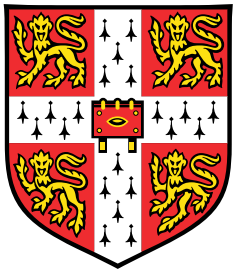
# Elucidating the function and biogenesis of small non-coding RNAs using novel computational methods & machine learning



**Dimitrios Vitsios**

EMBL - European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Selwyn College

December 2017

I would like to dedicate this thesis to my friends and to my family.

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Dimitrios Vitsios
December 2017

# Acknowledgements

Well…"it's been one hell of a ride". Reaching the end of this long journey, after so many ups and downs, I cannot but express my gratitude to all those who stood by me and helped me come through this challenging but highly rewarding and transforming experience.

First and foremost, I would like to thank my supervisor, Dr Anton Enright, for embracing me in the lab from the beginning, for his continuous guidance throughout my PhD and his insightful ideas that motivated me in all my projects. This work would not have been possible without the invaluable input from Matthew Davis as well. The countless discussions I had with him, his expertise and insight in biological problems and his persistence in detail guided me through the vast majority of my projects and helped me overcome many of the problems I encountered along the way.

I am also grateful to all other members of the Enright Group for their scientific guidance and support: Stijn van Dongen, Tommaso Leonardi, Elsa Kentepozidou, Adrien Leger and Jack Monahan. An extra special thanks to Anton, Mat, Adrien, Jack and Elsa for their significant contribution to the proofreading of this thesis.

I would also like to thank the members of my Thesis Advisory Committee for their valuable feedback all these years: Prof Donal O'Carroll, Prof Eric Miska and Dr Oliver Stegle.

Many people would suggest that you need to reach the bottom in order to raise again. Life throughout my PhD studies has certainly taught me that. I owe my biggest gratitude to all those friends and people around me that stood by me and kept me sane during those difficult times along the road. My greatest gratefulness goes to my friends Dimitris from Florina, Giorgos from Katerini and Giorgos from Kalamata. There are many other significant ones of course whose impact was determinant for me to persist and grow stronger from all these stiff circumstances. If it wasn't for them, I'm not sure this thesis would have come to an end now.

Finally, I want to express my greatest gratitude to Elsa, for standing by me, supporting me till the end and just…being in my life!

Alright, enough with the tears. Thank God for having been through this awesome though challenging experience and for having met so many important people in my life in the meantime.

Last but not least, I want to thank my mother and my sister for their support all these years and welcome our newest member, my newly born nephew Nicholas!

*Cambridge, 14th August 2017*
Dimitrios Vitsios

# Abstract

The discovery of RNA in 1868 by Friedrich Miescher was meant to be the prologue to an exciting new era in Biology full of scientific breakthroughs and accomplishments. Since then, RNAs have been proven to play an indispensable role in biological processes such as coding, decoding, regulation and expression of genes. In particular, the discovery of small non-coding RNAs and especially miRNAs, in *C. elegans* first and thereafter to almost all animals and plants, started to fill in the puzzle of a complex gene regulatory network present within cells. The aim of this thesis is to shed more light on the features and functionality of small RNAs. In particular, we will focus on the function and biogenesis of miRNAs and piRNAs, across multiple species, by employing advanced computational methods and machine learning.

We first introduce a novel method (Chimira) for the identification of miRNAs from sets of animal and plant hairpin precursors along with post-transcriptional terminal modifications that are not encoded by the genome. This method allows the characterisation of the prevalence of miRNA isoforms within different cell types and/or conditions. We have applied Chimira within a larger study that examines the effect of terminal uridylation in RNA degradation in oocytes and cells in either embryonic or adult stage. This study showed that uridylation is the predominant transcriptional regulation mechanism in oocytes while it does not retain the same functionality on mRNAs and miRNAs, both in embryonic and adult cells.

We then move on to a large-scale analysis of small RNA-Seq datasets in order to identify potential modification signatures across specific conditions and cell types or tissues in Human and Mouse. We extracted the full modification profiles across 461 samples, unveiling the high prevalence of modification signatures of mainly 1 to 4 nucleotides. Additionally, samples of the same cell type and/or condition tend to cluster together based on their miRNA modification profiles while miRNA gene precursors with close genomic proximity showed a significant degree of co-expression. Finally, we elucidate the determinant factors in strand selection during miRNA biogenesis as well as update the miRBase annotation with corrected miRNA isoform sequences.

Next, we introduce a novel computational method (mirnovo) for miRNA prediction from RNA-Seq data with or without a reference genome using machine learning. We demonstrate its efficiency by applying it to multiple datasets, including single cells and RNaseIII deficient samples, supporting previous studies for the existence of non-canonical miRNA biogenesis pathways. Following this, we explore and justify a novel piRNA biogenesis pathway in Mouse which is independent of the MILI enzyme. Finally, we explore the efficiency of CRISPR/Cas9 induced editing of miRNA targets based on the computationally predicted accessibility of the targeted regions in the genome.

We have publicly released two web-based novel computational methods and one online resource with results regarding miRNA biogenesis and function. All findings presented in this study comprise another step forward within the journey of elucidation of RNA functionality and we believe they will be of benefit to the scientific community.

# Table of contents

# List of figures

# List of tables

# Chapter 1

# Introduction

In 1868, Johannes Friedrich Miescher, a Swiss physician and biologist, was the first scientist to isolate nucleic acid and particularly Ribonucleic acid (RNA) molecules (Dahm, 2005; Miescher and Schmiedeberg, 1896). It wasn't until 1939 though when RNA was formally shown to be playing a role in protein synthesis (Caspersson and Schultz, 1939). In 1956, the first structure of an RNA crystal was determined via X-ray crystallography (Rich and Davies, 1956) and a few years later, in 1965, the first transfer RNA (tRNA) sequence (77 nucleotides in length) from yeast was characterised (Holley et al., 1965). The discovery of retroviruses and reverse transcriptase in the early 1970s revealed the potential of transcribing RNA back to DNA sequences (Baltimore, 1964; Mizutani and Temin, 1970). Around 1990, the discovery of RNA interference (Fire et al., 1998; Napoli et al., 1990) through small interference RNAs (siRNAs) as an innate gene silencing mechanism was meant to establish the foundation for deciphering the complex regulatory network within cells. About a decade later, in 2001, the identification of 22 nt long RNAs or microRNAs (miRNAs), first in *C. elegans* and shortly after in other animals as well as plants (Baulcombe, 2002; Lagos-Quintana et al., 2002; Pasquinelli et al., 2000; Reinhart et al., 2000, 2002; Tuschl et al., 1999), defined in a more solid way the landscape of gene regulation via small RNA molecules. Since then, small RNAs and in particular miRNAs have been found to play an indispensable role in RNA silencing and post-transcriptional regulation of gene expression (Ambros, 2004; Bartel, 2004) while some of them can act as disease response biomarkers (Wu and Mo, 2009) or even as tumour antagonists (Liu et al., 2014a).

## 1.1 MicroRNAs

Non-coding regulators such as miRNAs have been a significant avenue of research since their discovery and the realisation that they are both widespread in animals and plants and

also often highly conserved (Altuvia et al., 2005; Lee et al., 2007; Zhang et al., 2006). The main mode of regulation by miRNAs in animals is translational repression and degradation of target transcripts (Baulcombe, 2004; Kai and Pasquinelli, 2010; Lim et al., 2005; Pratt and MacRae, 2009). This regulation involves the binding of a mature 19-22nt miRNA to a target transcript through direct formation of a double stranded duplex driven by complementarity between the miRNA and the target site (Lewis et al., 2003). This binding event is initiated through the so-called *seed region* (Lewis et al., 2005, 2003) of the miRNA (nucleotides 2-8), which requires for the most part perfect complementarity.

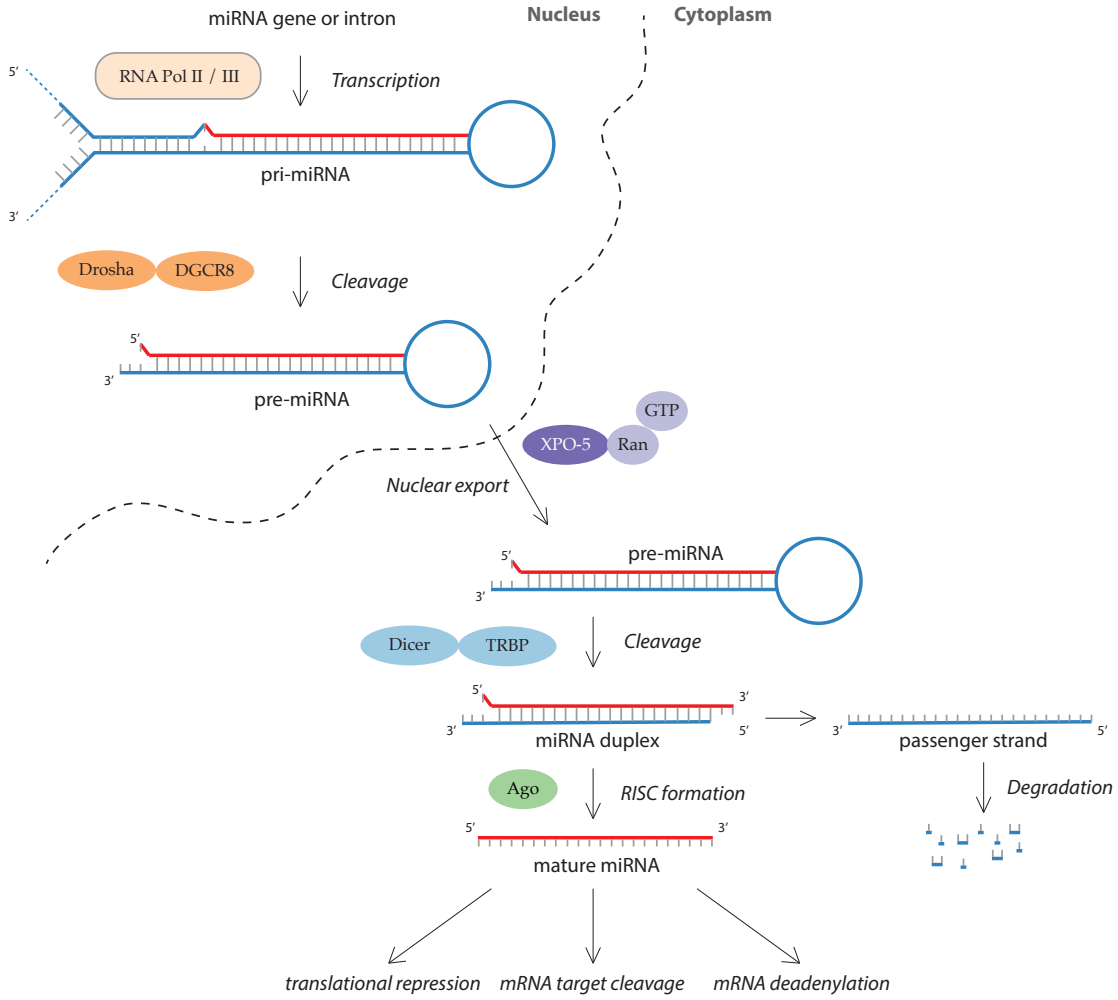



**Fig. 1.1** A schematic representation of the miRNA processing pathway in animals, leading to miRNA maturation. miRNA primary transcripts are transcribed by RNA Pol II into pri-miRNAs which incorporate a hairpin loop. The hairpin precursor is processed by the Drosha and DGCR8 proteins into pre-miRNAs and then exported from the nucleus to the cytoplasm by Exportin-5 (XPO5). Subsequently, the loop is processed (excised) by Dicer, followed by the separation of the mature and the passenger (star) strand. The mature miRNA then binds to the AGO proteins forming the RISC complex, which is able to induce mRNA target cleavage, de-adenylation and/or translational repression.

### 1.1.1 Biogenesis of miRNAs in animals

The biogenesis of miRNAs in animals (Figure 1.1) is now relatively well characterised (Krol et al., 2010). They are encoded by long non-coding transcripts or as passengers in the introns and UTRs of protein-coding transcripts. They are formed as 70-120nt stem-loop structures on the host molecule and are recognised and excised by enzymes including Drosha and DGCR8. The resulting cleaved hairpin molecule is referred to as a miRNA precursor and these pre-miRNAs are exported from the nucleus to the cytoplasm by Exportin 5 (XPO5) where they enter the RNA silencing machinery. The enzyme Dicer with cofactors excises a double-stranded duplex from the pre-miRNA which is unwound. In general, one of the strands is degraded (the passenger strand) and the other strand becomes a mature miRNA capable to load onto the RNA induced silencing complex (RISC) which can subsequently induce silencing of target transcripts.

### 1.1.2 Biogenesis of miRNAs in plants

miRNA biogenesis in plants differs from animal biogenesis mainly in the steps of nuclear processing and export (Ha and Kim, 2014). Plant primary miRNAs (pri-miRNAs) are mainly transcribed by RNA polymerase II and their length is highly variable (Axtell et al., 2011; Chang et al., 2012; Voinnet, 2009). However, homologues of Drosha and DGCR8 are not found in plants and unlike in animals, plant miRNA processing is completed in the nucleus. Plants are equipped with a Dicer-Like1 (DCL1) enzyme which processes most pri-miRNAs by sequential cleavage. The RNA-binding protein Dawdle (DDL) interacts with DCL1 and stabilizes pri-miRNAs in nuclear foci called dicing bodies (D-bodies). The double-stranded RNA-binding protein Hyponastic-Leaves1 (HYL1), DCL1, the zinc-finger protein Serrate (SE) and the nuclear cap-binding complex form a complex and process pri-miRNAs (Figure 1.2). Following processing, the miRNA-miRNA* duplex is 2′-O-methylated at the 3′ end by Hua-Enhancer1 (HEN1), which blocks uridylation and decay of miRNAs. Then, pre-miRNAs or mature miRNAs are exported to the cytoplasm by Hasty (HST), the plant homologue of exportin 5 (XPO5). In the cytoplasm, miRNAs are loaded onto cytoplasmic Argonaute (AGO) proteins, with AGO1 playing the most important role in the miRNA pathway.

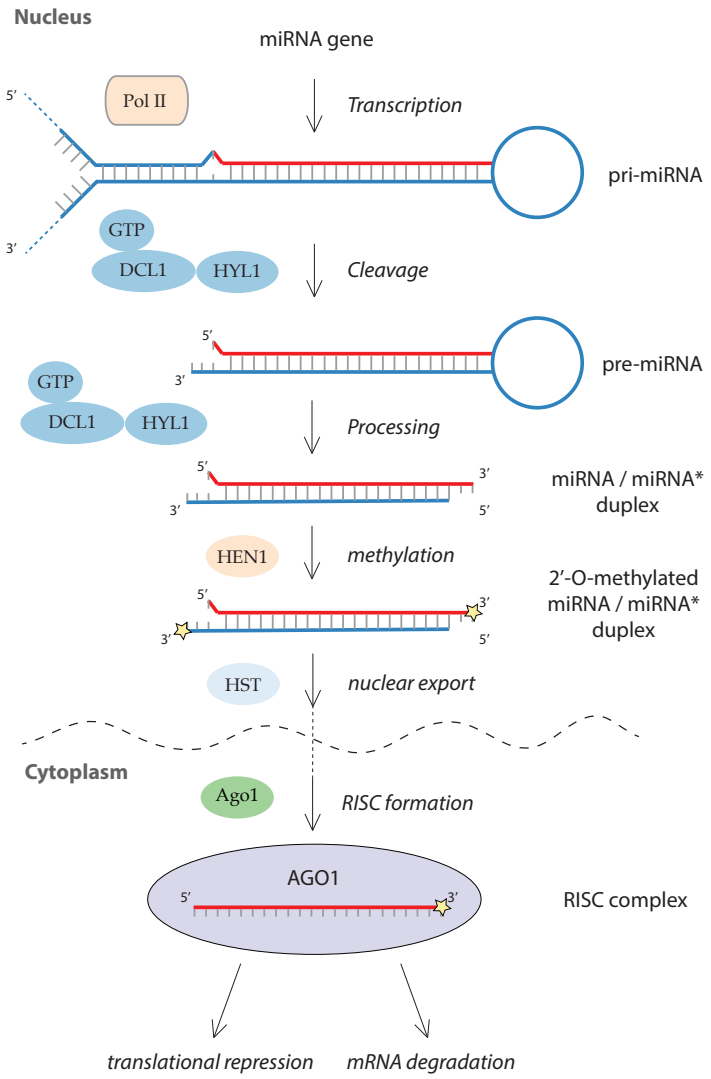



**Fig. 1.2** Schematic representation of miRNA biogenesis pathway in plants. Plant miRNAs are processed by a Dicer homologue, called Dicer-Like1 (DCL1) which is expressed only in the nucleus of plant cells. miRNA cleavage takes place exclusively inside the nucleus. Then, the 3' overhangs of the miRNA-miRNA* duplex are 2'-O-methylated by an RNA methyl-transferase protein called Hua-Enhancer1 (HEN1) and subsequently exported to the cytoplasm by a protein called Hasty (HST). There, they disassemble and the mature miRNA is loaded onto cytoplasmic Argonaute proteins (Ago1), able to target mRNA transcripts.

### 1.1.3 miRNA function

miRNAs are essential for most cellular functions in animals and plants (Ambros, 2004; Ameres and Zamore, 2013; Chen, 2005; Wienholds and Plasterk, 2005), while their dysregulation has been associated with multiple diseases (Ardekani and Naeini, 2010; Korpal et al., 2011; Mendell and Olson, 2012; Mraz and Pospisilova, 2012). We are going to present here

the main functionality of miRNAs in normal cells as well as their implication in diseases, such as cancer and heart disease.

The main function of miRNAs is directly associated with gene regulation. miRNAs are capable of binding to complementary parts of one or more messenger RNAs (mRNAs) and leading them to degradation or suppressing translation. In animals, miRNAs are almost always complementary to a site in the 3' UTR while in plants they are usually complementary to coding regions of mRNAs (Wang et al., 2004). In plants, miRNAs bind to mRNA target sites via perfect or near perfect complementarity, promoting cleavage of the RNA. In animals, miRNAs recognise their targets with partial complementarity. More specifically, the part of the miRNA that needs to be perfectly complementary to the mRNA target region is called the *seed region* (usually nucleotides 2 to 8). The mode of action of miRNAs is mainly inhibiting translation of the target mRNA into protein (Williams, 2008). Additionally, they may speed up de-adenylation of mRNAs that contributes to their degradation (Eulalio et al., 2009).

The way translational repression is accomplished is not fully understood yet. There has been a lot of debate around whether mRNA degradation, translational inhibition or a combination of both are responsible for suppressing translation of mRNA transcripts. Some studies in *Zebrafish* (Bazzini et al., 2012) and *D. melanogaster* (Djuranovic et al., 2012) have shown though that translational repression is predominantly caused by the disruption of translation initiation and is not related to mRNA de-adenylation.

Furthermore, miRNA functionality can be distinguished based on different kinetic signatures of the mechanisms they are involved in (Morozova et al., 2012). Specifically, there are nine mechanisms of miRNA action based on their kinetics (Morozova et al., 2012):

- *60S Ribosomal unit joining inhibition*
- *Cap-40S initiation inhibition*
- *Co-translational nascent protein degradation*
- *Elongation inhibition*
- *mRNA cleavage*
- *mRNA decay (destabilisation)*
- *Ribosome drop-off (premature termination)*
- *Sequestration in P-bodies*

- *Transcriptional inhibition through microRNA-mediated chromatin reorganization followed by gene silencing.*

Animal miRNAs may target diverse genes (Enright et al., 2003; Lewis et al., 2003; Stark et al., 2003). However, genes involved in fundamental functions of cells, such as gene expression, seem to be under selection since they have relatively fewer miRNA target sites in their sequence (Stark et al., 2005). In addition to the aforementioned modes of action, miRNAs have also been implicated in developmental processes, such as embryogenesis, differentiation, organogenesis and growth control (Alvarez-Garcia and Miska, 2005). Moreover, their dysregulation is associated with multiple diseases.

### 1.1.4 miRNA activity in human disease

There have been efforts in the past to manually catalogue all known relationships between miRNA malfunction and human disease (Jiang et al., 2008). Among those, miRNAs are believed to function in tumour suppression or as oncogenes, although their role in this derangement mechanism in cells has yet to be determined.

Early studies identified that several miRNA genes are located in cancer-associated genomic regions and their relationship with oncogenesis is very variable, depending on the type of cancer (Blenkiron and Miska, 2007; Croce and Calin, 2005; Kumar et al., 2007; Lagana et al., 2010; Lamy et al., 2006; Shah and Calin, 2014). Other studies showed that the miRNA expression profile is disrupted in cancer (Lan et al., 2015) while miRNA overexpression due to the disruption may lead to the development of tumours (He et al., 2015). Moreover, single-nucleotide polymorphisms (SNPs) in miRNA genes and/or their targets may be either enhancing suppression of tumour development or inducing oncogenic activity (Mishra et al., 2008).

Several types of epigenetic alterations have long been associated with many types of cancer (Camps et al., 2008; Gee et al., 2010; Iacobuzio-Donahue, 2009; Jansson and Lund, 2012; Korpal et al., 2011). Hypermethylation has been observed to induce silencing of miRNA genes in breast cancer and colorectal cancer (Lehmann et al., 2008; Toyota et al., 2008) while DNA hypomethylation in ovarian cancer enhanced the expression of oncogenic miRNAs (Iorio et al., 2007). Furthermore, it has been shown that miRNA silencing induced by methylation of miRNA promoter regions is directly related with breast cancer development and metastasis (Wee et al., 2012). Apart from DNA methylation, other types of epigenetic modifications such as histone acetylation have been associated with the reduction of antioncogenic miRNA expression in breast cancer cells (Scott et al., 2006).

More recent studies on various types of cancer including oesophageal squamous cell cancer (Song et al., 2014), cervical cancer (Juan et al., 2014) and urothelial bladder cancer (Network et al., 2014) have revealed notably different miRNA expression profiles in normal vs. tumour samples from the same cell or tissue. These studies have actually defined specific miRNAs as potential biomarkers for cancer diagnosis. This led to the development of numerous preclinical models, mainly in mouse xenografts and primates (Krützfeldt et al., 2005; Lanford et al., 2010), aiming to study the inhibition of oncogenic miRNAs (oncomiRs) that are over-expressed in various cancer types. These approaches rely on using chemically modified antisense miRNAs in order to target oncomiRs and silence their activity. However, there are still many obstacles in this approach related to overcoming the cellular barriers and ensuring targeted delivery of the therapeutic agent.

miRNAs have also been identified to play an essential role during the development of the heart (Chen et al., 2008; Zhao et al., 2007). Changes in expression levels of specific miRNAs in diseased human hearts imply their involvement with cardiomyopathies (Tatsuguchi et al., 2007; Thum et al., 2007; Van Rooij et al., 2006). Additionally, several specific miRNAs have been identified to play a fundamental role in regulating key factors of cardiogenesis, cardiac conductance and the hypertrophic growth response (Care et al., 2007; van Rooij et al., 2007; Xiao et al., 2007; Yang et al., 2007; Zhao et al., 2005).

Furthermore, miRNAs seem to be involved with the development and function of the nervous system (Maes et al., 2009). More particularly, the activity of neural miRNAs (such as miR-124, miR-132 and miR-134) has been associated with various stages of synaptic development, including dendritogenesis, synapse formation (Amin et al., 2015) and synapse maturation (Schratt, 2009). Additionally, some studies have identified altered miRNA expression in psychiatric disorders such as schizophrenia and bipolar disorder as well as major depression and anxiety disorders (Beveridge et al., 2010; Feng et al., 2009; Hommers et al., 2015). Another intriguing example is the involvement of miR-96 in progressive hearing loss in mice (Lewis et al., 2016, 2009). A single base change in the seed region of miR-96 leads to a drastic alteration of the mRNA repertoire targeted by it thus leading to either upregulation or downregulation of target genes. This perturbation induces progressive hair cell degeneration and eventually hearing loss (Lewis et al., 2009).

Finally, among the numerous areas of function miRNAs are involved in, it has been discovered that cellular miRNAs may play a role in the mammalian virus-host interactions by limiting virus replication (Lecellier et al., 2005). On the other hand, viruses such as the Epstein-Barr virus (EBV) can infect cells and generate their own viral miRNAs (Pfeffer et al., 2005). These may be used by the virus to manipulate both cellular and viral

gene expression, as a defence mechanism to the attacks of the host's miRNAs and/or as an enhancement factor of their replication potential (Skalsky and Cullen, 2010).

Taking into account all the above, it is evident without a doubt that miRNAs have an ubiquitous activity not only in normal cell function but also in human related diseases. That makes the effort for complete elucidation of their function (or malfunction respectively) even more imperative.

### 1.1.5 miRNA modifications

The expression of miRNA genes undergoes several stages until complete maturation. A large part of this process is defined by the genome-encoded sequence. However, research in recent years has revealed that the addition of non-templated sequences to the 3' ends of miRNAs plays a significant role in multiple functions including miRNA stability and function (Song et al., 2015). We are going to present here some of the most frequent modifications found in animal and plant miRNAs along with the enzymes they are mostly associated with. This section will serve as a reference for the miRNA modification analyses that will be presented later (Chapters 3 & 4).

#### GLD-2 (implicated in adenylation)

Germ Line Development 2 or GLD-2 is a cytoplasmic poly-A polymerase which adds successive adenosine monoposphate (AMP) monomers to the 3' end of specific RNAs, forming a poly-A tail. This process is known as poly-adenylation.

#### TUTases (implicated in uridylation)

Transferases are a class of enzymes that induce the transfer of specific functional groups, such as methyl or ketone groups, from one molecule to another. Terminal uridylyltransferases or TUTases, as their common name is, are associated with 3' RNA uridylation in plants, animals and fungi. Their function has been associated with several processes such as biogenesis of miRNAs, regulation of gene expression and cell proliferation (Hagan et al., 2009; Heo et al., 2012; Lim et al., 2014).

#### Adenylation

Adenylation is one of the most prevalent miRNA modifications. It has been found both in plant and animal miRNAs and its presence is implicated in functions like RNA stability or degradation. For instance, a study in *Populus trichocarpa* (Lu et al., 2009) has shown that a

significant portion of plant miRNAs are adenylated and consequently get degraded slower than other non-adenylated miRNAs. In this case adenylation contributes to miRNA stabilisation thus altering the relative proportions of miRNAs and target transcripts with implications in overall gene expression. Similarly, a study in human hepatocytes and mouse livers has shown that 3' terminal adenylation of miR-122, which is highly abundant in this tissue, by GLD-2 leads to stabilisation of this miRNA in liver (Figure 1.3).

On the other hand, it has been shown that 3'-adenylation of miR-21, a miRNA well known for its crucial role in cancer and other diseases, by the non-canonical poly(A) polymerase PAPD5 leads to its degradation (Boele et al., 2014). In fact, the pathway of miR-21 degradation via adenylation is a general feature of tumours across a wide range of tissues as well as of other proliferative diseases like psoriasis (Boele et al., 2014). Finally, another published work suggests that 3' adenylation of animal miRNAs may be modulating miRNA targeting effectiveness, potentially through interference during loading of the modified miRNA onto the RNA-induced silencing complex (RISC) (Burroughs et al., 2010).

### Uridylation

Uridylation of miRNAs is ubiquitous and conserved across many species including *Droshopila Melanogaster* (Berezikov et al., 2011), vertebrates (Burroughs et al., 2010; Kim et al., 2010) and plants (Yu et al., 2005). 3′ uridylation of mature small RNAs was first found in plants (Yang et al., 2006; Yu et al., 2005). Arabidopsis miRNAs are 2'-O-methylated on the 3′ end by the methyltransferase HUA ENHANCER 1 or *HEN1* (Figure 1.3). The *HEN1* mutants then undergo 3' truncation and uridylation that negatively affects their abundance. Since then, the addition of 3' non-templated uridine residues in small RNAs like miRNAs and piRNAs has been demonstrated in many more species like C. elegans (Billi et al., 2012; Kamminga et al., 2012), zebrafish (Kamminga et al., 2010), Chlamydomonas reinhardtii (Ibrahim et al., 2010), mouse (Jones et al., 2012; Kirino and Mourelatos, 2007a) and human cells (Thornton et al., 2014).

Uridylation affects not only mature miRNAs but also their precursors. The let-7 precursors were the first ever discovered precursors to undergo uridylation (Heo et al., 2008). In human embryonic stem cells, TUT4, which is a tutase, in conjunction with the RNA-binding protein Lin28 induce uridylation on the pre-let-7 transcript. The generated tailing deters Dicer from processing the pre-miRNA duplex and thus may facilitate its decay (Heo et al., 2008, 2009). Thanks to high throughput sequencing, it has been revealed that 3' uridylation is not an exclusive property of the pre-let-7 miRNAs but also expands to several other pre-miRNAs (Heo et al., 2009; Kim et al., 2015; Liu et al., 2014b; Newman et al.,

2011), with its function impacting either pre-miRNA degradation or enhancing the efficient processing of miRNAs.

In differentiated cells, uridylation of let-7 is a key factor for the biogenesis of a certain class of miRNAs. Specifically, pre-miRNAs belonging to the Group II of miRNAs (Heo et al., 2012) acquire a shorter (1nt) 3' overhang after Drosha processing instead of a 2nt overhang which is encountered in the most prevalent class of miRNAs (Group I). Thus, these pre-miRNAs require an extra nucleotide prior to Dicer processing. This is accomplished via TUTases, TUT4, TU7 and TUT2 which act redundantly and make the pre-miRNA susceptible to Dicer processing via mono-uridylation.

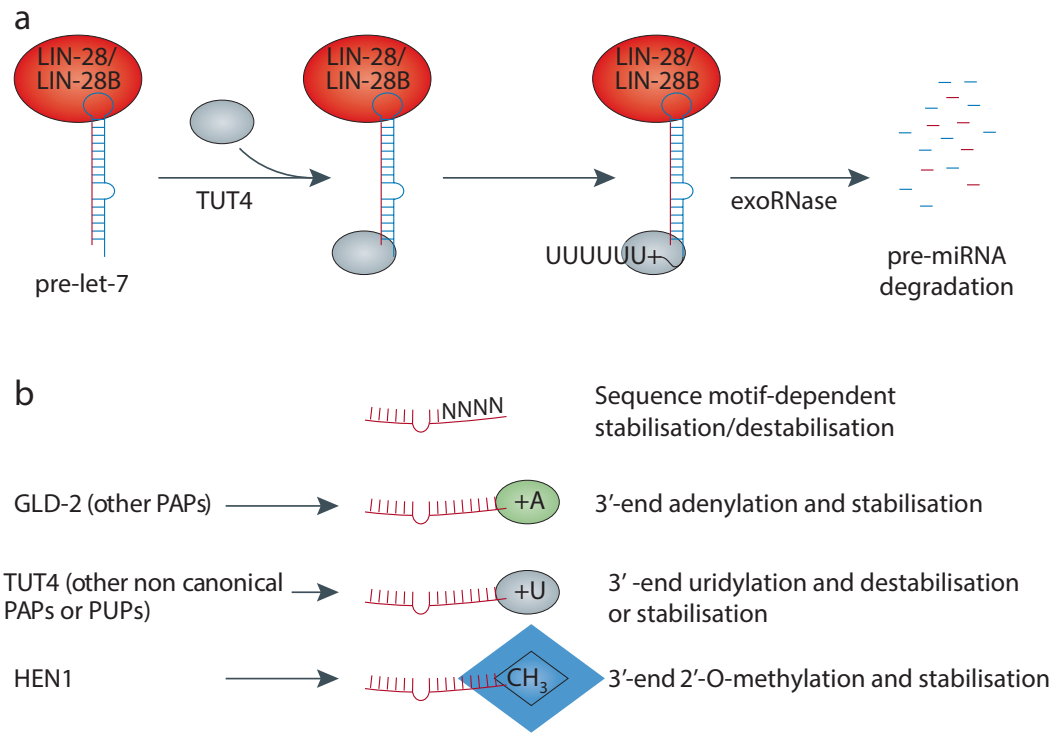



**Fig. 1.3** 3' terminal miRNA modifications act as stability regulators: the post-transcriptional addition of non-genome-encoded nucleotides to the 3′ end of pre-miRNAs or mature miRNAs affects their stability or abundance. a) The RNA-binding protein LIN-28 promotes uridylation of pre-let-7 in *C. elegans* and mammalian cells by recruiting TUT4, which adds multiple uracil resides to the 3′ end of RNA molecules. Poly-uridylation of pre-let-7 prevents Dicer processing and induces precursor degradation. b) 3' terminal modifications (in particular adenylation and uridylation) are affecting RNA stability: miRNAs are marked either for degradation or are protected against exonucleolytic activity. This largely depends on the specific miRNAs and the tissue. For instance, in liver cells, a single adenine added to the 3′ end of miR-122 prevents trimming and protects the miRNA against degradation (Katoh et al., 2009). miRNA methylation at the 3′ end by a methyltransferase (HEN1) prevents uridylation and degradation in plants (Li et al., 2005). In *D. melanogaster*, miRNAs that are loaded onto Argonaute 2 instead of Argonaute 1 are methylated at the 3′ end (Czech et al., 2009; Ghildiyal et al., 2010; Okamura et al., 2009), a modification that is likely to increase their stability. Figure adapted from: (Krol et al., 2010).

#### ADAR edits

Adenosine deaminases acting on RNA (ADAR) are enzymes that can bind to double stranded RNA (dsRNA) and convert adenosine (A) to inosine (I) by deamination (Samuel, 2011). ADAR protein is a RNA-binding protein, which functions in RNA-editing via post-transcriptional modification of mRNA transcripts. The conversion from $A$ to $I$ in the RNA disrupts the normal $A$:$U$ pairing which destabilises the RNA. Inosine is structurally similar to that of guanine ($G$) thus $I$ is binding to cytosine ($C$). In RNA, $I$ functions the same as $G$ in both translation and replication. Most editing sites are found in non-coding regions of RNA such as untranslated regions (UTRs), *Alu* elements and long interspersed nuclear element (LINEs), a class of transposable elements.

### 1.1.6 Identification of miRNA targets

miRNAs have been recognised as key factors in gene regulation for various biological pathways. They way this is accomplished is via targeting mRNA transcripts. In animals, the mature miRNA guides the RNA induced silencing complex (RISC) to the target site and binds to it. The binding site is located in most of the cases at the 3' UTR of the target transcripts (Figure 1.4) and binding specificity is defined by sequence complementarity of the seed region (nucleotides 2 to 8) of the miRNA with the target site (Lewis et al., 2005). In other cases, where there is imperfect complementarity of the seed region with the target, further complementarity of a region close to the 3' end of miRNAs may compensate for successful binding (Bartel and Chen, 2004). While the majority of miRNA-target binding occurs at the 3' UTR of transcripts, there have been studies revealing target sites located within the exons of protein coding genes (Lewis et al., 2005; Tay et al., 2008) and even in 5' UTRs (Lee et al., 2009b). However, these cases are not so common and it is also thought that the effect of miRNA driven regulation in these occasions is moderated due to ribosomes competing with the RISC complex on the same regions. (Bartel, 2009).

The binding of miRNAs to their target messenger RNAs (mRNAs) negatively affects their translation or causes degradation of the mRNA transcript. Specifically, the translation of the target transcript may be inhibited by completely blocking ribosome assembly and the initiation process itself or by promoting ribosomal drop-off and degradation of the nascent peptide. Additionally, the target mRNA can be de-adenylated and de-capped thus leading to degradation (Fabian et al., 2010; Giraldez et al., 2006).

Several computational methods have been developed to predict miRNA target candidates. The majority of these methods are searching for sequence complementarity between mature miRNAs and the 3' UTRs of mRNA transcripts. The set of features used by each

tool to classify candidate target sites is very similar, although each method may follow a different weighing scheme for each factor. The most important features taken into account are complementarity between the seed region of the miRNA and the mRNA target site and free energy of the miRNA-target duplex, potentially also taking into account regions surrounding the target site. Finally, some methods may explore for target site conservation in order to improve accuracy.

Two of the most popular algorithms, which were also among the first to be introduced in the field of miRNA target prediction, are miRanda (Enright et al., 2003) and TargetScan (Agarwal et al., 2015; Chiang et al., 2010; Friedman et al., 2009; Fromm et al., 2015; Garcia et al., 2011; Grimson et al., 2007; Lewis et al., 2005).

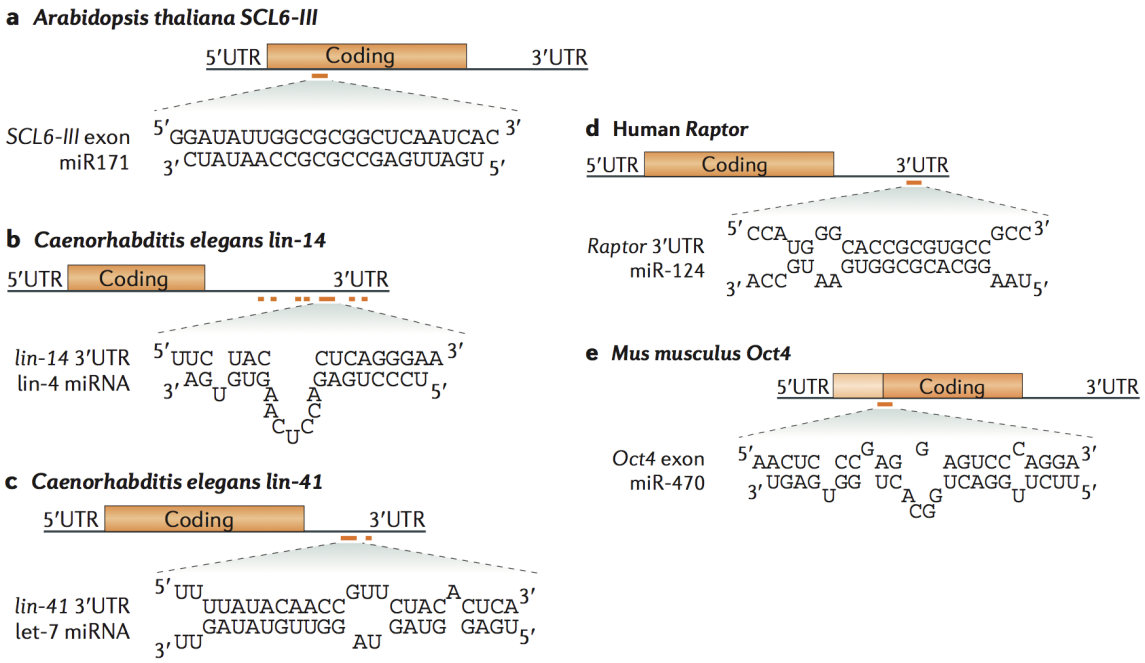



**Fig. 1.4** Examples of functional miRNA target sites in animals and plant based on different degrees of base pairing along the miRNA-target duplex: a) In plants, miRNAs usually pair nearly perfectly with target sites. For example, in *A. thaliana*, miR171 regulates SCL6-III mRNA through a fully complementary site in its coding region (Llave et al., 2002). b) In animals, partial pairing between miRNAs and their target sites is the most common case. Perfect pairing between the seed region (nucleotides 2–8) of the miRNA and a target in the 3' untranslated region (UTR) is the most frequent motif (Beitzinger et al., 2007; Wightman et al., 1993). Additional bars under the targeted transcript indicate other target sites, which may not necessarily exhibit perfect seed complementarity c) In some cases, extensive pairing of the 3′ end of the miRNA to the target sequence may compensate for the absence of perfect seed pairing (Reinhart et al., 2000; Slack et al., 2000). d) In other cases, where there is not essentially any seed pairing, sequences from the main body of the miRNA and closer to the 3' end may regulate binding to target sites via base complementarity (Shin et al., 2010). e) Finally, there have been found cases where the target site overlaps an exon junction thus inducing regulation through the coding region (Tay et al., 2008). Figure adapted from: (Pasquinelli, 2012).

## 1.2 Sequencing technologies & miRNA analysis tools

Sequencing refers to the process of determining the exact order of nucleotides within a DNA or RNA molecule. All sequence-based technologies are taking advantage of the properties of nucleic acids as efficient information carriers (Church et al., 2012). More specifically, DNA and RNA molecules can transfer information from one strand to the complementary one via the rules of base pairing (Adenine with Thymine or Uracil and Guanine with Cytosine). In addition, double strands of DNA are able to separate and re-hybridise under high-low temperature cycles or enzymatic treatment and this enables efficient amplification of information contained within a single DNA molecule. These properties combined with various chemically modified nucleotides and protein catalysts still remain the basis for the majority of methods for sequence identification and quantification (Morozova and Marra, 2008).

One of the first methods in sequencing was the Sanger sequencing, which was based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication (Sanger and Coulson, 1975; Sanger et al., 1977). This method was developed in 1977 by Frederick Sanger and his colleagues and was established as the state-of-the-art method in sequencing for around 30 years. The advent of Next-Generation Sequencing methods though in the early 2000s, progressively replaced previous methods due to its high accuracy, speed and cost-effectiveness, thus enabling a rapid acceleration in biological and medical research.

### 1.2.1 Next-Generation Sequencing (NGS)

Next-generation sequencing (NGS), also known as high-throughput sequencing, is a term describing a number of different modern sequencing technologies including:

- Illumina (former Solexa) sequencing
- Ion torrent: Proton / PGM sequencing
- SOLiD sequencing
- Roche 454 sequencing

These technologies allow us to sequence DNA and RNA molecules much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionised the study of genomics and molecular biology. We will present briefly in the next section the methodology behind Illumina Sequencing, which is currently the most widely used and is also the primary sequencing technique used for the datasets analysed in this thesis.

### 1.2.2 Illumina Sequencing

In NGS, a huge amount of short reads can be sequenced at once via massive parallel sequencing. This can be achieved by cleaving, first of all, the input sample into short sections. The length of these sections varies and primarily depends on the sequencing machine used. In Illumina sequencing in particular, 100-150bp reads are prepared.

The general methodology of Illumina sequencing follows three main steps: amplification, sequencing and analysis (Figure 1.5). First, DNA is purified and chopped into fragments. These fragments are ligated with adapters, indices and/or other types of molecular modifications so that they can be uniquely identifiable through the different stages of the procedure. Then, DNA molecules are loaded onto a specialised chip and a phase called 'cluster generation' begins. Each chip is equipped with thousands of oligos (short synthetic DNA sequences) that are able to attach to free DNA fragments via complementarity. This step produces eventually around a thousand copies for each DNA fragment. At that stage, the chip is provided with primers and modified nucleotides which force the primers to add only one nucleotide at a time along with fluorescent tags. At each round a new nucleotide from the original DNA fragments is synthesised and a picture of the chip is taken with a camera. This image is analysed computationally and a base is detected based on the wavelength of the fluorescent tag. After each round, the 3' terminal blocking groups and the fluorescent substance are removed so that they don't contaminate the next round of synthesis. This process continues until the whole DNA molecule is sequenced.

NGS has four main advantages over classical Sanger sequencing:

- accuracy
- speed
- cost
- sample size

Sequence duplicates are intrinsic to NGS methods, as each read is amplified before sequencing. Since NGS is so much quicker and cheaper, it is possible to retrieve more duplicates via Polymerase chain reaction (PCR) than with Sanger sequencing. This allows to obtain greater coverage, which means higher accuracy and reliability of sequencing, even if individual reads were less accurate for NGS. On the other hand, Sanger sequencing has the advantage of being able to produce much longer sequence reads. However, the parallel nature of NGS allows to compensate for this 'defect' by constructing longer reads from many contiguous short reads.

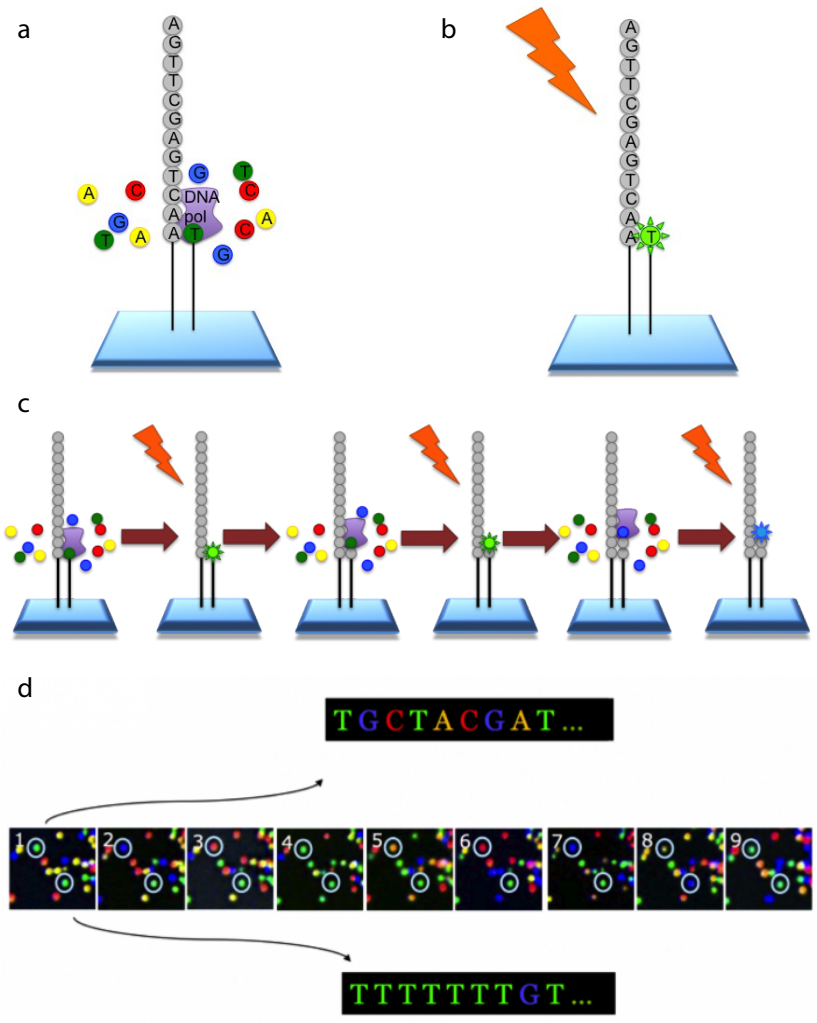



**Fig. 1.5** Illustration of Illumina Sequencing methodology. Input reads of 100-150bp are prepared and amplified with Polymerase Chain Reaction (PCR), thus creating many copies of the same read. They are then split into single strands to be sequenced, based on the following procedure: a) a flow cell is flooded with nucleotides and DNA polymerase. These nucleotides are fluorescently tagged, with a distinct colour corresponding to each base. They also have a terminating sequence to make sure that only one base is added at a time, b) an image is taken of the flow cell and a fluorescent signal at each location detects the base that has been added. Both the terminating sequences and the fluorescent substance are removed allowing the next base to be added and preventing contamination of the image of the next base from the current fluorescent signal, c) the same process is repeated for every incoming nucleotide, one at a time, and taking images of the fluorescent signal, d) as soon as imaging is complete for all bases, each image is analysed at each site through computational image analysis allowing the detection of the sequenced bases and eventually leading to the re-construction of sequences of the same length. Illustration adapted from: https://www.ebi.ac.uk/training.

### Detection of miRNAs using NGS

Large-scale cloning and sequencing of small RNAs using capillary sequencing allowed the initial detection of large sets of animal miRNAs (Landgraf et al., 2007). However, the advent of next-generation sequencing (NGS) allowed these molecules to be rapidly de-

tected in different tissues and organisms. The primary repository for miRNAs is the miRBase database (Griffiths-Jones et al., 2008). Initially, miRNAs were usually confirmed by northern blot or similar assay prior to their inclusion in miRBase. However, the advent of large-scale NGS studies has meant that it is impractical to confirm every single sequence detected via targeted amplification. Given that the genome is replete with putative stem-loop structures and that small RNA sequencing detects many short molecules and degradation products, there are many putative miRNA sequences in miRBase which may in fact not be canonical miRNAs but instead may be other functional ncRNAs or the degradation products of longer molecules. We are going to address this issue later in Chapter 3, by suggesting lists of mis-annotated miRNA sequences in miRBase based on their coverage profiles obtained from Next Generation Sequencing data.

### 1.2.3 Small RNA-Seq analysis tools

The vast amount of data generated by NGS techniques requires novel and efficient methods for their analysis. There have been several tools published in previous studies for performing small RNA-Seq analysis, each providing a different set of features. Some of these tools are web based while others offer stand-alone versions, requiring though many dependencies in some cases. The features they provide vary from mere miRNA quantification to Gene Ontology/pathway analysis and identification of terminal or internal modifications. Later in Chapter 2, we are going to present Chimira, our novel method for identification of miRNA modifications (5'-, 3'-terminal, ADAR edits and Single Nucleotide Polymorphisms or SNPs). Thus, we will present here, as a reference, several representative small RNA-Seq analysis tools that are either offering a diversified set of features, have integrated some kind of functionality for modifications identification or have been developed as web-server applications.

- **CAP-miRSeq** (Sun et al., 2014)

  Supported genomes: *need to be installed manually by the end-user.*

  Although it requires the installation of a virtual machine software package and the import of the developed Linux virtual environment, the whole setup is easy and straightforward. However, applying any of the provided tools involves an extra overhead of downloading the genome and/or annotation files and creating manually the configuration files. Besides, with regards to the modifications identification, CAP-miRSeq only allows the detection of single nucleotide variants (SNV) and does not support 3'/5' modifications or ADAR edits detection.

- **CPSS** (Zhang et al., 2012)

  Supported genomes: *H. sapiens, M. musculus, R. norvegicus, P. troglodytes, G. gallus, B. taurus, C. lupus familiaris, P. abelii, S. scrofa, D. rerio.*

  The CPSS web server allows miRNA isoforms detection, supporting 3', 5' modifications and SNPs. However, the extracted results do not directly provide the modifications information in a fully quantitative format. Specifically, all detected isoforms are displayed just as an aligned stack of sequences and a summary table is provided only for the total number of modifications in either of the 3' or 5' ends with no extra information about their content or exact positions. Besides, no ADAR edits detection is supported. Apart from that, CPSS's interface allows upload of only one file at a time. File upload is very slow (e.g. uploading 50MB requires over 40min on average) and even though a script is provided for pre-trimming/cleaning of the input files the post-processed files' size is not significantly reduced. Thus, the web server is not practically usable even for small sample files.

- **MAGI** (Kim et al., 2014)

  Supported genomes: *H. sapiens.*

  MAGI web server allows alignment against only the human genome. Although it uses web workers for downsizing the input files before upload, total upload time is not decreased significantly compared to other methods (see Chapter 2 - Time Benchmarking). In addition, input files need to follow a specific naming scheme (with group annotation) and have to be de-compressed before upload, which is very impractical for large datasets in terms of local disk space requirements. In this case again, no modifications information is extracted and the user cannot query the results interactively.

- **OASIS** (Capece et al., 2015)

  Supported genomes: *H. sapiens, M. musculus, D. melanogaster, D. Rerio, C. elegans.*

  OASIS does not offer any tools for modifications identification. The tools provided by this platform are not integrated in a very coherent manner since for a single dataset a new job has to be launched in order to perform either sRNA detection, differential expression or classification analysis. Moreover, no information is provided about the progress of each task apart from an e-mail upon initiation or completion of a job. Output results cannot be visualized in a queryable manner and only five genomes are supported.

- **seqBuster** (Pantano et al., 2009)

  Supported genomes: *H. sapiens*.

  seqBuster is a stand-alone tool that is able to infer miRNA modifications and it also offers basic 3' adapter removal. However, it is no longer supported in its original form as a web-server. Its current stand-alone version requires a lot of dependencies and installation is not sufficiently documented. We were not able to test this tool due to lack of straightforward documentation.

- **UEA sRNA workbench** (Stocks et al., 2012)

  Supported genomes: *all miRBase annotated species*.

  UEA sRNA workbench is a suite of tools for small RNA-Seq data analysis in animals and plants. It provides tools for quality checking, normalisation and differential expression of small RNA-Seq samples. Additionally, it can predict miRNAs from high-throughput sequencing data as well as miRNA targets. Based on our experience while testing this method, there is a small overhead of installing the application, compared to any web-server application. Moreover, analysis is not always straightforward since it requires setting up a local working directory tree and manually installing the dependencies for each of the available tools. Finally, UEA sRNA workbench does not support identification of microRNA modifications. However, it certainly consists a robust platform with a rich set of features for the analysis of small RNA data.

### 1.2.4 Methods for novel miRNA prediction

Apart for mere identification of known miRNAs, the discovery and annotation of novel miRNAs has been a challenge for many years. Several tools have been developed in the past that perform novel miRNA prediction. Traditionally, these tools attempt to associate mature miRNAs with their hairpin precursors and define features on them based on their computationally predicted secondary folding. All these methods require a reference genome for mapping the small RNA sequences and extracting the precursor sequence of each miRNA and potentially extra flanking genomic sequence for the folding analysis. We also found three methods that do not require a reference genome but none of them is supported or functional (see also Chapter 4). Later in this thesis, we are going to introduce mirnovo, a novel method that we have developed and which is able to predict miRNAs with or without a reference genome from small RNA-Seq and single-cell data using machine learning methods. As a reference, we cite in Table 1.1 the main features of some of

the most popular tools that have been previously published in the field of novel miRNA prediction in comparison with the novel method (mirnovo) that we are going to introduce in the fourth Chapter of this thesis.

**Table 1.1** Comparison of novel miRNA prediction tools based on features availability. Cells with a *checkmark* indicate a supported feature while cells with a *dash* denote the lack of the feature. Web-server applications that are no longer available are denoted as 'offline'. Stand-alone tools that fail to be installed or throw an exception during runtime are flagged with the 'crashes' property.

| Method | Web-server | Stand-alone | Animals | Plants | Genomic features | Prediction without genome |
|---|---|---|---|---|---|---|
| miRDeep2 | - | ✓ | ✓ | - | ✓ | - |
| miRanalyzer | *offline* | ✓ | ✓ | - | ✓ | - |
| miRTRAP | *offline* | - | ✓ | - | ✓ | - |
| mirTools | ✓ | - | ✓ | ✓ | ✓ | - |
| miRDeep-P | *offline* | ✓ | - | ✓ | ✓ | - |
| MIReNA | *offline* | - | ✓ | ✓ | ✓ | - |
| miReader | - | *crashes* | ✓ | ✓ | - | ✓ |
| MirPlex | - | *crashes* | ✓ | ✓ | - | ✓ |
| ***mirnovo*** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

## 1.3 Machine learning

Around two centuries ago, western human civilisation and society started changing dramatically thanks to a revolution of machines, the so called *"Industrial Revolution"*. The emergence of steam engine allowed us to overcome the limitations of muscle power, both human and animal, and generate massive amounts of useful energy (Brynjolfsson and McAfee, 2014). Similarly, in our times, we may be witnessing the beginning of a second revolution of machines, only in that case the setting under which they are acting is cognitive rather than physical (Brynjolfsson and McAfee, 2014). This revolution is no other than the recent explosion of machines that are able to 'learn' and imitate human-like cognitive tasks, thanks to the rapid (re-)emergence and expansion of the field of machine learning.

The idea of machine learning, which is essentially making "computers able to learn without being explicitly programmed" was already introduced in 1959 by Arthur Samuel (Samuel, 1959). The field started off by focusing initially on pattern recognition problems but eventually evolved into entailing any algorithm that can learn from data and make predictions based on them. Common machine learning applications include spam e-mail filtering, optical character recognition (OCR) and object identification with computer vision. Advances in computational power and resources in recent years led eventually to an aggressive expansion of machine learning in even more applications, either scientific or commercial. In fact, 2016 was marked by the grand hype surrounding machine learning, especially via the emergence of deep learning.

The field of deep learning is actually based on a method already known since 1943, the artificial neural networks or ANNs (McCulloch and Pitts, 1943). ANNs have been inspired by biological neural networks and are essentially a collection of interconnected units with weighted activation signals, able to transmit information from one layer of nodes or 'neurons' to another. The novelty introduced by deep learning had to do with the introduction of multiple hidden layers in the learning network (Bengio, 2012; Hinton et al., 2006), thus its characterisation as '*deep learning*'.

Along with deep learning methods, machine learning entails several other algorithms that have been successfully tested and used over various applications. Such methods, among others, include Logistic Regression, Support Vector Machines, Gradient Boosting, Bayesian Networks and Random Forests. The applications of these algorithms have been interspersed across a multitude of fields, from finance and linguistics to speech recognition and game playing. Bioinformatics is also among the most prominent fields where machine learning is being applied. The vast amount of information aggregated in the last one or two decades with regards to biological sequences, processes and systems resembles a gold mine that seeks for competent researchers to find its hidden secrets and invaluable treasures. We are going to present shortly in one of the next sections some machine learning successes in the area of Bioinformatics. Moreover, in the fourth chapter we are going to introduce a novel method that is aiming to boost research in the field of novel miRNA prediction, with the aid of machine learning.

All this rapid expansion of machine learning applications demonstrates that an 'intellectual' revolution of machines may be taking place in our era, aiming towards the ultimate goal of artificial intelligence (Brynjolfsson and McAfee, 2014). Aside from the multiple ethical or philosophical questions that may arise due to fast evolution of machine learning, humanity could benefit a great deal from this evolution, subject to worldwide collaboration and sensible reasoning. What exactly is machine learning though? We will

go through some theoretical basics, methods and applications of machine learning, specifically in Bioinformatics, within the next few sections.

### 1.3.1 Approaches

Machine learning methods are classified into three main categories, based on the type of data that is available to the learning system. These are:

- *Supervised learning*: where available data are labelled, i.e they include both example inputs and known outputs aiming to find the optimal mapping from inputs to outputs,
- *Unsupervised learning*: where the aim is to find hidden patterns or structure from unlabelled data, and
- *Reinforcement learning*: where the computer program is constantly improving its performance via feedback received by a dynamic environment which interacts with it.

Alternatively, we can categorise machine learning methods based on the desired type of output the system is aiming for. Specifically, we can distinguish four main categories in that case:

- *Classification*: where input data points need to be divided into two or more classes (supervised learning),
- *Regression*: which estimates the relationships among variables and returns a continuous output rather than a discrete one (also supervised learning),
- *Clustering*: where inputs need to be divided into groups but in this case data is not labelled (thus unsupervised learning), and finally
- *Dimensionality reduction*: which tries to simplify input data by representing them with a reduced number of dimensions (e.g. PCA).

We employ unsupervised learning in the form of 'clustering' in various analyses within this work. However, we are going to focus mainly in supervised learning (Figure 1.6) since this is the predominant approach adopted in Chapter 4, yielding the high success rate in the predictions of our novel method.

With regards to supervised methods, a typical learning task involves four distinct stages:

1. Collect labelled data, i.e. a data set that you know the answer to each data point.
2. Train the machine learning algorithm on that data set (training set).
3. Collect data that you want to make predictions or inferences for (test set).
4. Make predictions on the test set using the algorithm which has been pre-trained on the training set.

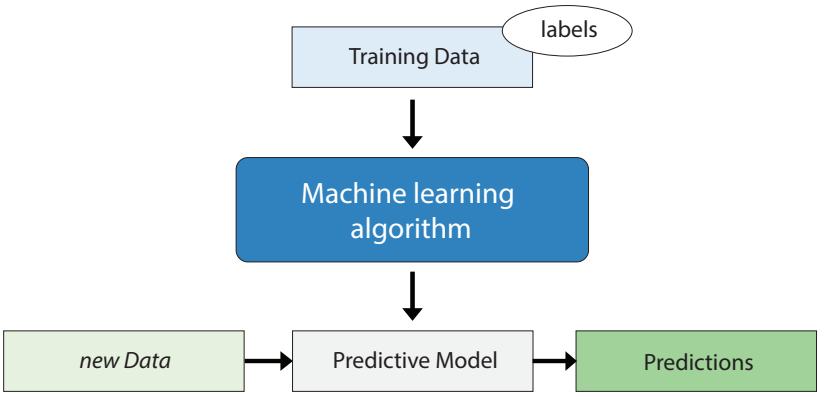



**Fig. 1.6** Simplistic representation of a supervised learning task. A set of labelled input data is required for the training of the algorithm. The learning algorithm analyses the training data and infers a prediction model that can be used for classifying new examples into one of the classes defined by the input labels. The performance of a learning algorithm is assessed by its ability to generalise from the training data and allow accurate prediction of the class for unseen instances.

The repertoire of available machine learning techniques is very rich and diverse. We are going to present briefly here some of the most popular approaches in machine learning.

### Support Vector Machines

Support Vector Machines (SVMs) are a class of supervised learning models that can be used for classification and regression analysis. The fundamental idea behind an SVM model is that input examples are represented as points in $n$-dimensional space ($n = 1, 2, 3, ...$) and a clear gap as wide as possible has to be found in order to divide optimally the examples of different classes represented in the input data. After the model has been built, new examples can be mapped into the same space and assigned to a class based on which side of the gap they fall into.

The SVM algorithm was originally invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. In 1992, it was Vladimir N. Vapnik again, along with Bernhard E. Boser, Isabelle M. Guyon, who suggested the design of non-linear classifiers by applying the kernel trick to maximum-margin hyperplanes (Boser et al., 1992), which implicitly

maps input data points into high-dimensional feature spaces. The current standard incarnation (soft margin) was proposed by Corinna Cortes and Vapnik in 1993 and published in 1995 (Cortes and Vapnik, 1995).

SVMs belong to a family of generalized linear classifiers and can be interpreted as an extension of the perceptron algorithm (Meyer et al., 2003). Data points are viewed as $p$-dimensional vectors and the SVM algorithm is searching for a $p-1$-dimensional hyperplane that represents the largest separation, or margin, between the two classes that need to be distinguished. In this case, SVM acts as a linear classifier on data points defined in a finite dimensional space.

However, in many cases the sets to discriminate are not linearly separable in a space with finite number of dimensions. It was proposed then that the original finite-dimensional space be mapped into a much higher- or infinite-dimensional space that could make the separation easier (Boser et al., 1992). This could be achieved by replacing the dot product of the linear SVM classifier by a non-linear kernel function. This allows the SVM algorithm to fit the maximum-margin hyperplane in a transformed feature space and classify non-linearly separable data points.

### Artificial Neural Networks

An artificial neural network (ANN) is a learning algorithm that is inspired by the structure of biological neural networks. An ANN is based on a collection of interconnected units called artificial neurons that can transmit signals to each other via their connecting edges (analogous to synapses in a biological brain). Neurons are organised into layers that may perform distinct kinds of transformations on their input signals. Typically, a neural network is comprised of three layers (Figure 1.7a): i) the input layer, which includes the nodes that receive the input signals, ii) the hidden layer, which applies some kind of transformation in the input signals and iii) the output layer, which extracts the output signals. However, neural networks are not restricted to only one hidden layer but they may include multiple layers that impose various transformations on the input signals. These networks belong to a sub-domain of ANNs called Deep Learning.

### Deep Learning

Advances in hardware and particularly the broad expansion of multi-core graphics processing units (GPUs) along with the exponential growth and availability of big data in recent years has ignited the emergence of a special field of machine learning, which is called deep learning. GPUs are well-suited for the matrix/vector multiplication calcula-

tions which are vastly employed by machine learning algorithms (Chellapilla et al., 2006; Oh and Jung, 2004). The speed-up in algorithm training using modern GPUs can be measured by orders of magnitude, reducing running times from weeks to days (Cireşan et al., 2010; Raina et al., 2009). At the same time, a large amount of data is constantly being generated by humans (e.g. in social media), sensors (e.g. aeroplanes and cars) and more specialised machines (e.g. X-ray devices) in a digital format that makes them an invaluable resource from various kinds of machine learning applications.

Deep learning specifically studies deep neural networks, which are essentially neural networks with more than one hidden layers (Figure 1.7b) and usually over ten layers or even orders of magnitude more than that. Deep neural networks scale efficiently by absorbing huge amounts of data and creating even more accurate models as training data increases, in contrast with all other machine learning methods whose performance plateaus after a certain input size (Ng, 2016). A fundamental role in the emergence of this field played mainly three eminent Computer Science Professors: Yann LeCun, Yoshua Benzio and Geoffrey Hinton (LeCun et al., 2015). Deep learning has already found exceptionally successful applications in fields like computer vision, image recognition and speech recognition (Lee et al., 2009a) while it is also constantly expanding into Computational Biology as well (Angermueller et al., 2016).

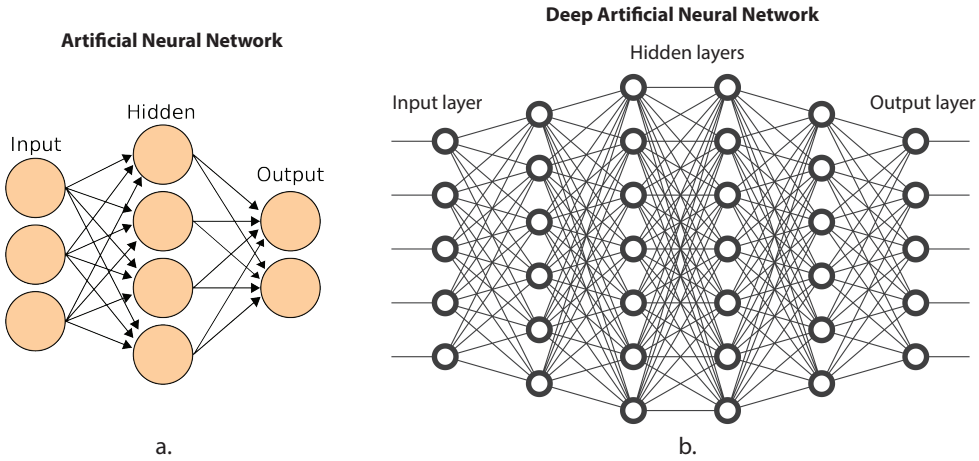



**Fig. 1.7** Graphical representation of neural networks. a) A classical neural network: there is one input layer for the incoming signals, one output layer for the outcoming signals and another one in between (the hidden layer) that is transforming the inputs into output signals. b) A deep learning network: instead of having just one hidden layer, a deep network is equipped with multiple hidden layers that impose a series of transformations to the input signals in order to extract the output features.

#### Decision Trees

Another popular machine learning method used for either classification or regression tasks is the decision tree. It is one of the most widely used and efficient methods for inductive inference, i.e. the process of deducing a general conclusion from specific examples. More specifically, learning with decision trees involves a top-down induction process, where input data start at the root of a tree and are recursively examined at each node of the tree until a leaf node is found.

Checks at each node of the tree are rule-based comparisons of the input data point features with values derived from the trained model. During training, the variable that ensures the ”best split” at each node is selected as the criterion for splitting the test data later on for this particular node. The metrics for measuring the ”best split” generally assess the homogeneity of the target variable within the subsets. Some of the most well-known metric for measuring the quality of a split are *gini impurity*, *information gain* and variance reduction. For instance, *gini impurity* (which we will employ in Chapter 4 for the training of a Random Forest classifier) is the degree of misclassification of input data at the child nodes of any given node.

Eventually, the output of the tree is an output class (for classification tasks) or a continuous number (for regression analyses). Using multiple decision trees during training leads to the construction of Random Forests, a popular machine learning algorithm that we are going to present in the next section.

### 1.3.2 Random Forests

Random forests are an ensemble learning method that employ multiple decision trees at training time and are used for classification and regression, among other tasks (Figure 1.8). The *randomness* in the name of the method refers to the fact that each tree is analysing a random subset of the entire dataset and that features at each node that are used for rule-based splits are also selected randomly. Additionally, the multitude of trees used during training form a *forest*, thus justifying the full name of the method.

Each of the trees is independently trained and at each node the feature variable that assures the largest split of data, over a smaller random set of all features, is assessed. In the end, each tree extracts an output class or continuous numerical value. The final result of the Random Forests is the mode of the classes for classification tasks (i.e. the most frequent output class among all decision trees) or the mean numerical value across all predictions of individual trees, for regression analyses. In this way, random forests resolve the issue of overfitting, which is commonly found in decision trees, since they are trained using the

average output of multiple training processes and as such can be more efficiently applied to independent datasets as well.

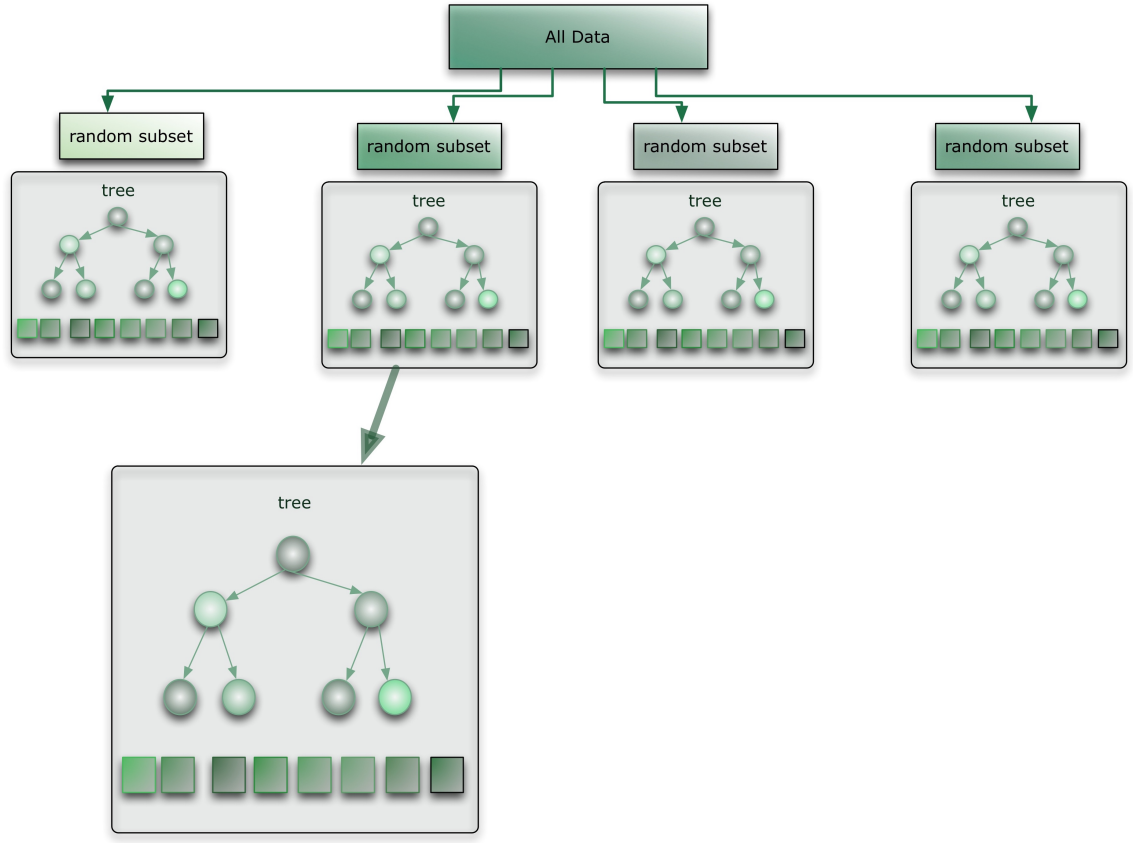



**Fig. 1.8** Illustration of a Random Forests learner. Random Forests are constructed by combining multiple individual tree learners, thus they are characterised as an *ensemble* method. Each tree is trained independently over a sub-sample of the entire dataset. At each node of a tree, a small subset of features is selected at random and the split of the next level is driven by the variable which optimises this split. In the end, outputs from all trees are weighted and averaged in order to build the overall Random Forests classifier. Figure adapted from: http://blog.citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics.

### 1.3.3 Applications in Bioinformatics

Machine learning has numerous applications emerging in the field of Bioinformatics. The vast amount of biological data accumulated in recent years require highly sophisticated methods to be analysed. Traditional methods that are based on explicit programming of a set of commands cannot be implemented in practice in order to decipher the information hidden in the huge pile of biological data. Instead, machine learning offers the possibility of automatic (or semi-automatic) learning from the data that enables researchers to advance rapidly in various sub-domains of Bioinformatics. Some of the most representative

areas in Bioinformatics that employ machine learning for data processing and analysis are: genomics, proteomics, systems biology and text mining (Inza et al., 2010).

### Genomics

Genomics is the field of molecular biology that focuses on the assembly and analysis of the function and structure of entire genomes. This is accomplished through the adoption of high throughput DNA sequencing techniques and bioinformatics methods, including machine learning. One of the first applications of machine learning in genomes was gene prediction, i.e. the determination of the location of protein-coding genes within a given DNA sequence (Mathé et al., 2002). Typically, gene prediction is performed by aligning an input DNA sequence to large databases of sequences whose genes have been previously discovered and annotated. However, in order to accelerate the biological interpretation of the perpetually and exponentially increasing raw data, machine learning methods need to be used. Features can be defined based on known genes in order to train a learning algorithm that can then predict novel genes from unseen sequences (Mathé et al., 2002). Additionally, machine learning has been used for the problem of multiple sequence alignment (alignment of multiple DNA or amino acid sequences) in order to trace shared evolutionary history among similar sequences (Larranaga et al., 2006).

### Proteomics

Proteomics refers to the large-scale study of proteins (long strings of amino acids) and their functions (Anderson and Anderson, 1998; Blackstock and Weir, 1999). The functionality of proteins is determined by protein folding in which they conform into a three-dimensional structure. Protein folding undergoes four stages of transition: primary structure (a flat string of amino acids), secondary structure (alpha helices and beta sheets), tertiary and quartenary structure. The main focus in this sub-field is the secondary structure due to its primary role to determining the subsequent tertiary and quartenary structures and eventually the overall functionality of the protein. Prior to machine learning, researchers needed to predict protein structure by manual analysis of amino acid sequences (Larranaga et al., 2006; Yang et al., 2016). The first work in this field came from Pauling and Corey when they predicted the hydrogen bond configurations of a protein from a polypeptide chain (Pauling et al., 1951). However, predicting the true structure of a protein manually is not only an extremely expensive and time-intensive process but also practically infeasible given the amount of data available. That made the adoption of machine learning techniques for automatic feature learning and prediction imperative. Today, the state-of-the-art models are

able to predict protein secondary structure with an accuracy of 82-84%, very close to the theoretical limit of 88-90% (Wang et al., 2016; Yang et al., 2016). These methods are based on Deep Convolutional Neural Fields (DeepCNF), a type of artificial neural network used for Deep Learning.

### Systems Biology

Systems Biology studies complex interactions of simpler biological components, such as DNA, RNA, proteins and metabolites, within a system in order to infer emergent behaviours. In this regard, machine learning has been applied in order to identify transcription factor binding sites using Markov chain optimisation (Larranaga et al., 2006). Moreover, probabilistic graphic models, a machine learning technique for inferring the relationship between different variables, has been widely used in complex interaction systems, such as metabolic pathways and signal transduction networks (Larranaga et al., 2006). Finally, genetic and regulatory networks have been extensively studied with the use of another class of machine learning methods, the genetic algorithms, which mimic biological evolution by applying some kind of natural selection process to the data.

### Text mining

The exponential rate of sequencing data harvesting and the subsequent increase of biological publications have made it practically impossible to track all the information generated in the last few years. Employing new methods to extract information from all accumulated publications and databases, a task known as knowledge extraction, has become imperative. Machine learning contributes to this direction mainly through Natural Language Processing techniques. The learning algorithm is fed with a set of input data, such as published manuscripts, aiming to generate new biological knowledge. For instance, drug development requires first of all exhaustive examination of information available in biological databases and scientific journals. However, there is rarely a unique resource integrating all available information for an entity, e.g. a protein. Text-mining makes it possible to parse large sets of scientific texts in order to complete the annotations extracted from databases for all required entities (such as sub-cellular localisation of a protein and large-scale protein interaction analysis), thus enabling development of advanced therapeutic methods.

## 1.4 Aims of the thesis

In this thesis we wanted to elucidate some unexplored parts of the powerful world of small non-coding RNAs and in particular of miRNAs and piRNAs. The first aim of this work was to build a novel computational method that is able to capture post-transcriptional modifications in miRNAs, either 5'-, 3'-terminal or internal (such as ADAR-edits). The development of such a tool was essential due to the lack of other tools providing the same type of information. Additionally, we have already seen the importance of some types of modifications for miRNA stabilisation and maturation (Heo et al., 2012; Katoh et al., 2009). Thus, we wanted to perform for the first time a large-scale analysis of small RNA datasets in order to extract the global landscape of miRNA modifications.

That consists the second main aim of the thesis, which is finding potential distinct modification patterns across specific conditions as well as their prevalence and features in different cell types and/or tissues. Moreover, by aggregating large amounts of data we wanted to explore other important features of miRNAs such as their co-expression, depending on their genomic location or cell type, regulation of miRNA clusters by common sets of transcription factors and also determine the specifics of the strand selection mechanism during miRNA maturation.

Considering the growth of high-throughput sequencing and the emergence of single-cell biology, discovery and annotation of novel miRNAs was another intriguing challenge we wanted to address. There have been some tools implemented in the past performing novel miRNA prediction with three of them claiming to support genome-free prediction (Jha and Shankar, 2013; Kuenne et al., 2014; Mapleson et al., 2013). However, with regards to the latter ones, they are either no longer supported, only identify known miRNAs or utilise very stringent criteria for miRNA prediction which are not typical for miRNAs (e.g. requirement for the detection of miRNA products from both strands). Thus, we wanted to explore the potential of developing a novel method for miRNA prediction that would be able to function either with or without a reference genome. Additionally, we wanted to apply this method to interesting datasets such as single-cell data to discover novel miRNA candidates, as well as to datasets examining non-canonical miRNA biogenesis pathways, dependent on different sets of enzymes (e.g. Drosha, Dicer or both).

Finally, this thesis also includes the computational analysis that was conducted as part of three collaborative projects. The aim of the first of these projects was to determine the role of uridylation or other modifications in the regulation of transcript expression in mouse oocytes and in adult cells or embryonic stem cells. The second collaborative project involved exploring the biogenesis landscape of piRNAs in mice. Specifically, the aim was to

explore the existence of alternative piRNA biogenesis pathways that are not dependent on both MIWI2 and MILI proteins, which is the prevalent case. The last collaboration includes the preliminary analysis of the profile of all targets of a single miRNA in *D. Melanogaster*, after mutagenesis induced by CRISPR/Cas9. We also wanted to correlate editing efficiency with the accessibility profile of each of those sites as part of this project.

The results of the work presented in this thesis have been published in six papers so far (one currently under review). We hope that the outcome of this endeavour will highlight additional features in miRNA and piRNA biogenesis and function and will push the boundaries of small RNA research even further.

# Chapter 2

# Analysis of small RNA sequencing data and microRNA modifications

*The results from this chapter have been published in the following papers*:

**1.** "Chimira: analysis of small RNA sequencing data and microRNA modifications"
**DM Vitsios** and AJ Enright.
***Bioinformatics***, Volume 31, p.3365-3367, doi: 10.1093/bioinformatics/btv380 (2015).

**2.** "mRNA 3' uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome."
M Morgan*, C Much*, M DiGiacomo, C Azzi, I Ivanova, **DM Vitsios**, J Pistolic, P Collier, P Moreira, V Benes, AJ Enright and D O'Carroll.
***Nature***, Volume 548, p.347-351, doi: 10.1038/nature23318 (2017).

## 2.1 Chimira

### 2.1.1 Introduction

Small RNA sequencing data are among the most straightforward types of Next Generation Sequencing (NGS) data to analyse. However, many laboratories that generate such data still struggle to develop or apply efficient computational pipelines for the analysis and interpretation of these data. Additionally, in recent years it has been reported that many miRNAs go through post-transcriptional alterations that modify their 3' end, mainly via mono/poly-Uridylation (Heo et al., 2012, 2009) or poly-Adenylation (Lu et al., 2009). Such modifications are believed to impart significant functional changes to miRNAs. Indeed, other modifications and/or editing events have also been observed to occur in several other

studies (Burroughs et al., 2010; Li et al., 2012; Yu and Chen, 2010). These findings from previous studies are a strong indication that the functional roles of small RNAs in different conditions may be greatly influenced by such modifications. Hence, we wanted to explore the full profile of all modifications and/or edits that can be identified in small RNA-Seq data, starting with the development of a novel method.

A method for identifying miRNA modifications could be implemented by aligning small RNA sequences first against their hairpin precursors. The alignment region spanning each miRNA can then be analysed to detect bases in the miRNA sequence that could not possibly have derived from the precursor it aligns to. These unalignable nucleotides are likely either: i) base-calling errors, ii) single nucleotide polymorphisms, iii) ADAR edits or iv) other post-transcriptional miRNA modifications (e.g. via TUTases). Base-calling errors are pseudo-random depending on the platform used and usually more likely to occur towards the 3' end of sequences (Vitsios et al., 2017).

In order to study this diverse pool of possible miRNA post-transcriptional modifications, we eventually developed Chimira (Vitsios and Enright, 2015). This is a cohesive web-based platform for the processing and analysis of small RNA NGS data allowing simultaneous detection of 3', 5' and internal miRNA modifications. The web-server version of Chimira can be found here: http://wwwdev.ebi.ac.uk/enright-dev/chimira.

### 2.1.2 Input

Our method, Chimira, accepts FASTQ or FASTA files as an input, containing adapter and/or barcode stripped small RNA-Seq data (Figure 2.1). The user is provided with a simple system for uploading each sample and replicates and selects among available run options. Additionally, Chimira provides a limited 3' adapter cleaning functionality using *reaper* (Davis et al., 2013) supporting different adapters for each input sample. Finally, the system provides a simple interface for computationally determining likely 3' sequencing adapters in case the user does not have this information available.

Chimira supports mapping of small RNA-Seq data against 209 species specific sets of precursors overall, which are already registered in miRBase (Griffiths-Jones et al., 2008). In order to optimise and speed-up the analysis, *tally* (Davis et al., 2013) is used for de-duplicating the uploaded sequence fragments. *Tally* dramatically reduces the size of input sequence files by collapsing identical sequences into a single entry while storing the total read depth.

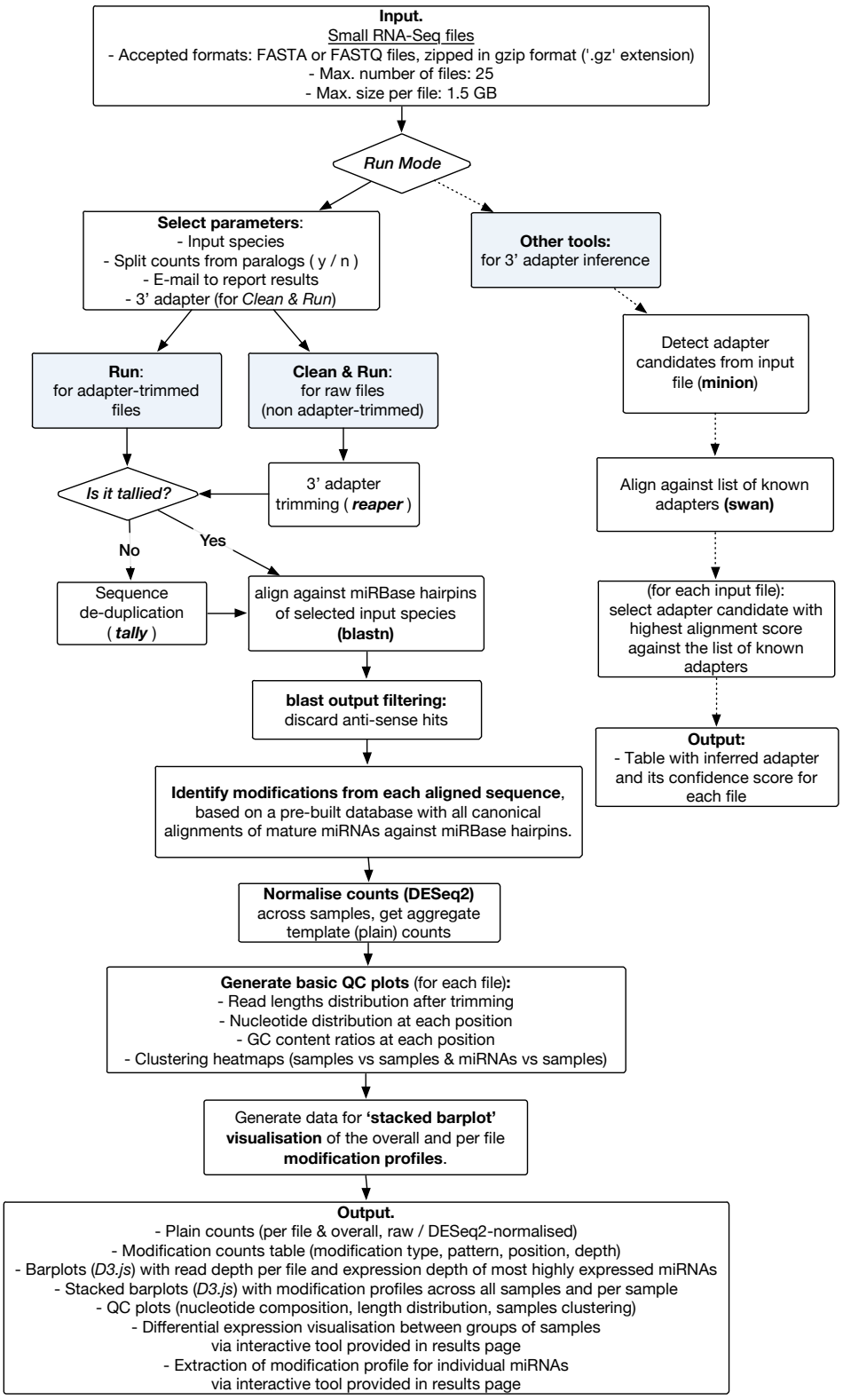



**Fig. 2.1** Chimira pipeline workflow.

Upon a successful file upload and species selection, a new job is submitted to a high-performance computing cluster and the user can follow its progress via a real-time analysis console. In the next section, we will go through the methodological details behind Chimira.

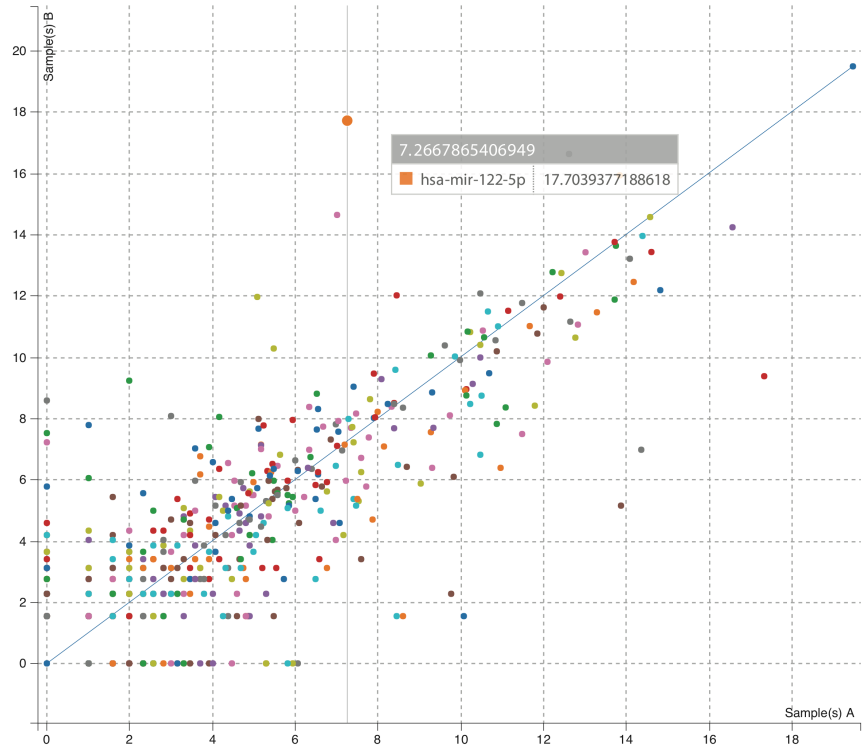



**Fig. 2.2** Differential expression analysis tool: interactive scatterplot of the differential expression of miRNAs between a heart and a liver tissue sample. A single miRNA is highlighted showing its identifier (hsa-miR-122-5p) and the log2 normalised counts in the two samples. As expected, hsa-miR-122-5p expression is significantly skewed towards the liver sample.

### 2.1.3 Methodology

Chimira provides two types of miRNA quantification: *"plain counts"* and *"modifications"*. The *"plain counts"* mode refers to the quantification of miRNA molecules that are expressed in any form (either template or modified) in each of the input samples. Input sequences are first mapped against miRBase using BLASTn (Boratyn et al., 2013) allowing up to two mismatches for each sequence. BLASTn output is then filtered so that antisense matches are discarded. The extracted counts are normalised across all samples using DESeq2 (Love et al., 2014). In cases where a small RNA sequence identically matches to more than one precursor sequences (i.e. miRNA paralogues) the user can choose between using only the first optimal alignment or assigning counts fractionally with equal weights between the identified paralogues.

Furthermore, Chimira integrates basic functionality for comparative analysis of input samples based on their expression profiles. Specifically, differential expression of plain counts (either raw or normalised) between two samples or sets of samples can be visualised through an interactive scatterplot that allows the user to view the miRNA identifier and the different expression levels at each point of the plot (an example is shown in Figure 2.2). Moreover, the user can compare the expression levels of the top-10, 20 or 50 most highly expressed miRNAs in two samples (or sets of samples).

The second and arguably most important functionality of Chimira is the identification of base modifications with miRNAs. This mode quantifies those sequence segments that are found within an input sequence read but which cannot be explained by the underlying reference genome sequence. More specifically, the types of modifications being identified by Chimira include:

- **3'-modifications**: any non-templated sequence in a window that starts at the 5th nt upstream of the 3' end of each miRNA and ends at the 6th nt downstream of the 3' end.
- **5'-modifications**: any non-templated sequence in a window that starts at the 8th nt upstream of the 5' end of each miRNA and ends at the 5th nt downstream of the the 5' end.
- **Internal modifications**: SNPs, ADAR edits and any other non-templated sequence. In order for a modification to be classified as a SNP, an arbitrary 70% value is used as a threshold for the ratio of the modified counts to the overall counts.

We should note here that SNPs reported by Chimira are reflecting the inherent genomic variance of the input samples and are not the effect of enzymes altering the content of miRNAs at the post-transcriptional stage. Specifically, SNP detection is based on the idea that miRBase annotated miRNAs represent the predominant consensus sequence as identified across several samples and conditions. Thus, the identification of single-nucleotide variances in input miRNA sequences by Chimira is the direct effect of an underlying genomic variance that is already present in the sequenced individuals/animals.

In the example shown (Figure 2.3), uridylation and adenylation are the most prevalent modification types in the 1st nucleotide after the 3' end of the miRNAs, while C modifications are highly enriched exactly at the 3' end. ADAR editing is the predominant modification type amongst the internal modifications followed by a moderately expressed C-SNP, 11nt upstream of the 3' end (index position: 11).

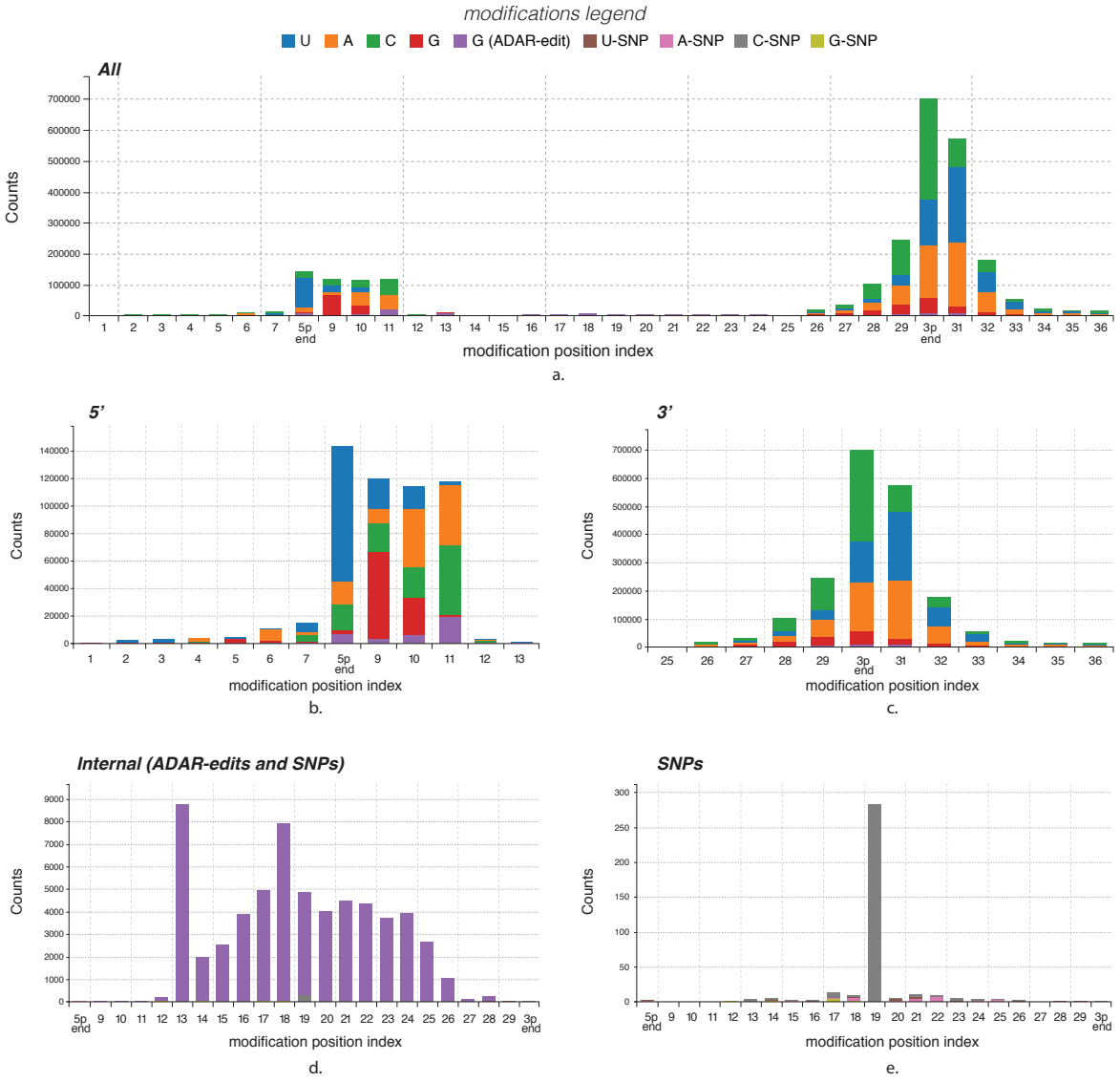



**Fig. 2.3** Aggregated modification profile from 12 Heart, Liver and Brain tissue samples in H. sapiens, as detected by Chimira: a. Global profile b. 5'-Modifications c. 3'-Modifications d. Internal modifications (ADAR edits and SNPs) e. Internal modifications (SNPs). The x-axis corresponds to the index positions across a miRNA molecule. They y-axis corresponds to the raw counts of the identified modification patterns. The start of a miRNA on the x-axis is at index '8' (5' end) while its end is at index '30' (3' end).

It is worth noting that the window lengths being used for identification of 3' and 5' modifications include nucleotide positions also within the original miRNA sequence to better distinguish all possible modifications from multiple miRNA variants originating from the same precursor but with different length mature products. Modification types are inferred from BLAST alignments of input sequences aligned to their hairpin precursors by examining the content of alignment mismatches returned. In order to decipher the correct modification position a reference database has been built initially for all supported

genomes, containing canonical alignments between mature miRNAs and their hairpin precursors. Based on these data each of the identified modification patterns is assigned a position index (Table 2.1) in order to build the full depth-wise modification profile. Chimira also allows the display of the modification profiles across all the samples supplied by the user for a specific selected miRNA. Finally, all counts (plain and modifications) can be downloaded for further analysis as separate files as soon as the processing of a set of samples is complete.

**Table 2.1** Index positions of all modifications relative to the 5'/3' ends of the miRNAs. The directionality of all modification patterns is always considered to be from the 5' to the 3' end.

| Modification type | Modification position | Description |
|---|---|---|
| *3'* | 0 | Modification pattern starts from the 3' end of the miRNA |
| *3'* | +k / -k | Modification pattern starts k nucleotides downstream / upstream of the 3' end of the miRNA |
| *5'* | 0 | Modification pattern starts from the 5' end of the miRNA |
| *5'* | +k / -k | Modification pattern starts k nucleotides upstream / downstream of the 5' end of the miRNA |
| *Internal* | 0 | Modification pattern is precisely at the 5' end of the miRNA |
| *Internal* | +k | Modification pattern is k nucleotides downstream of the 5' end of the miRNA |

### 2.1.4 Validating Chimira against previously published work

In order to validate Chimira's performance we applied it into several example datasets from previously published studies, which explored miRNA modifications. Here, I provide two examples: i) ADAR edits [Adenosine-to-Inosine or A-to-I modification, (Vesely et al., 2014)] and ii) uridylation changes upon TUT4/7 knockout (Liu et al., 2014b).

**Table 2.2** Top five most abundant miRNAs, as identified by Chimira and in a previous study (Vesely et al., 2014).

| **Chimira top miRNAs** | ***(Vesely et al., 2014)* top miRNAs** |
|---|---|
| mmu-miR-378a-3p | mmu-miR-378a-3p |
| mmu-miR-9-5p | mmu-miR-9-5p |
| mmu-miR-127-3p | mmu-miR-127-3p |
| mmu-miR-183-5p | mmu-miR-182-5p |
| mmu-miR-182-5p | mmu-miR-183-5p |

### miRNA expression and ADAR editing validation

We initially present a comparison of Chimira's results with those obtained by a published study regarding ADAR editing events (Vesely et al., 2014). The aim of this validation test is to examine Chimira's efficiency in adequately and accurately quantifying miRNAs and identifying ADAR edits within miRNA sequences. This study specifically examined the effect of ADAR2 enzyme knockout on A-to-I editing and miRNA expression in the mouse brain. First of all, mmu-miR-378a was reported as the most abundant miRNA. Additionally, no significant change was observed in the expression of the five most abundant miRNAs, which also made up 48% of all identified miRNAs. Chimira's respective results (presented in Tables 2.2, 2.3 and Figure 2.4) are comparable or exactly the same as the results of the examined study.

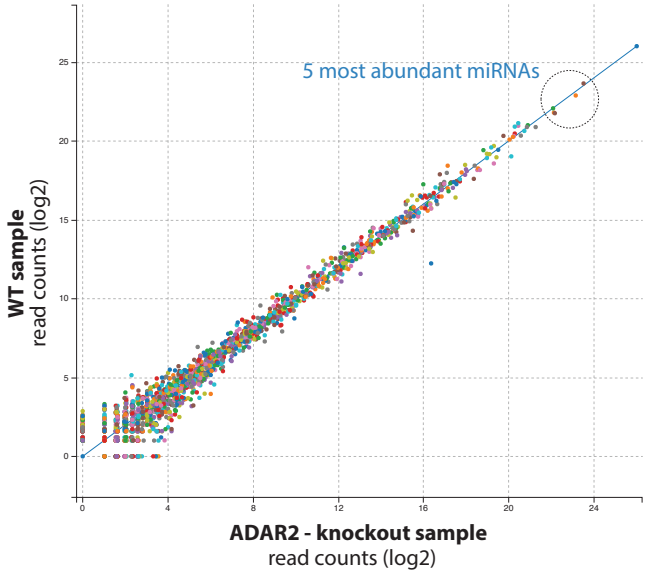



**Fig. 2.4** Differential expression of all miRNAs between wild-type and ADAR2-knockout samples, as returned by Chimira. The expression of the 5 most abundant miRNAs doesn't change significantly between the two conditions.

**Table 2.3** Comparison of the top five most abundant miRNAs depth ratios.

| | Chimira | (Vesely et al., 2014) |
|---|---|---|
| Total depth | 126208757 | - |
| 5 most abundant miRNAs depth | 57220251 | - |
| 5 most abundant miRNAs ratio | 45.3% | 48% |

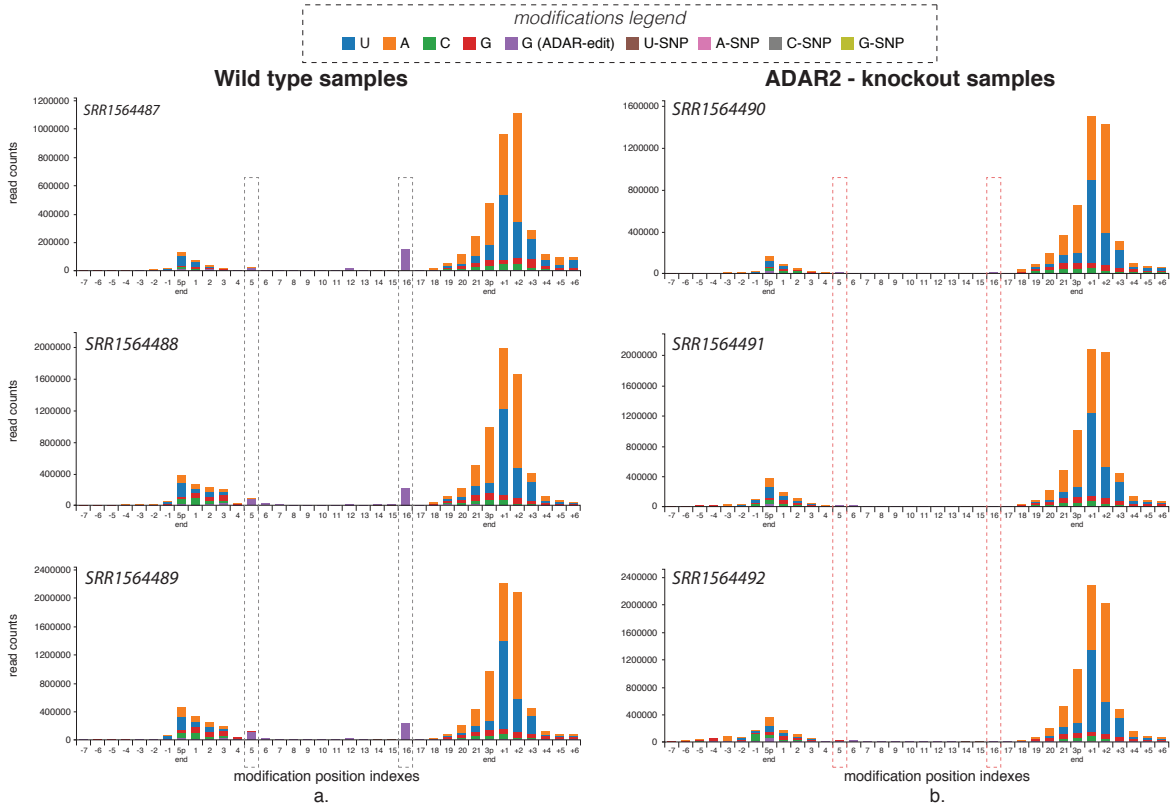



**Fig. 2.5** Visualisation of overall dropout of ADAR editing events in the ADAR knockout samples (b) compared to the wild-type ones (a). The *SRR*** accession numbers correspond to *Run* IDs of replicates from NIH's Sequence Read Archive (SRA).

With regards to identification of modifications, the study reported an overall decrease in ADAR editing events in the ADAR2 enzyme knockout samples, as expected. Chimira is also capturing this global ADAR editing dropout as seen in the global modifications profiles across all samples (Figure 2.5). Additionally, Chimira is successfully identifying ADAR edit events for individual miRNAs as we can see in the comparison of the results returned by the study (Vesely et al., 2014) and Chimira for a representative list of 5 miRNAs (Table 2.4). It is worth noting that for one particular miRNA (*mmu-let-7e-5p*) ADAR editing levels seem to be higher in the knockout samples, based on both the original analysis (Vesely et al., 2014) and the analysis performed by Chimira (both referring to the exact same data).

**Table 2.4** Comparison of significant editing events in a list of 5 miRNAs (detected in all three replicates).

| | Chimira | | | (Vesely et al., 2014) | | |
|---|---|---|---|---|---|---|
| **miRNA** | Position | % editing (WT) | % editing (KO) | Position | % editing (WT) | % editing (KO) |
| mmu-mir-378a-3p | 16 | 6.08 | 0.12 | 16 | 6.3 | 0.2 |
| mmu-mir-379-5p | 5 | 44.5 | 10.8 | 5 | 46.4 | 11.7 |
| mmu-let-7e-5p | 19 | 28.3 | 41.2 | 19 | 29.1 | 42.7 |
| mmu-mir-3099-3p | 7 | 87 | 80.7 | 7 | 79.8 | 66.1 |
| mmu-mir-421-3p | 14 | 14.3 | 3.6 | 14 | 10.9 | 3.3 |

However, *mmu-let-7e-5p* expression represents only 0.02% of the total read depth in the analysed samples, thus this signal can be attributed to noise due to low expression levels and not to an emerging biological effect in the absence of ADAR enzymes. Overall, we can confirm for this dataset a high consistency between Chimira's results and the published findings, thus justifying the validity of our method in terms of accurately extracting both miRNA expression counts and modifications.

### TUT4/7 knockout validation

As a second validation for Chimira we present its performance in detecting the drop of uridylation levels in TUT4/7 enzyme knockout samples. For this type of validation, we have used a dataset with one wild-type and one TUT4/7 knockout sample (Liu et al., 2014b). We first plotted the global profile of 3' terminal uridylation (Figure 2.6) in the two conditions. We can observe dramatic depletion (10-fold decrease) of uridylation at all positions around the 3' ends of miRNAs in the TUTase knockout samples.

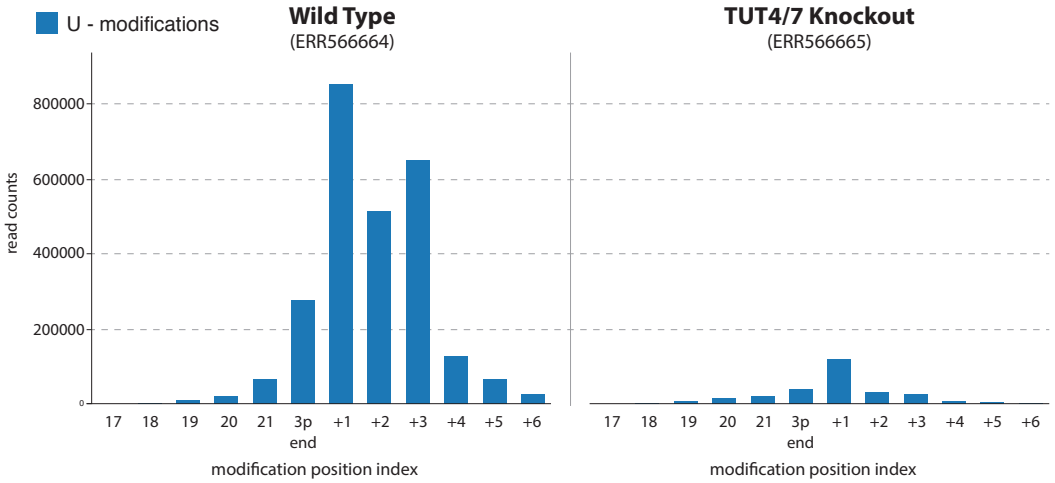



**Fig. 2.6** Uridylation levels in the WT and TUT4/7 samples (Liu et al., 2014b) - accession number: PRJEB6759, as extracted by Chimira (both plots are on the same scale and counts have been normalised across the two samples).

Moreover, after processing Chimira's modification counts across all detected miRNA isoforms, we plotted the change in uridylation levels between the wild-type and knockout samples for each miRNA individually (Figure 2.7). We can observe again that uridylation levels are highly skewed towards the wild-type sample, as previously shown in the global 3' uridylation profile.

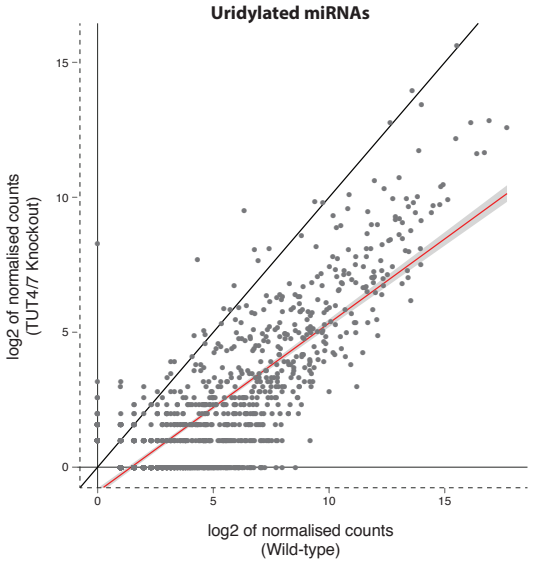



**Fig. 2.7** miRNA uridylation levels across the WT and TUT4/7 Knockout samples (log2 of DESeq2 normalised counts). Normalisation has been performed based on read counts of all miRNAs, either templated or modified.

Following these successful validations, we are going to present in detail the main modes of function provided by Chimira. In the second half of this chapter (section 2.2) we are going to apply Chimira into a larger study that explores the impact of 3' terminal uridylation on the Mouse transcriptome.

### 2.1.5 Plain counts analysis

Chimira's first mode of function is called *plain counts* quantification. The *plain counts* analysis is run by default in either 'Run' or 'Run & Clean' mode. The latter requires trimming of the input files before miRNAs quantification. This is achieved using the *reaper* utility (Davis et al., 2013) and the adapter sequence or sequences file provided by the user. As soon as an input file is clean from the adapter sequences, its reads are uniquified with *tally* (Davis et al., 2013) and a FASTA file is created recording only one entry for each distinct sequence, accompanied with the respective depth count. This format conversion reduces dramatically the size of the input files and consequently the complexity of the alignment

process that follows. The input content is also analysed in order to extract various QC plots (read lengths distribution, nucleotide distribution at each read position and GC content). Input files are then aligned against all miRBase hairpin precursors for the species that has been selected by the user. Alignment, allowing up to two mismatches, is performed using the Standard Nucleotide Blast (BLASTn, v2.2.24+) and output is being filtered to discard any anti-sense hits. If the user has selected to split the multi-mapped read counts to their paralogues, the output from BLASTn is processed in order to identify all paralogues and assign fractionally to them the correct read count values from the multi-mapped hits, using equal weights. If the user has not enabled this option, then only the first BLAST hit is used to assign depth. Finally, the extracted counts are post-processed so that they can be visualised and queried in D3.js and C3.js enabled charts (see e.g. Section 2.1.7), providing also the option to the users to download them in raw format .

### 2.1.6 Modification analysis

Chimira's second and most important feature is the calling of miRNA modifications. Modification analysis is also a default component of the 'Run' and 'Run & Clean' modes. This step is performed after the BLAST alignment has been complete. BLAST's output is parsed to identify all the mismatches of each hit with the associated subject hairpin precursor. The subject/query start and end indexes are retained and are used to infer the position of the detected mismatches. In order to locate the position of every mismatch across the canonical hairpin precursor sequence, relative to the mature miRNAs, a database has been built for all 209 supported species containing the canonical alignments of all mature miRNAs with their respective hairpin precursors, including information concerning the indexes of the alignment for each case. Based on this information and the depths for each of the detected mismatches, modifications are assigned a specific modification type and position. Thus, a collection of modifications is assembled eventually containing all identified 3', 5' and internal modifications. SNPs are identified in cases where the respective mismatch is found in at least 70% of the associated reads. ADAR edits are called from all A to G transitions provided they occur in at least 90% of all associated sequences. Finally, the extracted modification counts are being post-processed for visualisation purposes and are bundled in more user-friendly raw formatted files, downloadable by the user.

### 2.1.7 Quality-Control (QC) visualisation

Chimira generates basic QC plots regarding the read length and nucleotide distribution of input sequences (Figures 2.8, 2.9 & 2.10). In addition, plots with the total miRNA

counts per sample and the top-10 most highly expressed miRNAs across all samples are also provided.

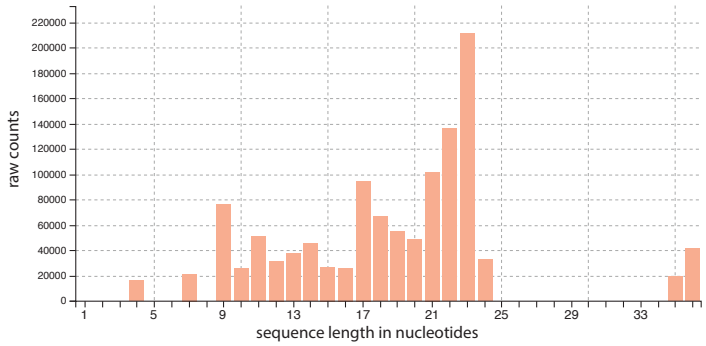



**Fig. 2.8** QC plot generated by Chimira: read lengths distribution after trimming.

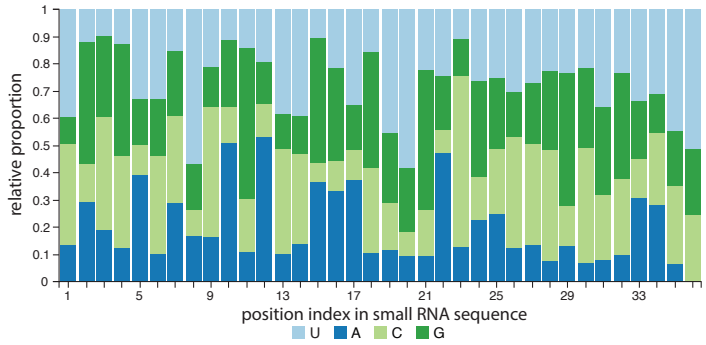



**Fig. 2.9** QC plot generated by Chimira: nucleotide distribution at each position.

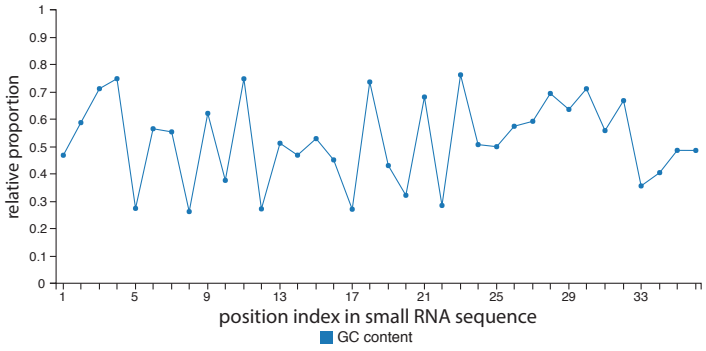



**Fig. 2.10** QC plot generated by Chimira: GC content ratios at each position.

### 2.1.8 3' adapter detection feature

In case the 3' adapter associated with a FASTQ file is not known, Chimira is able to suggest possible adapter candidates through the 'Other tools' section. This feature is based on

the tool *minion* (Davis et al., 2013) which uses De Bruijn graphs in order to infer potential adapter sequences based on two criteria: a) frequency of occurrence b) attachment to multiple different prefixes. After a list of potential adapters has been created, it is being aligned with *swan* (Davis et al., 2013) against a list of known adapters that we have compiled from various popular sequencing machines and protocols. The best candidate for each input file is selected based on the alignment score of a potential adapter with a known adapter, its length and the end indexes of the alignment.

After an optimised selection process is performed based on the aforementioned criteria a single adapter sequence is suggested for each input file, with a 'Degree of Alignment' score denoting its alignment score with a known adapter sequence. In almost all cases, a suggested adapter with a 'Degree of Alignment' of 100% can be safely used as the correct adapter. This can still be true also for cases where the score is 95% or even less, since the suggested adapter may just be a subsequence of a known adapter.

In case none of the adapter candidates aligns with any of the verified adapters over a certain threshold ratio (85% by default), a sequence inferred by *minion* may still be suggested as the adapter of the input samples. However, the Degree of Confidence in that case is 0% and the suggested adapter should be cross-checked manually. This cross-validation might also be needed in case the score is 100% but the length of the suggested adapter is less than 15nt or the score is e.g. less than 90%. In any case, the trimming efficiency from the use of a particular adapter can be evaluated by inspection of Chimira's stacked modifications profiles and QC plots after a run has been complete.

### 2.1.9 Methods

#### Chimira back-end

All input files are uploaded and stored on a server using one of the fastest academic networks in Europe. The file content is validated and an error message is displayed on the user's browser window if it does not comply with the allowed input specifications. In any other case, its size is further estimated and based on performance evaluation data acquired form previous training datasets the required resources for submission to a cluster are allocated. After this quality control and pre-processing step, Chimira submits a new queued job to the EMBL-EBI High Performance Computing Cluster. The progress of the process can be viewed at all times from an analysis console window that is available on Chimira's progress page, displayed right after a job has been launched. The pipeline Chimira uses for the core analysis is based on Perl (v5.16.0) and R (v3.1.2). A separate thread provided by the Perl API handles each file. Moreover, multiple other threads are launched during the

process to initiate cascaded parallel processes for quality control, statistical analysis and merging of various output results, among other tasks. Thread synchronisation and data integrity is coordinated and assured during the run. Upon the completion of each job the user is redirected to the results page where he can browse through the output, download all the extracted data (e.g. counts, modifications, etc.) and query the results using the interactive tools provided.

### Time benchmarking

We are providing here the benchmarking results from the comparison of Chimira with other already published web servers (Oasis - Capece et al., 2015 and MAGI - Kim et al., 2014). We tested two different aspects from each application: a) upload time and b) execution time, using three different datasets of increasing input size (Figure 2.11).

### Dependencies

Chimira has been developed as a web-application with a core pipeline based in Perl and R. The full list of library/tool dependencies of Chimira along with their recommended versions is shown in Table 2.5.

Chimira is supported by all popular web browsers (e.g. Chrome, Safari, Firefox) that run on personal computers and is also accessible by any JavaScript enabled browser on mobile devices.

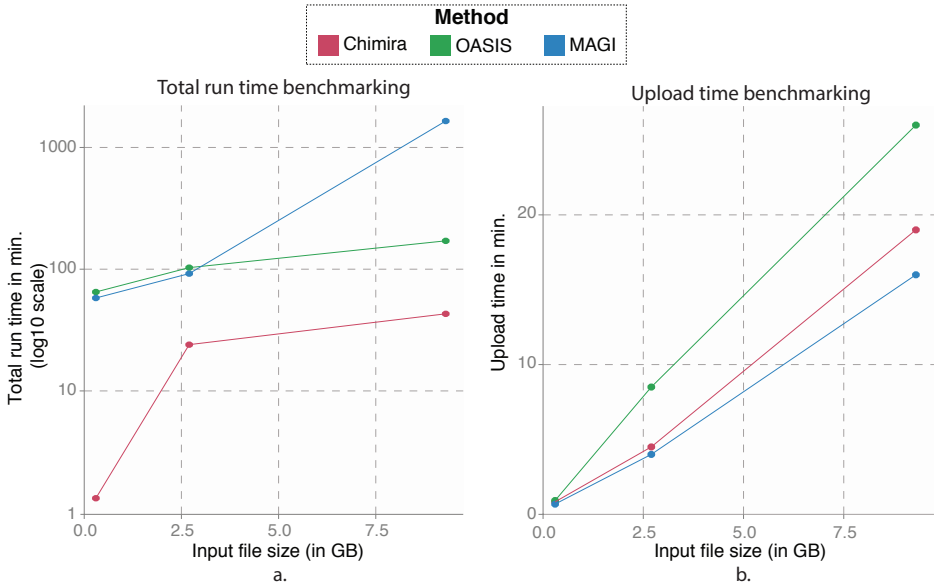



**Fig. 2.11** Scatterplots for: a) total run times and b) upload times of Chimira, MAGI and OASIS methods for three different input datasets of small, medium and large size. Y-axis in sub-figure (a) has been normalised to a log10 scale.

**Table 2.5** Chimira dependencies and recommended versions

| Library / Tool | Version |
|---|---|
| miRBase | Release 21 |
| BLASTn | 2.2.24+ |
| Reaper | 15-065 |
| Minion | 15-065 |
| R | 3.1.2 |
| Perl | 5.16.0 |
| Perl CGI | 3.59 |
| PHP | 5.3.3 |
| JavaScript | >= 1.7 |
| Fine-uploader (JavaScript plugin) | 5.1.3 |
| jQuery | 1.10.2 |
| jQuery UI | 1.11.4 |
| D3.js | 3.5.0 |
| C3.js | 0.4.8 |
| DataTables (jQuery plugin) | 1.10.0 |

## 2.2 3' terminal uridylation impact on the Mouse transcriptome

### 2.2.1 Background

It is already known that in the early stages of female zygotic development in mammals there is complete lack of transcription at all stages of the growth of oocytes (Tadros and Lipshitz, 2009). Growing oocytes undergo maturation through sequential stages of growth (Svoboda et al., 2015): primary, secondary, early antral or late antral and increasing cell size (Figure 2.12). At the end of the maturation process, mature oocytes upon ovulation are capable of supporting fertilisation as well as development (Eppig and Schroeder, 1989). This competence of mature oocytes is largely driven by the maternal mRNAs which are directly deposited into the oocytes. So, despite the inherent lack of any transcriptional activity, gene expression in oocytes is already active and specifically instructed by the maternally-derived transcriptome (Ma et al., 2013; Pan et al., 2005). Additionally, maternal mRNAs at all stages of the growing oocytes have a high degree of stability (Brower et al., 1981; De Leon et al., 1983) but what's even more striking is that each stage of maturation is associated with a distinct transcriptome (Pan et al., 2005). Thus, growing oocytes not only have the ability to retrieve mRNAs from the maternal transcriptome but they can also regulate the expression of the derived transcripts using a selective degradation mechanism.

The mechanism oocytes employ to drive transcripts degradation is not known. In the next section we try to address this problem.

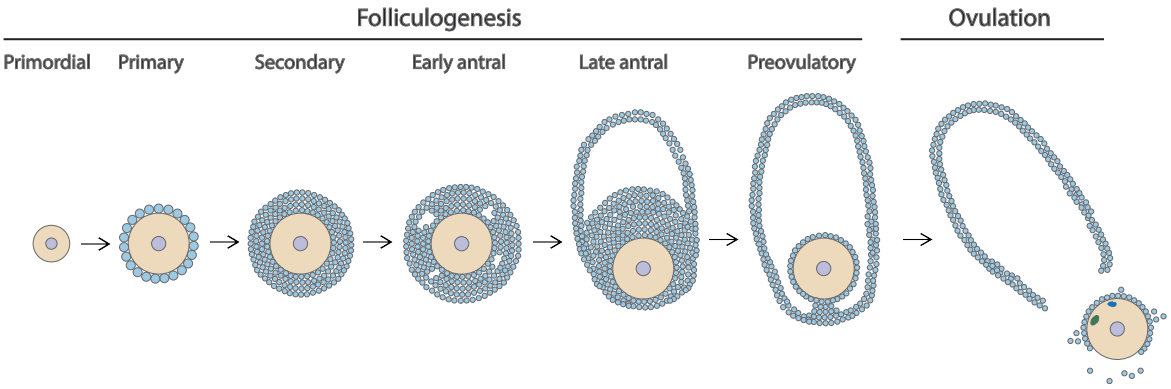



**Fig. 2.12** Sequential stages of oocyte maturation from folliculogenesis to ovulation.

### 2.2.2 Results

In an effort to elucidate the transcriptional regulation machinery in oocytes we collaborated with the O'Carroll Lab. The results of this work with regards to mRNA modifications and gene expression have been conducted and extracted by members of the O'Carroll Lab and Dr. Anton J Enright. My work towards this project has focused on extracting the modification profiles for miRNAs using Chimira and trying to identify the degree to which a modification is affecting the expression of miRNA molecules across somatic or embryonic cells.

To begin with, it has already been discovered that poly-A tail length and 3' terminal uridylation of mRNAs are playing a predominant role for mRNA turnover and degradation (Lim et al., 2014; Mullen and Marzluff, 2008; Rissland and Norbury, 2009). Specifically, for transcripts with a poly-A tail below 25 nucleotides, there has been observed destabilisation of Poly(A)-binding protein (PABP) binding (Baer and Kornberg, 1983; Eliseeva et al., 2013) which in turn allows binding of the terminal uridylyl transferases 4 (TUT4) and TUT7 (TUT4/7) and subsequent uridylation of mRNAs. This mechanism is present in around one fifth of transcripts in human cells lines, and it has been proved that the addition by TUT4/7 enzymes of extra nucleotides at the 3' end of transcripts with shorter than normal poly-A tails aids notably to their decay (Chang et al., 2014; Lim et al., 2014).

In order to examine the respective modification profile in oocytes, which is dependent on TUT4/7 enzymes, the collaborators prepared samples from the late antral/preovulatory stage oocytes (GV oocytes) and confirmed that TUT4 and TUT7 are expressed at each stage of maturation. Meanwhile, three libraries from somatic tissues (bone marrow, liver

and mouse embryonic fibroblasts or MEFs) and another one from embryonic stem cells (ESCs) were prepared in order to compare the 3' terminal profiles in all these cells to those from oocytes. For all samples, the poly-A tail length and 3' terminal uridylation of mRNAs were identified using TAIL-seq (Chang et al., 2014). Terminal uridylation levels varied across the samples however the poly-A tail length with the highest frequency was shorter in GV oocytes than in all other cases (around 68 nucleotides instead of 78). Moreover, GV oocytes had the highest relative ratio of oligo- to mono-uridylation among all examined cell types. These findings indicate that uridylation, in the form of short fragments, is more prevalent in oocytes than in somatic cells and thus may have a distinguishing functional role.

This hypothesis was tested by examining the effect of TUT4/7 double knockout in GV oocytes. What was observed was a complete incompetence of oocytes to support early embryonic development in the absence of the TUTase enzymes. This may be explained by either failure of oocytes to complete meiosis I or the incapability of supporting fertilisation upon successful completion of the meiosis I stage. So, we can safely presume that TUT4/7's role is indispensable for oocyte growth and specifically for successful transition from the meiosis I to meiosis II phase of cell division. In addition, gene expression changes were studied between the TUT4/7 wild-type and knockout conditions. Many genes were observed to be deregulated and specifically the large majority of them where upregulated. This is expected, since the loss of an essential component of RNA degradation mechanism, such as uridylation via TUT4/7, stabilises transcripts and allows them to avoid decay. So, it was shown in this way that the maternal transcriptome in oocytes is largely regulated and defined by the TUT4/7 enzymes, via the addition of oligo-uridine fragments at the 3' ends of transcripts.

We have seen so far that 3' uridylation induced by TUTases is a prevalent mechanism within the RNA degradation pathway in oocytes. A possible assumption could be that this mechanism is so fundamental that it may be ubiquitous in other cell types and tissues. Thus, the collaborators sought to examine if uridylation has the same functionality in other cell types, as that seen in oocytes, and more specifically in embryonic stem cells (ESCs), mouse embryonic fibroblasts (MEFs), liver cells and bone marrow (BM) cells. They prepared wild-type and TUT4/7 knockout samples for each cell type. In all knockout samples, there was observed a dramatic decrease or complete depletion of uridylation levels (both oligo- and mono-uridylation). However, they observed that the loss of TUT4/7 did not have any effect in ESC differentiation/pluripotency or in proliferation of ESCs and MEFs. Moreover, TUT4/7 deletion in somatic cells (liver and bone marrow cells) did not have any phenotypic effect, since mice appeared to be healthy up to several months after

knockout induction. So, in all cases, the loss of TUT4/7 did not have any detrimental effect in any of the embryonic or somatic cells.

### Uridylation effects on miRnome

The next step of this analysis was to investigate whether uridylation or its loss has any appreciable effects on miRNA expression. It is already known that miRNA expression is completely suppressed in oocytes (Ma et al., 2010; Suh et al., 2010). Thus, we focused on examining the effect of TUT4/7 knockout in embryonic and somatic cells (ESCs, MEFs, liver cells and bone marrow cells). For this type of analysis, small RNA libraries for each cell type where prepared by the collaborators from the O'Carroll Lab.

We first extracted plain miRNA counts from the small RNA-seq samples using Chimira (Vitsios and Enright, 2015). Chimira cleans input sequences from their 3' adapters and maps them against all mouse hairpin precursors (miRBase Relase 21, Griffiths-Jones et al. (2008)), allowing up to two mismatches. Counts of multi-mapped reads were assigned only to the first optimal alignment call returned by BLASTn. In all cases, control and experimental (TUTase knockout) samples were normalised using the DESeq2 package (Love et al., 2014). Furthermore, we also extracted miRNA modification counts using Chimira. Modification analysis was then restricted to pure modification events, i.e. mono-nt or poly-nt, where nt can be any of the U, A, C or G (poly-nt modifications refer to sequences of two or more identical nucleotides). We also collapsed counts from all other miRNA variants with their respective unmodified miRNA counts.

The first part of our analysis was to confirm the TUT4/7 knockout effect by examining the uridylation profiles between the wild-type and knockout conditions across the selected cell types. We noticed clearly in all cases that the expression of a large proportion of uridylated miRNAs is skewed towards the wild-type condition (Figure 2.13). These miRNAs (highlighted in green) show statistically significant differential expression (fold-change > 2 and P-value < 0.05) between the two conditions and demonstrate the depletion of uridylation levels in the absence of TUTase enzymes.

Meanwhile, we wanted to assess any implications of TUT4/7 knockout in other modification events including adenylation, cytidylation and guanylation to assess the extent of perturbations in transcriptional machinery of TUT4/7-deficient cells. Following a similar approach, we extracted the profiles for these modification patterns (mono- or poly- nucleotide modifications of type A, C or G) and assessed the significance for each differential expression profile, using a negative binomial Wald test. Based on these profiles we observed that TUT4/7 knockout has a significant impact only in miRNA terminal uridylation (Figure 2.14) and specifically leads to decrease in uridylation levels. On the other hand,

adenylation seems to follow a pattern of modest increase in the TUT4/7 knockout samples. This could be a slight indication that the adenylation mechanism is not obstructed by the TUT4/7 enzymes in the knockout samples thus allowing for higher terminal adenylation of miRNA transcripts. However, this signal is not statistically significant in any of the examined cell types, thus no conclusion can be made about its validity or real biological functionality. All other modification types (cytidylation and guanylation) remain either stable or their change is not statistically significant in the event of TUT4/7 knockout and also their presence could mostly be attributed to sequencing biases (Vitsios et al., 2017).

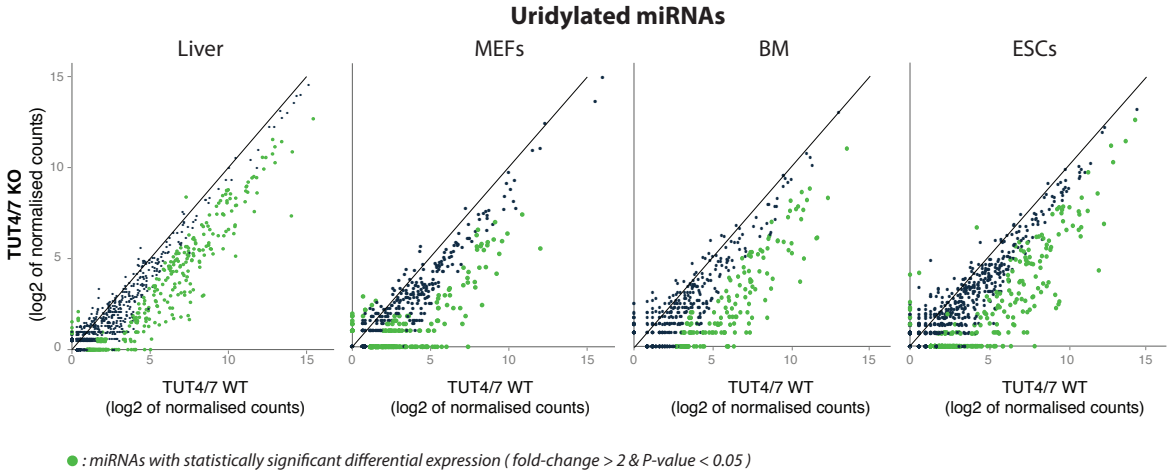



**Fig. 2.13** Expression of uridylated miRNAs in TUT4/7 wild-type (WT) and knockout (KO) samples across liver cells, mouse embryonic fibroblasts (MEFs), bone marrow (BM) cells and embryonic stem cells (ESCs). MicroRNAs with statistically significant differential expression between the WT and KO conditions are coloured in green. Two to four biological replicates have been used for each cell type.

We now wanted to determine if those changes in uridylation levels in the absence of TUT4/7 have any effect in miRNA degradation and overall expression. We observed that there was a very modest impact on overall miRNA expression levels in the four cell types examined (Figure 2.15). Specifically, there were only few cases of miRNAs that demonstrated statistically significant differential expression levels in the two conditions, including mainly miRNAs of the let-7 family. Another feature observed in these results is that in embryonic stem cells there was a mild increase of the expression of let-7 miRNAs in the absence of TUT4/7 while in the other three somatic cell types (liver, bone marrow and MEFs) the opposite effect was seen. This result can be attributed to the expression of LIN28a protein-coding gene in ESCs, which acts as an inhibitor of pre-let-7 processing via TUT4/7-mediated oligo-uridylation (Ali et al., 2012; Hagan et al., 2009; Heo et al., 2008; Viswanathan et al., 2008). On the other hand, in liver, bone marrow and MEFs, where LIN28

is not expressed, TUT4/7 normally acts as an enhancing factor for pre-let-7 processing into mature miRNAs thus justifying the mild drop in let-7 expression in the absence of TUTases.

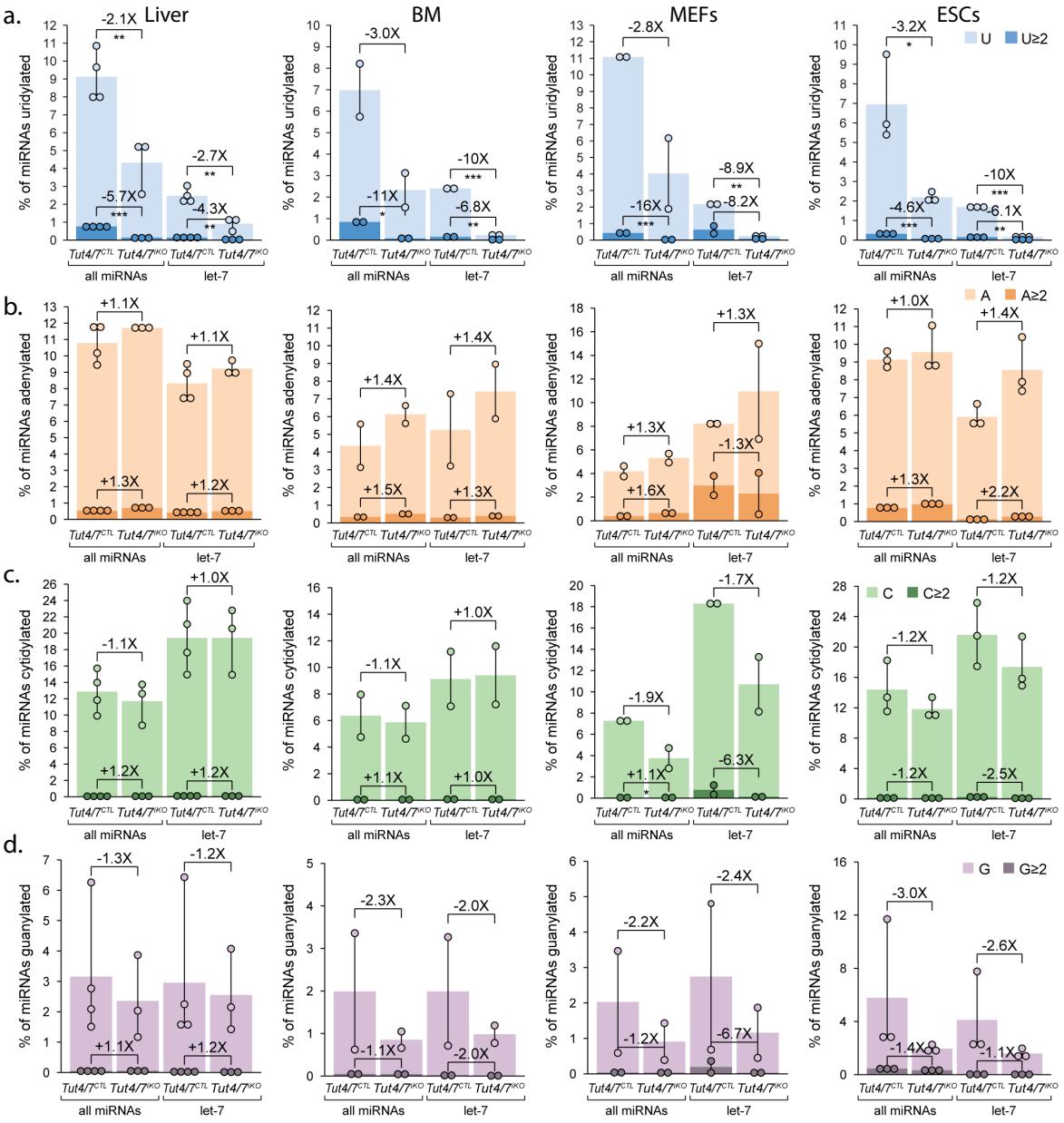



**Fig. 2.14** TUT4/7 knockout has a significant impact only in miRNA terminal uridylation. 3' terminal (a) uridylation, (b) adenylation, (c) cytidylation and (d) guanylation levels for all miRNAs and let-7 family members in Tut4/7 wild-type and knockout samples from liver, bone marrow (BM), mouse embryonic fibroblasts (MEFs) and embryonic stem cells (ESCs). Data points (circles) represent distinct replicates, with vertical lines indicating the range of observed values. The average value in all cases is denoted by the height of each bar. The fold changes in modification frequency between the wild-type and TUT4/7 knockout conditions is shown as well as their significance scores using a (negative binomial) Wald test, from the DESeq2 statistical package (*: P-value < 0.05, **: P-value < 0.01, ***: P-value < 0.001).

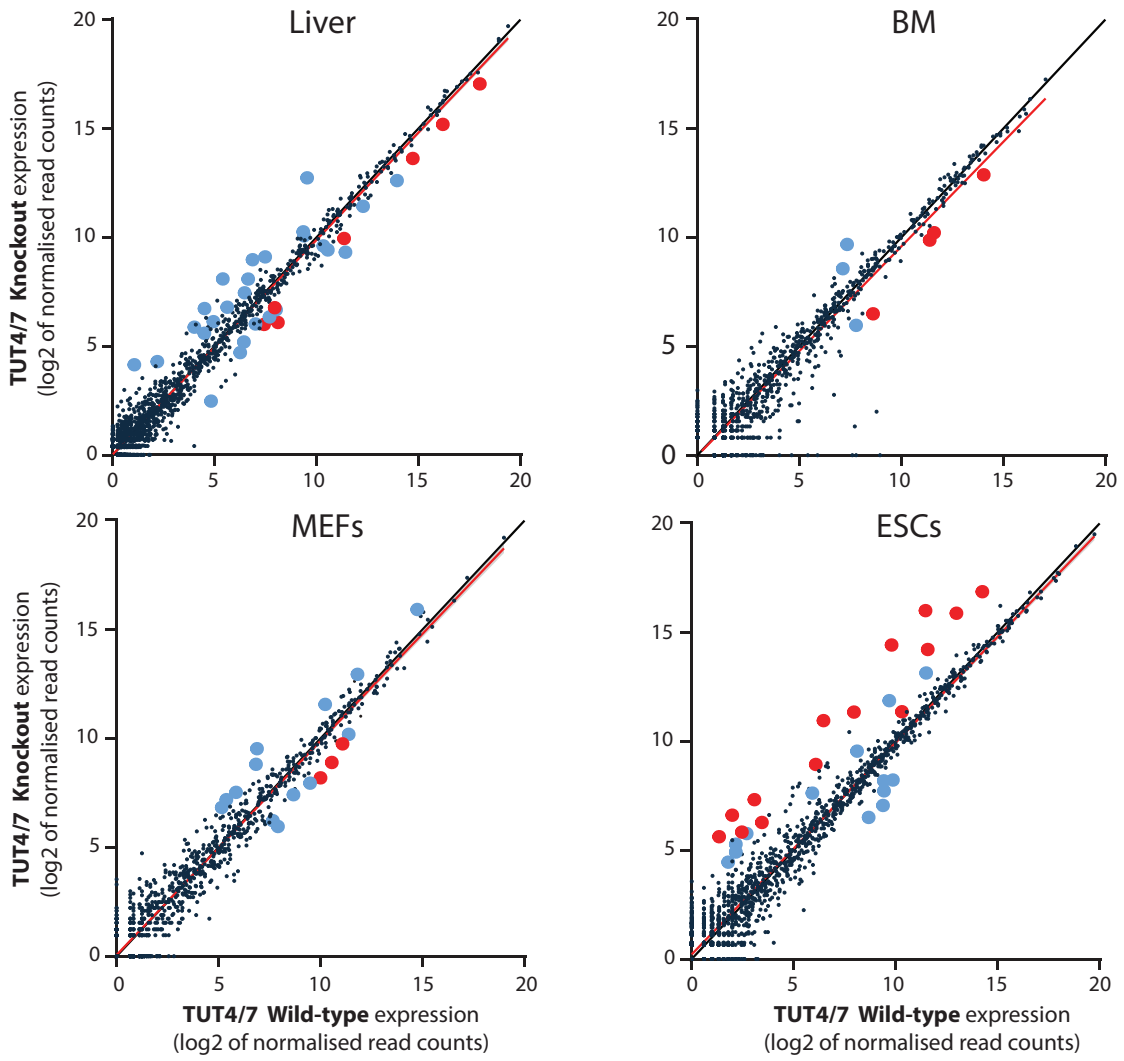



**Fig. 2.15** Differential expression of miRNAs in TUT4/7 wild-type (WT) and knockout (KO) samples across liver cells, mouse embryonic fibroblasts (MEFs), bone marrow (BM) cells and embryonic stem cells (ESCs). MicroRNAs with statistically significant differential expression between the WT and KO conditions are coloured in red (for the let-7 family miRNAs) and blue for the rest of miRNAs. Two to four biological replicates have been used for each cell type.

Finally, we wanted to examine the extent of uridylation occurring in the 3' ends of miRNAs based on which strand (5' or 3') of the precursor duplex they originate from. Thus, we profiled uridylyl-modifications across all miRNAs in liver, bone marrow, MEFs and ESCs both in the wild-type and knockout conditions. Both 5' and 3' miRNAs show a drop in uridylation levels in the absence of TUT4/7. However, we noticed that uridylation levels in the wild-type condition are higher (around 2-fold) in the 3' miRNAs compared to the 5'

miRNAs. This result may indicate that since 3' miRNAs have their 3' ends exposed right after Drosha processing (while for 5' miRNAs this only happens after Dicer processing), they may be modified in multiple stages, thus yielding higher levels of overall modifications.

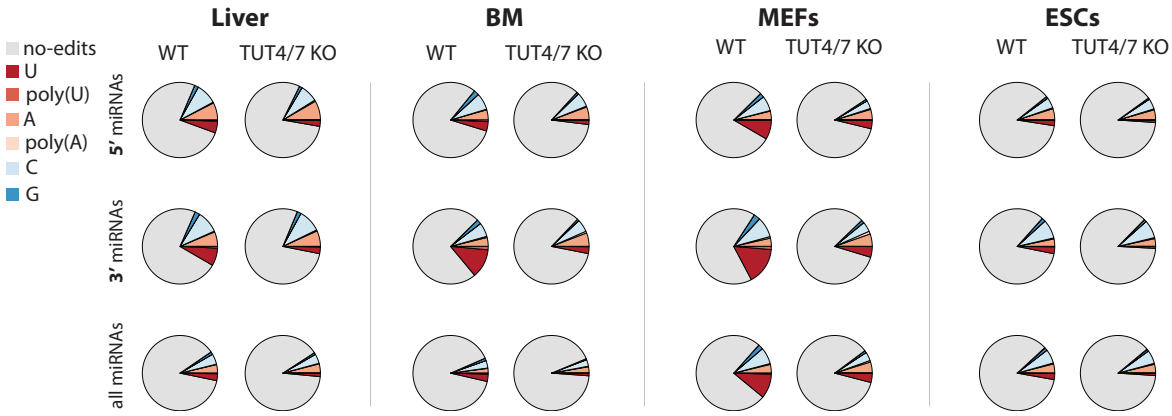



**Fig. 2.16** miRNA modifications based on the precursor strand of origin during maturation. Distribution of modification events (U/poly-U, A/poly-A, C and G) across 5'-strand miRNAs, 3'-strand miRNAs and all miRNAs in wild-type and TUT4/7 knockout samples from liver, bone marrow (BM), mouse embryonic fibroblasts (MEFs) and embryonic stem cells (ESCs). 3' miRNAs are more extensively modified than 5' miRNAs in all cell types, potentially due to the fact that their 3' ends are exposed already for processing at an earlier stage during maturation than the 3' ends of 5' miRNAs.

## 2.3 Conclusion

We have developed a novel method, Chimira, for accurate identification of modification events and their positions across the entire miRNA length (3'/5' ends and main body). The web-server version of Chimira is available here: http://wwwdev.ebi.ac.uk/enright-dev/chimira. After validation of Chimira, we applied it for a larger analysis exploring the effect of uridylation in oocytes and somatic cells in Mouse. In summary, we can conclude that the loss of TUT4/7 leads to a consequent reduction of terminal mRNA and miRNA uridylation without affecting any other types of modification. However, these changes in terminal modifications brought only modest alterations to the repertoire of expressed miRNAs and similarly did not have a notable impact on gene expression in any of the cell types or tissues analysed. Thus, in somatic cells, uridylation via TUT4/7 is not essential as a degradation mechanism, in contrast to oocytes where we demonstrated its indispensable role in mRNA transcripts regulation.

Chimira is provided publicly as a web-application with an intuitive interface (Appendix A) and its efficiency and speed have been demonstrated. The traffic recorded on Chimira's website has reached around 1,000 unique users since its initial release in June 2015. More than 2,500 sessions of analyses have been performed with Chimira, with a variable number

of runs submitted in each session. Chimira has also been cited in 21 papers (November 2017) and we expect it to be of benefit to the small RNA community for the years to come.

# Chapter 3

# Large-scale study on miRNA biogenesis, function and epi-transcriptomic features

*The results from this chapter have been published in the following paper*:
"Large-scale analysis of microRNA expression, epi-transcriptomic features and biogenesis"
**DM Vitsios**, MP Davis, S van Dongen, AJ Enright.
***Nucleic Acids Research***, Volume 45, p.1079-1090, doi: 10.1093/nar/gkw1031 (2017).

## 3.1 Introduction

Chimira has proven to be a great asset in the identification of miRNA modifications, as described in the previous Chapter. Subsequently, we wanted to explore the prevalence of miRNA modifications on a larger scale. Thus, in this chapter, we aim to extract the miRNA modification profiles from a wide range of datasets. This will allow us to infer the most prevalent patterns in epi-transcriptomic modifications of miRNAs as well as other characteristics associated with their biogenesis, giving indications as to the features that may drive these attributes.

The advent of Next-Generation Sequencing (NGS) technologies has made it a relatively straightforward task to detect these molecules and their relative expression via sequencing. However, even though NGS has greatly increased our power to detect and catalogue miRNA expression, these data are usually complex and are processed differently from laboratory to laboratory. Hence, while there are currently over 850 deposited small RNA

sequencing datasets in ENA (Leinonen et al., 2010) and GEO (Barrett et al., 2013), there is not a comprehensive database or catalogue of where and when these miRNAs have been detected. Additionally, as each experiment has been processed with different criteria and filters the results may be difficult or impractical to compare directly. We sought to address these issues by building a comprehensive catalogue of miRNA expression from large numbers of previously published small RNA sequencing datasets for both Human and Mouse, for which raw FASTQ data are available.

In this analysis, we focus on Human and Mouse for which the majority of data are available. We reanalyse sequencing data from 461 samples into a coordinated catalogue of microRNA expression. We use this to perform large-scale analyses of miRNA function and biogenesis in order to further expand our understanding of miRNA function in animals. These analyses include global expression comparison, co-expression of miRNA clusters and the prediction of miRNA strand-specificity and underlying constraints. Additionally, we report for the first time a global analysis of miRNA epi-transcriptomic modifications and assess their prevalence across tissues, samples and families. Finally, we report a list of potentially mis-annotated miRNAs in miRBase based on their coverage profiles.

For each dataset we have performed automated barcode demultiplexing, 5'/3' adapter detection using de Bruijn graph analysis followed by adapter excision and computational size selection (15-32nts). Additionally, some samples require the removal of poly-A or poly-C tracts. This data pre-processing step has been performed by a pipeline which was based on the already published pipelines *Kraken* (Davis et al., 2013) and *Chimira* (Vitsios and Enright, 2015). Each dataset has been mapped to known miRNA precursor sequences using a single computational pipeline (*Chimira*). This pipeline not only represents a cohesive platform for the collation and analysis of small RNA NGS data but also allows the detection of events such as 5'/3' modification of miRNAs via enzymes such as terminal uridylyl transferases [TUTases, (Heo et al., 2009)] or adenosine deaminase RNA (ADAR) editing (Blow et al., 2006). The raw count data obtained was normalised and annotated according to each experiment, providing a comprehensive catalogue of miRNA expression in Human and Mouse together with a variety of complementary data that can assist us in the analysis of miRNA function and biogenesis.

Finally, the results have been collated into a comprehensive online repository of miRNA expression and features such as modifications and RNA editing events, which is available at: http://wwwdev.ebi.ac.uk/enright-dev/miratlas. This resource is also accompanied with tools for advanced queries to extract miRNA modification and/or expression patterns across multiple conditions.

## 3.2 Results

A total of 52 NGS datasets were obtained from both ENA and GEO covering in total 461 biological samples including biological replicates (Table 3.1). For each dataset, FASTQ raw data were downloaded and annotation information was manually curated according to tissue, cell type, disease state or cell line. More specifically, dataset annotation was generated manually based on the information that is available in the original databases for each dataset. The curated annotation classes may refer to either a cell line/type/tissue (e.g. liver) or a condition/disease (e.g. cancer). In case both a cell line/type/tissue and a condition/disease are provided for a dataset, only the condition/disease information is used for the annotation of that dataset. Additional information is provided in the *miratlas* repository as well as links to the original resources.

These raw data served as the foundation for all subsequent analyses described below. Of particular note in the case of small RNA datasets is that the molecule being sequenced is usually shorter than the sequence read obtained from an NGS experiment. This means that most captured sequences contain both small RNA sequence and some amount of the 3' sequencing adapter. In general, input datasets used for this analysis have been prepared by different experimental protocols using a variety of barcodes, 3' adapters and/or 5' adapters. Thus, it was imperative first of all to infer the read geometry of each input dataset in order to later clean the sequences from barcodes/adapters and further process the samples. Thus, we developed a pipeline (Figure 3.1) that is deciphering firstly the presence or not of a barcode sequence in the input samples by looking for enrichment of any sequence of 3-6nt long at the 5' end of the first 2 million sequences of an input sample file. Inference of the 3' adapter was accomplished through the command-line version of the 3' adapter detection feature of *Chimira* (Vitsios and Enright, 2015), which integrates *minion* and *swan* (Davis et al., 2013) and is based on 3' de Bruijn graph assembly. In that case though, the position of the suggested adapter relative to the input sequences is also defined and thus the inferred adapter may either be a 5' or 3' adapter. In case the suggested adapter sequence did not match at least 90% with a known Illumina adapter sequence (without any mismatches), input files were also manually checked in order to identify any potential sequences that were attached to already known highly expressed miRNAs, such as the let-7 miRNAs. Datasets from ENA/GEO that were detected with ambiguous adapter sequences or barcode annotation were excluded from the analysis. Eventually, we compiled a set of 52 datasets with a well characterised read geometry that we used for our analysis.

Following successful inference of the 3' adapter sequence across 52 datasets we remove the adapter sequences using *reaper* (Davis et al., 2013). Finally, these adapter purged

sequences (representing small RNAs and contaminants) were de-duplicated, using *tally*, such that each sequence was only represented once in the final input FASTA file accompanied with its respective coverage depth. These cleaned and de-duplicated sequences were the primary input into the miRNA analysis pipeline (*Chimira*). This pipeline automatically scans each sequence against all known miRBase precursor sequences from a selected species and detects the likely miRNA, which arm of the precursor it originated from (5'/3') and searches for non-canonical nucleotides which may be the result of editing and/or modification by enzymes such as Tutases. All miRNA counts, annotations and features detected are stored in a MySQL database for further analysis.

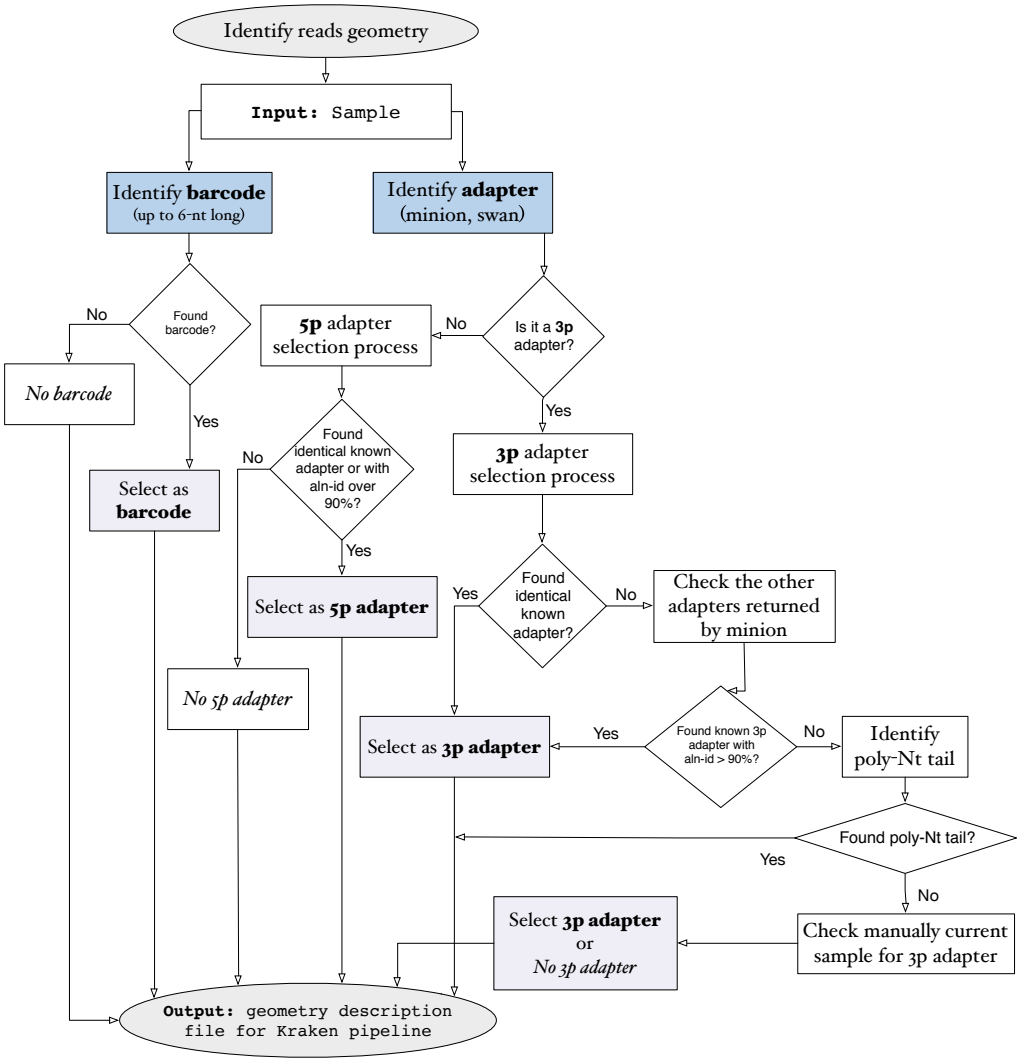



**Fig. 3.1** Flowchart for the detection of barcodes, 5' and/or 3' adapters from small RNA-Seq samples with complex read geometry.

**Table 3.1** Comprehensive table of all examined datasets accompanied with annotation about the data source, genome, number of samples and tissue/cell type of origin or condition.

| NGS Dataset ID | Source | Genome | Num of samples | Tissue / Cell type |
|---|---|---|---|---|
| E-GEOD-15229 | ArrayExpress | Homo sapiens | 31 | B cells |
| E-GEOD-16579 | ArrayExpress | Homo sapiens | 12 | Cancer |
| E-GEOD-18381 | ArrayExpress | Homo sapiens | 12 | Cancer |
| E-GEOD-20664 | ArrayExpress | Homo sapiens | 3 | Leukemia |
| E-GEOD-21191 | ArrayExpress | Homo sapiens | 3 | Germ cells |
| E-GEOD-21279 | ArrayExpress | Homo sapiens | 15 | Liver |
| E-GEOD-23090 | ArrayExpress | Homo sapiens | 6 | B cells |
| E-GEOD-30286 | ArrayExpress | Mus musculus | 19 | Brain |
| E-GEOD-31069 | ArrayExpress | Homo sapiens | 4 | MCF-7 |
| E-GEOD-31225 | ArrayExpress | Mus musculus | 8 | Fibroblast |
| E-GEOD-31667 | ArrayExpress | Mus musculus | 4 | Fibroblast |
| E-GEOD-32055 | ArrayExpress | Mus musculus | 12 | Brain |
| E-GEOD-32109 | ArrayExpress | Homo sapiens | 6 | Lymphoma |
| E-GEOD-32253 | ArrayExpress | Homo sapiens | 4 | Breast milk |
| E-GEOD-33584 | ArrayExpress | Homo sapiens | 4 | Fibroblast |
| E-GEOD-33665 | ArrayExpress | Homo sapiens | 30 | Cancer |
| E-GEOD-34494 | ArrayExpress | Homo sapiens | 3 | Fibroblast |
| E-GEOD-36236 | ArrayExpress | Homo sapiens | 31 | Cancer |
| E-GEOD-37616 | ArrayExpress | Homo sapiens | 35 | Cancer |
| E-GEOD-37686 | ArrayExpress | Homo sapiens | 10 | ESCs |
| E-GEOD-37710 | ArrayExpress | Homo sapiens | 3 | Cancer |
| E-GEOD-38916 | ArrayExpress | Homo sapiens | 4 | HEK293 |
| E-GEOD-39086 | ArrayExpress | Homo sapiens | 4 | HEK293 |
| E-GEOD-39162 | ArrayExpress | Homo sapiens | 15 | Cancer |
| E-GEOD-39841 | ArrayExpress | Homo sapiens | 34 | Cancer |
| E-GEOD-40819 | ArrayExpress | Homo sapiens | 14 | Placenta |
| E-GEOD-44568 | ArrayExpress | Homo sapiens | 6 | HEK293 |
| PRJDB2583 | ENA | Homo sapiens | 10 | Cancer |
| PRJDB2585 | ENA | Homo sapiens | 3 | Saliva |
| PRJDB2675 | ENA | Homo sapiens | 6 | Multiple |
| PRJDB2807 | ENA | Mus musculus | 8 | Liver |
| PRJDB745 | ENA | Homo sapiens | 2 | Cancer |
| PRJEB6759 | ENA | Mus musculus | 2 | MEFs |
| PRJNA176037 | ENA | Mus musculus | 3 | Liver |
| PRJNA177892 | ENA | Homo sapiens | 3 | Cancer |
| PRJNA178259 | ENA | Homo sapiens | 2 | Cancer |
| PRJNA190003 | ENA | Mus musculus | 1 | Brain |
| PRJNA193184 | ENA | Mus musculus | 5 | Germ cells |
| PRJNA198453 | ENA | Mus musculus | 3 | MSCs |
| PRJNA200090 | ENA | Mus musculus | 3 | Bone |
| PRJNA210360 | ENA | Homo sapiens | 2 | Blood |
| PRJNA218007 | ENA | Homo sapiens | 10 | Lymphoma |
| PRJNA218834 | ENA | Homo sapiens | 2 | Tuberculosis |
| PRJNA219216 | ENA | Mus musculus | 2 | Liver |
| PRJNA219428 | ENA | Homo sapiens | 13 | Plasma |
| PRJNA219492 | ENA | Homo sapiens | 1 | Liver |
| PRJNA222704 | ENA | Mus musculus | 4 | Cancer |
| PRJNA232648 | ENA | Mus musculus | 8 | Brain |
| PRJNA258408 | ENA | Mus musculus | 12 | Serum |
| PRJNA262814 | ENA | Mus musculus | 5 | Serum |
| PRJNA267543 | ENA | Mus musculus | 1 | Soleus muscle |
| PRJNA80147 | ENA | Mus musculus | 18 | Multiple |
| **Total** | | | **461** | |

### 3.2.1 Global analysis of microRNA expression

In order to validate the initial results and to evaluate how well the automated small RNA analysis performs we normalise the count data using DESeq2 (Love et al., 2014), providing each dataset as a distinct condition at the design formula of the DESeq2 normalisation method. We then perform sample-to-sample unsupervised clustering based on co-expression correlation analysis. This allows us to explore both the sample to sample variation of miRNAs and to identify clusters of miRNAs which are significantly overrepresented in certain datasets. Additionally, it allows us to identify groups of samples with very similar miRNA profiles. Our aim is hence to explore miRNA expression across this heterogeneous pool of data and to characterise patterns among datasets of similar or different conditions. This analysis (Figures 3.2 and 3.3) clearly demonstrates clustering of both miRNAs and samples across the datasets.

For miRNAs, the data clearly show a disparity between highly tissue specific and ubiquitously expressed miRNAs (Figure 3.4A). For example, the let-7 family of miRNAs are among the most abundant and widespread detected miRNAs as expected, together with miR-21, miR-191 and miR-92a. Some highly expressed miRNA clusters also show distinct expression, including the miR-106b-25 cluster and the miR-17-92 cluster. Two miRNAs, hsa-miR-147a and hsa-miR-518a-5p, were expressed only in placenta tissue samples, which may imply that their functionality is exclusively influencing embryonic development in humans. Moreover, six miRNAs (hsa–miR–3689b–3p, hsa–miR–5707, hsa–miR–4534, hsa–miR–5583–5p, hsa–miR–3529–3p, hsa–miR–603) are expressed only in a particular dataset from lymphoma cell lines. The miR-302 cluster, thought to be important for pluripotency and cell-cycle regulation was among the most specifically expressed clusters, being predominantly expressed only in ESCs and in brain cancer. Overall, however these data are complex and it is convenient to instead perform pairwise clustering of miRNAs and samples separately to better detect significant commonalities and differences between miRNAs in one analysis and samples in the second analysis. This functionality is available in the web-based interface of *miratlas*.

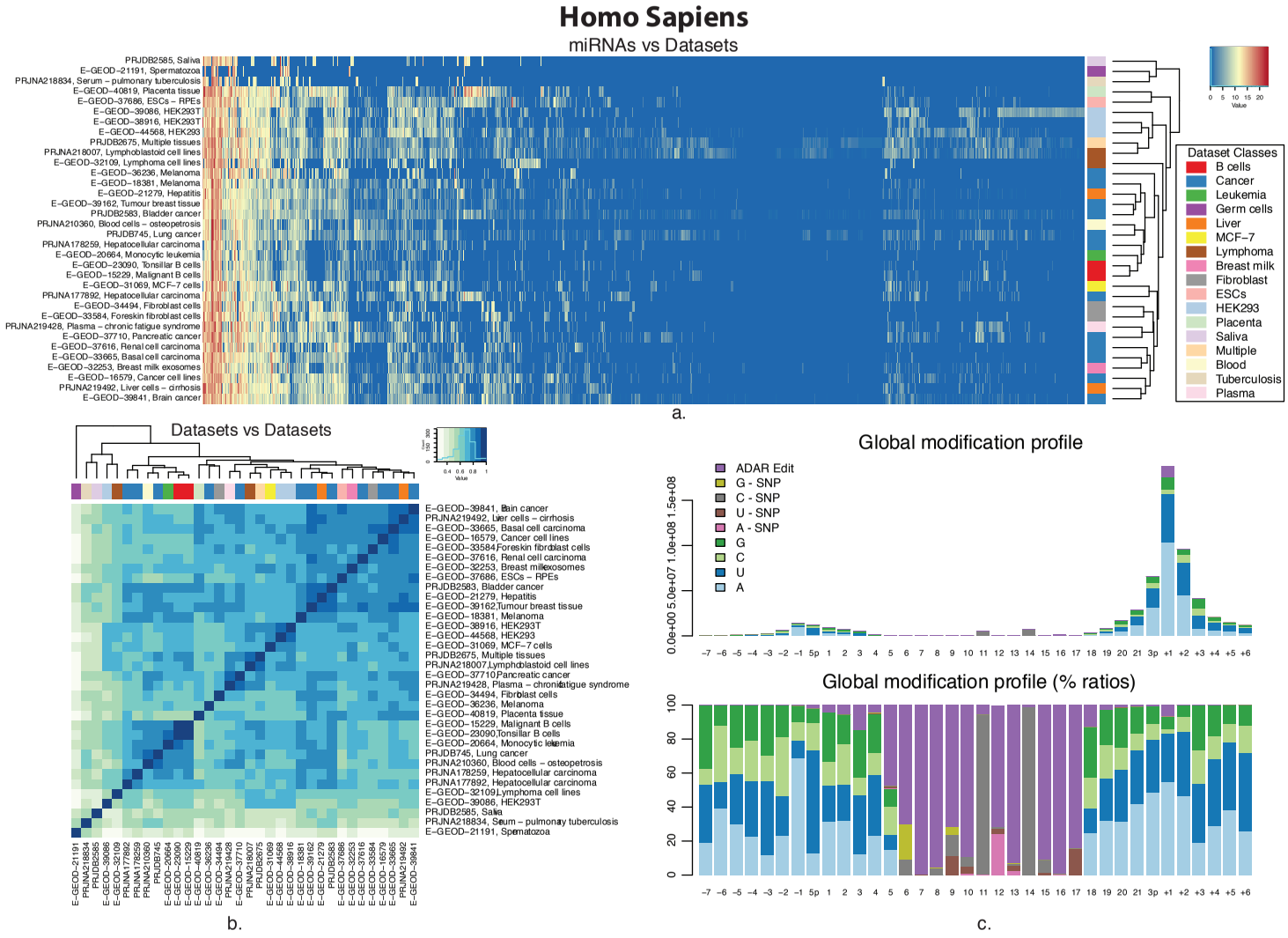



**Fig. 3.2** Global miRNA expression and modification profiles across all 34 human datasets. (a): miRNA expression profile, (b): Sample to sample clustering based on the global miRNAs expression, (c): Aggregate modification profiles from the human datasets positioned with reference to a mature miRNA sequence.

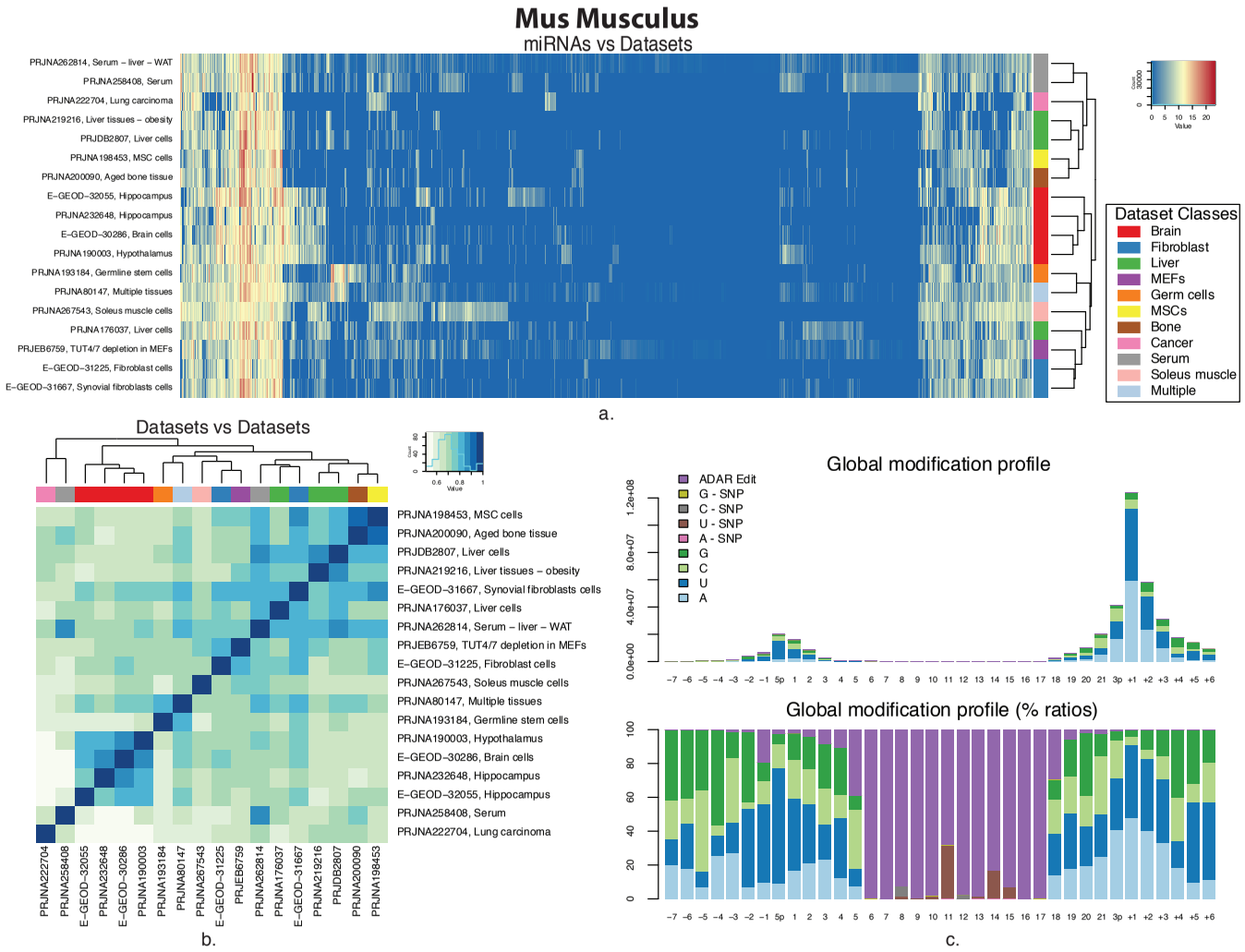



**Fig. 3.3** Global miRNA expression and modification profiles across all 18 mouse datasets. (a): miRNA expression profile, (b): Sample to sample clustering based on the global miRNAs expression, (c): Aggregate modification profiles from the mouse datasets positioned with reference to a mature miRNA sequence.

For sample to sample correlations (Figures 3.2, 3.3) we observe specific groups of tissues and conditions clustering together for example cancer cell lines, B-cells and other similar tissues. For some tissues and cell types multiple experiments from different sources are available. These would ideally have extremely correlated results with differences being explained by differing NGS platforms or experimental strategies. We observe on average 0.79 Pearson correlation of miRNA counts across 7 human datasets where the same tissue or cell type has been profiled (0.82 respectively across 11 samples in Mus Musculus). In contrast, taking random comparisons of different tissues resulted in an average correlation of (0.68 in human and 0.69 in mouse). Clearly, although RNA extraction protocols, sequencing platform and sample treatment account for variation between samples from the same tissue, the correlations remain highly significant ($P \leq 0.018$ for human and $P \leq 0.005$ for mouse).

Three sample types show much lower expression than others (Saliva, Spermatozoa and Serum from pulmonary tuberculosis). These samples do not cluster effectively as they are difficult to normalise due to low sequence counts. In these cases it is likely that the correlation observed is spurious and primarily due to low-counts and/or contamination with RNA degradation products. However, the spermatozoa sample likely has low counts due to the previously observed paucity of small RNAs detectable in sperm (Krawetz et al., 2011; Suh et al., 2010). Clearly, one of the most defined features of the miRNA expression level correlation within Human and Mouse is due to the fact that many miRNAs are co-expressed from the same host transcript. We next sought to explore the expression of miRNAs while taking into consideration their genomic context and likely transcriptional unit.

### 3.2.2 microRNA Clusters derived from genomic proximity

It is well known that many groups of miRNAs are encoded by a single transcript (coding or non-coding). These miRNA clusters are usually predicted by virtue of their close proximity on the genome. Previous computational studies have suggested that miRNA hairpins lying within 10kb (Saini et al., 2007) are likely to be co-transcribed. We sought to update these findings from earlier studies, based on EST and cDNA data, with the data described here. Additionally we use both the genomic location and also miRNA co-expression analysis to re-evaluate these predictions and to generate novel miRNA clusterings. For this analysis we assess the accuracy of genomic clusters of miRNAs predicted using different genomic distance thresholds and miRNA co-expression as a measure of their co-regulation.

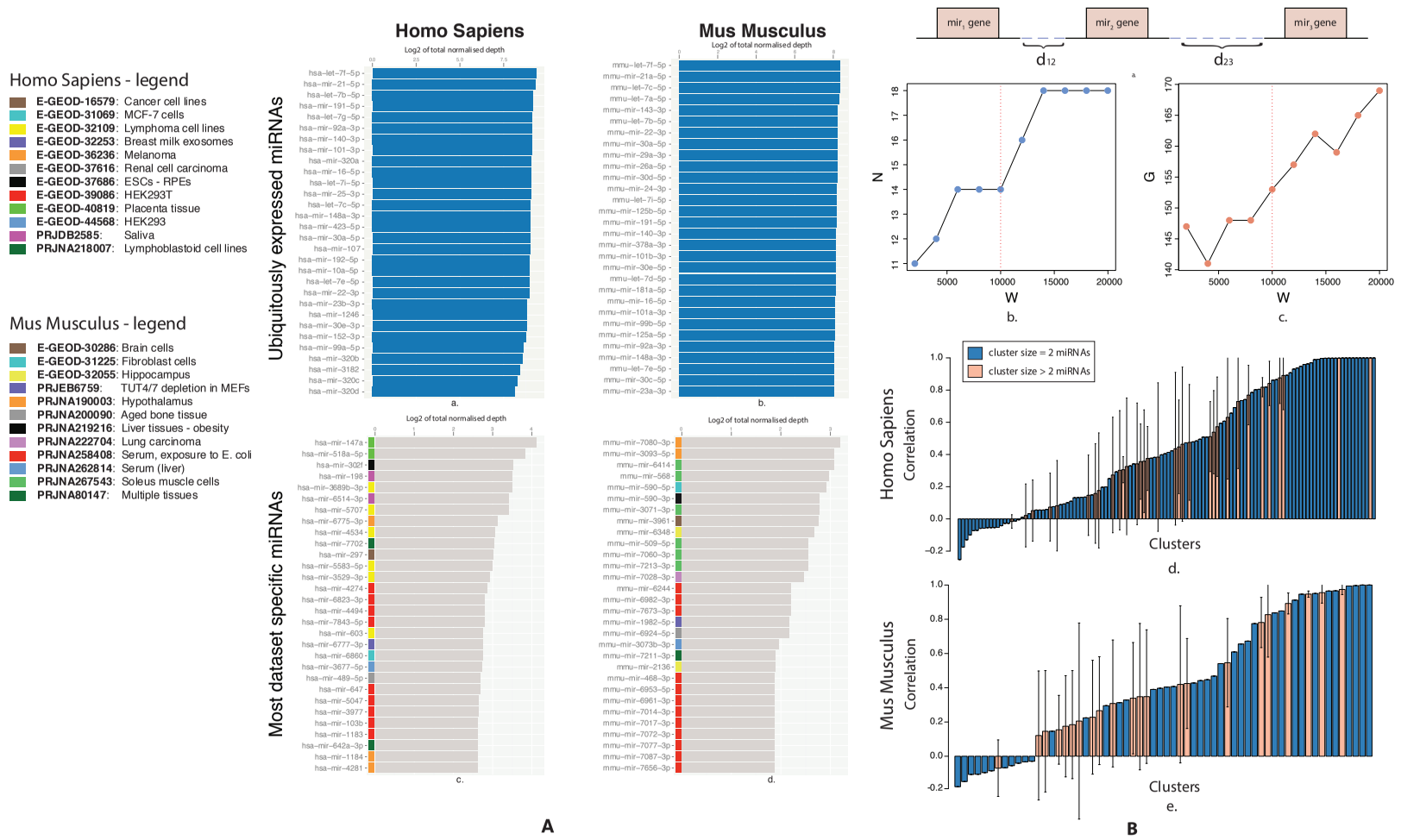



**Fig. 3.4** MicroRNA expression prevalence and genomic clusters definition. A: List of top-30 most expressed miRNAs, which are present in all (a) human and (b) mouse datasets. List of top-30 most expressed miRNAs, present only in a single (c) human or (d) mouse dataset, B: Genomic clusters definition, window selection and expression analysis. (a): 3 miRNA genes located at different positions at the genome with a distance of d12 and d23 kb, respectively, between them. (b): Number of clusters (N) with negative correlation for different cluster window sizes (W). (c): Number of genomic clusters (G) for different cluster window sizes (W). (d), (e): Average intrinsic correlation of miRNA expression within all human and mouse genomic clusters, using a cluster window of 10kb.

We first define all possible miRNA genomic clusters using a custom window of size $W$. Genomic clusters are defined as follows (Figure 3.4B.a): let $mir_1$, $mir_2$ and $mir_3$ be three miRNA genes in neighbouring locations on the genome without any other miRNA genes interfering at the genomic space between $mir_1$ and $mir_3$. Then, $mir_1$, $mir_2$ belong to the same genomic cluster ($GC_1$) if and only if: $d_{12} \leq W$. $mir_3$ also belongs to $GC_1$ if and only if $d_{23} \leq W$ ($d_{13}$ may be greater than $W$ but it will be less than or equal to $2W$). Thus, a genomic cluster may contain pairs of miRNAs whose distance is greater than W but for each miRNA there is at least another miRNA in that cluster that is closer to it less than $W$ base pairs. The $W$ parameter has been selected as large as possible while still retaining the number of clusters with negative intrinsic correlation at a relatively low level (Figure 3.4B.b,c). Based on these criteria and the relevant literature (Griffiths-Jones et al., 2008; Saini et al., 2007), we assign 10k bp as our window size for further analysis. This value produces a total of 153 genomic clusters in human and 92 clusters in mouse.

After the genomic clusters have been constructed, we calculate the average correlation of miRNA co-expression within each genomic cluster (Figure 3.4B.d,e). The number of clusters with positive intrinsic correlation compared to those with negative correlation is statistically significant ($P \leq 10^{-5}$), based on a model that is constructed as the average consensus of 10 runs with random genomic cluster assignments to the miRNAs of our study. We additionally observe that 33.3% of all genomic clusters in human datasets demonstrate a significant average intra-cluster correlation of $> 0.7$ ($P \leq 2.7 \times 10^{-6}$). Interestingly, there are 18 clusters in the human datasets and 12 in the mouse datasets that have non-significant negative correlation values (−0.3 to 0.0). In these instances the small RNAs detected likely are transcribed from separate transcriptional units, products of alternative splicing, possibly mis-annotated RNA degradation products or under some other form of complex regulation. One interesting example with poor expression correlation is the cluster containing hsa-miR-1306 and hsa-miR-3618. These miRNAs are products of the DGCR8 transcript with the miR-3618 hairpin present in the 5' UTR being processed by the microprocessor complex as part of DGCR8's complex transcriptional control mechanism (Triboulet et al., 2009).

### 3.2.3 Clusters derived from miRNA co-expression

Another way to explore the clustering of miRNAs is to look for functional clustering of miRNAs based solely on their co-expression. The assumption here is that miRNAs with high expression correlation are likely to be involved in similar biological systems. We expect that clusters defined in this manner should show considerable overlap with clusters derived from the genomic proximity analysis above. However, we may also be able to

identify groups of miRNAs encoded by transcripts at different genomic loci that still exhibit correlated expression of their host transcripts and may well be functionally linked. In order to generate all miRNA functional clusters, we first got miRNA expression counts with loci-specific information at the genomic level using *SequenceImp* (Davis et al., 2013). We then created a correlation matrix with the co-expression of all miRNAs detected in this study. This matrix defines a weighted graph and weights of its edges correspond to the correlation of expression between pairs of miRNAs. We then clustered this graph using *MCL* (van Dongen and Abreu-Goodger, 2012), setting the value of the Pearson filtering threshold to 0.8 and the inflation parameter to 1.4.

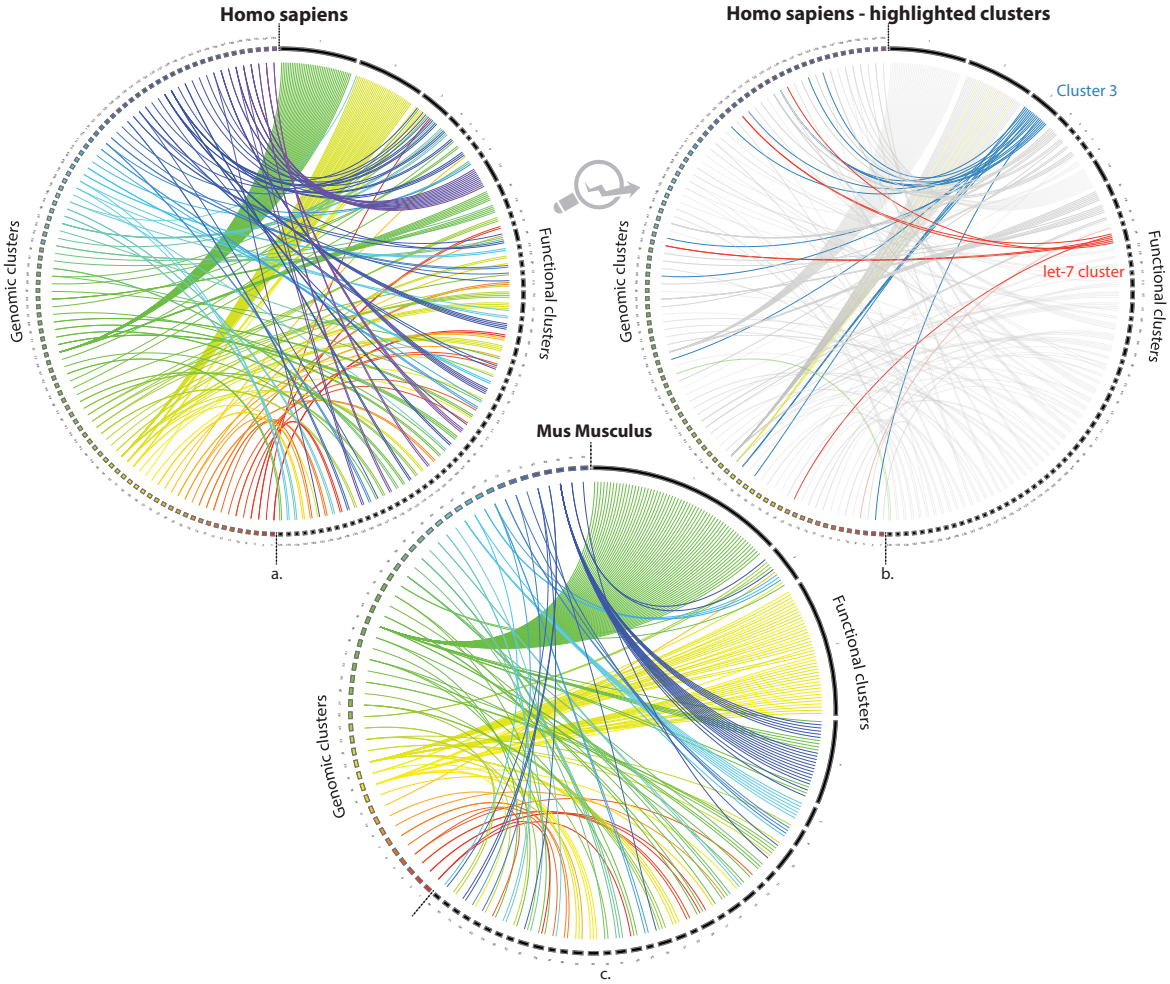



**Fig. 3.5** Associations of functional miRNA clusters with the respective genomic clusters. (a): Human samples, (b) Human samples, highlighting two clusters of co-expressed miRNAs with distant genomic origins, (c): Mouse samples. Each functional cluster is denoted by a black coloured arc with a numeric id. The length of the arc is proportional to the size of the cluster it represents. MicroRNAs that don't have any genomic cluster assignment have been omitted from this analysis for the sake of clarity of the figure. Genomic clusters correspond to the arcs of fixed length, coloured with a non-black hue, and they are sorted in a clockwise order based on their proximity at the genome.

As expected, results show (Figure 3.5) significant overlap between clusters derived from proximity and those derived by expression correlation. However, we also observe a subset of functional links between groups of miRNAs expressed at different genomic loci with significant expression correlation. For example transcriptional cluster 3 (Figure 3.5a) is comprised of a number of genomic clusters including those on chrX, 19 and 13. Close inspection of these transcriptionally linked clusters (Figure 3.6) indeed indicates a preponderance of EGR1 transcription factor motifs, coupled with SP1 and NRF1. These results indicate that the high degree of transcriptional correlation observed between these three genomic clusters is a result of their being driven by the same transcriptional inputs. The high majority of the rest of transcriptional clusters with divergent genomic origin content contain either miRNA paralogues or let-7 family miRNAs, in both human and mouse.

### 3.2.4 Calling and analysis of the prevalence of microRNA modifications

In the past few years it has been widely demonstrated that miRNAs go through post-transcriptional alterations that can modify their 3' ends, mainly via mono- or poly-uridylation (Heo et al., 2012, 2009). Such epi-transcriptomic alterations can have tremendous regulatory impact including how the small RNA machinery in the cell processes these molecules or whether or not they are degraded. In this study, we present for the first time a global profile of miRNA modifications occurring at both 3' and 5' ends. In order to identify the modifications in both ends of each miRNA we have employed additional analysis steps where all primary miRNA sequences are mapped against miRNA precursors using *Chimira* (Vitsios and Enright, 2015). *Chimira* scans the aligned regions in order to detect bases in the miRNA sequence that are not encoded in the genomic sequence. All extracted modification patterns are associated with the exact location of the modification relative to the original sequence. These unalignable nucleotides can be any of the following classes: i) base-calling errors, ii) single nucleotide polymorphisms or iii) post-transcriptional miRNA modifications (e.g. via tutases). Base-calling errors are pseudo-random, platform-dependent and are more likely to occur at the 3' end of a sequencing read, although at relatively low frequencies. SNPs are easier to detect as they will be present in a significant fraction of all reads observed. Finally, modifications such as uridylation or ADAR editing can be detected due to their being highly skewed towards particular modifications (e.g. mono-U, poly-U or A $\rightarrow$G).

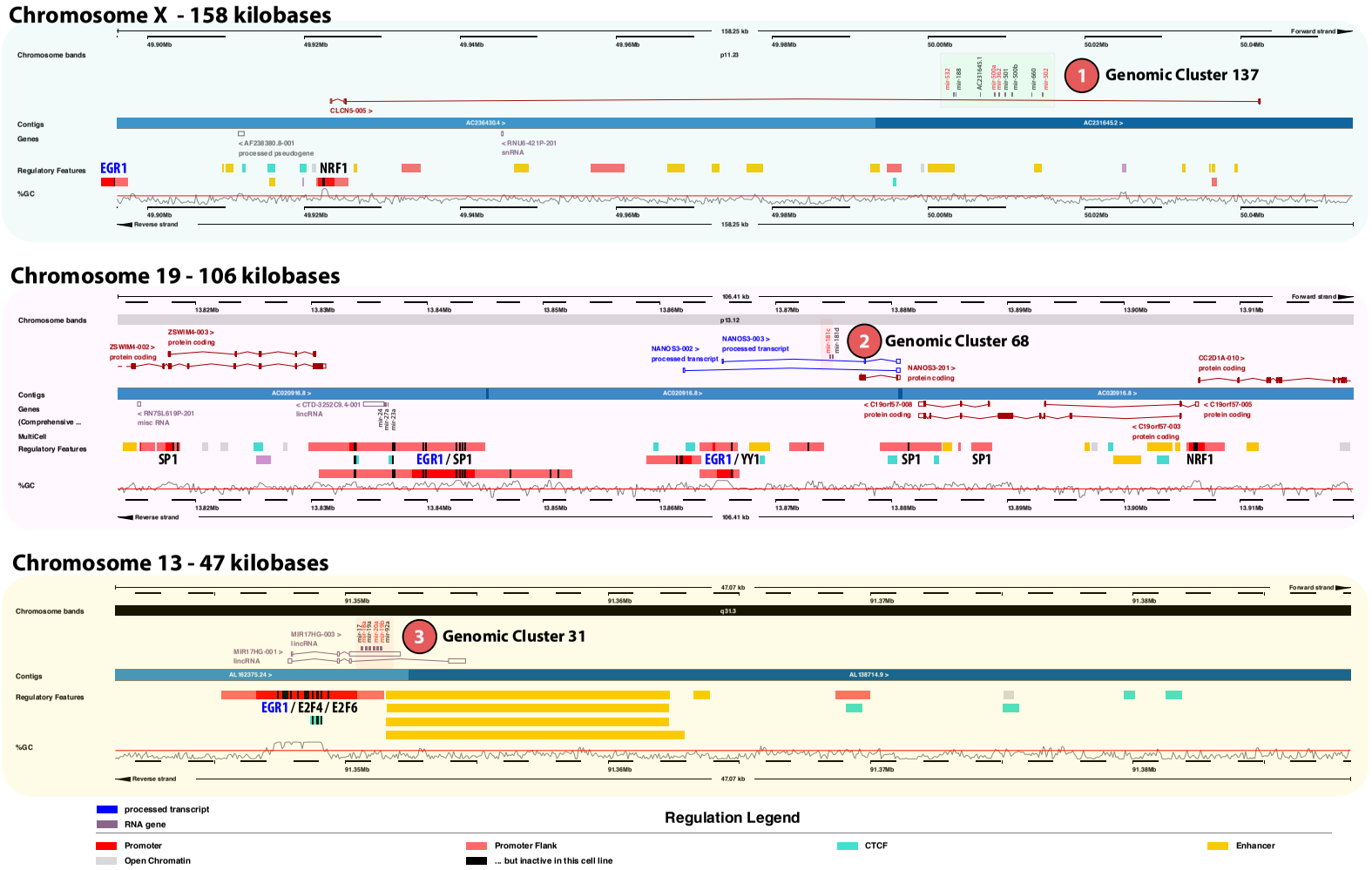



**Fig. 3.6** miRNA *functional* cluster from human (i.e. cluster of miRNAs that are co-expressed across the examined cell types/conditions) whose miRNAs originate from distant genomic locations, and specifically chromosomes 13, 19 and X. Transcriptional correlation between those miRNAs can be attributed though to co-regulation by similar sets of transcription factors. In this case, transcription factors EGR1, coupled with SP1 and NRF1 seem to regulate the expression of these miRNAs.

Overall, we find that 3' modifications are far more prevalent than detected 5' modifications (Figures 3.2, 3.3). In total, 95 human (4.4%) and 142 mouse (7.8%) miRNAs showed on average significant levels of 3' modification (i.e. more than 25%). Similarly, 23 human (1.1%) and 24 mouse (1.3%) miRNAs showed on average significant levels (i.e. more than 25%) of 5' modification. Mono and dinucleotide additions are the most common modifications, although longer modifications were observed too, albeit at lower frequencies (Figure 3.7a,b). In both Human and Mouse we observe a preponderance of Adenosine and Uracil modifications (Figures 3.2, 3.3) suggesting that both adenylation and uridylation by TUTases are likely the primary modifications made to miRNAs at least in animal systems. Both cytoplasmic adenylation by GLD-2 (Katoh et al., 2009) and terminal uridylation by Tut4/Tut7 have been reported before as important for miRNA stability and degradation (Heo et al., 2012). However, in this study we performed the first large scale detection and analysis of these events across animal tissues.

In order to investigate the significance of the presence of 3' Guanine and Cytosine modifications, we performed an analysis in 12 human samples from mRNA-Sequencing experiments that were derived from Illumina Sequencing instruments to identify whether these G:C modifications may result from known sequencing biases present in the instrument. To evaluate this we assume that G:C sequencing biases for mRNA samples will be largely similar to those obtained from small RNA sequencing. However, we would not expect any terminal modifications to occur within sequencing reads derived from exonic mRNA sequence, any non-genomic nucleotides observed are more likely to be sequencing errors. In order to extract potential sequencing bias profiles, we filtered the reads from 12 human samples sequenced by Illumina, retaining only those that were at least 10nt shorter than the maximum length among all the reads, which is the length that occurs more frequently among the reads of the sample. This filtering process allows us to retain only the reads that may correspond to 3' exons of actual mRNA transcripts. The filtered reads were then aligned against the 3' human exons of the reference database we had constructed allowing the identification of sequence artefacts that are appended to the 3' end of the transcripts and probably represent sequencing artefacts.

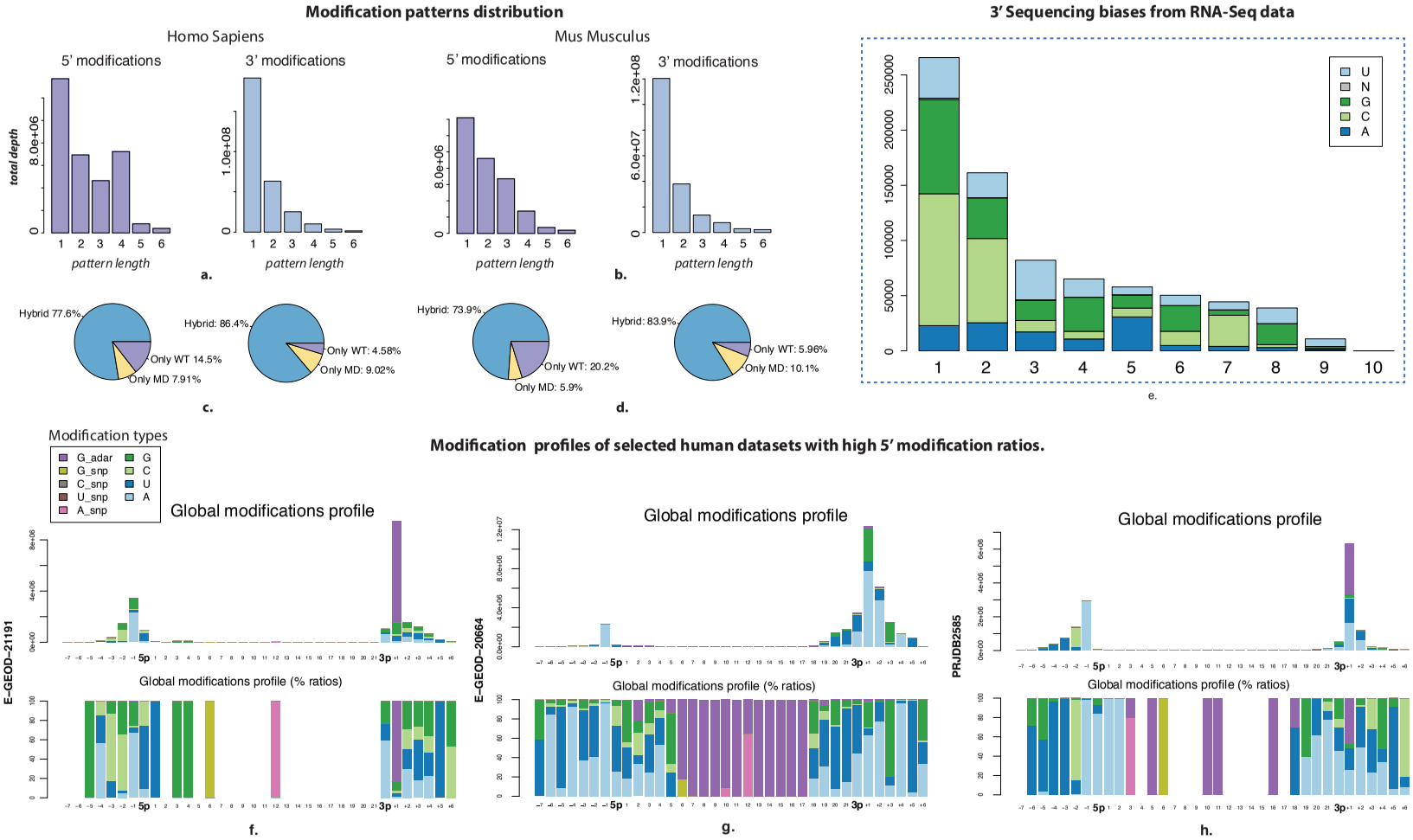



**Fig. 3.7** Summary of modification patterns, sequencing biases and stand-out datasets with high levels of 5' modifications. (a), (b): Normalised depth of 5'/3' modification patterns for various lengths across all human and mouse datasets (not to scale). (c), (d): miRNA distribution into Wildtype (WT), Modified (MD) and Hybrid (WT & MD) variant classes across all (c) human and (d) mouse datasets. (e): Sequencing bias profiles at 3' end, induced by Solexa sequencing from 12 human RNA-Seq samples. (f)-(h): Modification profiles of selected human datasets with high 5' modification ratios. (f): E-GEOD-21191, spermatozoa samples, (g): E-GEOD-20664, monocytic leukemia samples and (h): PRJDB2585, saliva samples.

The derived profile of sequencing biases is very rich in Gs and Cs (Figure 3.7e) and greater than 65% of all observed errors for mono and dinucleotide errors. With regards to the datasets that have been used in our large-scale miRNA analysis, some of them are derived from Illumina Sequencing instruments, while others are derived from different types of instruments and another significant percentage among them do not provide in their annotation any information about the sequencing instruments that have been used during the experiment. The lack of annotation makes it difficult to computationally model and filter these likely G:C biases, however the data suggest that for the most part they are largely sequencing artefacts. Besides, this strongly suggests that the observed A:U enrichments are highly unlikely to be due to such sequencing artefacts and instead represent valid biological effects.

The prevalence of 5' modifications is far lower than that observed for 3' changes. Although some tRNAs are known to have 5' modifications, we are not aware of any reported biochemical experiments of 5' modification of small RNAs. For both human and mouse however a preponderance of 5' A and U modifications are observed but are extremely rare as compared to 3' modifications. It has already been reported that the 5' ends of miRNAs are generally post-processing stable in contrast with the 3' ends (Hibio et al., 2012). Additionally, addition of 5' nucleotides would dramatically alter the targeting of a miRNA loaded into the RISC complex. This may explain the lower count numbers and also the lower variability of the 5' modifications in comparison with the 3' modifications.

However, certain datasets, among those from spermatozoa, monocytic leukemia and saliva samples (Figure 3.7f-h) as well as two cancer datasets (E-GEOD-39841: brain cancer and E-GEOD-36236: skin cancer; profiles available in *miratlas*) exhibit a high ratio of 5' modifications, especially at the first nt upstream to the 5' end. These modifications are capable of redefining the seed region patterns of the modified miRNAs and consequently change the repertoire of the mRNAs that are being targeted by them. This may be affecting the functionality of some or all of these tissues by causing irregularities related with disease conditions. These initial observations though need further investigation in order to truly validate their biological effect. More specifically, it would be imperative to first predict miRNA targets [e.g. via Sylamer (van Dongen and Abreu-Goodger, 2012)] in the datasets with high 5' modification levels. This search should include both the canonical mature miRNA sequences and the 5'-modified ones, since each of them is associated with a different set of mRNA targets, let them be *canonical-targets* and *modified-targets*, respectively. Then, if 5'-modifications were biologically active, we would expect that *canonical-targets*'s expression is increased while expression of *modified-targets* is decreased, both at a statistically significant level. Thus, we propose as future research the extraction of both small

RNA-Seq and RNA-Seq data from the identified datasets (Figure 3.7f-h), predict all relevant miRNA targets, extract mRNA expression data and finally assess the enrichment/depletion levels of mRNAs that are being targeted by the canonical or 5'-modified miRNAs.

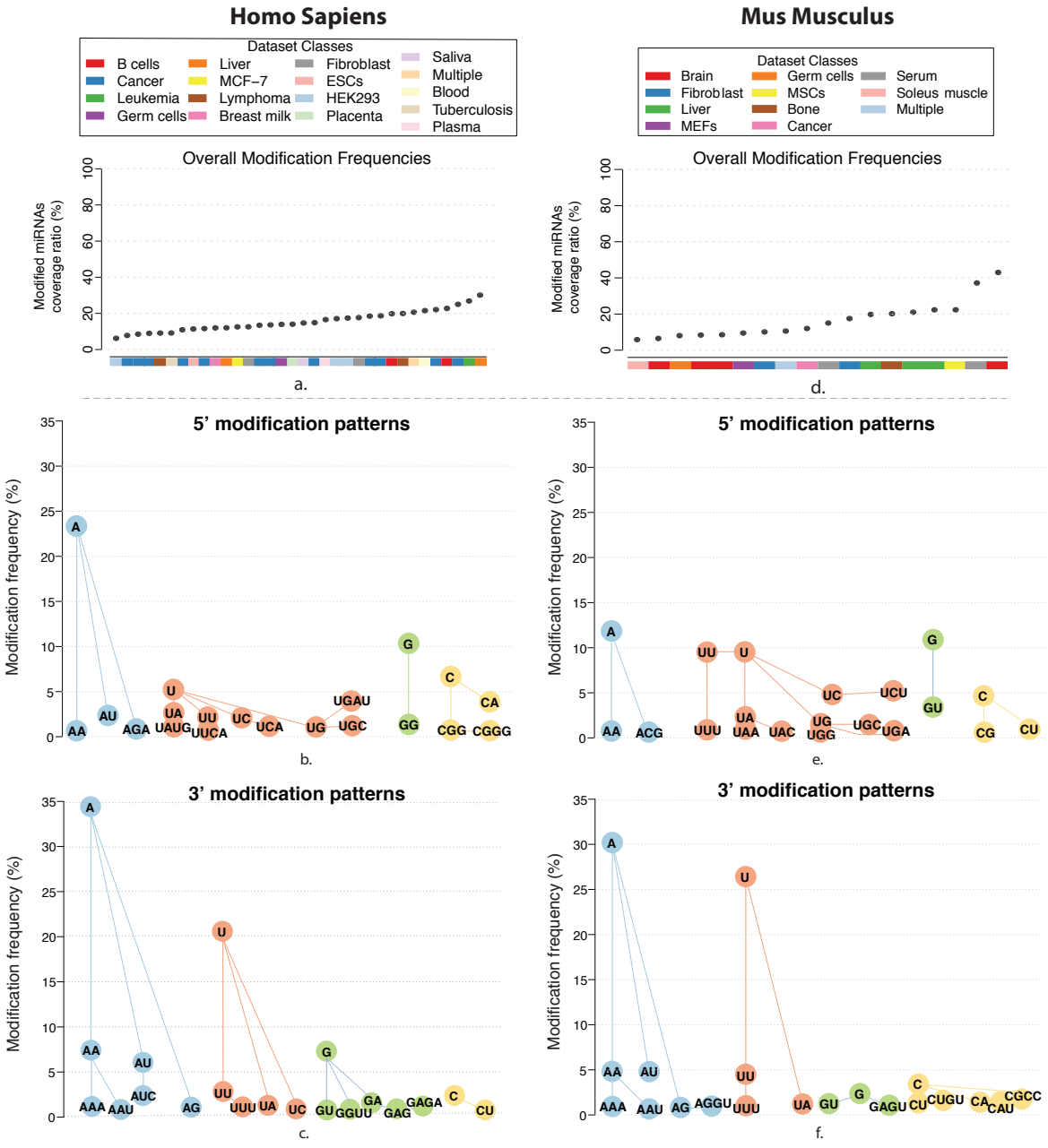



**Fig. 3.8** Overall extent of modification events and most dominant patterns. (a),(d): Modification ratios coverage across all human and mouse datasets. (b), (e): Prevalence of top-20 most frequent modifications patterns at the 5' end of miRNAs in human and mouse. (c),(f): Prevalence of of top-20 most frequent modifications patterns at the 3' end of miRNAs in human and mouse.

In the majority of cases, modifications affect less than or around 20% of the total expression depth while there are very few cases that they reach 30% or 40% of the total depth (Figure 3.8a,d). Moreover, the prevalence of 3' modifications is significantly higher than 5' modifications (Figure 3.8). A strong enrichment for 3' A and U modifications is observed in both human and mouse which is in agreement with previous studies implicating tutase enzymes. For 5' modifications a smaller enrichment is observed for 5' Adenylation, however this enrichment is only observed to be significant in Human samples and a corresponding shift is not observed in mouse. The mild enrichment for 5' Adenosine is puzzling and possibly reflects the presence of 5'Methyl adenosine sites known to be important for primary miRNA processing.

We then explored the distribution of adenylated and uridylated variants across all datasets (Figures 3.9 and 3.10). We focused on the expression of the let-7 family miRNAs and we observed a high resemblance of their modification profiles for these particular variants. Only mir-98 has a markedly different profile, potentially due its lower expression compared to other members of the let-7 family. We also projected the modification distribution of a set of highly expressed miRNAs. Some of these profiles show high similarity with the respective let-7 profiles while others demonstrate a relatively low modification depth (Figures 3.9 and 3.10). This finding suggests that the frequency of modification events is not always associated with miRNA abundance but may be driven by other factors related to a particular condition, cell type or tissue. Several dataset types also tend to cluster together based on the overall expression of different types of variants in both species (Figures 3.9 and 3.10) or the modification frequencies of the most highly modified miRNAs (Figure 3.11). However, although overall expression of let-7 miRNAs is very similar across samples of the same cell type or condition, the expression of individual let-7 variants (e.g. adenlyated, guanylated) seems to deviate even for samples of the same annotation class.

For ADAR editing events, we observe an enrichment for brain in both Human and Mouse (Figure 3.11), in correspondence with previous studies (Blow et al., 2006). Additionally, we observed an enrichment in serum and some cancer samples of non neuronal origin (data available in *miratlas*). The rate of ADAR editing observed in brain samples is 2% and it occurs most predominantly in the seed region of miRNAs, also in line with previous studies. We also observed that two cancer samples from human and mouse have very similar profiles and that is also the case for another pair of serum samples from the two species (data available in *miratlas*). This may imply that ADAR edits for those particular conditions are preserved across these two species.

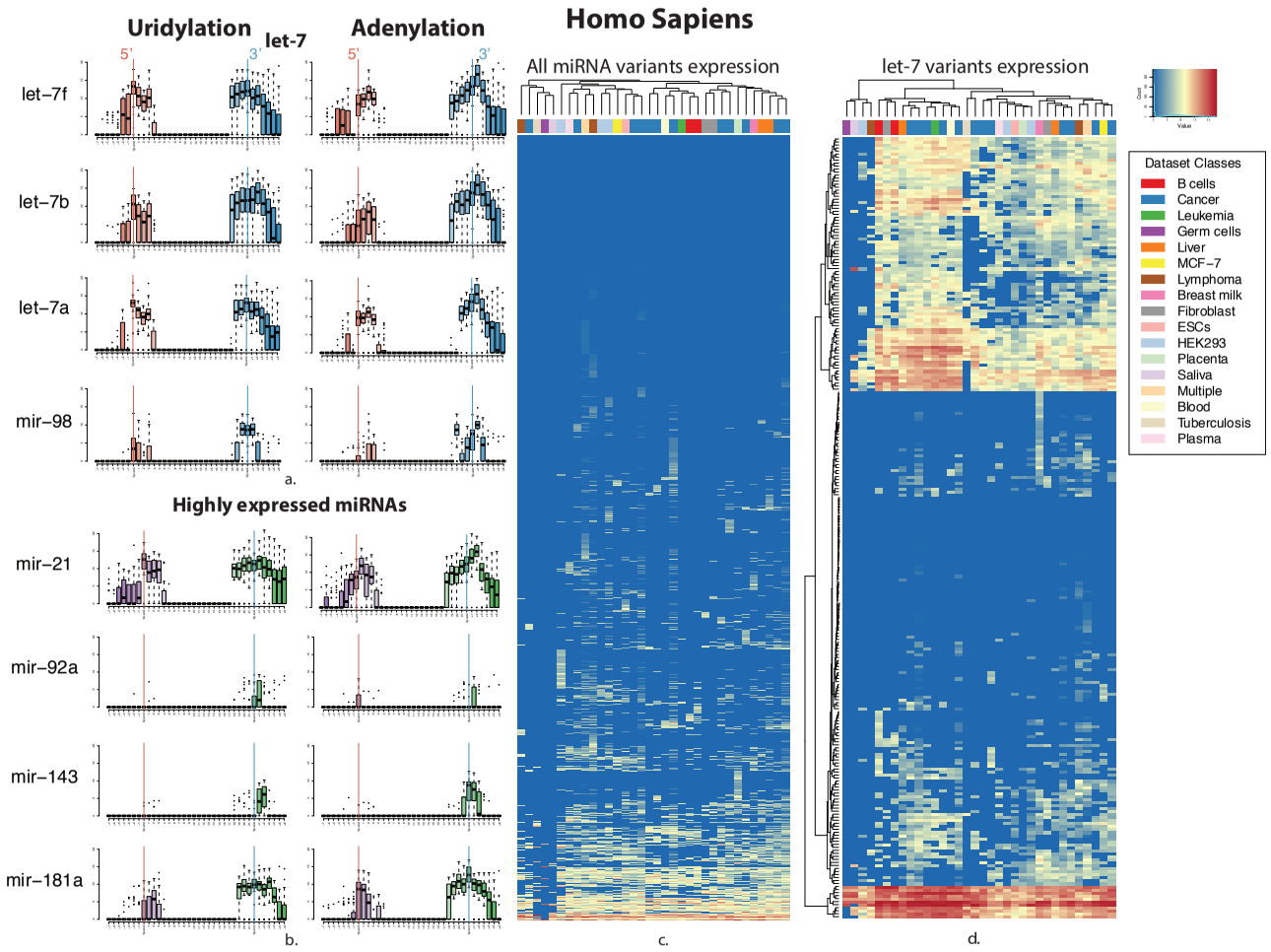



**Fig. 3.9** Modification analysis in human based on collapsed modification variants (i.e. wildtype and mono/poly-: adenylated, uridylated, gunaylated, cytidylated variants). (a): Distribution profiles of uridylated and adenylated miRNA variants for a subset of the let-7 family across all human datasets. (b): Distribution profiles of uridylated and adenylated variants for a set of 4 highly expressed miRNAs across all human datasets. (c): Expression profiles of all miRNA variants with collapsed modifications across all human datasets. (d): Expression profiles of the let-7 family miRNA variants with collapsed modifications across all human datasets.

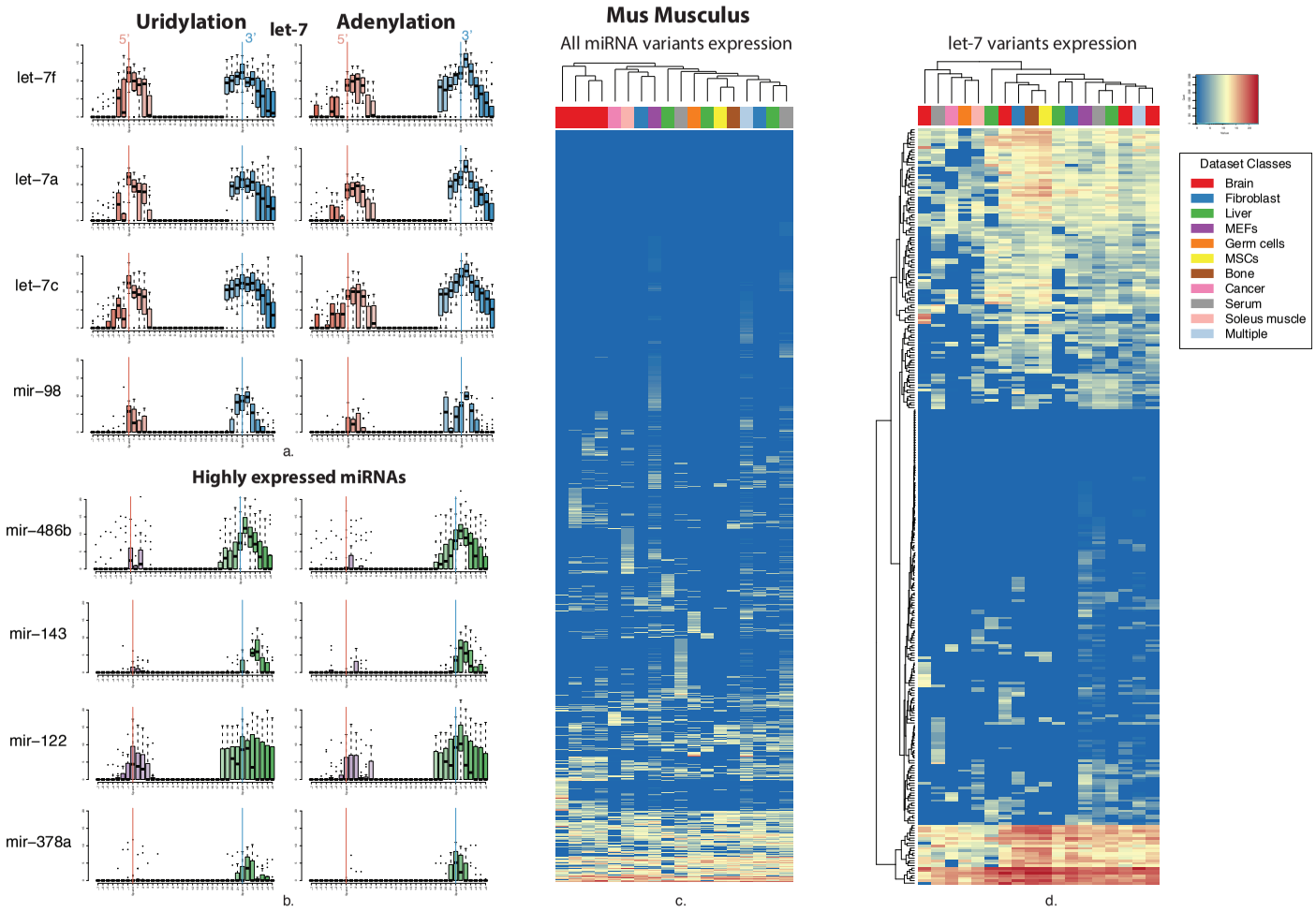



**Fig. 3.10** Modification analysis in mouse based on collapsed modification variants (i.e. wildtype and mono/poly-: adenylated, uridylated, gunaylated, cytidylated variants). (a): Distribution profiles of uridylated and adenylated miRNA variants for a subset of the let-7 family across all mouse datasets. (b): Distribution profiles of uridylated and adenylated variants for a set of 4 highly expressed miRNAs across all mouse datasets. (c): Expression profiles of all miRNA variants with collapsed modifications across all mouse datasets. (d): Expression profiles of the let-7 family miRNA variants with collapsed modifications across all mouse datasets.

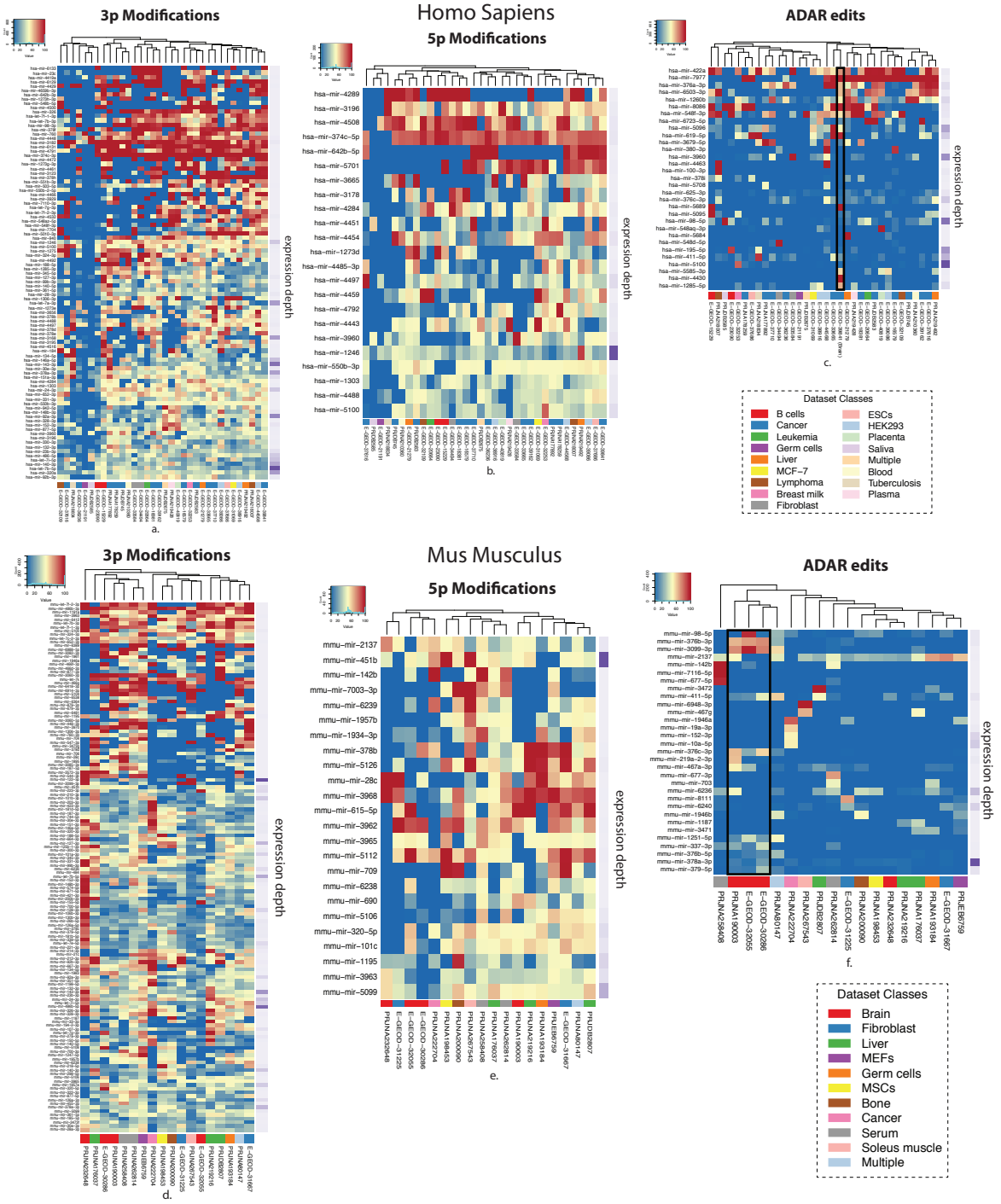



**Fig. 3.11** Modifications distribution for highly modified miRNAs in human/mouse. (a)/(d): 3' modification ratios across all human/mouse datasets for the miRNAs with average edit ratio > 25%. (b)/(e): 5' modification ratios across all human/mouse datasets for the miRNAs with average edit ratio > 25%.. (c)/(f): ADAR edit ratios across all human/mouse datasets for the miRNAs with average edit ratio > 3% and depth over the median total depth of each sample.

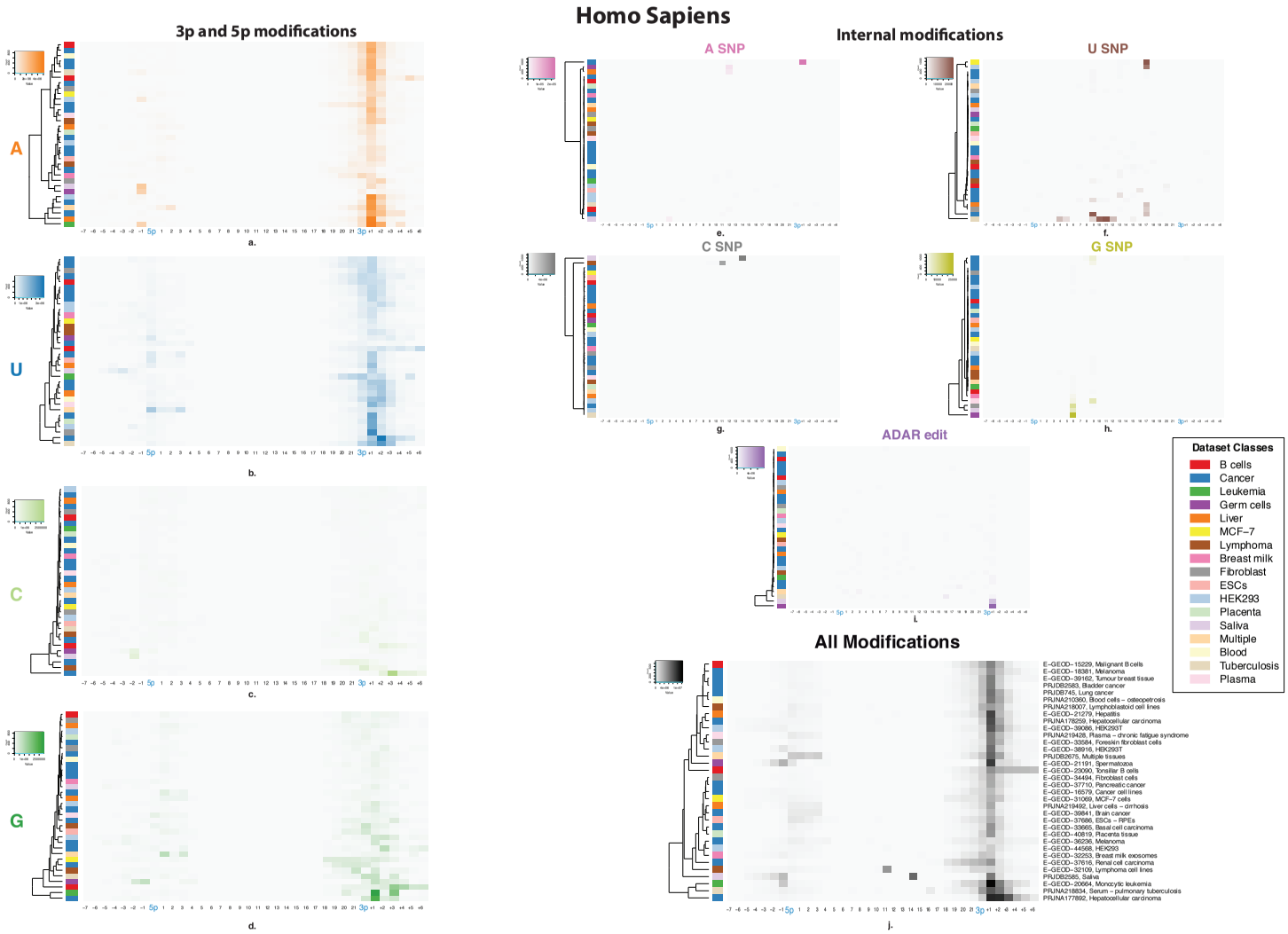



**Fig. 3.12** Aggregate modification profiles across all human datasets for different modification types: (a) Adenylation, (b) Uridylation , (c) Cytidylation , (d) Guanylation , (e) A-SNPs, (f) U-SNPs , (g) C-SNPs, (h) G-SNPs, (i) ADAR edits, (j) All modifications.

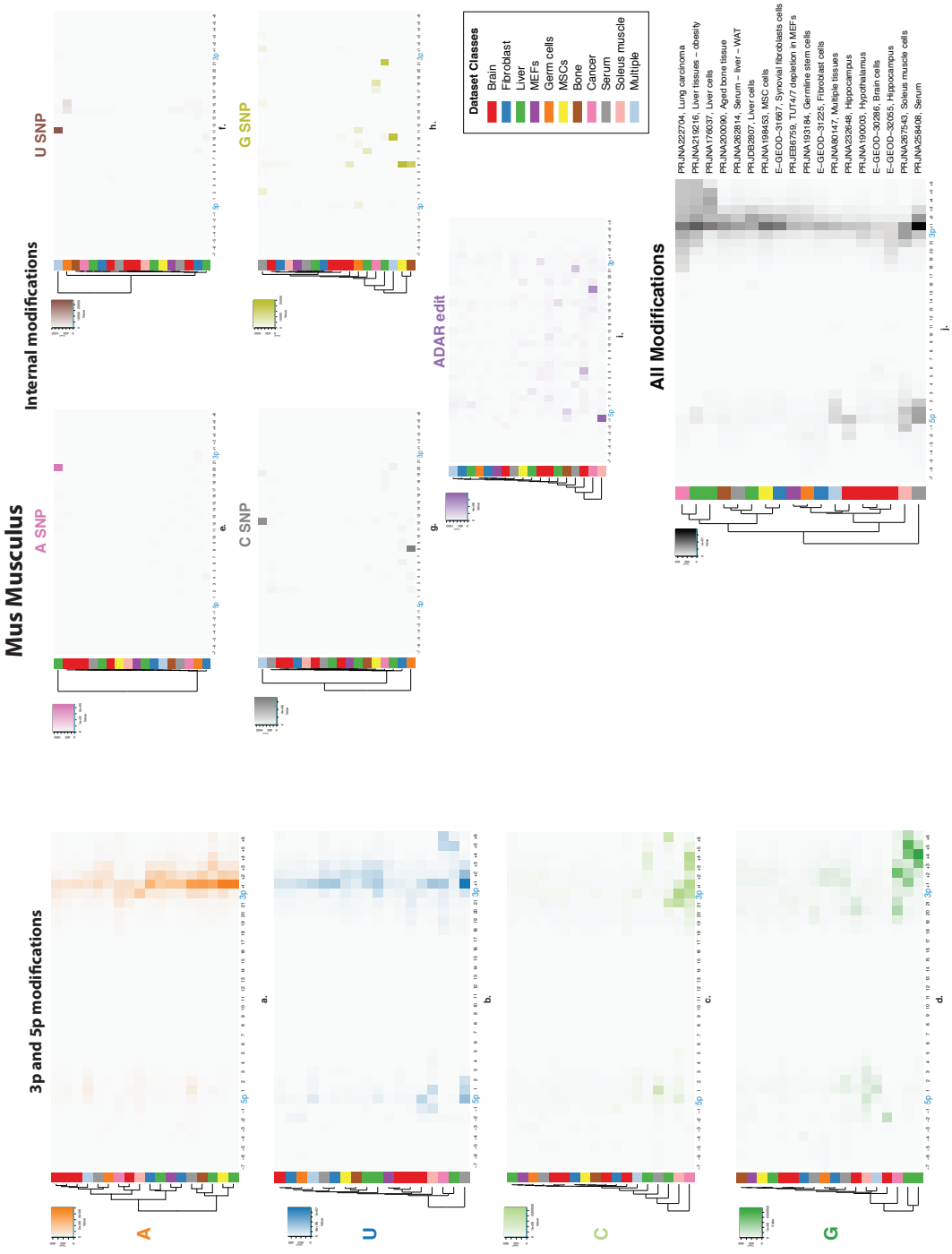



**Fig. 3.13** Aggregate modification profiles across all mouse datasets for different modification types: (a) Adenylation, (b) Uridylation , (c) Cytidylation , (d) Guanylation , (e) A-SNPs, (f) U-SNPs , (g) C-SNPs, (h) G-SNPs, (i) ADAR edits, (j) All modifications.

Finally, we have built the global maps of modification expression for all distinct types of modification and for aggregate variants (Figure 3.12, Figure 3.13). We confirm again that adenylation and uridylation are the most predominant modification types and they tend to occur significantly more frequently at the 3' end of miRNAs , both in human and mouse.

### 3.2.5 MicroRNA strand-specificity analysis and characterisation

The dataset we obtained is extremely useful for exploring other aspects of miRNA biogenesis. In particular, because we obtain sequencing counts for both the 5' and 3' strands from a miRNA precursor we can use these data to globally explore mature strand selection of miRNAs. During the miRNA maturation process in general only one strand of the miRNA duplex is assembled into RNA-induced silencing complex (RISC) while the complementary strand is degraded. This phenomenon has been studied in the past and the prevailing theory is that the asymmetry in the selection of the dominant miRNA strand may be explained by the difference in the stability of the bonds of the miRNA duplex at 5' ends of each strand. This hypothesis has been proved experimentally for a small number of miRNAs (Schwarz et al., 2003). However, there is no global analysis so far that evaluates and models strand selection for miRNAs. We sought to both test these hypotheses and extract a global model of strand-selection for miRNAs based on the Gibbs free energies (ΔG) of the bonds present in the double stranded pre-miRNA.

During the formation of a double stranded RNA molecule, low ΔG values indicate that the reaction can occur spontaneously and lead to a stable form. Conversely, high ΔG values, calculated with reference to a ds-RNA segment, indicate high likelihood for that segment to unwind without the intervention of an external energy source.

For all miRNAs, we calculated the ΔG free energies (in kcal/mol) for short double stranded segments of their hairpin structures around the 5' end of the miRNA from each strand. We tested a variety of definitions for these segments. In each case, the window used for the definition of the segments focuses on a ds-RNA region of the hairpin, starting upstream, downstream or right at the 5' end of each mature miRNA and extending for N ($N \geq 1$) nt overall towards the 3' end of the miRNA. Specifically, we calculated the ΔG for each segment starting at the 5' end of the 5' mature product ($\Delta G_1$) and at the 5' end of the 3' mature product ($\Delta G_2$) and set their difference as $\Delta\Delta G = \Delta G_2 - \Delta G_1$ (Figure 3.14a.i).

Based on expression data from this analysis, all let-7 family miRNAs turn out to be very highly 5'-strand specific. Looking closely at the secondary structure of the let-7 family hairpin precursors, let us assume that the 5' end regions of the 5' miRNA products are more unstable than the corresponding ends of the 3'-miRNA products (e.g. due to prevalence of A:U bonds, gaps or wobbles). So, based on the conventions for the calculated free energies

we used before, we would expect that $\Delta G1 > \Delta G2$, since $\Delta G1$ refers to a more unstable structure. As a result, we would expect that $\Delta\Delta G < 0$ for the 5'-strand specific miRNAs, $\Delta\Delta G > 0$ for the highly 3p-strand specific miRNAs and $\Delta\Delta G \approx 0$ for the non strand specific miRNAs.

In order to test our hypothesis, we classified all miRNAs based on their strand specificity. For this analysis, only miRNAs with two mature products, one for each strand of the hairpin precursor, have been taken into account. We first calculated the expression ratio of each miRNA strand product using the following formula:

$$expression\_ratio_{(arm)} = \frac{counts_{(arm)}}{counts_{(arm)} + counts_{(compl\_arm)}}$$

where:

- $(arm,\ compl\_arm) = (3p,\ 5p)\ or\ (5p,\ 3p)$
- $counts_{(arm)}$ : is the total normalised depth of the *arm* mature miRNA product across all datasets and
- $counts_{(compl\_arm)}$ : is the total normalised depth of the *compl_arm* mature miRNA product, at all possible loci of the genome.

Based on the *expression_ratio* scores calculated using the formula above, we grouped all miRNA precursors into three groups:

1. Highly 5'-strand specific: $a \leq expression\_ratio_{5p} \leq b$
2. Highly 3'-strand specific: $a \leq expression\_ratio_{3p} \leq b$
3. Non strand specific: $0.4 \leq (expression\_ratio_{5p}\ ||\ expression\_ratio_{3p}) \leq 0.6$

for different sets of increasing strand specificity thresholds:
$(a, b) = \{\ (0.7,\ 0.85),\ (0.85,\ 0.93),\ (0.93, 0.97),\ (0.97,\ 1)\ \}$

We then test our hypothesis by calculating the $\Delta\Delta G$ values for all three types of strand specific groups with reference to a different segment of the ds-RNA hairpin structure each time. We have used increasing strand specificity thresholds for the highly 5' and 3' strand specific groups in order to examine if there is any shift in the $\Delta\Delta G$ values as the strand specificity criteria become more stringent. Moreover, we checked if the $\Delta\Delta G$ values from each strand specific group were distinguishable for the other groups implying that free energies calculated for a specific window of a ds-RNA hairpin segment are correlated with

the strand selection process. Gibbs free energies have also been calculated for an additional group of 1000 'random' miRNAs. This group of 'random' miRNAs is formed by selecting randomly 10 non-strand specific miRNAs identified in our study and generating for each of them 100 permutations of their hairpin precursor sequences, permitting only permutations that fold into hairpin-like structures in the end.

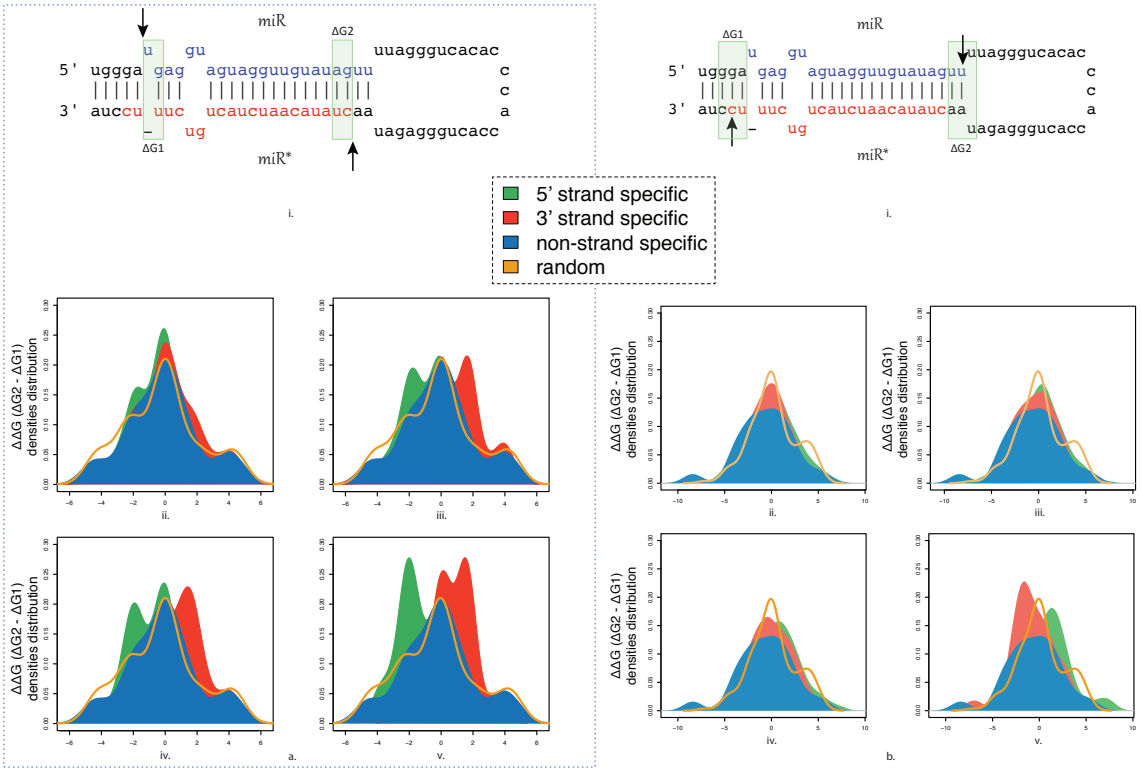



**Fig. 3.14** Separation of strand specificity classes based on free energies difference. (a): Optimal classes segregation based on free energy calculations at the 5' ends of the two potential strand products. (a.i): Window used for the free energy calculations that leads to a clear separation of the strand specificity classes based on the energy difference: ΔΔG = ΔG2 - ΔG1. The start of the window is set to the first nucleotide of the 5' end of both strands and the window extends for an extra nucleotide towards the rest of each miRNA strand. (b): Free energy calculations at a 3nt long ds-RNA segment adjacent to the optimal window. (b.i): 3-nt long window starting with the 2-nt overhang at the 3' end of each strand used for the ΔG calculations in the hairpin sequences. The start of the window is set to the first nucleotide of the 2-nt overhang at the 3' end of each strand and the window extends for another two nucleotides towards the 3' end of the whole hairpin in both cases. (a),(b).ii-v: ΔΔGs densities for the groups of the highly 5'-strand specific, highly 3'-strand specific, non-strand specific miRNAs and random miRNAs, by progressively setting stricter criteria of strand specificity, based on the $expression\_ratio_{5p/3p}$ values (Eq. 1). (ii): $0.7 \leq expression\_ratio_{5p/3p} < 0.85$, (iii): $0.85 \leq expression\_ratio_{5p/3p} < 0.93$, (iv): $0.93 \leq expression\_ratio_{5p/3p} < 0.97$, (v): $0.97 \leq expression\_ratio_{5p/3p} \leq 1$.

After rigorous testing, we identified that there is a strong separation between the highly 5'-strand specific, highly 3'-strand specific and the non strand specifics groups of miRNAs only when the ΔΔG is calculated using a window that contains the first N=2 nucleotides

of each strand (Figure 3.14a.i). In this case, the ΔΔG distribution complies remarkably well with the assumption we have made earlier with regards to the expected ΔΔG values for different types of miRNAs in terms of their strand specificity (Figure 3.14a.ii-v). Additionally, the group of random miRNAs, that is used as a control to examine the variance of ΔΔG across a large number of hypothetical hairpins, follows quite precisely the ΔΔG profile of the non-strand specific miRNAs that originate from real hairpins. Hence, these results indicate that, in general, the stability of the first 2nt at the 5' end of each strand plays the most crucial role for mature miRNA strand selection.

Furthermore, we made another observation that refers to the ΔΔG calculations for a window of 3nt, starting at the 3' end of each miRNA strand and extending for an extra nt in both sides (Figure 3.14b.i). This window contains nucleotides that are not present in the ds-RNA that is extracted to the cytoplasm but exist only in the hairpin precursor. However, we can notice again that there is a quite clear, although milder than in the previous case, separation of the miRNAs based on the strand specificity of their mature products (Figure 3.14b.ii-v). In this case though, ΔΔG distribution for the highly 5'-strand and 3'-strand specific is reversed compared to the previous distribution. That may indicate that the unstable 2nt long ds-RNA segment at the 5' end of each strand is reinforced by the adjacent 3nt long ds-RNA segment that contains the 2nt overhang at the 3' end of the complementary strand. So, the asymmetry of miRNA duplexes in their 5' ends is balanced by an opposite asymmetry in their 3' ends which contributes to the preservation of the hairpin structure energy equilibrium.

### 3.2.6 Detection of mis-annotated miRNAs

The data obtained clearly shows that miRNAs have distinct patterns of expression, modification, strand-selection and genomic localisation. The many hundreds of thousands of miRNA to precursor alignments obtained from NGS data also allow us to detect miRNAs which do not appear to illustrate the hallmarks of well characterised miRNAs. Previous reports have described many such molecules present in the miRBase database and suggest they represent mis-annotated sequences likely derived from other non-coding RNAs or degradation products of longer molecules (e.g. tRNAs).

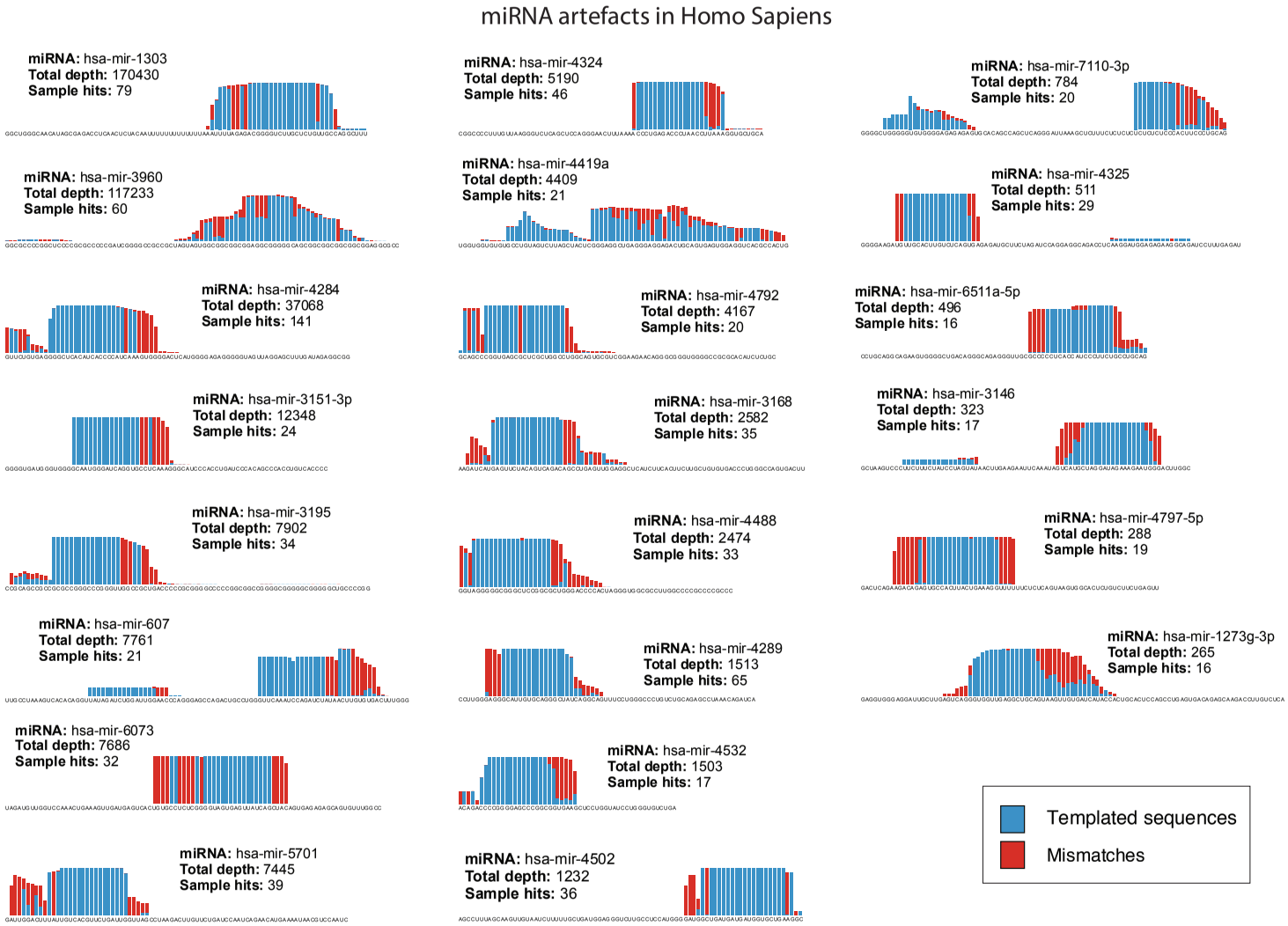



**Fig. 3.15** Coverage profiles of miRNAs that have been detected in human samples as potential artefacts. miRNAs with reads in less than 15 samples (5% of all samples) have been excluded from this analysis.

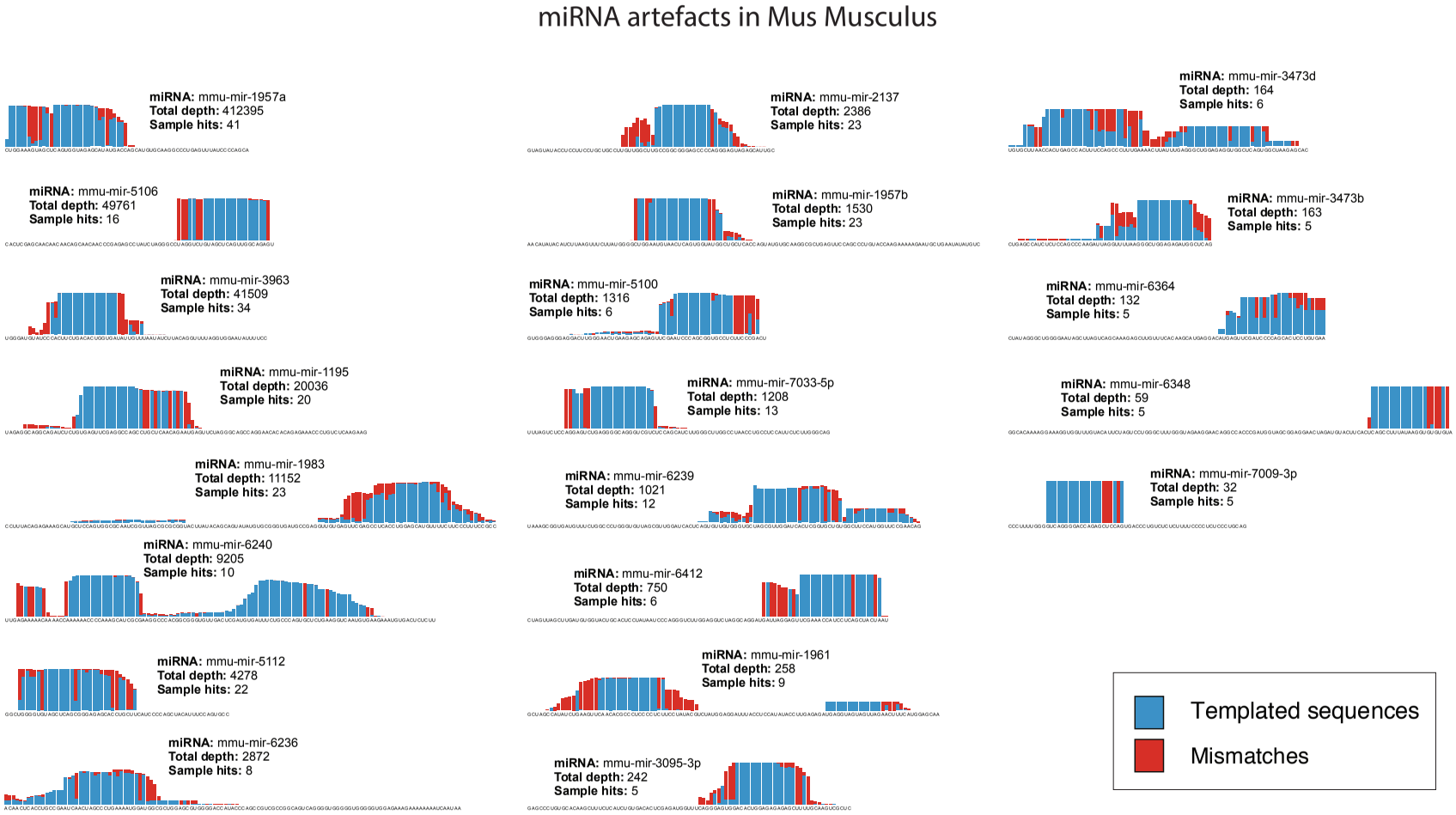



**Fig. 3.16** Coverage profiles of miRNAs that have been detected in mouse samples as potential artefacts. miRNAs with reads in less than 5 samples (5% of all samples) have been excluded from this analysis.

We used the alignment data obtained to automatically scan for miRNAs whose alignments and modification profiles did not fit those of well characterised miRNAs. More specifically, from the entire set of miRNAs for each species, we retained only those that were expressed in at least 5% of all datasets of that species. We then parsed all coverage profiles and highlighted those that had 4 or more nucleotide positions with a mismatch ratio >70%. The filtered list of modification profiles was then compared manually to the profiles of the control miRNAs in order to retain only the miRNAs with a non-canonical modification profile.

We identified 22 Human and 21 Mouse miRNAs whose profiles clearly differ from miRNAs such as let-7 (Figure 3.15 & Figure 3.16). Scanning this set of miRNAs against miRBase shows that 11 of 43 identified miRNAs show similarity to annotated non miRNA molecules in the Rfam database (Nawrocki et al., 2014). These miRNAs together with a comparison of miRNAs whose provenance is well established, e.g. via northern blot (Figure 3.17), is also available in *miratlas*.

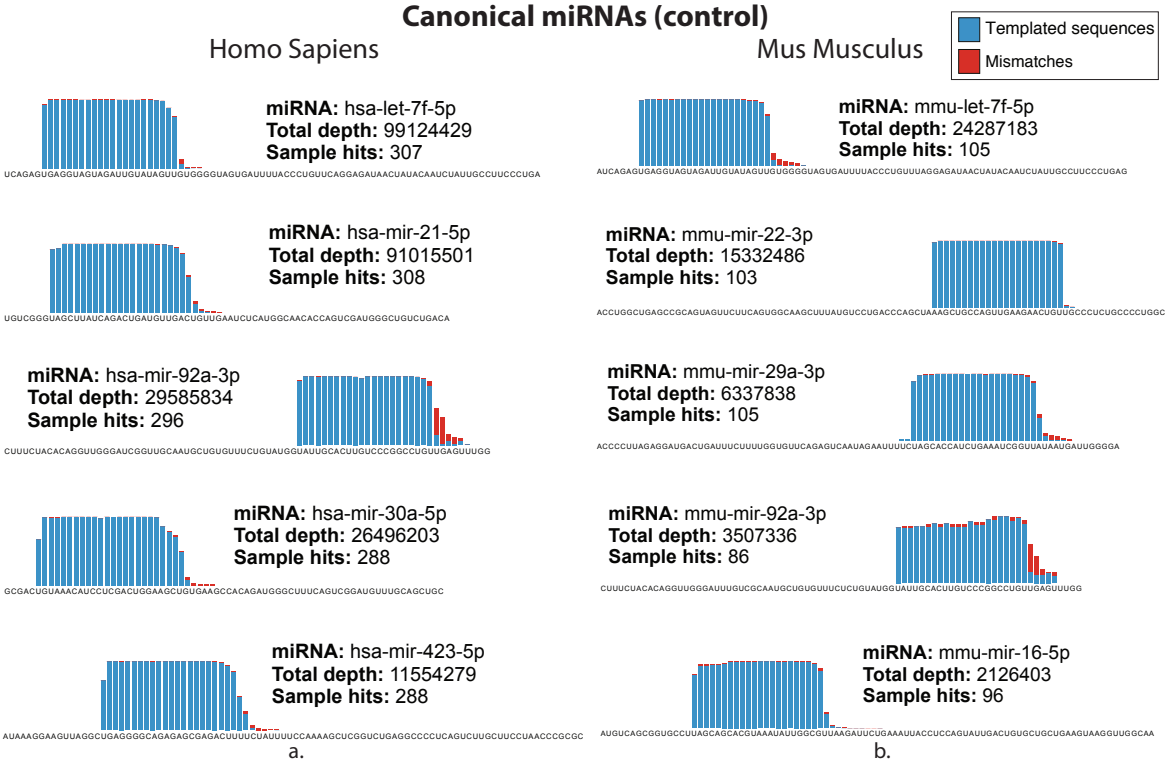



**Fig. 3.17** Coverage profiles of miRNAs that are highly expressed in: (a) all 34 human datasets and (b) all 18 mouse datasets of this study and have validation via northern blot in miRBase. These profiles serve as the control reference for the detection of potential miRNA artefacts from the analysis of miRNA coverage profiles.

Finally, we wanted to examine if the prevalent form of miRNAs expressed in the *miratlas* datasets is equivalent with the miRBase canonical annotation. Specifically, we are interested in miRNAs whose predominant form detected in our data was longer or shorter than the annotated version. Thus, we reanalysed all human and mouse *miratlas* registered samples and extracted the average length of the expressed template miRNA sequences across them, normalised by their overall expression depth. We then calculated the difference in length between each expressed miRNA and its corresponding annotated sequence in miRBase (Figure 3.19). We detected that for both Human and Mouse the prevalent form of expressed miRNAs is on average 1nt shorter than the accepted canonical sequence in miRBase. We also detect a small number of miRNAs which appear to be longer than their annotated mature sequence. Examples of both are shown (Figure 3.18).

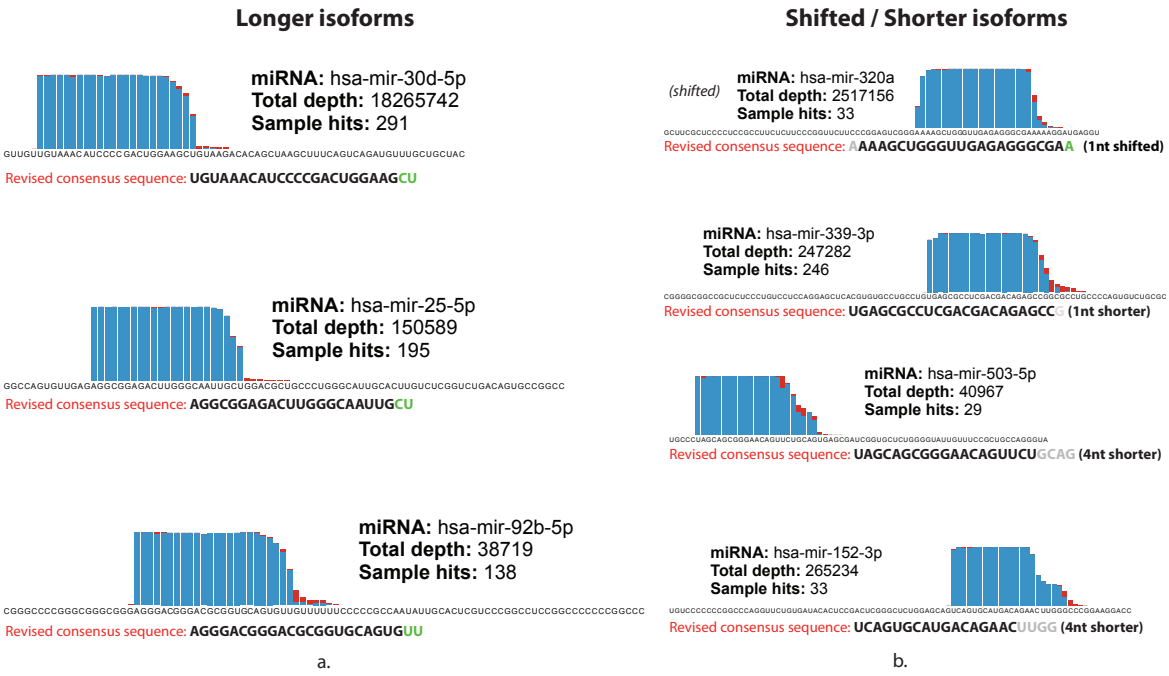



**Fig. 3.18** Mature miRNAs with revised consensus sequences. (a): miRNAs with prevalent isoforms 2nt longer than the miRBAse annotated ones. (b): miRNAs with shifted or shorter prevalent isoforms compared to the miRBAse annotated ones.

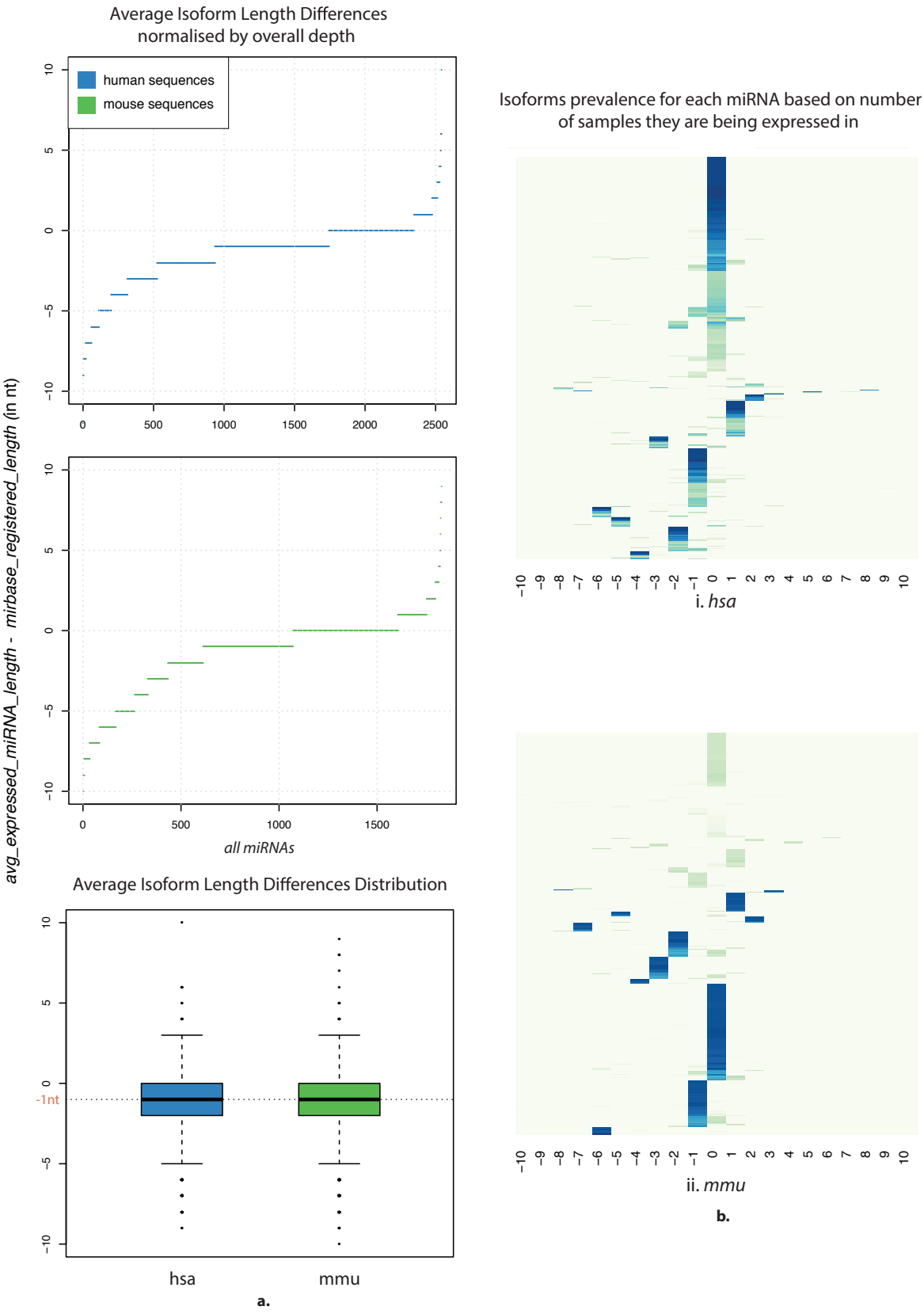



**Fig. 3.19** Length difference (in nt) between expressed miRNAs vs. miRBase annotated miRNA sequences.

## 3.3 Conclusion

In this chapter, we presented a comprehensive analysis of miRNA expression across multiple tissues and cell lines in Human and Mouse. These data are derived from high-throughput sequencing experiments from public resources. We have used these data to build a comprehensive miRNA expression dataset for Human and Mouse that takes into account both expression levels and detected modifications to miRNAs (e.g. 3' uridylation or ADAR editing). This combined data resource allowed us to explore the complex features of miRNA transcription across tissues and to group miRNAs into clusters based on their expression correlation. Additionally, we used these data to explore the likely transcriptional coupling of miRNAs in co-expressed clusters. We explored in detail, for the first time, the prevalence of both 5' and 3' nucleotide modifications to miRNAs and showed that mono and dinucleotide 3' terminal modifications are the primary modifications observed in both human and mouse, with ADAR editing mostly restricted to brain and cancer cell types.

We have also used these data to build a thermodynamic model for how the mature strand of a miRNA precursor is selected by exploring structural constraints around the ends of miRNA precursors derived from large-scale NGS data. Finally, we have suggested updated miRNA annotations based on the most prevalent isoforms derived from our expression data and we highlighted some inconsistencies in miRBase, the official miRNA repository.

Since we are able to detect mis-annotated miRNAs from high-throughput sequencing data we were intrigued to explore the possibility of predicting novel miRNAs by solely inspecting their sequencing profiles. This idea motivated us to the next project, presented in the following Chapter, which addresses the problem of novel miRNA prediction from small RNA-Seq data.

# Chapter 4

# Genome free discovery of miRNAs from small RNA-Seq with machine learning

*The results from this chapter have been published in the following paper*:
"Genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests"
**DM Vitsios**, E Kentepozidou, L Quintais, E Benito-Gutiérrez, S van Dongen, MP Davis & AJ Enright.
***Nucleic Acids Research***, Volume 45, p.e177, doi: 10.1093/nar/gkx836 (2017).

## 4.1 Introduction

The identification and annotation of novel miRNAs from various species, either animals or plants, has been a challenge in the field of small non-coding RNAs for many years. This remains an open problem, particularly given the growth of high-throughput sequencing, cell sorting and single cell biology. While a large number of miRNAs have already been annotated, there may well be large numbers of miRNAs that are expressed in very particular cell types and remain elusive. Sequencing allows us to quickly and accurately identify the expression of known miRNAs from small RNA-Seq data. Traditionally, novel miRNA prediction was based on the identification of short sequences, mapping such sequences to the genome, and searching for those loci that may produce the characteristic hairpin structure of a pre-miRNA via analysis of derived structural features. However, we sought to

explore the possibility of predicting novel miRNAs with high accuracy without requiring a reference genome in the process.

Our initial hypothesis is that features of microRNA (miRNA) sequences, derived from their biogenesis may be sufficient to predict miRNAs *de novo*, i.e. without using a reference genome. These 'biogenesis' features (Figure 4.1) are clearly evident when one interrogates large numbers of miRNA sequencing datasets from multiple species. In brief, miRNAs usually have a well-defined 5' end and a more flexible 3' end with the possibility of 3' tailing events, such as uridylation.

In order to perform genome-free feature analysis of miRNA sequences, one needs to take an input set of small RNA sequences and globally group them into clusters of related sequence. These clusters may then be aligned and filtered. This alignment allows a consensus sequence to be constructed and biogenesis features to be assessed. The advantages of *de novo* discovery of miRNAs purely from sequencing data are readily apparent: i) it does not require a reference genome, ii) removing the genomic mapping and RNA secondary structural analysis allows for faster computation and iii) it will produce a smaller set of novel candidate sequences, should one want to do genomic feature analysis later.

To this end, we have developed a new method, mirnovo, which allows for prediction of novel miRNAs in animals and plants, with or without a reference genome. This method performs comparably to existing tools, however is simpler to use with reduced run time. Its performance and accuracy has been tested on multiple datasets, including species with poorly assembled genomes, RNaseIII (Drosha and/or Dicer) deficient samples and single cells (at both embryonic and adult stage). This method is available as both a web-application (http://wwwdev.ebi.ac.uk/enright-dev/mirnovo) and a stand-alone tool (https://github.com/dvitsios/mirnovo).

## 4.2 Main methodology

The main methodology behind mirnovo, either with or without a reference genome, lies in graph-based clustering of read-read similarities from raw FASTQ files (Figure 4.2). Although there have been reported three tools in the past for "genome-free" discovery of microRNAs (Kang and Friedländer, 2015), they are either not supported anymore (Jha and Shankar, 2013), not predicting novel miRNAs (Kuenne et al., 2014) and/or using extremely restrictive approach for novel miRNA prediction, e.g. requirement for detection of both strands (Mapleson et al., 2013), which is not common for miRNAs. Our approach to the discovery of miRNAs is novel in the sense of combining large-scale RNA-Seq with machine learning methods using fine-grained feature engineering.

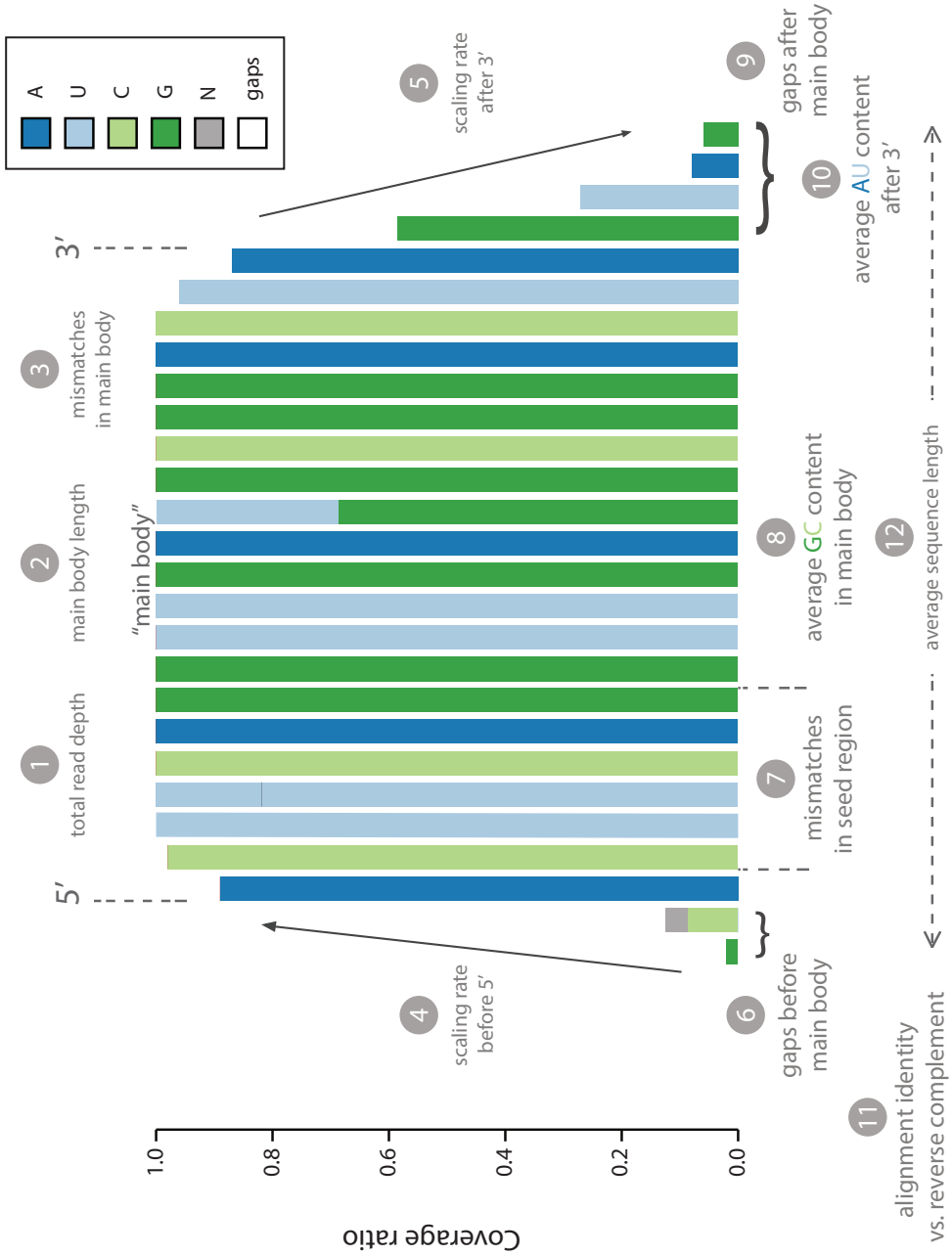



**Fig. 4.1** Coverage (biogenesis) features definition for each cluster of similar sequences.

Input files for *mirnovo* are one gzipped (.gz) FASTQ file for each run from either bulk or single-cell small RNA-Sequencing data. Input sequences may have already been pre-cleaned from their 3' adapters otherwise a 3' adapter sequence needs to be provided by the user. In the latter case, the 3' adapter from input data is removed with *reaper* and the cleaned sequences are de-duplicated with *tally* (Davis et al., 2013).

Following de-duplication, sequences are clustered together into groups based on their similarity using *vsearch* (Rognes et al., 2016). Subsequently, clusters are filtered based on the minimum number of isoform variants they contain and their overall depth. After a consensus sequence has been calculated for each cluster, all clusters are aligned against Rfam (Nawrocki et al., 2014) and aligned hits are retained to be reported separately as potential tRNAs or rRNAs. Due to inconsistencies in the initial clustering by *vsearch*, an extra refinement step has been introduced in order to merge clusters with highly similar consensus sequences. This refinement step is performed using *cd-hit* (Fu et al., 2012), based on 7-mer searches and an 0.85 alignment identity threshold.

Next, mirnovo performs multiple-sequence alignment within each cluster using *muscle* (Edgar, 2004), it assesses the new consensus sequences of the merged clusters and maps them against miRBase (Griffiths-Jones et al., 2008) in order to identify known miRNAs (if the input species has annotated miRNAs in miRBase). At this step of the workflow, mirnovo is calculating a set of features for each of the refined clusters. Then, a Random Forests-based model predicts known and novel miRNAs. After this step, the consensus sequences of all identified known and/or novel miRNAs are mapped again against the reference genome. The most stable hairpins, in terms of $\Delta$G free energy, around these sequences are selected and genomic features are calculated for each hairpin candidate. Eventually, up to 5 hairpins are reported as paralog precursors for each mature miRNA in case the calculated free energies of these secondary structures are below a certain empirically defined threshold.

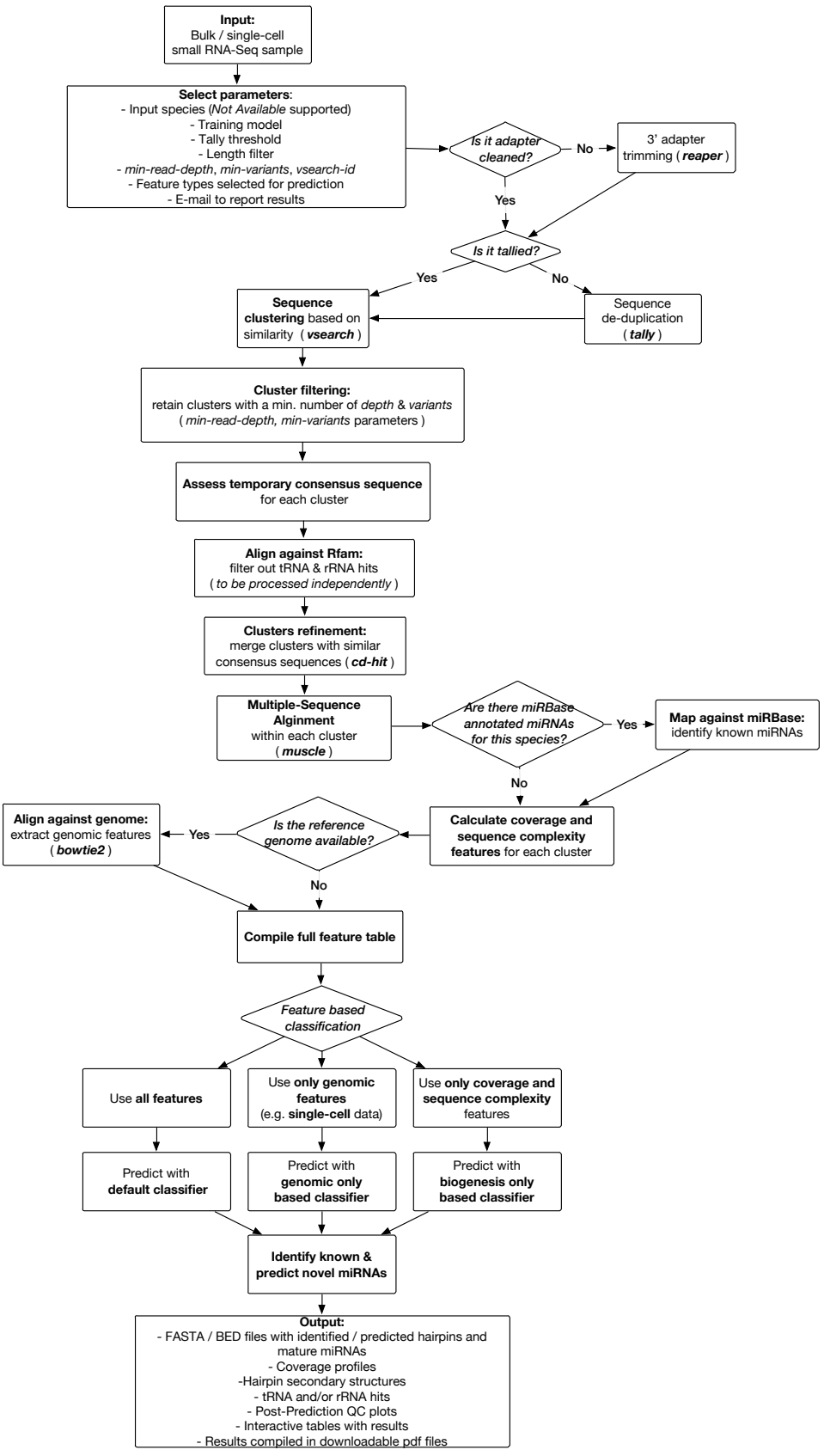



**Fig. 4.2** Mirnovo pipeline workflow.

### 4.2.1 Features definition

In order to characterise each of the small RNA clusters generated at the first step of the *mirnovo* process, we compiled a set of 33 features, grouped into three categories: 12 *coverage profile* features, a set of 12 *sequence complexity* features and finally 9 *genomic* features. The full set of features used for classification and prediction is described as follows:

- 12 coverage profile features: total read depth, main body length, mismatches in main body, scaling rate before 5', scaling rate after 3', gaps before main body, mismatches in seed region, average GC content in main body, gaps after main body, average AU content after 3', alignment identity against the potential reverse complement and average sequence length.
- 12 sequence complexity features: A+T skew (ats), C+G skew (gcs), CpG skew (cpg), complexity (cwf) by Wootton & Federhen (Wootton and Federhen, 1993), entropy (ce), complexity as compression ratio using gzip (cz), complexity as Markov model size of N ∈ {2,3} (cm2, cm3), Trifonov's complexity (Trifonov, 1990) with order N ∈ {2,3} (ct2, ct3) and linguistic complexity with order N ∈ {2,3} (cl2, cl3).
- 9 genomic features (hairpin folding retrieved using *RNAfold* from the Vienna package (Lorenz et al., 2011)): hairpin size estimate, mature miRNA distance from stem loop, loop size estimate, number of loops in hairpin, minimum free energy of secondary structure, 'majority' brackets in the entire folding (prevalent of the two distinguishing bracket directions, i.e. most frequent between '**(**' and '**)**' ), miRNA bracket discrepancy ($K/N$, where $N$ is the total number of brackets in the miRNA and $K$ is the number of 'majority' brackets), miRNA bracket fraction ($K/N$, where $N$ is the miRNA length and $K$ is the number of 'majority' brackets) and number of unmatched nucleotides from the mature miRNA sequence.

Calculation of genomic features is enabled only when the input species genome is integrated in *mirnovo*. Based on these features, *mirnovo* makes predictions for known and novel mature miRNAs using a pre-trained classifier. Predicted sequences are aligned against the genome to detect possible paralogs and results are prepared for visualisation and download.

### 4.2.2 Machine learning model selection & training

In order to predict performance of various machine learning models we have used a labelled set of feature instances derived from 65 mouse samples and applied 10-fold cross-validation on it. This allowed us to assess the bias-free predictability of each prediction

model which demonstrates how well the model will perform in general given an independent dataset. The models whose performance was tested included Support Vector Machines, Gradient Boosting Method and Random Forests (Figure 4.3). The most efficient approach in terms of discriminative power (based on the Area Under Curve -AUC- scores), with or without using the genomic features, turned out to be Random Decision Forests (or Random Forests). Hence, we selected this method to be integrated into *mirnovo* as the primary prediction algorithm.

The Random Forests implementation we used was the one provided by the *randomForest* R package. In order to fine-tune our model we tested its performance independently for various numbers of randomly selected predictors (*mtry*) and numbers of trees (*ntrees*). Optimal performance was obtained for $mtry = 6$ and $ntrees = 2000$ and thus these parameters were selected for the training of each classifier (Figure 4.3). Samples used for training of the animal and plant species models were downloaded from ENA (Leinonen et al., 2010). Overall, we have trained models for 8 animal and 7 plant species using 2-66 samples with labelled data in each case and 433 samples overall (Tables 4.1 & 4.2).

More specifically, we have trained individual models for 8 animal species (*Apis mellifera*, *Bos taurus*, *Caenorhabditis elegans*, *Canis familiaris*, *Danio rerio*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*) and 7 plant species (*Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Glycine max*, *Hordeum vulgare*, *Medicago truncatula*, *Oryza sativa japonica*, *Solanum lycopersicum*). These models are offered for optimised results when the input files originate from one of those species. Additionally, we have created two universal models, for animals and plants respectively, that can be used generically for any species belonging to one of the two kingdoms. These models have been created by sampling data-points from the entire dataset of small RNA clusters from the aforementioned animal and plant species.

The 10-fold cross-validation demonstrated accuracy measures of 84.4-96.5% without a reference genome using a model built from animal species (Figure 4.5). Interestingly, miRNA predictions on plant sequences still managed accuracy between 70.7-82.9%, despite their differences in biogenesis compared to animals (Chen, 2005). We also extracted the importance score for each feature used during the 10-fold cross-validation. Inspection of the feature importance scores for the accuracy of predictions (Figure 4.4) yields some of the coverage features (read depth, average sequence length of mature sequence, average GC content and average AT content after 3' end) as the most critical ones for correct classification, in both animals and plants. This strongly suggests that miRNA identification can be largely driven solely by inspection of their biogenesis features, without requiring extra information from the secondary structure of their precursors. Moreover, we can observe that genomic features play a more predominant role in animals than in plants, most likely

because of the high variability of secondary structures of miRNA precursors in plants. This variability in plant miRNAs can be seen in the high variance of their feature importance scores, in contrast with the lower variance of the respective animal features.

These initial results confirmed that without integrating any information from the genome it is still possible to reliably identify both known and novel miRNAs directly from sequencing data. Addition of 9 extra genomic features does improve accuracy, but not by as much as expected. There was a 1.65% (±1.53%) and 0.7% (± 2.79%) improvement of prediction accuracy for animals and plants respectively (absolute scores: 85.9-97.9%, 71.4-88.7% respectively). We also built a universal-animal and a universal-plant model by sampling data points (refined sequence clusters) from each respective pool of species such that they can be uniformly used by any species originating from these kingdoms (Figure 4.6). Obtained accuracy for these universal models was 89.7 or 92% for animals, and 71.4 or 71.8% for plants, without and with a reference genome, respectively.

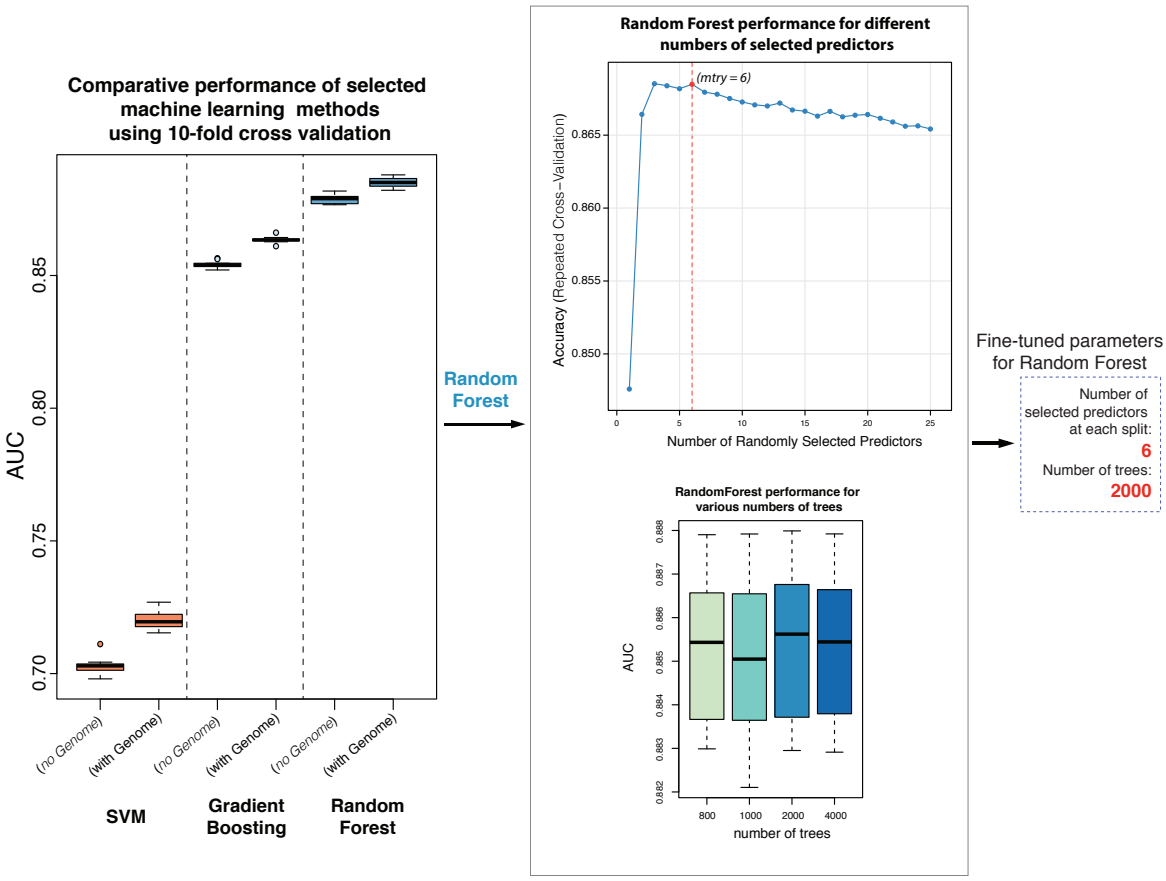



**Fig. 4.3** Machine Learning method selection & fine-tuning.

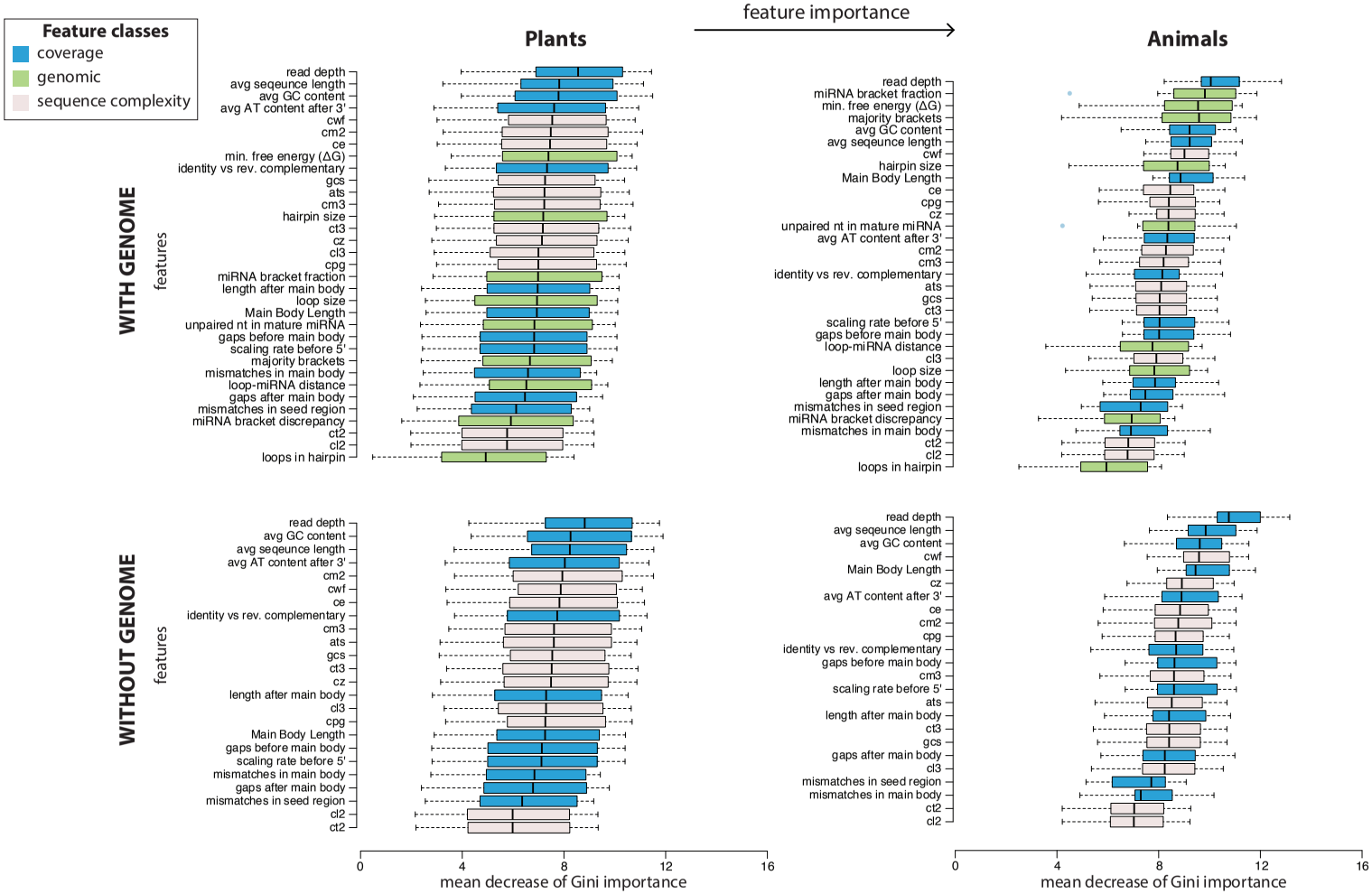



**Fig. 4.4** Feature importance scores across 8 animal and 7 plant training models, based on *Gini* scores (*Gini* importance measures the average gain of purity by splits on a given variable at a node of a tree classifier). The importance of each feature is measured by eliminating it from the set of features, making the predictions and then calculating the average decrease in *Gini* scores at all nodes of all trees of the random forest. The highest the decrease is the most important the feature is for splitting the data.

**Table 4.1** Number of samples (from ENA) used for training of each of the *animal* species training models.

| Animal Species | Number of Samples |
|---|---|
| Apis mellifera (ame) | 13 |
| Bos taurus (bta) | 62 |
| Caenorhabditis elegans (cel) | 35 |
| Canis familiaris (cfa) | 21 |
| Drosophila melanogaster (dme) | 23 |
| Danio rerio (dre) | 20 |
| Homo sapiens (hsa) | 66 |
| Mus musculus (mmu) | 65 |
| universal-animals | 306 |

**Table 4.2** Number of samples (from ENA) used for training of each of the *plant* species training models.

| Plant Species | Number of Samples |
|---|---|
| Arabidopsis thaliana (ath) | 18 |
| Chlamydomonas reinhardtii (cre) | 3 |
| Glycine max (gma) | 59 |
| Hordeum vulgare (hvu) | 6 |
| Medicago truncatula (mtr) | 2 |
| Oryza sativa japonica (osa) | 26 |
| Solanum lycopersicum (sly) | 13 |
| universal-plants | 127 |

In general, *mirnovo* can analyse datasets from any species, without requiring a reference genome or miRBase annotated miRNAs. The option '– Not Available –' should be used in this case in the place of the *Input species* input parameter. However, even higher accuracy can be achieved by integrating the genomic features into prediction. Thus, mirnovo has integrated genomic support for 67 species. This means that for those species, the full set of coverage profile, sequence complexity and genomic features can be compiled in order to identify known miRNAs and predict novel ones. Additionally, mirnovo supports miRNA identification and prediction for another 160 species with miRBase annotated miRNAs, but lacking genomic feature support. The command-line version of our method though allows the user to build and integrate into the identification process any custom reference genome.

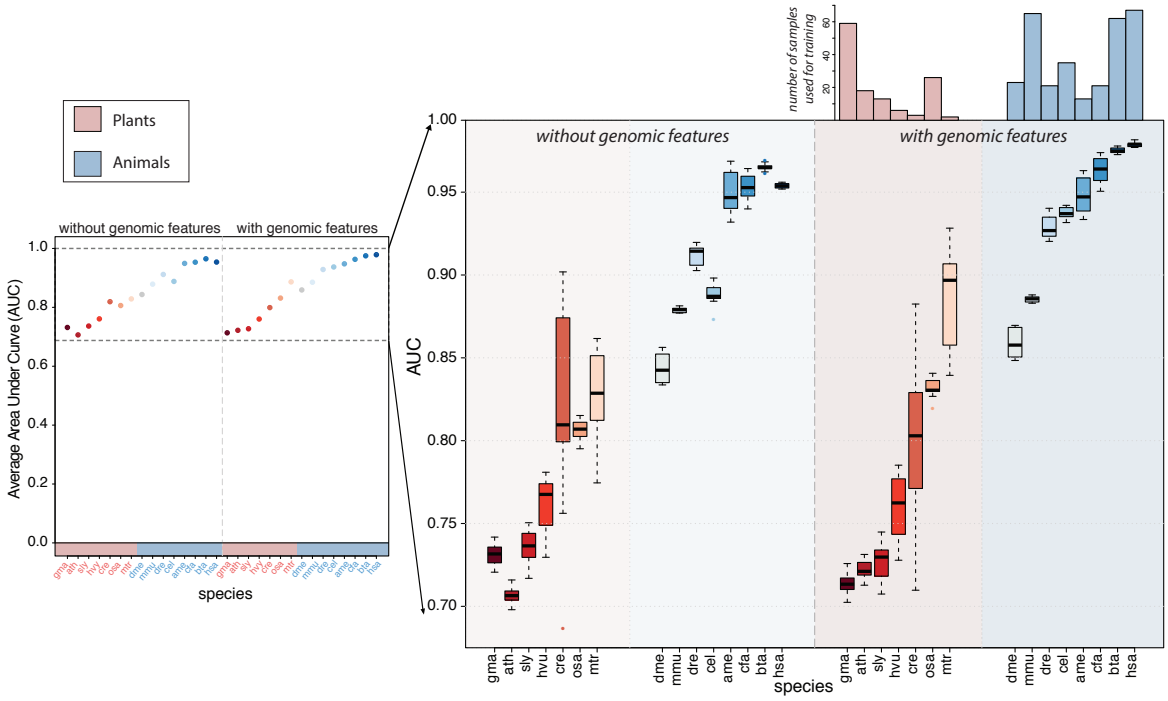



**Fig. 4.5** Training Model Performance with 10–fold cross–validation across 7 Plant and 8 Animal Species.

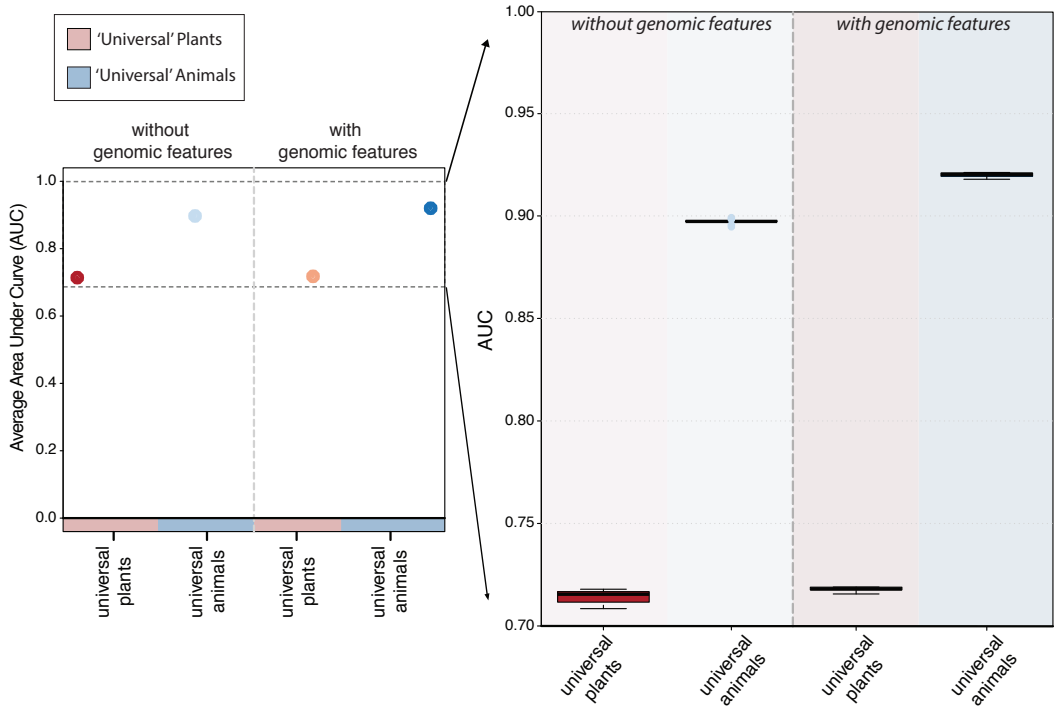



**Fig. 4.6** Training Model Performance with 10–fold cross–validation across the universal-plant & universal-animal models.

### 4.2.3 Parameter specification and output

MicroRNA prediction is performed by default using all 24 biogenesis features and the 9 genomic features (in case the reference genome is available). However, it is also possible to completely disable genomic features, by checking the '*Disable genomic features*' option, or use exclusively the genomic features for prediction, by checking the '*Use only genomic features for prediction*' option. Furthermore, mirnovo offers a set of mainly three parameters in order to facilitate sequence clustering and boost correct classification of predicted miRNAs. Specifically, when analysing samples with high read depth and high sequence complexity (i.e. high number of generated clusters at the initial sequence clustering of input data with *vsearch*), we noticed that in some cases predictions contain an unexpectedly high number of novel miRNAs, sometimes even higher than the number of predicted known miRNAs (Figure 4.7). In order to resolve this issue we introduced, first of all, the '*Reduce input sequence complexity*' option which allows the user to filter out unique sequences from the input file with a total read depth below a certain threshold. For instance, by using a tally-threshold of x3, all unique sequences from the tallied file with a maximum number of 3 reads will be discarded from the rest of the analysis. Following the initial sequence clustering, additional filtering is possible by retaining only those clusters that have a total depth at least equal to the *min-read-depth* value and a number of isoforms at least equal to the *min-variants* parameter value.

With regards to *mirnovo*'s output, the results from each job contain first of all FASTA files for the predicted known and novel miRNAs (both for the mature products and their respective hairpin precursors), and for any tRNA and/or rRNA identified hits. Additionally, BED files with genomic coordinates of predicted hairpins are provided along with coverage profiles for each mature miRNA and also the secondary structures of each identified hairpin paralog. Furthermore, each job is associated with a table of performance measures with regards to the machine learning predictions.

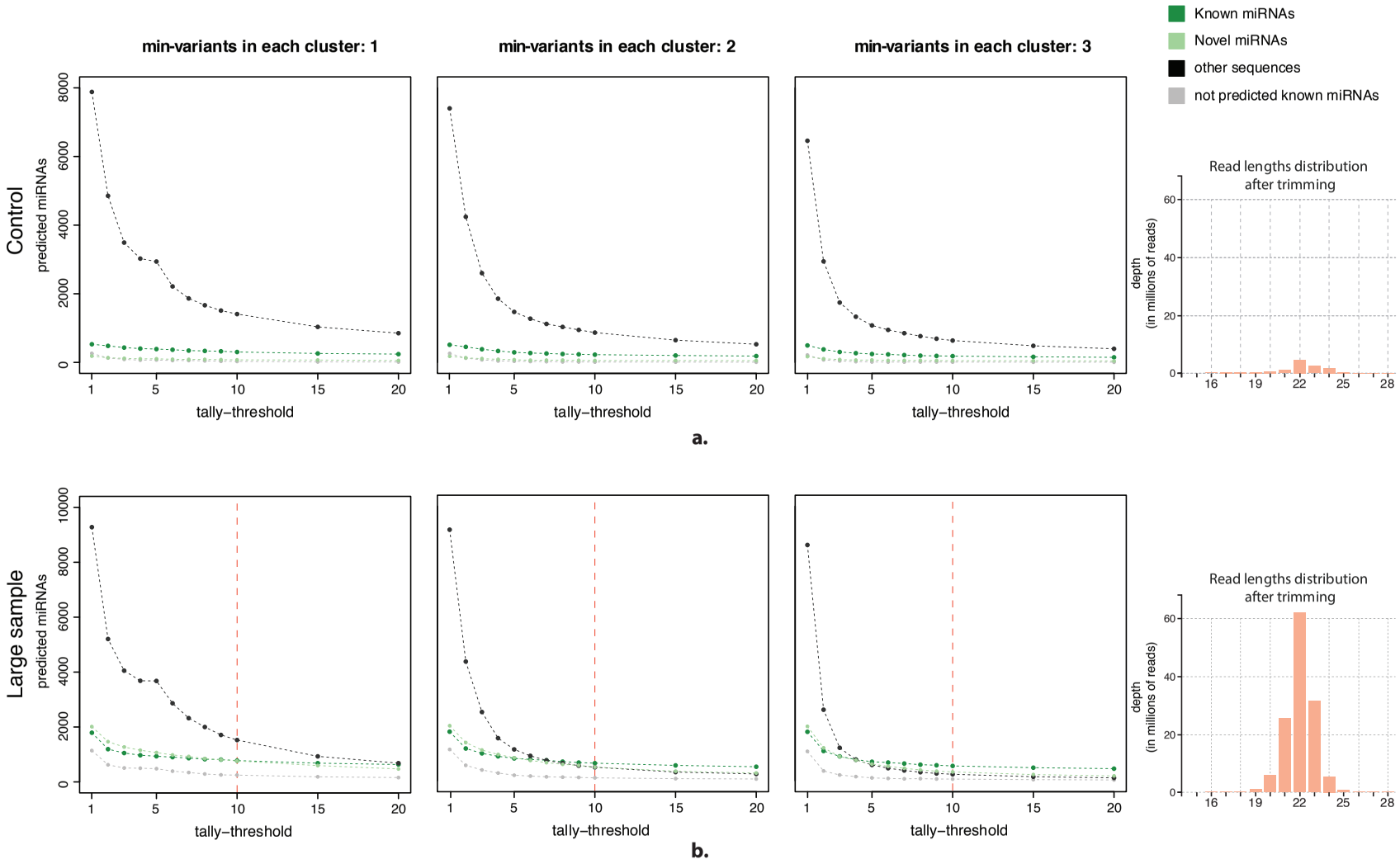



**Fig. 4.7** Mirnovo machine learning performance across a (a) medium-size control sample and a (b) large sequence complexity sample for various cut-off filter values regarding the minimum number of variants in each cluster and the tally-threshold.

The reported measures are:

$Precision = \frac{TP}{TP + FP}$,

$Sensitivity = \frac{TP}{P}$ and

$Specificity = \frac{TN}{N}$, where:

- $TP$: is the number of predicted known miRNAs,
- $FP$: is the number of predicted novel miRNAs,
- $P$: is the number of all known miRNAs contained in the input data (based on the miRBase annotation),
- $TN$: is the number of (correctly) predicted non-miRNA sequences and
- $N$: is the number of all non-miRNA sequences contained in the input data (based on the miRBase annotation).

Additionally, predictions are accompanied with a Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curve, which demonstrate the performance of the machine learning method with regards to correctly identifying known and novel miRNAs, respectively. Finally, the distribution of all feature values (coverage, sequence complexity and genomic) for each class of predicted miRNA/non-miRNA sequences is visualised and made available as post-prediction QC box-plots.

In the next sections we are going to present the results from applying mirnovo into large scale datasets, Drosha/Dicer-dependent samples and single cells in order to showcase its performance in all these cases as well as highlight the findings and new insights into miRNA biogenesis retrieved from each one of them.

## 4.3 Large-scale benchmarking & mirnovo applications

### 4.3.1 miRNA prediction from the GEUVADIS dataset

We first applied mirnovo into a large-scale benchmarking test using all human samples of the GEUVADIS dataset (Lappalainen et al., 2013). The majority of datasets were run using the default *mirnovo* parameters (*length filter*: 16-28nt, *min-read-depth*: 5, *min-variants*: 1, *vsearch-id*: 0.9). The '*Reduce input sequence complexity*' option with a tally-threshold of x3 was

used only for 2% of all datasets in order to reduce sequence complexity within the samples and thus optimise the initial sequence clustering with *vsearch*. The initial run was performed without using the human reference genome (Figure 4.8). The obtained accuracy reached an average score of 92.14% while sensitivity and novel prediction rate were at 69.07% and 34.62%, respectively. After introducing the reference genome and corresponding genomic features (Figure 4.9), performance gets notable improvement since accuracy and sensitivity rise to 95.51% and 78.8% while the novel prediction rate falls to 18.63%. This implies that the use of genomic features is boosting the clarification in prediction of real miRNAs while at the same time keeping the number of false positive hits among the novel miRNAs at a relatively low level.

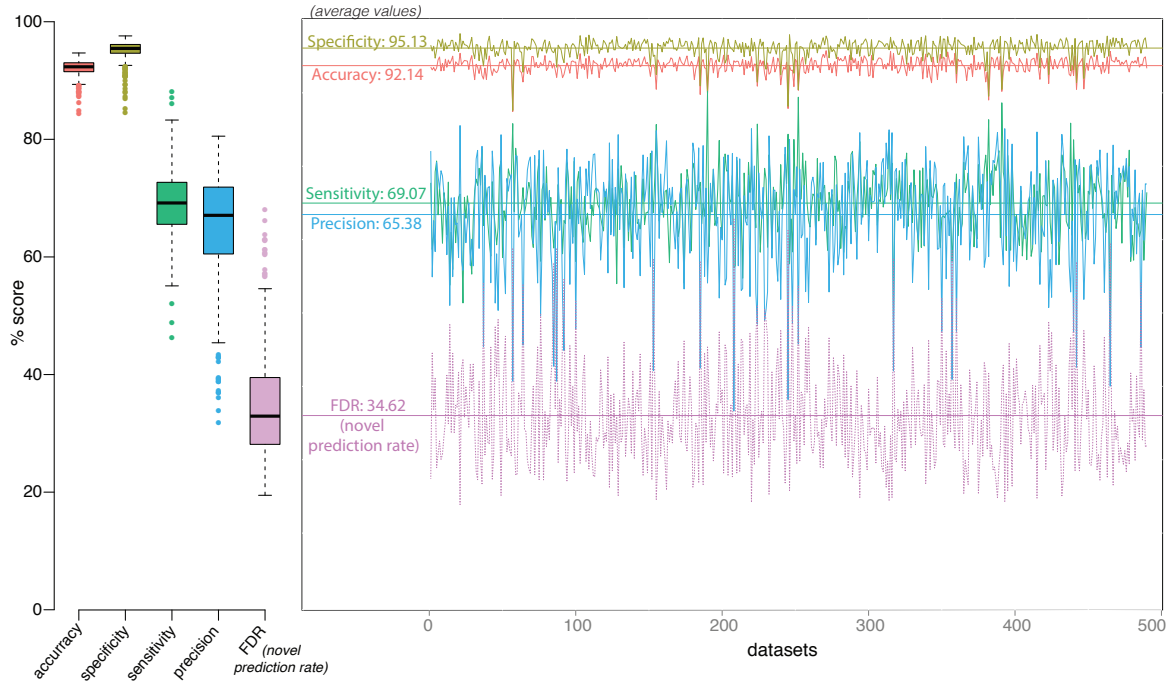



**Fig. 4.8** Machine learning prediction performance of *mirnovo* for the GEUVADIS Dataset (**without a reference genome**).

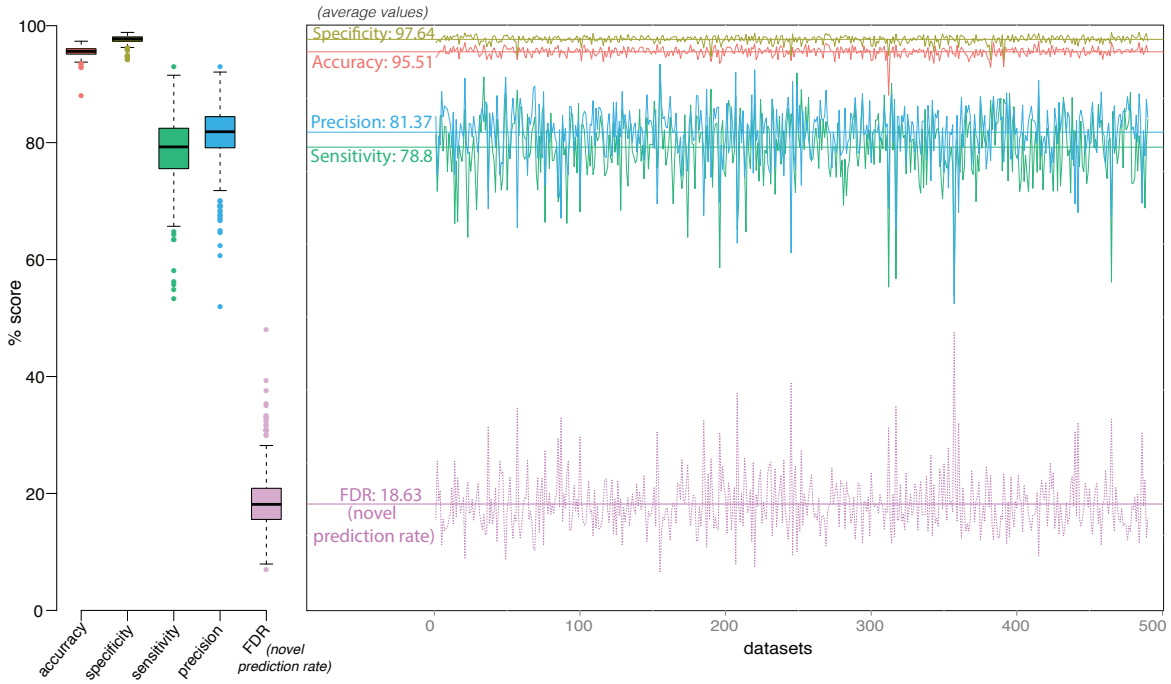



**Fig. 4.9** Machine learning prediction performance of *mirnovo* for the GEUVADIS Dataset (**with a reference genome**).

Moreover, we wanted to evaluate *mirnovo*'s performance in comparison with miRDeep2 (Friedländer et al., 2012), one of the most widely used methods for miRNA prediction. We ran miRDeep2 by always providing the human reference genome, all known human hairpins and mature miRNAs as well as a list of other known miRNAs from another two species. Mirnovo was tested both with and without the reference genome in separate runs (Figure 4.10, 4.11). We observed that mirnovo outperforms miRDeep2 in 92% of the cases for known mature miRNAs identification and in all cases for novel miRNAs prediction, when using a reference genome. In the absence of a reference genome, mirnovo performs equivalently with miRDeep2 in terms of predicting known miRNAs, however we also notice a rise in novel prediction rate, which probably includes more false positive hits. In general, we observed that the addition of genomic features in the prediction algorithm improves only slightly the sensitivity of the final results but has however a notable impact in improving precision, thus reducing the number of falsely predicted novel miRNAs (Figure 4.12). The final outcome included 2,414 predicted novel mature miRNAs originating from 3,173 hairpin precursors, including any detected paralogs (Appendix C: Supplementary Data S1). Expression of novel miRNAs was fairly balanced across all samples, and quite similar with known miRNAs expression, while the lengths of the high majority of predicted novel miRNAs were within the range of 20-23nt (Figure 4.13).

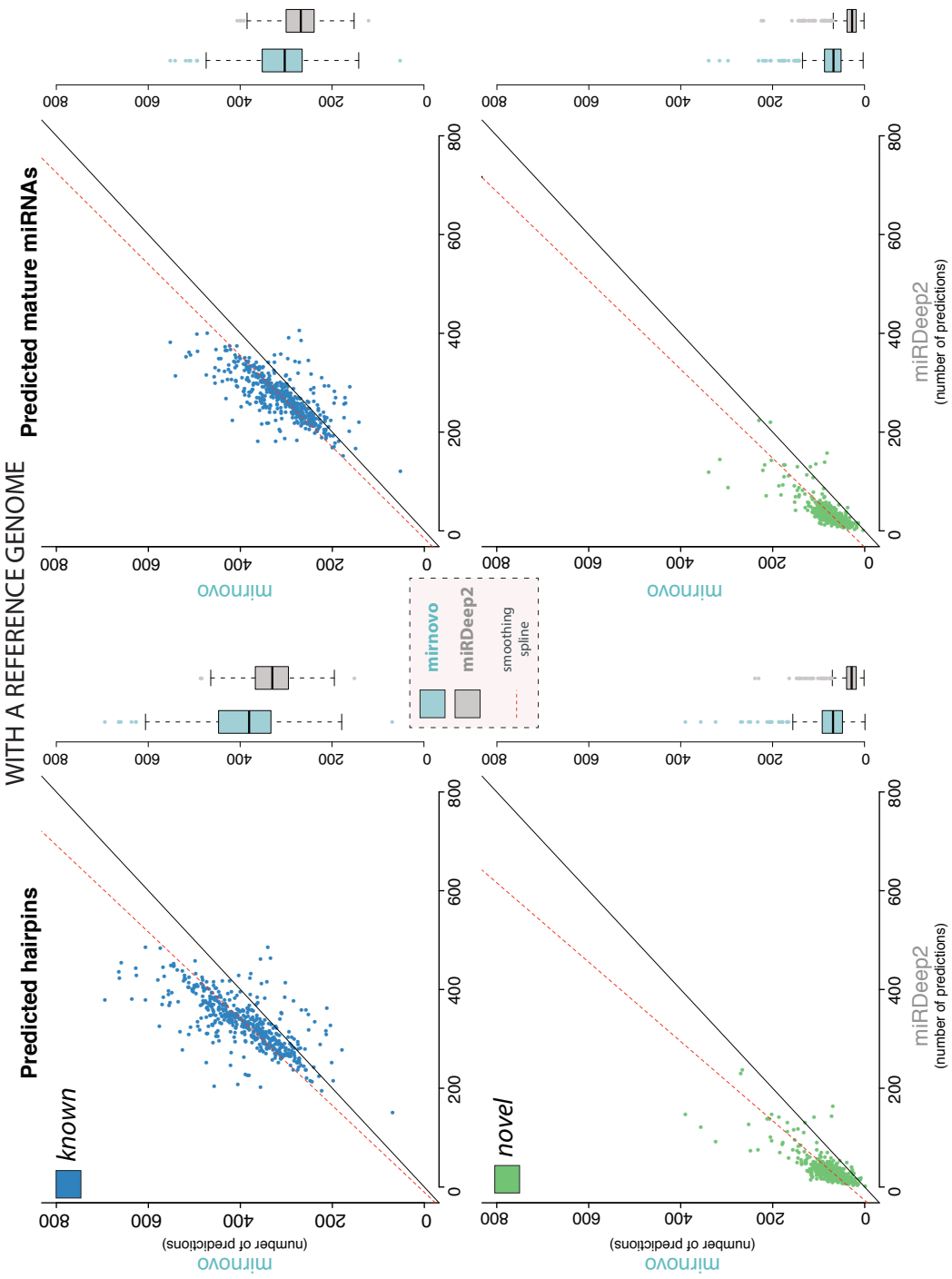



**Fig. 4.10** mirnovo vs miRDeep2 benchmark across the GEUVADIS dataset. miRDeep2 was run using a reference genome, all known human hairpins and mature miRNAs and mature miRNAs from 2 extra species (*D. melanogaster* and *C. briggsae*). mirnovo was run **using a reference genome**.

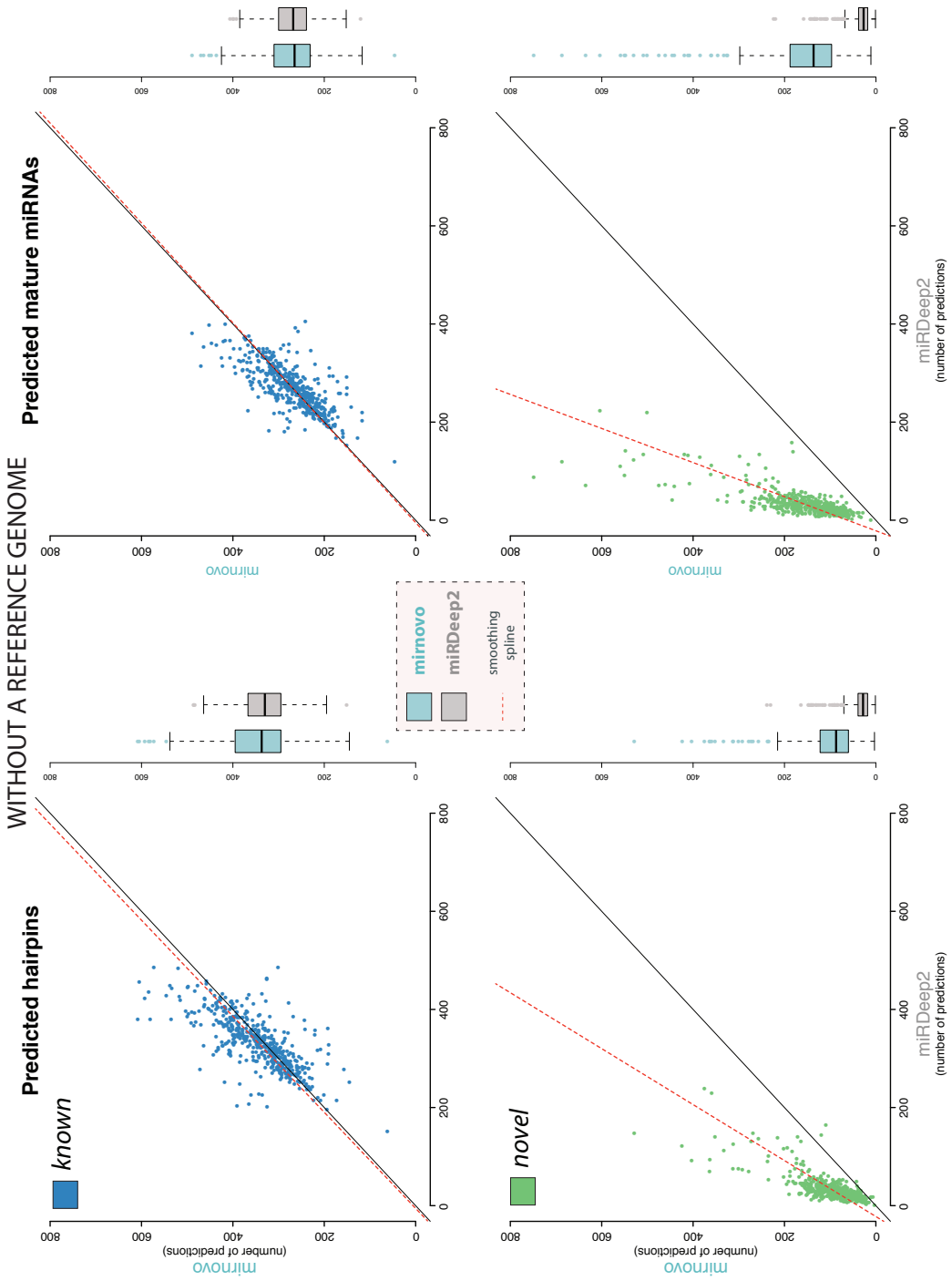



**Fig. 4.11** mirnovo vs miRDeep2 benchmark across the GEUVADIS dataset. miRDeep2 was run using a reference genome, all known human hairpins and mature miRNAs and mature miRNAs from 2 extra species (*D. melanogaster* and *C. briggsae*). mirnovo was run **without a reference genome**.

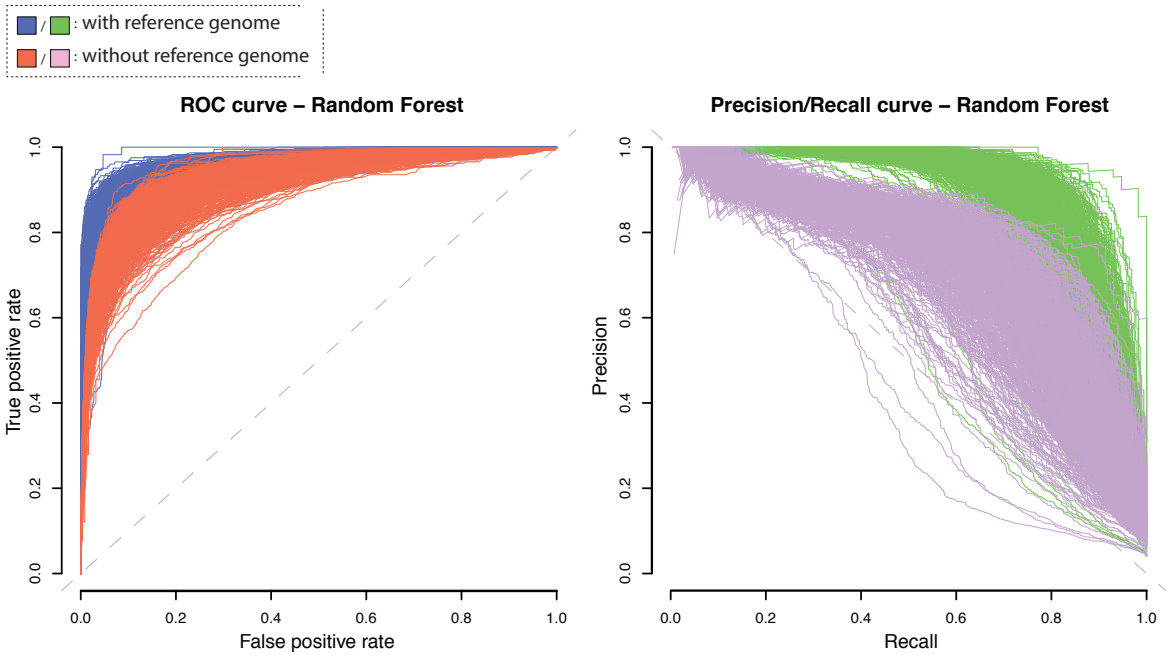



**Fig. 4.12** ROC and Precision-Recall (PR) curves for mirnovo's prediction performance across all samples from the GEUVADIS dataset, with or without using a reference genome.

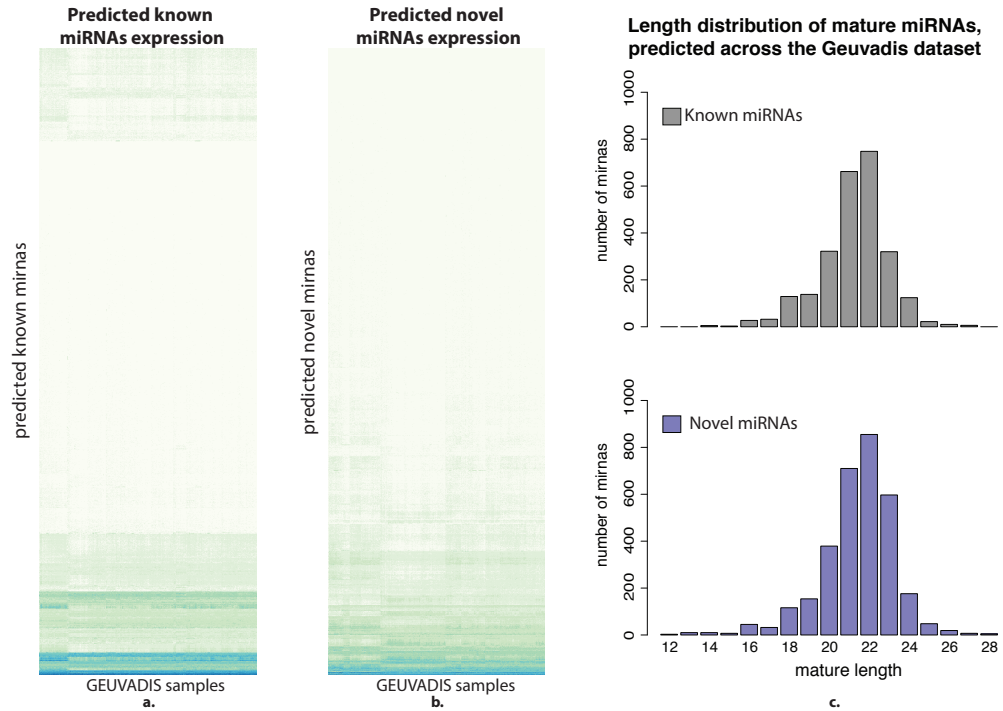



**Fig. 4.13** Expression of predicted (a) known and (b) novel miRNAs across all samples from the GEUVADIS dataset and (c) length distribution of predicted mature novel miRNAs.

We also ran mirnovo and miRDeep2 across 8 animal species (Figure 4.14) and again observed that mirnovo performs better than miRDeep2 in the majority of cases for predicting both known and novel miRNAs. In terms of execution time, based on the benchmarking run using the GEUVADIS dataset, mirnovo was on average x2.5 faster than miRDeep2 and the mean execution time across all samples was around 43min, as compared to 1hr 49min for miRDeep2 (Figure 4.15).

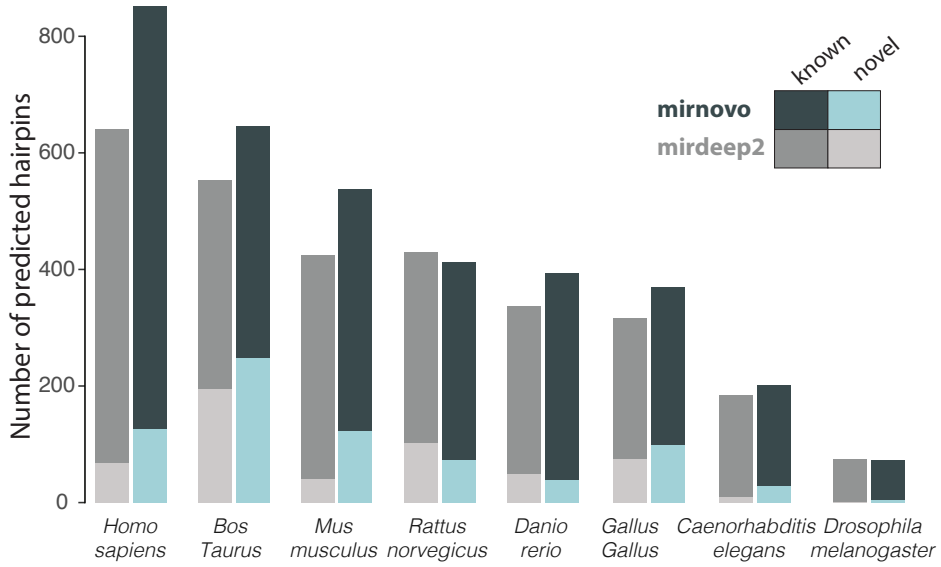



**Fig. 4.14** mirnovo vs miRDeep2 prediction performance in samples from 7 model organisms.

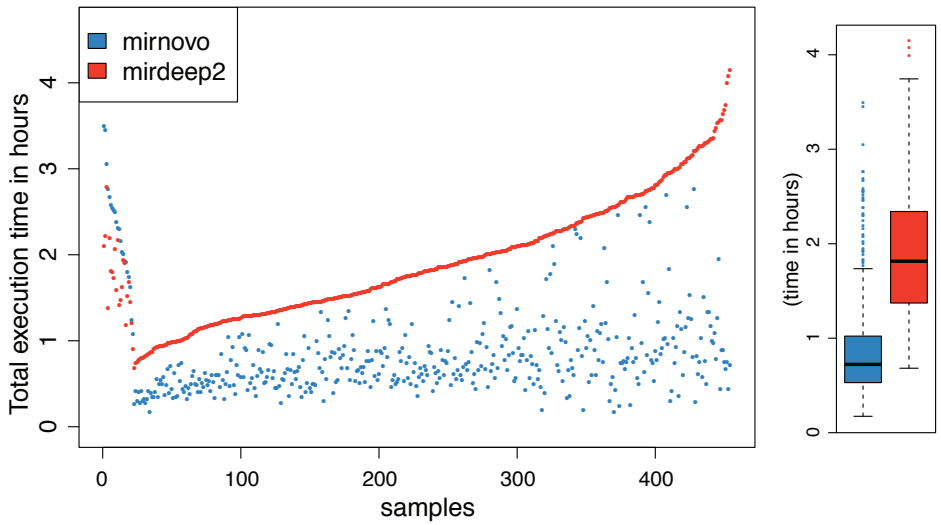



**Fig. 4.15** mirnovo vs miRDeep2 'time' benchmarking across the GEUVADIS dataset.

### 4.3.2 Prediction performance in species with incomplete genome assemblies

Mirnovo provides for the first time the ability to predict miRNAs from species without a reference genome or with poorly assembled genomes with very high accuracy. We assessed mirnovo's performance on 7 samples from 5 different moth species without fully assembled genomes (Quah et al., 2015), 2 of which do not have any miRBase annotation (Table 4.3). Mirnovo was able to retrieve known miRNAs from all species with miRBase annotation (*B. mori*, *H. melpomene melpomene*, *H. melpomene rosina*) along with hundreds of novel miRNAs (Appendix C: Supplementary Data S2-S6) . Additionally, mirnovo predicted 119 and 192 miRNAs from the *C. ohridella* and *P. aegeria* samples, respectively, which do not have any miRBase annotated miRNAs (Appendix C: Supplementary Data S7, S8). Among all predicted novel miRNAs, *C. ohridella* and *P. aegeria* were the species with the highest majority of miRNAs aligning with paralogs from other species registered in miRBase. This is expected since the other 3 moth species have been studied more extensively in the past and already have miRNA entries in miRBase. A small proportion of novel miRNAs were predicted without any genomic evidence, based solely on features derived from their coverage profiles. That demonstrates another mirnovo's strength to infer miRNAs in species without a reference genome, enabling research on non-coding RNAs for an amplitude of non-model organisms.

In order to retrieve these results, we first identified the 3' adapter sequence for each sample using *minion* (Davis et al., 2013) and where possible confirmed it with the relevant manuscript or database methods. Each sample was analysed using *mirnovo* with either default or custom set of parameters. The sample ids that were analysed are:

- SRR035544 & SRR035546 (*GSE17965, PMID: 20199675*),
- SRR062599 (*GSE23292, PMID: 200023292*),
- SRR062600 (*GSE23292, PMID: 21266089*),
- SRR1663190 & SRR1663191 (*GSE63644, PMID: 25576364*) and
- SRR035545 (*GSE17965, PMID: 200017965*).

A relevant genome was used for each sample (*Bombyx mori*: GCA-000151625.1, *Heliconius melpomene*: Hmel2 v2-0 Release-20151013, *Cameraria ohridella*: k51, *Pararge aegeria*: k51) and a *Drosophila Melanogaster* (dme) training model for miRNA predictions. To find the orthologues, novel mature miRNA sequences were compared to all miRBase sequences (v21) using *swan* (v17-096) requiring at least a 90% identity match (*–key-value* parameter).

**Table 4.3** Mirnovo predictions for known and novel hairpins and mature miRNAs across 7 samples from moth and butterfly species.

| Sample | known hairpins | known mature miRNAs | novel hairpins | novel mature miRNAs | novel paralogs from other species in miRBase |
|---|---|---|---|---|---|
| Bombyx mori (Whole body) | 89 | 75 | 64 | 59 | 2 |
| Bombyx mori (Anterior silkgland) | 78 | 72 | 91 | 80 | 0 |
| Bombyx mori (Posterior silkgland) | 55 | 51 | 60 | 40 | 0 |
| Heliconius melpomene melpomene | 41 | 40 | 114 | 118 | 12 |
| Heliconius melpomene rosina | 44 | 43 | 89 | 98 | 8 |
| Cameraria ohridella | - | - | 119 | 120 | 32 |
| Pararge aegeria | - | - | 192 | 191 | 29 |

### 4.3.3 MicroRNA prediction in RNase III-deficient cells

Novel miRNAs are predicted based on features consistent with their processing by the small RNA biogenesis machinery. Hence, if they are real miRNAs, one would expect to observe their dysregulation when key miRNA biogenesis enzymes are missing or are mutated. We tested this hypothesis using published experimental data from Drosha, XPO5 and Dicer knockout samples (Kim et al., 2016). These enzymes are responsible for cleavage of miRNA primary transcripts, their nuclear export and processing into functional mature miRNAs respectively. Samples were normalised using the same strategy that was suggested in the original manuscript. Specifically, we normalised the wild-type, Drosha and XPO5 knockout samples based on the read counts of *miR-320a-3p* across all replicates, since its expression is independent of Drosha. The Dicer knockout samples were respectively normalised based on the combined tRNA and rRNA levels of the WT samples, which should remain unaffected in the knockout samples as well.

We first predicted known and novel human hairpins from the wildtype (WT) samples. Then, we aligned all the WT and Knockout (KO) samples with *Chimira* (Vitsios and Enright,

2015) against the predicted known hairpins that were predicted by *mirnovo* (Figures 4.16a,b and 4.17a,b). In fact, in order to get expression data from all samples with reference to the hairpins that were predicted by *mirnovo*, we expanded the already published method *Chimira*. That was necessary because inherent sequence clustering steps (initial and refined) of the *mirnovo* pipeline may be imperfect in some cases and thus affect, even at a low level, the yielded expression data. The additional feature of *Chimira* allows alignment against a custom reference species that can be uploaded as a set of FASTA files by the user (e.g. FASTA files with known and/or novel hairpins predicted by *mirnovo*). All uploaded files are merged and sequences with an alignment identity over 0.90 are collapsed. As an additional functionality, *Chimira* is able to generate coverage profiles of each identified mature miRNA and the secondary structure of the corresponding hairpin reference hit, using the Vienna package (Lorenz et al., 2011).

Our data verified the observed minor effect of XPO5 knockout in miRNA expression, since miRNAs are still being expressed, just in lower levels in some cases. The Dicer knockout, as expected, leads to notable decrease in miRNA expression. The absence of Drosha is verified to be the most critical one since it results in extensive depletion of the majority of miRNAs, in accordance with the results reported in the original paper (Kim et al., 2016).

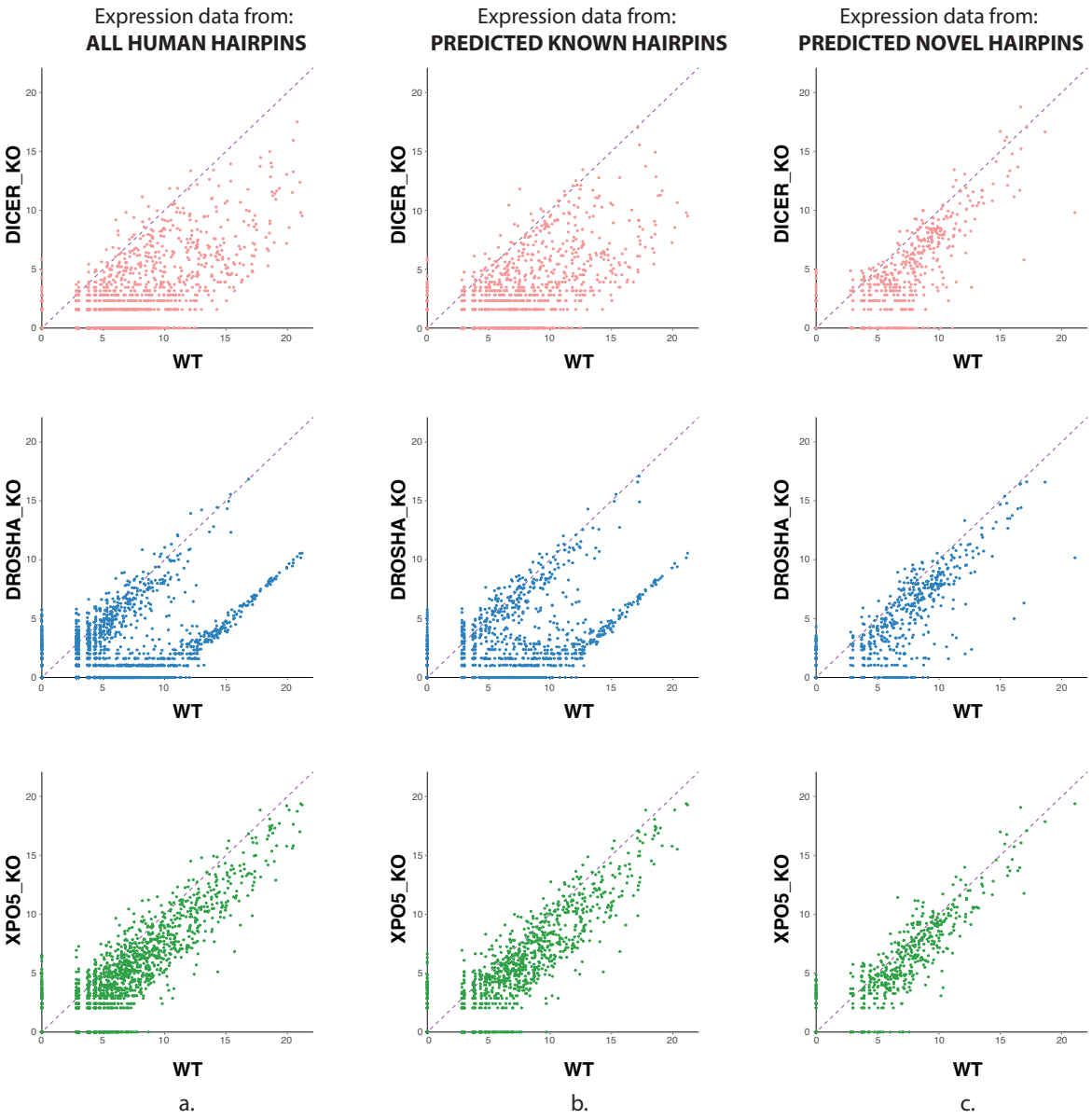



**Fig. 4.16** Expression data across wild-type and Drosha/Dicer/XPO5 knockout conditions after alignment against different sets of hairpins predicted by mirnovo (pairwise plots): a) all known human hairpins, b) known human hairpins predicted by mirnovo and c) novel human hairpins predicted by mirnovo.

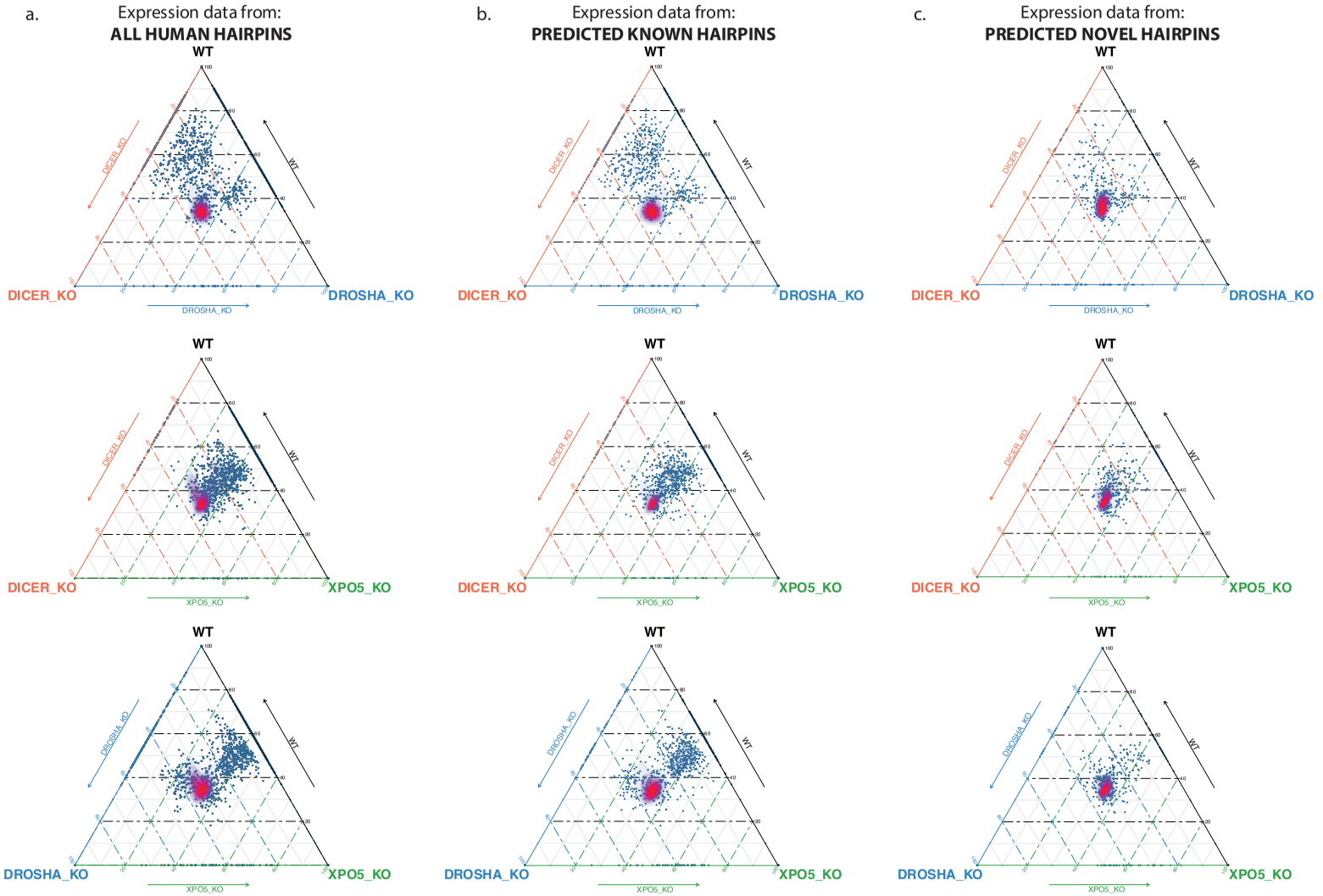



**Fig. 4.17** Expression data across wild-type and Drosha/Dicer/XPO5 knockout conditions after alignment against different sets of hairpins predicted by mirnovo (ternary plots): a) all known human hairpins, b) known human hairpins predicted by mirnovo and c) novel human hairpins predicted by mirnovo.

For our novel miRNA predictions, we aligned all samples against the list of predicted novel hairpins (Figures 4.16c and 4.17c). We then assessed which miRNAs where differentially expressed (fold-change $> 2$ and $P < 0.05$) between the WT and KO conditions and found three sets of novel miRNAs, dependent on different types of enzymes each of them (Figure 4.18 and Appendix C: Supplementary Data S9). Overall, we have found 40 novel miRNAs that were significantly differentially expressed both in Drosha and Dicer knockout samples ( Appendix C: Supplementary Data S10, S11). This implies that this set of novel miRNAs is dependent on the two most important enzymes for miRNA biogenesis (Drosha and Dicer) and thus they should be following the canonical biogenesis pathway. Moreover, we noticed that 25 novel miRNAs were dependent only by Dicer and 33 were Drosha-only dependent (Appendix C: Supplementary Data S12). This finding comes in accordance with previous studies (Cheloufi et al., 2010; Cifuentes et al., 2010; Kim et al., 2016; Ruby et al., 2007) that some miRNAs may be dependent on only one type of enzyme (either Drosha or Dicer) and/or originate from other structured non-coding RNAs (Kim et al., 2016). These results, again, provide validation that mirnovo is predicting molecules likely to be processed by the canonical biogenesis machinery yet can also identify those miRNAs which are independent of one or more of the key enzymes.

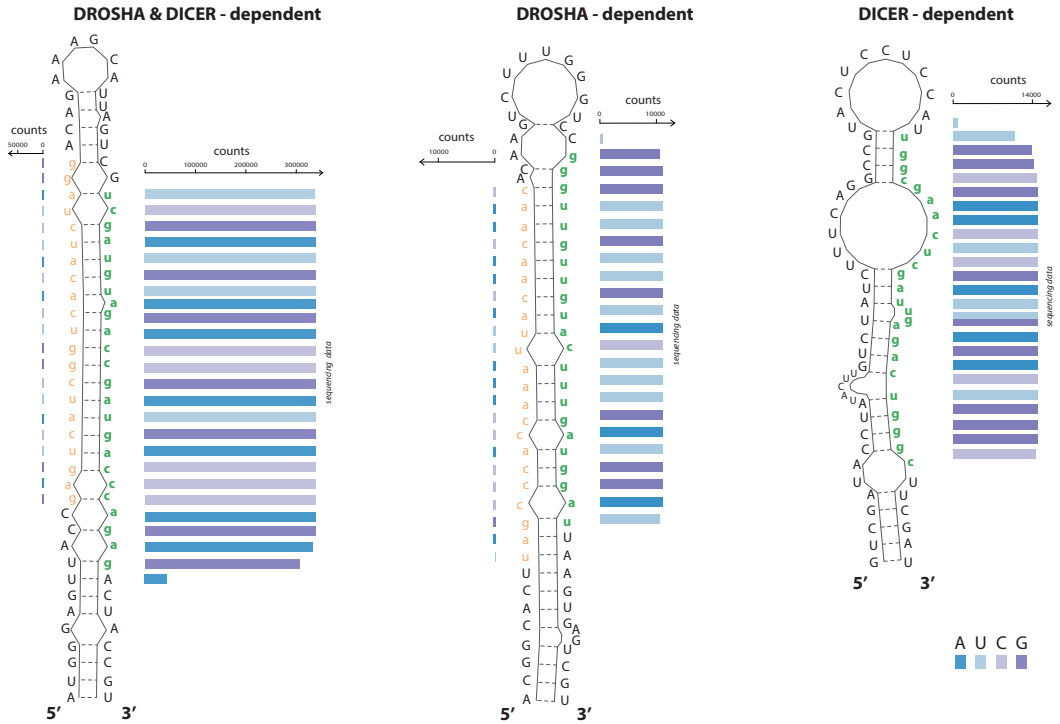



**Fig. 4.18** Examples of predicted novel miRNAs in human cell line (HCT116) by mirnovo, dependent on both Drosha and Dicer, Drosha only or Dicer only, respectively.

### 4.3.4 MicroRNA prediction from single-cell RNA-Seq data

Recently, single-cell RNA sequencing has become both tractable and an extremely active topic of research. Given that some miRNAs have been shown to be extremely cell-type specific, such datasets represent an important area for novel miRNA discovery. Hence, we wished to assess the performance of mirnovo in analysis of single-cell small RNA-Seq. We initially attempted prediction using all sets of features (coverage, sequence complexity and genomic) but the extracted coverage profiles and sequence complexity scores were distorting predictions due to high noise of input data. We then tried making our predictions using only the genomic features and observed a clear improvement in accuracy scores, thus we followed this approach for the analysis of single-cell data. This proves to be another useful feature of mirnovo, since the user is always able to switch off certain sets of features in order to make their predictions based on the specific requirements, quality or noise of input data.

We re-analysed 204 scRNA-Seq (single-cell) samples overall from HEK cells, naive human embryonic stem cells (hESCs) and primed hESCs (Faridani et al., 2016). Embryonic stem cells are referred to as naive in pre-implantation embryos and they turn into primed cells during post-implantation development (Nichols and Smith, 2009). Mirnovo predicted 4,747 novel hairpins overall from these samples, 356 of which had also been predicted from the GEUVADIS dataset on human lymphoblastoid cell lines. We then aligned all samples against the predicted set of hairpins using *Chimira* (Vitsios and Enright, 2015), and obtained mature miRNA expression data for each cell sample. Novel miRNA expression is quite balanced between the two types of ESCs, with a small group of miRNAs being down-regulated during the transition from naive to primed ESCs (Figure 4.19). On the other hand, both types of embryonic stem cells show notable differentiation in expression compared to an adult cell type, which is the HEK cells. Interestingly, highly similar expression patterns can be observed with regards to known miRNA expression across these cell types (Figures 4.19 and 4.20).

We observed that novel miRNA expression varies across different states of pluripotency and/or development in *Homo Sapiens*, with a more significant difference observed between embryonic stem cells versus fully differentiated cell types. We performed hierarchical clustering for all cells based on their novel miRNAs expression. We identified 5 major groups of cells with similar novel miRNA signature (Figure 4.20b). Two of those groups were exclusively comprised of HEK cells and two groups were primarily populated by ESC naive and ESC primed cells, respectively. Finally, the last group consisted of cells from all 3 cell types. Hierarchical clustering of these cells based on known miRNA expression also yield similar grouping of the samples based on their cell type (Figure 4.20a).

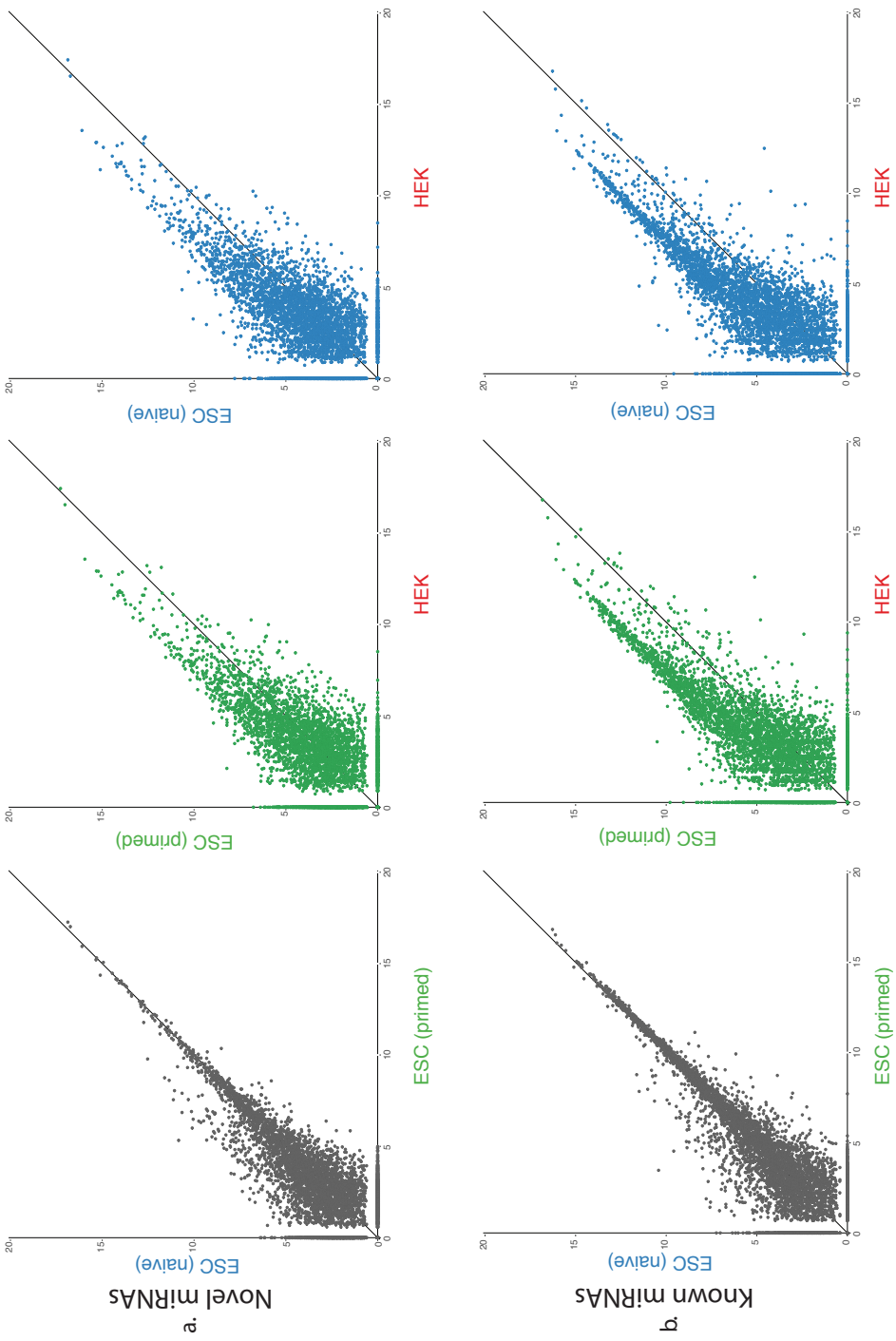



**Fig. 4.19** miRNA expression in single cells (naive ESCs, primed ESCs, HEK cells) based on mirnovo predictions. a) Normalised expression of **novel** miRNAs, predicted by mirnovo and quantified by Chimira, across the three sample conditions in pairwise plots, b) Normalised expression of **known** miRNAs, predicted by mirnovo and quantified by Chimira, across the three sample conditions in pairwise plots.

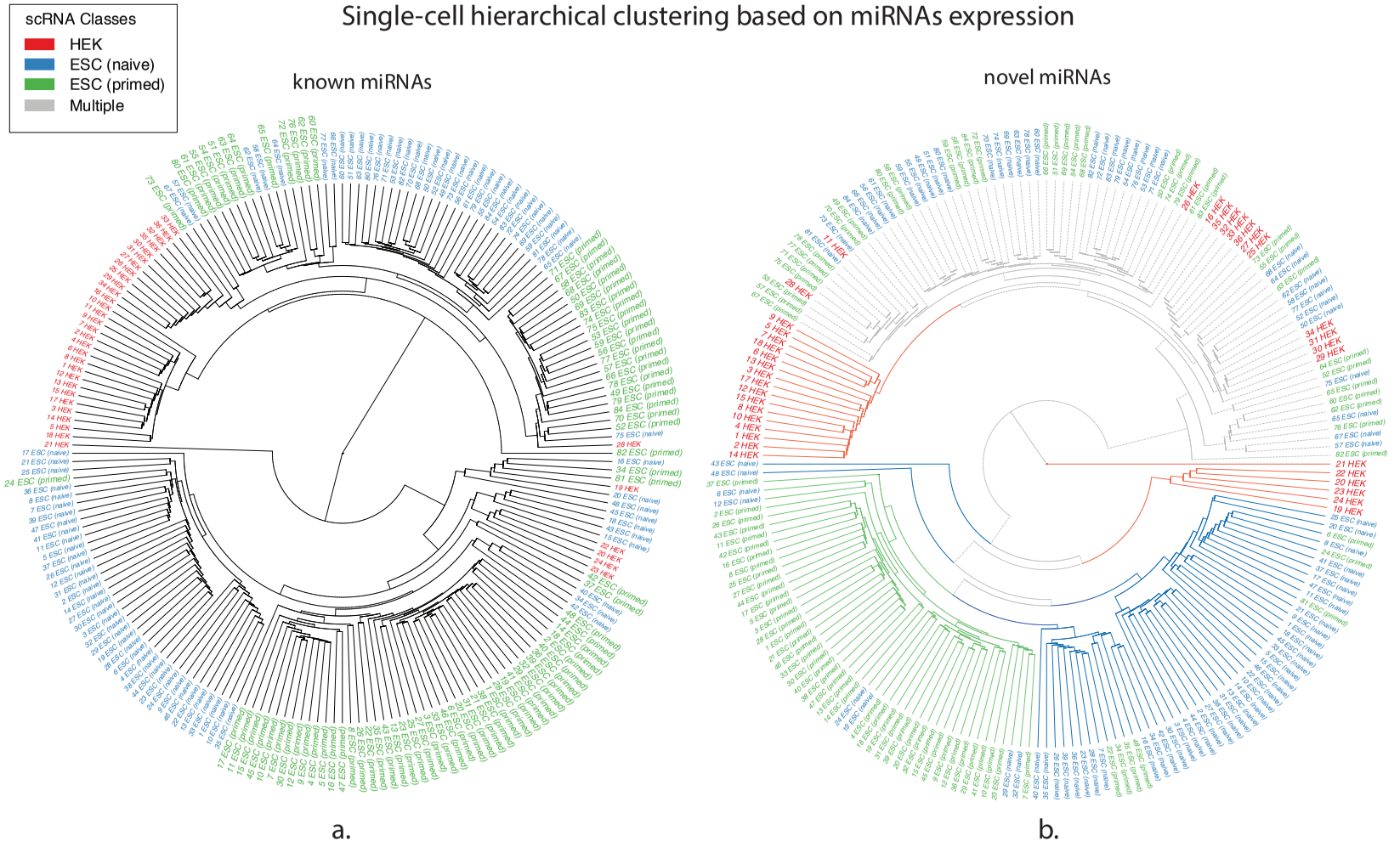



**Fig. 4.20** Hierarchical clustering of primed, naive human ESCs and HEK cells based on a) **known** and b) **novel** miRNA expression, inferred by mirnovo and quantified by Chimira.

This finding illustrates that individual cells may expose a unique novel miRNA signature that is characteristic of the cell type of origin while other cells may show a lower degree of differentiation and thus retain a more generic miRNA expression profile, regardless of their cell type.

## 4.4 Methods

### 4.4.1 Implementation and availability

Mirnovo's jobs are submitted to the EMBL-EBI high performance computing cluster. Each job process is extensively parallelised with multiple threads taking over calculations and processing over different subsets of the entire data and over different subtasks of the entire process. The job's progress is visualised in real-time through a console window at mirnovo's progress page in the browser.

Mirnovo is available as a stand-alone package besides the web-server version (Appendix B). The downloadable bundle contains all necessary scripts and binaries for execution of mirnovo, providing separate versions for either Mac OSX or Linux platforms. The only required dependencies for the local machine are: Perl (v5.24.1), Python (v2.7.10) and R (v3.2.2), with the recommended versions in parentheses.

### 4.4.2 Alignment against a reference genome

When the reference genome is available, following the miRNA prediction step, the consensus sequences of all identified known and/or novel miRNAs are mapped against the genome using *bowtie2* (Langmead and Salzberg, 2012). The selected parameters for the *bowtie2* call are as follows: *-k* 1, *-D* 20, *-R* 3, *-N* 1, *-L* 20, *-i* S,1,0.50 *–rdg* 1,1 *–rfg* 1,1.

### 4.4.3 Refined mature miRNA quantification with Chimira

Mirnovo is able to extract both hairpins and mature miRNAs in the output along with count data for the latter case. Due to inherent imperfection of sequence clustering steps during the mirnovo run, in some cases incorrect groupings of miRNAs may affect, even at a low level, the yielded expression data. In order to resolve this issue we have expanded Chimira, a method that we previously published in our lab. In this case, Chimira serves as a mirnovo extension, allowing the user to upload a custom set of hairpin sequences, e.g. known and/or novel hairpins predicted by mirnovo. All uploaded files are merged and sequences with an alignment identity over 0.90 are collapsed. Chimira then aligns

the input files against the merged reference sequences set to extract the mature miRNA expression counts. Additionally, Chimira is able to generate coverage profiles for each of the identified mature miRNAs and the secondary structures of the corresponding hairpin precursors (Lorenz et al., 2011).

### 4.4.4 mirnovo vs miRDeep2 benchmarking

miRDeep2 was always provided with the human reference genome, all known human hairpins, all known human mature miRNA sequences and also all mature miRNAs from two extra species (*D. melanogaster* and *C. briggsae*) for additional diversity. Mirnovo was tested both with and without the reference genome. With regards to the time benchmarking, mirnovo is a highly-parallelised multi-threaded method while miRDeep2 is serially processed. Thus, we wanted for the benchmarking to reflect the run time experienced by the end user. Both methods ran on HPC clusters consisting of 32-processor nodes equipped with the Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz CPU model. Mirnovo was run using the default number of hosts that is selected for each job (-n=3) while miRDeep2 was run using -n=1 (assigning -n=3 hosts to miRDeep2 proved to be slightly slower -data not shown- most likely due to synchronisation latency among the hosts, and thus one host was eventually assigned for benchmarking of the miRDeep2 runs). Both methods were provided with 8GB of memory (-M 8192).

### 4.4.5 Analysis of single-cell RNA-Seq data

Processing of single-cell RNA-Seq data follows the same core pipeline as regular small RNA-Seq data processing. The only exception is that due to high innate noise of single-cell data, coverage and sequence complexity features are not taken into consideration at the final classification step, and thus predictions are inferred by models that have been pre-trained solely based on the genomic features. Thus, in order to make predictions from single-cell data the option '*Use only genomic features for prediction*' needs to be enabled.

## 4.5 Conclusion

We have demonstrated that machine learning based, genome-free discovery of miRNAs is possible from small RNA sequencing in animal and plant species. Our approach has similar levels of accuracy to the most widely used previously published tool, which utilises genomic information (miRDeep2). Additionally, our approach exceeds miRDeep2's performance when genome information is available and does so at a significantly lower compu-

tational cost. This approach has been extensively validated using multiple species, training sets and 10-fold cross validation. Our method has been validated using large-scale datasets and miRNA biogenesis mutant datasets that elucidate potential novel miRNA biogenesis pathways, based on their dependency on different types of RNaseIII enzymes. We have also demonstrated the possibility of discovering novel miRNA candidates from single-cell data, despite their inherent noise, and thus further enable the discovery of novel miRNA molecules associated with very particular cell types and/or conditions.

Moreover, we observed a higher degree of consistency in predicting novel miRNAs in animals than in plants, in terms of the features with the most discriminative power, which complies with the presence of more diverse miRNA biogenesis mechanisms in plants. However, miRNA predictions in plants still managed high levels of accuracy and thus mirnovo can serve additionally as a formidable and easy-to-use resource for researchers of the plants community.

Finally, mirnovo, is simple to install as a command-line tool and may also be used as a user-friendly web-based method. Given the quality of results obtained without genome data, we believe this method could have an important role for miRNA discovery in non-model organisms. We believe that mirnovo represents a significant new contribution to the miRNA field and in particular to the prediction of novel miRNAs.

# Chapter 5

# MILI-independent piRNA biogenesis

*The results from this chapter have been published in the following paper*:
"A MILI-independent piRNA biogenesis pathway empowers partial germline reprogramming"
L Vasiliauskaitė, **DM Vitsios**, RV Berrens, C Carrieri, W Reik, AJ Enright & D O'Carroll.
***Nature Structural & Molecular Biology***, Vol. 24, p.604–606, doi: 10.1038/nsmb.3413 (2017).

## 5.1 Introduction

### 5.1.1 Transposable Elements

In the late 1940s, the world was still reeling amidst the turbulences and remaining debris in the aftermath of World War II. Meanwhile, at the same time, Barbara McClintock, a prominent American geneticist, was quietly discovering one of the most fundamental elements in Genetics, the Transposable Elements (McClintock, 1950). Transposable Elements (TEs) or 'jumping genes' as they are often called, are sequences of DNA that can change their location (jump) in the genome. The importance of TEs was largely dismissed for decades thereafter, since many scientists believed that these elements had no function at all in the genome and thus were referred to as 'junk DNA'. However, around 1965, Prof McClintock was again among the leading scientists suggesting that TEs were actually playing a regulatory role, defining the repertoire of active genes by turning on/off respective areas in the genome. Since then, TEs have been extensively studied although their function has yet to be defined definitively. Some studies report that TEs have a positive impact in genome evolution (Brandt et al., 2005) while others underline their role as mutagens and inducers of genomic instability via insertional mutagenesis (Deininger et al., 2003).

There are two classes of transposable elements, retrotransposons or Class I TEs and DNA transposons or Class II TEs. Class I TEs employ a "*copy-and-paste*" mechanism where they are initially transcribed into RNA, then reverse transcribed into cDNA sequences and finally inserted into the genome at various target sites. On the other hand, Class II TEs follow a "cut-and-paste" mechanism which does not involve an RNA intermediate but are rather catalysed by several transposase enzymes that allow them to translocate in the genome. Retrotransposons are highly abundant in eukaryotes. For instance, in maize, as a representative example of the plants kingdom, they comprise 49-78% of the genome (Sanmiguel and Bennetzen, 1998). In mammals, almost half of the genome (45-48%) is transposable elements and in human around 42% of the genome consists of retrotransposons while DNA transposons comprise only 2-3% of the entire genome (Kazazian Jr and Moran, 1998). The three major categories of Class I TEs are:

- TEs with long terminal repeats (LTRs, e.g. IAP family elements): which encode reverse transcriptase
- Long interspersed nuclear elements (LINEs, LINE-1s, or L1s): which also encode reverse transcriptase and are transcribed by RNA polymerase II
- Short interspersed nuclear elements; theses do not encode reverse transcriptase and are transcribed by RNA polymerase III.

The versatile nature of TEs enables them to relocate randomly in the genome and inflict similarly unexpected changes in gene expression. Thus, many organisms have developed various defence mechanisms against TEs in order to limit or more rationally regulate their activity. Some prokaryotic organisms, like bacteria, have developed defence mechanisms against TEs based on short repeats interspersed throughout their genome. These repeats are complementary to previously encountered viruses or mobile genetic elements and thus are able to target them and limit their replication. This natural mechanism, which is called CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats), has been adapted into an artificial technique for targeted genome editing, CRISPR/Cas9 (Jansen et al., 2002). Another almost ubiquitous defence system found in organisms from unicellular organisms and prokaryotes to plants and animals involves the RNA interference (RNAi) pathway aided by AGO proteins (Aravin et al., 2003; Djikeng et al., 2001; Hamilton and Baulcombe, 1999; Makarova et al., 2009). DNA methylation has also been an important factor for regulating transposable elements expression in some species of plants (Mosher et al., 2008) and animals (Aravin et al., 2008; Brandt et al., 2005) since it affects the chromatin structure and subsequently drives gene activation based on chromatin region accessibility.

In Animals, the main defence mechanism against transposable elements employs a class of small non-coding RNAs called piwi-interacting RNAs (piRNAs). This class of small RNAs is the most highly expressed one found in animal cells (Seto et al., 2007). However, their expression is restricted to germ-line cells and in mammals they seem to be indispensable exclusively to males (Siomi et al., 2011). piRNAs interact with other proteins to form RNA-protein complexes that are able to target complementary TEs and silence their expression. This mechanism is of great importance in germ-cells since it helps protect the genetic information that is going to be passed onto the offspring. We are going to elaborate more on the biogenesis features and functionality of piRNAs in the next three sections.

### 5.1.2 Piwi-interacting RNAs (piRNAs)

It was only a decade ago that piRNAs emerged as a novel class of small non coding RNAs. Multiple studies simultaneously confirmed the existence of piRNAs in mouse and rat cells (Aravin et al., 2006; Girard et al., 2006; Grivna et al., 2006; Lau et al., 2006; Watanabe et al., 2006). piRNAs are distinct from other non-coding RNA classes such as miRNAs and siRNAs in that their length is 26-31 nt and their production is independent of Dicer proteins (Vagin et al., 2006). Moreover, piRNA sequences demonstrate strong bias for a 5'-Uridine which indicates the involvement of RNAse III enzymes in their processing (Aravin et al., 2003). Like other small non-coding RNAs, while their 5' end contains a monophosphate group, their 3' end exhibits a distinct 2'-O-methylation (Kirino and Mourelatos, 2007b). piRNAs are organised into clusters throughout the genome. Cluster size may vary from 1kb to 100kb and the number of piRNAs located within ranges from just a dozen to several thousands of piRNAs (O'Donnell and Boeke, 2007). Though their sequences do not show any degree of conservation, they are highly syntenic and piRNA cluster positions are particularly conserved across species (Girard et al., 2006; Malone and Hannon, 2009). In mammals, piRNAs have been found in both testes (Aravin et al., 2006) and ovaries (Tam et al., 2008) although their role is fairly dispensable in females (Siomi et al., 2011). Intriguingly, piRNAs can be found in both the nucleus and the cytoplasm, suggesting their potential involvement in various and diverse functional pathways.

### 5.1.3 piRNA biogenesis in mammals and associated PIWI proteins

The current model of piRNA biogenesis in mammals includes two main pathways (Figure 5.1). In the primary biogenesis pathway, long transposon transcripts are exported from the nucleus into the cytoplasm and cleaved into fragments by the endonuclease Zucchini. This is believed to generate the 5' ends of primary piRNAs (Aravin et al., 2006; Ipsaro et al.,

2012; Nishimasu et al., 2012). The primary piRNAs bind to Mili protein molecules forming an RNA-protein complex that can target transposons or other primary transcripts (Aravin et al., 2007). This is the first step of the secondary biogenesis pathway and is also referred to as the 'ping-pong' cycle. The resulting anti-sense transcript has a 5' terminal Adenine following cleavage 10nt downstream from the 5'-Uridine end of the sense transcript. This is due to the so-called ping-pong signature, a 10nt overlap between the 5' ends of piRNAs associated with Mili and Miwi2 in mouse testes, which was first observed in *Drosophila Melanogaster* (Brennecke et al., 2007; Gunawardane et al., 2007). The antisense transcripts are then loaded onto Miwi2 and trimmed into secondary piRNAs. This allows the Miwi2-secondary piRNA complex to target sense transposons that are later loaded onto Mili and processed into transposon derived piRNAs. Thus, piRNAs follow a closed pathway of amplification that justifies the "ping-pong" annotation of this biogenesis cycle.

This biogenesis model comes with some limitations though. First of all, adult testes lack the ping-pong mechanism for piRNA biogenesis (Beyret et al., 2012). Furthermore, Mili and Miwi2 are not physically compartmentalised in the same granules (Aravin et al., 2009). Finally, Miwi2 proteins are no longer present in the adult stages of spermatogenesis, where pachytene piRNAs have already been formed (Carmell et al., 2007; Girard et al., 2006; Kuramochi-Miyagawa et al., 2008).

Another biogenesis mechanism which has been discovered only in *D. melanogaster* is called 'phasing' (Han et al., 2015; Mohn et al., 2015). This pathway comprises the nuclear branch of piRNA biogenesis in *Drosophila* since it only occurs in the nucleus of cells. Based on this pathway, piRNAs are generated at periodic intervals of approximately 27 nt from longer transcripts of the genome, so they have a 'phased' coverage profile. There is currently no evidence of phasing occurring in other organisms.

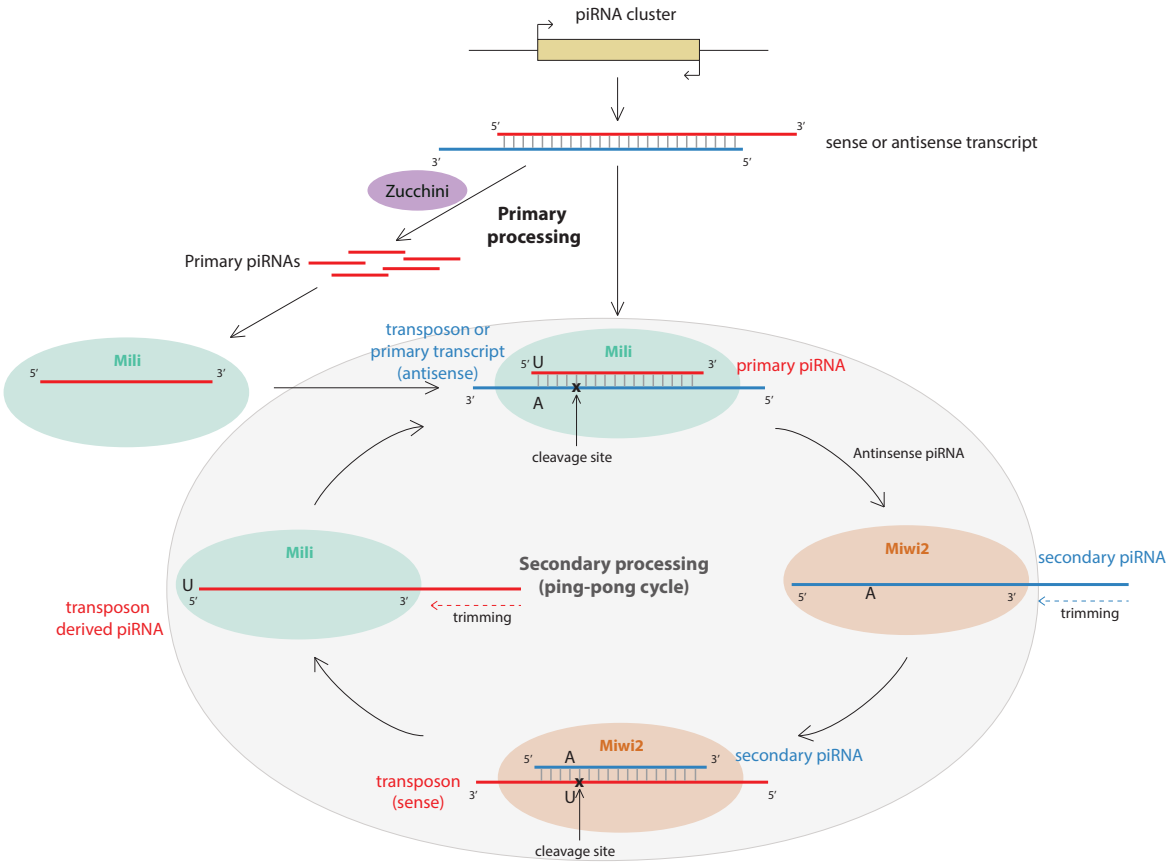



**Fig. 5.1** piRNA biogenesis model in mammals. Primary precursors are transcribed from piRNA clusters by an unidentified RNA polymerase and exported to the cytoplasm. These long precursors are excised by the endonuclease Zucchini into shorter fragments (around 26-31nt). This part represents the primary processing pathway of piRNA biogenesis. Subsequently, primary piRNAs bind to MILI ribo-nucleoproteins (RNPs) and target antisense transposons or primary transcripts. Eventually, a secondary piRNA is generated by cleavage of the antisense transcript, 10nt away from the 5' end of the primary piRNA. The generated secondary piRNA is then bound by MIWI2 RNP. This targets sense transposons and eventually creates a transposon derived RNA that can bind to MILI RNPs and thus ignite the same procedure from the beginning. This cycle of piRNA amplification represents the secondary piRNA processing pathway and is usually referred to as "ping-pong" cycle.

## 5.2 Results

### 5.2.1 Overview

In mice, piRNAs are bound to the cytoplasmic RNA endonuclease MILI, which plays an integral part for the amplification of secondary piRNAs, effectively inducing transposon repression (De Fazio et al., 2011). Subsequently, secondary piRNAs guide the activity of the PIWI protein MIWI2 through sequence complementarity. This induces de novo DNA methylation and silencing of certain classes of Transposable Elements, specifically LINE1

and IAP elements (Aravin et al., 2008, 2007; Carmell et al., 2007; De Fazio et al., 2011; Kuramochi-Miyagawa et al., 2008). Previous studies of MILI or MIWI2 deficiency in mice have shown inefficient DNA methylation and depression of LINE1s and IAPs, leading eventually to meiotic arrest (Aravin et al., 2007; Carmell et al., 2007; Kuramochi-Miyagawa et al., 2004). Furthermore, Miwi2 deficiency causes gradual loss of germ cells in mice and leads to full aspermatogenesis by 9 months after MIWI2 knockout (Carmell et al., 2007; De Fazio et al., 2011). However, when Mili deficiency was tested in mice by the O'Carroll lab, it showed that around 50% of tubules still remain spermatogenic even 1 year after birth. This observation motivated an effort to investigate whether MIWI2 can still play a role in germ cell reprogramming in the complete absence of MILI protein. The experimental work for this project was carried out at the O'Carroll Lab while I was responsible for the piRNA bioinformatic analysis in the Enright Lab.

### 5.2.2 DNA methylation subject to piRNA pathway

For the purpose of this study, the Mili-/Miwi2-knockout mouse samples that were used for the analysis had been previously produced in the O'Carroll Lab (De Fazio et al., 2011; Di Giacomo et al., 2013). The collaborators first performed whole-genome bisulfite sequencing on undifferentiated spermatogonia to determine the methylation patterns in the absence of Mili and Miwi2 proteins. It is already known that both MILI and MIWI2 disruption affect significantly LINE1 and IAP elements (Aravin et al., 2007; Kuramochi-Miyagawa et al., 2008; Molaro et al., 2014). Preliminary analysis of those samples yielded different degrees of impact from MILI or MIWI2 disruption in DNA remethylation. Specifically, MIWI2 deficiency had a more significant impact in genome remethylation than MILI. Overall, the collaborators identified 1,704 and 258 loci in the mouse genome, whose methylation is dependent on MIWI2 and MILI function, respectively.

Using this dataset of loci as a starting point, we moved on to investigate whether piRNAs are associated with their methylation. If there is a connection, piRNAs should induce silencing of these loci via methylation, by guiding ribo-nucleoproteins (RNPs) to these targets through sequence complementarity. With a fully functional piRNA pathway, i.e. in the presence of MIWI2 and MILI, we would expect that piRNAs are complementary to the genomic sequences defined by these loci. We sought to address this at the beginning of our analysis in the subsequent section.

### 5.2.3 piRNA bioinformatics analysis

In order to assess the degree of complementarity between piRNAs and the previous loci, we mapped the sequences from our mouse allele samples against these sets of loci, by quantifying the normalised number of aligned fragments within each region. Alignment results demonstrated that indeed piRNAs from both RNPs were complementary to the respective sets of loci (Figure 5.2). We only profiled the uniquely mapping reads, to exclude transposon-associated piRNAs, and we found that complementarity of remaining piRNAs against these regions is still intact. These results confirm our hypothesis that methylation of the two sets of loci is in fact directed by the piRNA pathway. Moreover, since MIWI2 affects a notably larger number of genomic regions with regards to their methylation, its presence seems to be of greater significance than MILI's in the canonical piRNA pathway.

Since MILI is less important than MIWI2 in piRNA biogenesis, we can hypothesise that even in the absence of MILI, there may still be a functional piRNA biogenesis pathway which is dependent exclusively on MIWI2. In order to test this hypothesis, small RNA samples from wild-type and Mili-knockout fetal testes were produced by the O'Carroll Lab. An initial assessment of the length distribution of expressed transcripts in the two conditions showed that 16% of all piRNAs (sequences of 25-32nt) remained expressed in the Mili-knockout samples (Figure 5.3a,b). When we examined the differential expression between all unique piRNA sequences expressed in WT and Mili-knockout samples, we noticed that no novel transcripts were generated in the absence of Mili, compared to the wild-type condition (Figure 5.3c). This observation implies that the loss of the Mili enzyme is not causing extensive perturbation in the transcriptional machinery but rather preserves the main piRNA biogenesis pathway intact, albeit at a lower level of production.

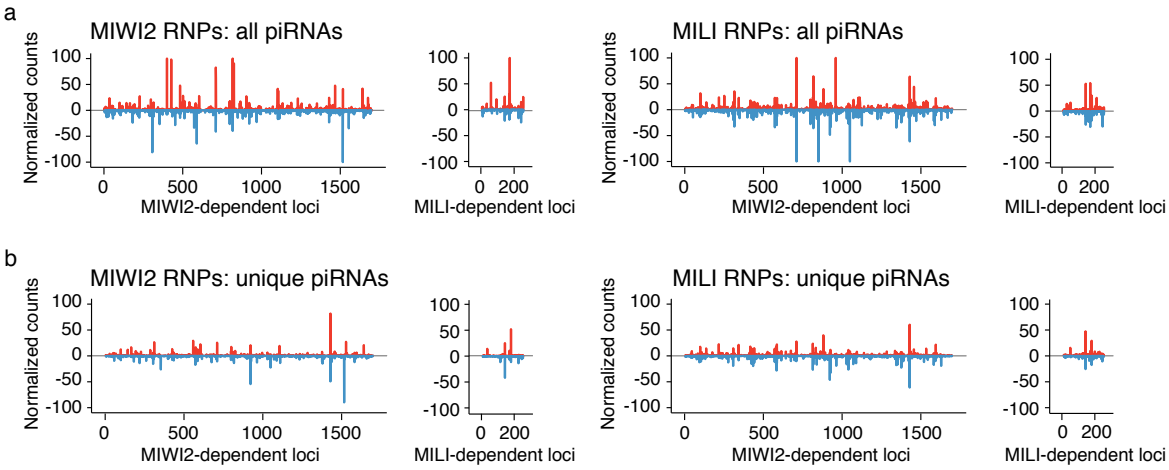



**Fig. 5.2** (a) All piRNAs and (b) uniquely mapping piRNAs from MIWI2 and MILI RNPs mapped to loci whose methylation is dependent upon MIWI2 and MILI, respectively. Positive (red) and negative (blue) values indicate sense and antisense piRNAs, respectively. Normalised counts refer to the number of aligned fragments within each region, divided by the size of the region per kb. Graphs have been generated by averaging values obtained from two sample replicates.

We then wanted to examine the degree of alignment of all piRNAs to the MIWI2 and MILI-dependent loci, in the wild-type and Mili-knockout conditions. We observed that piRNAs originating from the 1,704 MIWI2-dependent loci are still present in the Mili-knockout sample, though at lower levels (Figure5.4). By quantifying the densities per kilobase for the piRNAs that are cognate to this set of 1,704 loci we noticed a 2-fold (P-value < 0.05) and 3-fold (non-significant) reduction in multiply or uniquely mapping piRNAs, respectively (Figure5.5). On the other hand, piRNAs cognate to the 258 MILI-dependent loci showed a much higher statistically significant (P-value < 0.005) reduction especially in the case of unique mappers. Specifically, we observed a dramatic 18-fold decrease in piRNA expression of uniquely mapping piRNAs in the Mili-knockout samples as opposed to the wild-type condition. This result confirms that the loss of MILI protein extensively disrupts piRNA production from the loci whose methylation is dependent upon MILI, while there is a much milder effect in piRNAs which are dependent only on MIWI2.

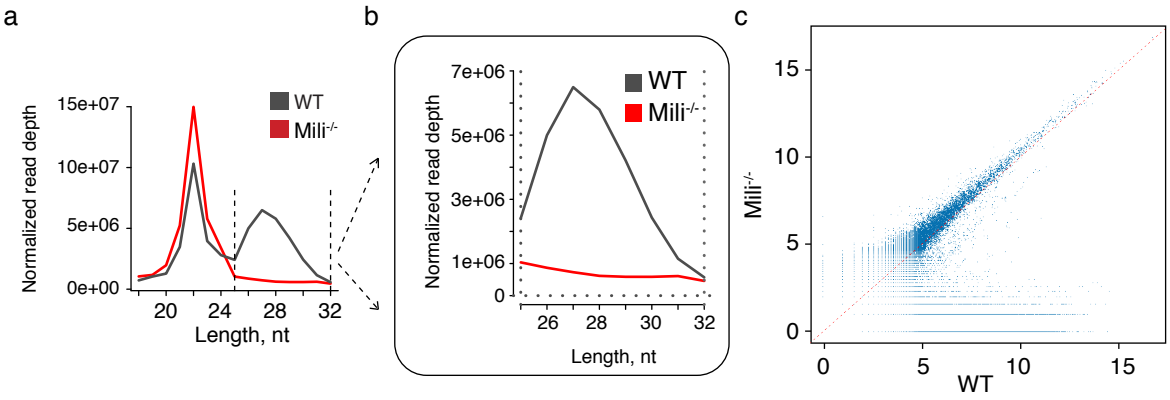



**Fig. 5.3** Length distribution of (a) small RNAs and (b) piRNAs only in wild-type (WT) and Mili–knockout fetal testes samples. (b) Differential piRNA expression in WT and Mili–knockout fetal testes samples. Blue dots represent read counts from individual (unique) piRNA sequences in log2 scale.

Next, we analysed the effect of Mili-knockout in the expression of the well-studied IAP and LINE1 transposable element families, where piRNAs can also originate from (Aravin et al., 2007; Carmell et al., 2007; Kuramochi-Miyagawa et al., 2004). We noticed a notably high drop in expression of those elements in the absence of the MILI protein (Figure5.6a,b,d). Moreover, we sought to identify evidence for the existence of the secondary piRNA biogenesis pathway. Thus, we examined the prevalence of the 1U-10A pattern and 10nt overlaps (ping-pong signature) across piRNA transcripts in both the wild-type and Mili-knockout samples. We noticed that the 1U-10A pattern is significantly decreased (P-value < 0.0005) and additionally the ping-pong signature is no longer present in the Mili-knockout samples (Figure5.6c, e). As a result, we may assume that piRNA biogenesis in the absence of MILI is not directed at all by the ping-pong amplification cycle but may rather be dependent on the primary biogenesis pathway. Instead, there may be an alternative biogenesis pathway, including the phasing mechanism. We are going to look into this hypothesis in our next step of our analysis, in order to decipher any hidden piRNA biogenesis mechanisms in mice.

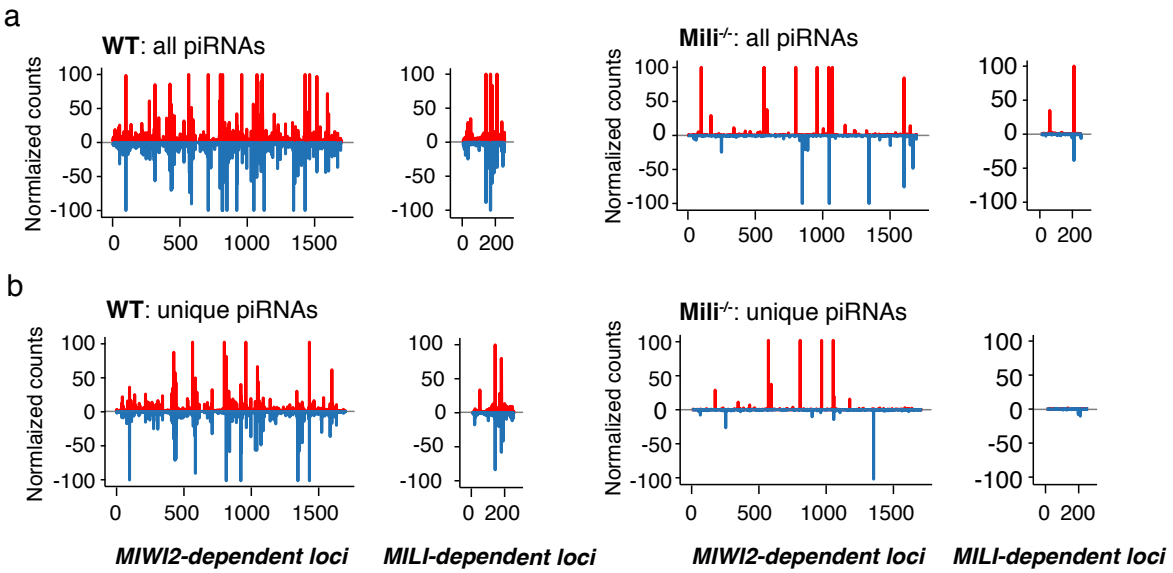



**Fig. 5.4** (a) All piRNAs and (b) uniquely mapping piRNAs from WT and Mili–knockout fetal testes samples mapped to loci whose methylation is dependent upon MIWI2 and MILI, respectively. Positive (red) and negative (blue) values indicate sense and antisense piRNAs, respectively. All graphs in a,b have been generated after averaging of the values across two replicates.

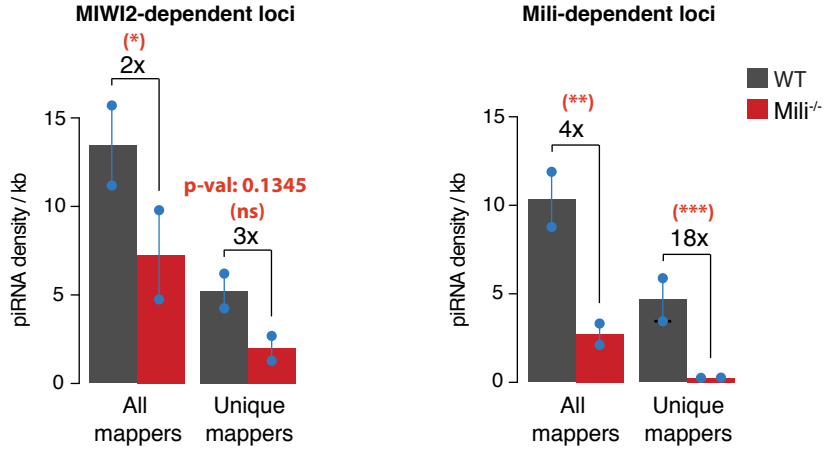



**Fig. 5.5** Quantification of all or uniquely mapping piRNA densities per kb across 1,704 MIWI2- and 258 MILI-dependent loci. Significance was assessed using the BootstRatio algorithm (Clèries et al., 2012). * indicates a P-value of < 0.05, ** indicates a P-value of < 0.005, *** indicates a P-value of < 0.0005.

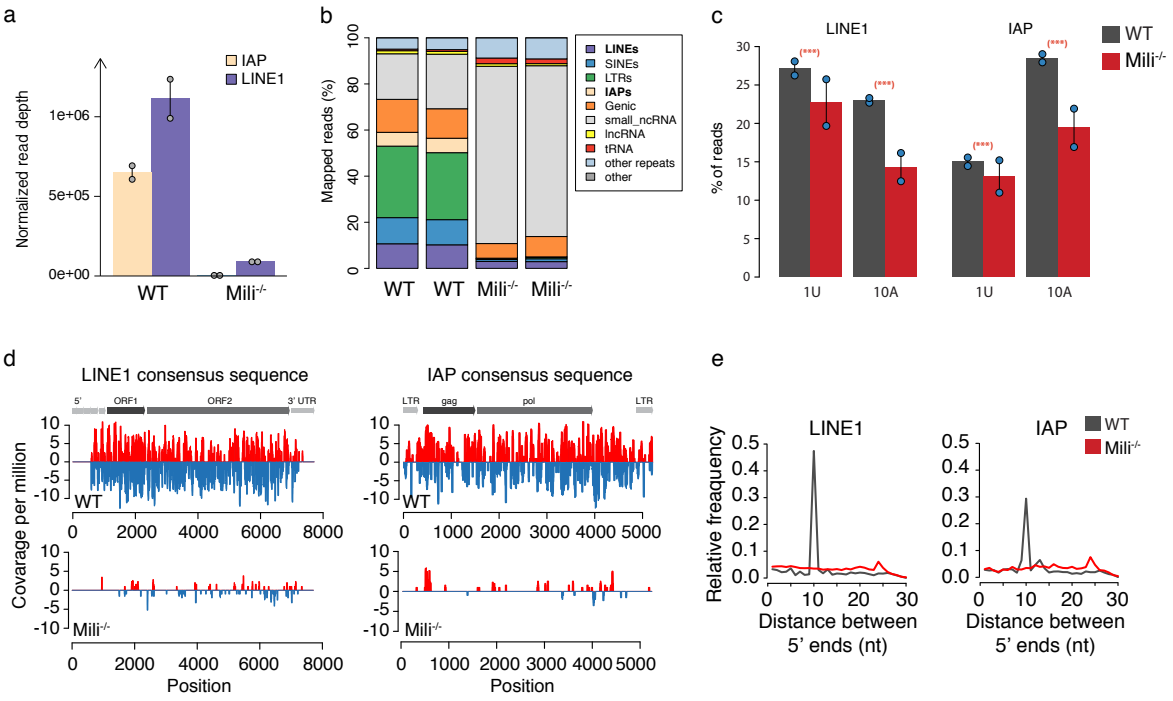



**Fig. 5.6** (a) Quantification of LINE1 and IAP associated piRNAs in wild-type (WT) and Mili-knockout fetal testes. Individual data points represent values from two biological replicates. (b) Distribution of non-coding transcripts expression across the WT and Mili-knockout samples. (c) Percentage of LINE1 and IAP associated piRNAs in WT and Mili-knockout fetal testes with a U at the first position (1U) without an A at the position 10 and an A at position 10 (10A) without a U at position 1 is shown. Individual data points again represent values from two biological replicates. (d) piRNAs from WT and Mili-knockout fetal testes mapped to LINE1 and IAP consensus sequences allowing up to three mismatches. Positive (red) and negative (blue) values indicate sense and antisense piRNAs, respectively. An averaged value from biological duplicates is shown. (e) Ping-pong analysis of LINE1 and IAP associated piRNAs in WT and Mili-knockout fetal testes. The frequency of the distance between 5' ends of complementary piRNAs from LINE1 (right) and IAP (left) elements is shown (averaging across two replicates). Significance in (c) was assessed using BootstRatio algorithm (Clèries et al., 2012). * indicates a P-value of < 0.05, ** indicates a P-value of < 0.005, *** indicates a P-value of < 0.0005.

In this last section, we are going to look into the potential presence of phasing, as an alternative biogenesis mechanism for piRNAs in mice. Phasing, as a piRNA biogenesis pathway, has so far been discovered only in *Drosophila melanogaster* (Han et al., 2015; Mohn et al., 2015). piRNA complementary targets are first cleaved into long transcripts. These transcripts are then further processed into sequential fragments which are generated from splicing of the longer transcript at periodic intervals of approximately 27 nucleotides. We sought to determine if we could find such periodicity in the piRNA clusters that are cognates of the MIWI2 or MILI-dependent loci. Specifically, we wanted to look for enrichment of any nucleotide distance between consecutive piRNA clusters in the mouse genome. Thus, we aggregated the distances of consecutive piRNA clusters, weighted ei-

ther uniformly, as an average score or as a product value (Figure5.7), in order to see if there is any prevalent distance peak among the entire set of recorded neighbouring distances.

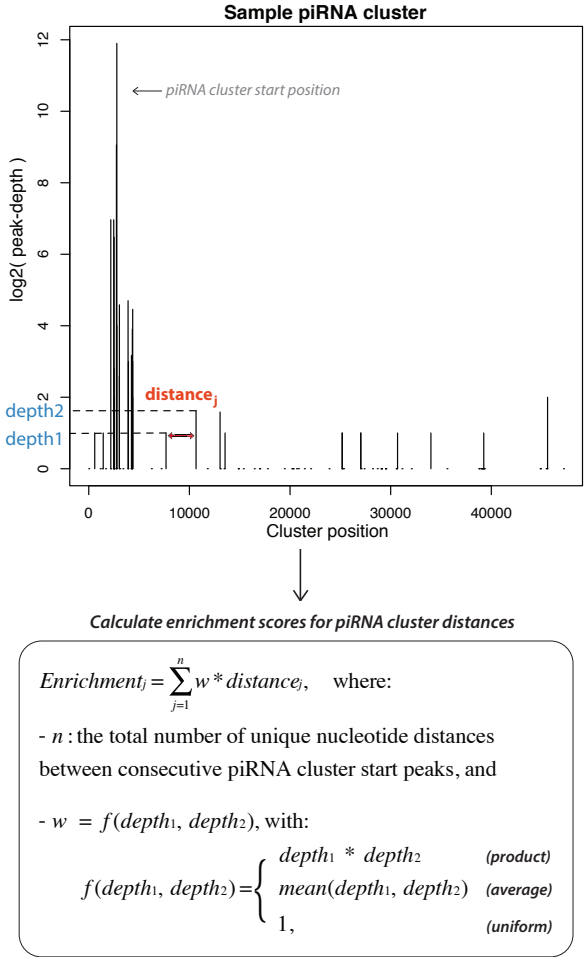



**Fig. 5.7** Phasing calculations approach across all piRNA clusters. For each pair of consecutive peaks the distance is first calculated. A weight $w$ is then multiplied with the calculated distance. The types of weights used are either: (a) product of depths ($w = depth_1 * depth_2$), (b) average value of depths ($w = \frac{depth_1 + depth_2}{2}$) or (c) uniform ($w = 1$). All weighted unique distances are then aggregated in order to be used later for peak enrichment profiling analysis.

At first, we noticed a large peak at the 36-nt distance but only in one wild-type replicate and for the MIWI2-dependent loci (Figure5.8). Further investigation of this peak revealed that it actually corresponds to a single pair (outlier) of over-expressed consecutive piRNA clusters. Apart from that, we observed a distance enrichment at or around 38 nucleotides (Figure5.8), in both replicates of the MIWI2-dependent loci and for one replicate of the MILI-dependent loci (only in the Mili-knockout condition in both cases). This finding may not present strong evidence for the existence of phasing in mice. However, it does represent support for the existence of this pathway at periodic intervals of around 38nt. Addi-

tionally, this pattern is more prevalent in the MIWI2-dependent loci, since it is observed in both replicates even though not at the exact same distance peaks. Thus it may be possible that an innate phasing pathway in mice is responsible for the biogenesis of piRNAs identified in the samples analysed in this study, despite the loss of the MILI protein.

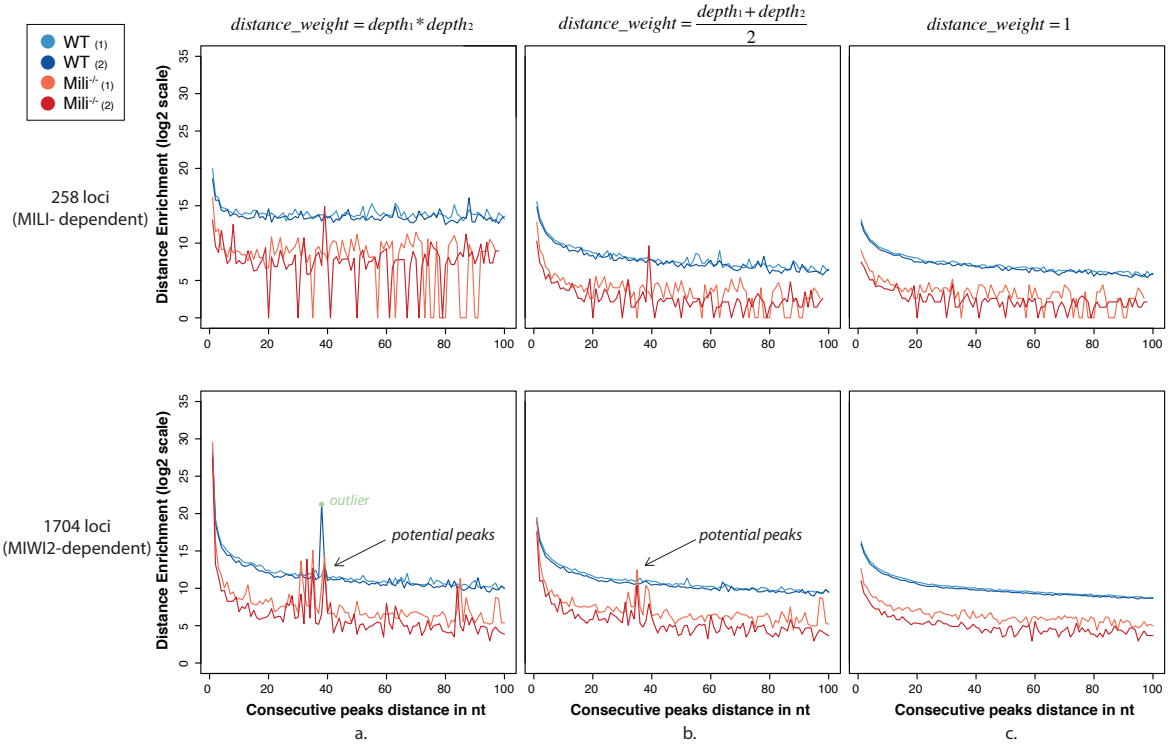



**Fig. 5.8** Calculations over the 1,704 MIWI2 and 258 MILI-dependent loci for phasing detection as a potential piRNA biogenesis mechanism. Blue graphs depict weighted distance enrichment across the wild-type replicates while red graphs correspond to the Mili-knockout replicates. The enrichment of unique distance values between piRNA clusters has been calculated using three different sets of weights: (a) $w = depth_1 * depth_2$, (b) $w = \frac{depth_1 + depth_2}{2}$ and (c) $w = 1$.

## 5.3 Conclusion

Experimental data from Bisulfite sequencing of mouse allele samples produced by the O'Carroll Lab (De Fazio et al., 2011; Di Giacomo et al., 2013) revealed the presence of novel loci of piRNA clusters, which are either MIWI2 or MILI-dependent. We investigated the effect of MILI knockout on piRNA expression, with reference to these loci and other known transposable elements (IAPs and LINE1s). We observed a 84% drop in piRNA expression (24-32nt long sequences). This included potentially both the LTR and IAP elements and the new sets of piRNA clusters, whose expression was verified in the first step of our analysis. However, the remaining 16% of piRNAs that were still expressed in the

Mili-knockout samples indicates that piRNA biogenesis may not be dependent at all on the Mili enzyme in some cases. Indeed, we showed that some piRNA clusters from the 1,704 MIWI2-dependent loci were still expressed in the absence of Mili. This suggests the existence of a non-canonical piRNA biogenesis pathway. Moreover, when looking at the differential expression between the wild-type and Mili-knockout samples we did not notice any novel transcripts exclusively present in the knockout condition. This indicates that the loss of the MILI enzyme is not inducing extensive perturbation in transcriptional machinery. Finally, we attempted to explain MILI-independent piRNA biogenesis via a *Drosophila*-like phasing mechanism. We did not find a strong enrichment signal for any unique nucleotide distance. However, a distance enrichment of around 38nt may explain the origin of piRNAs in the absence of Mili, especially of those originating from the MIWI2-dependent loci.

## 5.4 Methods

### 5.4.1 Data pre-processing and cleaning

All samples were initially aligned against Rfam (Nawrocki et al., 2014) in order to filter out tRNA sequences (tRNA hits with an alignment identity score > 90% were excluded from the rest of the analysis). Analysis of the filtered samples was then performed using SequenceImp (Davis et al., 2013). Input reads were first trimmed from the 3' adapter with reaper (using default configuration for read geometry without barcode) and de-duplicated with tally, which are both part of the Kraken suite of tools (Davis et al., 2013). The length distribution of all cleaned reads of between 18 and 32 nucleotides was recorded in order to check for depletion of piRNA sequences between the wild type and Mili-knockout conditions.

### 5.4.2 Normalisation across samples

In all cases, normalisation was performed based on the total number of reads of transcripts that remained unchanged between the two conditions(Figure5.9). Specifically, the types of transcripts that were used for normalisation (based on the official Ensembl genebuild annotation) were: miRNAs, rRNAs, snRNAs, snoRNAs, processed_transcripts, aa_tRNAs, Mt_tRNAs, other_tRNAs, miscRNAs and RNA_repeats. Significance was assessed using the BootstRatio algorithm (Clèries et al., 2012).

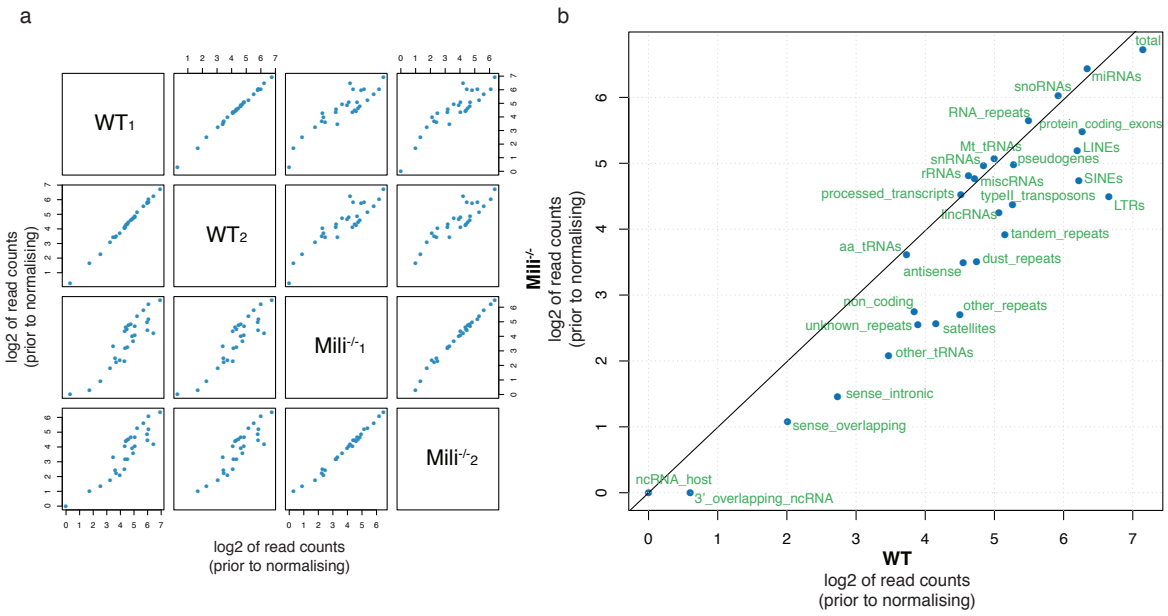



**Fig. 5.9** (a) Replicates from both the wildtype and Mili-knockout conditions show strong resemblance within each condition type allowing for efficient evaluation of statistical significance of the results. (b) Classes of transcripts that retained unchanged expression between the wildtype and Mili-knockout conditions have been selected as a reference for the normalisation of expression of all transcripts in both conditions.

### 5.4.3 Length filtering and piRNA quantification

Cleaned reads were later filtered by length (retaining only 24-32 nucleotide long sequences) and aligned against the Mouse genome (Ensembl release 66) allowing up to 2 mismatches and reporting up to 20 hits per sequence, when analysing for all mappers. In the case of uniquely mapped sequences, the *bowtie* call from the *SequenceImp* pipeline was adjusted using the parameter $-m = 1$ (parameter $-k$ was set to 1 for both the unique and all mappers cases). BAM output files from the alignment step were intersected with BED files containing the coordinates of 1,704 and 258 loci, whose methylation is dependent on MIWI2 and MILI, respectively. piRNA counts within each locus were calculated as the average number of fragments aligning against the locus, divided by the size of the locus region in 1 kb units. Expression densities were restricted to the interval (-100, 100) in order to filter out outliers from 3 overexpressed loci and thus increase densities resolution for all loci.

With regards to the piRNA differential expression analysis, a custom database of all 26-31 nucleotide long unique sequences found across all wild type and Mili-knockout replicates was initially built. Each sample replicate was then aligned against this database and expression levels of all matching sequences were quantified between the two conditions. As for the quantification of LINE1 and IAP repeats, the analysis was performed using the 'features' step of SequenceImp for repeat elements, allowing up to 3 mismatches and cor-

recting the read counts to the number of genome mapping reads. The ping-pong signatures and 1U-10A content of the LINE1 and IAP elements were also calculated as part of this step using the same method.

# Chapter 6

# Insights from miRNA targets editing in *D. Melanogaster* with CRISPR/Cas9

*Results from this chapter have been published in the following paper*:
"In situ functional dissection of RNA cis-regulatory elements by multiplex CRISPR/Cas9 genome engineering"
Q Wu*, Q Ferry*, Y Michaels, TA Baeumler, **DM Vitsios**, O Habib, R Arnold, X Jiang, S Maio, BR Steinkraus, M Tapia, P Piazza, N Xu, GA Holländer, TA Milne, JS Kim, AJ Enright, AR Bassett, Fulga T.
***Nature Communications***, Volume 8, p.2109, doi: 10.1038/s41467-017-00686-2 (2017).

## 6.1 Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR) are short, partially palindromic repeated DNA sequences found in the genomes of prokaryotic organisms. In bacteria and archaea, CRIPSR is an endogenous adaptive immunity mechanism which protects them against viruses and plasmids. This defence mechanism has three steps of action. First, a segment of the invading nucleic acid is cleaved by a Cas protein complex and integrated into the CRISPR locus as a novel spacer sequence. Then, the modified CRISPR locus is transcribed into crRNAs (CRISPR RNAs) with the aid of another Cas protein and tracrRNAs. Finally, mature crRNAs are incorporated into a third Cas complex that can target the foreign nucleic acid. Specifically, the crRNA guides the complex to the invading nucleic acid due to its complementarity with part of it and then the Cas proteins cleave and eventually degrade it.

The CRISPR mechanism was first described in 1987 (Ishino et al., 1987). Research on that field grew gradually for the next several years and it was only shown in 2012 that CRISPR could be used for genome editing in human cell cultures (Jinek et al., 2013). Since then, CRISPR has been established as the most accurate, cheap and efficient genome engineering/editing tool and it has been used in a wide range of organisms including mice (Wang et al., 2013), monkeys (Guo and Li, 2015), zebrafish (Hwang et al., 2013), fruit flies (Gratz et al., 2013) and yeast (DiCarlo et al., 2013; Liu et al., 2016; Zhang et al., 2014) and recently even in human (Cyranoski, 2016).

## 6.2 Overview

In this study, we briefly collaborated with Dr. Fulga's Lab from the University of Oxford in order to extract any patterns of regulation related with CRISPR/Cas9 induced gene editing. Our collaborators had first applied CRISPR/Cas9 in order to edit genes in the *Drosophila Melanogaster* genome, using S2R+ cell lines. The targeted genes were all coding and non-coding mir-184 targets (88 overall, Appendix D) that were computationally predicted by miRanda (Enright et al., 2003) and TargetScan (Agarwal et al., 2015) or experimentally verified in the lab (Hong et al., 2009; Kertesz et al., 2007). These targets will be referred as MREs (microRNA Response Elements) for the rest of this work.

Custom single guide RNAs (sgRNAs) were designed to target the defined MREs and cause cleavage at the respective areas of the genome. CRISPR-MIT (http://crispr.mit.edu/) was used for the design of a distinct sgRNA for each of the targeted MREs, taking into account both NGG and NAG protospacer adjacent motifs (PAMs). DNA cleavage induced by CRISPR/Cas9 eventually causes various types of gene disruptions including deletions and/or insertions at or around the target sites, through the normally occurring DNA repair mechanisms. This means that by applying the CRISPR/Cas9 method at the same target sequence in multiple sites we could retrieve in the end a pool of sequences that correspond to either wild type reads or to one or multiple types of mutants created by CRISPR/Cas9 (Figure 6.1). DNA and RNA was then extracted from the S2R+ cells and gDNA and cDNA libraries were prepared (2 replicates for each).These libraries were processed in order to contain only 150 nt long amplicons around each of the MRE sites. In the end, all amplicons were PCR amplified, polled together and sequenced.

The aim of this study is to analyse the diverse pool of MRE mutants in order to elucidate the extent and structure of CRISPR/Cas9 effect in a range of 88 different targets. This analysis will also incorporate a correlation study of the CRISPR efficacy with genome

accessibility based on computational prediction of the secondary structure of genomic sequence segments.

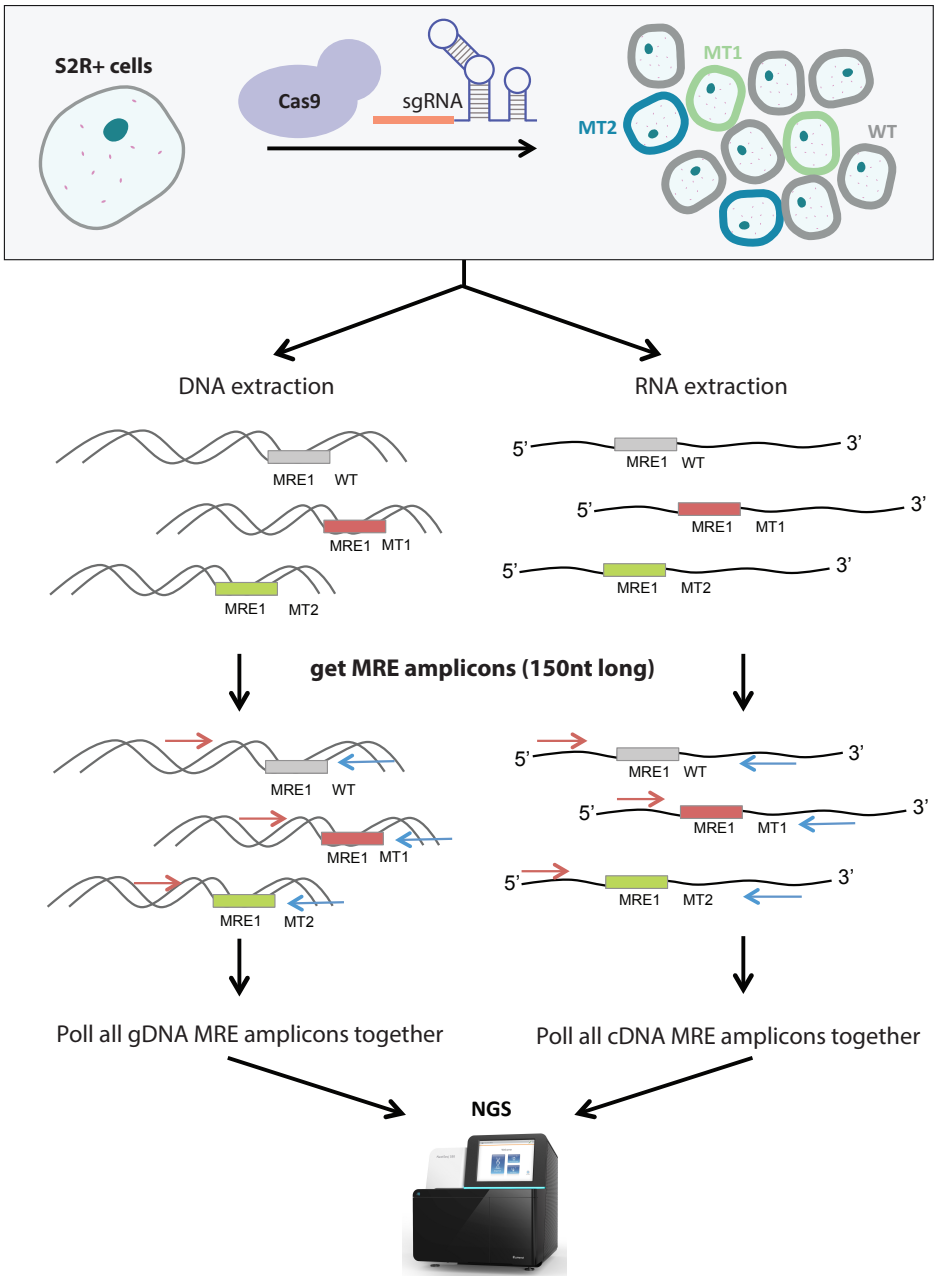



**Fig. 6.1** Experimental design for the deletion of all mir-184 targets in S2R+ cells using CRISPR/Cas9 and sequencing of the constructed gDNA and cDNA libraries.

## 6.3 Results

We retrieved four samples overall from Next Generation Sequencing: two sample replicates (A, B) for the *gDNA* MRE amplicons and another two replicates (C, D) for the *cDNA* amplicons. Each sample contains sequences (either wild type or modified) from all 88 MREs. Another 48 barcoded calibration sequences had also been added in order to verify the dilution of the samples along the experimental process. These sequences served only as experimental validation for the collaborators' work and were not included in the rest of the computational analysis.

Due to the complexity of experimental design we first needed to efficiently de-multiplex all sequences from input samples and assign them to the right MRE class or calibration sequence. In the next section, we are going to go through the methodology we followed for efficient sequence de-multiplexing. Following this, we will continue with the analysis of the CRISPR/Cas9 induced edited miRNA targets.

### 6.3.1 Demultiplexing and reads classification

In order to demultiplex each of the libraries we have applied the following steps:

1. Align library reads against all 136 sequences of interest (88 MRE amplicons + 48 calibration reads) using the Needleman-Wunsch algorithm. The Needleman-Wunsch (Needleman, 1970) algorithm was the ideal option since, as a global alignment technique, it can capture alignments having potentially big gaps (which we expect to have due to CRISPR/Cas9 activity).

2. Following alignment, each read is either assigned to a specific MRE bin or is classified as calibration read.

3. The barcode of each calibration read is checked (4 or 3 nt long) and the read is assigned to a specific calibration bin.

4. For each MRE/calibration, reads with a number of mismatches over 10 nt are discarded since they actually correspond to unsuccessful alignments.

5. All filtered reads of each bin are aligned with BLAST (McGinnis and Madden, 2004) against the respective wild type sequence of the bin in order to discard any remaining sequencing artefacts.

We observed that a proportion of reads in some of the MREs after initial classification was actually mis-classified, most likely due to sequencing artefacts. That is why we introduced *Step 5* in the demultiplexing process, which is to make sure that the highest majority of mis-classified reads from *Step 1* are discarded from the rest of the analysis (Figure 6.2a). Following that step, we computed the read depth for each MRE, normalised by read depth across all libraries (Figure 6.2b), to assess the quality of our input data.

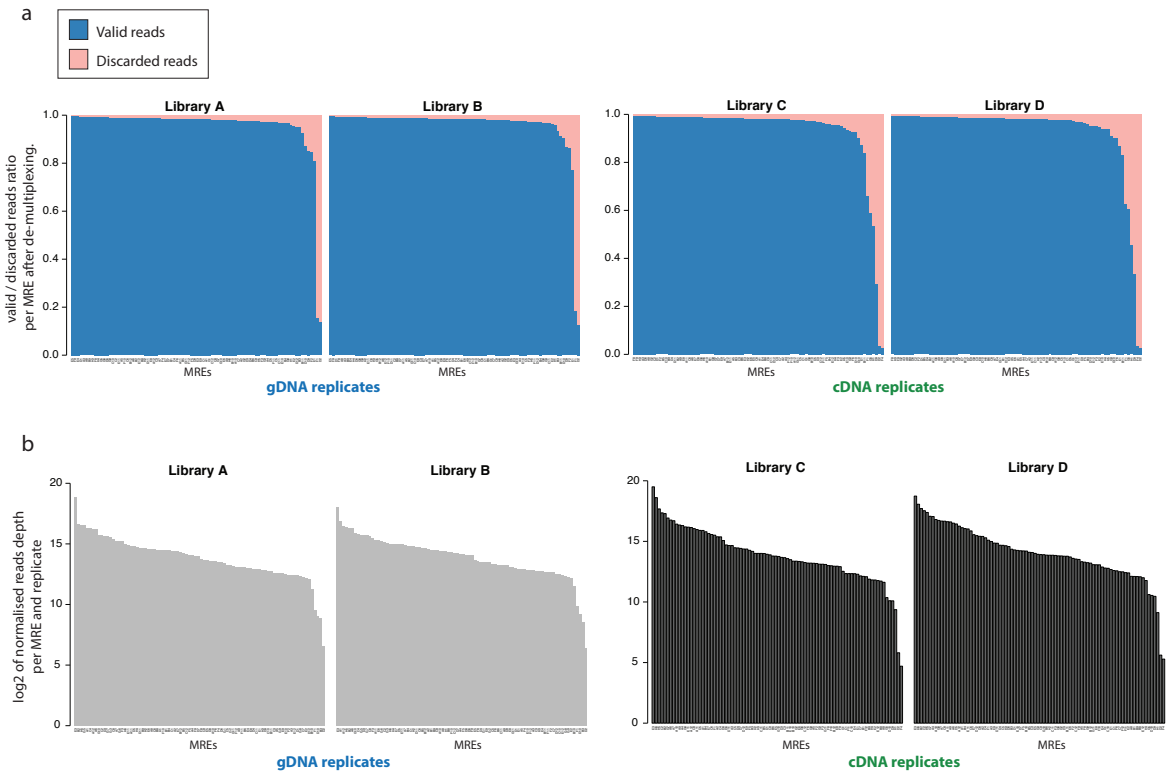



**Fig. 6.2** a) Quality check for final binning after de-multiplexing of all library replicates, b) Normalised (log2) read depth for all MREs and library replicates after refined de-multiplexing.

Following assignment of all reads into a read class (either MRE or calibration read) we analysed the distribution of variant types for all MREs across the gDNA and cDNA replicates. First, reads were classified into four categories: wild-type and another three types of variants (full seed deletion, partial seed deletion or insertion/deletion outside seed region) based on the position of the insertions/deletions caused by CRISPR/Cas9 relative to the seed region (Figure 6.3). We observe that the variant classes with CRISPR/Cas9 deletions in the seed region or outside (and close to) it are comparable in size. Moreover, there is a higher ratio of mutant reads in the cDNA replicates compared to the gDNA library (Figure 6.3). However, if we look at the individual per MRE mutant read ratios in the gDNA and cDNA libraries (Figures 6.4 and 6.5) we can see that ratios in the two libraries vary

for each MRE. This is a strong evidence that although CRISPR/Cas9 is targeting the same seed sequence in all these MREs, the effect it induces is quite variable. Thus, we can make the hypothesis that CRISPR's efficiency may be regulated by other factors, for instance accessibility of the target. Finally, if we consider as mutant reads only those with at least one deletion or insertion inside the seed region, after comparing the edits in gDNA and cDNA libraries we observed a clear shift of mutants towards the cDNA library (and respectively wild-type reads towards the gDNA library). This confirms that, overall, the majority of MREs have been effectively targeted by CRISPR/Cas9 and are actively transcribed after the induced mutation events (Figure 6.6).

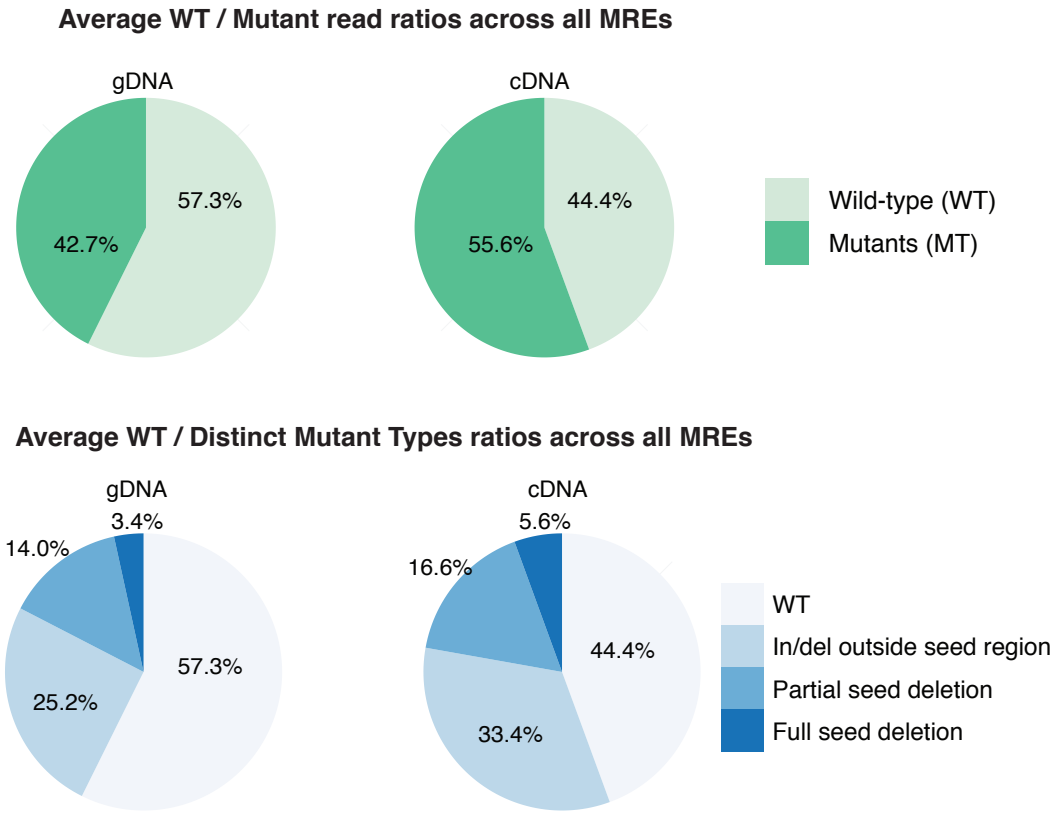



**Fig. 6.3** Summarised ratios of WT and MT reads per MRE, based on the relative CRISPR effect position to the seed region.

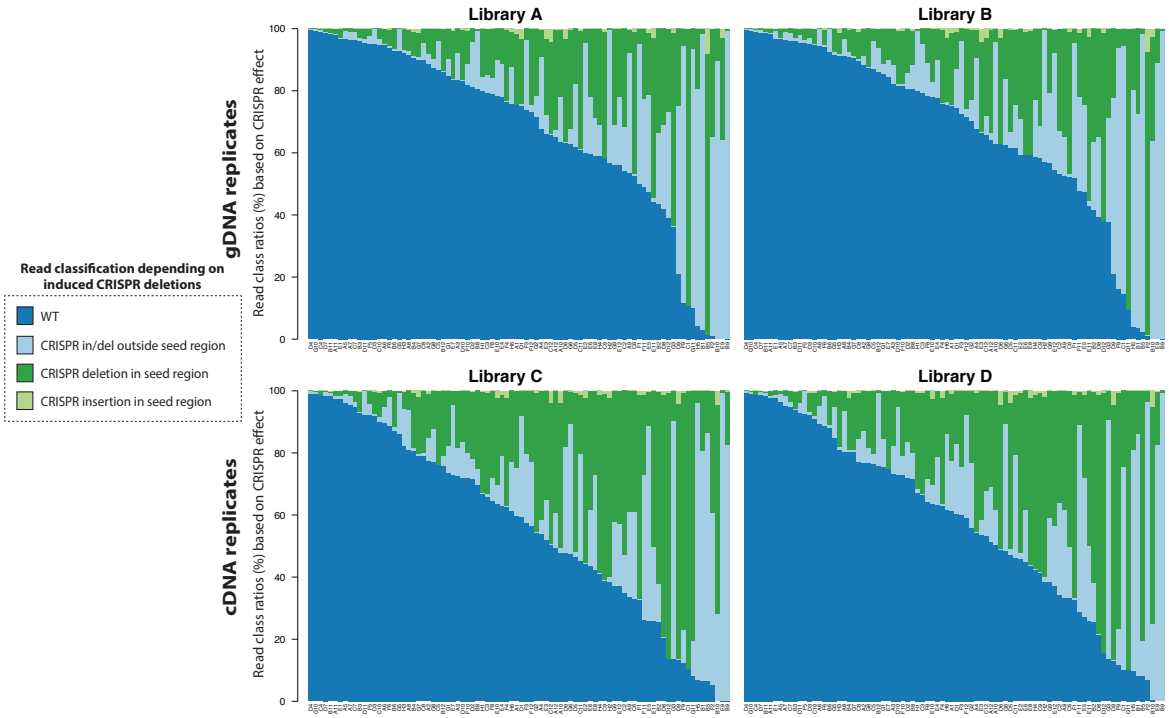



**Fig. 6.4** Reads classification per MRE into WT and 3 classes of mutants based on the position and type of CRISPR effect.

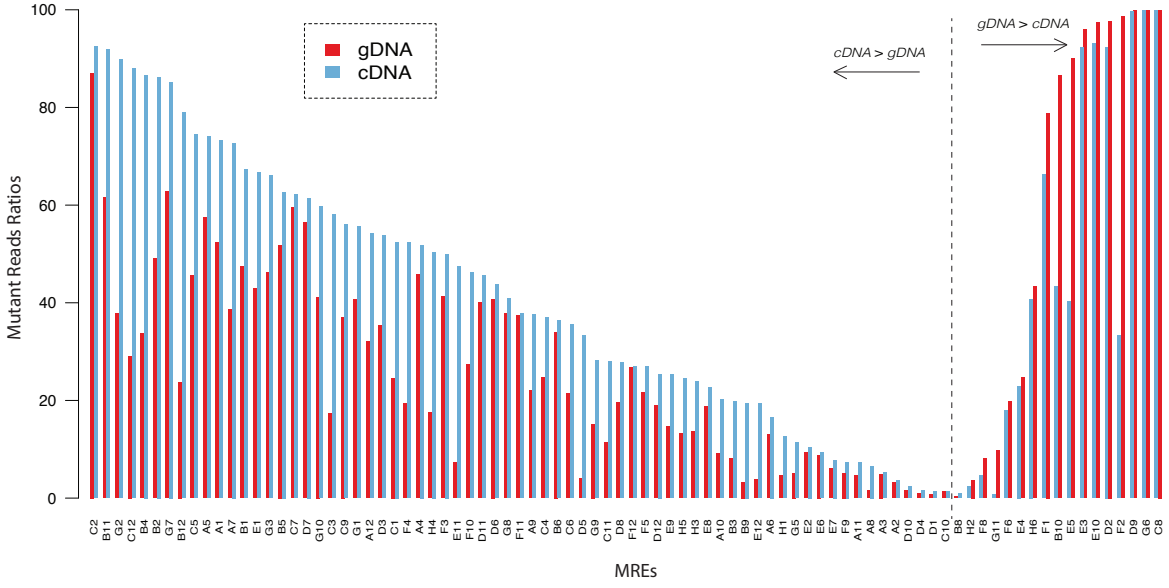



**Fig. 6.5** Mutant respective reads ratios in gDNA and cDNA libraries, for each MRE.

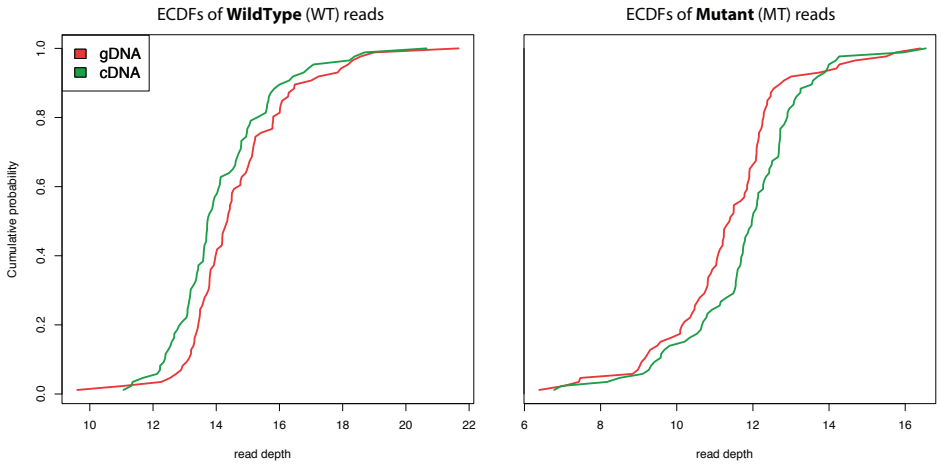



**Fig. 6.6** Shift of cumulative distribution functions for the numbers of WT and MT reads in the gDNA and cDNA libraries. *ECDF*: empirical cumulative distribution function.

### 6.3.2 Deletion profiles across read regions

Our analysis showed an overall enrichment of mutant reads in the cDNA library and at the same time a variance in the mutant ratio change across all MREs. Next, we attempted to study the CRISPR/Cas9 edit effect across distinct regions of the amplicons. It is worth noting here that all the analysis has been performed using the cDNA amplicons directionality. This means that orientation depicted in all deletion profiles in this chapter is the reverse of the MREs orientation at the genome. So, e.g. the MRE 3p binding region at the genome (i.e. the genomic region downstream to the MRE) is located upstream to the seed region in all plots. However, for the rest of the analysis the regions upstream and downstream to seed region are based on the cDNA amplicons directionality. Using this convention, reads are segregated into six distinct segments (Figure 6.7)a:

- *Distal Upstream region* (*DU*): up to 50 nt long, upstream to the Proximal Upstream region
- *Proximal Upstream Region* (*PU*): up to 15nt long, upstream to the Seed region (the MRE 3p binding region at the genome)
- *Seed Region* (*SD*): the seed (6 to 15nt long, depending on MRE target length)
- *Proximal Downstream Region* (*PD*): 14 nt long, downstream to Seed region - referring to the rest of the mature miRNA sequence
- *Medium Downstream Region* (*MD*): 15nt long, downstream to PD region

- *Distal Downstream Region* (*DD*): up to 50 nt long, downstream to MD.

Overall deletion ratios have been calculated for each MRE and for each of the defined regions. These ratios represent the number of *mutant* sequences of an MRE divided by all identified sequences of this MRE. Insertions were not taken into consideration for the rest of the analysis due to low representation across all recorded insertion/deletion events (Figure 6.4). We can observe a strong enrichment of deletions at the seed region in the cDNA library (Figure 6.7). Additionally, the regions that are most highly enriched in deletions after the seed region are the *Proximal Downstream* (*PD*), the *Medium Downstream* (*MD*) and *Proximal Upstream* (*PU* or genomic MRE 3p binding region). This is expected since CRISPR/Cas9 induces a cut at a strand it extends this cut towards one direction or the other (Wu et al., 2014). So, there are some MREs where the directionality of the cut is from 5' to 3' and in this case if the cut is made on the seed region it is possible to extend to the *PD* or even to the *MD Region*.

Similarly, when the directionality of the cut is inverse (3' to 5') then the *PU Region* is also likely to contain many deletions. This can also be verified if we calculate the correlations of deletion ratio fold changes across all regions from the gDNA to the cDNA libraries and for each MRE (Figure 6.8). We can see that the *Seed Region* is strongly correlated with both the *Proximal Downstream* and the *Medium Downstream* regions. There is also a less strong correlation of the *Seed Region* with the *Proximal Upstream Region* implying that CRISPR/Cas9 favours more the 5' to 3' cut directionality (cDNA coordinates) in this set of targets. If we convert these regions into genomic coordinates then we observe that regions upstream to the seed region are the ones that have the higher deletion ratios after the *Seed Region* while the MRE 3p biding region is less affected by CRISPR/Cas9.

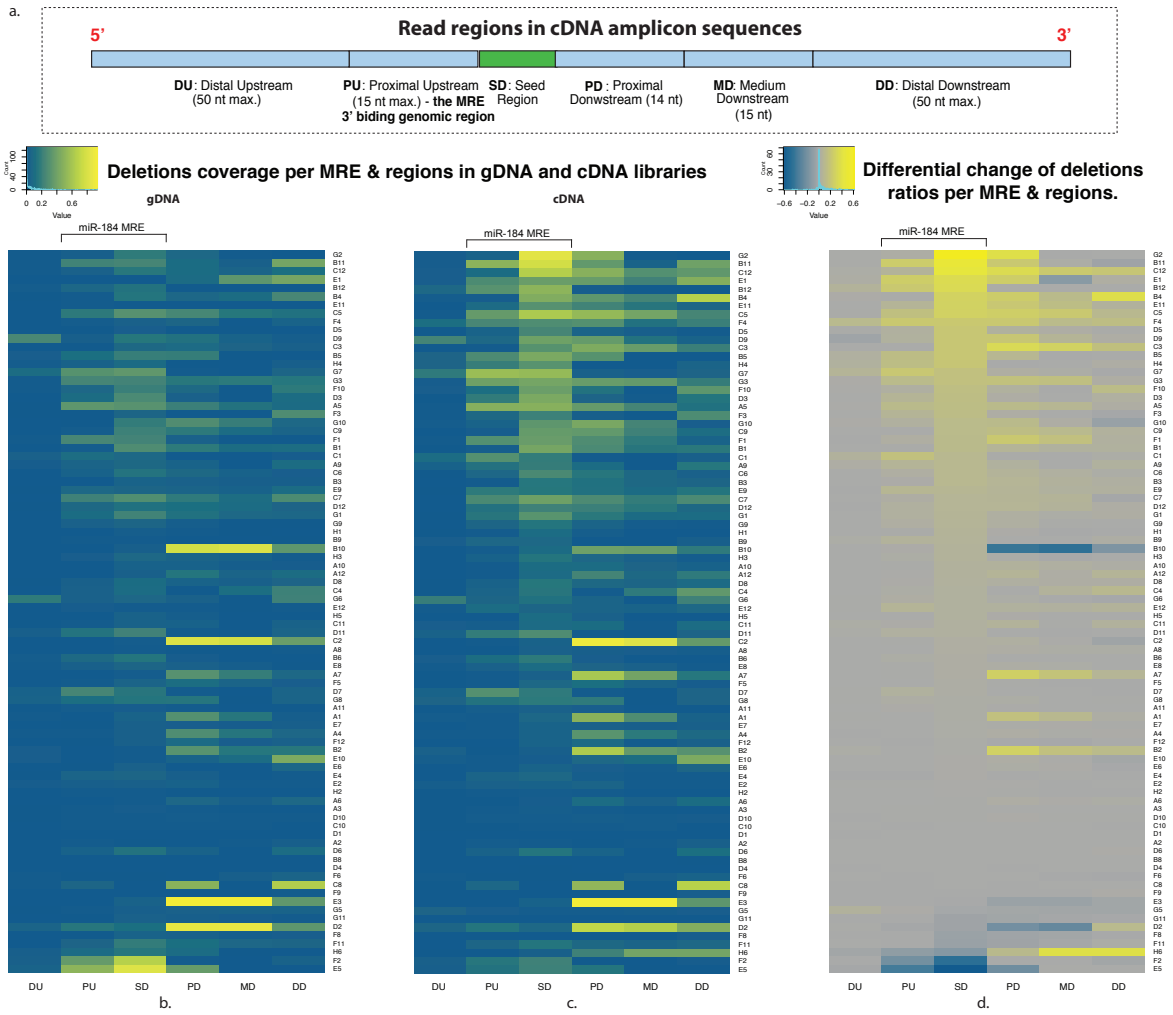



**Fig. 6.7** a) Definition of regions across the amplicons with reference to the seed region. Average deletion ratios in read regions of the gDNA (b) and cDNA (c) libraries. Differential change (cDNA-gDNA) of deletion ratios across all MREs (d). The directionality of each MRE amplicon corresponds to cDNA sequences, so it is the reverse of the genomic directionality.

Furthermore, deletion ratios shift towards the cDNA library for all regions but to a different extent (Figure 6.9). The *Seed Region*, *Proximal Upstream* and *Proximal Downstream* regions show the most evident shift. Similarly, if we segregate the reads only into seed, upstream and downstream regions, the seed has the most predominant shift among all regions followed by the *Downstream Region*.

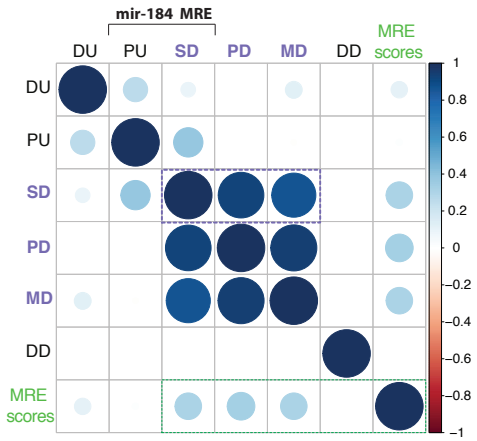



**Fig. 6.8** Correlations of deletion ratio fold changes between all regions and MRE scores (see Func. 1).

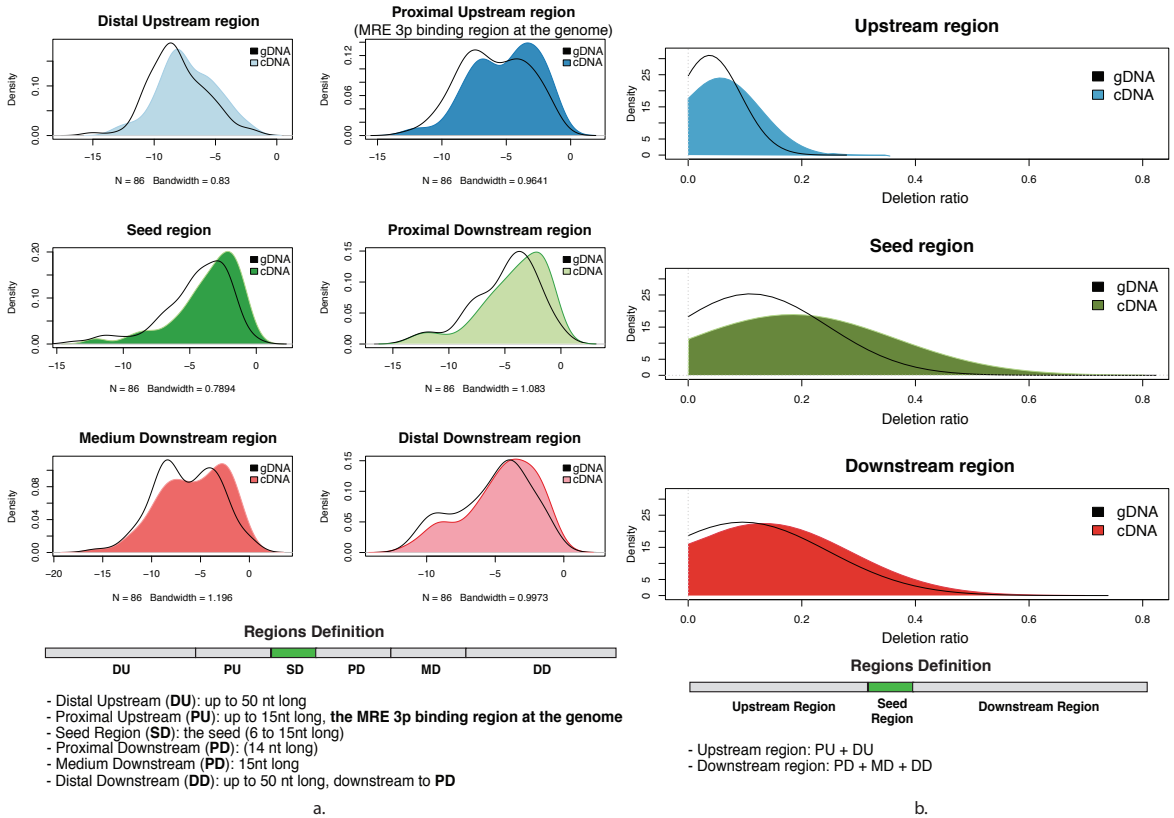



**Fig. 6.9** Deletion ratio densities from all MREs across (a) 5 (b) and 3 distinct regions. The Upstream and Downstream annotations of the regions refer to the cDNA amplicons directionality.

We have observed, so far, quite high variability in the deletion ratio changes between the gDNA and cDNA library. We may also confirm this variance by looking at individual coverage profiles of targeted MREs (Figure 6.10). These profiles depict the normalised coverage obtained from all reads of each MRE. In order to study the enrichment of deletions

towards the cDNA library we introduced an enrichment score for each MRE, which we refer to as *MRE score*, based on the depths of WT and MT reads in the gDNA and cDNA libraries:

$$MRE_{score} = \frac{\frac{cDNA_{(MT)}}{gDNA_{(MT)}}}{\frac{cDNA_{(WT)}}{gDNA_{(WT)}}} \qquad (Func.\ 1)$$

MT (mutant) reads are considered that have at least one insertion or deletion in the seed region. Based on the value of the MRE score an MRE is considered as active or inactive. More specifically:

- if $MRE_{score} > 1 \Longrightarrow$ the MRE is active (actively transcribed)
- if $MRE_{score} \leq 1 \Longrightarrow$ the MRE is inactive (not extensively transcribed).

Using these *MRE scores* we can clearly distinguish the MREs where CRISPR induced insertions/deletions have been integrated effectively into the genome and are being transcribed at a higher level (i.e. higher *MRE scores*) or expressed at lower levels (i.e. lower *MRE scores*). *MRE score* calculations have shown that the majority of MREs are active (*MRE score* > 1) (Figure 6.11a). However, the *MRE scores* distribution seems very similar both for the coding and non-coding MREs (Figure 6.11b).

Finally, in order to confirm the validity of our results, we assessed the quality of our replicates. MRE scores calculated from any pair of replicates have a Pearson correlation of at least 0.99 and a P-value < 0.001 (Figure 6.12), thus confirming the consistency of all library replicates.

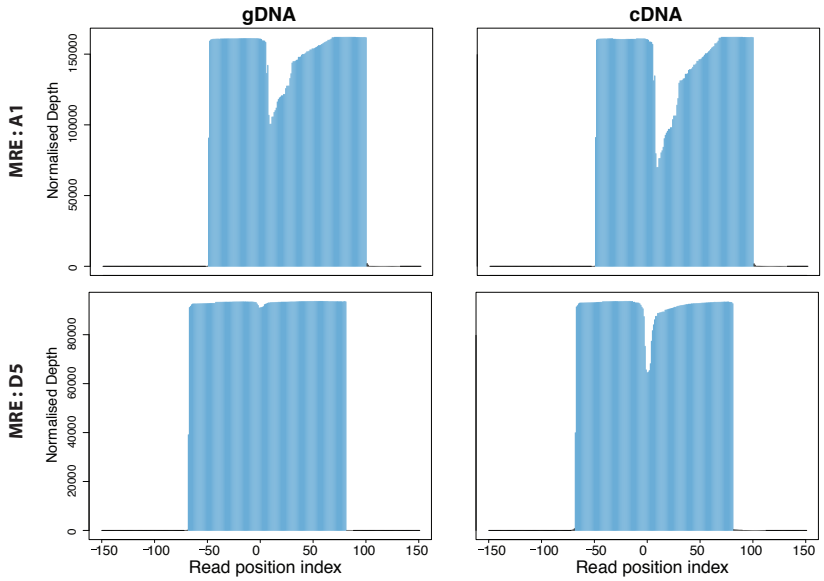



**Fig. 6.10** Example coverage profiles of individual MREs support the assumption of high variability in deletion ratio change between the gDNA and cDNA library.

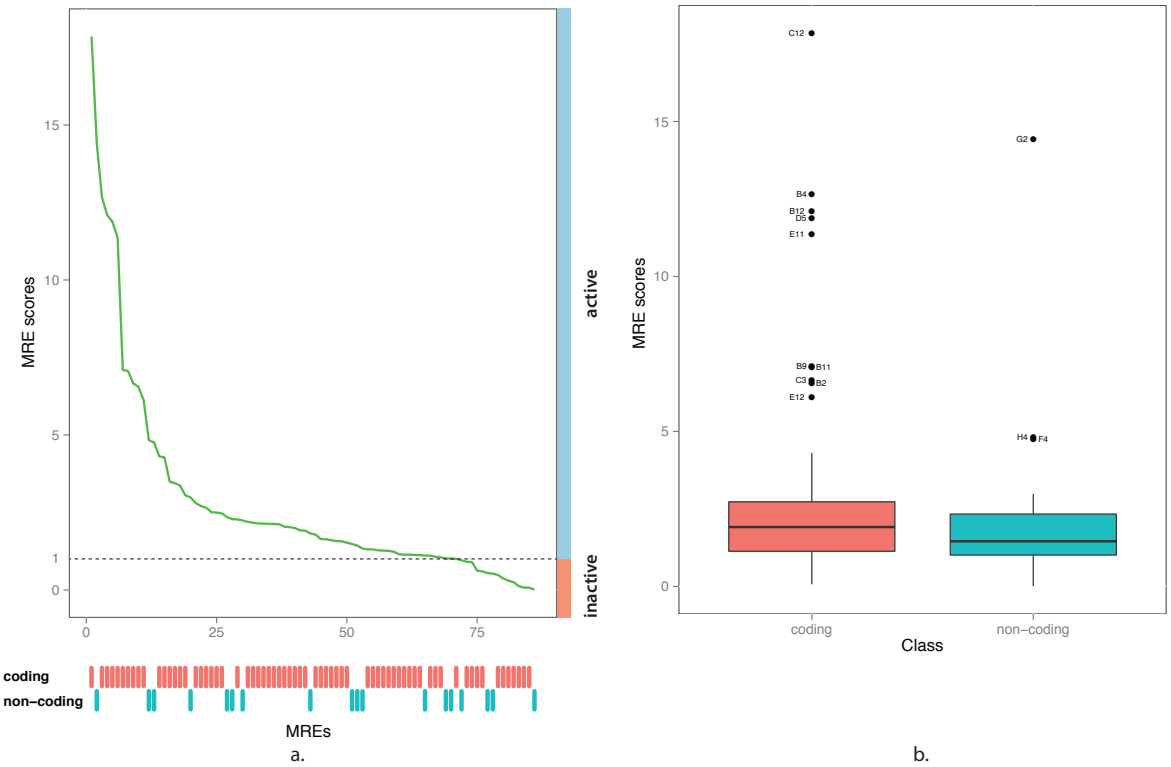



**Fig. 6.11** a) Ranking of all MRE scores associated with their MRE class, b) MRE scores distribution across MRE classes (coding and non-coding).

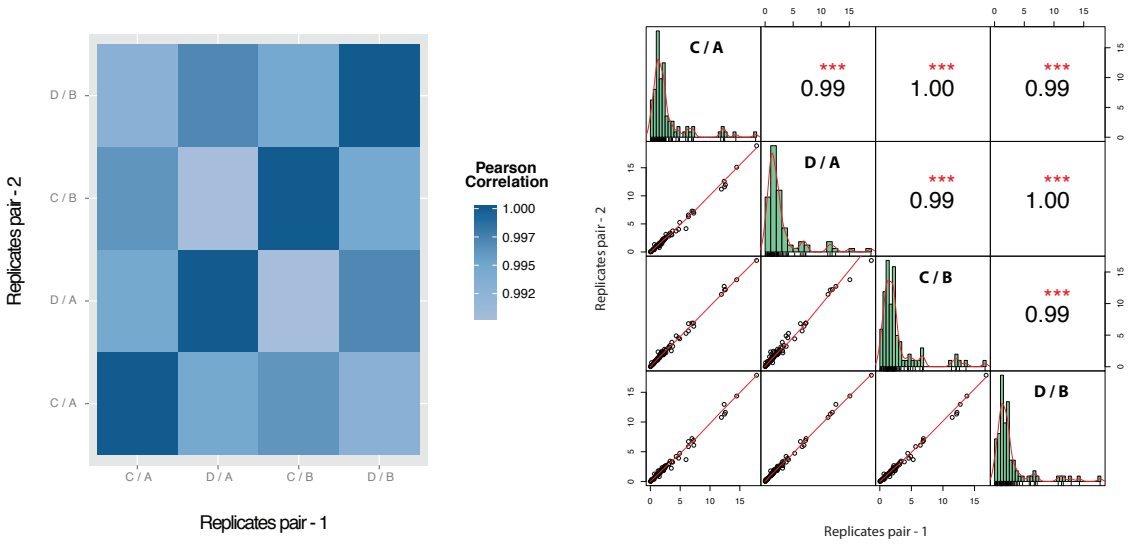



**Fig. 6.12** Pearson correlation between MRE scores, calculated for each pair of library replicates (left) along with the P-values of each correlation score (right). ***: P-value < 0.001.

### 6.3.3 Single-nucleotide resolution deletion profiles

As a further step for studying the deletions induced by CRISPR/Cas9 we calculated the deletion ratios at each nucleotide position of every MRE amplicon (Figure 6.13), thus attaining the highest possible resolution. These profiles allow the inference of the directionality of the cut by visual inspection. We can see that for most MREs there is a deletion 'band' that starts from either the 5' or 3' end of the MRE target and extends to the 3' or 5' end, respectively. Deletion ratios are higher (more intense colour) at the start of the deletion band and attenuate towards the end of these deletion 'bands'.

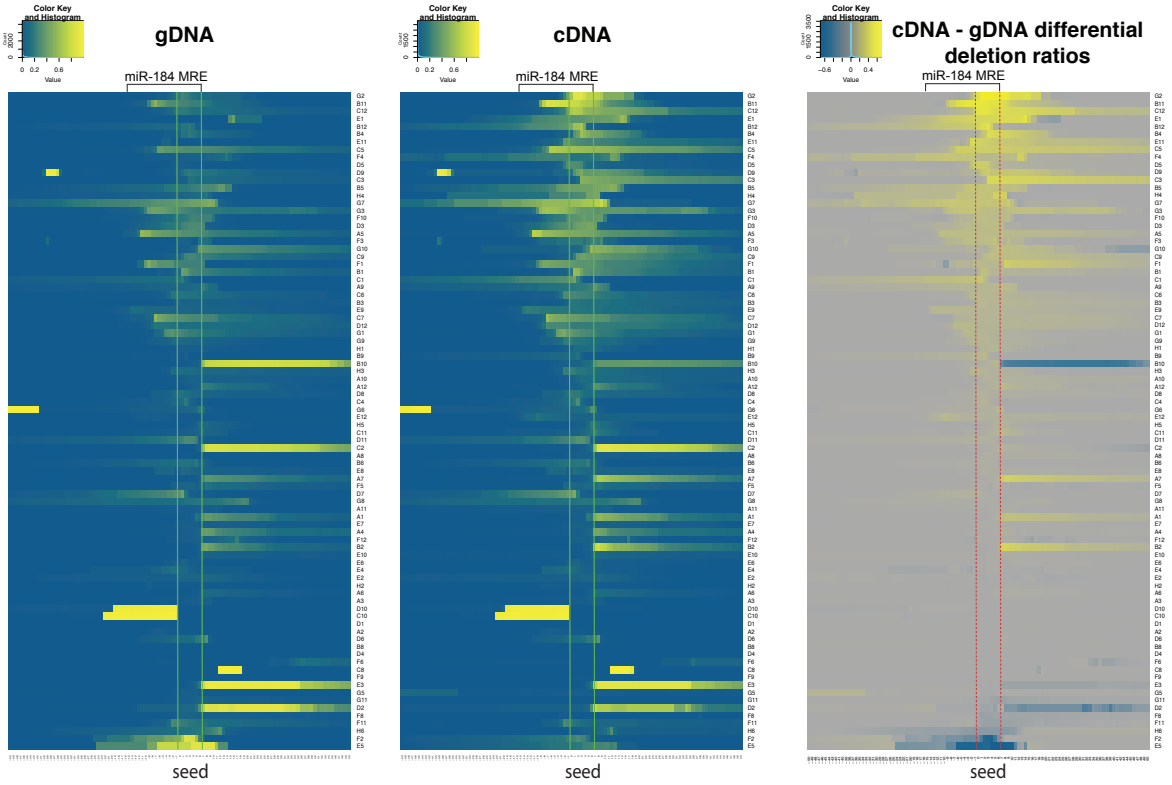



**Fig. 6.13** Deletion ratios at each nucleotide position of the gDNA (a) and cDNA (b) libraries, (c): Differential change (cDNA-gDNA) of deletion ratios at each nucleotides position across all MREs.

Subsequently, we calculated the average sum of deletion ratios at each position and we noticed two peaks downstream to the start of the seed region (Figure 6.14). The first peak is very close to the start of the seed region and the second peak is positioned at index 6, which is the end of the large majority of MRE seeds in this study, having 7nt long seeds. Thus, deletion ratios are higher at or close to the ends of the MRE seeds. Moreover, we can see that deletion profiles differ slightly between the coding and non-coding MRE classes. Non-coding MREs show a higher restriction of deletions inside or close the seed region while coding MREs demonstrate a decreasing deletion trace that extends upstream

(less) and downstream (more) to the seed region. Besides, non-coding MREs show higher deletion enrichment in genomic DNA than coding MREs (greater average negative sum). Finally, we can observe that deletion ratios are higher upstream to the genomic MRE and lower downstream to it.

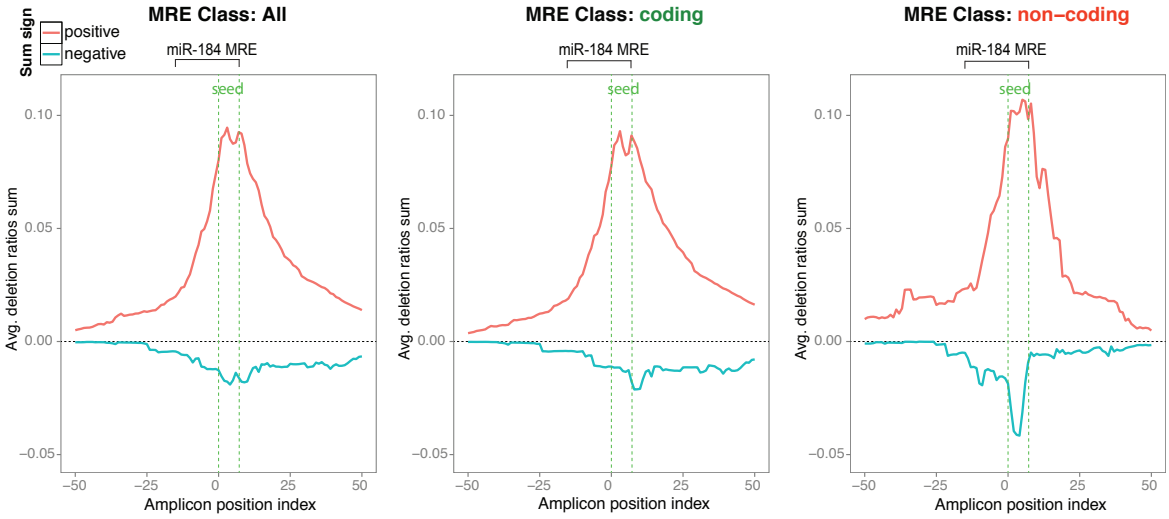



**Fig. 6.14** Average sum of deletion ratios between the cDNA and gDNA MRE read depths, at each nucleotide position of a 100nt window, centered at the start of the Seed Region (index: 0).

Single-nucleotide resolution deletion profiles allow us to study in depth all mutation types that occur across the examined amplicons. As a first step, we used this data in order to define two additional custom types of *MRE scores*. The first type of *MRE score* (Type I) is defined as the log2 fold-change between gDNA in cDNA libraries of the reads that have a single deletion at each position of the amplicon (Figure 6.15). We notice that Type I MRE scores are enriched at the seed region and diminish gradually as we move away from the seed. We also need to mention that there is a slight drop in MRE scores at index 6 of the Seed Region. This is just due to the fact that some of the examined MRE targets are 6 nt long and their sequence corresponds to the 0 to 5 indexes of the seed.

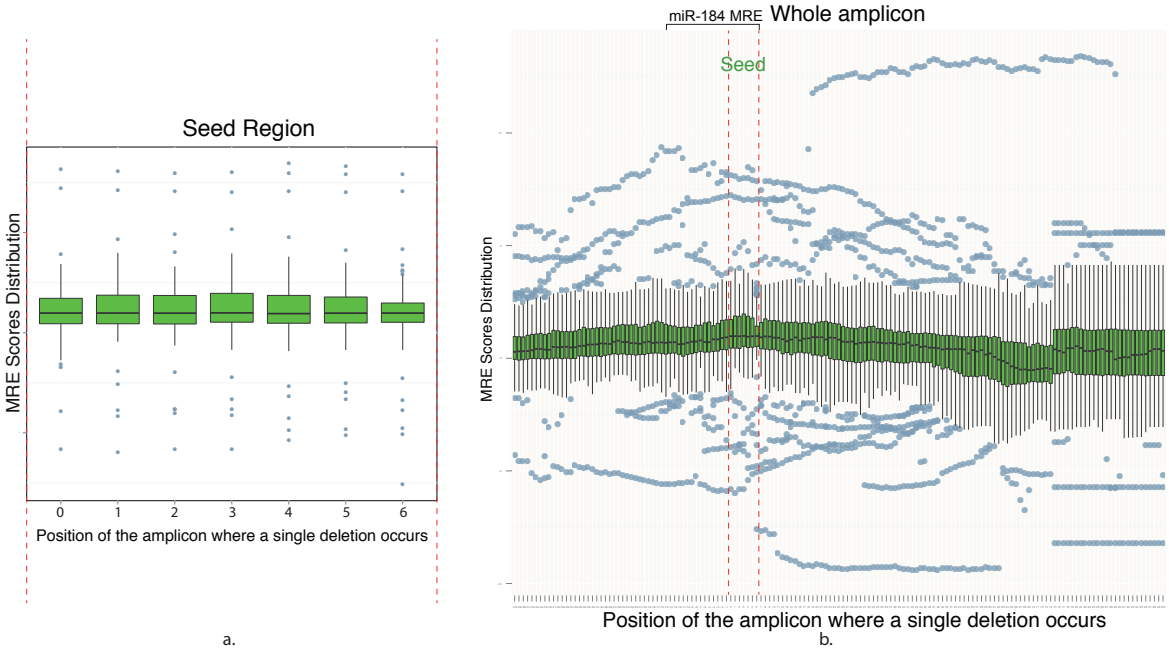



**Fig. 6.15** Custom (Type I) MRE scores distribution for single deletions at each position of the *Upstream* (a), *Seed* (b), *Downstream* (c) regions and the whole amplicon (d).

The second additional type of MRE score (Type II) attempts to capture the length variance of the mutation types induced by CRISPR/Cas9. These *MRE scores* are calculated as the log2 fold-change of reads that have a certain (exact) number of deleted base-pairs anywhere at a region of interest (Figure 6.16). We can observe at these profiles that mutation types with 6 or 7 deletions are the most predominant mutation type for the Seed region. However, the other mutation types follow with comparable scores. This means that the full length of each target hasn't been deleted by CRISPR in all cases. Furthermore, we notice a low variance of *MRE-Type II scores* in the *Upstream region*, which implies that mutation types of variant lengths (from 1 to 50 nt) change with similar frequency between the gDNA and cDNA libraries. Finally, the downstream region shows an enrichment of *MRE Type II scores* for mutation types of length > 26 nt. This means that the genomic region that is upstream to the MRE targets is also extensively deleted by CRISPR/Cas9, along with the seed regions.

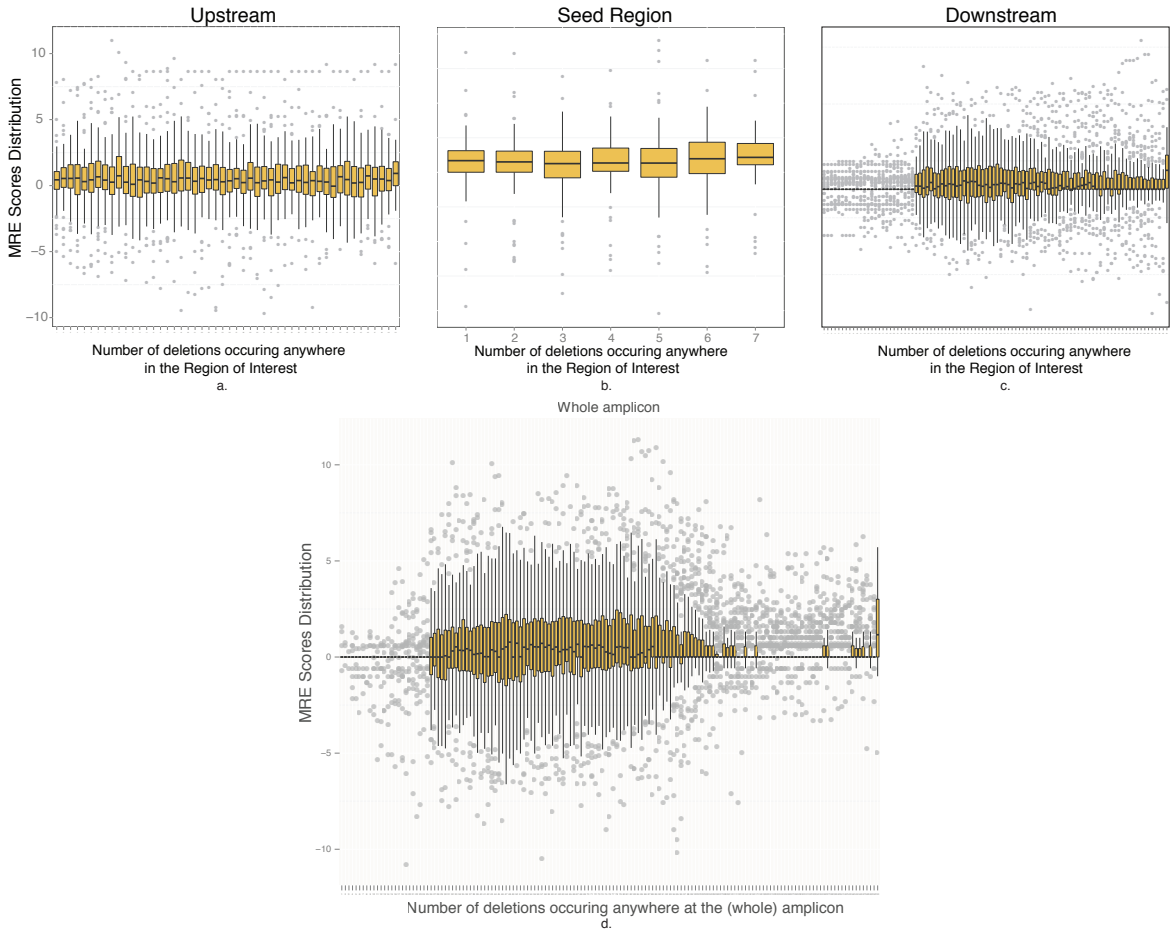



**Fig. 6.16** Custom (Type II) MRE scores distribution for different (exact) numbers of deletions occurring anywhere within the *Upstream* (a), *Seed* (b), *Downstream* (c) regions and the whole amplicon (d).

### 6.3.4 Accessibility Analysis of MRE targets

From the previous analysis so far, we have observed a high variability of the CRISPR effect across all MREs. This variability may be explained by various factors, including accessibility of the target sequence. In this section we will try to correlate CRISPR efficiency, as it is captured by MRE enrichment scores, with target accessibility predicted computationally using the *Vienna package* (Lorenz et al., 2011). In order to do so, we first folded each cDNA amplicon (200 nt long, centered at the seed region) using the *RNAplfold* tool from the *Vienna package*. We have used three window sizes for assembling the secondary structure of each amplicon sequence: 50, 100 and 150 nt. Accessibility for each nucleotide is assessed as the likelihood that this nucleotide and its neighbouring 6 nts are unpaired, based on the structure that has been predicted using each window size. As a control for the accessibility assessment, for each MRE amplicon we have used 10,000 shuffles of the original amplicon

and calculated the average and standard deviation of accessibility at each nucleotide (Figure 6.17a). The final accessibility score is the Z score accessibility (Figure 6.17b), which is the distance of the accessibility of the original sequence from the average accessibility of the shuffled sequences at each nucleotide, in units of standard deviation of the shuffled sequences accessibilities.

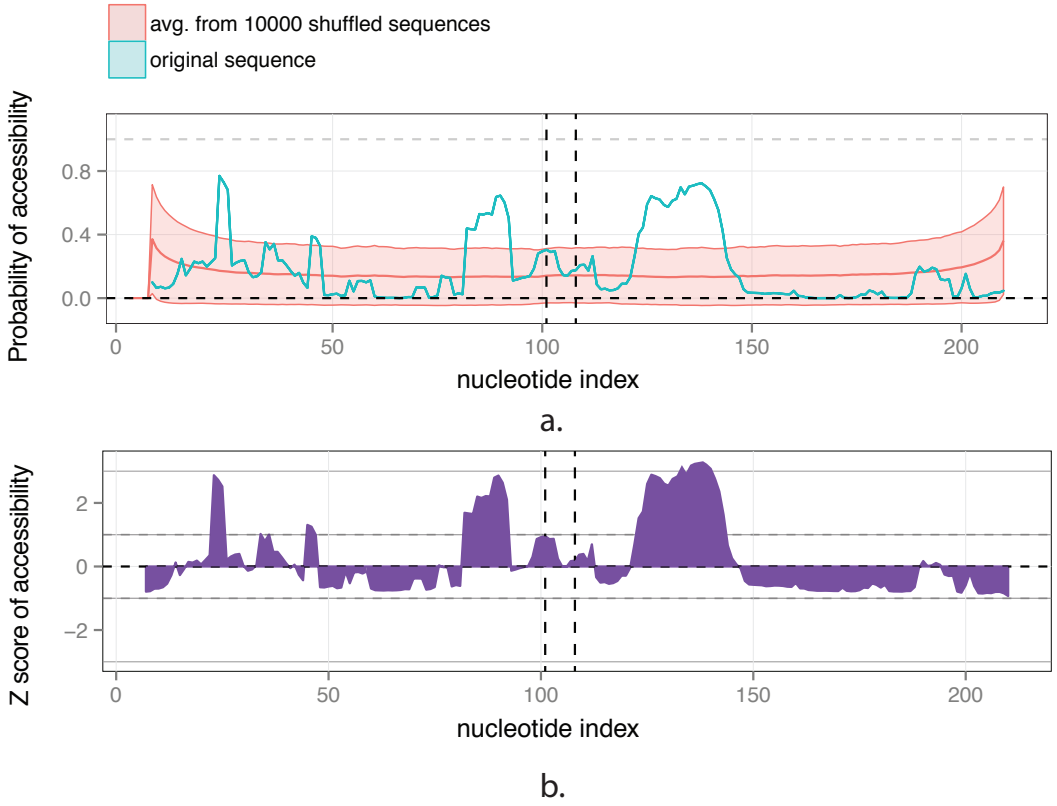



**Fig. 6.17** a) Example results from probability assessment of the accessibility at each nucleotide position of the original MRE sequence and the avg. of 10000 shuffled sequences. b) Z score of accessibility probability at each nucleotide position in reference to the mean and standard deviation of accessibilities of the shuffled sequences.

We have tested our method to known examples, such as the let-7 precursor, and the retrieved accessibility scores, calculated within a 200nt long sequence around the precursor, capture with very good precision the structure of the precursor and specifically the stems and the loop parts (Figure 6.18).

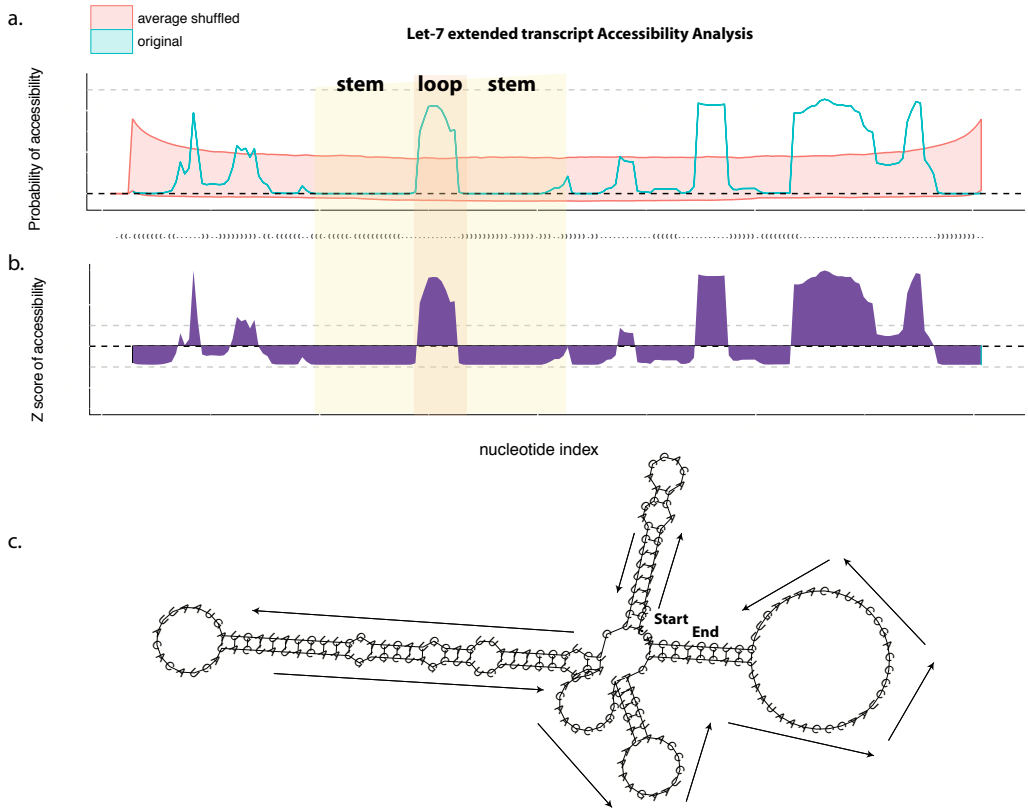



**Fig. 6.18** a) Probability of accessibility of each nucleotide calculated for a 200nt sequence around the let-7 precursor and 10,000 shuffled control sequences. b) Z-score accessibility: distance of the accessibility of the original sequence from the average accessibility of the shuffled sequences at each nt, in units of standard deviation of the shuffled sequences accessibilities. c) Computational prediction of structure and accessibility of a 200nt long sequence window around the let-7 precursor.

### 6.3.5 Association of accessibility profiles with classes of enriched MREs

We have applied our accessibility assessment method to all MRE amplicons with canonical targets (59 overall), in order to assure a fixed centre seed region for all amplicons, and calculated the Z score accessibilities at each nucleotide position. We then clustered (hierarchical clustering) the accessibility profiles for different sub-regions of each amplicon and correlated the overall accessibility profile with the MRE enrichment scores. MRE scores have been classified into 4 classes based on their values:

- High, if: $MRE_{score} > 4$
- Medium, if: $MRE_{score} > 2$
- Low, if: $MRE_{score} > 1$

- Reversed, if: $MRE_{score} < 1$

and the regions that correlation analysis has been performed involved the following sub-regions of each amplicon:

- *Whole Amplicon length*
- *Seed region*
- *Extended Seed* (Seed region ± 6 nt)
- *Downstream*
- *Upstream*
- *Downstream section* (25nt downstream to the seed)
- *Upstream section* (25 nt upstream to the seed)
- *Seed & Downstream* (seed region & 25nt downstream to the seed)
- *Upstream & Seed* (25 nt upstream to the seed & seed region)

Apart from the hierarchical clustering we also performed a Principal Component Analysis for each case in order to see if MRE score classes can be separated adequately based on their respective accessibilities Figure 6.19. By inspecting each of the retrieved heatmaps and PCA plots for all window sizes (50, 100 and 150 nt), we have observed that the only region that demonstrates a pattern of correlation between accessibility and MRE scores is the *Upstream section* (25 nt upstream to the seed). Of course, separation is not accurate for all members of each class. However, we can get a fairly clear separation between the Reversed MRE score class and the non-Reversed ones, which imply a segregation between CRISPR deletions enrichment either at gDNA or cDNA.

This region (*Upstream section*) corresponds to the MRE 3p binding region at the genome. Based on the previous analyses (Figure 6.8, 6.9, 6.14), we observed that deletion ratios are more predominant downstream to the cDNA amplicon seed region (*Downstream section*). So, we could potentially assume that the region that is critical for CRISPR's cutting efficiency is the MRE 3p biding region. High accessibility in this area leads to MRE scores enrichment in most cases and it is also very likely that deletions starting at the seed region also expand upstream to the seed region (in genomic coordinates), thus explaining the high deletion ratios in that area as well.

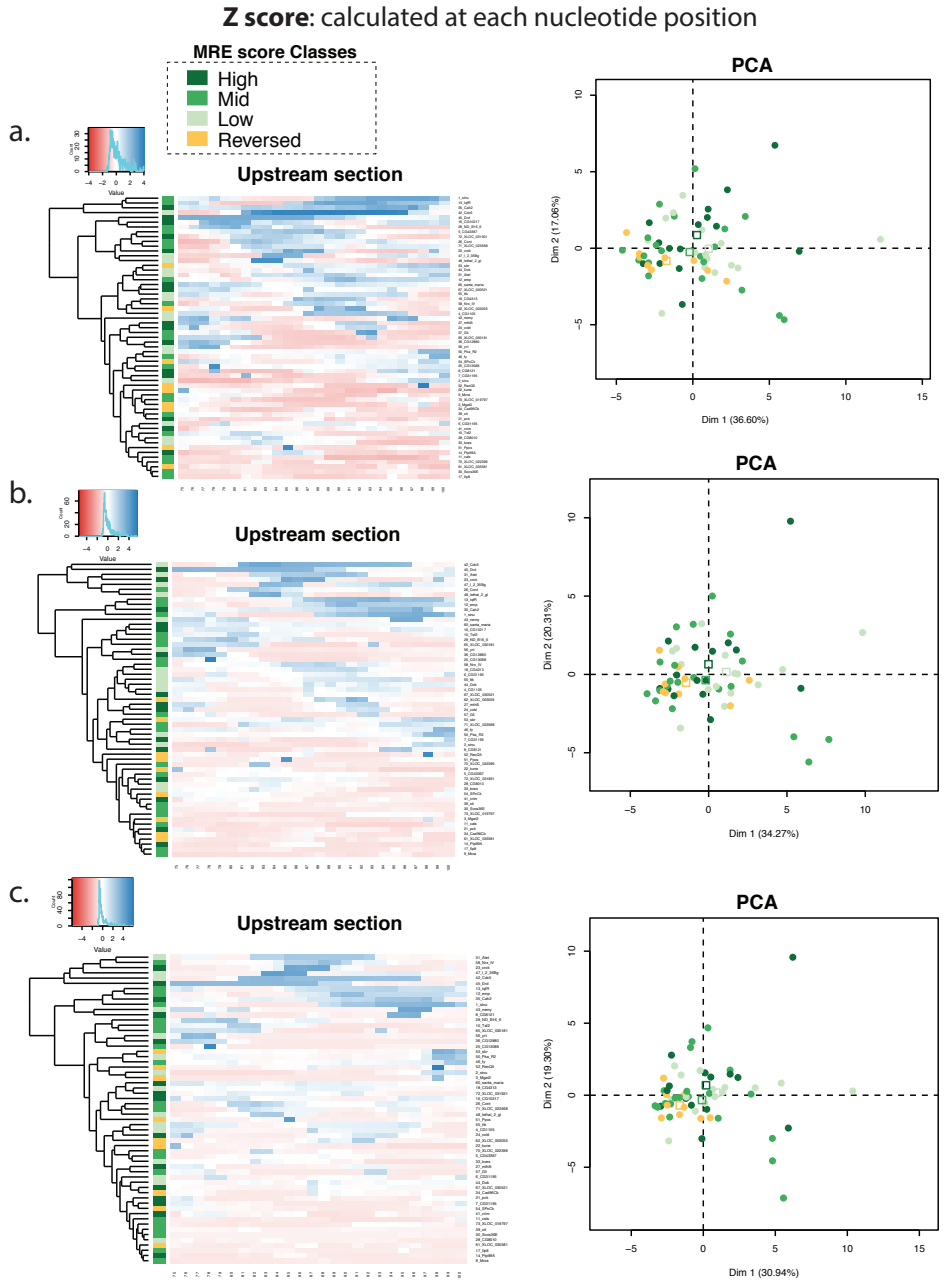



**Fig. 6.19** Hierarchical clustering (left) and Principal Component Analysis (right) of the canonical MREs based on their accessibility profiles, in association with their MRE score class for different window sizes used for the computational prediction of accessibility: a) 50 nt, b) 100 nt, c) 150 nt. Accessibility profiles contain the Z score at each nucleotide position.

Based on the original accessibility analysis, we computed a smoothed version of accessibility profiles. In order to do that, we have collapsed every 5 columns from the original profiles into a single one and assigned as a Z score for the new column either the median or the sum of the collapsed Z scores. We noticed again (Figure 6.20, 6.21) that the highest

correlation of accessibility with MRE scores enrichment is achieved at the Upstream section region, with the Seed included in some cases. In conclusion, we suggest that the Seed Region cutting efficiency may be regulated by the MRE 3p binding region's -and potential Seed's too- accessibility.

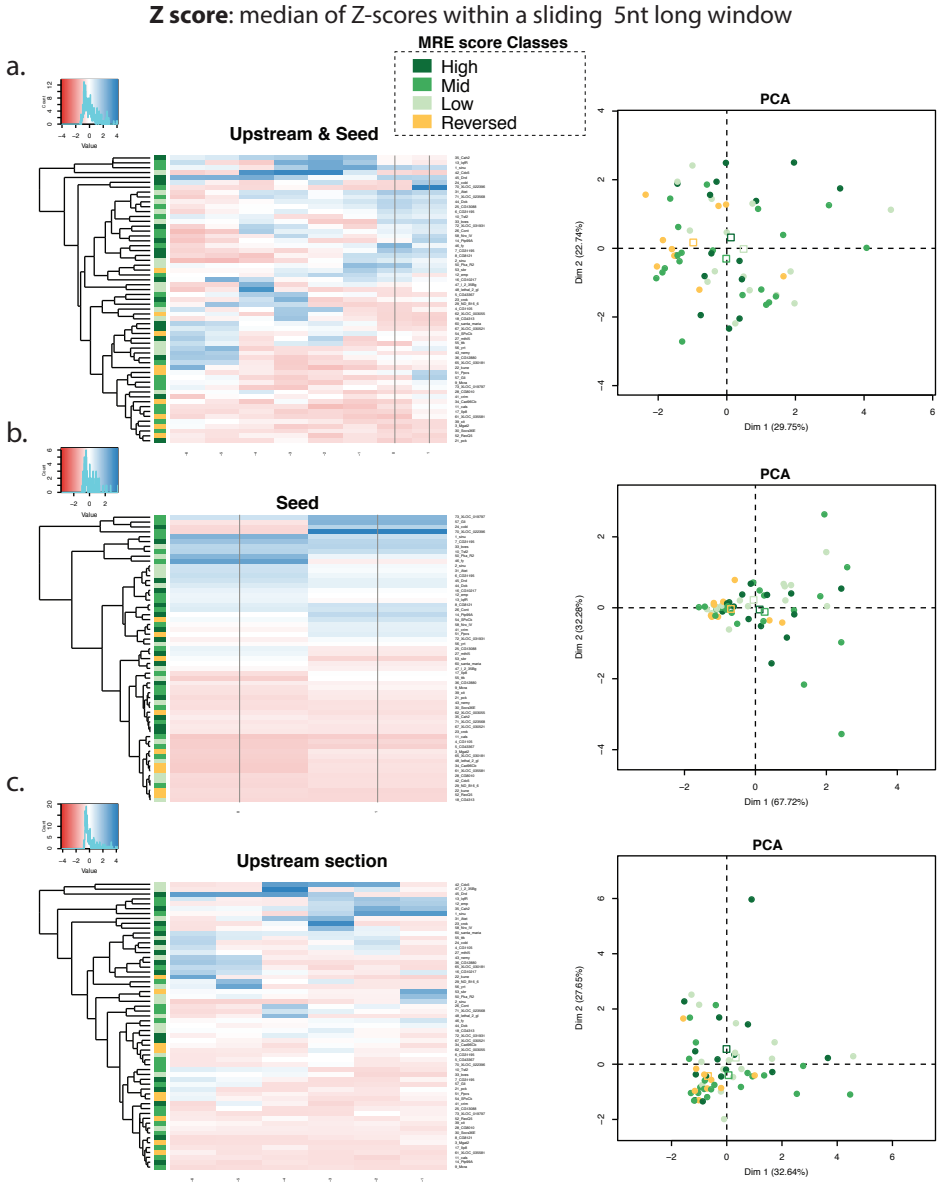



**Fig. 6.20** Hierarchical clustering (left) and Principal Component Analysis (right) of the canonical MREs based on their accessibility profiles, in association with their MRE score class for different window sizes used for the computational prediction of accessibility: a) 50 nt, b) 100 nt, c) 150 nt. Accessibility profiles contain the median of Z scores from every 5 columns of the original profiles.

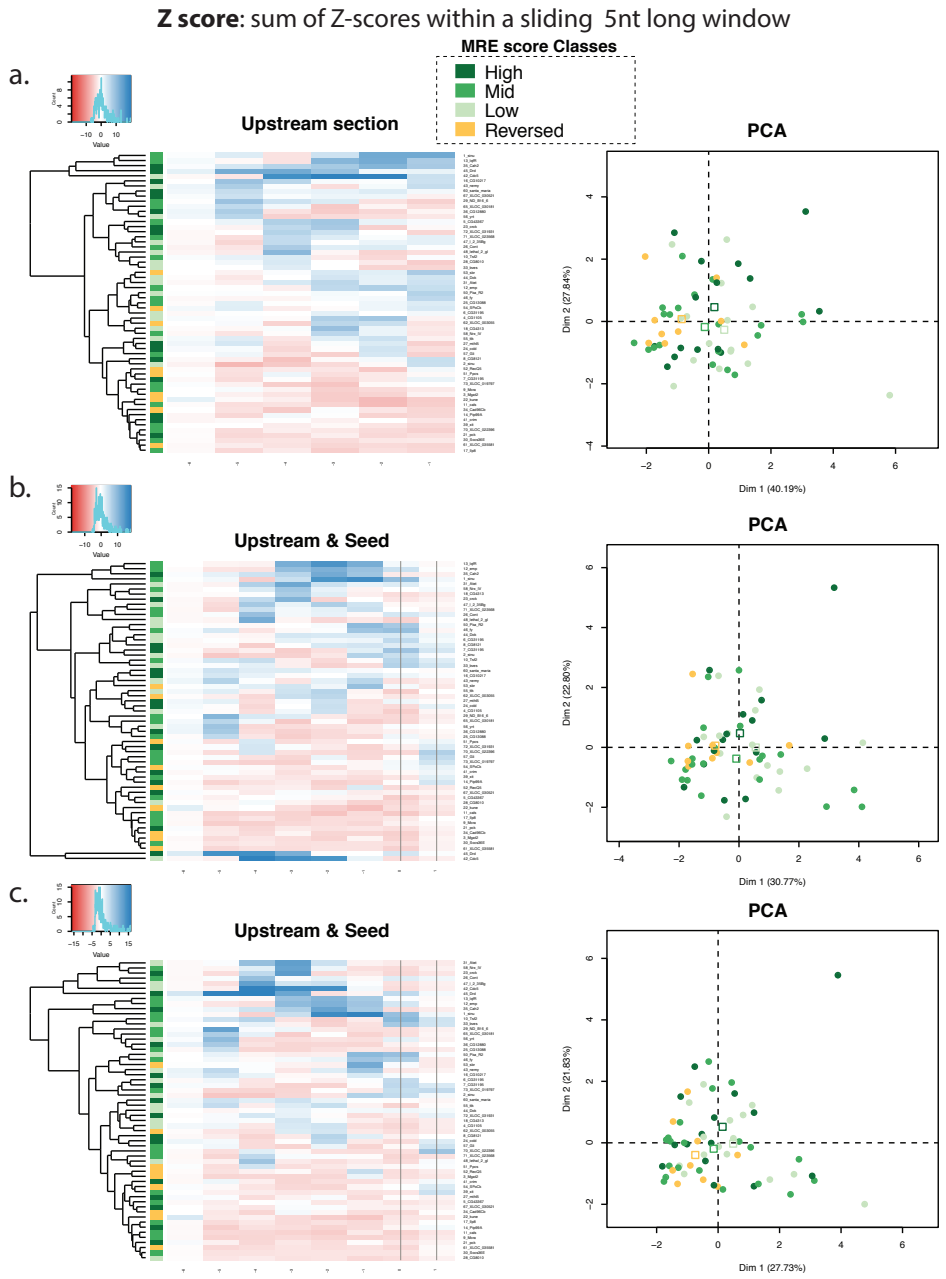



**Fig. 6.21** Hierarchical clustering (left) and Principal Component Analysis (right) of the canonical MREs based on their accessibility profiles, in association with their MRE score class for different window sizes used for the computational prediction of accessibility: a) 50 nt, b) 100 nt, c) 150 nt. Accessibility profiles contain the sum of Z scores from every 5 columns of the original profiles.

## 6.4 Conclusion

Based on the results presented in this chapter we concluded that CRISPR/Cas9 introduces deletions predominantly in the seed region of the target, along with regions immediately

upstream or downstream to it. Thus, its function may not always target exclusively the desired regions but may affect other parts of the genome as well. Moreover, we observed a high variability in CRISPR/Cas9 efficiency even though the sequence of the targeted regions are the same. This implies that there may be other factors that affect CRISPR/Cas9 functionality with regards to editing a region. Thus, we assessed the correlation of computationally predicted accessibility with target (MRE) enrichment scores and we noticed that the MRE 3' binding region may be playing the most prevalent role in the accessibility of the target by the CRISPR/Cas9 mechanism. As a future work, we could suggest integrating more accessibility data in our study, derived from ATAC-Seq, Chip-Seq and SHAPE-Seq (for single-nt resolution accessibility assessment) experiments using the same cell lines of *D. melanogaster*. Additionally, we could take into account the variability of the designed sgRNAs for each target.

# Chapter 7

# Discussion

## 7.1 Conclusions

The advent of Next-Generation Sequencing and the rapid development of bioinformatic analysis tools in recent years has enabled the acceleration of progress in biological research. A significant part of recent discoveries has been devoted to small non-coding RNAs, such as miRNAs and piRNAs. The discovery that these classes of small RNAs take part into fundamental processes of animal and/or plant cells, such as differentiation, embryogenesis and gene regulation, has ignited a notable scientific endeavour in order to decipher the secrets of the world of small RNAs. In this thesis, we shed more light into novel pathways and features of miRNA and piRNA biogenesis as well as introduced two novel methods for the analysis and prediction of miRNAs.

First, we presented Chimira, a novel method for miRNA quantification and identification of 5'-,3'-terminal and internal modifications (including ADAR-edits and SNPs). This work was inspired by previous studies that have shown the important role of modifications, such as uridylation and adenylation, in miRNA biogenesis and stabilisation (Heo et al., 2012; Katoh et al., 2009). The method that we developed is provided publicly as a web-application with a user-friendly interface and its efficiency and speed were demonstrated.

Next, we applied Chimira into a large study investigating the impact of 3' terminal uridylation on the Mouse transcriptome. The outcome of this collaborative work was that the transcriptome in oocytes is regulated by extensive uridylation of the maternally deposited transcripts. On the other hand, changes in uridylation levels did not have a significant impact in either mRNA or miRNA levels in adult somatic cells or embryonic stem cells.

In chapter three, we presented a large-scale analysis of deposited small RNA datasets in order to elucidate hidden miRNA biogenesis features as well as explore the landscape of miRNA post-transcriptional modifications. We saw that datasets from similar cell types or tissues tend to cluster together based on miRNA expression and also based on their modification profile in several cases. With regards to miRNA expression, it was shown for the first time in such a large scale that miRNAs located in proximal regions within the genome are expressed simultaneously. In cases of observed co-expression with no genomic proximity, it was also shown that miRNAs belonging to distant genomic locations are in fact regulated by the same sets of transcription factors.

Furthermore, we found a high variability in the modification profiles across different datasets. 3' terminal modifications are the most predominant modification type with a high prevalence of patterns with 1 to 4 nucleotides. We also discovered, though at a reduced level, the presence of various 5' modification patterns that may be affecting the miRNA targets repertoire by changing the seed region of the modified miRNAs. The large amount of miRNA expression data we retrieved from the analysed datasets allowed us to explore the rules that regulate strand selection during miRNA maturation. We confirmed that the 5' ends of both the mature and the star miRNA products are responsible for defining the strand to be selected. Moreover, we discovered that it is actually only the first two nucleotides at the 5' end of each strand that play the most predominant role in strand selection. Finally, we extracted coverage profiles for numerous miRNAs and suggested several mis-annotated miRNAs from miRBase that appear to have non-canonical coverage profiles and thus may not be real miRNAs.

This last part of the third chapter was the leading idea for the next project, analysed in Chapter 4. By inspecting multiple miRNA coverage profiles, we observed that a typical miRNA has a well-processed 5' end, some 3' tailing and very few modifications/SNPs inside the main body of the mature sequence and especially not in the seed region. This motivated us to explore the possibility of identifying miRNAs based only on features derived from the coverage profile, thus not requiring a reference genome. To this end, we developed mirnovo, a novel machine learning based method that is able to predict known and novel miRNAs with or without a reference genome with very high accuracy. Its performance exceeded the performance of miRDeep2, which is currently the state-of-the-art tool. Additionally, we applied mirnovo to various large scale analyses and retrieved interesting insights about Drosha and/or Dicer independent miRNA biogenesis as well as discovered novel miRNAs from single-cell small RNA-Seq data.

In the last two chapters of this thesis, we presented the work that was done as part of two collaborative projects. Specifically, in Chapter 5, we explored alternative biogenesis

pathways for piRNAs in mice. We confirmed first that MIWI2 plays a more important role in piRNA biogenesis than the MILI protein. Additionally, we discovered that in the absence of MILI, piRNAs are still expressed, though at lower levels. This MILI-independent biogenesis pathway for piRNAs in mice might be explained with a Drosophila-like phasing mechanism. We did not find a very strong confirmation for the existence of phasing from our data, however we extracted a distance enrichment of 38nt between piRNA clusters that may explain piRNA expression despite the loss of MILI.

In the sixth and final chapter we presented an exploratory analysis of multiple target sites of a single miRNA in *D. melanogaster*, which had been edited by CRISPR/Cas9. We observed a high variability of CRISPR/Cas9 editing effect across the targets examined, despite their common sequence. Thus, we attempted to associate editing efficiency with computationally predicted accessibility of the targets in the genome. Eventually, we found a correlation, though not very strong, between the computationally predicted accessibility of the targets and efficient integration of edited sites in the genome.

## 7.2 Future research

The work conducted as part of this thesis yielded several insights into small RNA biogenesis and function. However, it also provides the ground to extend and build upon in order to perform further and/or improved analyses.

One of the first future analyses that we are suggesting would entail the analysis of miRNA modifications in very specific dataset conditions, e.g. cancer data from a certain cell type/tissue. In chapter 3 we explored the extent of modifications across several datasets derived from various conditions. This allowed us to get a macroscopic overview of modification events and establish some general rules that apply to the high majority of cases, such as prevalence of adenylation, uridylation and patterns of up to 4 nucleotides long. However, we believe that focusing on an 'isolated' dataset of control (healthy) and experimental (cancer or other disease) samples would allow us to assess the extent to which miRNA modifications regulate or contribute to perturbation of normal cell function, as has been shown in previous studies for specific cases (Boele et al., 2014; Li et al., 2012). To this end, we also believe that it is worth providing a stand-alone version of Chimira as well, in order to facilitate labs willing to perform very large-scale analyses across various conditions and/or organisms.

Furthermore, it would be very interesting to associate miRNA post-transcriptional modifications with other factors. A very interesting approach would be to examine if and to what extent DNA methylation may be affecting the abundance and motif repertoire of

modifications. Additionally, a very crucial analysis would be to analyse miRNA modifications in relation to mRNA expression. The regulation network that should be studied in that case may be extremely complex. However, we could focus on the most prevalent modification patterns (mono-adenylation and mono-uridylation) and build upon there trying to detect any correlation with changes in mRNA expression or vice-versa.

Our novel miRNA prediction method, mirnovo, exhibited notably high accuracy levels either with or without using a genome. One improvement that we can suggest is to introduce an extra de-noising/filtering step either prior to or after prediction when analysing for single-cell data. This will probably reduce the overhead of validating too many novel miRNA candidates from this type of data, which are already characterised with significantly high noise compared to bulk small RNA-Seq data.

Moreover, it is thought that miRNAs can be grouped into six groups based on their distinct biogenesis characteristics (Kim et al., 2016). Some miRNAs require both Drosha and Dicer during maturation and others need to be mono-uridylated prior to Dicer processing. In addition, other miRNAs may be dependent only on Drosha or Dicer while others may originate from spliced-out introns or from other structured non-coding RNAs. Processing of each miRNA from a different biogenesis pathway leads to a certain degree of variability when it comes to the coverage profiles retrieved from sequencing. For instance, Drosha independent miRNAs usually exhibit some 5' tailing which is not found in canonical miRNAs, whose 5' end is well-processed instead. In this regard, we could extend our method so that it integrates different training models for different types of miRNAs based on their biogenesis in order to eliminate incorrect classification of predicted miRNAs as much as possible.

The current Random Forest classifier employed by mirnovo has proven to be highly efficient, achieving levels of accuracy over 95% and sensitivity/precision of at least 80%. One improvement for the model would be to seek for a minimum set of features and an optimal tree depth for the algorithm to converge, assuming that the new model attains at least an equivalent performance with the current model. This would allow for faster training of the Random Forests, faster prediction calling in unseen data and reduced file size of the trained classifiers per species. Additionally, in Chapter 3 we demonstrated the importance of the first 2nt in mature miRNA selection from the miRNA precursor duplex. Thus, it would also be very interesting to import into the set of features used by mirnovo the difference of free energies ($\Delta\Delta G$) of the 2nt-long duplexes at the 5' ends of each strand product of the duplex. Introducing this feature could potentially enhance mirnovo's predictive performance even further.

With regards to improving the classification step of mirnovo, we also examined the performance of a few deep learning networks using the *h2o.ai* framework (https://www.h2o.ai). The achieved accuracy was 5-10% lower than the respective accuracy of the Random Forest classifier. However, the broad and diverse repertoire of networks available for deep learning definitely offers a lot of space for exploration of numerous network structures of increasing complexity that may even outperform the current Random Forest classifier.

Finally, the emergence and progressive growth of Nanopore Sequencing in recent years offers an abundance of new opportunities for designing and executing efficiently large-scale experiments without requiring massive equipment or budget. Of course, there is still a long way to cover until Nanopore Sequencing reaches the extremely high accuracy returned by Illumina Sequencing, which is the state-of-the-art sequencing technique. However, this is an ongoing endeavour and at the same time an opportunity for future work in order to design and develop new and more accurate classifiers for Nanopore Sequencing data. Additionally, classifiers could be trained to detect directly post-transcirptional modifications in miRNAs or other RNA molecules and even some types of DNA modifications, such as methylation. This would allow us to design and conduct computational experiments of really high complexity that could elucidate even further the complex regulatory mechanisms within cells.

# Chapter 8

# List of Publications

*(derived from this thesis)*

1. "Genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests"

   **DM Vitsios**, E Kentepozidou, L Quintais, E Benito-Gutiérrez, S van Dongen, MP Davis & AJ Enright.

   ***Nucleic Acids Research***, Volume 45, p.e177, doi: 10.1093/nar/gkx836 (2017).

2. "mRNA 3' uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome"

   M Morgan*, C Much*, M DiGiacomo, C Azzi, I Ivanova, **DM Vitsios**, J Pistolic, P Collier, P Moreira, V Benes, AJ Enright and D O'Carroll.

   ***Nature***, Volume 548, p.347-351, doi: 10.1038/nature23318 (2017).

3. "Large-scale analysis of microRNA expression, epi-transcriptomic features and biogenesis"

   **DM Vitsios**, MP Davis, S van Dongen, AJ Enright.

   ***Nucleic Acids Research***, Volume 45, p.1079-1090, doi: 10.1093/nar/gkw1031 (2017).

4. "A MILI-independent piRNA biogenesis pathway empowers partial germline reprogramming"

   L Vasiliauskaitė, **DM Vitsios**, RV Berrens, C Carrieri, W Reik, AJ Enright & D O'Carroll.

   ***Nature Structural & Molecular Biology***, Volume 24, p.604–606, doi: 10.1038/nsmb.3413 (2017).

5. ”In situ functional dissection of RNA cis-regulatory elements by multiplex CRISPR/Cas9 genome engineering”

   Q Wu*, Q Ferry*, Y Michaels, TA Baeumler, **DM Vitsios**, O Habib, R Arnold, X Jiang, S Maio, BR Steinkraus, M Tapia, P Piazza, N Xu, GA Holländer, TA Milne, JS Kim, AJ Enright, AR Bassett, Fulga T.

   ***Nature Communications***, Volume 8, p.2109, doi: 10.1038/s41467-017-00686-2 (2017).

6. ”Chimira: analysis of small RNA sequencing data and microRNA modifications”

   **DM Vitsios** and AJ Enright.

   ***Bioinformatics***, Volume 31, p.3365-3367, doi: 10.1093/bioinformatics/btv380 (2015).

# References


Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microrna target sites in mammalian mrnas. *elife*, 4:e05005.

Ali, P. S. S., Ghoshdastider, U., Hoffmann, J., Brutschy, B., and Filipek, S. (2012). Recognition of the let-7g mirna precursor by human lin28b. *FEBS letters*, 586:3986–3990.

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M. J., Tuschl, T., and Margalit, H. (2005). Clustering and conservation patterns of human micrornas. *Nucleic acids research*, 33(8):2697–2706.

Alvarez-Garcia, I. and Miska, E. A. (2005). Microrna functions in animal development and human disease. *Development*, 132(21):4653–4662.

Ambros, V. (2004). The functions of animal micrornas. *Nature*, 431(7006):350.

Ameres, S. L. and Zamore, P. D. (2013). Diversifying microrna sequence and function. *Nature reviews. Molecular cell biology*, 14(8):475.

Amin, N. D., Bai, G., Klug, J. R., Bonanomi, D., Pankratz, M. T., Gifford, W. D., Hinckley, C. A., Sternfeld, M. J., Driscoll, S. P., Dominguez, B., et al. (2015). Loss of motoneuron-specific microrna-218 causes systemic neuromuscular failure. *Science*, 350(6267):1525–1529.

Anderson, N. L. and Anderson, N. G. (1998). Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, 19(11):1853–1861.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, 12(7):878.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M. J., Kuramochi-Miyagawa, S., Nakano, T., et al. (2006). A novel class of small rnas bind to mili protein in mouse testes. *Nature*, 442(7099):203–207.

Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small rna profile during drosophila melanogaster development. *Developmental cell*, 5(2):337–350.

Aravin, A. A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K. F., Bestor, T., and Hannon, G. J. (2008). A pirna pathway primed by individual transposons is linked to de novo dna methylation in mice. *Molecular cell*, 31(6):785–799.

Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G. J. (2007). Developmentally regulated pirna clusters implicate mili in transposon control. *Science*, 316(5825):744–747.

Aravin, A. A., Van Der Heijden, G. W., Castañeda, J., Vagin, V. V., Hannon, G. J., and Bortvin, A. (2009). Cytoplasmic compartmentalization of the fetal pirna pathway in mice. *PLoS Genet*, 5(12):e1000764.

Ardekani, A. M. and Naeini, M. M. (2010). The role of micrornas in human diseases. *Avicenna journal of medical biotechnology*, 2(4):161.

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

Baer, B. W. and Kornberg, R. D. (1983). The protein responsible for the repeating structure of cytoplasmic poly(a)-ribonucleoprotein. *The Journal of cell biology*, 96:717–721.

Baltimore, D. (1964). In vitro synthesis of viral rna by the poliovirus rna polymerase. *Proceedings of the National Academy of Sciences*, 51(3):450–456.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2013). Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995.

Bartel, D. P. (2004). Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297.

Bartel, D. P. (2009). Micrornas: target recognition and regulatory functions. *cell*, 136(2):215–233.

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

Baulcombe, D. (2002). An rna microcosm. *Science*, 297(5589).

Baulcombe, D. (2004). Rna silencing in plants. *Nature*, 431(7006):356.

Bazzini, A. A., Lee, M. T., and Giraldez, A. J. (2012). Ribosome profiling shows that mir-430 reduces translation before causing mrna decay in zebrafish. *Science*, 336(6078):233–237.

Beitzinger, M., Peters, L., Zhu, J. Y., Kremmer, E., and Meister, G. (2007). Identification of human microrna targets from isolated argonaute protein complexes. *RNA biology*, 4(2):76–84.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer.

Berezikov, E., Robine, N., Samsonova, A., Westholm, J. O., Naqvi, A., Hung, J.-H., Okamura, K., Dai, Q., Bortolamiol-Becet, D., Martin, R., et al. (2011). Deep annotation of drosophila melanogaster micrornas yields insights into their processing, modification, and emergence. *Genome research*, 21(2):203–215.

Beveridge, N. J., Gardiner, E., Carroll, A., Tooney, P., and Cairns, M. (2010). Schizophrenia is associated with an increase in cortical microrna biogenesis. *Molecular psychiatry*, 15(12):1176.

Beyret, E., Liu, N., and Lin, H. (2012). pirna biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism. *Cell research*, 22(10):1429–1439.

Billi, A. C., Alessi, A. F., Khivansara, V., Han, T., Freeberg, M., Mitani, S., and Kim, J. K. (2012). The caenorhabditis elegans hen1 ortholog, henn-1, methylates and stabilizes select subclasses of germline small rnas. *PLoS genetics*, 8(4):e1002617.

Blackstock, W. P. and Weir, M. P. (1999). Proteomics: quantitative and physical mapping of cellular proteins. *Trends in biotechnology*, 17(3):121–127.

Blenkiron, C. and Miska, E. A. (2007). mirnas in cancer: approaches, aetiology, diagnostics and therapy. *Human molecular genetics*, 16(R1):R106–R113.

Blow, M. J., Grocock, R. J., van Dongen, S., Enright, A. J., Dicks, E., Futreal, P. A., Wooster, R., and Stratton, M. R. (2006). Rna editing of human micrornas. *Genome biology*, 7(4):R27.

Boele, J., Persson, H., Shin, J. W., Ishizu, Y., Newie, I. S., Søkilde, R., Hawkins, S. M., Coarfa, C., Ikeda, K., Takayama, K.-i., et al. (2014). Papd5-mediated 3′ adenylation and subsequent degradation of mir-21 is disrupted in proliferative disease. *Proceedings of the National Academy of Sciences*, 111(31):11467–11472.

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., and Zaretskaya, I. (2013). Blast: a more efficient report with usability improvements. *Nucleic acids research*, 41:W29–W33.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM.

Brandt, J., Schrauth, S., Veith, A.-M., Froschauer, A., Haneke, T., Schultheis, C., Gessler, M., Leimeister, C., and Volff, J.-N. (2005). Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene*, 345(1):101–111.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete small rna-generating loci as master regulators of transposon activity in drosophila. *Cell*, 128(6):1089–1103.

Brower, P. T., Gizang, E., Boreen, S. M., and Schultz, R. M. (1981). Biochemical studies of mammalian oogenesis: synthesis and stability of various classes of rna during growth of the mouse oocyte in vitro. *Developmental biology*, 86:373–383.

Brynjolfsson, E. and McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Burroughs, A. M., Ando, Y., de Hoon, M. J., Tomaru, Y., Nishibu, T., Ukekawa, R., Funakoshi, T., Kurokawa, T., Suzuki, H., Hayashizaki, Y., 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):1398–1410.

Camps, C., Buffa, F. M., Colella, S., Moore, J., Sotiriou, C., Sheldon, H., Harris, A. L., Gleadle, J. M., and Ragoussis, J. (2008). hsa-mir-210 is induced by hypoxia and is an independent prognostic factor in breast cancer. *Clinical cancer research*, 14(5):1340–1348.

Capece, V., Garcia Vizcaino, J. C., Vidal, R., Rahman, R.-U., Pena Centeno, T., Shomroni, O., Suberviola, I., Fischer, A., and Bonn, S. (2015). Oasis: online analysis of small rna deep sequencing data. *Bioinformatics*, 31(13):2205–2207.

Care, A., Catalucci, D., Felicetti, F., Bonci, D., Addario, A., Gallo, P., Bang, M.-L., Segnalini, P., Gu, Y., Dalton, N. D., et al. (2007). Microrna-133 controls cardiac hypertrophy. *Nature medicine*, 13(5):613.

Carmell, M. A., Girard, A., van de Kant, H. J., Bourc'his, D., Bestor, T. H., de Rooij, D. G., and Hannon, G. J. (2007). Miwi2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Developmental cell*, 12(4):503–514.

Caspersson, T. and Schultz, J. (1939). Pentose nucleotides in the cytoplasm of growing tissues. *Nature*, 143(3623):602–3.

Chang, H., Lim, J., Ha, M., and Kim, V. N. (2014). Tail-seq: genome-wide determination of poly(a) tail length and 3' end modifications. *Molecular cell*, 53:1044–1052.

Chang, S.-S., Zhang, Z., and Liu, Y. (2012). Rna interference pathways in fungi: mechanisms and functions. *Annual review of microbiology*, 66:305–323.

Chellapilla, K., Puri, S., and Simard, P. (2006). High performance convolutional neural networks for document processing. In *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft.

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):584–589.

Chen, J.-F., Murchison, E. P., Tang, R., Callis, T. E., Tatsuguchi, M., Deng, Z., Rojas, M., Hammond, S. M., Schneider, M. D., Selzman, C. H., et al. (2008). Targeted deletion of dicer in the heart leads to dilated cardiomyopathy and heart failure. *Proceedings of the National Academy of Sciences*, 105(6):2111–2116.

Chen, X. (2005). Microrna biogenesis and function in plants. *FEBS letters*, 579(26):5923–5931.

Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., Baek, D., Johnston, W. K., Russ, C., Luo, S., Babiarz, J. E., et al. (2010). Mammalian microrna: experimental evaluation of novel and previously annotated genes. *Genes & development*, 24(10):992–1009.

Church, G. M., Gao, Y., and Kosuri, S. (2012). Next-generation digital information storage in dna. *Science*, page 1226355.

Cifuentes, D., Xue, H., Taylor, D. W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G. J., Lawson, N. D., et al. (2010). A novel mirna processing pathway independent of dicer requires argonaute2 catalytic activity. *Science*, 328(5986):1694–1698.

Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, 22(12):3207–3220.

Clèries, R., Galvez, J., Espino, M., Ribes, J., Nunes, V., and de Heredia, M. L. (2012). Bootstratio: a web-based statistical analysis of fold-change in qpcr and rt-qpcr data using resampling methods. *Computers in biology and medicine*, 42(4):438–445.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Croce, C. M. and Calin, G. A. (2005). mirnas, cancer, and stem cell division. *Cell*, 122(1):6–7.

Cyranoski, D. (2016). Crispr gene-editing tested in a person for the first time. *Nature*.

Czech, B., Zhou, R., Erlich, Y., Brennecke, J., Binari, R., Villalta, C., Gordon, A., Perrimon, N., and Hannon, G. J. (2009). Hierarchical rules for argonaute loading in drosophila. *Molecular cell*, 36(3):445–456.

Dahm, R. (2005). Friedrich miescher and the discovery of dna. *Developmental biology*, 278(2):274–288.

Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A. J. (2013). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63(1):41–49.

De Fazio, S., Bartonicek, N., Di Giacomo, M., Abreu-Goodger, C., Sankar, A., Funaya, C., Antony, C., Moreira, P. N., Enright, A. J., and O'Carroll, D. (2011). The endonuclease activity of mili fuels pirna amplification that silences line1 elements. *Nature*, 480(7376):259–263.

De Leon, V., Johnson, A., and Bachvarova, R. (1983). Half-lives and relative amounts of stored and polysomal ribosomes and poly(a) + rna in mouse oocytes. *Developmental biology*, 98:400–408.

Deininger, P. L., Moran, J. V., Batzer, M. A., and Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Current opinion in genetics & development*, 13(6):651–658.

Di Giacomo, M., Comazzetto, S., Saini, H., De Fazio, S., Carrieri, C., Morgan, M., Vasiliauskaite, L., Benes, V., Enright, A. J., and O'Carroll, D. (2013). Multiple epigenetic mechanisms and the pirna pathway enforce line1 silencing during adult spermatogenesis. *Molecular cell*, 50(4):601–608.

DiCarlo, J. E., Norville, J. E., Mali, P., Rios, X., Aach, J., and Church, G. M. (2013). Genome engineering in saccharomyces cerevisiae using crispr-cas systems. *Nucleic acids research*, 41(7):4336–4343.

Djikeng, A., Shi, H., Tschudi, C., and Ullu, E. (2001). Rna interference in trypanosoma brucei: cloning of small interfering rnas provides evidence for retroposon-derived 24-26-nucleotide rnas. *RNA*, 7(11):1522–1530.

Djuranovic, S., Nahvi, A., and Green, R. (2012). mirna-mediated gene silencing by translational repression followed by mrna deadenylation and decay. *Science*, 336(6078):237–240.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797.

Eliseeva, I. A., Lyabin, D. N., and Ovchinnikov, L. P. (2013). Poly(a)-binding proteins: structure, domain organization, and activity regulation. *Biochemistry. Biokhimiia*, 78:1377–1391.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). Microrna targets in drosophila. *Genome biology*, 5(1):R1.

Eppig, J. J. and Schroeder, A. C. (1989). Capacity of mouse oocytes from preantral follicles to undergo embryogenesis and development to live young after growth, maturation, and fertilization in vitro. *Biology of reproduction*, 41(2):268–276.

Eulalio, A., Huntzinger, E., Nishihara, T., Rehwinkel, J., Fauser, M., and Izaurralde, E. (2009). Deadenylation is a widespread effect of mirna regulation. *RNA*, 15(1):21–32.

Fabian, M. R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mrna translation and stability by microrna. *Annual review of biochemistry*, 79:351–379.

Faridani, O. R., Abdullayev, I., Hagemann-Jensen, M., Schell, J. P., Lanner, F., and Sandberg, R. (2016). Single-cell sequencing of the small-rna transcriptome. *Nature Biotechnology*.

Feng, J., Sun, G., Yan, J., Noltner, K., Li, W., Buzin, C. H., Longmate, J., Heston, L. L., Rossi, J., and Sommer, S. S. (2009). Evidence for x-chromosomal schizophrenia associated with microrna alterations. *PloS one*, 4(7):e6121.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., et al. (1998). Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *nature*, 391(6669):806.

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):37–52.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mrnas are conserved targets of microrna. *Genome research*, 19(1):92–105.

Fromm, B., Billipp, T., Peck, L. E., Johansen, M., Tarver, J. E., King, B. L., Newcomb, J. M., Sempere, L. F., Flatmark, K., Hovig, E., et al. (2015). A uniform system for the annotation of vertebrate microrna genes and the evolution of the human micrornaome. *Annual review of genetics*, 49:213–242.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A., and Bartel, D. P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microrna. *Nature structural & molecular biology*, 18(10):1139–1146.

Gee, H. E., Camps, C., Buffa, F. M., Patiar, S., Winter, S. C., Betts, G., Homer, J., Corbridge, R., Cox, G., West, C. M., et al. (2010). hsa-mir-210 is a marker of tumor hypoxia and a prognostic factor in head and neck cancer. *Cancer*, 116(9):2148–2158.

Ghildiyal, M., Xu, J., Seitz, H., Weng, Z., and Zamore, P. D. (2010). Sorting of drosophila small silencing rnas partitions microrna* strands into the rna interference pathway. *RNA*, 16(1):43–56.

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Van Dongen, S., Inoue, K., Enright, A. J., and Schier, A. F. (2006). Zebrafish mir-430 promotes deadenylation and clearance of maternal mrnas. *science*, 312(5770):75–79.

Girard, A., Sachidanandam, R., Hannon, G. J., and Carmell, M. A. (2006). A germline-specific class of small rnas binds mammalian piwi proteins. *Nature*, 442(7099):199–202.

Gratz, S. J., Cummings, A. M., Nguyen, J. N., Hamm, D. C., Donohue, L. K., Harrison, M. M., Wildonger, J., and O'Connor-Giles, K. M. (2013). Genome engineering of drosophila with the crispr rna-guided cas9 nuclease. *Genetics*, 194(4):1029–1035.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). mirbase: tools for microrna genomics. *Nucleic acids research*, 36(suppl 1):D154–D158.

Grimson, A., Farh, K. K.-H., 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):91–105.

Grivna, S. T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small rnas in mouse spermatogenic cells. *Genes & development*, 20(13):1709–1714.

Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M. C. (2007). A slicer-mediated mechanism for repeat-associated sirna 5'end formation in drosophila. *Science*, 315(5818):1587–1590.

Guo, X. and Li, X.-J. (2015). Targeted genome editing in primate embryos. *Cell research*, 25(7):767.

Ha, M. and Kim, V. N. (2014). Regulation of microrna biogenesis. *Nature reviews. Molecular cell biology*, 15:509–524.

Hagan, J. P., Piskounova, E., and Gregory, R. I. (2009). Lin28 recruits the tutase zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells. *Nature structural & molecular biology*, 16:1021–1025.

Hamilton, A. J. and Baulcombe, D. C. (1999). A species of small antisense rna in posttranscriptional gene silencing in plants. *Science*, 286(5441):950–952.

Han, B. W., Wang, W., Li, C., Weng, Z., and Zamore, P. D. (2015). pirna-guided transposon cleavage initiates zucchini-dependent, phased pirna production. *Science*, 348(6236):817–821.

He, Y., Lin, J., Kong, D., Huang, M., Xu, C., Kim, T.-K., Etheridge, A., Luo, Y., Ding, Y., and Wang, K. (2015). Current state of circulating micrornas as cancer biomarkers. *Clinical chemistry*, 61(9):1138–1155.

Heo, I., Ha, M., Lim, J., Yoon, M.-J., Park, J.-E., Kwon, S. C., Chang, H., and Kim, V. N. (2012). Mono-uridylation of pre-microrna as a key step in the biogenesis of group ii let-7 microrna. *Cell*, 151(3):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*, 32:276–284.

Heo, I., Joo, C., Kim, Y.-K., Ha, M., Yoon, M.-J., Cho, J., Yeom, K.-H., Han, J., and Kim, V. N. (2009). Tut4 in concert with lin28 suppresses microrna biogenesis through pre-microrna uridylation. *Cell*, 138(4):696–708.

Hibio, N., Hino, K., Shimizu, E., Nagata, Y., and Ui-Tei, K. (2012). Stability of mirna 5′ terminal and seed regions is correlated with experimentally observed mirna-mediated silencing efficacy. *Scientific reports*, 2:996.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., and Zamir, A. (1965). Structure of a ribonucleic acid. *Science*, 147(3664):1462–1465.

Hommers, L. G., Domschke, K., and Deckert, J. (2015). Heterogeneity and individuality: microrna in mental disorders. *Journal of neural transmission*, 122(1):79–97.

Hong, X., Hammell, M., Ambros, V., and Cohen, S. (2009). Identification of microrna targets by immunopurification of ago1 rnps: selection for a distinct class of targets. *PNAS*, 106:15085–90.

Hwang, W. Y., Fu, Y., Reyon, D., Maeder, M. L., Tsai, S. Q., Sander, J. D., Peterson, R. T., Yeh, J. J., and Joung, J. K. (2013). Efficient genome editing in zebrafish using a crispr-cas system. *Nature biotechnology*, 31(3):227–229.

Iacobuzio-Donahue, C. A. (2009). Epigenetic changes in cancer. *Annual Review of Pathological Mechanical Disease*, 4:229–249.

Ibrahim, F., Rymarquis, L. A., Kim, E.-J., Becker, J., Balassa, E., Green, P. J., and Cerutti, H. (2010). Uridylation of mature mirnas and sirnas by the mut68 nucleotidyltransferase promotes their degradation in chlamydomonas. *Proceedings of the National Academy of Sciences*, 107(8):3906–3911.

Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., and Lozano, J. A. (2010). Machine learning: an indispensable tool in bioinformatics. *Methods in molecular biology (Clifton, N.J.)*, 593:25–48.

Iorio, M. V., Visone, R., Di Leva, G., Donati, V., Petrocca, F., Casalini, P., Taccioli, C., Volinia, S., Liu, C.-G., Alder, H., et al. (2007). Microrna signatures in human ovarian cancer. *Cancer research*, 67(18):8699–8707.

Ipsaro, J. J., Haase, A. D., Knott, S. R., Joshua-Tor, L., and Hannon, G. J. (2012). The structural biochemistry of zucchini implicates it as a nuclease in pirna biogenesis. *Nature*, 491(7423):279–283.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in escherichia coli, and identification of the gene product. *Journal of bacteriology*, 169:5429–5433.

Jansen, R., Embden, J., Gaastra, W., Schouls, L., et al. (2002). Identification of genes that are associated with dna repeats in prokaryotes. *Molecular microbiology*, 43(6):1565–1575.

Jansson, M. D. and Lund, A. H. (2012). Microrna and cancer. *Molecular oncology*, 6(6):590–610.

Jha, A. and Shankar, R. (2013). mireader: Discovering novel mirnas in species without sequenced genome. *PloS one*, 8(6):e66857.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2008). mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic acids research*, 37(suppl_1):D98–D104.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). Rna-programmed genome editing in human cells. *eLife*, 2:e00471. Original DateCompleted: 20130207, Original DateCompleted: 20140206.

Jones, M. R., Blahna, M. T., Kozlowski, E., Matsuura, K. Y., Ferrari, J. D., Morris, S. A., Powers, J. T., Daley, G. Q., Quinton, L. J., and Mizgerd, J. P. (2012). Zcchc11 uridylates mature mirnas to enhance neonatal igf-1 expression, growth, and survival. *PLoS genetics*, 8(11):e1003105.

Juan, L., Tong, H.-l., Zhang, P., Guo, G., Wang, Z., Wen, X., Dong, Z., and Tian, Y.-p. (2014). Identification and characterization of novel serum microrna candidates from deep sequencing in cervical cancer patients. *Scientific reports*, 4.

Kai, Z. S. and Pasquinelli, A. E. (2010). Microrna assassins: factors that regulate the disappearance of mirnas. *Nature structural & molecular biology*, 17(1):5–10.

Kamminga, L. M., Luteijn, M. J., Den Broeder, M. J., Redl, S., Kaaij, L. J., Roovers, E. F., Ladurner, P., Berezikov, E., and Ketting, R. F. (2010). Hen1 is required for oocyte development and pirna stability in zebrafish. *The EMBO journal*, 29(21):3688–3700.

Kamminga, L. M., Van Wolfswinkel, J. C., Luteijn, M. J., Kaaij, L. J., Bagijn, M. P., Sapetschnig, A., Miska, E. A., Berezikov, E., and Ketting, R. F. (2012). Differential impact of the hen1 homolog henn-1 on 21u and 26g rnas in the germline of caenorhabditis elegans. *PLoS genetics*, 8(7):e1002702.

Kang, W. and Friedländer, M. R. (2015). Computational prediction of mirna genes from small rna sequencing data. *Frontiers in bioengineering and biotechnology*, 3.

Katoh, T., Sakaguchi, Y., Miyauchi, K., Suzuki, T., Kashiwabara, S.-i., 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):433–438.

Kazazian Jr, H. H. and Moran, J. V. (1998). The impact of l1 retrotransposons on the human genome. *Nature genetics*, 19(1):19–24.

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

Kim, B., Ha, M., Loeff, L., Chang, H., Simanshu, D. K., Li, S., Fareh, M., Patel, D. J., Joo, C., and Kim, V. N. (2015). Tut7 controls the fate of precursor micrornas by using three different uridylation mechanisms. *The EMBO journal*, 34(13):1801–1815.

Kim, J., Levy, E., Ferbrache, A., Stepanowsky, P., Farcas, C., Wang, S., Brunner, S., Bath, T., Wu, Y., and Ohno-Machado, L. (2014). Magi: a node. js web service for fast microrna-seq analysis in a gpu infrastructure. *Bioinformatics*, 30(19):2826–2827.

Kim, Y.-K., Heo, I., and Kim, V. N. (2010). Modifications of small rnas and their associated proteins. *Cell*, 143(5):703–709.

Kim, Y.-K., Kim, B., and Kim, V. N. (2016). Re-evaluation of the roles of drosha, exportin 5, and dicer in microrna biogenesis. *Proceedings of the National Academy of Sciences*, 113(13):E1881–E1889.

Kirino, Y. and Mourelatos, Z. (2007a). The mouse homolog of hen1 is a potential methylase for piwi-interacting rnas. *RNA*, 13(9):1397–1401.

Kirino, Y. and Mourelatos, Z. (2007b). Mouse piwi-interacting rnas are 2'-o-methylated at their 3' termini. *Nature structural & molecular biology*, 14:347–348.

Korpal, M., Ell, B. J., Buffa, F. M., Ibrahim, T., Blanco, M. A., Celià-Terrassa, T., Mercatali, L., Khan, Z., Goodarzi, H., Hua, Y., et al. (2011). Direct targeting of sec23a by mir-200s influences cancer cell secretome and promotes metastatic colonization. *Nature medicine*, 17(9):1101–1108.

Krawetz, S. A., Kruger, A., Lalancette, C., Tagett, R., Anton, E., Draghici, S., and Diamond, M. P. (2011). A survey of small rnas in human sperm. *Human reproduction*, page der329.

Krol, J., Loedige, I., and Filipowicz, W. (2010). The widespread regulation of microrna biogenesis, function and decay. *Nature Reviews Genetics*, 11(9):597–610.

Krützfeldt, J., Rajewsky, N., Braich, R., Rajeev, K. G., et al. (2005). Silencing of microrna in vivo with'antagomirs'. *nature*, 438(7068):685.

Kuenne, C., Preussner, J., Herzog, M., Braun, T., and Looso, M. (2014). Mirpipe: quantification of micrornas in niche model organisms. *Bioinformatics*, 30(23):3412–3413.

Kumar, M. S., Lu, J., Mercer, K. L., Golub, T. R., and Jacks, T. (2007). Impaired microrna processing enhances cellular transformation and tumorigenesis. *Nature genetics*, 39(5).

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T. W., Isobe, T., Asada, N., Fujita, Y., Ikawa, M., Iwai, N., Okabe, M., Deng, W., et al. (2004). Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development*, 131(4):839–849.

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T. W., et al. (2008). Dna methylation of retrotransposon genes is regulated by piwi family members mili and miwi2 in murine fetal testes. *Genes & development*, 22(7):908–917.

Lagana, A., Russo, F., Sismeiro, C., Giugno, R., Pulvirenti, A., and Ferro, A. (2010). Variability in the incidence of mirnas and genes in fragile sites and the role of repeats and cpg islands in the distribution of genetic material. *PloS one*, 5(6):e11166.

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):735–739.

Lamy, P., Andersen, C. L., Dyrskjot, L., Torring, N., Orntoft, T., and Wiuf, C. (2006). Are micrornas located in genomic regions associated with cancer? *British journal of cancer*, 95(10):1415.

Lan, H., Lu, H., Wang, X., and Jin, H. (2015). Micrornas as potential biomarkers in cancer: opportunities and challenges. *BioMed research international*, 2015.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., et al. (2007). A mammalian microrna expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–1414.

Lanford, R. E., Hildebrandt-Eriksen, E. S., Petri, A., Persson, R., Lindow, M., Munk, M. E., Kauppinen, S., and Ørum, H. (2010). Therapeutic silencing of microrna-122 in primates with chronic hepatitis c virus infection. *Science*, 327(5962):198–201.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359.

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., et al. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics*, pages 86–112.

Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P., and Kingston, R. E. (2006). Characterization of the pirna complex from rat testes. *Science*, 313(5785):363–367.

Lecellier, C.-H., Dunoyer, P., Arar, K., Lehmann-Che, J., Eyquem, S., Himber, C., Saïb, A., and Voinnet, O. (2005). A cellular microrna mediates antiviral defense in human cells. *Science*, 308(5721):557–560.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee, C.-T., Risom, T., and Strauss, W. M. (2007). Evolutionary conservation of microrna regulatory circuits: an examination of microrna gene complexity and conserved microrna-target interactions through metazoan phylogeny. *DNA and cell biology*, 26(4):209–218.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009a). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM.

Lee, I., Ajay, S. S., Yook, J. I., Kim, H. S., Hong, S. H., Kim, N. H., Dhanasekaran, S. M., Chinnaiyan, A. M., and Athey, B. D. (2009b). New class of microrna targets containing simultaneous 5′-utr and 3′-utr interaction sites. *Genome research*, 19(7):1175–1183.

Lehmann, U., Hasemeier, B., Christgen, M., Müller, M., Römermann, D., Länger, F., and Kreipe, H. (2008). Epigenetic inactivation of microrna gene hsa-mir-9-1 in human breast cancer. *The Journal of pathology*, 214(1):17–24.

Leinonen, R., Akhtar, R., Birney, E., Bonfield, J., Bower, L., Corbett, M., Cheng, Y., Demiralp, F., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Hunter, C., Jang, M., Leonard, S., Lin, Q., Lopez, R., Maguire, M., McWilliam, H., Plaister, S., Radhakrishnan, R., Sobhany, S., Slater, G., Ten Hoopen, P., Valentin, F., Vaughan, R., Zalunin, V., Zerbino, D., and Cochrane, G. (2010). Improvements to services at the european nucleotide archive. *Nucleic acids research*, 38:D39–D45.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *cell*, 120(1):15–20.

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):787–798.

Lewis, M. A., Buniello, A., Hilton, J. M., Zhu, F., Zhang, W. I., Evans, S., Van Dongen, S., Enright, A. J., and Steel, K. P. (2016). Exploring regulatory networks of mir-96 in the developing inner ear. *Scientific reports*, 6:23363.

Lewis, M. A., Quint, E., Glazier, A. M., Fuchs, H., De Angelis, M. H., Langford, C., Van Dongen, S., Abreu-Goodger, C., Piipari, M., Redshaw, N., et al. (2009). An enu-induced mutation of mir-96 associated with progressive hearing loss in mice. *Nature genetics*, 41(5):614–618.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects mirnas and sirnas from a 3′-end uridylation activity in arabidopsis. *Current biology*, 15(16):1501–1507.

Li, S.-C., Tsai, K.-W., Pan, H.-W., Jeng, Y.-M., Ho, M.-R., and Li, W.-H. (2012). Microrna 3'end nucleotide modification patterns and arm selection preference in liver tissues. *BMC systems biology*, 6(2):S14.

Lim, J., Ha, M., Chang, H., Kwon, S. C., Simanshu, D. K., Patel, D. J., and Kim, V. N. (2014). Uridylation by tut4 and tut7 marks mrna for degradation. *Cell*, 159:1365–1376.

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., et al. (2005). Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, 433(7027):769.

Liu, G., Sun, Y., Ji, P., Li, X., Cogdell, D., Yang, D., Parker Kerrigan, B. C., Shmulevich, I., Chen, K., Sood, A. K., et al. (2014a). Mir-506 suppresses proliferation and induces senescence by directly targeting the cdk4/6–foxm1 axis in ovarian cancer. *The Journal of pathology*, 233(3):308–318.

Liu, J.-J., Kong, I. I., Zhang, G.-C., Jayakody, L. N., Kim, H., Xia, P.-F., Kwak, S., Sung, B. H., Sohn, J.-H., Walukiewicz, H. E., et al. (2016). Metabolic engineering of probiotic saccharomyces boulardii. *Applied and environmental microbiology*, 82(8):2280–2287.

Liu, X., Zheng, Q., Vrettos, N., Maragkakis, M., Alexiou, P., Gregory, B. D., and Mourelatos, Z. (2014b). A microrna precursor surveillance system in quality control of microrna synthesis. *Molecular cell*, 55:868–879.

Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002). Cleavage of scarecrow-like mrna targets directed by a class of arabidopsis mirna. *Science*, 297(5589):2053–2056.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

Lu, S., Sun, Y.-H., and Chiang, V. L. (2009). Adenylation of plant mirnas. *Nucleic Acids Research*, 37(6):1878–1885.

Ma, J., Flemr, M., Stein, P., Berninger, P., Malik, R., Zavolan, M., Svoboda, P., and Schultz, R. M. (2010). Microrna activity is suppressed in mouse oocytes. *Current biology*, 20(3):265–270.

Ma, J.-Y., Li, M., Luo, Y.-B., Song, S., Tian, D., Yang, J., Zhang, B., Hou, Y., Schatten, H., Liu, Z., et al. (2013). Maternal factors required for oocyte developmental competence in mice: transcriptome analysis of non-surrounded nucleolus (nsn) and surrounded nucleolus (sn) oocytes. *Cell cycle*, 12(12):1928–1938.

Maes, O. C., Chertkow, H. M., Wang, E., and Schipper, H. M. (2009). Microrna: implications for alzheimer disease and other human cns disorders. *Current genomics*, 10(3):154–168.

Makarova, K. S., Wolf, Y. I., Van der Oost, J., and Koonin, E. V. (2009). Prokaryotic homologs of argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biology direct*, 4(1):29.

Malone, C. D. and Hannon, G. J. (2009). Small rnas as guardians of the genome. *Cell*, 136:656–668.

Mapleson, D., Moxon, S., Dalmay, T., and Moulton, V. (2013). Mirplex: A tool for identifying mirnas in high-throughput srna datasets without a genome. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 320(1):47–56.

Mathé, C., Sagot, M.-F., Schiex, T., and Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic acids research*, 30(19):4103–4117.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36:344–355.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

McGinnis, S. and Madden, T. L. (2004). Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, 32(suppl_2):W20–W25.

Mendell, J. T. and Olson, E. N. (2012). Micrornas in stress signaling and human disease. *Cell*, 148(6):1172–1187.

Meyer, D., Leisch, F., and Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, 55(1):169–186.

Miescher, F. and Schmiedeberg, O. (1896). Physiologisch-chemische untersuchungen über die lachsmilch. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 37(2):100–155.

Mishra, P. J., Mishra, P. J., Banerjee, D., and Bertino, J. R. (2008). Mirsnps or mir-polymorphisms, new players in microrna mediated regulation of the cell: Introducing microrna pharmacogenomics. *Cell Cycle*, 7(7):853–858.

Mizutani, S. and Temin, H. M. (1970). An rna-dependent dna polymerase in virions of rous sarcoma virus. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 35, pages 847–849. Cold Spring Harbor Laboratory Press.

Mohn, F., Handler, D., and Brennecke, J. (2015). pirna-guided slicing specifies transcripts for zucchini-dependent, phased pirna biogenesis. *Science*, 348(6236):812–817.

Molaro, A., Falciatori, I., Hodges, E., Aravin, A. A., Marran, K., Rafii, S., McCombie, W. R., Smith, A. D., and Hannon, G. J. (2014). Two waves of de novo methylation during mouse germ cell development. *Genes & development*, 28(14):1544–1549.

Morozova, N., Zinovyev, A., Nonne, N., Pritchard, L.-L., Gorban, A. N., and Harel-Bellan, A. (2012). Kinetic signatures of microrna modes of action. *RNA*, 18(9):1635–1655.

Morozova, O. and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264.

Mosher, R. A., Schwach, F., Studholme, D., and Baulcombe, D. C. (2008). Polivb influences rna-directed dna methylation independently of its role in sirna biogenesis. *Proceedings of the National Academy of Sciences*, 105(8):3145–3150.

Mraz, M. and Pospisilova, S. (2012). Micrornas in chronic lymphocytic leukemia: from causality to associations and back. *Expert review of hematology*, 5(6):579–581.

Mullen, T. E. and Marzluff, W. F. (2008). Degradation of histone mrna requires oligouridylation followed by decapping and simultaneous degradation of the mrna both 5' to 3' and 3' to 5'. *Genes & development*, 22:50–65.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. *The plant cell*, 2(4):279–289.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., et al. (2014). Rfam 12.0: updates to the rna families database. *Nucleic acids research*, page gku1063.

Needleman, S. (1970). Needleman-wunsch algorithm for sequence similarity searches. *J Mol Biol*, 48:443–453.

Network, C. G. A. R. et al. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315.

Newman, M. A., Mani, V., and Hammond, S. M. (2011). Deep sequencing of microrna precursors reveals extensive 3′ end modification. *RNA*, 17(10):1795–1803.

Ng, A. (2016). Deep learning: What's next. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1–1. International Foundation for Autonomous Agents and Multiagent Systems.

Nichols, J. and Smith, A. (2009). Naive and primed pluripotent states. *Cell stem cell*, 4(6):487–492.

Nishimasu, H., Ishizu, H., Saito, K., Fukuhara, S., Kamatani, M. K., Bonnefond, L., Matsumoto, N., Nishizawa, T., Nakanaga, K., Aoki, J., et al. (2012). Structure and function of zucchini endoribonuclease in pirna biogenesis. *Nature*, 491(7423):284–287.

O'Donnell, K. A. and Boeke, J. D. (2007). Mighty piwis defend the germline against genome intruders. *Cell*, 129:37–44.

Oh, K.-S. and Jung, K. (2004). Gpu implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314.

Okamura, K., Liu, N., and Lai, E. C. (2009). Distinct mechanisms for microrna strand selection by drosophila argonautes. *Molecular cell*, 36(3):431–444.

Pan, H., O'Brien, M. J., Wigglesworth, K., Eppig, J. J., and Schultz, R. M. (2005). Transcript profiling during mouse oocyte development and the effect of gonadotropin priming and development in vitro. *Developmental biology*, 286(2):493–506.

Pantano, L., Estivill, X., and Martí, E. (2009). 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):e34–e34.

Pasquinelli, A. E. (2012). Micrornas and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature reviews. Genetics*, 13:271–282.

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., et al. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory rna. *Nature*, 408(6808):86.

Pauling, L., Corey, R. B., and Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grässer, F. A., van Dyk, L. F., Ho, C. K., Shuman, S., Chien, M., et al. (2005). Identification of micrornas of the herpesvirus family. *Nature methods*, 2(4):269.

Pratt, A. J. and MacRae, I. J. (2009). The rna-induced silencing complex: a versatile gene-silencing machine. *Journal of Biological Chemistry*, 284(27):17897–17901.

Quah, S., Hui, J. H., and Holland, P. W. (2015). A burst of mirna innovation in the early evolution of butterflies and moths. *Molecular biology and evolution*, 32(5):1161–1174.

Raina, R., Madhavan, A., and Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880. ACM.

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., et al. (2000). The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *nature*, 403(6772):901.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). Micrornas in plants. *Genes & development*, 16(13):1616–1626.

Rich, A. and Davies, D. R. (1956). A new two stranded helical structure: polyadenylic acid and polyuridylic acid. *Journal of the American Chemical Society*, 78(14):3548–3549.

Rissland, O. S. and Norbury, C. J. (2009). Decapping is preceded by 3' uridylation in a novel pathway of bulk mrna turnover. *Nat. Struct. Mol. Biol.*, 16:616–623.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584.

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

Saini, H. K., Griffiths-Jones, S., and Enright, A. J. (2007). Genomic analysis of human microrna transcripts. *Proceedings of the National Academy of Sciences*, 104(45):17719–17724.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2):206–226.

Samuel, C. E. (2011). *Adenosine Deaminases Acting on RNA (ADARs) and A-to-I Editing*, volume 353. Springer Science & Business Media.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, 94(3):441IN19447–446IN20448.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467.

Sanmiguel, P. and Bennetzen, J. L. (1998). Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Annals of Botany*, 82(suppl 1):37–44.

Schratt, G. (2009). micrornas at the synapse. *Nature reviews. Neuroscience*, 10(12):842.

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

Scott, G. K., Mattie, M. D., Berger, C. E., Benz, S. C., and Benz, C. C. (2006). Rapid alteration of microrna levels by histone deacetylase inhibition. *Cancer research*, 66(3):1277–1281.

Seto, A. G., Kingston, R. E., and Lau, N. C. (2007). The coming of age for piwi proteins. *Molecular cell*, 26(5):603–609.

Shah, M. Y. and Calin, G. A. (2014). Micrornas as therapeutic targets in human cancers. *Wiley Interdisciplinary Reviews: RNA*, 5(4):537–548.

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

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):246–258.

Skalsky, R. L. and Cullen, B. R. (2010). Viruses, micrornas, and host interactions. *Annual review of microbiology*, 64:123–141.

Slack, F. J., Basson, M., Liu, Z., Ambros, V., Horvitz, H. R., and Ruvkun, G. (2000). The lin-41 rbcc gene acts in the c. elegans heterochronic pathway between the let-7 regulatory rna and the lin-29 transcription factor. *Molecular cell*, 5(4):659–669.

Song, J., Song, J., Mo, B., and Chen, X. (2015). Uridylation and adenylation of rnas. *Science China Life Sciences*, 58(11):1057–1066.

Song, Y., Li, L., Ou, Y., Gao, Z., Li, E., Li, X., Zhang, W., Wang, J., Xu, L., Zhou, Y., et al. (2014). Identification of genomic alterations in oesophageal squamous cell cancer. *Nature*, 509(7498):91.

Stark, A., Brennecke, J., Bushati, N., Russell, R. B., and Cohen, S. M. (2005). Animal micrornas confer robustness to gene expression and have a significant impact on 3′ utr evolution. *Cell*, 123(6):1133–1146.

Stark, A., Brennecke, J., Russell, R. B., and Cohen, S. M. (2003). Identification of drosophila microrna targets. *PLoS biology*, 1(3):e60.

Stocks, M. B., Moxon, S., Mapleson, D., Woolfenden, H. C., Mohorianu, I., Folkes, L., Schwach, F., Dalmay, T., and Moulton, V. (2012). The uea srna workbench: a suite of tools for analysing and visualizing next generation sequencing microrna and small rna datasets. *Bioinformatics*, 28(15):2059–2061.

Suh, N., Baehner, L., Moltzahn, F., Melton, C., Shenoy, A., Chen, J., and Blelloch, R. (2010). Microrna function is globally suppressed in mouse oocytes and early embryos. *Current Biology*, 20(3):271–277.

Sun, Z., Evans, J., Bhagwate, A., Middha, S., Bockol, M., Yan, H., and Kocher, J.-P. (2014). Cap-mirseq: a comprehensive analysis pipeline for microrna sequencing data. *BMC genomics*, 15(1):423.

Svoboda, P., Franke, V., and Schultz, R. M. (2015). Chapter nine-sculpting the transcriptome during the oocyte-to-embryo transition in mouse. *Current topics in developmental biology*, 113:305–349.

Tadros, W. and Lipshitz, H. D. (2009). The maternal-to-zygotic transition: a play in two acts. *Development*, 136(18):3033–3042.

Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M., and Hannon, G. J. (2008). Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes. *Nature*, 453:534–538.

Tatsuguchi, M., Seok, H. Y., Callis, T. E., Thomson, J. M., Chen, J.-F., Newman, M., Rojas, M., Hammond, S. M., and Wang, D.-Z. (2007). Expression of microrna s is dynamically regulated during cardiomyocyte hypertrophy. *Journal of molecular and cellular cardiology*, 42(6):1137–1141.

Tay, Y., Zhang, J., Thomson, A. M., Lim, B., and Rigoutsos, I. (2008). Micrornas to nanog, oct4 and sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455(7216):1124.

Thornton, J. E., Du, P., Jing, L., Sjekloca, L., Lin, S., Grossi, E., Sliz, P., Zon, L. I., and Gregory, R. I. (2014). Selective microrna uridylation by zcchc6 (tut7) and zcchc11 (tut4). *Nucleic acids research*, 42(18):11777–11791.

Thum, T., Galuppo, P., Wolf, C., Fiedler, J., Kneitz, S., van Laake, L. W., Doevendans, P. A., Mummery, C. L., Borlak, J., Haverich, A., et al. (2007). Micrornas in the human heart. *Circulation*, 116(3):258–267.

Toyota, M., Suzuki, H., Sasaki, Y., Maruyama, R., Imai, K., Shinomura, Y., and Tokino, T. (2008). Epigenetic silencing of microrna-34b/c and b-cell translocation gene 4 is associated with cpg island methylation in colorectal cancer. *Cancer research*, 68(11):4123–4132.

Triboulet, R., Chang, H.-M., LaPierre, R. J., and Gregory, R. I. (2009). Post-transcriptional control of dgcr8 expression by the microprocessor. *RNA*, 15(6):1005–1011.

Trifonov, E. N. (1990). Making sense of the human genome. *Structure and methods: proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics held at the State University of New York at Albany, June 6-10, 1989/edited by RH Sarma & MH Sarma.*

Tuschl, T., Zamore, P. D., Lehmann, R., Bartel, D. P., and Sharp, P. A. (1999). Targeted mrna degradation by double-stranded rna in vitro. *Genes & development*, 13(24):3191–3197.

Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P. D. (2006). A distinct small rna pathway silences selfish genetic elements in the germline. *Science (New York, N.Y.)*, 313:320–324.

van Dongen, S. and Abreu-Goodger, C. (2012). Using mcl to extract clusters from networks. *Bacterial Molecular Networks: Methods and Protocols*, pages 281–295.

Van Rooij, E., Sutherland, L. B., Liu, N., Williams, A. H., McAnally, J., Gerard, R. D., Richardson, J. A., and Olson, E. N. (2006). A signature pattern of stress-responsive micrornas that can evoke cardiac hypertrophy and heart failure. *Proceedings of the National Academy of Sciences*, 103(48):18255–18260.

van Rooij, E., Sutherland, L. B., Qi, X., Richardson, J. A., Hill, J., and Olson, E. N. (2007). Control of stress-dependent cardiac growth and gene expression by a microrna. *Science*, 316(5824):575–579.

Vesely, C., Tauber, S., Sedlazeck, F. J., Tajaddod, M., von Haeseler, A., and Jantsch, M. F. (2014). Adar2 induces reproducible changes in sequence and abundance of mature microrna in the mouse brain. *Nucleic acids research*, 42:12155–12168.

Viswanathan, S. R., Daley, G. Q., and Gregory, R. I. (2008). Selective blockade of microrna processing by lin28. *Science (New York, N.Y.)*, 320:97–100.

Vitsios, D. M., Davis, M. P., van Dongen, S., and Enright, A. J. (2017). Large-scale analysis of microrna expression, epi-transcriptomic features and biogenesis. *Nucleic acids research*, 45:1079–1090.

Vitsios, D. M. and Enright, A. J. (2015). Chimira: analysis of small rna sequencing data and microrna modifications. *Bioinformatics*, page btv380.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microrna. *Cell*, 136:669–687.

Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F., and Jaenisch, R. (2013). One-step generation of mice carrying mutations in multiple genes by crispr/cas-mediated genome engineering. *cell*, 153(4):910–918.

Wang, S., Peng, J., Ma, J., and Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports*, 6.

Wang, X.-J., Reyes, J. L., Chua, N.-H., and Gaasterland, T. (2004). Prediction and identification of arabidopsis thaliana microrna and their mrna targets. *Genome biology*, 5(9):R65.

Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., Sasaki, H., Minami, N., and Imai, H. (2006). Identification and characterization of two novel classes of small rna in the mouse germline: retrotransposon-derived sirna in oocytes and germline small rna in testes. *Genes & development*, 20(13):1732–1743.

Wee, E., Peters, K., Nair, S., Hulf, T., Stein, S., Wagner, S., Bailey, P., Lee, S., Qu, W., Brewster, B., et al. (2012). Mapping the regulatory sequences controlling 93 breast cancer-associated mirna genes leads to the identification of two functional promoters of the hsa-mir-200b cluster, methylation of which is associated with metastasis or hormone receptor status in advanced breast cancer. *Oncogene*, 31(38):4182.

Wienholds, E. and Plasterk, R. H. (2005). Microrna function in animal development. *FEBS letters*, 579(26):5911–5922.

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):855–862.

Williams, A. (2008). Functional aspects of animal microrna. *Cellular and Molecular Life Sciences*, 65(4):545–562.

Wootton, J. C. and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry*, 17(2):149–163.

Wu, H. and Mo, Y.-Y. (2009). Targeting mir-205 in breast cancer. *Expert opinion on therapeutic targets*, 13(12):1439–1448.

Wu, X., Kriz, A. J., and Sharp, P. A. (2014). Target specificity of the crispr-cas9 system. *Quantitative biology (Beijing, China)*, 2:59–70.

Xiao, J., Luo, X., Lin, H., Zhang, Y., Lu, Y., Wang, N., Zhang, Y., Yang, B., and Wang, Z. (2007). Microrna mir-133 represses herg k+ channel expression contributing to qt prolongation in diabetic hearts. *Journal of Biological Chemistry*, 282(17):12363–12367.

Yang, B., Lin, H., Xiao, J., Lu, Y., Luo, X., Li, B., Zhang, Y., Xu, C., Bai, Y., Wang, H., et al. (2007). The muscle-specific microrna mir-1 regulates cardiac arrhythmogenic potential by targeting gja1 and kcnj2. *Nature medicine*, 13(4):486.

Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., and Zhou, Y. (2016). Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings in bioinformatics*, page bbw129.

Yang, Z., Ebright, Y. W., Yu, B., and Chen, X. (2006). Hen1 recognizes 21–24 nt small rna duplexes and deposits a methyl group onto the 2′ oh of the 3′ terminal nucleotide. *Nucleic acids research*, 34(2):667–675.

Yu, B. and Chen, X. (2010). Analysis of mirna modifications. *Plant MicroRNAs: Methods and Protocols*, pages 137–148.

Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R. W., Steward, R., and Chen, X. (2005). Methylation as a crucial step in plant microrna biogenesis. *Science*, 307(5711):932–935.

Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., and Anderson, T. A. (2006). Conservation and divergence of plant microrna genes. *The Plant Journal*, 46(2):243–259.

Zhang, G.-C., Kong, I. I., Kim, H., Liu, J.-J., Cate, J. H., and Jin, Y.-S. (2014). Construction of a quadruple auxotrophic mutant of an industrial polyploid saccharomyces cerevisiae strain by using rna-guided cas9 nuclease. *Applied and environmental microbiology*, 80(24):7694–7701.

Zhang, Y., Xu, B., Yang, Y., Ban, R., Zhang, H., Jiang, X., Cooke, H. J., Xue, Y., and Shi, Q. (2012). Cpss: a computational platform for the analysis of small rna deep sequencing data. *Bioinformatics*, 28(14):1925–1927.

Zhao, Y., Ransom, J. F., Li, A., Vedantham, V., von Drehle, M., Muth, A. N., Tsuchihashi, T., McManus, M. T., Schwartz, R. J., and Srivastava, D. (2007). Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking mirna-1-2. *Cell*, 129(2):303–317.

Zhao, Y., Samal, E., and Srivastava, D. (2005). Serum response factor regulates a muscle-specific microrna that targets hand2 during cardiogenesis. *Nature*, 436(7048):214.

# Appendix A

# *Chimira*: user interface and functionality

Chimira is an online tool for analysing large amounts of small RNA-Seq data and acquiring their modifications profiles. It allows the mapping of input sequences to miRBase in order to decipher miRNA expression content and modifications related with the input sequences (3'/5'-modifications, ADAR edits and SNPs). This functionality is provided by a user-friendly interface in a web-app (Figure A.1).

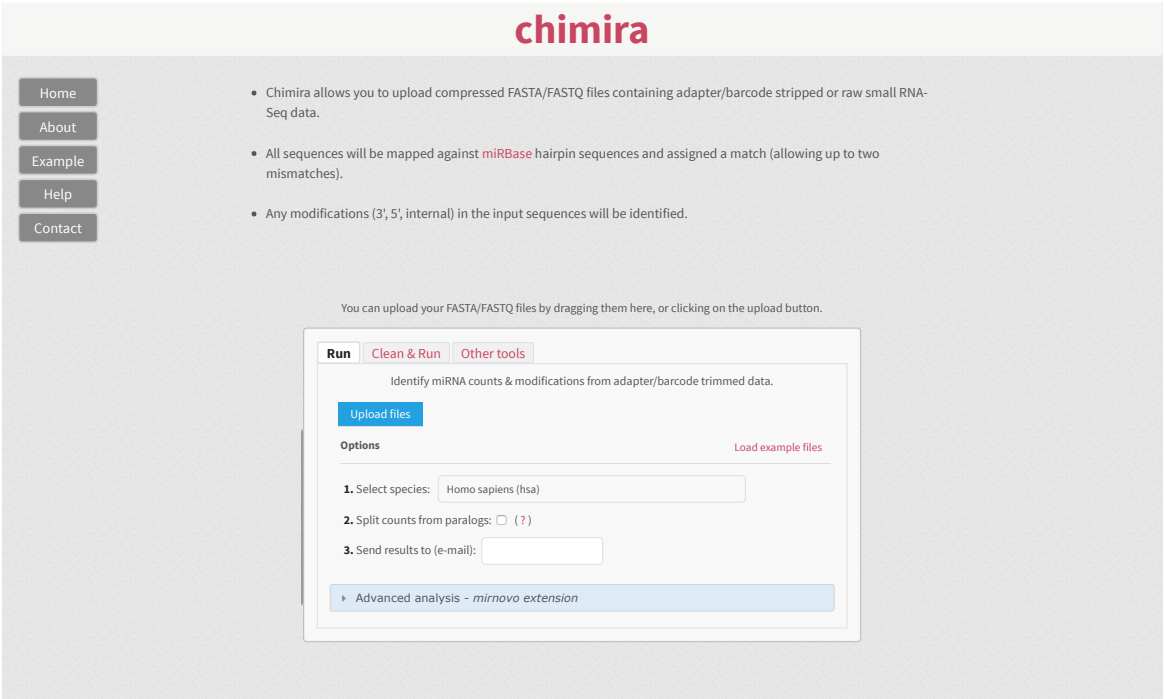



**Fig. A.1** Chimira's homepage. Three modes of operation are provided: *Run*, for adapter-trimmed files, *Clean & Run*, for raw files (non adapter-trimmed), and *Other tools*, for 3' adapter inference.

Chimira offers 3 main modes of operation:

1. **Run** mode: Chimira expects as an input FASTQ or FASTA files containing adapter and/or barcode trimmed small RNA-Seq data. It is important that the input file sequences have been cleaned properly from any adapters/barcodes that were used during sequencing, so that the extracted modification profiles are free from sequencing noise of this type and results are reliable.

2. **Clean & Run** mode: this mode should be used when input files contain 3' adapters (barcodes should have already been removed). Input files may all have the same 3' adapter (in that case a common adapter sequence should be provided). However, different adapters for each file are supported, in which case a file containing the adapter sequences for each file should be provided.

3. **Other tools** mode: the user can upload raw FASTQ/FASRA files and search for the 3' adapter in their sequences. This process is making use of *minion* and *swan* (Davis et al., 2013) and provides in the end a suggested adapter for each of the input samples, along with its alignment score against a database of verified adapters.

Chimira provides two types of miRNAs identification: *Plain Counts* and *Modifications*. *Plain Counts* refers to the quantification of the miRNA molecules that are expressed in any form in each of the input samples. *Modifications*, on the other hand, refers to the quantification of any sequence segments that are part of the input sequences and cannot be justified by the genomic sequence of reference.

The output from each type of analysis contains interactive visualisations based on D3.js that summarise the results along with query-able tables (based on jQuery DataTables) with the raw output data (Figures A.2, A.3).

Apart from the initial mapping of input sequences to miRNAs and the identification of modification patterns and counts, Chimira provides a set of tools for further analysis of the results. More specifically, the web-server version of Chimira provides:

- an interactive interface for the differential expression visualisation between two specific samples or sets of samples,

- the identification of the most highly expressed miRNAs within two samples (or sets of samples) and

- the projection of the modifications profile for specific miRNAs across all samples.

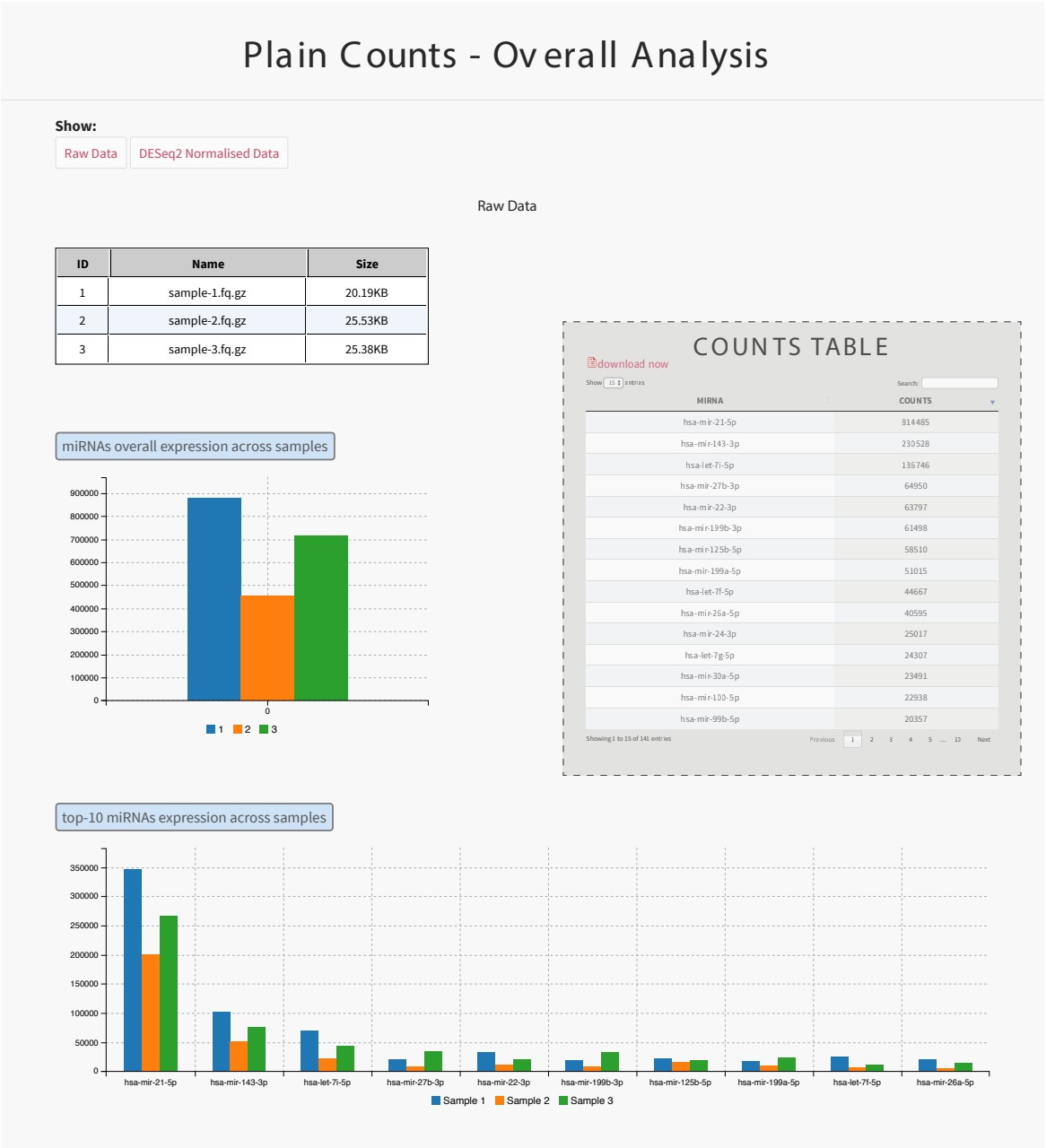
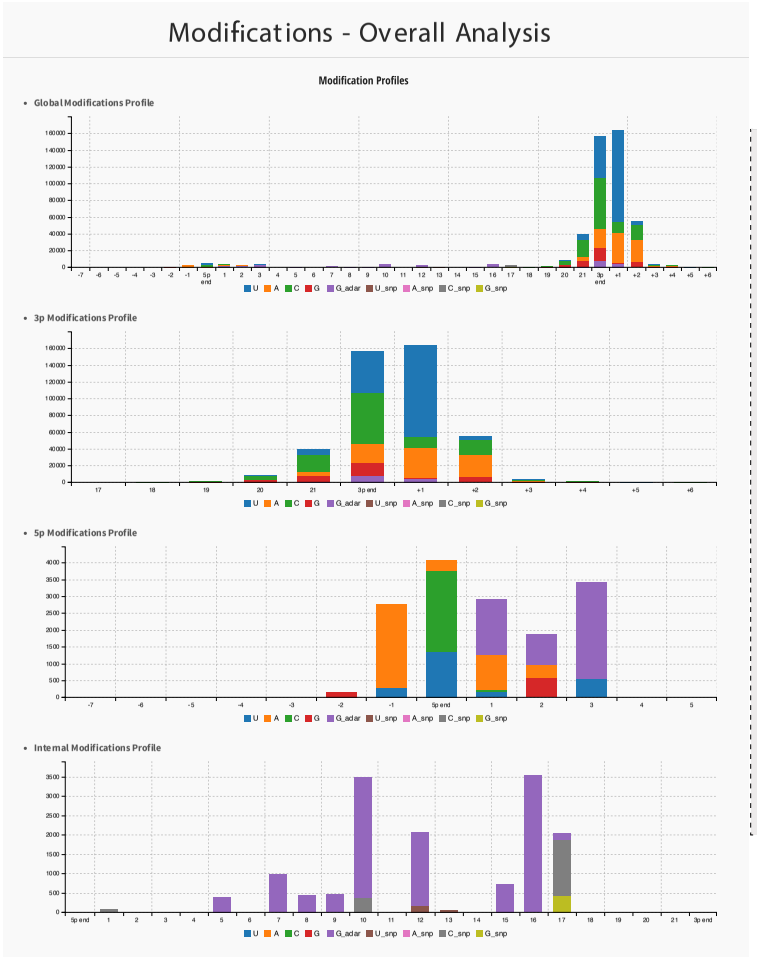



**Fig. A.2** Chimira's snapshot with the overall results from the *Plain Counts* analysis of 3 samples.



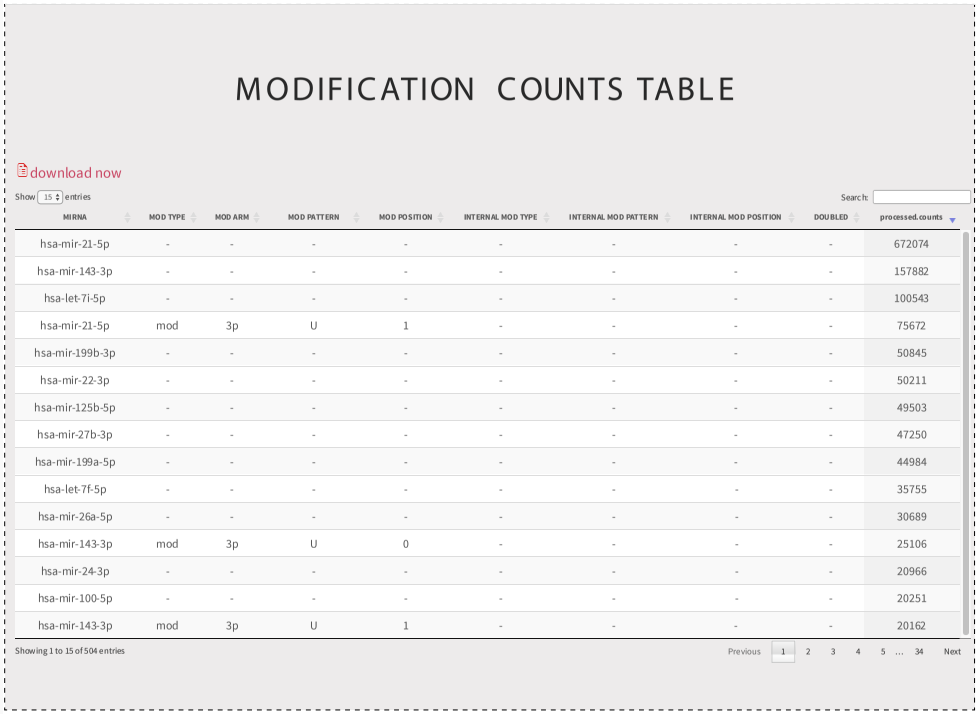




**Fig. A.3** Chimira's snapshot with the overall results from the *Modifications* analysis of 3 samples.

# Appendix B

# *mirnovo*: standalone version tutorial

Source code from the mirnovo standalone package is available on GitHub: https://github.com/dvitsios/mirnovo

## B.1 Installation

1. Download mirnovo_pkg (.tar.gz file) from either Linux or MACOSX folder (depending on the Operating System on your machine).
2. Untar pkg: tar xvzf mirnovo_pkg_[linux | macosx].tar.gz
3. cd mirnovo_pkg_[linux | macosx]

**Dependencies**

- Python (tested with v2.7.10)
- Perl (tested with v5.24.1)
- R (tested with v3.2.2) required libraries: png, ROCR, randomForest
- Unix utilities: wget, gunzip, tar, convert (pre-installed in most distributions).

## B.2 Configuration

***[Important Note]***: mirnovo comes with no pre-installed training models and/or genomes. You need to download at least one training model (and optionally a genome) prior to run.

## B.3 Run

***[Important Note]***: you need to call mirnovo.pl from inside the *bin/* directory.
Output is stored under the *tmp/* directory. A custom output sub-dir name may be defined using the *-o* option.

- **Basic example run:**

```
cd bin
./mirnovo.pl -i ../example_file.tallied.gz -g hsa -t hsa -o example_run
```

- **Example run without a reference genome (-g NA option):**

```
./mirnovo.pl -i ../example_file.tallied.gz -g NA -t hsa -o example_run
```

- **Example run without generating pdf files with coverage profiles and hairpins (–disable-pdf option):**
  (allows for faster execution time, especially for large files)

```
./mirnovo.pl -i ../example_file.tallied.gz -g hsa -t hsa -o example_run --disable-pdf
```

## B.4 Download / Install reference genome

```
cd bin
./download_genome.pl [genome_id]
```

e.g.: *./download_genome.pl dme*

For more info see:
http://wwwdev.ebi.ac.uk/enright-dev/mirnovo-standalone-pkg/Genome-Annotation-1.0
(see README file).

## B.5 Download Training models

```
cd bin
./download_training_model.pl [model_id]
```

e.g.: *./download_training_model.pl universal_animals*

All trained models are available here:
http://wwwdev.ebi.ac.uk/enright-dev/mirnovo-standalone-pkg/Training-Models-1.0

## B.6 Quantification of known and novel miRNAs with Chimira

Mirnovo is able to predict both hairpins and mature miRNAs, providing count data in the latter case.

However, inherent sequence clustering steps (initial and refined) of the mirnovo pipeline may be imperfect in some cases and thus affect, even at a low level, the yielded expression data.

Thus, in order to extract even more accurate expression data we have expanded Chimira, a method that was previously published in our lab (Vitsios and Enright, 2015).

In that case, Chimira serves as a mirnovo extension, allowing the user to upload a custom set of hairpin sequences (e.g. known and/or novel hairpins predicted by mirnovo) and then align their input files against this reference set to get mature miRNA expression counts.

All uploaded reference files are merged and sequences with an alignment identity over 0.90 are collapsed. As an additional functionality, Chimira is able to generate coverage profiles of each identified mature miRNA and the secondary structure of the corresponding hairpin reference hit.

Precomplied binaries are provided with the tool, specifically for the MAC OS X and Linux platforms: *vsearch, muscle, blastn, blastall, fasta_formatter, cdhit, bowtie2, twoBitToFa, twoBitInfo, faToTwoBit, RNAfold, RNAplot.*

# Appendix C

# Repository links with predicted miRNAs by mirnovo

- **Supplementary Data S1**: predicted novel miRNAs in GEUVADIS (Lappalainen et al., 2013).

  *Link*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S1.xlsx

- **Supplementary Data S2-S6**: predicted novel miRNAs in Moth species (with coverage profiles and secondary structures) - (*part 1*).

  *Bombyx mori (Whole body)*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S2.pdf

  *Bombyx mori (Anterior silkgland)*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S3.pdf

  *Bombyx mori (Posterior silkgland)*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S4.pdf

  *Heliconius melpomene melpomene*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S5.pdf

  *Heliconius melpomene rosina*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S6.pdf

- **Supplementary Data S7, S8**: predicted novel miRNAs in Moth species (with coverage profiles and secondary structures) - (*part 2*).

  *Cameraria ohridella*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S7.pdf

  *Pararge aegeria*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S8.pdf

- **Supplementary Data S9**: predicted novel miRNAs in Human, dependent on either Drosha, Dicer or XPO5.

  *Link*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S9.xlsx

- **Supplementary Data S10-S12**: lists of predicted novel miRNAs in Human (with coverage profiles and secondary structures), dependent on either Drosha, Dicer or both.

  *Drosha & Dicer dependent*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S10.pdf

  *Dicer-only dependent*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S11.pdf

  *Drosha-only dependent*:
  https://github.com/dvitsios/mirnovo-predicted-miRNAs/raw/master/Supplementary_Data_S12.pdf

# Appendix D

# List of MREs in *D. melanogaster*, edited by CRISPR/Cas9

**Table D.1** Complete list of miRNA Responsive Elements (MREs) that were edited by CRISPR/Cas9 in *D. melanogaster* S2R+ cell lines. The targeted genes are all coding and non-coding mir-184 targets.

| Sample name | Gene name | Amplicon size | Chromosome | Start index | End index |
|---|---|---|---|---|---|
| A1 | Sinu (MRE1) | 179 | 3L | 5553583 | 5553762 |
| A2 | Sinu (MRE2) | 185 | 3L | 5552688 | 5552872 |
| A3 | Mga2 | 188 | 3R | 25841614 | 25841801 |
| A4 | CG1105 | 186 | 3R | 2890119 | 2890304 |
| A5 | CG1332 | 190 | 3L | 4223355 | 4223544 |
| A6 | CG31195 (MRE1) | 192 | 3R | 16584236 | 16584427 |
| A7 | CG31195 (MRE2) | 188 | 3R | 16583078 | 16583265 |
| A8 | CG8121 | 181 | 3R | 5162712 | 5162892 |
| A9 | Mcr | 180 | 2L | 8074699 | 8074884 |
| A10 | Tsf2 | 180 | 3L | 12523643 | 12523822 |
| A11 | cals | 183 | 4 | 1135886 | 1136068 |
| A12 | emp | 171 | 2R | 20864066 | 20864236 |
| B1 | Iqfr (epsin-like) | 181 | 3R | 18244429 | 18244609 |
| B2 | Ptp99A | 198 | 3R | 25310745 | 25310942 |
| B3 | Sema-1b | 194 | 2R | 13560745 | 13560938 |
| B4 | CG10217 | 174 | 3R | 19586478 | 19586651 |
| B5 | CG14059 | 170 | 3L | 17022973 | 17023142 |
| B6 | CG4313 | 180 | X | 1948946 | 1949125 |
| B7 - (no amplicons) | CG6583 | 197 | 2L | 12171486 | 12171680 |
| B8 | CG7713 | 170 | 3R | 13621764 | 13621933 |
| B9 | Pck | 187 | X | 1365986 | 1366172 |
| B10 | CG1298 | 179 | 2R | 1560073 | 1560251 |
| B11 | CG17218 | 180 | 2L | 12173779 | 12173958 |

Table D.1 Continued from previous page

| Sample name | Gene name | Amplicon size | Chromosome | Start index | End index |
|---|---|---|---|---|---|
| B12 | CG2813 | 182 | 2L | 575216 | 575397 |
| C1 | CG13088 | 182 | 2L | 8490302 | 8490483 |
| C2 | CG1084 | 182 | 3R | 212546 | 212727 |
| C3 | CG6965 | 181 | 3R | 7709684 | 7709864 |
| C4 | CG8010 | 180 | X | 19160736 | 19160915 |
| C5 | CG3446 | 184 | X | 6243435 | 6243618 |
| C6 | CG15154 | 182 | 2L | 18139584 | 18139765 |
| C7 | Atet | 180 | 2L | 4344915 | 4345094 |
| C8 | BetaInt-nu | 193 | 2L | 21057312 | 21057504 |
| C9 | Bves | 182 | X | 20948104 | 20948285 |
| C10 | Cad96Cb | 200 | 3R | 21049878 | 21050077 |
| C11 | Cah2 | 182 | 3L | 12176184 | 12176365 |
| C12 | CG12880 | 198 | 3R | 23517836 | 23518033 |
| D1 | CG14785 | 182 | X | 1349278 | 1349459 |
| D2 | CG31495 | 180 | 3R | 9664108 | 9664287 |
| D3 | CG4542 (MRE1) | 185 | X | 6717054 | 6717238 |
| D4 | CG4542 (MRE2) | 170 | X | 6717199 | 6717368 |
| D5 | CG6038 | 177 | 3L | 11703607 | 11703783 |
| D6 | CG6905 | 176 | 3L | 358050 | 358225 |
| D7 | CG8776 | 176 | 2R | 8547656 | 8547831 |
| D8 | Dok | 199 | X | 7222337 | 7222535 |
| D9 | drd | 176 | X | 15030878 | 15031053 |
| D10 | fy | 220 | 2L | 8401476 | 8401695 |
| D11 | l(2)35Bg | 185 | 2L | 15037140 | 15037324 |
| D12 | l(2)gl | 194 | 2L | 10637 | 10830 |
| E1 | Oseg1 | 189 | 3L | 8400337 | 8400524 |
| E2 | Pka-R2 | 186 | 2R | 5884633 | 5884818 |
| E3 | Ppcs | 170 | 3R | 14966053 | 14966222 |
| E4 | RecQ5 | 177 | 3L | 14623330 | 14623506 |
| E5 | Sbr | 180 | X | 10727188 | 10727367 |
| E6 | SPoCk | 196 | 3L | 22779195 | 22779390 |
| E7 | ttk | 180 | 3R | 27560302 | 27560481 |
| E8 | yrt | 192 | 3R | 9255295 | 9255486 |
| E9 | Gli | 200 | 2L | 15756083 | 15756282 |
| E10 | Nrx-IV | 180 | 3L | 12149551 | 12149730 |
| E11 | Lac | 180 | 2R | 8352674 | 8352853 |
| E12 | CG12789 | 200 | 2L | 7445489 | 7445688 |
| F1 | XLOC_035581 | 180 | 4 | 668555 | 668734 |
| F2 | XLOC_003055 | 199 | 2L | 21799331 | 21799529 |
| F3 | XLOC_001425 | 182 | 2L | 9973580 | 9973761 |
| F4 | XLOC_030523 | 173 | 3R | 886117 | 886289 |

Table D.1 Continued from previous page

| Sample name | Gene name | Amplicon size | Chromosome | Start index | End index |
|---|---|---|---|---|---|
| F5 | XLOC_030181 | 187 | 3R | 26541912 | 26542098 |
| F6 | XLOC_030521 (MRE1) | 192 | 3R | 882972 | 883163 |
| F7 | XLOC_030521 (MRE2) | 192 | 3R | 882972 | 883163 |
| F8 | XLOC_000338 | 180 | 2L | 1989625 | 1989804 |
| F9 | XLOC_022242 | 198 | 3L | 8995929 | 8996126 |
| F10 | XLOC_022396 | 200 | 3L | 10405100 | 10405299 |
| F11 | XLOC_023568 | 180 | 3L | 19415111 | 19415290 |
| F12 | XLOC_031931 | 180 | 3R | 11379049 | 11379228 |
| G1 | XLOC_019797 | 180 | 3L | 19548116 | 19548295 |
| G2 | XLOC_005141 | 180 | 2L | 11958578 | 11958757 |
| G3 | XLOC_001669 (del1) | 234 | 2L | 12024216 | 12024449 |
| G4 | XLOC_001669 (del2) | 234 | 2L | 12024216 | 12024449 |
| G5 | XLOC_001669 (MRE5) | 304 | 2L | 12024352 | 12024653 |
| G6 | CR44786-RA | 191 | 2L | 21821339 | 21821529 |
| G7 | Ctr1B | 192 | 3R | 4144590 | 4144781 |
| G8 | CG9422 | 192 | 2R | 2580385 | 2580576 |
| G9 | CG5850 | 187 | 2L | 9960878 | 9961064 |
| G10 | comm | 185 | 3L | 15715296 | 15715480 |
| G11 | CG11594 | 196 | 3L | 4025155 | 4025350 |
| G12 - (no amplicons) | Atg4 | No | primer | designed | |
| H1 | CR44033 | 182 | 3R | 17360424 | 17360605 |
| H2 | Kaz-m1 | 181 | 3R | 21841477 | 21841657 |
| H3 | CG4705 | 220 | 2L | 11091442 | 11091661 |
| H4 | MRE23 | 181 | 3R | 24414696 | 24414876 |
| H5 | Dp1 | 100 | 2R | 14302334 | 14302433 |
| H6 | Ubc12 | 196 | 3L | 8190762 | 8190957 |