DISSERTATIONS IN
# HEALTH SCIENCES

Liisa Heikkinen

# *Computational Analysis of Small Non-coding RNAs in Model Systems*

Publications of the University of Eastern Finland
*Dissertations in Health Sciences*

UNIVERSITY OF
EASTERN FINLAND

LIISA HEIKKINEN

# *Computational Analysis of Small Non-coding RNAs in Model Systems*

To be presented by permission of the Faculty of Health Sciences, University of Eastern Finland for public examination in Auditorium L22, Snellmania, Kuopio, on Saturday, March 15th 2014, at 12

Publications of the University of Eastern Finland
Dissertations in Health Sciences
Number 219

Department of Neurobiology, A.I. Virtanen Institute, Faculty of Health Sciences, University of Eastern Finland
Kuopio
2014

Kopijyvä Oy
Kuopio, 2014
Finland



Distributor:
University of Eastern Finland
Kuopio Campus Library
P.O.Box 1627
FI-70211 Kuopio, Finland
http://www.uef.fi/kirjasto

ISBN (print): 978-952-61-1385-2
ISBN (pdf): 978-952-61-1386-9
ISSN (print): 1798-5706
ISSN (pdf): 1798-5714
ISSN-L: 1798-5706

Author's address: A. I. Virtanen Institute for Molecular Sciences
University of Eastern Finland
KUOPIO
FINLAND

Supervisors: Research Director Garry Wong, Ph.D.
A. I. Virtanen Institute for Molecular Sciences
University of Eastern Finland
KUOPIO
FINLAND

Professor Mikko Kolehmainen, Ph.D.
Department of Environmental Science
University of Eastern Finland
KUOPIO
FINLAND

Docent Markus Storvik, Ph.D.
School of Pharmacy
University of Eastern Finland
KUOPIO
FINLAND

Reviewers: Associate Professor Cecilia Sarmiento, Ph.D.
Department of Gene Technology
Tallinn University of Technology
TALLINN
ESTONIA

Sam Griffiths-Jones, Ph.D.
Faculty of Life Sciences
University of Manchester
MANCHESTER
UNITED KINGDOM

Opponent: Docent Emily Knott, Ph.D.
Department of Biological and Environmental Science
University of Jyväskylä
JYVÄSKYLÄ
FINLAND

Heikkinen, Liisa
Computational Analysis of Small Non-coding RNAs in Model Systems
University of Eastern Finland, Faculty of Health Sciences
Publications of the University of Eastern Finland. Dissertations in Health Sciences Number 219. 2014. 64 p.

ISBN (print): 978-952-61-1385-2
ISBN (pdf): 978-952-61-1386-9
ISSN (print): 1798-5706
ISSN (pdf): 1798-5714
ISSN-L: 1798-5706

## ABSTRACT

The modulation of gene expression by small non-coding RNAs (ncRNAs) is a recently discovered regulatory mechanism in eukaryotes. This thesis aims to deepen the understanding of small ncRNA biology by using computational approaches. The main focus is on microRNAs (miRNAs) which constitute a large family of small ncRNAs that have emerged as key post-transcriptional regulators of gene expression. miRNAs are predicted to control most of the protein-coding genes and a large number of cellular pathways appear to be modulated by miRNAs. First, we aim to gain novel information about the transcriptional regulation of miRNAs. By using established sequence motif discovery tools, we identify a novel conserved sequence element GANNNNGA, which is found upstream of all miRNAs in nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. This motif may have a role in miRNA transcriptional or post-transcriptional regulation, or it may serve as a recognition factor for miRNA biogenesis. Secondly, we develop a novel tool, mirSOM, based on self-organizing map, for predicting miRNA targets in *C. elegans*. As mirSOM applies unsupervised learning, it avoids bias towards the characteristics of the small set of available, experimentally verified positive and negative target sites. In comparison with seven other miRNA target prediction tools, mirSOM works best in finding the verified true and false miRNA-target gene relationships, suggesting that miRNA target prediction can be improved by the use of machine learning methods. Thirdly, *de novo* sequencing of the genome and transcriptome of *Panagrellus redivivus* is accomplished, where we annotate the complement of *P. redivivus* miRNAs, thus providing a novel powerful resource for comparative genomics in nematode phylum. Finally, by using deep sequencing of small RNAs, we profile the miRNA expression specific to human embryonic stem cells (hESCs). For the first time, we also report the discovery of microRNA-offset RNAs (moRNAs) in hESCs and present the specific expression patterns of moRNAs in hESCs. This finding is a step towards understanding the complex network of small ncRNAs maintaining the unique characteristics of stem cells.

In conclusion, this thesis provides novel resources for the research of small ncRNAs and highlights the benefits of using computational analysis and bioinformatics in generating testable biological hypotheses and in advancing our knowledge.

National Library of Medicine Classification: QU 26.5; QU 58.7; QU 460
Medical Subject Headings: Small non-coding RNA, miRNA, computational biology, bioinformatics, machine learning, neural network model, high-throughput RNA-sequencing, *Caenorhabditis elegans*

Heikkinen, Liisa
Bioinformatiikan menetelmiä pienten ei-koodaavien RNA-molekyylien analysointiin biologisissa mallisysteemeissä
Itä-Suomen yliopisto, terveystieteiden tiedekunta
Publications of the University of Eastern Finland. Dissertations in Health Sciences Numero 219. 2014. 64 s.

ISBN (print): 978-952-61-1385-2
ISBN (pdf): 978-952-61-1386-9
ISSN (print): 1798-5706
ISSN (pdf): 1798-5714
ISSN-L: 1798-5706

## TIIVISTELMÄ

Pienillä ei-koodaavilla RNA-molekyyleillä on tärkeä tehtävä geenien ja sitä kautta useiden eri biologisten prosessien säätelyssä. Tässä tutkimuksessa pyrittiin laskennallisen biologian avulla etsimään uutta tietoa pienistä ei-koodaavista RNA-molekyyleistä, erityisesti mikro-RNA:ista (miRNA), jotka on jo jonkin aikaa tunnettu lähetti-RNA:n hiljentäjinä aitotumallisissa eliöissä. Lisätäksemme tietämystä miRNA-geenien säätelymekanismeista, tutkimme niiden ylävirta-alueita vakiintuneilla motiivien tunnistukseen tarkoitetuilla bioinformatiikka-työkaluilla nematodeissa *Caenorhabditis elegans* ja *Caenorhabditis briggsae*. Löysimme ennestään tuntemattoman sekvenssi-motiivin, GANNNNGA, joka esiintyy lähellä jokaisen miRNA-geenin alkua tutkituissa lajeissa. Löydetyllä motiivilla voi olla rooli joko miRNA:n transkriptiossa tai transkriptin myöhemmässä säätelyssä, tai se voi toimia miRNA-geenin tunnisteena genomissa. Kehitimme myös uuden, koneoppimista hyödyntävän menetelmän miRNA-kohdegeenien ennustamiseen. Koska miRNA-kohdegeenejä tunnetaan vain muutamia, käytimme algoritmia joka ei tarvitse oikeat ja väärät mallit sisältävää opetusjoukkoa, vaan perustuu ohjaamattomaan oppimiseen. Kyseinen menetelmä, mirSOM, hyödyntää neuroverkko-arkkitehtuuria nimeltään itse-organisoituva kartta (SOM). Vertailussa seitsemän muun miRNA-kohdegeenien ennustusohjelman kanssa mirSOM löytää tunnetut oikeat kohdegeenit ja hylkää väärät suurimmalla varmuudella, mikä viittaa siihen että miRNA-kohdegeenien ennustamista voidaan parantaa kone-oppimismenetelmien avulla. Nematodin *Panagrellus redivivus* genomi-projektissa sekvensoimme pienet RNA:t ja annotoimme miRNAomin. Lopputulos tarjoaa uuden, arvokkaan resurssin nematodien vertailevalle genomiikalle. Lopuksi, käyttämällä pienten RNA:iden syväsekvensointia, määritimme ihmisen embryonaalisissa kantasoluissa (hESC) esiintyvät miRNA:t ja niiden ekspressioprofiilin. Lisäksi raportoimme miRNA-geenien viereisestä alueesta syntyvien miRNA 'off-set' RNA -molekyylien (moRNA) löytymisestä hESC-soluista sekä tiettyjen moRNA sekvenssien kantasoluspesifisen ekspression. Tämä löytö on askel kohti embryonaalisten kantasolujen uniikkeja ominaisuuksia ylläpitävän monimutkaisen säätelyverkon ymmärtämistä.

Tämä väitöskirjatutkimus tarjoaa uusia resursseja pienten ei-koodaavien RNA-molekyylien tutkimukseen ja korostaa laskennallisen analyysin ja bioinformatiikan merkitystä luotaessa testauskelpoisia biologisia hypoteeseja ja sitä kautta tiedon lisäämisessä.

Luokitus: QU 26.5; QU 58.7; QU 460
Yleinen Suomalainen asiasanasto: mikro-RNA, bioinformatiikka, neuroverkot, koneoppiminen, sekvensointi, sukkulamadot, kantasolut

*To Seppo, Hilla and Suvi*

# Acknowledgements

This work was carried out in the Department of Biosciences and in the Department of Neurobiology, AI Virtanen Institute for Molecular Sciences at the University of Kuopio / University of Eastern Finland during the years 2007-2014. It has been possible only because of the guidance, contribution, and support from many different people and funding from several sources.

First and foremost, I would like to express my deepest gratitude to my principal supervisor Research Director Garry Wong. He has supported me throughout this work with patience and motivation, while allowing me the space to carry on my own way. His wise advice and positive outlook have saved my day dozens of times.

I am very grateful to my supervisor Professor Mikko Kolehmainen for his invaluable insights and suggestions which really aided in creating mirSOM. Many thanks to Docent Markus Storvik for being such a great help in practicalities and encouraging me, in particular when I had just started this work.

I am indebted to PhD Suvi Asikainen for answering all my questions about wet lab and expanding my knowledge of biology. Working with Suvi in several small RNA projects has been fascinating and fun.

I would like to express my deep appreciation to Associate Professor Cecilia Sarmiento and Dr Sam Griffiths-Jones for pre-reviewing my dissertation and for their valuable and sagacious feedback. Sincere thanks to Docent Emily Knott for accepting the invitation to act as the opponent in my thesis.

Many thanks to all my co-authors in the manuscripts in this thesis: Thanks to Jagan Srinivasan, Adler R. Dillman, Ali Mortazavi, Marissa Macchietto, Merja Lakso, Kelley Fracchia and Igor Antoshechkin for valuable co-work in the *P. redivivus* genome project. Additional thanks to Professor Paul W. Sternberg for a chance to visit his worm laboratory and WormBase in Caltech. Thanks to Juuso Juhila, Frida Holm, Jere Weltner, Ras Trokovic, Milla Mikkola, Sanna Toivonen, Diego Balboa, Riina Lampela, Katherine Icay and Timo Tuuri for their contribution in the hESC small RNA project. Especially, thanks to Professor Outi Hovatta from Karolinska Institute, Professor Timo Otonkoski and Docent Iiris Hovatta from Helsinki University for their time and invaluable advice during that project.

Many thanks to all my fellow group members over these years in the Wong lab for creating a pleasant working atmosphere. Especially, thanks to Vuokko Aarnio, Martina Rudgalvyte and Juhani Peltonen for keeping the spirit up these days.

Finally, I would like to express my heartfelt thanks to my friends and family - life would be so boring without you! Thanks to Hilla and Suvi for bringing so much joy into my life. Thank you, Seppo, just for being there and for understanding while I have been pursuing my research dream.

This work was made possible through the financial support from Saastamoinen Foundation, the Finnish Cultural Foundation Central Fund, the Finnish Cultural Foundation North Savo Regional Fund, Biocenter Finland, Doctoral Program in Molecular Medicine and Faculty of Health Sciences at University of Eastern Finland.

Kuopio, February 2014

Liisa Heikkinen

# List of the original publications

This dissertation is based on the following original publications:

I Heikkinen L, Asikainen S and Wong G. Identification of phylogenetically conserved sequence motifs in microRNA 5' flanking sites from *C. elegans* and *C. briggsae*. *BMC Molecular Biology 9:105, 2009.*

II Heikkinen L, Kolehmainen M and Wong G. Prediction of microRNA targets in *C. elegans* using a self-organizing map. *Bioinformatics, 27(9):1247-1254, 2011.*

III Srinivasan J, Dillman A R, Macchietto M G, Heikkinen L, Lakso M, Fracchia K M, Antoshechkin I, Mortazavi A, Wong G and Sternberg P W. The draft genome and transcriptome of *Panagrellus redivivus* are shaped by the harsh demands of a free-living lifestyle. *Genetics, 193(4):1279-95, 2013*.

IV Asikainen S*, Heikkinen L*, Juhila J, Holm F, Weltner J, Trokovic R, Mikkola M, Toivonen S, Balboa D, Lampela R, Icay K, Tuuri T, Otonkoski T, Wong G, Hovatta, O. MicroRNA-offset RNAs are abundantly and specifically expressed in human embryonic stem cells. Submitted. *indicates equal contribution.

# Contents

# Abbreviations

| | |
|---|---|
| 3′ UTR | 3 prime untranslated region |
| 5′ UTR | 5 prime untranslated region |
| AGO | Argonaute protein |
| bp | base pair |
| *C. briggsae* | *Caenorhabditis briggsae* |
| *C. elegans* | *Caenorhabditis elegans* |
| CLIP-Seq | cross-linking immunoprecipitation-high-throughput sequencing |
| DNA | deoxyribonucleic acid |
| endo-siRNA | endogenous siRNA |
| exo-siRNA | exogenous siRNA |
| GFP | green fluorescent protein |
| hESC | human embryonic stem cell |
| kbp | kilo base pair |
| mESC | mouse embryonic stem cell |
| miRNA | microRNA |
| moRNA | miRNA-offset RNA |
| mRNA | messenger RNA |
| ncRNA | non-coding RNA |
| NGS | next-generation sequencing |
| nt | nucleotide |
| PCR | polymerase chain reaction |
| piRNA | Piwi-interacting RNA |
| *P. redivivus* | *Panagrellus redivivus* |
| piRISC | piRNA-induced silencing complex |
| Pol II | RNA Polymerase II |
| Pol III | RNA Polymerase III |
| pre-miRNA | miRNA precursor |
| pre-mRNA | precursor mRNA |
| pri-miRNA | miRNA primary precursor |
| qPCR | quantitative real-time PCR |
| RACE | rapid amplification of cDNA ends |
| RISC | RNA-induced silencing complex |
| RNA | ribonucleic acid |
| RNAi | RNA interference |
| RNA-Seq | RNA sequencing |
| rRNA | ribosomal RNA |
| SILAC | stable isotope labeling by amino acids in cell culture |
| siRNA | short interfering RNA |
| snRNA | small nuclear RNA |
| snoRNA | small nucleolar RNA |
| SOM | self-organizing map |
| TE | transposable element |
| TFBS | transcription factor binding site |
| tRNA | transfer RNA |
| TSS | transcription start site |
| TUT | terminal uridyl transferase |

# 1 Introduction

Small non-coding RNAs (ncRNAs) are functional RNA molecules that are shorter than 200 nucleotides (nt) and are not translated into proteins. They make up much of the RNA content of a cell and are involved in essential regulatory mechanisms in most eukaryotic organisms (reviewed in Aalto and Pasquinelli, 2012). Many small ncRNA families are established and remain under active investigation, while novel classes of small ncRNAs are continuously discovered and their biogenesis pathways and functions are being introduced (Lee, Feinbaum and Ambros, 1993; Wightman et al., 1993; Fire et al., 1998; Aravin, Hannon and Brennecke, 2007; Taft et al., 2009; Shi et al., 2009; Djebali et al., 2012). One of the best understood classes of small ncRNAs are microRNAs (miRNAs) which down-regulate gene expression by targeting the messenger RNA (mRNA) for translational inhibition, degradation, deadenylation or destabilization (reviewed in Bartel, 2004; Shukla et al., 2011). Acting at the post-transcriptional level, miRNAs may alter the expression of a significant portion of protein-encoding genes and affect nearly every cellular pathway (Boehm and Slack, 2006; Hwang and Mendell, 2006; Friedman et al., 2009; Ambros, 2011). Coupled with the fact that miRNAs and their functions are widely conserved, this implies that these tiny molecules are an ancient and essential part of the gene regulatory network (Sempere et al., 2006; Christodoulou et al., 2010).

Since the discovery of the first miRNA, *lin-4*, in *Caenorhabditis elegans* twenty years ago (Lee, Feinbaum and Ambros, 1993; Wightman et al., 1993), enormous advances in understanding miRNA biology have been made, including identification of over 20,000 miRNA genes in over 200 species (Kozomara and Griffiths-Jones, 2011), specification of multiple miRNA biogenesis pathways (reviewed in Winter et al., 2009) and revealing the principles of miRNA target regulation (reviewed in Bartel, 2009). Less well understood are the transcriptional regulation of miRNA genes and the whole repertoire of the targets each miRNA regulates. Understanding the miRNA transcription and determining their regulators and targets, however, are crucial in order to identify the specific role for each miRNA in gene regulatory networks. Further, while the genome sequence and the majority of miRNAs of several model organisms like humans and *C. elegans* are presently known, sequencing of additional species and annotation of their small RNAs is needed for identifying conserved functional elements related to miRNAs and in enhancing the understanding of their functions. Because the expression of many miRNAs and the gene pathways they regulate are cell type specific, profiling of miRNA expression in different developmental stages and cell types will further aid in elucidating their functions.

In this thesis, novel information about different fields of miRNA biology was gained with computational methods. In Publication I, new insights into transcriptional regulation of miRNA genes were investigated by examining the miRNA upstream regions of *C. elegans* and *C. briggsae* for conserved sequence motifs. A novel motif, GANNNNGA, was found with conserved frequency distribution upstream of all miRNAs in these two nematodes. The function of this motif is not yet elucidated, but it may have a role in miRNA transcriptional or post-transcriptional regulation or it may serve as a recognition factor for miRNA biogenesis. Publication II shows how unsupervised learning can be applied to predict miRNA targets and introduces a novel tool for miRNA target prediction in *C. elegans*. The *de novo* sequencing of the genome, transcriptome, and small RNAs of *Panagrellus redivivus* reported in Publication III provides a powerful resource for comparative genomics. It is the first free-living worm genome sequenced not belonging to *Caenorhabditis* family, thus highlighting the common features with the genome of *C. elegans*. In Publication IV, the miRNA profile of human embryonic stem cells (hESCs) was characterized using small RNA deep sequencing data. For the first time, also microRNA-offset RNAs (moRNAs) were observed in hESCs, and their specific expression patterns in comparison to human fibroblasts were characterized.

# 2 Review of the literature

## 2.1 MIRNAS

miRNAs are single-stranded, small, ~22 nucleotides (nt), ncRNAs, which act as guide molecules in post-transcriptional gene repression (reviewed in Bartel, 2004; Shukla et al., 2011). miRNAs associate with specific Argonaute (AGO) family proteins in RNA-induced silencing complex (RISC) which they guide to cognate mRNA, causing silencing of the target by the AGO protein (Hutvagner and Zamore, 2002; Mourelatos et al., 2002). The core element in the recognition of the target is the miRNA "seed", which covers nucleotides 2-8 from the miRNA 5′ end and typically has a perfect, or near perfect, match with the target mRNA 3′ untranslated region (3′ UTR) (Lee, Feinbaum and Ambros, 1993; Wightman et al., 1993; Reinhart et al., 2000). Since each miRNA has hundreds of putative targets, they may alter the expression of most of the protein-coding genes (Brennecke et al., 2005; Lim et al., 2005; Xie et al., 2005; Friedman et al., 2009). As many miRNAs and their target genes are well conserved in eukaryotic organisms (Pasquinelli et al., 2000; Chen and Rajewsky, 2006a; Friedman et al., 2009), miRNAs are regarded as a vital and ancient component of genetic regulation. A growing body of evidence shows that miRNAs have an important role in a wide range of biological processes, including developmental timing, cell proliferation and differentiation, cell death and metabolic control (reviewed in Boehm and Slack, 2006; Hwang and Mendell, 2006; Ambros, 2011). Consequently, mutation in miRNA sequence or dysfunction of miRNA biogenesis may cause many diseases, such as cancer, cardiovascular disease or metabolic disorders (Ono, Kuwabara and Han, 2011; Rottiers and Näär, 2012; Zhong, Coukos and Zhang, 2012). Differential expression of miRNAs between different cell types and tissues makes them ideal biomarkers for detection of diseases and targets for therapeutic intervention (reviewed in Broderick and Zamore, 2011; Nana-Sinkam and Croce, 2012). Moreover, it has been recently discovered that miRNAs can also act in post-transcriptional up-regulation and transcriptional silencing of protein coding genes (Kim et al., 2008; Vasudevan, 2012). Indeed, although miRNAs mostly work in the cytoplasm, a subset of them is predominantly found in nucleus where they are transported back after maturation. In the nucleus, miRNAs can regulate gene expression by binding with high complementarity to gene promoter regions (Castanotto et al., 2009; Weinmann et al., 2009; Liao et al., 2010). The first example of promoter targeting miRNA in human cells was *miR-373*, which can activate E-Cadherin (CDH1) and cold-shock domain-containing protein C2 (CSDC2). Both of these genes contain putative *miR-373* target sites with at least 80% sequence complementarity in their promoters, and it has been shown that the activation of

these genes depends on Dicer and involves recruitment of RNA Polymerase II (Pol II) at the promoter region (Place et al., 2008).

## 2.2 DISCOVERY OF MIRNAS

Simultaneous efforts of Victor Ambros' and Gary Ruvkun's laboratories in the early 1990s led to the discovery of the first miRNA *lin-4* in *C. elegans* (Lee, Feinbaum and Ambros, 1993; Wightman et al., 1993). They reported that *lin-4* gene, that was known to control developmental timing in *C. elegans* (Chalfie et al., 1981), did not encode a protein but instead they noticed two very short transcripts, 61 and 22 nt long. The longer RNA molecule was predicted to form a stem loop and proposed to be the precursor of the shorter one. The Ambros and Ruvkun laboratories also found that these short RNAs were complementary to a repeated sequence in the 3' UTR of *lin-14* mRNA, a region which was earlier shown to be required for the normal down-regulation in *lin-14* protein level during *C. elegans* development (Wightman et al., 1991), and postulated that *lin-4* down-regulates the translation of *lin-14* mRNA to protein via an antisense RNA-RNA interaction.

After the finding of *lin-4*, it took seven years before the second miRNA, *let-7*, was reported by Ruvkun's laboratory (Reinhart et al., 2000). Like *lin-4*, also *let-7* regulates developmental timing in *C. elegans*, but while *lin-4* appeared to be worm specific, *let-7* sequence and its temporal regulation function were found to be highly conserved across species (Pasquinelli et al., 2000). This observation led to increased interest in miRNAs, and very soon, in year 2001, a landmark set of ~100 miRNA genes were reported to be found in worms, flies and mammals (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Today, the latest release 20 (June 2013) of miRNA curation database miRBase (Kozomara and Griffiths-Jones, 2011) contains 24521 hairpin precursor miRNAs expressing 30424 mature miRNA products, identified in 206 species including animals, plants, unicellular algae and viruses (Table 1). In addition to *de novo* miRNA studies, novel miRNAs are discovered continuously in the widely studied model species like human and *C. elegans*, and there is no consensus estimate on the upper limit for their amount.

## 2.3 MIRNA EVOLUTION

It seems that miRNAs as a class of gene regulators have been present very early in the animal evolution, perhaps since the last common ancestor of eukaryotes about a billion years ago (Axtell, Westholm and Lai, 2011; Tarver, Donoghue and Peterson, 2012). Studies of the conservation of miRNAs across early branching animal phyla have revealed several characteristics of miRNA evolution in animals (reviewed in Berezikov, 2011). Paralogous miRNA genes, which have significant sequence homology and often identical seed regions

*Table 1.* The miRNA count for selected species in miRBase release 20 (June 2013).

| **Species** | **Common name** | **miRNA hairpins** | **mature miRNAs** |
|---|---|---|---|
| *Homo sapiens* | human | 1872 | 2578 |
| *Mus musculus* | house mouse | 1186 | 1908 |
| *Ciona intestinalis* | vase tunicate | 348 | 550 |
| *Danio rerio* | zebrafish | 346 | 255 |
| *Arabidopsis thaliana* | thale cress | 298 | 337 |
| *Drosophila melanogaster* | fruit fly | 238 | 426 |
| *Caenorhabditis elegans* | roundworm | 223 | 368 |
| *Chlamydomonas reinhardtii* | unicellular green alga | 50 | 85 |
| *Human herpesvirus-5* | human cytomegalovirus | 15 | 26 |

with each other, are called miRNA families (Ambros et al., 2003). These families have likely arisen through gene duplications during evolution (Hertel et al., 2006). On the bilaterian lineage, there are up to thirty miRNA families conserved in all species (Hertel et al., 2006; Prochnik et al., 2007; Christodoulou et al., 2010), but only one of them, *mir-100*, is also present in cnidarians, suggesting that *mir-100* appeared ~650 million years ago, in the origin of multicellularity (Grimson et al., 2008; Wheeler et al., 2009; Griffiths-Jones et al., 2011). Thus it seems that there was an explosion of miRNAs in the first bilaterian animals with clear body structures including head and tail, upside and downside (Hertel et al., 2006). The next increase of the miRNA count is observed in the vertebrate lineage, and a further increase in placental mammals (Hertel et al., 2006; Heimberg et al., 2008). This expansion of the miRNA repertoire suggests that increased miRNA-mediated gene regulation may contribute to the development of complex, organ-containing animals (Sempere et al., 2006). Also the establishment of tissue identities has been closely coupled with miRNA evolution in bilateria (Christodoulou et al., 2010).

Novel miRNAs continuously evolve in organisms, and once integrated into a gene regulatory network, the new miRNA is only rarely lost (Heimberg et al., 2008). Based on comparative genomics studies, several molecular mechanisms for miRNA genesis and evolution have been suggested (Liu et al., 2008; de Wit et al., 2009). Local gene duplication is the main route for expansion of the miRNA repertoire, and it is typically followed by changes in the duplicate miRNA sequence like mutations in the seed area or seed shifting (Liu et al., 2008; Grimson et al., 2008; Wheeler et al., 2009). Also suggested is a mechanism whereby an up- or downstream genomic region of the original hairpin mutates so that a novel hairpin can be formed, and a novel miRNA can be expressed from the fresh stem of

this hairpin (de Wit et al., 2009). Switching the effective miRNA strand and antisense transcription may also contribute to the evolution of miRNA genes (Liu et al., 2008).

## 2.4 MIRNA PATHWAY IN ANIMALS

The canonical miRNA processing pathway was described already in the very beginning of miRNA focused research. In addition, a variety of alternative pathways have been described over the past few years (reviewed in Winter et al., 2009). This chapter introduces the many different ways a miRNA can be processed from the genome and describes how additional small RNAs, miRNA-offset RNAs, derive from the miRNA genomic loci.

### 2.4.1 The canonical miRNA pathway

Canonical animal miRNAs are generated through a two-step processing pathway (Figure 1). The primary transcript, pri-miRNA, is usually several kilobases (kb) long and contains a local, imperfectly paired stem loop structure (Lee et al., 2002; Bracht et al., 2004). Drosha-DGCR8 complex (Drosha-Pasha in invertebrates) initiates miRNA maturation by precise cleavage of the stem loop embedded in the pri-miRNA (Lee et al., 2003). The ~55-70 nt long miRNA precursor (pre-miRNA) hairpin is then transported to cytoplasm by one of the nuclear transport receptors, exportin-5 (Yi et al., 2003), where it is subsequently processed into ~22-nt RNA duplex by Dicer (Grishok et al., 2001; Hutvagner et al., 2001; Ketting et al., 2001; Knight and Bass, 2001). Following the two subsequent processing steps by RNase-III-type endonucleases, Drosha and Dicer, this small RNA duplex contains a two nucleotide overhang in the 3′ end of both strands. Typically, the accumulation of the duplex strands is asymmetric, and the strand that accumulates to a higher level is defined as the guide strand or the mature miRNA, while its less abundant partner is referred to as the passenger or the star strand, miRNA* (Winter et al., 2009). In general, the guide miRNA sequence is incorporated with Argonaute (AGO) protein into the RISC complex, while the passenger strand is degradated (Khvorova, Reynolds and Jayasena, 2003; Schwarz et al., 2003; Kim and Kim, 2012). miRNA guides the RISC complex to the target mRNA, which results in reduced protein production through a variety of mechanisms involving mRNA degradation, translational repression or polyA tail removal (Huntzinger and Izaurralde, 2011).

When selecting the miRNA strand, AGO proteins use sequence and structural information of the miRNA/miRNA* duplex (Khvorova, Reynolds and Jayasena, 2003; Schwarz et al., 2003; Czech et al., 2009; Hu et al., 2009). However, both strands of miRNA/miRNA* duplex can be simultaneously accumulated, and emerging evidence show that they can both act as active miRNAs (Okamura, Liu and Lai, 2009; Yang et al., 2011). Furthermore, Dicer cleavage of a miRNA hairpin precursor can generate also a third single

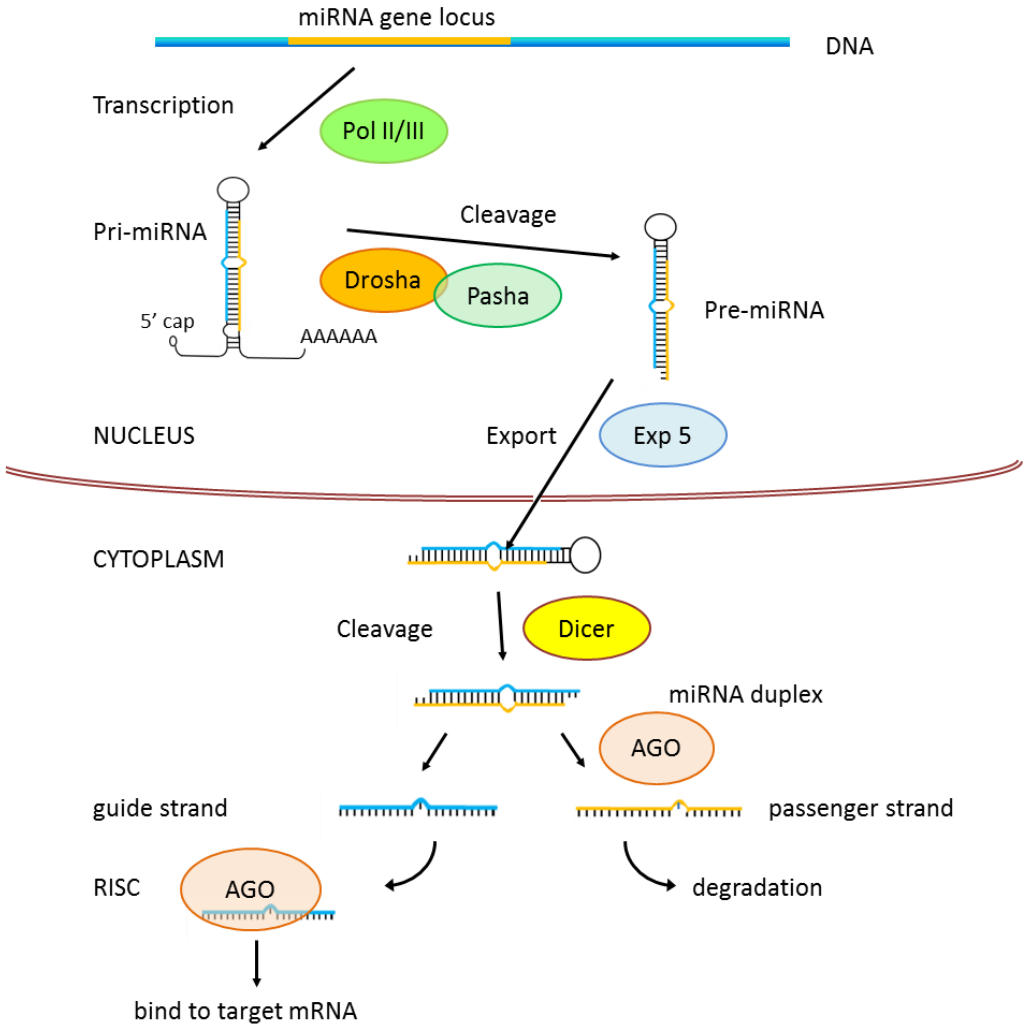


*Figure 1.* A simplified view of the canonical miRNA pathway in animals.

stranded small RNA from the intervening terminal loop, called loop-miR, which may be accumulated as high as the guide strand, incorporates into RISC and functions like mature miRNA (Okamura et al., 2013; Winter et al., 2013).

The observations, that both strands of the miRNA/miRNA* duplex can be functional and that the arm from which the dominant mature miRNA is processed can be species-specific or depend on tissue or developmental stage (Ro et al., 2007; Ruby et al., 2007; de Wit et al., 2009; Chiang et al., 2010), have led to re-nomenclature of miRNAs. When the mature miRNAs derived from the same precursor were earlier named as miRNA and miRNA*, since miRBase release 17 (2011) they are named according to the precursor arm from

whichthey derive. For example, *C. elegans* miRNAs *cel-miR-124-5p* and *cel-miR-124-3p* derive from the 5′ stem and 3′ stem of the precursor *cel-mir-124*, respectively. The previous name of *cel-miR-124-5p* is *cel-miR-124** where the star indicates that it is the minor product of this miRNA gene.

### 2.4.2 Mirtrons

Mirtrons are a recently found class of miRNAs derived from short introns and processed by non-canonical, Drosha-independent pathway (reviewed in Westholm and Lai, 2011). Mirtrons were first found in flies (Okamura et al., 2007; Ruby, Jan and Bartel 2007), and were later characterized in mammals (Berezikov et al., 2007; Babiarz et al., 2008; Ladewig et al., 2012; Sibley et al., 2012) and *C. elegans* (Chung et al., 2011; Jan et al., 2011). Mirtrons are typically spliced from pre-mRNAs, but also derive from non-coding transcripts (Jan et al., 2011). It is worth noting that, in addition to mirtrons, a large fraction of canonical miRNAs are located in introns and should not be confused with mirtrons (Kim, Han and Siomi, 2009).

The pre-miRNA precursor of a conventional mirtron contains the total sequence of its host intron and the hairpin ends correspond precisely to intron splice sites, where typically the "AG" acceptor site adopts a two nucleotide 3′ overhang to the hairpin, thus mimicking a Drosha product (Figure 2, Okamura et al., 2007; Ruby, Jan and Bartel, 2007). The mirtron precursor is shorter than the canonical pri-miRNAs since it comprises only the miRNA/miRNA* duplex and lacks the longer stem that mediates the cleavage by Drosha/DGCR8 complex. Thus, the mirtron pathway is initiated by splicing and intron lariat debranching by lariat debranching enzyme, Ldbr, and then merged with the canonical miRNA pathway to generate active regulatory miRNAs from the pre-miRNA hairpin precursor (Okamura et al., 2007; Ruby, Jan and Bartel, 2007).

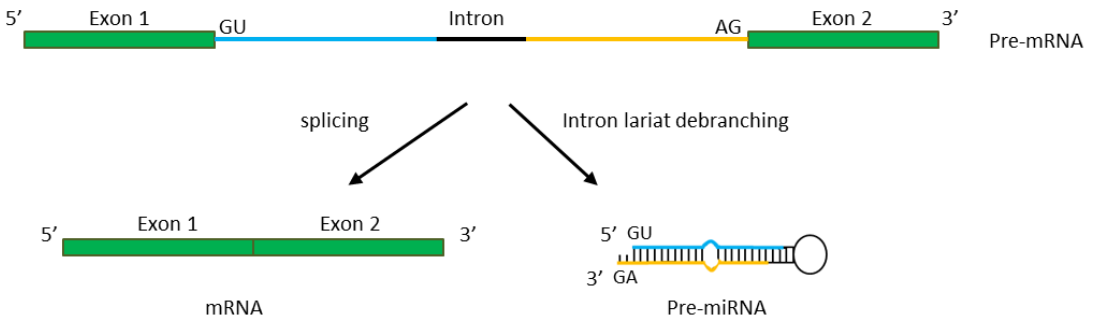


*Figure 2.* The conventional mirtron pathway.

In addition to the conventional mirtron loci, where both ends of the pre-miRNA are excised by the splicing reaction, the miRNA-generating loci can reside at one end of a longer intron. These loci are called 5′ tailed or 3′ tailed mirtrons, because they include an unstructured extension in either 5′ or 3′ end of the hairpin, respectively (Ruby, Jan and Bartel, 2007; Babiarz et al., 2008). Also the tailed mirtrons undergo splicing and debranching, after which the extra tail on the intermediate hairpin is trimmed away. The 3′ extension is trimmed by the RNA exosome, the major eukaryotic 3′/5′ exonuclease complex (Flynt et al., 2010). The trimming machinery of the 5′ extension is not yet known, but one potential candidate to remove the 5′ tails is XRN1/2, the major 5′/3′ exonuclease in eukaryotes (Babiarz et al., 2008).

### 2.4.3 Other non-canonical miRNA pathways

Like the mirtron pathway, most of the other non-canonical miRNA pathways also replace Drosha in the first cleavage step with some other cellular ribonuclease, while the generated pre-miRNA hairpin is processed in the canonical way. This type of strategy is used for example by miRNAs derived from small nucleolar RNAs (snoRNAs) and transfer RNAs (tRNAs) (Babiarz et al., 2008; Ender et al., 2008; Cole et al., 2009; Brameier et al., 2011). One exception is the pathway of human *mir-451*, which is the first known miRNA that is processed without Dicer (Cheloufi et al., 2010; Cifuentes et al., 2010; Yang et al., 2010). The primary precursor or *mir-451* is cleaved by Drosha/DGCR8, but the generated pre-miRNA contains only ~18 base pair (bp) of duplex stem, which is too short for Dicer cleavage. Instead, the *pre-mir-451* is cleaved by AGO2 and then processed to mature miRNA by exonuclease trimming. Recently, a subset of human intron derived miRNAs were found which do not follow the mirtron pathway described above. Instead, the pathway involves Drosha, but does not require its binding partner DGCR8, or Dicer (Havens et al., 2012).

### 2.4.4 miRNA-offset RNAs

moRNAs are a recently discovered class of ~20 nt small RNA molecules generated from the sequence immediately adjacent to the mature miRNA and miRNA* genomic loci (Figure 3). Initially, these molecules were observed among *Drosophila* high-throughput sequencing data (Ruby et al., 2007), and in mouse embryonic stem cells (mESCs) (Babiarz et al., 2008). They were characterized and named as moRNAs in a sequencing study of a simple chordate *Ciona intestinalis* (Shi et al., 2009), and have since been found in several human and mouse small RNA sequencing libraries (Langenberger et al., 2009; Meiri et al., 2010; Bortoluzzi et al., 2012; Zhou et al., 2012). Like miRNAs, moRNAs are also observed at specific developmental stages (Shi et al., 2009). Moreover, many miRNA precursors that express moRNAs are evolutionary old, and the moRNA sequences are also often conserved (Langenberger et al., 2009).

The moRNA processing pathway is not known. One end of each moRNA is probably determined by the Drosha cleavage of pre-miRNA, while the other, more variable end, may result from exonuclease digestion of the pri-miRNA (Ruby et al., 2007). On the other hand, several examples of 5′ and 3′ moRNA duplexes with ~2 nt 3′ overhangs refer to RNAse III processing, thus suggesting that extended hairpin regions on pri-miRNA transcript are cleaved via secondary Drosha processing (Shi et al., 2009). Both theories are consistent with the DGCR8-dependent and Dicer-independent biogenesis of moRNAs inferred from mutant mESC analysis (Babiarz et al., 2008).

moRNAs preferentially arise from the 5′ stem of the hairpin, regardless of the miRNA strand selection bias, and the expression level of moRNAs is not strictly correlated with the expression level of the mature miRNAs (Langenberger et al., 2009; Bortoluzzi et al., 2012). These observations suggest that miRNA and moRNA processing may be linked but is not necessarily interdependent, and thus provide evidence that moRNAs are not just random by-products of the miRNA pathway (Langenberger et al., 2009; Zhou et al., 2012). However, the function of moRNAs remains to be uncovered.

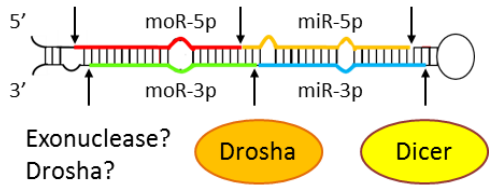


*Figure 3.* An extended miRNA hairpin containing moRNAs in its ends and the suggested moRNA processing machinery.

### 2.4.5 miRNA isoforms

Polymorphism of miRNA 5′ and 3′ ends and shifted sequence variants of same miRNA were observed already in early studies (Lagos-Quintana et al., 2002; reviewed in Ameres and Zamore, 2013). Recently, by using deep sequencing methods, the scale of miRNA heterogeneity has been found to be more prevalent than anticipated and these sequence variants are termed isomiRs (Morin et al., 2008). Imprecision in the Drosha and/or Dicer processing is proposed to be one of the most likely explanations for the miRNA end polymorphism where the isomiR end nucleotides match the genomic sequence (Ruby et al., 2006; Morin et al., 2008; Wu et al., 2009). Nucleotides differing from genomic DNA can also be added to either miRNA or pre-miRNA ends by specific entzymes after Drosha or Dicer cleavage (Landgraf et al., 2007; Morin et al., 2008; Burroughs et al., 2010). It has been shown, for example, that the majority of human *let-7* family members acquire a too short (1 nt) 3′

overhang after Drosha processing and are therefore mono-uridylated by terminal uridylyl transferases (TUTs) to elongate the overhang to two nucleotides, thus making the precursor an optimal substrate for Dicer cleavage (Heo et al., 2012). Also nucleotide substitutions in the sequence, mainly caused by adenosine to inosine RNA editing, identified as A-to-G changes, are frequent (Blow et al., 2006; Landgraf et al., 2007; Morin et al., 2008).

The heterogeneity of miRNAs increases their regulatory potential. A shift in the miRNA 5′-end may redefine its repertoire of targets (Chiang et al., 2010). Different nucleotides in miRNA 5′-end may also have effect on the thermodynamic stability of the miRNA duplex ends and thus change the preferentially accumulated miRNA strand (Hu et al., 2009). Different isomiRs may also be loaded into different AGO proteins (Burroughs et al., 2011). Extra nucleotides added to miRNA 3′-end are typically adenosines or uridines and they affect miRNA stability (Katoh et al., 2009) and targeting efficiency (Burroughs et al., 2010).

## 2.5 MIRNA GENES

The part of DNA from which the pri-miRNA is expressed is perceived as the miRNA gene. Since the total sequence for most of the miRNA genes is not verified, they are localized in the genome based on the alignment of their precursor hairpin. This chapter presents how miRNA genes are located in the genome and the current knowledge of their transcriptional and post-transcriptional regulation.

### 2.5.1 Genomic organization

miRNAs derive from single, stand-alone genes or clusters that contain multiple miRNA precursors encoded in tandem with close proximity (Lagos-Quintana et al., 2001; Lau et al., 2001). There are also a few cases where a single miRNA locus can give rise to two miRNAs with distinct seed sequences through bidirectional transcription (Stark et al., 2008; Tyler et al., 2008). A considerable fraction of miRNA genes are located in intergenic regions while some reside antisense to annotated genes. Most of the other miRNA genes are found within introns of protein-coding genes, or within introns of long non-coding RNA transcripts (Rodriguez et al., 2004). For example, in miRBase release 20 (2013), from 223 annotated *C. elegans* miRNA precursors, 30% are located within introns and 58% are located in intergenic area, while among the 1872 human miRNAs the percentages are 46% and 36%, respectively. A minority of miRNAs derive from exons of non-coding RNA, or from untranslated regions (Rodriguez et al., 2004). In some cases, miRNAs are located in either an exon or an intron depending on alternative splicing of the host transcript (Rodriguez et al., 2004; Kim and Kim, 2007).

### 2.5.2 miRNA clusters

A set of miRNAs that reside closely distributed in the genome is called a miRNA cluster (Lagos-Quintana et al., 2001; Lau et al., 2001). Even though the majority of miRNA genes are isolated, clustered miRNAs compose a significant fraction of all miRNAs. For example, when allowing at most 10 kilobase pairs (kbp) inter-miRNA distance, 38% of *C. elegans* and 25% of human miRNAs are located in clusters. The number of miRNA genes in a cluster varies between 2 to 10 in *C. elegans* and between 2 to 46 in humans (miRBase release 20, 2013). Often the clustered miRNA genes belong to the same miRNA family and thus have significantly similar sequences and often identical seed regions, but there are also clusters of miRNAs which share no sequence homology. Moreover, not all miRNAs in an organism that belong to the same miRNA family are necessarily found in the same cluster. Many miRNA clusters are conserved in closely related species such as human and mouse, or *C. elegans* and *C. briggsae*, and some clusters are shown to have special functions in biological processes (Suh et al., 2004; He et al., 2005; Massirer et al., 2012). For example, *miR-302/367* -cluster located in chromosome IV is highly expressed in hESCs (Suh et al., 2004). Because this cluster is not expressed in later developmental stages, it probably has a role in maintaining the self-renewal capability and pluripotency of embryonic stem cells (Suh et al., 2004; Morin et al., 2008). *miR-302/367* -cluster contains five miRNA precursors: *mir-302b,*

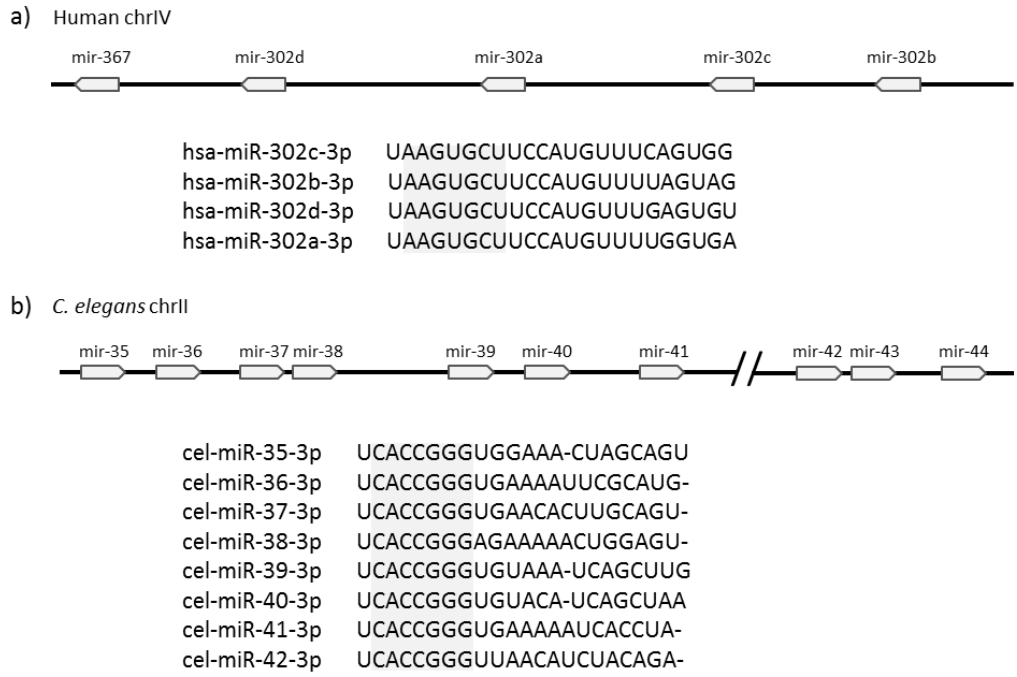


*Figure 4.* Schematic representation of miRNA clusters and miRNA families: a) Human *mir-302/367* –cluster and *mir-302* family, b) *C. elegans mir-35* family located in two clusters. Seed sequences are shaded.

*mir-302c, mir-302a, mir-302d* and *mir-367*, the first four of which belong to *mir-302* family (Figure 4a), but for *mir-367* there are no obvious paralogs in the human genome. Another example is the *mir-35* family in *C. elegans* comprising eight miRNA genes: *mir-35, mir-36, mir-37, mir-38, mir-39, mir-40, mir-41* and *mir-42*, which are located in two genomic clusters, spaced with 350 kbps in chromosome II (Figure 4b). The first cluster includes miRNA genes *mir-35-41*, while the second cluster includes *mir-42,* and two miRNA genes from other families: *mir-43* and *mir-44*. The *mir-35* family members express specifically in germline and deletion of these clusters causes embryonic lethality demonstrating an essential role for this miRNA family in embryonic development of nematodes (Alvarez-Saavedra and Horvitz, 2010). On the other hand, endogenous expression of single miRNA of the family, *mir-35*, is sufficient to sustain embryonic development (Alvarez-Saavedra and Horvitz, 2010), thus illustrating functional redundancy which has been observed in some cases among miRNA family members (Abbott et al., 2005; Miska et al., 2007).

**2.5.3 Transcriptional regulation**

Transcription of eukaryotic miRNAs, like transcription of protein coding genes, is carried out by Pol II. The evidence indicating that Pol II is the polymerase of miRNA transcription includes the discovery of pri-miRNA sequences that are capped and polyadenylated, and the observation that pri-miRNA expression levels are greatly reduced by $\alpha$-amanitin at concentrations that specifically inhibit Pol II (Lee et al., 2004). Further, many miRNA genes are shown to have the same type of promoters as protein coding genes, and thus are very likely to be transcribed by Pol II (Zhou et al., 2007; Ozsolak et al., 2008). However, a small fraction of miRNAs in human genome are shown to be transcribed by RNA Polymerase III (Pol III) (Borchert, Lanier and Davidson, 2006; Ozsolak et al., 2008).

miRNA genes that are not located in protein coding gene area probably derive from their own, independent transcription units (Bartel, 2004). Intron-embedded miRNAs are usually coordinately expressed with their host gene mRNA, implying that they generally are derived from a common transcript (Baskerville and Bartel, 2005). Splicing is not a prerequisite for intronic miRNA production: an unspliced intron can be cleaved by Drosha before splicing, not affecting the level of the mRNA (Kim and Kim, 2007). However, it has been shown that about one third of intronic miRNAs have distinct transcription initiation regions and expression level which is not correlated with the host gene expression level (Ozsolak et al., 2008; Isik, Korswagen and Berezikov, 2010; Monteys et al., 2010). miRNAs organized to clusters are transcribed together as polycistronic transcripts (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee et al., 2002; Ozsolak et al., 2008). However, the expression of clustered miRNAs is not always tightly correlated because of differences in post-transcriptional processing or stability (Sempere et al., 2004). Moreover, clustered miRNAs may transcribe independently of each other (Song and Wang, 2008).

### 2.5.3 miRNA promoter regions

The observation that miRNAs are transcribed by RNA Pol II suggests that miRNA transcription is subject to similar control mechanisms as transcription of protein-coding genes. The transcription start sites (TSS) for most of the miRNAs have not been mapped, but it has been shown in *C. elegans* that sequence fragments between 1 and 2 kbp upstream of the pre-miRNA hairpin in the genome are sufficient to rescue *lin-4*, *let-7* and *lsy-6* mutant phenotypes (Lee, Feinbaum and Ambros, 1993; Johnson, Lin and Slack, 2003). Thus all attempts to analyze miRNA promoters have this far focused on the area immediately upstream of the the pre-miRNA loci (Ohler et al., 2004; Zhou et al., 2007), or upstream of the few experimentially verified pri-miRNAs (Saini et al., 2007; Zhou et al., 2007). In these studies, many *cis*-elements essential for gene transcription are found in *C. elegans* and humans, including CT-repeat microsatellites and sequence motifs resembling initiator element (Inr), as well as CpG islands. On the other hand, TATA-box does not seem to be necessary for most miRNA genes in these species (Ohler et al., 2004; Zhou et al., 2007). On the whole, like protein-coding gene promoters, also miRNA promoters are found to consist of both very specific, non-conserved sequence elements regulating only few miRNAs (Johnson, Lin and Slack, 2003), as well as more common transcription factor binding sites (TFBS) shared with many miRNA promoters (Martinez et al., 2008; Ow et al., 2008). In addition to these experimentally validated *cis*-acting elements, also other motifs shared across promoters of independently expressed miRNAs have been computationally predicted on the genomic scale (Ohler et al., 2004; Zhou et al., 2007). Nevertheless, the regulatory capacity of these motifs has not yet been fully experimentally elucidated.

### 2.5.4 Post-transcriptional regulation of miRNAs

The expression of some miRNAs can also be regulated after transcription, during the Drosha and Dicer processing steps of the precursor. Post-transcriptional control of miRNA expression is reported to occur in tissue-specific (Obernosterer et al., 2006) and in development-specific manner (Thomson et al., 2006; Wulczyn et al., 2007). For example, the *let-7* miRNA is associated with the neuronal differentiation of embryonic stem cells in mammals (Wulczyn et al., 2007). The primary precursor pri-*let-7* is present in both undifferentiated and differentiated cells. On the other hand, the mature *let-7* is not present in undifferentiated ES cells but is induced after differentiation. It has been shown that in undifferentiated cells, the processing of pre-*let-7* is significantly inhibited by high levels of a broadly conserved RNA-binding protein, lin28 (Viswanathan, Daley and Gregory, 2008). Differentiation gradually represses the expression on lin28, and enables the maturation of *let-7*. Lin28 inhibits the maturation of *let-7* by recruiting TUTs to the pre-miRNA causing generation of a long single-stranded tail of Us at the precursor 3'-end which block further processing of the pre-*let-7* by Dicer (Heo et al., 2008). It has been shown that also in *C. elegans* lin-28 binds directly to pre-*let-7* and prevents Dicer processing in epithelial stem

cells (Lehrbach et al., 2010), suggesting that the *let-7*/lin28 regulatory switch might be as conserved as *let-7* itself.

## 2.6 MIRNA TARGET RECOGNITION

The mature miRNA joins specific AGO protein in RISC and guides it to the target mRNA. In animals, partial pairing of the miRNA with its target usually results in reduced protein expression through a variety of mechanisms involving mRNA degradation, translational repression or polyA tail removal (Huntzinger and Izaurralde, 2011). Because of the imperfect binding and the modest impact of an individual miRNA to its target gene expression, detection of miRNA genuine targets is a challenging task (Wightman et al., 1993; Doench and Sharp, 2004; Bartel, 2009). This chapter takes a look at the characteristics of miRNA target binding sites and methods that are used to predict, discover and validate these sites.

### 2.6.1 Characteristics of miRNA target sites

The founding members of the miRNA class, *lin-4* and *let-7*, were shown to act on the 3′ UTRs of the target gene transcripts (Lee, Feinbaum and Ambros, 1993; Wightman et al., 1993; Reinhart et al., 2000). The miRNA–target site duplexes located in the early studies were imperfect, containing mismatches, gaps, and G:U basepairs at various positions (Figure 5). Several of these sites included perfect match to nucleotides 2-8 from the miRNA 5′ end, a section that has since been found to be the core element in miRNA target site recognition (Stark et al., 2003; Lewis et al., 2003). This seven nucleotides long miRNA seed is the most conserved part of miRNAs among metazoans (Lewis et al., 2003; Lim et al., 2003), and many 3′ UTR elements which are shown to mediate posttranscriptional repression in invertebrates are perfectly complementary to miRNA seeds (Lai, 2002). In addition, miRNA-like regulation is most sensitive to disruption of seed pairing (Doench and Sharp, 2004; Brennecke et al., 2005), and pairing of the miRNA 5′ region has been shown to be sufficient to cause repression while the 3′ part of the miRNA is less critical (Doench and Sharp, 2004).

Indeed, in animals, most target mRNAs are regulated through 3′ UTR interactions and the vast majority of miRNAs form only partial duplexes with their targets which include a contiguous Watson-Crick base pairing with the miRNA seed area (Bartel, 2009). However, imperfect pairing of the 5′ seed area of the miRNA to a target site can be compensated by extensive miRNA 3′ end interactions to achieve repression functionality (Reinhart et al., 2000). Recently, 'centered sites' have been described, which lack both perfect seed pairing and 3′-compensatory pairing and instead the middle region, nucleotides 4-15, of the miRNA makes 11–12 contiguous base pairs with the target sequence (Shin et al., 2010).

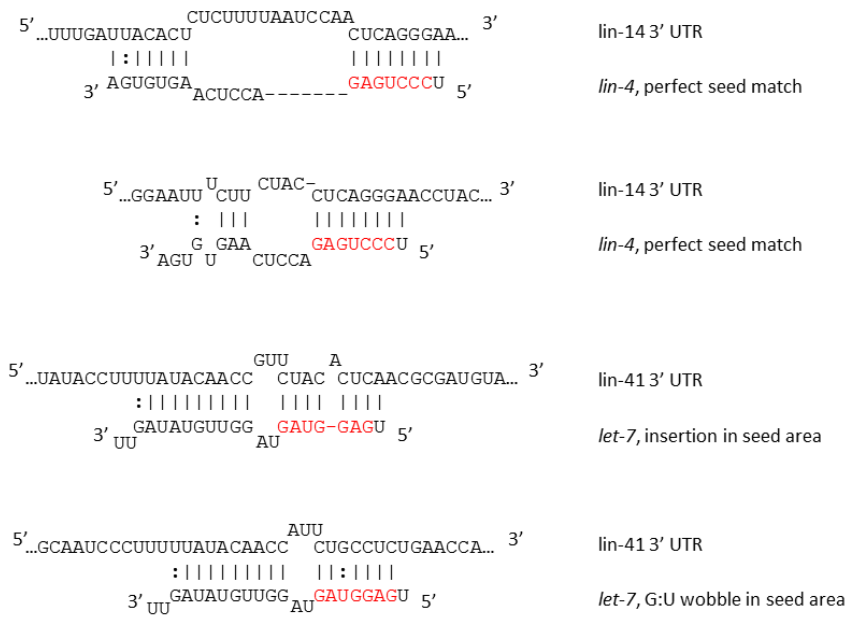


*Figure 5.* Examples of miRNA target sites. Seed sequence in red.

There are also examples of functional miRNA target sites that do not fit any of the patterns described above, and reside beyond the 3′ UTR (Zisoulis et al., 2010).

miRNA families and their target genes are often conserved in related species (Lewis et al., 2003; Brennecke et al., 2005; Krek et al., 2005; Xie et al., 2005; Friedman et al., 2009). For example, *let-7*, which is one of the most broadly conserved animal miRNAs, regulates developmental timing in *C. elegans* by downregulating the *lin-41* gene. This relationship is conserved in humans where the *lin-41* ortholog TRIM71 is similarly targeted by *let-7* (Lin et al., 2007). Another example is the targeting of RAS gene by *let-7* which is conserved from worms to humans (Johnson et al., 2005). Because the members of a miRNA family share the seed sequence, they are often presumed to have the same set of targets. Their 3′-end sequences, however, often diverge which also affects the targeting specificity. miRNA clusters which contain different miRNA families can target multiple different mRNAs, and it has been proposed, that these target mRNAs code for proteins with mutual interactions (Yuan et al., 2009).

It is difficult to establish general rules for miRNA-target interactions. Although conserved pairing to the miRNA seed region on its own can be sufficient for target gene down-regulation (Lewis et al., 2003; Brennecke et al., 2005; Krek et al., 2005; Alvarez-Saavedra and Horvitz, 2010), it has been also shown that perfect seed pairing is not a generally reliable predictor for miRNA-target interaction. For example, in *C. elegans* there are 14 predicted *lsy-6* target genes with perfect seed matched sites in their 3′ UTRs, but only

one of these genes responds to *lsy-6* (Didiano and Hobert, 2006). In addition, the secondary structure of the target mRNA likely contributes to recognition of the embedded miRNA target site (Kertesz et al., 2007). Many mRNAs contain several putative binding sites for the targeting miRNA, thus emphasizing the importance of synergistic binding (Doench and Sharp, 2004). Furthermore, boosting effect of combinatorial regulation by several different miRNAs has been demonstrated (Krek et al., 2005).

### 2.6.2 Biochemical methods for finding miRNA targets

Cross-Linking Immunoprecipitation-high-throughput sequencing (CLIP-Seq) is a recently developed technique used for screening RNA sequences that directly interact with a particular RNA-binding protein (Licatalosi et al., 2008). The idea is to sequence those sites in the mRNA that co-immunoprecipitate with RISC factors, mainly with AGO (Chi et al., 2009; Zisoulis et al., 2010). These studies have provided extensive data supporting seed pairing, conservation and structural accessibility as common features of miRNA target sites.. However, they also reveal new considerations, such as interaction of the RISC complex with coding exons, and many binding sites that do not follow the traditional miRNA target prediction rules (Chi et al., 2009; Zisoulis et al., 2010). A limitation of CLIP-Seq is that it does not guarantee the functionality of the identified binding sites (Thomson, Bracken and Goodall, 2011).

### 2.6.3 Validation of miRNA targets

miRNA targets can be validated using direct validation of specific miRNA:mRNA interactions, or using high-throughput experiments which provide an overview of changes in a large number of gene products. It has been suggested that up to 84% of miRNA mediated repression can be measured as decreased mRNA level (Guo et al., 2010), while some miRNA targeting occurs mostly at the translation level and only affects protein output. Thus, in order to get a complete view of miRNA mediated gene silencing, both mRNA and protein levels need to be studied.

The effect of miRNA expression to a specific gene can be observed at the protein level with western blot and at the mRNA expression level by quantitative real-time PCR (qPCR, Kuhn et al., 2008). Alternatively, reporter assays have been employed to demonstrate a direct link whereby expression of a reporter construct (Luciferase or Green Fluorescent Protein, GFP) carrying the 3′ UTR of the putative target gene will be altered through miRNA transfection. Direct effect of a miRNA can be demonstrated by the loss of regulation in constructs including mutated miRNA target sites (Kiriakidou et al., 2004; Kuhn et al., 2008; Thomson, Bracken and Goodall, 2011).

High-throughput techniques provide information about global effects of exogenous miRNA transfection or silencing of an endogenous miRNA. Degradation of target mRNAs caused by ectopic miRNA expression can be studied on genome-wide scale by microarrays

(Lim et al., 2005; Grimson et al., 2007). Today, next-generation sequencing (NGS) of RNA provides a digital readout of transcript levels and imparts a higher level of accuracy than microarray platforms by enhancing the detection of moderately changed transcripts and assessment of the different gene isoforms expressed (Xu et al., 2010). Global changes in protein levels in response to miRNA transfection or knockdown can be measured with stable isotope labeling by amino acids in cell culture (SILAC), where treated cells are labelled with heavy versions of amino acids (isotopes), making all newly synthesized proteins 'heavy' while the proteins present in the control cells (untreated) remain in the 'light' form. Quantitative mass spectrometry analyzes the ratios of the intensity of heavy versus light peptides (Baek et al., 2008; Selbach et al., 2008).

However, when the effect of miRNA differential expression is measured using high throughput methods, in addition to changes in direct miRNA target expression, a set of indirect changes are also measured. Consequently, it is hard to distinguish the direct miRNA target genes from the indirect ones. Thus, high-throughput methods provide a broad view of miRNA mediated expression, but are not as specific as direct miRNA target validation methods. Either way, validation of miRNA:mRNA interactions by ectopic expression of the miRNA at artificially high levels, may confirm an interaction that does not exist *in vivo* (Doench and Sharp, 2004). Hence, the expression levels of both miRNA and mRNA and the potentially competing binding sites of other miRNAs, should be considered when determining the endogenous regulation of the mRNA by the miRNA (Doench and Sharp, 2004).

### 2.6.4 Computational prediction of miRNA targets

The small number of validated miRNA target gene interactions and the imperfect sequence complementary of animal miRNAs with their targets make accurate computational prediction of targets within whole genome or transcriptome databases a challenging task. In the past decade, many different tools for miRNA target prediction have been developed using empirically derived conclusions about the miRNA recognition sequence as criteria. At first, methods were based mainly on strong Watson-Crick basepairing of the miRNA seed to a site in the 3′ UTR, conservation of that site in the 3′ UTRs of homologous genes in related species, and accessibility of the target site for miRNA binding (Enright et al., 2003; Lewis et al., 2003; John et al., 2004; Grün et al., 2005; Krek et al., 2005). Thereafter, more relaxed seed binding has been permitted, when supported with additional base pairing to 3′ end of miRNA (Friedman et al., 2009), and the boosting effect of multiple miRNA target sites in the same 3′ UTR is taken into account (Saetrom et al., 2007). Features concerning target site sequence context and its location in the 3′ UTR are also added to measure the effectiveness of miRNA binding (Grimson et al., 2007). While evolutionary conservation is an important factor to filter out false positive targets, the non-conserved target sites outnumber the conserved sites 10 to 1 and are also often functional (Farh et al., 2005). For

example, about 30% of verified mammalian miRNA target sites are species specific (Sethupathy et al., 2006) and about 40% of the verified miRNA targets in *C. elegans* are located in 3′ UTRs that align poorly between *C. elegans* and *C. briggsae*. To overcome this issue, tools that can be used on a single genome, even with custom small RNA and mRNA input are developed from sequence specific point of view (Rehmsmeier et al., 2004; Miranda et al., 2006). Examples of commonly used miRNA target prediction tools include TargetScan (Lewis et al., 2005; García et al., 2011; Jan et al., 2011), PicTar (Krek et al., 2005; Chen and Rajewsky, 2006b; Lall et al., 2006), miRanda (Enright et al., 2003; Betel et al., 2010), PITA (Kertesz et al., 2007) and RNA22 (Miranda et al., 2006; Loher and Rigoutsos, 2012).

Supervised machine learning has been recently applied for miRNA target prediction (Kim et al., 2006; Yousef et al., 2007; Wang and El Naqa, 2008). In these algorithms, the classifier should be trained with appropriate example sets of positive and negative miRNA target sites. A number of validated true positive target sites can be extracted from the TarBase database which at present hosts more than 65000 manually curated miRNA–gene interactions (Vergoulis et al., 2012). However, there are no verified sets of false miRNA target sites available, so the supervised algorithms often use a set of randomly generated artificial sequences as negative examples. Such random sets may contain also true target sites by chance or they may differ unrealistically from the positive target site, causing poor performance of the classifier on real test data sets (Hammell, 2010).

As a cost for predicting more true target sites by relaxing the rules for target site detection, the number of predicted target genes has increased from tens to hundreds per miRNA (Sethupathy et al., 2006). The various miRNA target prediction programs apply slightly different targeting rules and thus produce different lists of predicted targets. The degree of overlap between the predictions is poor and the false positive rate is high (Sethupathy et al., 2006; Bartel, 2009). For the best performing computational tools, the fraction of predicted targets that are experimentally detected as down regulated is about 60% (Baek et al., 2008; Selbach et al., 2008). When evaluated with experimentally verified miRNA targets, the general sensitivity of the tools is ~50%, and when only the conserved targets are considered, the sensitivity of the best performing tools increases up to 65% (Sethupathy et al., 2006; Alexiou et al., 2009).

## 2.7 IDENTIFICATION OF NOVEL MIRNAS AND QUANTIFICATION OF EXPRESSION

The first step in a systematic approach to identify the biological roles of miRNAs is to find the miRNA genes and to measure their expression profiles in different tissues and conditions. The main criteria applied in miRNA gene finding is detected expression of a ~22

nt mature miRNA sequence processed from one arm of a stem-loop precursor transcript (Lee and Ambros, 2001). In addition, the base-pairing in the miRNA region of the hairpin has to be extensive, including no large loops or asymmetric bulges (Ambros et al., 2003). This chapter introduces the basic principles used in miRNA gene prediction and methods applied in quantifying their expression.

### 2.7.1 Conventional approaches for miRNA gene finding

Millions of miRNA precursor -like hairpins can be predicted in a genome. For example, there are about 44 thousand hairpins in the genome of *C. elegans* (Pervouchine, Graber and Kasif, 2003) and 11 million hairpins in human genome (Bentwich et al., 2005). Not all the hairpins are miRNA precursors, and it is difficult to describe precisely how a proper pre-miRNA differs from the other genomic hairpin structures. Hence, all approaches to computationally predict miRNAs (reviewed in Gomes et al., 2013) apply learning from the structure of known miRNA hairpins. However, no verified information is available about hairpins that are not miRNA precursors which impedes the efforts remarkably.

The most common method to reduce the number of putative miRNA hairpins is to require evolutionary conservation. Prominent examples of algorithms applying the homology filter are MiRscan (Lim et al., 2003) and miRseeker (Lai et al., 2003). Other filters used are rejection of repeats and coding sequence mapping hairpins (Grad et al., 2003; Lai et al., 2003; Lim et al., 2003). After preliminary filtering, the good hairpin candidates are further classified to miRNA-like and non-miRNA hairpins using features concerning the sequence and structural properties of different regions of known miRNA stem–loops. Examples of rules used in the hairpin scoring are base-pairing in the mature miRNA section, base pairing in the other parts of the fold-back, minimum free energy, GC content, number of matches, mismatches, gaps and loops (Grad et al., 2003; Lai et al., 2003; Lim et al., 2003; Yousef et al., 2006; Wu et al., 2011). There are also algorithms presented for miRNA gene finding task that deviate from the conventional procedure. One example is target centered approach based on the idea that a functional miRNA must have targets. In this approach, authors identify highly conserved short motifs from mammalian 3′ UTR sequences and find several hundred conserved miRNAs that could bind these motifs, including more than hundred novel miRNAs (Xie et al., 2005). Another interesting approach is to search for miRNA sequence homologs inside a genome. This method is used to predict clustered miRNAs in human and rodents by scanning the regions close to previously identified miRNA precursors (Sewer et al., 2005).

Requiring sequence conservation of the miRNA gene helps to reduce the number of false positive predictions. However, miRNA evolution seems to proceed rapidly (Liang and Li, 2009) and the conservation based methods rarely are able to detect novel miRNAs, but instead verify candidates similar with the known ones (Bentwich et al., 2005). Also the strategies that reject parts of the genome based on an already annotated function have their

drawbacks: it has been shown that a sequence segment can have more than only one function. For example, eleven different mammalian miRNA precursors are shown to contain repeat sequences (Smalheiser and Torvik, 2005) and many miRNAs are derived from snoRNAs or tRNAs (Babiarz et al., 2008; Ender et al., 2008; Cole et al., 2009; Brameier et al., 2011; Ono et al., 2011).

### 2.7.2 Finding novel miRNAs from NGS data

Current NGS approaches like Illumina and SOLiD (Sequencing by Oligonucleotide Ligation and Detection), offer several advantages for exploration of small RNAs including high resolution, increased sequencing depth, and less complex experimental procedures. These technologies have made it possible to discover novel species-specific and low-abundant miRNAs, as well as the minor miRNA products (Bar et al., 2008; Morin et al., 2008; reviewed in Pritchard, Cheng and Tewari, 2012), as well as new classes of small RNAs, like transcription initiation RNAs, tiRNAs (Taft et al., 2009), and small RNAs that derive as fractions of other small non-coding RNAs like snoRNA-derived sdRNAs and tRNA-derived tRF RNAs (Martens-Uzunova et al., 2013).

At present, Illumina is the most common platform used for NGS (Illumina, 2014). The preparation of Illumina small RNA sequencing library is made by ligating specific RNA adapter sequences to the 5′ and 3′ ends of single RNA molecules in the size fractionated small RNA population. The sequences are further reverse transcribed, amplified by PCR and size selected. The library is deposited on a flow cell coated with single-stranded oligonucleotides that correspond to the ligated adapter sequences. The sequence fragments are bound to the surface of the flow cell and extended to make copies. The free end of a ligated fragment 'bridges' to a complementary oligo on the surface. Repeated bridge amplification makes up to 1000 copies of each sequence fragment, resulting in millions of unique clusters across the flow cell surface. The reverse strands are cleaved and washed away and clusters are sequenced base by base in parallel. During each sequencing cycle, a single fluorescently labeled nucleotide is added to the chain. The label serves as a terminator for polymerization, so each base incorporation is followed by high resolution imaging of the entire flow cell. The images show the physical location of a cluster and the identity of the base incorporated. After imaging, the fluorescent label is cleaved away to allow the incorporation of next nucleotide. This cycle is repeated base by base, generating a series of images each representing a single base extension at a specific cluster. The raw data of a sequencing experiment comprise a series of image files: for Illumina there is one image per cycle of nucleotide addition. These images are first interpreted to identify distinct clusters and the signal intensities of each cluster in each sequencing cycle. The signal intensities for each cluster are then used to call the bases and the quality for each base call, resulting to raw data reads along with their per base quality scores (Illumina, 2014).

Compared to earlier efforts to predict miRNA genes by using all the hairpins found in the genome, NGS technology gives new layers of information to begin with: the search for novel miRNAs can be limited to expressed sequences, and Dicer cleavage positions can be assessed by existence of the minor miRNA sequence. Starting with the genomic alignment of the sequenced reads that are not mapped to known miRNAs or other annotated RNA types, reads in the same locus are clustered together, and the corresponding genomic sequence, extended with some flanking regions, is extracted in order to include the full-length pre-miRNA sequence. The minimum free energy structures of the extracted sequences are determined and those with non-hairpin structures are rejected. Several structural features of the predicted hairpins are further inspected in order to distinguish high-confidence miRNA precursors from other sequences: the number of reads supporting the presence of the main mature miRNA sequence (10–20 are commonly used cutoffs), consistent processing of the 5′-end of the mature miRNA, reads supporting the presence of miRNA sequences from both arms of the predicted hairpin (miRNA and miRNA*) and strong base-pairing between them along with the correct 3′-overhang (Kozomara and Griffiths-Jones, 2011; Pritchard, Cheng and Tewari, 2012).

Since the application of deep sequencing to miRNA detection, various software tools have been developed to support the data related analysis (reviewed in Li et al., 2012). There are clear differences in performance of these methods: while some methods are highly accurate with a rather small set of predictions, there are also tools that may give thousands of predictions. Hence, it is difficult to assess the miRNA sequencing analysis tools based on their capacity to predict novel miRNAs, and the tools performance usually depends also on the species studied (Li et al., 2012). Examples of available miRNA sequencing analysis tools are miRDeep2 (Friedländer et al., 2012), miRanalyzer (Hackenberg, Rodriguez-Ezpeleta and Aransay, 2011) and SeqBuster (Pantano, Estivill and Marti, 2010). All these tools include the complete pipeline for miRNA analysis task: read preprocessing, counting of the known miRNA mapping reads and search for novel miRNAs. miRDeep2 and miRanalyzer pipelines include exclusion of sequences that align to other annotated RNAs in RNA family database, Rfam (Burge et al., 2012), while SeqBuster can apply any custom database when annotating the sequences. While miRDeep2 and SeqBuster can be applied to any genome, also those without any annotated miRNAs, miRanalyzer is restricted to preprocessed genomes for selected organisms. At present, SeqBuster is the only tool available which offers a deep characterization of isomiRs, for example the percentage of a specific miRNA with a specific type of modification, and a list of miRNAs that do not show any sequence variability (Pantano, Estivill and Marti, 2010).

### 2.7.3 miRNA expression profiling

The expression level of a single miRNA can be measured in the laboratory with Northern blotting or qPCR, which enable rapid detection of both miRNA and precursor (Lau et al.,

2001; Schmittgen et al., 2008). Global expression profiles for multiple miRNAs in a sample can be identified at one time point using miRNA-specific oligonucleotide microarrays (reviewed in Yin, Zhao and Morris, 2008) or deep sequencing. The basis for profiling the expression of known miRNAs from deep sequencing data is the alignment of the reads to the miRBase reference sequences. In order to detect the isomiRs, the reads must be aligned to pre-miRNA sequences, allowing a mismatch at least in the 3′ end of the reads. The mature miRNA read count is the sum of all reads that can be aligned to its location with a moderate overhang in both ends. Reads that map to multiple miRNAs of same miRNA family, are added to the read count of all the miRNAs in question. Once the read count for each miRNA in a library is calculated, it can be subsequently compared with its amount in any other small RNA sequencing sample to get the differentially expressed miRNAs (Pantano, Estivill and Marti, 2010; Hackenberg, Rodriguez-Ezpeleta and Aransay, 2011; Friedländer et al., 2012).

## 2.8 OTHER CLASSES OF SMALL NCRNAS

As a result of the development of NGS technologies and computational methods, the number of known small ncRNAs has expanded during the recent years. Several aspects of origin, structure, associated effector protein and biological function, have led to general recognition of three main classes of small ncRNAs: miRNAs, short interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs). Despite different biogenesis mechanisms they all serve as sequence specific guides for Ago family proteins in small RNA guided gene silencing. This chapter gives an overview of the main characteristics of animal siRNAs and piRNAs.

### 2.8.1 Short interfering RNAs

siRNAs are ~21 nt RNAs that are processed from long, fully complementary dsRNA through cleavage by Dicer (Fire et al., 1998; Tomari and Zamore, 2005). They can derive either from exogenous dsRNA (exo-siRNAs), or endogenously from RNA transcribed in the nucleus (endo-siRNAs). The first examples of RNA-interference (RNAi) were triggered by long exogenous dsRNA, which was cleaved in the cell into double stranded siRNA precursors by Dicer (Fire et al., 1998). Natural sources of exo-siRNAs are viruses containing dsRNA, while sources for endo-siRNAs are jumping genes, also known transposons, natural antisense transcripts and pseudogenes (Golden, Gerbasi and Sontheimer, 2008). After Dicer processing, siRNA duplex is separated and one strand is incorporated into the RISC complex. The guide siRNA strand directs RISC to perfectly complementary sites in the target mRNA, leading to cleavage of the target (Kim, Han and Siomi, 2009). In consequence of the requirement of perfect complementarity, almost all siRNAs silence the

same sequence from which they were derived. Thus, besides protecting a cell from intrusion of viruses, siRNAs are involved in maintenance of genome integrity by silencing transcription from undesired loci.

In *C. elegans*, the initial silencing trigger is amplified by RNA-dependent RNA polymerases (RdRPs). The interaction between primary siRNAs and the target mRNA is required for the recruitment of RdRPs which then catalyze the biogenesis of abundant secondary siRNAs by using the target mRNA as a template. Secondary siRNAs are 21 to 22 nts in length and are derived upstream of the trigger sequence (Pak and Fire, 2007; Sijen et al., 2007). The endo-siRNA classes defined in *C. elegans* are 22G-RNAs and 26G-RNAs, 22 and 26 nt long RNAs with a 5′ preference for guanosine (Ruby et al., 2006; Gu et al., 2009). The 26G-RNAs are primary siRNAs generated in the germline, while large population of 22G-RNAs are secondary siRNAs triggered by 26G-RNAs and 21U-RNAs (Vasale et al., 2010). However, 22G-RNAs can also be processed independently. Like 26G-RNAs, 22G-RNAs are also abundantly expressed in germline, and the majority of them target unique genome sequences, including half of the coding genes in *C. elegans* (Gu et al., 2009).

### 2.8.2 Piwi-interacting RNAs

piRNAs are distinct from siRNAs and miRNAs in their larger size (25–30 nt) and their Dicer-independent biogenesis (reviewed in Siomi et al., 2011). While miRNAs and siRNAs are broadly expressed in most cell and tissue types, the expression and function of piRNAs seem to be restricted to germ cells and to gonadal somatic cells (Aravin, Hannon and Brennecke, 2007). piRNAs are encoded by a small number of genomic clusters and all piRNAs within a cluster reside on the same strand. The genomic locations of the piRNA clusters and their promoter sequences are conserved in mammals, but the sequences of the single piRNAs are not. The organization of piRNAs as clusters suggests that there is a single stranded primary piRNA transcript encompassing the cluster. However, these hypothetical transcripts contain no significant secondary structures, and it is currently unknown how the primary piRNAs are processed from the piRNA clusters (Siomi et al., 2011).

The main function of piRNAs is to silence transposable elements (TEs) in the germ line. TEs are genomic parasites that can move to new places by insertion or transposition and hence disrupt genes and alter the genome (Kazazian, 2004). piRNAs associate with Piwi protein, which is the germ-line specific member of Ago protein family and unique to animals (Grimson et al., 2008). Piwi interacts with piRNAs forming piRNA-induced silencing complex (piRISC) that recognizes and silences complementary RNA targets (Siomi et al., 2011).

21U-RNAs are a class of diverse, autonomously expressed small RNAs described in *C. elegans*. They are 21 nt long and start with uracil but diverse in their remaining 20 nt (Ruby et al., 2006, Batista et al., 2008). They originate mostly from two broad and distinct regions

of chromosome IV. 21U-RNAs share two upstream sequence motifs which could either be promoter or processing signals. These motifs are conserved in *C. briggsae*, but the 21U-RNA sequences are not (Ruby et al., 2006). 21U-RNAs are expressed specifically in the germline, associate with worm Piwi Argonautes PRG-1 and PRG-2, which are required to maintain germline genomic integrity and are therefore suggested to be the piRNA counterparts in *C. elegans*. However, 21U-RNAs are shorter than mammalian or fly piRNAs and do not have significant roles in transposon silencing (Batista et al., 2008) and hence their exact function remains unknown.

# *3 Aims of the study*

The aim of this work was to develop and apply computational methods in order to deepen the understanding of miRNA biology. The specific aims were:

- To find novel insights related to the regulation of miRNA genes expression by examining the miRNA upstream sequences of *C. elegans* and *C. briggsae* for conserved sequence motifs.

- To enhance computational prediction of miRNA target genes with unsupervised machine learning.

- To increase the repertoire of available miRNAs and to provide an evolutionary perspective of nematode miRNAs by sequencing and annotating the miRNAome of *P. redivivus*.

- To gain additional knowledge concerning the small RNAome of hESCs and to further understand the role of moRNAs and their putative developmental stage-related expression by sequencing and profiling the miRNAs and moRNAs.

# 4 Materials and methods

This chapter describes the main bioinformatic workflows applied in the original publications. The software development tools and the main public data sources used throughout the studies are shortly listed at the end of the chapter.

## 4.1 MOTIF FINDING (I)

To find putative miRNA regulation motifs, we searched for short, conserved and over-represented patterns from the upstream sequences of all *C. elegans* and *C. briggsae* miRNAs as described in Publication I. Briefly, the genomic sequences of *C. elegans* and *C. briggsae* were downloaded from WormBase release WS160 and the sequences 1000 bp upstream from the beginning of miRNAs were extracted according to miRNA genomic coordinates in miRBase release 10.1. For the miRNA clusters, the 1000 bp upstream sequence was extracted only for the first miRNA in each cluster. The total of analyzed *C. elegans* miRNA upstream sequences was 100 of which 25 were located in protein coding gene regions. For *C. briggsae*, there were 95 miRNAs in all, of which 70 had an ortholog in *C. elegans*. The motif finding tools employed in the study were: POCO (Kankainen and Holm, 2005), Weeder (Pavesi, Mauri and Pesole, 2001), MEME (Bailey et al., 2006) and Gibbs Recursive Sampler (Thompson, Rouchka and Lawrence, 2003). POCO was the only tool that could find significant motifs occurring in every miRNA upstream sequence: these 65 alike patterns were aligned with ClustalW to get their consensus sequence (Chenna et al., 2003). Motif occurrences were counted with Visualize (Kankainen et al., 2006) or Weeder Motif Locator (Pavesi et al., 2006) and the position weight matrix for each motif was calculated from these results. The motif logos were created with WebLogo (Crooks et al., 2004) and matching TFBSs were searched from Transfac (Matys et al., 2006) and Jaspar (Vlieghe et al., 2006) databases with STAMP tool (Mahony, Auron and Benos, 2007).

## 4.2 SELF-ORGANIZING MAP (II)

The self-organizing map (SOM, Kohonen, 1995) is a neural network algorithm widely used to categorize large, high-dimensional datasets by mapping the data into a smaller dimension, typically into a two dimensional lattice of interconnected neurons. Each neuron of the trained SOM contains a reference model, which represents a local domain in the input space. As described in Publication II, we applied the SOM algorithm to predict *C. elegans* miRNA targets. Briefly, we trained a SOM with all 22 nt substrings extracted from

the 8980 experimentally verified *C. elegans* 3′ UTRs in WormBase release WS195. The structure of the SOM was set to a 32×32 lattice of interconnected neurons where each neuron contained a 4×22 model matrix initialized with a random sequence. The last eight nucleotides (15–22) in the neuron model matrix, representing the seed area, were weighted by multiplying the input sequence distance from the model matrix in nucleotides 15 and 22 with a factor of two and in nucleotides 16–21 with a factor of five. After unsupervised, competitive learning process, the weight matrix in each neuron of the SOM represented the position-specific frequency matrix of a mutually similar set of short 3′ UTR subsequences. The miRNA clusters in the trained SOM were defined as those neurons that contain at least one 7-mer seed matching sequence for one or more *C. elegans* miRNAs in miRBase release 15. The initial miRNA target prediction was performed by collecting for each miRNA those 22 nt substrings that map to its cluster neuron extracted from a more recent set of 12866 verified *C. elegans* 3′ UTR sequences in WormBase- release WS213. The initial *C. briggsae* miRNA target sites were predicted in a similar manner from the 1000 nt downstream sequences for all (total of 11851) genes that had a predicted ortholog in *C. elegans*. For each miRNA–target site pair in the initial prediction set the minimum free energy was calculated using stand-alone versions of RNAfold and RNAcofold (Hofacker, 2003).

## 4.3 GENERATION AND PREPROCESSING OF *P. REDIVIVUS* SMALL RNA LIBRARY (III)

The small RNA library of *P. redivivus* miRNAs was generated as described in Publication III. Briefly, small RNAs were isolated from mixed cultures of *P. redivivus* strain PS2298/MT8872 (Sternberg and Horvitz, 1981) and prepared for sequencing according to Illumina protocol. The library was size selected, and sequenced on Illumina Genome Analyzer IIx. Data was preprocessed by trimming the 3′ adapters and polyA tails. Further cleaning was made by discarding adapter dimers, as well as reads matching to *E. coli* genome. *P. redivivus* tRNAs were predicted using Aragorn (Laslett and Canback, 2004), and reads exactly mapping to these sequences were removed from the data set.

## 4.4 PREDICTION OF MIRNAS FROM NGS DATA (III)

The cleaned *P. redivivus* small RNA sequence reads shorter than 28 nt were mapped against its genome with Bowtie (Langmead et al., 2009), and the potential miRNA precursors were detected by clustering the reads in overlapping genomic loci (Publication III). To classify a sequence as a miRNA precursor hairpin, we used the following criteria: the minimum energy secondary structure of the sequence calculated with stand-alone version of RNAfold (Hofacker, 2003) is a hairpin, the most abundant read mapped to the sequence area is

located in the other arm of the hairpin and has at least ten occurrences, and there is strong base pairing between the mature miRNA and the opposite arm of the hairpin. The set of predictions was further supplemented with miRNAs found using miRDeep2 (Friedlander et al., 2012).

## 4.5 MIRNA ORTHOLOGY ANALYSIS (III)

The predicted miRNAs were searched for orthologs among the known miRNAs with the same seed in six nematode worms, fly and humans by using the sequence similarity of both mature miRNA and its precursor as criteria (Publication III). Briefly, all the 7-mer seed match miRNA pairs and the corresponding hairpin sequence pairs were aligned using stand-alone version of EMBOSS Needle (Rice, Longden and Bleasby, 2000) and the similarity of two sequences was measured by the ratio of the alignment score over the alignment length. To get the cut-off ratio for high similarity, all miRNA pairs that did not share the seed sequences were used as background, and the median value for their alignment score ratios was calculated. Alignments of the matching seed miRNAs 2-fold or more above background were considered to share high sequence similarity and thus to be orthologs. Hierarchical clustering of the orthology map was conducted with the R language and environment for statistical computing (Bates et al., 2013).

## 4.6 SEQUENCING OF HESC SMALL RNAS (IV)

Small RNA sequence libraries were generated for two in-house-derived hESC lines HS401 and HS181, and human primary foreskin fibroblast line HFF-1 (Hovatta et al., 2003; Unger et al., 2008; Hovatta et al., 2010) in order to analyze their miRNA and moRNA expression profiles as described in Publication IV. In brief, the small RNA libraries were prepared for sequencing according to Illumina protocol, size selected and sequenced on Illumina Genome Analyzer IIx. Data was preprocessed by trimming the 3′ adapters and polyA tails. Homopolymers and reads including adaptor dimers were discarded and reads that were shorter than 14 nt were removed. The trimmed reads were aligned to known human miRNAs in miRBase release 17 using Bowtie (Langmead et al., 2009). Reads that aligned exactly to other annotated human small RNA species: tRNAs, piRNAs, small nuclear RNAs (snRNAs), snoRNAs, ribosomal RNAs (rRNAs) or mitochondrial tRNAs (Mt_tRNAs), were discarded, as well as reads aligning exactly to human repeats.

## 4.7 PROFILING MIRNAS AND MORNAS FROM NGS DATA (IV)

The miRNA and moRNA expression profiles in hESCs were determined from the small RNA sequencing data as described in Publication IV. To assess the miRNA expression levels from the set of sequence reads that aligned to known human miRNA hairpins, we counted those reads that were located in the area of the mature miRNA loci, extended with two nucleotides both upstream and downstream. moRNAs were detected from the vicinity of miRNA precursors, and the number of moRNA reads were counted like miRNA reads. The differential expression analysis between hESCs and HFF-1 was performed using R Bioconductor package DESeq (Anders and Huber, 2010).

## 4.8 SOFTWARE DEVELOPMENT TOOLS

**Java** programming language was used for implementation of all the custom tools needed to retrieve and combine data uploaded from databases, and tools to perform the different data managing steps throughout the studies (Publications I-IV). Especially, java was used to implement the custom sequence clustering SOM in Publication II.

**PHP** programming language was used in implementing the web interface of the mirSOM tool.

## 4.9 DATA SOURCES

The main public data sources used for data retrieval and submission in the original publications are briefly described in Table 2.

*Table 2.* The main public data sources used.

| Database | Description | Reference |
|---|---|---|
| miRBase | MicroRNA database | Kozomara and Griffiths-Jones, 2011 |
| piRNABank | PIWI-interacting RNA (piRNA) repository | Sai Lakshmi and Agrawal, 2008 |
| RFam | RNA family database | Gardner et al., 2009 |
| Ensembl | Eukaryotic genome database | Flicek et al. 2013 |
| GenBank | An annotated collection of all publicly available DNA sequences | Benson et al., 2013 |
| RefSeq | NCBI Reference Sequence Database: A comprehensive and non-redundant set of genomic, transcript, and protein reference sequences | Pruitt et al., 2009 |
| UCSC | Genome database | Meyer et al., 2013 |
| WormBase | Worm genome database | Yook et al., 2012 |
| NCBI BioProject | A collection of genomics, functional genomics and genetics studies and links to their resulting datasets | Barrett et al., 2012 |
| NCBI SRA | The Sequence Read Archive, stores sequencing data from the next generation sequencing platforms | Wheeler et al., 2008 |
| Jaspar | Transcription factor binding site database | Vlieghe et al., 2006 |
| Transfac | Database on eukaryotic cis-acting regulatory DNA elements and trans-acting factors | Matys et al., 2006 |
| Repbase | Repetitive DNA element database | Jurka et al., 2005 |

# 5 Results

This chapter describes the key findings of the original publications in this thesis.

## 5.1 SHARED MOTIF UPSTREAM OF *C. ELEGANS* AND *C. BRIGGSAE* MIRNAS

In order to elucidate the factors related to the transcriptional regulation of miRNA genes, we searched, in Publication I, for conserved sequence motifs from the area 1000 bp upstream of all miRNAs independent of their genomic location. In addition to several motifs found also in earlier studies (Ohler et al., 2004; Zhou et al., 2007), we discovered one, previously unknown motif GANNNNGA with a conserved distribution in the upstream sequences of all miRNAs of *C. elegans* and *C. briggsae*. This motif appears to be a common factor for the upstream regions of all miRNAs in these two species and it is concentrated in the area near the beginning of the miRNA precursor start sites. The motif GANNNNGA is especially abundant upstream of two ancient miRNAs *mir-1* (*C. elegans* 23, *C. briggsae* 26 occurrences) and *mir-124* (*C. elegans* 26, *C. briggsae* 30 occurrences, expectation ~8 occurrences in 1 kbp). Of these, *mir-1* is intergenic, while *mir-124* is located in an intron of a protein coding gene. Furthermore, the 1000 bp upstream sequences of human and mouse *mir-1-1* and *mir-124-a1* contain nearly as many occurrences (*mir-1-1*: human 21, mouse 20; *mir-124-a1:* human 37 and mouse 19 occurrences) of the motif GANNNNGA as the corresponding sequences of *mir-1* and *mir-124* of *C. elegans* and *C. briggsae*.

## 5.2 SELF-ORGANIZING MAP PREDICTS MIRNA TARGETS IN *C. ELEGANS*

The SOM trained with all 22 nt substrings of *C. elegans* experimentally verified 3' UTRs resulted in a well-organized map of miRNA target site clusters (Publication II). Each 22 nt training sequence is assigned to one, most similar neuron of the SOM. According to the definition, the target site cluster for a miRNA contains all neurons which include at least one training sequence that matches its seed. 211 of 233 *C. elegans* miRNAs contain only one neuron in their cluster, while 21 miRNA clusters comprise two neurons and a cluster for one miRNA contains three neurons. When there is more than one neuron in a miRNA cluster, these neurons mainly are adjacent to each other. The total of miRNA cluster neurons is 160, of which 117 are linked to only one miRNA.

The initial set of putative target sites for a miRNA is found from its cluster neurons. Of the 20 experimentially verified true *C. elegans* miRNA target genes, 18 were present in the

cluster of the targeting miRNA, while the other two have a couple of sites clustered into neurons adjacent to their right miRNA cluster. Also the majority of the well characterized false target genes (9 of 13) were in the initial set of the miRNA target predictions, but the 'true' sites could be roughly separated from the 'false' sites by using an optimized free energy threshold. This threshold was further applied to all predictions in order to increase the specificity of the method. The sensitivity and specificity of mirSOM, calculated using the sets of known *C. elegans* miRNA true and false target genes, are 0.90 and 0.92, and it outperforms 7 other miRNA target prediction tools in reporting the verified *C. elegans* true targets and 6 tools in rejecting the known false targets.

## 5.3 *P. REDIVIVUS* MIRNAOME

In Publication III we reported the *de novo* sequencing of the small RNAs, along with the genome and transcriptome of *P. redivivus*. We sequenced 24 million reads from the small RNA library generated from mixed stage worms. From this data, we identified 248 miRNA precursors, 90% of which were supported by mature miRNAs from both of its arms. When grouped according to their 7-mer seeds, *P. redivivus* miRNAs fell in 171 families, of which up to 62% are phylogenetically conserved. When the conservation of the main miRNA product and the hairpin sequence were also required, at least one ortholog for 25% of the *P. redivivus* miRNAs were found. For 63% of the miRNA genes the main mature miRNA is located in the 3' arm of the hairpin, which is a phenomenon found to be typical for nematodes and flies (de Wit et al., 2009). Three miRNA loci are transcribed from both strands, giving rise to two independent miRNAs. There are also a few miRNAs that have multiple possible origins in the genome. The twenty most abundantly expressed miRNAs comprise almost 80% of the small RNA library. Ten of these miRNAs, *lin-4* and *miR-1* among them, have an ortholog in *C. elegans*. Seven abundant miRNAs, however, including the most highly expressed miRNA *prd-mir-7865* (34% of the reads), are specific for *P. redivivus*.

In comparison with *C. elegans*, the *P. redivivus* miRNAome is about the same size (248 vs. 223), while the number of different 7-mer seed families found in *P. redivivus* is smaller than in *C. elegans* (171 vs. 211). Nonetheless, half of the *C. elegans* miRNAs have the seed sequence conserved in *P. redivivus*, and 20% of the miRNAs are conserved up to the hairpin level. Also the number of miRNA clusters found in these two worms is close to each other (14 vs. 19). Interestingly, a cluster of seven miRNAs in *C. elegans* (*mir-35, mir-36, mir-37, mir-38, mir-39, mir-40, mir-41*) has a counterpart in *P. redivivus* where three miRNA hairpins, and three 7-mer seeds are conserved. On the other hand, another large miRNA cluster in *P. redivivus* (*mir-7941, mir-7889, mir-7944, mir-7966a, mir-7966b, mir-7905, mir-7912*) contains

seven miRNAs that are not conserved on the hairpin level, and only one seed of these miRNAs is conserved in one of the six other worms studied.

## 5.4 MIRNAS AND MORNAS IN HESCS

In Publication IV we sequenced the small RNAs of two in house-derived hESC lines: HS401 and HS181, and a human fibroblast line HFF-1, and characterized the miRNA and moRNA profiles. The set of miRNAs significantly over-expressed in the ES cell libraries included most of the miRNAs from the well-known pluripotency related *miR-302-367* and *miR-371/372/373* clusters and was thus in accordance with the earlier findings (Suh et al., 2004; Morin et al., 2008; Bar et al., 2008). We also observed additional evidence for ES cell related expression of the paralogous *miR-106a-363* and *miR-17-92* clusters, and the large C19MC cluster (Bar et al., 2008; Wilson et al., 2009).

Moreover, we found 350 candidate moRNAs from loci of 273 miRNA hairpins. Over 90% of the moRNAs were expressed only in ES cells, while less than 20% of them were detected in the skin cell library where they were associated with very low read counts. The median length of moRNAs was 20 nt, and similarly with miRNAs, moRNAs were also detected with variable length reads, isomoRs. The isomoR reads derived from the 5′ arm of the same miRNA hairpin were similar in their end, and isomoRs derived from the 3′ arm were similar in their start. Generally, the 5′ moRNAs ended in the last nucleotide before the start of the most abundant 5′ isomiRs, and the 3′ moRNAs started from the first nucleotide after the ending of the most abundant 3′ isomiRs. Almost all moRNAs (96%) in our fibroblast library were derived from miRNA 5′-end, whereas in the two hESC libraries over 40% derived from miRNA 3′-end. Although almost half of the detected moRNAs were present with only one read, there were 26 moRNAs whose most abundant isomoR was expressed with more than 50 reads. About 65% of all moRNAs were conserved in mammals, and derived from a conserved miRNA hairpin locus.

Typically moRNAs were detected with fewer reads than the miRNAs from the same stem of a hairpin, and there was no clear correlation between the expression levels of these two. Of the ten moRNAs most significantly over-expressed in ES cells, six derived from miRNAs which were also up-regulated in the hESC libraries. Interestingly, three of the four other moRNAs (*moR-421-5p, moR-517a-5p, moR-517c-5p*) in this set were expressed with more reads than the related mature miRNA in hESC libraries whereas only one moRNA (*moR-103a-2*) derived from a very abundant miRNA locus.

# 6 Discussion

## 6.1 THE ROLE OF MOTIF GANNNNGA

The expression pattern of many miRNAs is specific to developmental or metabolic stage and is also tissue-specific. Furthermore, the miRNA expression is altered during disease. The level of expression can be controlled at any step of the miRNA biogenesis, from transcription of pri-miRNA to turn-over of mature miRNAs. However, the regulation of miRNAs expression is poorly understood, because most of the miRNA TSSs and thus promoters are still uncharacterized. One approach for elucidating the regulatory factors involved is by searching for possible interaction sites from the miRNA proximal areas in the genome. For this purpose, we examined the upstream regions of all *C. elegans* and *C. briggsae* miRNAs in Publication I. The four motif finding tools used in the search gave as the result a couple of very frequent patterns whose consensus sequences were diverse, due to the different motif and background models used in the algorithms. One of the most significantly enriched sequence motifs observed was GANNNNGA, which was abundant and could be found from all studied *C. elegans* and *C. briggsae* miRNA upstream regions. There was no clear difference in its abundance when comparing the sequences upstream of intergenic miRNAs and the miRNAs located in the protein coding area. Moreover, this motif had a conserved distribution in the area surrounding the miRNA loci in these two worms, with the highest frequency in the close vicinity upstream of miRNA start sites.

The actual function of the motif GANNNNGA remains unclear. It does not match to any known TFBS and there seems to be only a weak correlation between the number of GANNNNGA motifs and the absolute expression level of miRNAs, suggesting a function as a co-factor site, or a recognition sequence for miRNA processing. The abundance and conservation of this motif in the upstream of two old and important miRNAs, *mir-1* and *mir-124* suggest a connection to miRNAs with global specialized function. The lack of verified pri-miRNA precursor sequences further prevents us from drawing conclusions about the nature of this motif. At present, the only *C. elegans* miRNA for which the full pri-miRNA sequence and transcription start site has been characterized is *let-7* (Bracht et al., 2004). There are two pri-*let-7* sequences detected with RACE, the longer one starts 1066 nt, and the shorter one 225 nt upstream of the pre-*let-7* start in the genome. The frequency peak of motif GANNNNGA is ~500 nt upstream of the pre-miRNAs, which is in the area of the longer pri-*let-7* transcript, meaning that the function of this motif could also be post-transcriptional. However, the TSS of the shorter pri-*let-7* transcript is located downstream of the motif peak occurrence, referring to a potential role on the level of miRNA

transcription. On the whole, motif GANNNNGA and other conserved motifs found by us and in earlier studies (Ohler et al., 2004; Zhou et al., 2007) from miRNA upstream sequences are potential candidates as regulators of miRNA processing. They may be useful for miRNA prediction and for elucidating the details of the regulation of miRNA biogenesis.

## 6.2 MACHINE LEARNING IN MIRNA TARGET PREDICTION

In spite of significant increase of experimentally validated miRNA target genes (Vergoulis et al., 2012), the majority of the targets are still unknown. Computational target prediction is hence often the only source for identification of putative miRNA targets and the development of these tools is crucial for a better understanding of miRNA function. The starting point for the study in Publication II was the idea of using unsupervised machine learning in order to enhance miRNA target prediction. The method developed, mirSOM, is founded on the basic knowledge about animal miRNA target sites: these sites are imperfectly complementary with the miRNA sequence where the most crucial part is the miRNA seed, these sites are preferentially located in the 3′ UTR of the mRNA and are accessible to miRNA binding. The idea behind mirSOM is, that the putative target sites of each miRNA can be clustered together based on their mutual similarity. The unsupervised learning algorithm used by mirSOM clusters the putative target sites for each miRNA objectively, using no knowledge about the verified true and false targets until the optimal total free energy threshold for the site accessibility is defined. Thus, unlike the supervised machine learning methods applied earlier for this purpose (Kim et al., 2006; Yousef et al., 2007; Wang and El Naqa, 2008), mirSOM avoids all bias towards the characteristics of the available, experimentally verified positive and negative target sites. In comparison with seven other miRNA target prediction tools, mirSOM works best in finding the verified true and false miRNA target relationships, suggesting that miRNA target prediction can be improved by the use of machine learning methods. SOM networks identifying targets in the five prime untranslated regions (5′ UTR) and coding regions can be constructed similarly to the approach used in mirSOM, and it can also be implemented in other organisms. mirSOM has been applied in a few *C. elegans* miRNA studies (Rudgalvyte et al., 2013; Taki et al., 2014), and bioinformatics researchers who work in developing novel methods for miRNA target prediction have been interested in mirSOM workflow. The main deficiency of mirSOM at the moment is that it includes only those *C. elegans* genes which had an annotated 3′ UTR sequence at the time it was developed, which may lead to no results for the query of a particular gene. This is, however, a common problem when using a published tool or database: not all of them can be continuously updated because of lack of resources.

## 6.3 COMMON FEATURES OF *P. REDIVIVUS* AND *C. ELEGANS* MIRNAOMES

The draft genome and transcriptome of *P. redivivus*, together with the complement of miRNAs provides a powerful resource for comparative genomics (Publication III). It is the first free-living worm genome sequenced that does not belong to *Caenorhabditis* family and thus highlights features that are common with the genome of *C. elegans* (Sternberg and Horvitz 1981; Dillman, Mortazavi and Sternberg, 2012). Especially it may reflect common constraints and adaptations resulting from the free living lifestyle in comparison with parasitism. The repertoire of the miRNAs of *P. redivivus* can further be used to study and confirm the principles behind miRNA evolution in the nematode lineage, by providing examples of previously described mechanisms of emergence of novel miRNAs, such as gene duplication (Liu et al., 2008; Grimson et al., 2008; Wheeler et al., 2009) and antisense transcription (Ruby et al. 2007; Stark et al., 2008; Tyler et al., 2008; de Wit et al., 2009). Our comparative analysis shows similarities in the miRNAomes of *P. redivivus* and *C. elegans*: several miRNAs are conserved in these worms, but there is also considerable diversity in *P. redivivus*. Conservation of the two important miRNAs *lin-4* and *let-7* suggests common mechanisms in controlling gene expression during development (Lee, Feinbaum and Ambros, 1993; Wightman et al., 1993; Reinhart et al., 2000). However, there are also some highly expressed and also well conserved miRNAs whose function remains unknown.

## 6.4 HESC SPECIFIC EXPRESSION OF MIRNAS AND MORNAS

For the first time, we reported expression of moRNAs in hESCs (Publication IV). In our data, consisting of two hESC and one fibroblast small RNA sequencing libraries, over 90% of the moRNAs were detected solely in the hESC libraries. While earlier studies (Langenberger et al., 2009; Meiri et al., 2010; Bortoluzzi et al., 2012) had found most of the moRNA reads from the 5′-stems of miRNA hairpins, in our hESC libraries we observed abundant moRNA expression from the 3′-stems as well, thus highlighting the unique characteristics of the small RNAome of pluripotent stem cells in comparison with differentiated cells. Although many abundant moRNAs derived from highly expressed miRNA precursors, there was no clear correlation between the expression levels of moRNA and miRNA pairs derived from the same miRNA precursor stem. This is in accordance with earlier studies (Langenberger et al., 2009; Bortoluzzi et al., 2012), and suggests that these two molecules may arise independently from each other, implicating that moRNAs are not by-products of miRNA processing but functional molecules as well. The functionality of moRNAs is further supported by the observation that they are typically derived from phylogenetically old miRNA loci and are conserved themselves (Langenberger et al., 2009; Bortoluzzi et al., 2012).

Intriguingly, many significant moRNAs arised from precursors in the hESC specific *miR-302-367* cluster, and none of these moRNAs have been found in earlier studies reporting moRNAs from humans (Langenberger et al., 2009; Meiri et al., 2010; Bortoluzzi et al., 2012). These moRNAs were not found from our fibroblast library either, suggesting that their expression is ES cell specific like the expression of their host miRNAs. Similarly, several moRNAs derived from the large C19MC cluster are significantly over-expressed in the hESCs, but were not detected in earlier studies. On the other hand, moRNAs expressed from the oncogenic *miR-17–92* cluster: *moR-20a-5p, moR-18a-5p, moR-92a-1-5p,* found in cancer cells (Bortoluzzi et al., 2012; Meiri et al., 2010) were abundant in hESCs, but not in our fibroblast library. This may indicate a pluripotency related function of these moRNAs, as has been earlier suggested also for the corresponding miRNAs (Wilson et al., 2009). Nevertheless, the functional characterization and the processing of moRNAs remain to be elucidated.

## 6.5 FUTURE PROSPECTS

Genome-wide initiatives aiming to identify all the functional sequence elements like the Encyclopedia of DNA Elements, ENCODE (NHGRI, 2013) will in the future aid the efforts to unravel the miRNA regulatory networks. They will, in genome scale, include the TFBSs, as well as miRNA binding sites and other functional elements in 3′ UTRs. In addition, many of these features will be annotated in different developmental stages and conditions. Some of the transcription factors studied will probably regulate miRNA genes, and many additional miRNA target genes will be verified. Nevertheless, motifs that are not necessarily TFBSs, like GANNNNGA found in Publication I, will probably not be elucidated. Computational miRNA target prediction methods, like mirSOM introduced in Publication II, will still be needed for finding the proper miRNA target sites in the area of RISC factor binding sites from immunoprecipitation studies. However, the novel verified binding sites will in turn aid to improve the approaches for miRNA target prediction.

With NGS technologies it is now possible to obtain perhaps the whole entity of miRNAs of a genome in one experiment, as was accomplished in this study for the nematode *P. redivivus* (Publication III). As more genomes and their miRNAomes become available, the more information is revealed about miRNA evolution and the relevance of a specific miRNA for a lineage can be speculated based on its conservation. Furthermore, as seen in this thesis (Publication IV), complex small RNA profiles of biological samples can be detected using NGS. Computational tools are invaluable in interpreting the sequencing data and have led to a number of important discoveries in the field of small RNAs. An example of the power of deep sequencing is the discovery of moRNAs which are specifically expressed in pluripotent cells. Elucidation of moRNA biogenesis and their role

in the cell are essential for future research, and will provide us deeper understanding of small RNAs in general and especially their function in stem cells.

# *7 Summary and conclusions*

The aim of this thesis was to gain additional knowledge about specific fields of miRNA biology by using available computational methods and by developing new ones. The main findings are:

- A novel conserved motif, GANNNNGA, found from the close upstream region of all *C. elegans* and *C. briggsae* miRNAs (Publication I). This motif may function in miRNA transcriptional or post-transcriptional regulation, or it may serve as a recognition factor for miRNA biogenesis.

- A miRNA target prediction tool, mirSOM, based on clustering of short 3′ UTR substrings with SOM (Publication II). As mirSOM applies unsupervised learning, it avoids bias towards the characteristics of the small set of available, experimentally verified target sites.

- The miRNAome of *P. redivivus* (Publication III). Together with its draft genome and transcriptome, the complement of *P. redivivus* miRNAs provides a novel powerful resource for nematode comparative genomics from an evolutionary perspective.

- The identification of miRNAs and moRNAs expressed in hESCs and their specific expression patterns in comparison to human fibroblast miRNAs and moRNAs (Publication IV). Although the processing and function of moRNAs remain to be elucidated, this finding is a step forward in understanding the complex network of small RNAs operating to maintain the unique characteristics of stem cells.

# *References*


Aalto, A. P. and Pasquinelli, A. E. (2012) Small non-coding RNAs mount a silent revolution in gene expression. *Current Opinion in Cell Biology.* 24(3). p.333-340.

Abbott, A. L., Alvarez-Saavedra, E., Miska, E. A., Lau, N. C., Bartel, D. P., Horvitz, H. R. and Ambros, V. (2005) The *let-7* MicroRNA family members *mir-48*, *mir-84*, and *mir-241* function together to regulate developmental timing in *Caenorhabditis elegans*. *Developmental Cell.* 9(3). p.403-414.

Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Reczko, M. and Hatzigeorgiou, A. G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics.* 25(23). p.3049-3055.

Alvarez-Saavedra, E. and Horvitz, H. R. (2010) Many families of *C. elegans* microRNAs are not essential for development or viability. *Current Biology.* 20(4). p.367–373.

Ambros, V. (2011) MicroRNAs and developmental timing. *Current Opinion in Genetics & Development.* 21(4). p.511-517.

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X. et al. (2003) A uniform system for microRNA annotation. *RNA.* 9(3). p.277-279.

Ameres, S.L., Zamore, P.D. (2013) Diversifying microRNA sequence and function. *Nature Reviews Molecular Cell Biology*. 14(8). p.475-488.

Aravin, A. A., Hannon, G. J. and Brennecke, J. (2007) The Piwi–piRNA pathway provides an adaptive defense in the transposon arms race. *Science*. 318(5851). p.761–764.

Axtell, M. J., Westholm, J.O. and Lai, E.C. (2011) Vive la difference: biogenesis and evolution of microRNAs in plants and animals. *Genome Biology.* 12(4). p.221.

Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P. and Blelloch, R. (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other microprocessor-independent, dicer-dependent small RNAs. *Genes & Development*. 22(20). p. 2773–2785.

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P. and Bartel, D. P. (2008) The impact of microRNAs on protein output. *Nature*. 455(7209). p.64-71.

Bailey, T. L., Williams, N., Misleh, C. and Li, W. W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*. 34(Web Server issue). p.W369-W373.

Bar, M., Wyman, S. K., Fritz, B. R., Qi, J., Garg, K. S., Parkin, R. K. et al. (2008) MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. *Stem Cells*. 26(10). p.2496-2505.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I. et al. (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*. 40(Database issue). p.D57-D63.

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

Bartel, D. B. (2009) MicroRNAs: Target Recognition and Regulatory Functions. *Cell*. 136(3). p.215-233.

Baskerville, S. and Bartel, D. P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*. 11(3). p.241–247.

Bates, D., Chambers, J., Dalgaard, P., Falcon, S., Gentleman, R., Hornik, K. et al. (2013) *R: A Language and Environment for Statistical Computing*. [Online] Available at: http://www.R-project.org [Accessed: 1 October 2013].

Batista, P. J., Ruby, J. G., Claycomb, J. M., Chiang, R., Fahlgren, N., Kasschau, K. D. et al. (2008) PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Molecular Cell*. 31. p.67-78.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2013) GenBank. *Nucleic Acids Research*. 41(Database issue). p.D36-D42.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O. et al. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics*. 37(7). p.766-770.

Berezikov, E. (2011) Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics*. 12(12). p.846-860.

Berezikov, E., Chung, W. J., Willis, J., Cuppen, E. and Lai, E. C. (2007) Mammalian mirtron genes, *Molecular Cell*. 28(2). p.328-336.

Betel, D., Koppal, A., Agius, P., Sander, C. and Leslie, C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*. 11(8). p.R90.

Blow, M., Grocock, R., van Dongen, S., Enright, A. J., Dicks, E., Futreal, P. A. et al. (2006) RNA editing of human microRNAs. *Genome Biology*. 7(4). p.R27.

Boehm, M. and Slack F. J. (2006) MicroRNA control of lifespan and metabolism. *Cell Cycle*. 5(8). p.837-840.

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

Bortoluzzi, S., Bisognin, A., Biasiolo, M., Guglielmelli, P., Biamonte, F., Norfo, R. et al. (2012) Characterization and discovery of novel miRNAs and moRNAs in JAK2V617F-mutated SET2 cells. *Blood*. 119(13). p.e120-e130.

Bracht, J., Hunter, S., Eachus, R., Weeks, P. and Pasquinelli, A. E. (2004) Trans-splicing and polyadenylation of *let-7* microRNA primary transcripts. *RNA*. 10(10). p.1586-1594.

Brameier, M., Herwig, A., Reinhardt, R., Walter, L. and Gruber, J. (2011) Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. *Nucleic Acids Research*. 39(2). p.675–686.

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

Broderick, J. A. and Zamore, P. D. (2011) MicroRNA therapeutics. *Gene Therapy*. 18(12). p.1104–1110.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P. et al. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*. 41(Database issue). p.D226-D232.

Burroughs, A. M., Ando, Y., de Hoon, M. J., Tomaru, Y., Nishibu, T., Ukekawa, R. et al. (2010) A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Research*. 20(10). p.1398–1410.

Burroughs, A. M., Ando, Y., de Hoon, M. J., Tomaru, Y., Suzuki, H., Hayashizaki, Y. and Daub, C. O. (2011) Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. *RNA Biology*. 8(1). p.158-177.

Castanotto, D., Lingeman, R., Riggs, A. D. and Rossi, J. J. (2009) CRM1 mediates nuclear-cytoplasmic shuttling of mature microRNAs. *Proceedings of the National Academy of Sciences of the United States of America*. 106(51). p.21655-21659.

Chalfie, M., Horvitz, H. R. and Sulston, J. E. (1981) Mutations that lead to reiterations in the cell lineages of *C. elegans*. *Cell*. 24(1). p.59–69.

Cheloufi, S., Dos Santos, C. O., Chong, M. M. and Hannon, G. J. (2010) A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*. 465(7298). p.584-589.

Chen, K. and Rajewsky, N. (2006a) Deep conservation of microRNA-target relationships and 3'UTR motifs in vertebrates, flies, and nematodes. *Cold Spring Harbor Symposia on Quantitative Biology*. 71. p.149-156.

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

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. and Thompson, J. D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*. 31(13). p.3497-3500.

Chi, S. W., Zang, J. B., Mele, A. and Darnell, R. B. (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature*. 460(7254). p.479–486.

Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., Baek, D. et al. (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes & Development*. 24(10). p.992-1009.

Christodoulou, F., Raible, F., Tomer, R., Simakov, O., Trachana, K., Klaus, S. et al. (2010) Ancient animal microRNAs and the evolution of tissue identity. *Nature*. 463(7284). p.1084-1088.

Chung, W. J., Agius, P., Westholm, J. O., Chen, M., Okamura, K., Robine, N. et al. (2011) Computational and experimental identification of mirtrons in *Drosophila melanogaster* and *Caenorhabditis elegans*. *Genome Research*. 21(2). p.286-300.

Cifuentes, D., Xue, H., Taylor, D. W., Patnode, H., Mishima, Y., Cheloufi, S. et al. (2010) A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science*. 328(5986). p.1694-1698.

Cole, C., Sobala, A., Lu, C., Thatcher, S. R., Bowman, A., Brown, J. W. et al. (2009) Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA*. 15(12). p.2147-2160.

Crooks, G., Hon, G., Chandonia, J. and Brenner, S. (2004) WebLogo: A sequence logo generator. *Genome Research*. 14(6). p.1188-1190.

Czech, B., Zhou, R., Erlich, Y., Brennecke, J., Binari, R., Villalta, C. et al. (2009) Hierarchical rules for Argonaute loading in *Drosophila*. *Molecular Cell*. 36(3). p.445–456.

de Wit, E., Linsen, S. E., Cuppen, E. and Berezikov, E. (2009) Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Research*. 19(11). p.2064-2074.

Didiano, D. and Hobert, O. (2006) Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature Structural & Molecular Biology*. 13(9). p.849–851.

Dillman, A. R., Mortazavi, A. and Sternberg, P.W. (2012) Incorporating genomics into the toolkit of nematology. *Journal of Nematology*. 44(2). p.191-205.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A. et al. (2012) Landscape of transcription in human cells. *Nature*. 489(7414). p.101-108.

Doench, J. G. and Sharp, P. A. (2004) Specificity of microRNA target selection in translational repression. *Genes & Development*. 18(5). p.504–511.

Ender, C., Krek, A., Friedländer, M. R., Beitzinger, M., Weinmann, L., Chen, W. et al. (2008) A Human snoRNA with MicroRNA-Like Functions. *Molecular Cell*. 32(4). p.519–528.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D. S. (2003) MicroRNA targets in *Drosophila*. *Genome Biology*. 5(1). p.R1.

Farh, K. K., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P. et al. (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755). p.1817-1821.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. and Mello, C. C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*. 391(6669). p.806-811.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S. et al. (2013) Ensembl 2013. *Nucleic Acids Research*. 41(Database issue). p.D48-55.

Flynt, A. S., Greimann, J. C., Chung, W. J., Lima, C. D. and Lai, E. C. (2010) MicroRNA biogenesis via splicing and exosome-mediated trimming in *Drosophila*. *Molecular Cell*. 38(6). p.900–907.

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W. and Rajewsky, N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*. 40(1). p.37-52.

Friedman, R. C., Farh, K. K., Burge, C. B. and Bartel, D. P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*. 19(1). p.92–105.

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S. et al. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Research*. 37(Database issue). p.D136-D140.

Golden, D. E., Gerbasi, V. R. and Sontheimer, E. J. (2008) An inside job for siRNAs. *Molecular Cell*. 31(3). p.309–312.

Gomes, C. P., Cho, J. H., Hood, L., Franco, O. L., Pereira, R. W. and Wang, K. (2013) A Review of Computational Tools in microRNA Discovery. *Frontiers in Genetics*. 4. p.81.

Grad, Y., Aach, J., Hayes, G. D., Reinhart, B. J., Church, G. M., Ruvkun, G. and Kim, J. (2003) Computational and experimental identification of *C. elegans* microRNAs. *Molecular Cell*. 11(5). p.1253-1263.

Griffiths-Jones, S., Hui, J. H., Marco, A. and Ronshaugen, M. (2011) MicroRNA evolution by arm switching. *EMBO Reports*. 12(2). p.172-177.

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P. and Bartel, D. P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*. 27(1). p.91-105.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N. et al. (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*. 455(7217). p.1193–1197.

Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I. et al. (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell*. 106(1). p.23–34.

Grün, D., Wang, Y. L., Langenberger, D., Gunsalus, K. C. and Rajewsky, N. (2005) microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Computational Biology*. 1(1). p.e13.

Gu, W., Shirayama, M., Conte, D. Jr., Vasale, J., Batista, P. J., Claycomb, J. M. et al. (2009) Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. *Molecular Cell*. 36(2). p.231-244.

Guo, H., Ingolia, N. T., Weissman, J. S. and Bartel, D. P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*. 466(7308). p.835-840.

Hackenberg, M., Rodriguez-Ezpeleta, N. and Aransay, A. M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research*. 39(Web Server issue). p.W132–W138.

Hammell, M. (2010) Computational methods to identify miRNA targets. *Semiars in Cell & Developmental Biology*. 21(7). p.738-744.

Havens, M. A., Reich, A. A., Duelli, D. M. and Hastings, M. L. (2012) Biogenesis of mammalian microRNAs by a non-canonical processing pathway. *Nucleic Acids Research*. 40(10). p.4626-4640.

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S. et al. (2005) A microRNA polycistron as a potential human oncogene. *Nature*. 435(7043). p.828-833.

Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C. and 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(8). p.2946-2950.

Heo, I., Ha, M., Lim, J., Yoon, M. J., Park, J. E., Kwon, S. C. et al. (2012) Mono-uridylation of pre-MicroRNA as a key step in the biogenesis of group II *let-7* microRNAs. *Cell*. 151(3). p.521–532.

Heo, I., Joo, C., Cho, J., Ha, M., Han, J. and Kim, V. N. (2008) Lin28 Mediates the Terminal Uridylation of *let-7* Precursor MicroRNA. *Molecular Cell*. 32(2). p.276-k84.

Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C. et al. (2006) The expansion of the metazoan microRNA repertoire. *BMC Genomics*. 7. p.25.

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

Hovatta, O., Jaconi, M., Töhönen, V., Béna, F., Gimelli, S., Bosman, A. et al. (2010) A teratocarcinoma-like human embryonic stem cell (hESC) line and four hESC lines reveal potentially oncogenic genomic changes. *PLoS One*. 5(4). p.e10263.

Hovatta, O., Mikkola, M., Gertow, K., Strömberg, A. M., Inzunza, J., Hreinsson, J. et al. (2003) A culture system using human foreskin fibroblasts as feeder cells allows production of human embryonic stem cells. *Human Reproduction*. 18(7). p.1404-1409.

Hu, H. Y., Yan, Z., Xu, Y., Hu, H., Menzel, C., Zhou, Y. H. et al. (2009) Sequence features associated with microRNA strand selection in humans and flies. *BMC Genomics*. 10. p.413.

Huntzinger, E. and Izaurralde, E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews Genetics*. 12(2). p.99–110.

Hutvagner, G., McLachlan, J., Pasquinelli, A. E., Bálint, E., Tuschl, T. and Zamore, P. D. (2001) A cellular function for the RNA-interference entzyme Dicer in the maturation of the *let-7* small temporal RNA. *Science*. 293(5531). p.834-838.

Hutvagner, G. and Zamore, P. D. (2002) A microRNA in a multiple-turnover RNAi enzyme complex. *Science*. 297(5589). p.2056-2060.

Hwang, H. W. and Mendell, J. T. (2006) MicroRNAs in cell proliferation, cell death and tumorigenesis. *British Journal of Cancer*. 94(6). p.776-780.

Illumina (2014) *Illumina homepage*. [Online] Available from: http://www.illumina. Com [Accessed: 2nd January 2014].

Isik, M., Korswagen and H. C., Berezikov, E. (2010) Expression patterns of intronic microRNAs in *Caenorhabditis elegans*. *Silence*. 1(1). p.5.

Jan, C. H., Friedman, R. C., Ruby, J. G. and Bartel, D. P. (2011) Formation, Regulation and Evolution of *Caenorhabditis elegans* 3′UTRs. *Nature*. 469(7328). p.97–101.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C. and Marks, D. S. (2004) Human MicroRNA Targets. *PLoS Biology*. 2(11). p.e363.

Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A. et al. (2005) RAS Is Regulated by the *let-7* MicroRNA Family. *Cell*. 120(5). p.635-647.

Johnson, S. M., Lin, S. Y. and Slack, F. J. (2003) The time of appearance of the *C. elegans let-7* microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Developmental Biology*. 259(2). p.364-379.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research*. 110(1-4). p.462-467.

Kankainen, M. and Holm, L. (2005) POCO: discovery of regulatory patterns from promoters of oppositely expressed gene sets. *Nucleic Acids Research*. 33(Web Server issue). p.W427-W431.

Kankainen, M., Pehkonen, P., Rosenstöm, P., Törönen, P., Wong, G. and Holm, L. (2006) POXO: a web-enabled tool series to discover transcription factor binding sites. *Nucleic Acids Research*. 34(Web Server issue). p.W534-W540.

Katoh, T., Sakaguchi, Y., Miyauchi, K., Suzuki, T., Kashiwabara, S., Baba, T. and Suzuki T. (2009) Selective stabilization of mammalian microRNAs by 3′ adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes & Development*. 23(4). p.433-438.

Kazazian, H. H. Jr. (2004) Mobile elements: drivers of genome evolution. *Science*. 303(5664). p.1626-1632.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. and Segal, E. (2007) The role of site accessibility in microRNA target recognition *Nature Genetics*. 39(10). p.1278–1284.

Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J. and Plasterk, R. H. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes & Development*. 15(20). p.2654-2659.

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

Kim, D. H., Saetrom, P., Snove, O. Jr. and Rossi, J. J. (2008) MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*. 105(42). p. 16230–16235.

Kim, S. K., Nam, J. W., Rhee, J. K., Lee, W. J. and Zhang, B. T. (2006) miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*. 7. p.411.

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

Kim, Y. and Kim, V. N. (2012) MicroRNA Factory: RISC Assembly from Precursor MicroRNAs. *Molecular Cell*. 46(4). p.384-386.

Kim, Y. K. and Kim, V. N. (2007) Processing of intronic microRNAs. *The EMBO Journal*. 26(3). p. 775–783.

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. and Hatzigeorgiou, A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes & Development*. 18(10). p.1165-1178.

Knight, S. W. and Bass, B. L. (2001) A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. *Science*. 293(5538). p.2269-2271.

Kohonen, T. (2001) *Self-Organizing Maps*. 3rd ed. Berlin: Springer.

Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*. 39(Database issue). p.D152–D157.

Krek, A., Grun, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J. et al. (2005) Combinatorial microRNA target predictions. *Nature Genetics*. 37(5). p.495–500.

Kuhn, D. E., Martin, M. M., Feldman, D. S., Terry, A. V., Nuovo, G. J. and Elton, T. S. (2008) Experimental validation of miRNA targets. *Methods*. 44(1). p.47–54.

Ladewig, E., Okamura, K., Flynt, A. S., Westholm, J. O. and Lai, E. C. (2012) Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Research*. 22(9). p.1634-1645.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*. 294(5543). p.853–858.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W. and Tuschl, T. (2002) Identification of tissue-specific microRNAs from mouse. *Current Biology*. 12(9). p.735–739.

Lai, E. C. (2002) Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genetics*. 30(4). p.363-364.

Lai, E. C., Tomancak, P., Williams, R. W. and Rubin, G. M. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biology*. 4(7). p.R42.

Lall, S., Grün, D., Krek, A., Chen, K., Wang, Y. L., Dewey, C. N. et al. (2006) A genome-wide map of conserved microRNA targets in *C. elegans*. *Current Biology*. 16(5). p.460-471.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A. et al. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*. 129(7). p.1401–1414.

Langenberger, D., Bermudez-Santana, C., Hertel, J., Hoffmann, S., Khaitovich, P. and Stadler, P. F. (2009) Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*. 25(18). p.2298-2301.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 10(3). p.R25.

Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*. 32(1). p.11-16.

Lau, N. C., Lim, L. P., Weinstein, E. G. and Bartel, D. P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*. 294(5543). p.858–862.

Lee, R. C. and Ambros, V. (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*. 294(5543). p.862–864.

Lee, R. C., Feinbaum, R. L. and Ambros, V. (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*. 75(5). p.843–854.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J. et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*. 425(6956). p.415-419.

Lee, Y., Jeon, K., Lee, J. T., Kim, S. and Kim, V. N. (2002) MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal*, 21(17). p.4663-4670.

Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H. and Kim, V. N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal*. 23(20). p.4051–4060.

Lehrbach, N. J., Armisen, J., Lightfoot, H. L., Murfitt, K. J., Bugaut, A., Balasubramanian, S. and Miska, E. A. (2010) LIN-28 and the poly(U) polymerase PUP-2 regulate *let-7* microRNA processing in *Caenorhabditis elegans*. *Nature Structural & Molecular Biology*. 16(10). p.1016–1020.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. and Burge, C. B. (2003) Prediction of mammalian microRNA targets. *Cell*. 115(7). p.787-798.

Li, Y., Zhang, Z., Liu, F., Vongsangnak, W., Jing, Q. and Shen, B. (2012) Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Research*. 40(10). p.4298-4305.

Liang, H. and Li, W. H. (2009) Lowly expressed human microRNA genes evolve rapidly. *Molecular Biology and Evolution*. 26(6). p.1195-1198.

Liao, J. Y., Ma, L. M., Guo, Y. H., Zhang, Y. C., Zhou, H., Shao, P. et al. (2010) Deep sequencing of human nuclear and cytoplasmic small RNAs reveals an unexpectedly

complex subcellular distribution of miRNAs and tRNA 3' trailers. *PLoS One*. 5(5). p.e10563.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W. et al. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*. 456(7221). p.464-469.

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J. et al. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*. 433(7027). p.769-773.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W. et al. (2003) The microRNAs of *Caenorhabditis elegans*. *Genes & Development*. 17(8). p.991-1008.

Lin, Y. C., Hsieh, L. C., Kuo, M. W., Yu, J., Kuo, H. H., Lo, W. L. et al. (2007) Human TRIM71 and Its Nematode Homologue Are Targets of *let-7* MicroRNA and Its Zebrafish Orthologue Is Essential for Development. *Molecular Biology and Evolution*. 24(11). p.2525-2534.

Liu, N., Okamura, K., Tyler, D. M., Phillips, M. D., Chung, W. J. and Lai, E. C. (2008) The evolution and functional diversification of animal microRNA genes. *Cell Research*. 18(10). p.985–996.

Loher, P. and Rigoutsos, I. (2012) Interactive exploration of RNA22 microRNA target predictions. *Bioinformatics*. 28(24). p.3322-3323.

Mahony, S., Auron, P. E. and Benos, P. V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Computational Biology*. 3(3). p.e61.

Martens-Uzunova, E. S., Olvedy, M. and Jenster, G. (2013) Beyond microRNA - Novel RNAs derived from small non-coding RNA and their implication in cancer. *Cancer Letters*. 340(2). p.201-211.

Martinez, NJ, Ow, M. C., Barrasa, M. I., Hammell, M., Sequerra, R., Doucette-Stamm, L. et al. (2008) A *C. elegans* genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes & Development*. 22(18). p.2535-2549.

Massirer, K. B., Perez, S. G., Mondol, V. and Pasquinelli, A. E. (2012) The *miR-35-41* family of microRNAs regulates RNAi sensitivity in *Caenorhabditis elegans*. *PLoS Genetics*. 8(3). p.e1002536.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A. et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*. 34(Database issue). p.D108-D110.

Meiri, E., Levy, A., Benjamin, H., Ben-David, M., Cohen, L., Dov, A. et al. (2010) Discovery of microRNAs and other small RNAs in solid tumors. *Nucleic Acids Research*. 38(18). p.6234-6246.

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M. et al. (2012) The UCSC Genome Browser database: extensions and updates 2013) *Nucleic Acids Research*. 41(D1). p. D64-D69.

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M. et al. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*. 126(6). p.1203-1217.

Miska, E. A., Alvarez-Saavedra, E., Abbott, A. L., Lau, N. C., Hellman, A. B., McGonagle, S. M. et al. (2007) Most *Caenorhabditis elegans* microRNAs are individually not essential for development or viability. *PLoS Genetics*. 3(12). p.e215.

Monteys, A. M., Spengler, R. M., Wan, J., Tecedor, L., Lennox, K. A., Xing, Y. and Davidson, B. L. (2010) Structure and activity of putative intronic miRNA promoters. *RNA*. 16(3). p.495-505.

Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A. L. et al. (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research*. 18(4). p.610-621.

Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L. et al. (2002) miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes & Development*. 6(6). p.720-728.

Nana-Sinkam, S. P. and Croce, C. M. (2013) Clinical applications for microRNAs in cancer. *Clinical Pharmacology & Therapeutics*. 93(1). p.98-104.

NHGRI (2013) *The ENCODE Project: ENCyclopedia Of DNA Elements*. [Online] Available at: <http://www.genome.gov/10005107> [Accessed 1 October 2013].

Obernosterer, G., Leuschner, P. J., Alenius, M. and Martinez, J. (2006) Post-transcriptional regulation of microRNA expression. *RNA*. 12(7). p.1161-1167.

Ohler, U., Yekta, S., Lim, L., Bartel, D. P. and Burge, C. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for MicroRNA gene identification. *RNA*. 10(9). p.1309-1322.

Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M. and Lai, E. C. (2007) The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in *Drosophila*. *Cell*. 130(1). p.89-100.

Okamura, K., Ladewig, E., Zhou, L. and Lai, E. C. (2013) Functional small RNAs are generated from select miRNA hairpin loops in flies and mammals. *Genes & Development*. 27(7). p. 778-792.

Okamura, K., Liu, N. and Lai, E. C. (2009) Distinct mechanisms for microRNA strand selection by *Drosophila* Argonautes. *Molecular Cell*. 36(3). p.431-444.

Ono, K., Kuwabara, Y. and Han, J. (2011) MicroRNAs and cardiovascular diseases. *The FEBS Journal*. 278(10). p.1619-1633.

Ono, M., Scott, M. S., Yamada, K., Avolio, F., Barton, G. J. and Lamond, A. I. (2011) Identification of human miRNA precursors that resemble box C/D snoRNAs. *Nucleic Acids Research*. 39(9). p.3879-3891.

Ow, M. C., Martinez, N. J., Olsen, P. H., Silverman, H. S., Barrasa, M. I., Conradt, B. et al. (2008) The FLYWCH transcription factors FLH-1, FLH-2, and FLH-3 repress embryonic expression of microRNA genes in *C. elegans*. *Genes & Development*. 22(18). p.2520-2534.

Ozsolak, F., Poling, L. L., Wang, Z., Liu, H., Liu, X. S., Roeder, R. G. et al. (2008) Chromatin structure analyses identify miRNA promoters. *Genes & Development*. 22(22). p.3172–3183.

Pak, J. and Fire, A. (2007) Distinct Populations of Primary and Secondary Effectors During RNAi in *C. elegans*. *Science*. 315(5809). p.241-244.

Pantano, L., Estivill, X. and Marti, E. (2010) SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Research*. 38(5). p.e34.

Pasquinelli, A. E., Reinhart, B. J. and Slack F. J., Martindale, M. Q., Kuroda, M. I., Maller, B. et al. (2000) Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature*. 408(6808). p.86–89.

Pavesi, G., Mauri, G. and Pesole, G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*. 17(suppl 1). p. S207-S214.

Pavesi, G., Mereghetti, P., Zambelli, F., Stefani, M., Mauri, G. and Pesole, G. (2006) MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Research*. 34(Web Server issue). p.W566-W570.

Pervouchine, D. D., Graber, J. H. and Kasif, S. (2003) On the normalization of RNA equilibrium free energy to the length of the sequence. *Nucleic Acids Research*. 31(9). p.e49.

Place, R. F., Li, L. C., Pookot, D., Noonan, E. J. and Dahiya, R. (2008) MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proceedings of the National Academy of Sciences of the United States of America*. 105(5). p.1608-1613.

Pritchard, C. C., Cheng, H. H. and Tewari, M. (2012) MicroRNA profiling: approaches and considerations. *Nature Reviews Genetics*. 13(5). p.358-369.

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

Pruitt, K. D., Tatusova, T., Klimke, W. and Maglott, D. R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*. 37(Database issue). p.D32-D36.

Rehmsmeier, M., Steffen, P., Hochsmann, M. and Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*. 10(10). p.1507-1517.

Reinhart, B. J., Slack F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie A. E. et al. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*. 403(6772). p.901–906.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*. 16(6). p.276-277.

Ro, S., Park, C., Young, D., Sanders, K. M. and Yan, W. (2007) Tissue-dependent paired expression of miRNAs. *Nucleic Acids Research*. 35(17). p.5944-5953.

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

Rottiers, V. and Näär, A. M. (2012) MicroRNAs in metabolism and metabolic disorders. *Nature Reviews Molecular Cell Biology*. 13(4). p.239-250.

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C. et al. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*. 127(6). p.1193-207.

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

Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P. and Lai, E. C. (2007) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Research*. 17(12). p.1850-1864.

Rudgalvyte, M., VanDuyn, N., Aarnio, V., Heikkinen, L., Peltonen, J., Lakso, M. et al. (2013) Methylmercury exposure increases lipocalin related (lpr) and decreases activated in blocked unfolded protein response (abu) genes and specific miRNAs in *Caenorhabditis elegans*. *Toxicology Letters*. 222(2). p.189-196.

Saetrom, P., Heale, B. S., Snøve, O. Jr, Aagaard, L., Alluin, J. and Rossi, J. J. (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Research*. 35(7). p.2333-2342.

Sai Lakshmi, S. and Agrawal, S. (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Research.* 36(Database issue). p.D173–D177.

Saini, H. K., Saini, H. K., Griffiths-Jones, S. and Enright, A. J. (2007) Genomic analysis of human microRNA transcripts. *Proceedings of the National Academy of Sciences of the United States of America*. 104(45). p.17719-17724.

Schmittgen, T. D., Lee, E. J., Jiang, J., Sarkar, A., Yang, L., Elton, T. S. and Chen, C. (2008) Real-time PCR quantification of precursor and mature microRNA. *Methods*. 44(1). p.31-38.

Schwarz, D. S., Hutvagner, G., Du, T., Xu, Z. S., Aronin, N. and Zamore, P. D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*. 115(2). p.199–208.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*. 455(7209). p.58-63.

Sempere, L. F., Cole, C. N., McPeek, M. A. and Peterson, K. J. (2006) The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *Journal of Experimental Zoology Part B Molecular and Developmental Evolution*. 306(6). p.575–588.

Sempere, L. F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E. and Ambros, V. (2004) Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biology*. 5(3). p.R13.

Sethupathy, P., Megraw, M. and Hatzigeorgiou, A. G. (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nature Methods*. 3(11). p.881-886.

Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M. J. et al. (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*. 6:267.

Shi, W., Hendrix, D., Levine, M. and Haley, B. (2009) A distinct class of small RNAs arises from pre-miRNA–proximal regions in a simple chordate. *Nature Structural & Molecular Biology*. 16(2). p.183–189.

Shin, C., Nam, J. W., Farh, K. K., Chiang, H. R., Shkumatava, A. and Bartel, D. P. (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular Cell*. 38(6). p.789–802.

Shukla, G. C., Singh, J. and Barik, S. (2011) MicroRNAs: Processing, Maturation, Target Recognition and Regulatory Functions. *Molecular and Cellular Pharmacology*, 3(3). p.83–92.

Sijen, T., Steiner, F. A., Thijssen, K. L. and Plasterk, R. H. (2007) Secondary siRNAs Result from Unprimed RNA Synthesis and Form a Distinct Class. *Science*. 315(5809). p.244-247.

Sibley, C. R., Seow, Y., Saayman, S., Dijkstra, K. K., El Andaloussi, S., Weinberg, M. S. and Wood, M. J. (2012) The biogenesis and characterization of mammalian microRNAs of mirtron origin. *Nucleic Acids Research*. 40(1). p.438-48.

Siomi, M. C., Sato, K., Pezic, D. and Aravin, A. A. (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Reviews. Molecular Cell Biology*. 12(4). p.246-258.

Smalheiser, N. R. and Torvik, V. I. (2005) Mammalian microRNAs derived from genomic repeats. *Trends in Genetics*. 21(6). p.322-326.

Song, G. and Wang, L. (2008) *MiR-433* and *miR-127* arise from independent overlapping primary transcripts encoded by the *miR-433–127* locus. *PLoS ONE*. 3(10). p.e3574.

Stark, A., Brennecke, J., Russell, R. B. and Cohen, S. M. (2003) Identification of *Drosophila* microRNA targets. *PLoS Biology*. 1(3). p.E60.

Stark, A., Bushati, N., Jan, C. H., Kheradpour, P., Hodges, E., Brennecke, J. et al. (2008) A single Hox locus in *Drosophila* produces functional microRNAs from opposite DNA strands. *Genes & Development*. 22(1). p.8-13.

Sternberg, P. W. and Horvitz, H. R. (1981) Gonadal cell lineages of the nematode *Panagrellus redivivus* and implications for evolution by the modification of cell lineage. *Developmental Biology*. 88(1). p.147–166.

Suh, M., Lee, Y., Kim, J. Y., Kim, S. K., Moon, S. H., Lee, J. Y. et al. (2004) Human embryonic stem cells express a unique set of microRNAs. *Developmental Biology*. 270(2). p.488-498.

Taft, R. J., Glazov, E. A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G. J. et al. (2009) Tiny RNAs associated with transcription start sites in animals. *Nature Genetics*. 41(5). p.572-578.

Taki, F. A., Pan, X. and Zhang, B. (2014) Chronic nicotine exposure systemically alters microRNA expression profiles during post-embryonic stages in *Caenorhabditis elegans*. *Journal of Cellular Physiology*. 229(1). p.79-89.

Tarver, J. E., Donoghue, P. C. and Peterson, K. J. (2012) Do miRNAs have a deep evolutionary history? *Bioessays*. 34(10). p.857-866.

Thompson, W., Rouchka, E. C. and Lawrence, C. E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Research*. 31(13). p.3580-3585.

Thomson, D. W., Bracken, C. P. and Goodall, G. J. (2011) Experimental strategies for microRNA target identification. *Nucleic Acids Research.* 39(16). p.6845-6853.

Thomson, J. M., Newman, M., Parker, J. S., Morin-Kensicki, E. M., Wright, T. and Hammond, S. M. (2006) Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes & Development*. 20(16). p.2202–2207.

Tomari, Y. and Zamore, P. D. (2005) Perspective: machines for RNAi. *Genes & Development*, 19:517–529

Tyler, D. M., Okamura, K., Chung, W. J., Hagen, J. W., Berezikov, E., Hannon, G. J. and Lai, E. C. (2008) Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes & Development*. 22(1). p.26-36.

Unger, C., Felldin, U., Nordenskjöld, A., Dilber, M. S. and Hovatta, O. (2008) Derivation of human skin fibroblast lines for feeder cells of human embryonic stem cells. *Current Protocols in Stem Cell Biology.* 5:1C.7.1–1C.7.10.

Vasale, J. J., Gu, W., Thivierge, C., Batista, P. J., Claycomb, J. M., Youngman, E. M. et al. (2010) Sequential rounds of RNA-dependent RNA transcription drive endogenous small-RNA biogenesis in the ERGO-1/Argonaute pathway. *Proceedings of the National Academy of Sciences of the United States of America*. 107(8). p.3582-3587.

Vasudevan, S. (2012) Posttranscriptional upregulation by microrna*s*. *Wiley Interdisciplinary Reviews: RNA*. 3(3). p. 311-330.

Vergoulis, T., Vlachos, I. S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M. et al. (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Research*. 40(Database issue). p.D222-D229.

Viswanathan, S. R., Daley, G. Q. and Gregory, R. I. (2008) Selective blockade of microRNA processing by Lin28. *Science*. 320(5872). p.97-100.

Vlieghe, D., Sandelin, A., De Bleser, P. J., Vleminckx, K., Wasserman, W. W., van Roy, F. and Lenhard, B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Research*. 34(Database issue). p.D95-D97.

Wang, X. and El Naqa, I. M. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*. 24(3). p.325-332.

Weinmann, L., Höck, J., Ivacevic, T., Ohrt, T., Mütze, J., Schwille, P. et al. (2009) Importin 8 is a gene silencing factor that targets argonaute proteins to distinct mRNAs. *Cell*. 136(3). p.496-507.

Westholm, J. O. and Lai, E. C. (2011) Mirtrons: microRNA biogenesis via splicing. *Biochimie*. 93(11). p.1897-1904.

Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S. and Peterson, K. J. (2009) The deep evolution of metazoan microRNAs. *Evolution & Development*. 11(1). p.50-68.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V. et al. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. 36(Database issue). p.D13-D21.

Wightman, B., Burglin, T. R., Gatto, J., Arasu, P. and Ruvkun, G. (1991) Negative regulatory sequences in the *lin-14* 3′-untranslated region are necessary to generate a temporal switch during *Caenorhabditis elegans* development. *Genes & Development*. 5(10). p.1813–1824.

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

Wilson, K. D., Venkatasubrahmanyam, S., Jia, F., Sun, N., Butte, A. J. and Wu, J. C. (2009) MicroRNA profiling of human-induced pluripotent stem cells. *Stem Cells and Development*. 18(5). p.749-758.

Winter J., Jung S., Keller S., Gregory R. I. and Diederichs S. (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature Cell Biology*. 11(3). p. 228-234.

Winter, J., Link, S., Witzigmann, D., Hildenbrand, C., Previti, C. and Diederichs, S. (2013) Loop-miRs: active microRNAs generated from single-stranded loop regions. *Nucleic Acids Research*. 41(10). p.5503–5512.

Wu, H., Ye, C., Ramirez, D. and Manjunath, N. (2009) Alternative processing of primary microRNA transcripts by Drosha generates 5′ end variation of mature microRNA. *PLoS One*. 4(10). p.e7566.

Wu, Y., Wei, B., Liu, H., Li, T. and Rayner, S. (2011) MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics*. 12:107.

Wulczyn, F. G., Smirnova, L., Rybak, A., Brandt, C., Kwidzinski, E., Ninnemann, O. et al. (2007) Post-transcriptional regulation of the *let-7* microRNA during neural cell specification. *The FASEB Journal*. 21(2). p.415-426.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K. et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*. 434(7031). p.338–345.

Xu, G., Fewell, C., Taylor, C., Deng, N., Hedges, D., Wang, X. et al. (2010) Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq. *RNA*. 16(8). p.1610-1622.

Yang, J. S., Maurin, T., Robine, N., Rasmussen, K. D., Jeffrey, K. L., Chandwani, R. et al. (2010) Conserved vertebrate *mir-451* provides a platform for Dicer-independent, Ago2-mediated microRNA biogenesis. *Proceedings of the National Academy of Sciences of the United States of America*. 107(34). p.15163-15168.

Yang, J. S., Phillips, M. D., Betel, D., Mu, P., Ventura, A., Siepel, A. C. et al. (2011) Widespread regulatory activity of vertebrate microRNA* species. *RNA*. 17(2). p.312–326.

Yi, R., Qin, Y., Macara, I. G. and Cullen, B. R. (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & Development*. 17(24). p.3011–3016.

Yin, J. Q., Zhao, R. C. and Morris, K. V. (2008) Profiling microRNA expression with microarrays. *Trends in Biotechnology*. 26(2). p.70-76.

Yook, K., Harris, T. W., Bieri, T., Cabunoc, A., Chan, J., Chen, W. J. et al, (2012) WormBase 2012: more genomes, more data, new website. *Nucleic Acids Research*. 40:D735–D741.

Yousef, M., Jung, S., Kossenkov, A. V., Showe, L. C. and Showe, M. K. (2007) Naïve Bayes for microRNA target predictions--machine learning for microRNA targets. *Bioinformatics*. 23(22). p.2987-2992.

Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L. C. and Showe, M. K. (2006) Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier. *Bioinformatics*. 22(11). p.1325–1334.

Yuan, X., Liu, C., Yang, P., He, S., Liao, Q., Kang, S. and Zhao, Y. (2009) Clustered microRNAs' coordination in regulating protein-protein interaction network. *BMC Systems Biology*. 3:65.

Zhong, X., Coukos, G. and Zhang L. (2012) miRNAs in human cancer. *Methods in Molecular Biology*. 822:295-306.

Zhou, H., Arcila, M. L., Li, Z., Lee, E. J., Henzler, C., Liu, J. et al. (2012) Deep annotation of mouse iso-miR and iso-moR variation. *Nucleic Acids Research*. 40(13). p.5864-5875.

Zhou, X., Ruan, J., Wang, G. and Zhang, W. (2007) Characterization and identification of microRNA core promoters in four model species. *PLoS Computational Biology*. 3(3). p.e37.

Zisoulis, D. G., Lovci, M. T., Wilbert, M. L., Hutt, K. R., Liang, T. Y., Pasquinelli, A. E. and Yeo, G. W. (2010) Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nature Structural & Molecular Biology*. 17(2). p.173–179.

**Liisa Heikkinen**

# *Computational Analysis of Small Non-coding RNAs in Model Systems*

Gene expression modulation by small non-coding RNAs is a recently discovered regulatory mechanism in eukaryotes. In this thesis, high-throughput genomic methods and bioinformatics were applied to study small RNA biology. The results include discovery of a novel sequence motif in miRNA promoters, the miRNAome of nematode Panagrellus redivivus, the first report of abundant expression of moRNAs in human embryonic stem cells, and a new bioinformatic method for miRNA target prediction. This thesis provides novel resources and bioinformatic methods for scientific investigation.

UNIVERSITY OF EASTERN FINLAND

Publications of the University of Eastern Finland

*Dissertations in Health Sciences*

ISBN 978-952-61-1385-2