF01465
mir-965	SF00276
mir-969	SF01177
mir-973	SF02559
mir-977	SF02138
mir-979	SF02359
mir-982	SF00484
mir-983	SF02729
mir-987	SF00533
mir-990	SF01255
mir-993b	SF00096
mir-997	SF01526
mir-99b	SF00096

mir-795	SF02647
mir-799	SF00949
mir-80	SF00950
mir-81	SF00493
mir-84a	SF01949
mir-87	SF00381
mir-874	SF01164
mir-878	SF00099
mir-880	SF00099
mir-883a	SF00099
mir-888	SF00099
mir-891a	SF01375
mir-892b	SF00099
mir-92	SF00009
mir-924	SF02626
mir-929	SF00013
mir-92d	SF00009
mir-933	SF02790
mir-937	SF01658
mir-93b	SF00038
mir-943	SF01765
mir-955	SF01729
mir-959	SF01002
mir-962	SF02625
mir-966	SF00065
mir-970	SF00258
mir-974	SF01535
mir-978	SF00682
mir-98	SF00057
mir-982a	SF00484
mir-984	SF01912
mir-988	SF00165
mir-991	SF01678
mir-994	SF01085
mir-998	SF00785
mir-9a	SF00267

Papers Published During This Work

- Amaral, A. J., Andrade, J., Matos, A. M., Foxall, R. B., Matoso, P., Guerra-Assunção, J. A., Santa-Marta, M., et al. (2012). A role for miR-34c-5p in T-cell activation and response to HIV infection. *In preparation.*

- Choudry, H., Schoedel, J., Camps, C., Saini, H. K., Loewy, E., Reczko, M., Guerra-Assunção, J. A., et al. (2012). Transcriptomic and Epigenetic profiling of hypoxic breast cancer cells reveals deregulation of non-coding RNAs, non poly-A RNAs and natural anti-sense transcripts. *In preparation.*

- Howe, K., Clark, M., Torroja, C., Torrance, J., Berthelot, C., Muffato, M., Collins, J., et al. (2012). The Zebrafish Reference Genome Sequence and its Relationship to the Human Genome. *Under Review.*

- Guerra-Assunção, J. A., & Enright, A. J. (2012). Large-scale analysis of microRNA evolution. BMC Genomics, 13(1), 218. doi:10.1186/1471-2164-13-218

- Parts, L., Hedman, A. K., Keildson, S., Knights, A. J., Abreu-Goodger, C., van de Bunt, M., Guerra-Assunção, J. A., et al. (2012). Extent, causes, and consequences of small RNA expression variation in human adipose tissue. PLoS genetics, 8(5), e1002704. doi:10.1371/journal.pgen.1002704

- Hu, M., Ayub, Q., Guerra-Assunção, J. A., Long, Q., Ning, Z., Huang, N., Romero, I. G., et al. (2012). Exploration of signals of positive selection derived from genotype-based human genome scans using re-sequencing data. Human Genetics, 131(5), 665-674. doi:10.1007/s00439-011-1111-9

- Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature, 477(7364), 289-294. doi:10.1038/nature10413

- Guerra-Assunção, J. A., & Enright, A. J. (2010). MapMi: automated mapping of microRNA loci. BMC Bioinformatics, 11, 133. doi:10.1186/1471-2105-11-133

References

Akita, T., Takuno, S. & Innan, H. (2012). Modeling evolutionary growth of a microRNA-mediated regulation system. *Journal of Theoretical Biology*, **311**, 54–65. 67

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M.J., Tuschl, T. & Margalit, H. (2005). Clustering and conservation patterns of human microRNAs. *Nucleic Acids Research*, **33**, 2697–2706. 84, 95

Ambros, V. (2003). A uniform system for microRNA annotation. *RNA*, **9**, 277–279. 28, 58

Ambros, V. & Horvitz, H.R. (1987). The lin-14 locus of Caenorhabditis elegans controls the time of expression of specific postembryonic developmental events. *Genes & Development*, **1**, 398–414. 4

Arteaga-Vázquez, M., Caballero-Pérez, J. & Vielle-Calzada, J.P. (2006). A family of microRNAs present in plants and animals. *The Plant Cell*, **18**, 3355–3369. 8

Artzi, S., Kiezun, A. & Shomron, N. (2008). miRNAminer: a tool for homologous microRNA gene search. *BMC Bioinformatics*, **9**, 39. 33, 49

Ason, B., Darnell, D.K., Wittbrodt, B., Berezikov, E., Kloosterman, W.P., Wittbrodt, J., Antin, P.B. & Plasterk, R.H.A. (2006). Differences in vertebrate microRNA expression. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 14385–14389. 90

Axtell, M.J., Westholm, J.O. & Lai, E.C. (2011). Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biology*, **12**, 221. 6

Bader, D.A., Moret, B.M. & Yan, M. (2001). A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *Journal of computational biology*, **8**, 483–491. 88

Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P. & Bartel, D.P. (2008). The impact of microRNAs on protein output. *Nature*, **455**, 64–71. 16

Barker, D. & Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Computational Biology*, **1**, e3. 71, 82

Barker, D., Meade, A. & Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics*, **23**, 14–20. 72, 82

Barroso-delJesus, A., Lucena-Aguilar, G., Sanchez, L., Ligero, G., Gutierrez-Aranda, I. & Menendez, P. (2011). The Nodal inhibitor Lefty is negatively modulated by the microRNA miR-302 in human embryonic stem cells. *The FASEB journal*, **25**, 1497–1508. 95

Bartel, D.P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, **136**, 215–233. 14

Bartel, D.P. & Chen, C.Z. (2004). Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature Reviews Genetics*, **5**, 396–400. 14, 16, 21, 67, 90, 114

Batuwita, R. & Palade, V. (2009). microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, **25**, 989–995. 47, 54, 56, 57

Berezikov, E., Cuppen, E. & Plasterk, R.H.A. (2006). Approaches to microRNA discovery. *Nature Genetics*, **38 Suppl**, S2–7. 33

Bhaskaran, M., Wang, Y., Zhang, H., Weng, T., Baviskar, P., Guo, Y., Gou, D. & Liu, L. (2009). MicroRNA-127 modulates fetal lung development. *Physiological genomics*, **37**, 268–278. 95

Bohnsack, M.T., Czaplinski, K. & Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, **10**, 185–191. 5

Bonnet, E., Wuyts, J., Rouzé, P. & Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917. 55

Borchert, G.M., Holton, N.W., Williams, J.D., Hernan, W.L., Bishop, I.P., Dembosky, J.A., Elste, J.E., Gregoire, N.S., Kim, J.A., Koehler, W.W., Lengerich, J.C., Medema, A.A., Nguyen, M.A., Ower, G.D., Rarick, M.A., Strong, B.N., Tardi, N.J., Tasker, N.M., Wozniak, D.J., Gatto, C. & Larson, E.D. (2011). Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. *Mobile Genetic Elements*, **1**, 8–17. 36

Brennecke, J., Stark, A., Russell, R.B. & Cohen, S.M. (2005). Principles of microRNA-target recognition. *PLoS Biology*, **3**, e85. 16

Brest, P., Lapaquette, P., Souidi, M., Lebrigand, K., Cesaro, A., Vouret-Craviari, V., Mari, B., Barbry, P., Mosnier, J.F., Hébuterne, X., Harel-Bellan, A., Mograbi, B., Darfeuille-Michaud, A. & Hofman, P. (2011). A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nature Genetics*, **43**, 242–245. 27

Cai, X., Hagedorn, C.H. & Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, **10**, 1957–1966. 4

Camin, J.H. & Sokal, R.R. (1965). A method for deducing branching sequences in phylogeny. *Evolution; international journal of organic evolution*, 311–326. 65

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., Allen, J.E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R.N., Bailey, T.L., Bansal, M., Baxter, L., Beisel, K.W., Bersano, T., Bono, H., Chalk, A.M., Chiu, K.P., Choudhary, V., Christoffels, A., Clutterbuck, D.R., Crowe, M.L., Dalla, E., Dalrymple, B.P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C.F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T.R., Gojobori, T., Green, R.E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T.K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S.P.T., Kruger, A., Kummerfeld, S.K., Kurochkin, I.V., Lareau, L.F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K.C., Pavan, W.J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J.F., Ring, B.Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S.L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C.A.M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S.L., Tang, S., Taylor, M.S., Tegner, J., Teichmann, S.A., Ueda, H.R., van Nimwegen, E., Verardo, R., Wei, C.L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S.M., Teasdale, R.D., Liu, E.T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J.S., Hume, D.A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno,

H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., FANTOM Consortium & RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005). The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563. 2

Chang, T.H., Horng, J.T. & Huang, H.D. (2008). RNALogo: a new approach to display structural RNA alignment. *Nucleic Acids Research*, **36**, W91–6. 39

Chatterjee, S., Fasler, M., Büssing, I. & Grosshans, H. (2011). Target-Mediated Protection of Endogenous MicroRNAs in C. elegans. *Developmental Cell*, **20**, 388–396. 10

Cheloufi, S., Dos Santos, C.O., Chong, M.M.W. & Hannon, G.J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*, **465**, 584–589. 6, 40

Chen, K. & Rajewsky, N. (2006). Natural selection on human microRNA binding sites inferred from SNP data. *Nature Genetics*, **38**, 1452–1456. 107, 114

Chi, S.W., Hannon, G.J. & Darnell, R.B. (2012). An alternative mode of microRNA target recognition. *Nature Structural and Molecular Biology*, **19**, 321–327. 20

Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E., Blelloch, R., Schroth, G.P., Nusbaum, C. & Bartel, D.P. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes & Development*. 13, 29, 107

Cifuentes, D., Xue, H., Taylor, D.W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G.J., Lawson, N.D., Wolfe, S.A. & Giraldez, A.J. (2010). A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science*, **328**, 1694–1698. 6

Cullen, B.R. (2004). Transcription and processing of human microRNA precursors. *Molecular Cell*, **16**, 861–865. 105

Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences*, **23**, 324–328. 71

Dang, Y., Yang, Q., Xue, Z. & Liu, Y. (2011). RNA interference in fungi: pathways, functions, and applications. *Eukaryotic cell*, **10**, 1148–1155. 8

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection: The Preservation of Favored Races in the Struggle for Life* . Penguin Classics. 26

Dasgupta, B., Jiang, T., Kannan, S., Li, M. & Sweedyk, E. (1998). On the complexity and approximation of syntenic distance. *Discrete Applied Mathematics*, **88**, 59–82. 84

De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271. 23, 67, 74, 82

Doench, J.G. & Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. *Genes & Development*, **18**, 504–511. 16

Dollo, L. (1893). The laws of evolution. *Bull. Soc. Bel. Geol. Paleontol*, **7**, 164–166. 22, 65

Drinnenberg, I.A., Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R. & Bartel, D.P. (2009). RNAi in budding yeast. *Science*, **326**, 544–550. 8, 49

Eck, R. & Dayhoff, M. (1966). *Atlas of Protein Sequence and Structure*. National Biomedical Resources Foundation. 65

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797. 39

Ehrlich, J., Sankoff, D. & Nadeau, J.H. (1997). Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics*, **147**, 289–296. 88

Ellwanger, D.C., Büttner, F.A., Mewes, H.W. & Stümpflen, V. (2011). The sufficient minimal set of miRNA seed types. *Bioinformatics*. 16, 111

Enard, W. & Pääbo, S. (2004). Comparative primate genomics. *Annual review of genomics and human genetics*, **5**, 351–378. 76

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640. 2

Ender, C. & Meister, G. (2010). Argonaute proteins at a glance. *Journal of Cell Science*, **123**, 1819–1823. 9

Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90. 71

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C. & Marks, D.S. (2003). MicroRNA targets in Drosophila. *Genome Biology*, **5**, R1. 4, 17

Esteller, M. (2011). Non-coding RNAs in human disease. *Nature Reviews Genetics*, **12**, 861–874. 27, 106

Fabian, M.R., Sonenberg, N. & Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annual review of biochemistry*, **79**, 351–379. 14, 15

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L. & Carrington, J.C. (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PLoS ONE*, **2**, e219. 8

Farh, K.K.H., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B. & Bartel, D.P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, **310**, 1817–1821. 16, 21

Farris, J. (1977). Phylogenetic analysis under Dollo's Law. *Systematic Biology*, **26**, 77–88. 22, 65

Felsenstein, J. (1983). Parsimony in systematics: biological and statistical issues. *Annual review of ecology and systematics*, **14**, 313–333. 22, 64, 81

Felsenstein, J. (1993). PHYLIP (phylogeny inference package), version 3.5 c. *Distributed by the author*. 81

Filipowicz, W., Bhattacharyya, S.N. & Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*, **2008**, 102–114. 20

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A.K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y.A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X.M., Harrow, J., Herrero, J., Hubbard, T.J.P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A. & Searle, S.M.J. (2011a). Ensembl 2012. *Nucleic Acids Research*. 28, 30, 54

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H.S., Rios, D., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y.A., Trevanion, S., Vandrovcova, J., Vilella, A.J., White, S., Wilder, S.P., Zadissa, A., Zamora, J., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X.M., Herrero, J., Hubbard, T.J.P., Parker, A., Proctor, G., Vogel, J. & Searle, S.M.J. (2011b). Ensembl 2011. *Nucleic Acids Research*, **39**, D800–6. 30, 68, 80, 89

Friedlander, M., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S. & Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology*, **26**, 407–415. 13, 54, 57

Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W. & Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, **40**, 37–52. 54

Friedman, R.C., Farh, K.K.H., Burge, C.B. & Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, **19**, 92–105. 16, 17, 18, 72, 122

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A. & Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural and Molecular Biology*, **18**, 1139–1146. 17, 18, 122

Georges, M., Coppieters, W. & Charlier, C. (2007). Polymorphic miRNA-mediated gene regulation: contribution to phenotypic variation and disease. *Current opinion in genetics & development*, **17**, 166–176. 104

Gerlach, D., Kriventseva, E.V., Rahman, N., Vejnar, C.E. & Zdobnov, E.M. (2009). miROrtho: computational survey of microRNA genes. *Nucleic Acids Research*, **37**, D111–7. 33, 49

Gerrits, A., Walasek, M.A., Olthof, S., Weersing, E., Ritsema, M., Zwart, E., van Os, R., Bystrykh, L.V. & de Haan, G. (2012). Genetic screen identifies microRNA cluster 99b/let-7e/125a as a regulator of primitive hematopoietic cells. *Blood*, **119**, 377–387. 95

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J. & Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, **312**, 75–79. 4, 14, 23, 67, 74, 77, 90, 111

Goff, L.A., Davila, J., Swerdel, M.R., Moore, J.C., Cohen, R.I., Wu, H., Sun, Y.E. & Hart, R.P. (2009). Ago2 Immunoprecipitation Identifies Predicted MicroRNAs in Human Embryonic Stem Cells and Neural Precursors. *PLoS ONE*, **4**, e7192. 107

Gong, J., Tong, Y., Zhang, H.M., Wang, K., Hu, T., Shan, G., Sun, J. & Guo, A.Y. (2012). Genome-wide identification of SNPs in microRNA genes and

the SNP effects on microRNA target binding and biogenesis. *Human mutation*, **33**, 254–263. 106

Graur, D. & Li, W.H. (2000). *Fundamentals of molecular evolution*. Sinauer Associates Inc. 105

Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N. & Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. *Nature*, **432**, 235–240. 5

Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Research*, **32**, D109–11. 28, 32

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S.R. (2003). Rfam: an RNA family database. *Nucleic Acids Research*, **31**, 439–441. 28

Griffiths-Jones, S., Grocock, R.J., Van Dongen, S., Bateman, A. & Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, **34**, D140–4. 17, 28, 29, 32

Griffiths-Jones, S., Saini, H.K., Van Dongen, S. & Enright, A.J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Research*, **36**, D154–8. 28, 32, 68, 72, 80

Grimson, A., Farh, K.K.H., Johnston, W.K., Garrett-Engele, P., Lim, L.P. & Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, **27**, 91–105. 16, 17, 18, 122

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G. & Mello, C.C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, **106**, 23–34. 5

Guerra-Assunção, J.A. & Enright, A.J. (2010). MapMi: automated mapping of microRNA loci. *BMC Bioinformatics*, **11**, 133. 33, 61, 68, 80

Guerra-Assunção, J.A. & Enright, A.J. (2012). Large-scale analysis of microRNA evolution. *BMC Genomics*, **13**, 218. 63, 89, 95

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696–704. 39

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.C., Munschauer, M., Ulrich, A., Wardle, G.S., Dewell, S., Zavolan, M. & Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141. 20

Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, **15**, 1153–1160. 74

Hall, I.M., Noma, K.I. & Grewal, S.I.S. (2003). RNA interference machinery regulates chromosome dynamics during mitosis and meiosis in fission yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 193–198. 8

Heimberg, A.M., Sempere, L.F., Moy, V.N., Donoghue, P.C.J. & Peterson, K.J. (2008). MicroRNAs and the advent of vertebrate morphological complexity. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 2946–2950. 1, 64, 69, 100

Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I.L., Stadler, P.F., of Bioinformatics Computer Labs 2004, S. & 2005 (2006). The expansion of the metazoan microRNA repertoire. *BMC Genomics*, **7**, 25. 2, 24, 69, 74, 91, 100

Hofacker, I., Fontana, W., Stadler, P. & Bonhoeffer, L. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*. 12, 35

Houbaviy, H.B., Murray, M.F. & Sharp, P.A. (2003). Embryonic stem cell-specific MicroRNAs. *Developmental Cell*, **5**, 351–358. 76

Hsu, S.D., Lin, F.M., Wu, W.Y., Liang, C., Huang, W.C., Chan, W.L., Tsai, W.T., Chen, G.Z., Lee, C.J., Chiu, C.M., Chien, C.H., Wu, M.C., Huang, C.Y., Tsou, A.P. & Huang, H.D. (2011). miRTarBase: a database

curates experimentally validated microRNA-target interactions. *Nucleic Acids Research*, **39**, D163–9. 72

Hu, M., Ayub, Q., Guerra-Assunção, J.A., Long, Q., Ning, Z., Huang, N., Romero, I.G., Mamanova, L., Akan, P., Liu, X., Coffey, A.J., Turner, D.J., Swerdlow, H., Burton, J., Quail, M.A., Conrad, D.F., Enright, A.J., Tyler-Smith, C. & Xue, Y. (2012). Exploration of signals of positive selection derived from genotype-based human genome scans using re-sequencing data. *Human Genetics*, **131**, 665–674. 27, 36, 62

Huang, J.C., Babak, T., Corson, T.W., Chua, G., Khan, S., Gallie, B.L., Hughes, T.R., Blencowe, B.J., Frey, B.J. & Morris, Q.D. (2007). Using expression profiling data to identify human microRNA targets. *Nature Methods*, **4**, 1045–1049. 19

Hubbard, T.J.P., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S. & Flicek, P. (2009). Ensembl 2009. *Nucleic Acids Research*, **37**, D690–7. 34

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., Tuschl, T. & Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, **293**, 834–838. 5

Jacob, F., Perrin, D., Sánchez, C. & Monod, J. (1960). Operon: a group of genes with the expression coordinated by an operator. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, **250**, 1727–1729. 66

Jacobson, A.B. & Zuker, M. (1993). Structural analysis by energy dot plot of a large mRNA. *Journal of molecular biology*, **233**, 261–269. 12

Jazdzewski, K., Murray, E.L., Franssila, K., Jarzab, B., Schoenberg, D.R. & de la Chapelle, A. (2008). Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 7269–7274. 9, 27, 106, 116

Jiang, M., Anderson, J., Gillespie, J. & Mayne, M. (2008). uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, **9**, 192. 41

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X. & Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, **35**, W339–44. 54

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C. & Marks, D.S. (2004). Human MicroRNA targets. *PLoS Biology*, **2**, e363. 17

Jordan, G.E. & Piel, W.H. (2008). PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics*, **24**, 1641–1642. 39

Karadag, O., Tufan, A., Yazisiz, V., Ureten, K., Yilmaz, S., Cinar, M., Akdogan, A., Erdem, H., Ozturk, M.A., Pay, S. & Dinc, A. (2012). The factors considered as trigger for the attacks in patients with familial Mediterranean fever. *Rheumatology international*. 72

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N.A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T.G., Oliver, P.L., McIntyre, R.E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C.A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I.J., Czechanski, A., Guerra-Assunção, J.A., Donahue, L.R., Reinholdt, L.G., Payseur, B.A., Ponting, C.P., Birney, E., Flint, J. & Adams, D.J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**, 289–294. 116, 121

Kensche, P.R., van Noort, V., Dutilh, B.E. & Huynen, M.A. (2008). Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *Journal of the Royal Society, Interface / the Royal Society*, **5**, 151–170. 24, 67, 71

Kersey, P.J., Lawson, D., Birney, E., Derwent, P.S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kähäri, A., Kinsella, R.J., Kulesha, E., Maheswari, U., Megy, K., Nuhn, M., Proctor, G., Staines, D., Valentin, F., Vilella, A.J. & Yates, A. (2009). Ensembl Genomes: Extending Ensembl across the taxonomic space. *Nucleic Acids Research*, **38**, D563–D569. 34, 68, 80

Kersey, P.J., Staines, D.M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J.C., Hughes, D.S.T., Keenan, S., Kerhornou, A., Koscielny, G., Langridge, N., McDowall, M.D., Megy, K., Maheswari, U., Nuhn, M., Paulini, M., Pedro, H., Toneva, I., Wilson, D., Yates, A. & Birney, E. (2011). Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Research*, **40**, D91–D97. 28, 30

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, **217**, 624–626. 104

Kluge, A.G. & Farris, J.S. (1969). Quantitative phyletics and the evolution of anurans. *Systematic Biology*, **18**, 1–32. 65

Kozomara, A. & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, **39**, D152–7. 6, 28, 29, 32, 80

Kurihara, Y. & Watanabe, Y. (2004). Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12753–12758. 6

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858. 4, 10, 11

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W. & Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. *Current biology*, **12**, 735–739. 11

Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A. & Tuschl, T. (2003). New microRNAs from mouse and human. *RNA*, **9**, 175–179. 11

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L.,

Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J. & International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921. 2, 30

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., Lin, C., Socci, N.D., Hermida, L., Fulci, V., Chiaretti, S., Foà, R., Schliwka, J., Fuchs, U., Novosel, A., Müller, R.U., Schermer, B., Bissels, U., Inman, J., Phan, Q., Chien, M., Weir, D.B., Choksi, R., De Vita, G., Frezzetti, D., Trompeter, H.I., Hornung, V., Teng, G., Hartmann, G., Palkovits, M., Di Lauro, R., Wernet, P., Macino, G., Rogler, C.E., Nagle, J.W., Ju, J., Papavasiliou, F.N., Benzing, T., Lichter, P., Tam, W., Brownstein, M.J., Bosio, A., Borkhardt, A., Russo, J.J., Sander, C., Zavolan, M. & Tuschl, T. (2007). A mammalian microRNA

expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414. 76

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25–R25. 34

Lau, N.C., Lim, L.P., Weinstein, E.G. & Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, **294**, 858–862. 4

Le Quesne, W.J. (1974). The uniquely evolved character concept and its cladistic application. *Systematic Biology*, **23**, 513–517. 22

Lee, I., Ajay, S.S., Yook, J.I., Kim, H.S., Hong, S.H., Kim, N.H., Dhanasekaran, S.M., Chinnaiyan, A.M. & Athey, B.D. (2009). New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. *Genome Research*, **19**, 1175–1183. 14

Lee, M.T. & Kim, J. (2008). Self containment, a property of modular RNA structures, distinguishes microRNAs. *PLoS Computational Biology*, **4**, e1000150. 54, 56

Lee, R.C. & Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science*, **294**, 862–864. 4

Lee, R.C., Feinbaum, R.L. & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854. 1, 4

Lee, Y., Jeon, K., Lee, J.T., Kim, S. & Kim, V.N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal*, **21**, 4663–4670. 4

Lee, Y.S., Nakahara, K., Pham, J.W., Kim, K., He, Z., Sontheimer, E.J. & Carthew, R.W. (2004). Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*, **117**, 69–81. 9

Lewis, B.P., Burge, C.B. & Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20. 4, 14, 17, 20, 29, 68, 84, 122

Lewis, M.A., Quint, E., Glazier, A.M., Fuchs, H., De Angelis, M.H., Langford, C., Van Dongen, S., Abreu-Goodger, C., Piipari, M., Redshaw, N., Dalmay, T., Moreno-Pelayo, M.A., Enright, A.J. & Steel, K.P. (2009). An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nature Genetics*, **41**, 614–618. 27, 106

Li, S.C., Liao, Y.L., Ho, M.R., Tsai, K.W., Lai, C.H. & Lin, W.c. (2012). miRNA arm selection and isomiR distribution in gastric cancer. *BMC Genomics*, **13 Suppl 1**, S13. 6

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C. & Darnell, R.B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469. 20

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B. & Bartel, D.P. (2003). The microRNAs of Caenorhabditis elegans. *Genes & Development*, **17**, 991–1008. 12

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S. & Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773. 20

Lund, E., Liu, M., Hartley, R.S., Sheets, M.D. & Dahlberg, J.E. (2009). Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos. *RNA*, **15**, 2351–2363. 67, 77

Maddison, W.P. & Maddison, D.R. (2008). Mesquite: A modular system for evolutionary analysis. *Evolution; international journal of organic evolution*, **62**, 1103–1118. 80

Mallory, A.C., Hinze, A., Tucker, M.R., Bouché, N., Gasciolli, V., Elmayan, T., Lauressergues, D., Jauvion, V., Vaucheret, H. & Laux, T. (2009). Redundant and specific roles of the ARGONAUTE proteins AGO1

and ZLL in development and small RNA-directed gene silencing. *PLoS genetics*, **5**, e1000646. 9

Maragkakis, M., Vergoulis, T., Alexiou, P., Reczko, M., Plomaritou, K., Gousis, M., Kourtis, K., Koziris, N., Dalamagas, T. & Hatzigeorgiou, A.G. (2011). DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association. *Nucleic Acids Research*. 29

Marco, A., Hui, J.H.L., Ronshaugen, M. & Griffiths-Jones, S. (2010). Functional shifts in insect microRNA evolution. *Genome Biology and Evolution*, **2**, 686–696. 6

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86. 71

Maselli, V., Bernardo, D.D. & Banfi, S. (2008). CoGemiR: A comparative genomics microRNA database. *BMC Genomics*, **9**, 457. 33, 49

Mathelier, A. & Carbone, A. (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, **26**, 2226–2234. 13, 54

Matranga, C., Tomari, Y., Shin, C., Bartel, D.P. & Zamore, P.D. (2005). Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell*, **123**, 607–620. 5

Mattick, J.S. (2009). The genetic signatures of noncoding RNAs. *PLoS genetics*, **5**, e1000459. 3

Mclysaght, A., Hokamp, K. & Wolfe, K.H. (2002). Extensive genomic duplication during early chordate evolution. *Nature Genetics*, **31**, 200–204. 67

Mencía, Á., Modamio-Høybjør, S., Redshaw, N., Morín, M., Mayo-Merino, F., Olavarrieta, L., Aguirre, L.A., del Castillo, I., Steel, K.P., Dalmay, T., Moreno, F. & Moreno-Pelayo, M.A. (2009). Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nature Genetics*, **41**, 609–613. 106

Mendes, N.D., Freitas, A.T. & Sagot, M.F. (2009). Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Research*, **37**, 2419–2433. 34

Milinkovitch, M.C., Helaers, R., Depiereux, E., Tzika, A.C. & Gabaldón, T. (2010). 2x genomes–depth does matter. *Genome Biology*, **11**, R16. 68, 69

Mimouni, N.K., Lyngsø, R.B., Griffiths-Jones, S. & Hein, J. (2009). An analysis of structural influences on selection in RNA genes. *Molecular Biology and Evolution*, **26**, 209–216. 106, 109

Miska, E., Alvarez-Saavedra, E., Abbott, A., Lau, N.C., Hellman, A., McGonagle, S., Bartel, D.P., Ambros, V. & Horvitz, H.R. (2007). Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. *PLoS genetics*, **3**, e215. 21, 27

Morgulis, A., Gertz, E.M., Schäffer, A.A. & Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of computational biology*, **13**, 1028–1040. 122

Mu, X.J., Lu, Z.J., Kong, Y., Lam, H.Y.K. & Gerstein, M.B. (2011). Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Research*, **39**, 7058–7076. 106

Murphy, D., Dancis, B. & Brown, J.R. (2008). The evolution of core proteins involved in microRNA biogenesis. *BMC Evolutionary Biology*, **8**, 92. 2, 9

Murray, M.J., Saini, H.K., Van Dongen, S., Palmer, R.D., Muralidhar, B., Pett, M.R., Piipari, M., Thornton, C.M., Nicholson, J.C., Enright, A.J. & Coleman, N. (2010). The two most common histological subtypes of malignant germ cell tumour are distinguished by global microRNA profiles, associated with differential transcription factor expression. *Molecular cancer*, **9**, 290. 95

Nadeau, J.H. & Sankoff, D. (1998). Counting on comparative maps. *Trends in genetics*, **14**, 495–501. 88, 95

Nadeau, J.H. & Taylor, B.A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 814–818. 88, 92

Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**, 443–453. 81, 87

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, **3**, 418–426. 105, 123

Nekrutenko, A. & Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, **13**, 667–672. 12

Okamura, K., Hagen, J., Duan, H., Tyler, D. & Lai, E. (2007). The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in Drosophila. *Cell*, **130**, 89–100. 6

Olena, A.F. & Patton, J.G. (2009). Genomic organization of microRNAs. *Journal of cellular physiology*, **222**, 540–545. 10, 24, 89

Ooi, C.H., Oh, H.K., Wang, H.Z., Tan, A.L.K., Wu, J., Lee, M., Rha, S.Y., Chung, H.C., Virshup, D.M. & Tan, P. (2011). A densely interconnected genome-wide network of microRNAs and oncogenic pathways revealed using gene expression signatures. *PLoS genetics*, **7**, e1002415. 24

Parts, L., Hedman, Å.K., Keildson, S., Knights, A.J., Abreu-Goodger, C., van de Bunt, M., Guerra-Assunção, J.A., Bartonicek, N., Van Dongen, S., Mägi, R., Nisbet, J., Barrett, A., Rantalainen, M., Nica, A.C., Quail, M.A., Small, K.S., Glass, D., Enright, A.J., Winn, J., MuTHER Consortium, Deloukas, P., Dermitzakis, E.T., McCarthy, M.I., Spector, T.D., Durbin, R. & Lindgren, C.M. (2012). Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS genetics*, **8**, e1002704. 62

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E. & Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89. 4, 33, 47

Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. (2008). Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, **18**, 1814–1828. 85, 90, 101

Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448. 81

Pellegrini, M. (2012). Using phylogenetic profiles to predict functional relationships. *Methods in molecular biology*, **804**, 167–177. 24

Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D. & Yeates, T. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285–4288. 24, 66

Pillai, R.S., Bhattacharyya, S.N. & Filipowicz, W. (2007). Repression of protein synthesis by miRNAs: how many mechanisms? *Trends in Cell Biology*, **17**, 118–126. 16

Piriyapongsa, J. & Jordan, I.K. (2007). A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE*, **2**, e203. 76

Piriyapongsa, J., Marino-Ramirez, L. & Jordan, I.K. (2007). Origin and evolution of human microRNAs from transposable elements. *Genetics*, **176**, 1323–1337. 36

Pritchard, C.C., Cheng, H.H. & Tewari, M. (2012). MicroRNA profiling: approaches and considerations. *Nature Reviews Genetics*, **13**, 358–369. 11

Provost, P., Silverstein, R.A., Dishart, D., Walfridsson, J., Djupedal, I., Kniola, B., Wright, A., Samuelsson, B., Rådmark, O. & Ekwall, K.

(2002). Dicer is required for chromosome segregation and gene silencing in fission yeast cells. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 16648–16653. 8

Quach, H., Barreiro, L.B., Laval, G., Zidane, N., Patin, E., Kidd, K.K., Kidd, J.R., Bouchier, C., Veuille, M., Antoniewski, C. & Quintana-Murci, L. (2009). Signatures of purifying and local positive selection in human miRNAs. *American journal of human genetics*, **84**, 316–327. 27

Qureshi, I.A. & Mehler, M.F. (2012). Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nature reviews Neuroscience*, **13**, 528–541. 7

Rayner, K.J., Esau, C.C., Hussain, F.N., McDaniel, A.L., Marshall, S.M., van Gils, J.M., Ray, T.D., Sheedy, F.J., Goedeke, L., Liu, X., Khatsenko, O.G., Kaimal, V., Lees, C.J., Fernández-Hernando, C., Fisher, E.A., Temel, R.E. & Moore, K.J. (2011). Inhibition of miR-33a/b in non-human primates raises plasma HDL and lowers VLDL triglycerides. *Nature*, **478**, 404–407. 84

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R. & Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, **403**, 901–906. 4

Ritchie, W., Flamant, S. & Rasko, J.E.J. (2009). Predicting microRNA targets and functions: traps for the unwary. *Nature Methods*, **6**, 397–398. 18

Ritchie, W., Gao, D. & Rasko, J.E.J. (2012). Defining and providing robust controls for microRNA prediction. *Bioinformatics*, **28**, 1058–1061. 13, 57

Rödelsperger, C. & Dieterich, C. (2010). CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. *PLoS ONE*, **5**, e8861. 85

Rodrigo, A.G. & Learn, G.H. (2001). *Computational And Evolutionary Analysis of HIV Molecular Sequences*. Springer. 123

Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L. & Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. *Genome Research*, **14**, 1902–1910. 10

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H. & Telford, M.J. (2011). A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 298–306. 39

Ruby, J., Jan, C. & Bartel, D. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature*, **448**, 83–86. 6

Saini, H.K., Enright, A.J. & Griffiths-Jones, S. (2008). Annotation of mammalian primary microRNAs. *BMC Genomics*, **9**, 564. 29, 86

Sankoff, D., Ferretti, V. & Nadeau, J.H. (1997). Conserved segment identification. *Journal of computational biology*, **4**, 559–565. 83, 88

Saxena, A. & Carninci, P. (2010). Whole transcriptome analysis: what are we still missing? *Wiley interdisciplinary reviews Systems biology and medicine*, **3**, 527–543. 3

Shabalina, S.A. & Koonin, E.V. (2008). Origins and evolution of eukaryotic RNA interference. *Trends in ecology & evolution*, **23**, 578–587. 1, 8

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**, 308–311. 31, 105

Smit, A., Hubley, R. & Green, P. (2004). Repeat Masker. *Website*. 36

Sneath, P.H. (1957). The application of computers to taxonomy. *Journal of general microbiology*, **17**, 201–226. 22

Sokal, R.R. & Sneath, P.H.A. (1963). Principles of numerical taxonomy. W.H. Freeman. 22

Sontheimer, E.J. (2005). Assembly and function of RNA silencing complexes. *Nature Reviews Molecular Cell Biology*, **6**, 127–138. 5

Stark, A., Brennecke, J., Russell, R.B. & Cohen, S.M. (2003). Identification of Drosophila MicroRNA targets. *PLoS Biology*, **1**, E60. 4

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D.A., Sommer, S.S. & Rossi, J.J. (2009). SNPs in human miRNA genes affect biogenesis and function. *RNA*, **15**, 1640–1651. 9

Susumo, O. (1970). *Evolution by gene duplication.*. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag. 100

Susumo, O. (1973). Ancient linkage groups and frozen accidents. *Nature*, **244**, 259–262. 88

Tang, F., Kaneda, M., O'Carroll, D., Hajkova, P., Barton, S.C., Sun, Y.A., Lee, C., Tarakhovsky, A., Lao, K. & Surani, M.A. (2007). Maternal microRNAs are essential for mouse zygotic development. *Genes & Development*, **21**, 644–648. 76, 95

Tanzer, A. & Stadler, P.F. (2004). Molecular evolution of a microRNA cluster. *Journal of molecular biology*, **339**, 327–335. 2, 24

Tanzer, A. & Stadler, P.F. (2006). Evolution of microRNAs. *Methods in molecular biology*, **342**, 335–350. 2, 24

Tanzer, A., Amemiya, C.T., Kim, C.B. & Stadler, P.F. (2005). Evolution of microRNAs located within Hox gene clusters. *Journal of experimental zoology Part B Molecular and developmental evolution*, **304**, 75–85. 24

Tay, Y., Zhang, J., Thomson, A.M., Lim, B. & Rigoutsos, I. (2008). MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, **455**, 1124–1128. 14

Thomas, M., Lieberman, J. & Lal, A. (2010). Desperately seeking microRNA targets. *Nature Structural and Molecular Biology*, **17**, 1169–1174. 19

Thornton, J.E. & Gregory, R.I. (2012). How does Lin28 let-7 control development and disease? *Trends in Cell Biology*, **22**, 474–482. 95

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A. & Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science*, **302**, 1212–1215. 19

Van Dongen, S. (2000). Graph clustering by flow simulation. *Ph.D. Thesis, University of Utrecht*. 71, 81

Van Dongen, S., Abreu-Goodger, C. & Enright, A.J. (2008). Detecting microRNA binding and siRNA off-target effects from expression data. *Nature Methods*, **5**, 1023–1025. 18

van Rooij, E., Quiat, D., Johnson, B.A., Sutherland, L.B., Qi, X., Richardson, J.A., Kelm, R.J. & Olson, E.N. (2009). A family of microRNAs encoded by myosin genes governs myosin expression and muscle performance. *Developmental Cell*, **17**, 662–673. 84

Vasudevan, S., Tong, Y. & Steitz, J.A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science*, **318**, 1931–1934. 14

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg,

N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. (2001). The sequence of the human genome. *Science*, **291**, 1304–1351. 2, 30

Vergoulis, T., Vlachos, I.S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T. & Hatzigeorgiou, A.G. (2011). TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Research*. 29

Vigorito, E., Perks, K.L., Abreu-Goodger, C., Bunting, S., Xiang, Z., Kohlhaas, S., Das, P.P., Miska, E.A., Rodriguez, A., Bradley, A., Smith, K.G.C., Rada, C., Enright, A.J., Toellner, K.M., Maclennan,

I.C.M. & Turner, M. (2007). microRNA-155 regulates the generation of immunoglobulin class-switched plasma cells. *Immunity*, **27**, 847–859. 4

Vilella, A., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. & Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, **19**, 327–335–327–335. 30

Vilella, A.J., Birney, E., Flicek, P. & Herrero, J. (2011). Considerations for the inclusion of 2x mammalian genomes in phylogenetic analyses. *Genome Biology*, **12**, 401. 68, 69

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell*, **136**, 669–687. 6

Wang, X. (2008). miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA*, **14**, 1012–1017. 30

Wang, Z. & Yang, B. (2010). *MicroRNA Expression Detection Methods*. Springer. 12

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. & Barton, G.J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191. 39

Wheeler, B.M. (2008). Automating the annotation and discovery of MicroRNA in multi-species high-throughput 454 Sequencing. *Ph. D. Thesis, North Carolina State University*. 69

Wheeler, B.M., Heimberg, A.M., Moy, V.N., Sperling, E.A., Holstein, T.W., Heber, S. & Peterson, K.J. (2009). The deep evolution of metazoan microRNAs. *Evolution & development*, **11**, 50–68. 65

Wightman, B., Ha, I. & Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, **75**, 855–862. 1, 4

Yoda, M., Kawamata, T., Paroo, Z., Ye, X., Iwasaki, S., Liu, Q. & Tomari, Y. (2009). ATP-dependent human RISC assembly pathways. *Nature Structural and Molecular Biology*, **17**, 17–23. 5

Yuan, Z., Sun, X., Liu, H. & Xie, J. (2011). MicroRNA genes derived from repetitive elements and expanded by segmental duplication events in mammalian genomes. *PLoS ONE*, **6**, e17666. 76

Zhang, L., Hou, D., Chen, X., Li, D., Zhu, L., Zhang, Y., Li, J., Bian, Z., Liang, X., Cai, X., Yin, Y., Wang, C., Zhang, T., Zhu, D., Zhang, D., Xu, J., Chen, Q., Ba, Y., Liu, J., Wang, Q., Chen, J., Wang, J., Wang, M., Zhang, Q., Zhang, J., Zen, K. & Zhang, C.Y. (2012). Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Research*, **22**, 107–126. 8

Zhang, R., Wang, Y.Q. & Su, B. (2008). Molecular evolution of a primate-specific microRNA family. *Molecular Biology and Evolution*, **25**, 1493–1502. 76

Zoffal, M. & Grewal, S. (2006). RNAi-mediated heterochromatin assembly in fission yeast. *Cold Spring Harbor symposia on quantitative biology*, **71**, 487–496. 8

Zorc, M., Skok, D.J., Godnic, I., Calin, G.A., Horvat, S., Jiang, Z., Dovc, P. & Kunej, T. (2012). Catalog of microRNA seed polymorphisms in vertebrates. *PLoS ONE*, **7**, e30737. 107