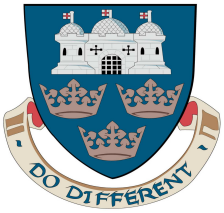
miRNA detection and analysis from high-throughput small RNA sequencing data



Claudia Paicu

School of Computing Sciences

University of East Anglia

A thesis submitted for the degree of

Doctor of Philosophy

December 2016

© This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there from must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

I would like to dedicate this thesis to my loving husband Andy, who gave me strength in the hard moments and reasons to smile at the end of the day. Thank you!

Acknowledgements

Firstly I would like to thank my supervisor, Vincent Moulton, and my co-supervisors Simon Moxon and Tamas Dalmay for the valuable advice and support that they gave me during my PhD.

I would also like to thank Irina Mohorianu and Matthew Stocks for their help and for sharing their knowledge with me.

I would like to thank all the biologists from Dr. Tamas Dalmay's laboratory, and especially Ping Xu, Aurore Coince, Martina Billmeier, Christopher Dacosta and Adam Hall, for providing sequencing data and for the opportunity to collaborate on their projects.

I would like to acknowledge Liviu Ciortuz for the encouragement and confidence he gave me to pursue a PhD in the first place.

I would like to thank the School of Computing Sciences at UEA and the Earlham Instiute for their financial support through the scholarship they provided, and for giving me the opportunity to undertake this work.

I am also grateful for all the help and advice I received from all of my colleagues throughout the years: Andrei, Alex, Awat, Bogdan, Dave, James and Matt.

Finally I would like to thank my husband Andy, my parents, Doinita and Cezar, my best friend, Raluca, my brother, Bogdan, and all of my friends, for their support, encouragement and understanding over the last three and a half years, without which I would not have succeeded.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or other institute of learning.

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged.

Publications

"miRCat2: Accurate prediction of plant and animal microRNAs from next-generation sequencing datasets", C Paicu, I Mohorianu, M Stocks, P Xu, A Coince, M Billmeier, T Dalmay, V Moulton and S Moxon; Bioinformatics (2017), accepted.

"The cytoskeleton adaptor protein ankyrin-1 is upregulated by p53 following DNA damage and alters cell migration", A E Hall, W-T Lu, J D Godfrey, A V Antonov, C Paicu, S Moxon, T Dalmay, A Wilczynska, P A J Muller and M Bushell; Cell Death and Disease (2016) 7, e2184; doi:10.1038/cddis.2016.91.

"Sulforaphane modulates microRNA expression in colorectal cancer cells to potentially implicate the regulation of the CDC25A, HMGA2 and MYC oncogenes", C A Dacosta, C Paicu, I Mohorianu, W Wang, P Xu, T Dalmay, Y Bao; Cancer Research (2017), submitted.

Abstract

Small RNAs (sRNAs) are a broad class of short regulatory non-coding RNAs. microRNAs (miRNAs) are a special class of ∼21-22 nucleotide sRNAs which are derived from a stable hairpin-like secondary structure. miRNAs have critical gene regulatory functions and are involved in many pathways including developmental timing, organogenesis and development in both plants and animals. Next generation sequencing (NGS) technologies, which are often used for identifying miRNAs, are continuously evolving, generating datasets containing millions of sRNAs, which has led to new challenges for the tools used to predict miRNAs from such data. There are several tools for miRNA detection from NGS datasets, which we review in this thesis, identifying a number of potential shortcomings in their algorithms.

In this thesis, we present a novel miRNA prediction algorithm, miRCat2. Our algorithm is more robust to variations in sequencing depth due to the fact that it compares aligned sRNA reads to a random uniform distribution to detect peaks in the input dataset, using a new entropy-based approach. Then it applies filters based on the miRNA biogenesis on the read alignment and on the computed secondary structure.

Results show that miRCat2 has a better specificity-sensitivity tradeoff than similar tools, and its predictions also contains a larger percentage of sequences that are downregulated in mutants in the miRNA biogenesis pathway. This confirms the validity of novel predictions, which may lead to new miRNA annotations, expanding and contributing to the field of sRNA research.

Contents

Nomenclature

A	adenine
aMFE	adjusted Minimum Free Energy
C	cytosine
CLI	Command Line Interface
Compl	complexity
DB	database
DE	differentially expressed
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
FC	fold change
FN	false negatives
FP	false positives
G	guanine
GFF	General Feature Format
GUI	Graphical User Interface
HD	high definition

HMDD	human miRNA-associated disease database
JVM	Java Virtual Machine
kb	kilobase
KLD	Kullback-Leibler divergence
MFE	minimum free energy
MIR	miRNA gene
miRNA	microRNA
mRNA	messenger RNA
MT	minimum total
MTC	median total count
ncRNA	non-coding RNA
NGS	next generation sequencing
NR	non-redundant
nt(s)	nucleotide(s)
piRNA	piwi-interacting RNA
pre-miRNA	miRNA precursor
pri-miRNA	primary miRNA transcript
RAM	random access memory
Red	redundant
RISC	RNA-induced silencing complex
RNA	ribonucleic acid
RNAi	RNA interference

RPM	reads per million
RR	Recall rate
rRNA	ribosomal RNA
RUD	random uniform distribution
SAM	Sequence Alignment/Map format
SFN	sulforaphane
siRNA	small interfering RNA
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
sRNA	small RNA
SVM	Support Vector Machines
T	thymine
TE	transposable elements
TN	true negatives
TP	true positives
tRNA	transfer RNA
U	uracil

Chapter 1

Introduction

Small RNAs are a broad class of short regulatory non-coding RNAs, with crucial roles in cell biology, which have been discovered fairly recently. MicroRNAs are a class of 22 nucleotide small RNAs which are derived from a stable hairpin-like secondary structure. They have important gene regulatory functions, hence they need to be identified and analysed. Existing miRNA prediction tools present various weaknesses, therefore the focus of the research presented in this thesis is the development of miRCat2, a new microRNA prediction algorithm in next generation sequencing data. In addition, a review of the most commonly used miRNA detection tools to date is presented. We give a detailed analysis of the results of miRCat2, presenting computationally verified novel predictions in tomato data. Moreover, we present a method of sRNA data analysis for identifying differentially expressed miRNAs between distinct conditions, work done in collaboration with biologists who have provided both small RNA sequencing data and experimental validation of the results. We now give an overview of the thesis.

Chapter 2. We present background information related to microRNAs, then we focus on their biogenesis and functions in the organism. We continue by describing next generation sequencing technologies, software tools and file formats commonly used by microRNA prediction algorithms, which are frequently referred to throughout this thesis.

Chapter 3. We provide the basics of the algorithms for the most commonly used miRNA prediction tools, with a focus on miRCat [1] and miRDeep2 [2], because they implement some features used by miRCat2 as well. We

then give a review of the performance of these tools, to create a clear image of the existing competition, presenting both their advantages and their issues. Based on this review, we choose the tools that we compare to miRCat2, to assess its performance: miRCat [1], miRDeep2 [2], miRPlant [3] and miReap (http://mireap.source-forge.net/). The results for this comparison are presented in the later chapters.

Chapter 4. We design and implement a new miRNA prediction algorithm, miRCat2, that is suitable for both plant and animal data. The algorithm was integrated into the UEA small RNA Workbench [4] with the help of Dr. Matthew Stocks. We present the new method used by miRCat2 to handle increasing depth of sequencing datasets, by implementing a peak selection algorithm, which provides the miRNA candidates. The peak approach was designed in collaboration with Dr. Irina Mohorianu. We then describe novel filters used on the selected reads, inspired from the miRNA biogenesis features. We go on by describing the secondary structure computation and the discriminative features searched on it. In this chapter we also provide a detailed description of performance assessment methods and novel miRNA verification methods used to test and benchmark our new algorithm.

Chapter 5. We tested miRCat2 on ten plant and animal model organisms and we present detailed results for three organisms from each Kingdom. Then we compare miRCat2 performance with miRCat [1], miRDeep2 [2], miRPlant [3] and miReap (http://mireap.source-forge.net/). To assess the performance of the tools, we have calculated their sensitivity and specificity (with miRBase [5] as reference). To better understand their predictions, we then computed the fold change of the expression levels of their results between wild type and mutants in the miRNA biogenesis pathway. For this experiment, amongst other five model organisms, we also make use of *A. thaliana* wildtype and DCL1 mutant data, which was sequenced by members of Dr. Tamas Dalmay's group (Dr. Ping Xu, Aurore Coince, Martina Billmeier). We then continue by computationally examining the miRCat2 novel predictions in the tomato dataset, on which miRCat2 obtained low specificity, to prove they are true miRNAs.

Chapter 6. We describe and apply a method of small RNA dataset analysis and identification of microRNA differential expression, which provides a wider

view on the area of research on miRNAs, by explaining the use for the annotated miRNAs and why it is important to have accurate novel miRNAs annotated. We first present tests for checking the quality of the constructed libraries, to ensure that the data is biologically accurate. Then we give an overview of normalisation methods and how to choose the most appropriate one, depending on the data. We then perform the differential expression analysis and report the miRNA sequences with changed expression levels. This work was done in collaboration with Dr. Irina Mohorianu, who developed the method and supervised the analysis I conducted, and with biologists who have provided both small RNA sequencing data and experimental validation of the results (Adam E. Hall, Christopher Dacosta and other members of Dr. Tamas Dalmay's group).

Chapter 7. We discuss the work presented in this thesis, summing up the key points of this research. We then specify possible future directions, extensions and improvements to this work.

Chapter 2

Background

2.1 Summary

In this chapter we give an introduction to DNA and RNA, focusing on small RNAs. We then give a detailed description of animal and plant microRNAs, describing their biogenesis, functioning mechanism and roles in biological systems. We continue by shortly presenting high throughput sequencing technologies, which are the bridge between biological and computational data. We then give a brief overview of microRNA features the sequencing data might present, which are essential for miRNA prediction algorithms. Finally, we present helper tools and file formats commonly used by such algorithms.

2.2 DNA and RNA

DNA (deoxyribonucleic acid) is a molecule containing hereditary material, present in all living organisms and many viruses. It is found in the nucleus of the cell and encodes genetic instructions for development and functioning of the cell. The information is organised into units called genes; it is stored using four chemical bases: guanine (G), adenine (A), thymine (T) and cytosine (C), their order in the sequence determining the information encoded [6]. Each base is attached to a sugar and a phosphate, together forming a nucleotide (or nt for short).

DNA is structured as two strands of nucleotides coiled around each other,

forming a 3D structured double helix [7]. To represent direction on a strand of DNA, the terms 5' (five prime) and 3' (three prime) are used, based on a chemical convention (the 5' and 3' carbons on the sugar). The 5' end represents the beginning of the strand, while the 3' end represents the end of the nucleotide sequence.

The two strands of DNA are called Watson and Crick strands. The Watson strand refers to the 5' to 3' top strand (5' $\rightarrow$ 3'), whereas the Crick strand refers to the 3' to 5' bottom strand (3' $\leftarrow$ 5'). The coding strand is defined as the strand of DNA that is sense to a gene of interest. The coding strand is gene dependent and will switch back and forth across a chromosome and it is complementary to the antisense strand. The two strands are bound to each other based on the Watson-Crick base-pairing (A - T, C - G) [7]. DNA can replicate, using one strand as a pattern to create a copy of the genetic material [8].

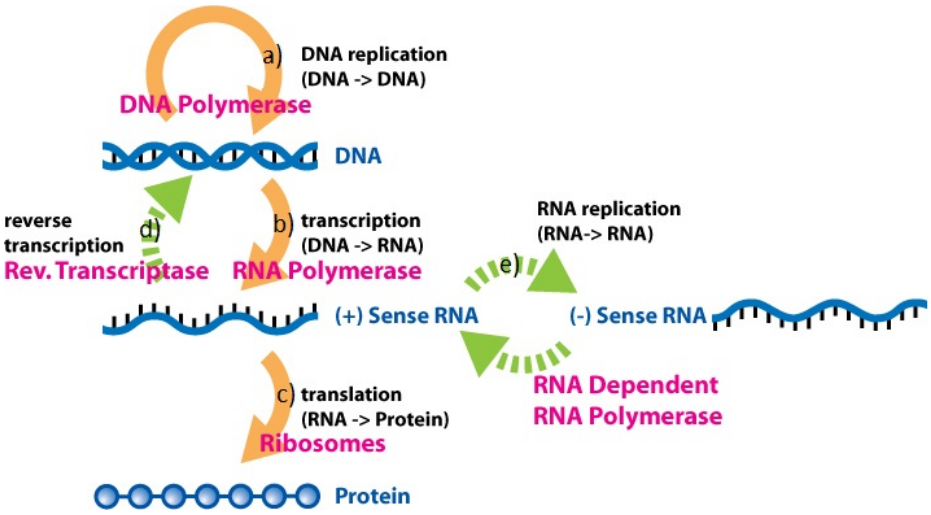


Figure 2.1: **Central dogma of molecular biology [9], presenting a) the replication of DNA, b) the transcription of DNA to RNA, c) the translation of RNA into proteins, d) the reverse transcription of RNA into DNA and e) RNA replication.** This summarises the flow of genetic information within a biological system. Unusual flow of information highlighted in green. (a) DNA is replicated to create a copy of itself. (b) Information is transferred from DNA to RNA through transcription. (c) RNA is transformed into proteins by translation. (d) Information is transferred from RNA to DNA through reverse transcription. (e) The information is copied from one RNA to another.

After DNA replication, the information is transferred inside of the cell nucleus to a similar molecule, RNA (ribonucleic acid). This process is called transcription

and it is part of the central dogma of molecular biology [10], which provides an explanation of the flow of genetic information within a biological system. RNA forms a key part of this dogma. A simplified representation of the central dogma of molecular biology is presented in Figure 2.1. Summarised, the central dogma of molecular biology states that DNA replicates to create a copy of a gene, which is then transcribed into RNA, which is transformed into proteins through translation. Reverse transcription (RNA to DNA) and RNA replication can also occur, but are less common, usually associated with viruses and virus infected cells [10].

Proteins are large complex molecules consisting of one or more long chains of amino acids. They perform most functions in the cell and are required for the structure and function of the cell, tissue and organ. They are involved in the catalysing of metabolic reactions, DNA replication, responding to stimuli, and transporting molecules from one location to another [6].

RNA has a crucial role in various biological processes, by participating in coding, decoding, regulation and gene expression. Like DNA, RNA is also formed as a sequence of nucleotides (guanine (G), adenine (A), uracil (U) and cytosine (C)), but it is more often found as a single-strand, often folded onto itself (into a secondary structure; A binds to U, C binds to G), rather than a paired double-strand. The RNA secondary structure is often stable on its own, which means it cannot easily jump out of the current state and fold into other conformations.

The RNA secondary structures can have various lengths and shapes, consisting of secondary structure motifs, which represent the building blocks through which the most complex three-dimensional RNA structures are constructed [6]. These motifs are presented in Figure 2.2. The motifs are: duplexes, which are regions where two strands are paired; single-stranded regions, representing a portion of nucleotides that are not paired; hairpins, which are structures comprised of a duplex and a loop (a bulge that binds the duplex on one of its ends); bulges, which are regions of unpaired nucleotides inside of a duplex, while all corresponding nucleotides on the opposite strand are paired to the nucleotides next to the bulge; mismatches, which occur when a pair of nucleotides from each strand do not match in a duplex, resulting in a symmetrical bulge; internal loops, which are bulges on both strands inside of a duplex, and can be symmetrical or asymmetrical

(having equal or unequal number of unpaired nucleotides on each strand).

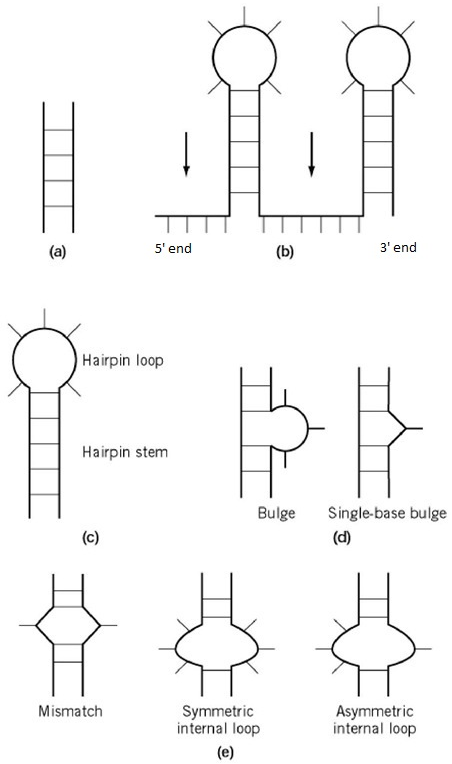


Figure 2.2: **RNA secondary structure motifs.** (a) Duplexes; (b) Single-stranded regions; (c) hairpins; (d) bulges; (e) mismatches and internal loops [11].

For representing direction on a strand of RNA, the same terms are used as for direction on DNA strands: 5' end for the beginning of the strand (left side), and 3' end for the end of the nucleotide sequence (right side) (see Figure 2.2, (b)).

The type of RNA containing the information for the synthesis of proteins is called messenger RNA (mRNA) because it carries the information, from the DNA, out of the nucleus, into the cytoplasm. Each sequence of three bases from the mRNA, called a codon, usually codes for one particular amino acid (which are the building units of proteins). In the cytoplasm, the process of translation is performed. A ribosome reads the information from the mature mRNAs, translating it into amino acids and then a transfer RNA (tRNA) assembles the protein, one amino acid at a time. The assembly continues until the ribosome encounters a "stop" codon (a sequence of three bases that does not code for an amino acid)

[6].

Another type of RNA is non-coding RNA (ncRNA), which is not translated into a protein [12], but is instead a functional molecule. Examples of RNAs belonging to this category include:

- transfer RNA (tRNA) and ribosomal RNA (rRNA) which are involved in the process of translation;
- microRNA (miRNA; 21-22 nt), small interfering RNAs (siRNA; 20-25 nt), piwi-interacting RNAs (piRNA; 29-30 nt) which are involved in gene regulation;
- small nuclear RNAs (snRNA), small nucleolar RNAs (snoRNA; 60-300 nt) involved in RNA processing.

Small RNA (sRNA) is the generic name for a broad class of short regulatory ncRNA. They usually have sequences of 19-28 nt in length and originate from a double-stranded RNA. sRNAs function at RNA level, inducing gene silencing by being loaded into Argonaute proteins (AGO) and targeting molecules through specific base-pairing in a mechanism called RNA interference (RNAi) [13] (see sections 2.3.1 and 2.3.2 for details). The RNAi machinery is conserved in most eukaryotes and mediated by different types of sRNAs: siRNAs, miRNAs and piRNAs. Eukaryotes are organisms consisting of a cell or cells in which the genetic material is DNA in the form of chromosomes contained within a distinct nucleus. Eukaryotes include all living organisms except bacteria, blue-green algae, and other primitive micro-organisms. RNAi is involved in almost all eukaryotic cellular processes, including host immunity and pathogen virulence [14].

2.3 What are microRNAs?

MicroRNAs (miRNAs) are a class of non-coding sRNAs that are derived from a longer, structured primary transcript (precursor) in the shape of a hairpin [15, 16], as illustrated in Figure 2.3. They are found in eukaryotes (e.g. animals, plants, green algae) and some viruses, and recently miRNA-like sRNAs were also discovered in fungi [17–20]. miRNAs function in post-transcriptional silencing of genes [15, 16, 21]. These tiny, ∼22-nt RNAs need to be identified and analysed because of their important cellular functions in gene regulation, where they control many pathways including developmental timing, hematopoiesis (formation of blood cel-

lular components), organogenesis and development, apoptosis (programmed cell death), cell proliferation (cell division) and tumourigenesis [22–27]. Therefore, miRNAs are absolutely essential to the health and development of plants and animals.

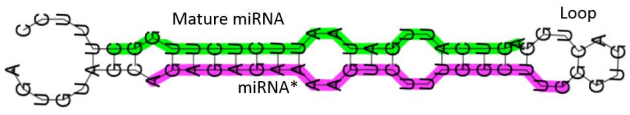


Figure 2.3: miRNA hairpin-like secondary structure.

It is believed that miRNAs have evolved from initial RNAi machinery as a defence mechanism against foreign genetic material inflicted by organisms such as viruses [28]. In time, miRNAs have specialised in the fine-tuning of gene expression, allowing organisms to develop complex traits. It has been shown that the complexity of an organism is directly proportional to the fraction of non coding genes of the genome, in mammals the amount of protein coding genes being only ∼1% [29, 30]. miRNAs continue to evolve, and there are a large amount of species-specific miRNA genes and gene families in a diverse range of organisms, including human and primates, with a relatively low rate of loss of the conserved miRNA families [31].

miRNAs bind to mRNA targets based on the Watson-Crick base-pairing (fully or partially, nucleotides pairing A-U, C-G) [7]. Plant miRNAs have near-perfect complementarity to their targets and function by cleaving them [27, 32] (see section 2.3.2 for details). In animals, only the 6-8 nt long region (miRNA nucleotides 2 to 8), known as the 'seed sequence', at the 5' end of the miRNA, will typically bind to the target, leading to translational repression [32–34] (see section 2.3.1 for details). The different modes of action of miRNAs in the two kingdoms, together with the fact that there is no seeming correspondence between plant and animal miRNA sequences, suggest that miRNAs evolved independently in the plant and animal kingdoms, after their most recent common ancestor (which is thought to have been unicellular), in an example of convergent evolution [28, 29, 31]. Even so, the presence of miRNAs in all plant and animal species suggests early origins in both lineages, facilitating the developmental patterning needed for multicellu-

lar organisms [16].

miRNAs are encoded by endogenous genes (MIR) (originating from within the organism/cell), the majority located in intergenic regions (>1 kilobases (kb) away from annotated/predicted protein coding genes). MIRs are often transcribed in a similar way to protein-coding genes. However, a considerable proportion of MIRs are not independent transcription units. Instead, they are embedded in either intronic or exonic sequences of known genes, both in the sense or antisense orientation (from one DNA strand or its complement) [15, 35]. An intronic region is the nucleotide sequence within a gene that is removed by RNA splicing, whereas an exonic region represents the nucleotide sequence encoded by a gene that remains present within the final mature RNA product of that gene. In addition, a few miRNAs are produced from transposable elements (TE) in Arabidopsis and rice [36]. A TE is a DNA sequence that can change its position within a genome, sometimes creating or reversing mutations (via reverse transcription of DNA) and altering the cell's genome size [37].

Many miRNAs have been found in close proximity to other miRNAs, forming clusters [15, 35] and several of them are perfectly conserved among species (orthologue miRNA genes) [38, 39]. Orthologues are genes in different species that evolved from a common ancestral gene by speciation. Orthologues of miRNAs differ only by a few nts and usually retain the same function in the course of evolution. However, miRNA hairpins differ significantly outside of the miRNA and miRNA* (the complement of the miRNA) regions, as their structure is rather more important than the sequence. For example, miRNA families such as let-7, lin-4, miR-1, miR-34, miR-60, and miR-87, are highly conserved between invertebrates and vertebrates [35, 40–43].

To date, thousands of MIRs have been identified and stored in miRBase [5] (http://www.mirbase.org/). The miRBase database is a searchable database of published miRNA sequences and annotations. Each entry in the miRBase represents a predicted hairpin, portion of a miRNA transcript (precursor), with information on the location and sequence of the mature miRNA sequence.

Animal miRNAs were first discovered in 1993 in nematodes (*Caenorhabditis elegans*), when *lin-4* was identified to belong to a new class of sRNAs with regulatory functions [44]. In 2000, *let-7* was reported in nematodes [45] and

shortly after, the same miRNA was found to have similar function in human [42]. In plants, the first miRNAs were discovered in 2002, when the miRNA families miR156 through miR171 were reported in *Arabidopsis thaliana* [46]. Since then, a lot of research has focused on understanding the miRNA biogenesis and function, as well as on the detection and annotation of new miRNA genes in a variety of animals, plants and viruses.

To detect new miRNAs, we need to understand their features and what differentiates them from other sRNAs (e.g. siRNA, snoRNA, piRNA). Some important features of miRNAs can be extracted by observing the process through which these sRNAs are generated in the cell.

2.3.1 miRNA biogenesis and roles in animals

The biogenesis of miRNAs in animals can be described sequentially (see Figure 2.4):

(a) endogenous miRNA genes (MIRs) are transcribed by the enzyme RNA polymerase II to generate a primary transcript (pri-miRNA) [47, 48]. Alternatively, a host gene can be transcribed, containing the miRNA in its intronic region. pri-miRNA are sometimes several kilobases long and contain one or several local hairpin structures [15];

(b) the first processing step ('cropping') is mediated by the Drosha-DGCR8 complex [49, 50]. First, the DGCR8/Pasha protein assists Drosha in substrate recognition [51, 52], for which both the double stranded structure around the cleavage site and the terminal loop are vital [53]. Next, Drosha cleaves the site located approximately two helical turns (∼22 nt) from the terminal loop [53]. The product of this nuclear processing step is a ∼70-nt pre-miRNA (precursor), which possesses a short stem-loop plus a ∼2-nt 3' overhang [15];

(c) a nuclear export factor (Exportin-5) recognises this structure as a signature motif and exports it into the cell cytoplasm [54–56];

(d) the Dicer protein participates in the second processing step ('dicing') [57–60]. Dicer is a highly conserved protein that is found in almost all eukaryotic organisms, originally found to function in generating siRNAs [57, 58, 60], that are similar in size to miRNAs (21-25 nts). Humans, mice and nematodes

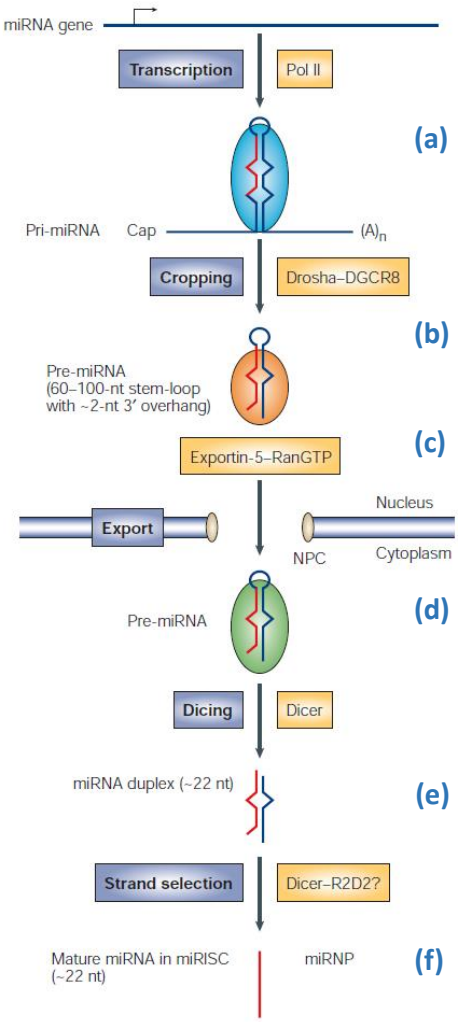


Figure 2.4: **Model for microRNA biogenesis in animals.** Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology [15], copyright 2005.

each possess only one Dicer gene [32, 61], while insects possess two Dicer genes, only one processing miRNAs (the other being involved in RNAi) [62, 63]. The role of Dicer during miRNA biogenesis is to cut the hairpin loop-region and produce ∼22 nts miRNA duplexes [15, 16, 64];

(e) the duplex does not persist in the cell for long and shortly after dicing is separated [15];

(f) usually one strand is selected as the mature miRNA (most often the 5' end), whereas the other strand (miRNA*) is degraded. The relative thermody-

namic stability of the two ends of the duplex determines which strand is to be selected [65], although in some cases both miRNA and miRNA* are stable and functional [64, 66].

After the mature miRNA is produced, it can target multiple transcripts and vice versa (one transcript can be targeted by multiple miRNAs) [32]. miRNA targeting in animals occurs in the following way: after the miRNA/miRNA* duplex separation, Dicer associates with proteins which are part of the Argonaute protein family [67–69], having a central role in RNA silencing processes. Dicer facilitates the transfer of the selected miRNA to AGO, the mature miRNA being incorporated into the RNA-induced silencing complex (RISC), sometimes referred to in the literature as miRISC [70]. Bound by AGO proteins, the miRNA guides the complex to complementary mRNA sequences to repress their expression.

The major determinant for AGO binding to its target mRNA is a 6-8 nt region at the 5' end of the miRNA (miRNA nucleotides 2 to 8), known as the miRNA 'seed' region [33]. AGO associates with this region to create the 'seed'. Functional target sites are usually located in the 3' UTR of a mRNA [33]. When perfect complementarity of the target to the seed region of the miRNA occurs, it is often referred to as 'canonical binding' [34]. In the event of seed region mismatches or bulges, 3' supplementary binding (additional pairing in the miRNA nucleotides 12 to 16) and 3' compensatory binding (extensive complementarity in the miRNA 3' region) can occur and is referred to as 'non-canonical binding' (see Figure 2.5) [71].

Once the miRISC complex is bound to a target, translational inhibition is initiated through two mechanisms: translational repression [72, 73] and then mRNA degradation through decapping and deadenylation [32–34, 74] (see Figure 2.5). Translation repression means that miRISC prevents translation of the target mRNA into a functional protein sequence, while mRNA degradation, refers to the decay of the mRNA molecule, initiated by miRNA targeting [32, 72, 73, 75]. The process of mRNA decapping consists of removing the 5' cap structure on the RNA, which leads to rapid degradation of the molecule [76]. Through deadenylation, the poly(A) tail (stretch of RNA that has only A bases, necessary for mRNA stability) of the mRNAs gradually gets shorter, mRNAs with shorter poly(A) tails being translated less and degraded sooner [77]. Both mechanisms

lead to reduced translation and therefore reduced protein production, although translational repression does not change the mRNA expression levels within the cell [72, 73].

If there is a high complementarity with the whole sequence, and not just a seed match, then the target is cleaved, rather than translationally repressed. However this happens more often in plants and extremely rarely in animals [69, 78].

At a genome-level, animal miRNA targeting is a very complex mechanism and is likely to involve a large network of mutually interacting components. On one hand, the regulation of a target is generally combinatorial, the mRNA expression depending on a combination of multiple miRNAs being involved. On the other hand, a certain miRNA can target various mRNA sequences [32, 79].

Because the region used to create the seed is so short, more than half of all protein-coding genes in mammals are regulated by miRNAs [80]. In human, the expression of >60% of protein-coding genes is controlled by miRNAs [81].

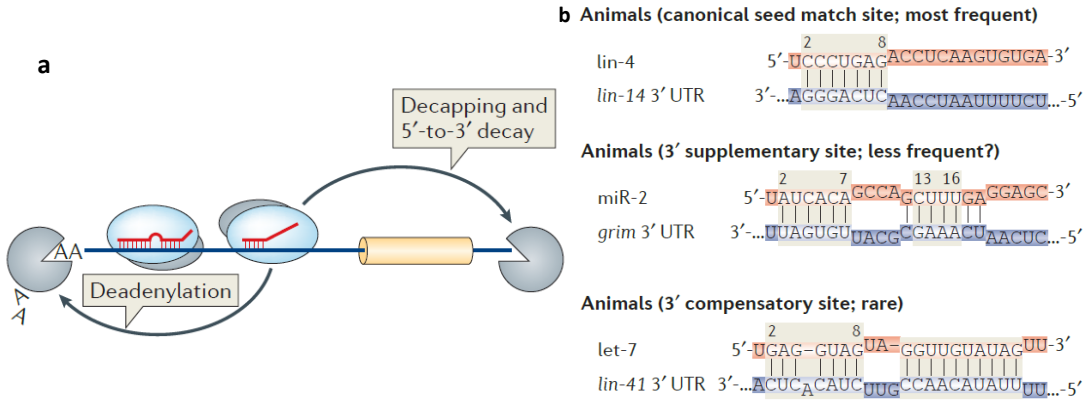


Figure 2.5: **miRNA translationally repress their targets in animals.** a) miRNA-directed translational repression via deadenylation, decapping and 5' to 3' decay. b) The seed sequence is the major determinant for target binding. In case of imperfect seed matches, additional pairing can occur for the miRNA nucleotides 12 to 16 or an extensive complementarity in the miRNA 3 region. Adapted with permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology [80], copyright 2013.

miRNAs have crucial roles in developmental stages, and especially they facilitate early development in a broad range of organisms (eg.: fish [74, 82], insects [83], mammals [84]).

For example, miR-430 was proved to directly regulate ∼160 mRNAs in zebrafish embryos, and it was also estimated to directly regulate several hundred target mRNAs during early zebrafish development [74]. The miR-10 family directly targets Hox genes in *N. tilapia*, which are a family of transcription factors that function during embryogenesis [82]. let-7 and miR-125 function in metamorphic processes in fly, the loss of these miRNAs resulting in temporal delays in wing development and maturation of neuromuscular junctions in adult abdominal muscles [83]. miR-9 and miR-124 were found to be involved in brain development in zebrafish [85], mouse [43], rat and monkey [84].

miRNAs are also involved during adulthood in important processes such as caste determination in honey bees (miR-184) [86]. miR-206 (part of miR-1 family) has roles in ensuring proper organ functioning in *C. elegans* [41], adult mouse and human heart [43], and miR-122 was shown to be specifically expressed in mouse liver [43] and zebrafish [85].

miRNAs are also critical in tumourigenesis and tumour suppression in many tissues, their activity being reported in many types of cancer. Cancer is typically caused by uncontrolled proliferation and the inappropriate survival of damaged cells, which results in tumour formation. Many regulatory factors switch on or off genes that direct cellular proliferation and differentiation, miRNAs being amongst them. In fact, half the annotated human miRNAs are associated with cancer [87].

In a study on human carcinomas, miR-21 was reported to be overexpressed in glioblastoma (brain tumour) [88], miR-17/20/92 was found to be involved in lung and breast cancer, all three members of the miRNA cluster accelerating lymphomagenesis when overexpressed [23, 88]. miR-218-2 is consistently downregulated in colon, stomach, prostate, and pancreas cancers [88]. microRNA-34a is tumour suppressive in brain tumours and glioma stem cells [89]. mir-125b-1, located on chromosome 11, was found to be deleted in a subset of patients with breast, lung, ovarian and cervical cancer [87]. Patients who were diagnosed with a common form of adult leukaemia, often have deletions or downregulation of two clustered miRNA genes, mir-15a and mir-16-1 [24, 90].

miRNAs also conduct other processes. In human, inhibition of nineteen miRNA families, such as miR-95, 124 and 125 caused a decrease in cell growth, while inhibition of miR-21 and miR-24 resulted in a profound increase in cell

growth [22]. Similarly, miRNAs function to increase (miR-1d, 7, 148, etc) or decrease (miR-214 in human, miR-14 in *D. melanogaster*) the level of apoptosis [22, 24].

miRNAs are also responsible with adaptation to stress in almost any tissue, reaction to disease [25] and ageing [91]. For example, miR-195, miR-1 and miR-133 play roles in almost all cardiovascular diseases, while miR-126 is associated with vascular inflammation [25]. let-7 and miR-9 are associated with Alzheimer's disease [92]. There exist databases for a comprehensive list of disease-associated miRNAs and more information on them: human miRNA-associated disease database (HMDD) [1] [25]; miR2Disease [2] [93].

Because of their important functions, identifying miRNAs is crucial. miRNA profiling might aid early stage cancer and disease diagnosis, which is essential in many cases for treatment efficiency. Presently, researchers focus on using miRNA-expression signatures to classify cancers, by defining miRNA markers that predict favourable prognosis [24, 94–97]. The discovery that serum, plasma [94] and saliva [96] contain a large amount of stable miRNAs derived from various tissues and organs, has facilitated the research for non-invasive biomarkers for early tumour detection.

Moreover, by discovering and understanding the miRNA functions, new treatments can be explored for diseases associated with them [98–100].

2.3.2 miRNA biogenesis and roles in plants

The biogenesis and functions of miRNAs were primarily discovered by studying *Arabidopsis thaliana*, a flowering plant [46], although the multitude of other 21 to 24 nt RNAs found in plants sometimes complicated their initial classification. In plants, miRNAs are generated in a similar way to animals, in a stepwise manner [101], with some important differences:

(a) the MIR is transcribed by RNA polymerase II [102] to generate the pri-miRNA, which contains the miRNA hairpin [16, 36].

(b) pri-miRNAs are processed to precursor miRNAs (pre-miRNAs), containing a

[1] http://210.73.221.6/hmdd

[2] http://www.mir2disease.org/

stem-loop structure with 2-nt 3' overhangs at the end of stem by a dicer-like1 enzyme (DCL1) in the nucleus [46, 101, 103–105] . Homologues (descendent from a common ancestral gene) of Drosha and DGCR8/Pasha have not been found in plants, suggesting that the Drosha dependent stepwise processing mode applies only to animal cells, in plants its role being assumed by DCL1 [101, 105].

In plants, four Dicer-like genes have been found in *A. thaliana* [106], each having distinct roles: DCL1 generates miRNAs, DCL2 generates siRNAs associated with virus defence, DCL3 generates siRNAs that guide chromatin modification, and DCL4 generates trans-acting siRNAs that regulate vegetative phase change [104, 105, 107]. Five Dicer genes were discovered in poplar and six in rice [108], suggesting that the number of Dicer-like genes has increased in plants during their evolution. This may reflect the differing threats and defence strategies that plants and mammals use; plants do not have an immune system, therefore they rely on Dicers to defend them against a multitude of viruses and transposons [108].

pri-miRNAs are usually processed by DCL1, however, besides DCL1, its homolog DCL4 has also been shown to generate miRNAs from some pri-miRNAs in *A. thaliana* [109]. In rice, the coordinative action of DCL1 and DCL3 was reported to be required for the production of some 24-nt miRNAs [110]. This result suggests the potential divergence of miRNA biogenesis in different plant species [36, 108].

The structures of pri-miRNAs are essential for recognition by DCL1; the structure should present an imperfectly paired lower stem (∼15 nt below the miRNA/miRNA* duplex) for the initial stem-loop cleavage of pri-miRNAs. The loop is also crucial for efficient processing [111–113].

While in animals the length and structure of the pre-miRNA hairpin is fairly consistent, in plants it is much more variable, the precursors being quite diverse in structure, with variable positioning of the miRNA/miRNA* duplex and reaching lengths from 60 up to 300 nts [114];

(c) after the pre-miRNA is formed, DCL1 excises the miRNA/miRNA* duplex from pre-miRNAs. DCL1 is a nuclear protein, which indicates that mature ∼22-nt miRNAs are generated in the nucleus in plants [103];

(d) the sRNA methyltransferase protein hua enhancer1 (HEN1) adds a methyl group to the 3' end of the miRNA/miRNA* duplex to stabilise them [36, 115].

(e) the duplex is then transported from the nucleus to the cytoplasm with the assistance of HASTY (HST) [116, 117], a homologue of Exportin-5, where it is separated and gives rise to the mature miRNA [21, 36].

In plants, miRNAs mainly function through their effector protein AGO, which cleaves the target RNA and/or inhibits its translation. Similar to animals, the strand of the miRNA/miRNA* duplex with a lower 5'-end thermostability is preferentially loaded into AGO as the mature miRNA [118–120].

Plant miRNAs need a much higher degree of complementarity to recognise their targets, usually across the length of their entire sequence. This leads in most cases to AGO-induced endonucleolytic cleavage of the mRNA target [16] (cleaving a nucleotide chain into two parts at an internal point), followed by mRNA degradation. This process is presented in Figure 2.6. The cleavage site is located at nucleotides 10 and 11 of the miRNA, counted from the miRNA 5' end (see Figure 2.6). Target cleavage is considered the predominant pathway for miRNA-mediated repression of gene expression in plants [16], but translational repression has also been observed to a lesser extent [27, 32].

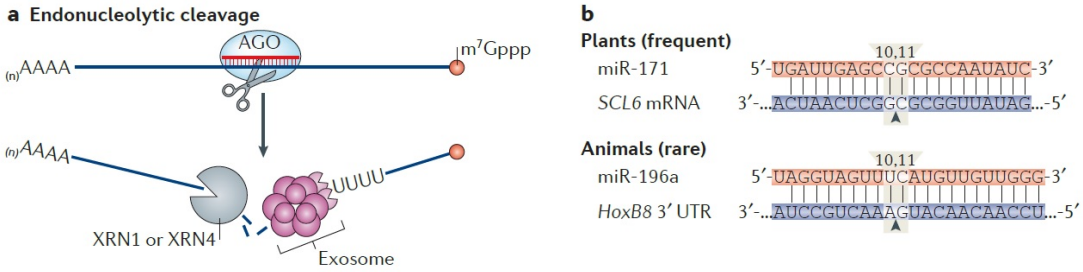


Figure 2.6: **miRNAs lead to deadenylation in plants.** a) miRNAs direct target cleavage (slicing). The XRN4 enzyme in plants, together with the exosome, subsequently degrade the sliced mRNA fragments. b) miRNA-directed cleavage of mRNAs requires extensive complementarity between the miRNA and its target site. The cleavage site is located at nucleotides 10 and 11 of the miRNA, counted from the miRNA 5' end. Adapted with permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology [80], copyright 2013.

miRNAs have roles in many developmental processes including root initiation and organ development (leaf, vain, flower, seed) [27]. For example, miR166 is

involved in leaf development, reduced miR166 levels resulting in abnormal leaf shapes and sizes [27]. miR397 is involved in lignification (process of becoming woody as a result of the deposition of lignin in the cell walls), a process more critical in woody plants, such as *Populus* [121]. miR172 is involved in proper specification of organs during flower development, plants that over-express miR172 having floral defects, such as the absence of petals and sepal transformation into carpels [122]. Over-expression of miR319 results in plants with uneven leaf shape and delayed flowering time [123], while over-expression of miR159a results in male sterility [124].

miRNAs also regulate phase transition. Plants usually undergo the following developmental phases: germination, vegetative growth, reproductive growth and flowering [125]. Two evolutionary highly conserved miRNAs, miR156 and miR172, have been identified as key components of plant phase changing. miR156 promotes the transition from juvenile to adult and to flowering, while miR172 targets mRNAs that encode proteins that have been shown to regulate both the transition to flowering and flower development [125].

The importance of miRNAs during plant development has been proved in an experiment depriving several *A. thaliana* plants of genes central to miRNA function, including DCL1, AGO1, HEN1, and HYL1. Severe mutations resulted in early embryonic arrest, and even partial loss-of-function mutants resulted in many defects, including abnormalities in floral organogenesis, leaf morphology (shape, structure, size), and auxiliary meristem initiation (growing tips of roots and shoots) [126]. This suggests that plants cannot develop into functional adults without proper miRNA regulation.

Additionally, miR172 family miRNAs were reported to be involved in metabolism and sex determination in maize [127].

miRNAs also are responsible for diverse responses to stress: biotic - viruses or bacteria [26, 128] and abiotic - drought, salt, cold, oxidative, nutrient deficiency [36, 129].

miRNAs have crucial roles in adaptive responses to abiotic stress. miR168, miR171, and miR396 were found to be responsive to high salinity, drought, and cold stress in *Arabidopsis* [128]. miR393 was upregulated by cold, dehydration and salinity treatments, while miR389a was downregulated by all of the stress

treatments. miR395 was increased upon sulphate starvation, showing that miRNAs can be induced by environmental factors and not only by developmental processes [128]. The expression of miR397 and miR169 were upregulated under cold stress in *Arabidopsis*, *Populus*, and *Brachypodium*, while miR172 is significantly downregulated in wheat in response to heat stress [128]. Expression levels of miR156g, miR157d, miR172a,b, etc. increased under low-oxygen stress [128]. 21 miRNAs belonging to 11 miRNA families have been identified to be upregulated under UV-B stress [130]. Plant can experience mechanical stress, such as when branches or stems are bent by wind or gravity. Testing mechanical stress, miR156, miR162, miR164, etc. were downregulated but miR408 was upregulated by tension and compression [131].

Plants have shown change in miRNA expression levels under biotic stress, i.e. when infected by pathogenic bacteria, viruses, nematodes and fungi. In *Arabidopsis*, the first miRNA discovered to play a role in defence against pathogens was miR393, a miRNA which induced resistance against bacteria [132]. In a study about the endemic rust fungus *Cronartium quercuum* when infecting loblolly pine, twenty-six miRNAs were identified to take part in the defence against it. Infection with this fungus causes fusiform rust disease, which is characterised by stem and/or branch galls. Results show that miRNAs produced around the fungal infection at the gall immunises the uninfected stem and may provide protection ahead of the spreading infection [133]. bra-miR158 and bra-miR1885 were greatly upregulated when *Brassica rapa* was infected by Turnip mosaic virus [26, 134].

Because of their important roles in plants, profiling miRNAs and understanding their functions could help researchers develop better crop strains and improve food quality. Pathogens can have a big impact on crop production, they spread quickly and are difficult to treat once a plant is infected [135]. miRNAs are essential in developing new crop strains with pathogen-resistance. Also, by studying miRNAs we can improve plant resistance in hard conditions, such as draught and soil nutrient deficiency and optimise the quantity and quality of the food produced.

2.3.3 Mirtrons

Mirtrons are a subtype of miRNAs, derived from short introns of the mRNA encoding host genes. Although many miRNAs are also located in introns, miRNAs are differentiated by the fact that they are Drosha dependent and they are derived from longer introns. Just like regular miRNAs, mirtrons need to be identified and analysed, because they share the same roles in the biological processes.

In animals, mirtrons arise from the short introns, where the miRNA/miRNA* sequences are at the splice junctions. Mirtrons are an alternative way to Drosha-Dicer miRNA biogenesis: the spliced debranched introns with hairpin structures equivalent to pre-miRNAs enter the miRNA processing pathway to produce mature miRNAs, avoiding Drosha-mediated pri-miRNA cleavage [136].

Mirtrons also appear in plants. All the miRNAs in plants are derived from the sequential DCL1 cleavages from pri-miRNA to give pre-miRNA (or miRNA precursor), but the mirtrons bypass the DCL1 cleavage and enter as pre-miRNA in the miRNA maturation pathway [137].

2.4 Detecting miRNAs from high throughput sequencing data

2.4.1 High throughput sequencing technologies

High-throughput sequencing (HTS), also known as next generation sequencing (NGS) [138] is the technology that enables DNA and RNA data collection. HTS captures millions of DNA and RNA fragments and outputs them as sequences in a digital format, easy for processing (sequencing libraries). Over the last few years new technologies in this field have rapidly evolved, including the development of robust protocols for generating these sequencing libraries and building effective new approaches for data-analysis. HTS has dramatically accelerated biological and biomedical research, by enabling the comprehensive analysis of genomes, transcriptomes and interactomes to become inexpensive, routine and widespread.

The first modern sequencing methods were developed in 1977, when Maxam

and Gilbert developed a chemical method [139] and Sanger, Nicklen and Coulson developed the dideoxy method [140], which enabled the first sequencing of a complete DNA molecule [141]. Most HTS technologies still rely on the Sanger biochemistry [140, 141], having shown continuous growth in DNA sequencing capacity and speed. This exponential growth is reflected in the growth of GenBank, the nucleotide sequence database [142], an annotated collection of publicly available DNA sequences (see Figure 2.7).

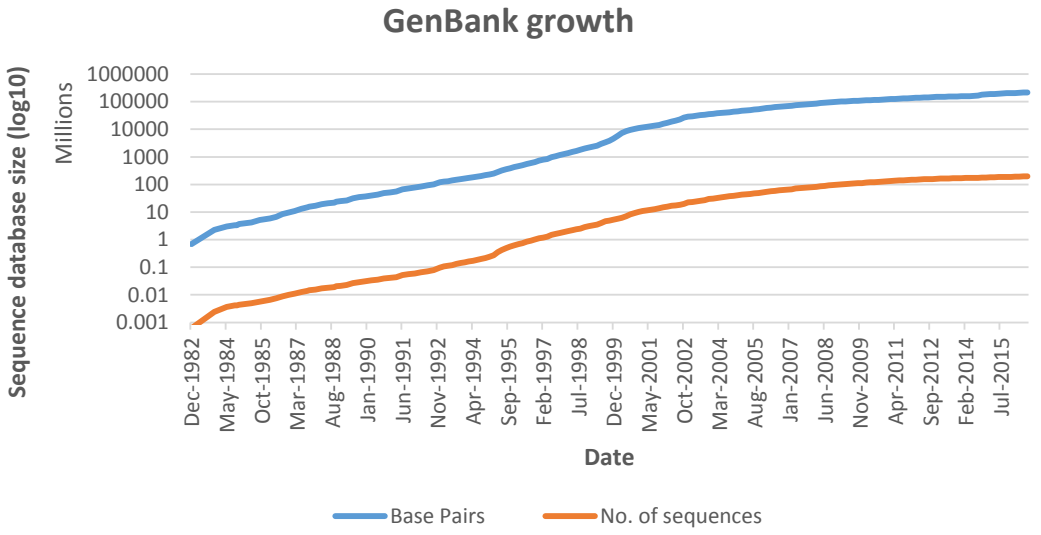


Figure 2.7: Growth of the nucleotide sequence database since 1981, data taken from ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt. The number of published nucleotide sequences and the total number of base pairs of sequence (log10 scale) are plotted versus the date of publication.

There are three main HTS platforms that offer massively parallel DNA sequencing and are widely used at present [143]: the Illumina platform [1] [144], PacBio RS II [145] and the Nanopore MinION Sequencing [146].

The Illumina platform [144, 147] routinely generates sequences of 51 bps (and up to 250 bps), the read-lengths being limited by multiple factors that cause signal decay and dephasing. The dominant error type is substitution, average raw error-rates being on the order of 11.5%, with higher accuracy bases having error rates of 0.1% or less [148].

[1] http://www.illumina.com/pages.ilmn?ID=203

The PacBio sequencing platform [145] performs real-time sequencing and offers longer read lengths than previous sequencing technologies (over 10 kb), making it well-suited for unsolved *de novo* genome assemblies, transcriptome, and epigenetics research. However, it has a much lower throughput than the Illumina platform [145].

The MinIon platform [146] identifies DNA bases by measuring the changes in electrical conductivity generated as DNA strands pass through a biological pore. It is portable and suitable for real-time applications, offering read lengths up to a few hundred thousand base pairs. However, it has higher error rates, its accuracy ranging 65%88%. Another drawback is that its throughput flowcell run is not very stable at the moment, ranging from below 0.1 GB to 1 GB of raw sequence data .

The HTS technology most commonly used at present is Illumina, which is the most advantageous for generating sRNA libraries, as it offers a good trade-off between precision and cost efficiency [148].

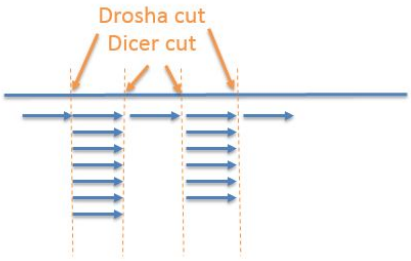
2.4.2 miRNA prediction from HTS data

The rapid development of HTS technology is posing challenges for bioinformatics in areas including data storage, increased memory for processing, sequence quality scoring, alignment, assembly and data release. HTS data can be used to detect miRNAs and their precursors, by providing millions of sRNA reads from only one biological sample.

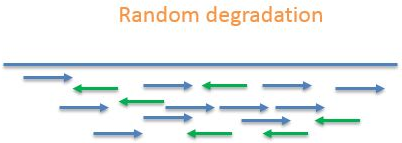
miRNA prediction from HTS data is not a new field of research. Several algorithms have been published since the HTS technology was developed. Early miRNA prediction methods, such as miRCat [1] and miRDeep [149], were designed when sequencing depth was low. Initial algorithms were run on tens of thousands of sequences, whereas nowadays, as HTS datasets are rapidly growing, they have to deal with tens of millions of reads [150, 151]. The large datasets have led to new challenges for the tools used to analyse such data, which struggle with the 'noise' in the datasets, lowing their accuracy, and also in terms of execution time and memory requirements. Higher depth of sequencing leads to more noise: trying to capture more sequences, technologies have an increased chance

of picking up shorter reads, that are not relevant, artificially created reads, that are not actually present in the biological sample, or they artificially increase the expression levels of lowly expressed reads.

After the datasets are collected, the sRNA samples can be processed to predict miRNAs. Properties of miRNAs, observed from their biogenesis, can be identified in the sRNA samples.



(a) miRNA reads distribution, depicting clear Drosha and Dicer cleavage around the mature and star sequences.

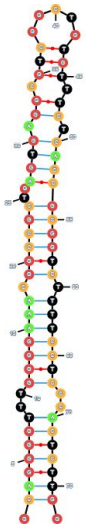


(b) Alignment of reads corresponding to a random degradation.

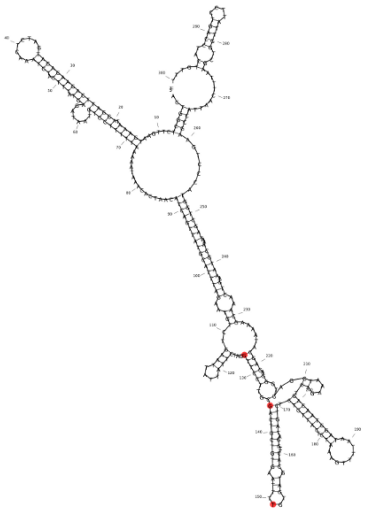
Figure 2.8: **Model for alignment of reads representing (a) a miRNA reads distribution and (b) a random degradation.** Different colours and direction of the arrows represent strand origin (mapping to sense or anti-sense).

First of all, from the miRNA biogenesis we can conclude that the alignments of small RNAs to the hairpin should be consistent with Dicer/Drosha processing, as seen in Figure 2.8. The miRNA processing machinery is very precise, creating patterns that can be used in identifying the miRNAs. The reads corresponding to the miRNA and miRNA* location should be more abundant than the nearby sequences, should have a clear cut on both sides (overlapping sequences having start and end position very close to each other) and should originate from the same strand. In the opposite case of a random degradation, reads can be derived from both strands and show a more dispersed alignment to the genome, overlapping inconsistently and having a uniform distribution.

Secondly, the miRNA precursors should have a stable hairpin-like structure, without many gaps or additional loops on the structured stem region, as seen in Figure 2.9.



(a) A valid miRNA precursor (hsa-mir-2110).



(b) Secondary structure that does not present pre-miRNA-like features.

Figure 2.9: **Examples of secondary structures depicting (a) a valid miRNA precursor in shape of a hairpin (hsa-mir-2110) and (b) a secondary structure that does not present hairpin-like features (a 300bp region of intron 1 of the FTO gene, http://tesla.pcbi.upenn.edu/savor/).**

2.4.3 Tools used by miRNA detection algorithms

Alignment tools

Most miRNA detection algorithms use helper alignment tools to get the alignment of sequences to their reference genome. Some of the most commonly used ones (that we also use) are:

(a) Bowtie 2 [152] is a command-line tool that takes a collection of FASTA files (see Section 6.3.2) for a reference genome and creates a series of index files; the indexing technique used in Bowtie is the key to its speed and memory efficiency. Once it creates the index, it can be queried any number of times. These files are then used to align short reads to the reference genome. Bowtie 2 searches for the best alignment of each read to the reference genome and outputs the results in SAM or BAM format (see Section 6.3.2);

(b) PatMaN [153] identifies all occurrences for a short sequence within a genome-sized database, constructing a single keyword tree of all the query sequences. Once the tree is constructed, each sequence in the target database is evalu-

ated base by base and compared to a list of partial matches.

Folding algorithms

To obtain the precursor secondary structure and test for its characteristic features, miRNA detection tools resort to folding algorithms. The most efficient and accurate are the ones in the ViennaRNA package [154].

All tools in the ViennaRNA package output the folded structure in the dot-bracket notation and its minimum free energy (MFE). In general, the free energy can be thought of as the energy released by folding a completely unfolded RNA molecule. Conversely, it can be thought of as the amount of energy that must be added to unfold a folded RNA. The minimum free energy structure of a sequence is the secondary structure that is calculated to have the lowest possible value of free energy that can be formed with that particular sequence of nucleotides. In the ViennaRNA package, it is calculated using dynamic programming [154].

The ViennaRNA package contains many programs, but for miRNA detection the most frequently used are:

(a) RNAfold - calculates minimum free energy (MFE) secondary structures of RNAs;

(b) RNALfold - calculates all locally stable secondary structures of a long RNA sequence with a maximal base pair span. It is a practical way of "scanning" very large genomes for short RNA structures;

(c) RNAcofold - calculates secondary structures of two RNAs. It allows one to specify two RNA sequences which are then folded to form a dimer structure (two similar sequences linked together).

RANDFold

RANDFold [155] is another useful tool that many algorithms have incorporated in their routine. RANDFold has proved that the majority of the microRNA sequences clearly exhibit a folding free energy that is considerably lower than that for shuffled sequences, indicating a high tendency in the sequence towards a stable secondary structure [155].

RANDFold takes a sequence, shuffles it and refolds it many times, then compares the MFE values of the original secondary structure with the values obtained

by the random shuffling of the original sequences. It then computes a p-value which gives some statistical confidence for whether or not the structure comes from a real miRNA precursor (a p-value closer to 0 indicates that the precursor has pre-miRNA properties).

Annotation databases

(a) miRBase [1] [5] - first established in 2002, miRBase is now the central online repository for miRNA nomenclature, sequence data and annotation. The database has the following main functions [156, 157]: (i) provides a consistent nomenclature scheme, assigning names to novel miRNA genes prior to their publication; (ii) acts as a repository for all published miRNA sequence data, annotation, references and links to other resources (see Figure 2.10). It also facilitates online searching and bulk download of all miRNA data; (iii) provides human-readable and computer-parsable annotation of miRNA sequences; (iv) provides a link to miRNA target predictions and validations [156, 157].

miRBase has continually grown since its inception, encouraging users to submit their results and edit new pages in the database. miRBase has grown from 15,172 loci in 142 species (release 16, October 2010) to 24,521 loci in 206 species (release 20, June 2013) [5]. Therefore, maintaining the quality of the miRNA sequence dataset is a significant challenge.

miRBase mapped reads from multiple public sRNA deep-sequencing experiments (downloaded from databases Gene Expression Omnibus [158] and Short Read Archive [159]) to miRNAs in miRBase and developed a web interface to view these mappings (see Figure 2.10). The user can view all read data associated with a given miRNA annotation, filter reads by experiment and count, and search for miRNAs by tissue-specific and stage-specific expression [157]. This was used to discriminate between true miRNAs and other RNA, creating a set of high confidence miRNAs.

To be annotated as high confidence, a locus must meet a set of criteria, such as: (i) at least 10 reads must map with no mismatches both to the miRNA and miRNA* sequence; (ii) the most abundant reads from each

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

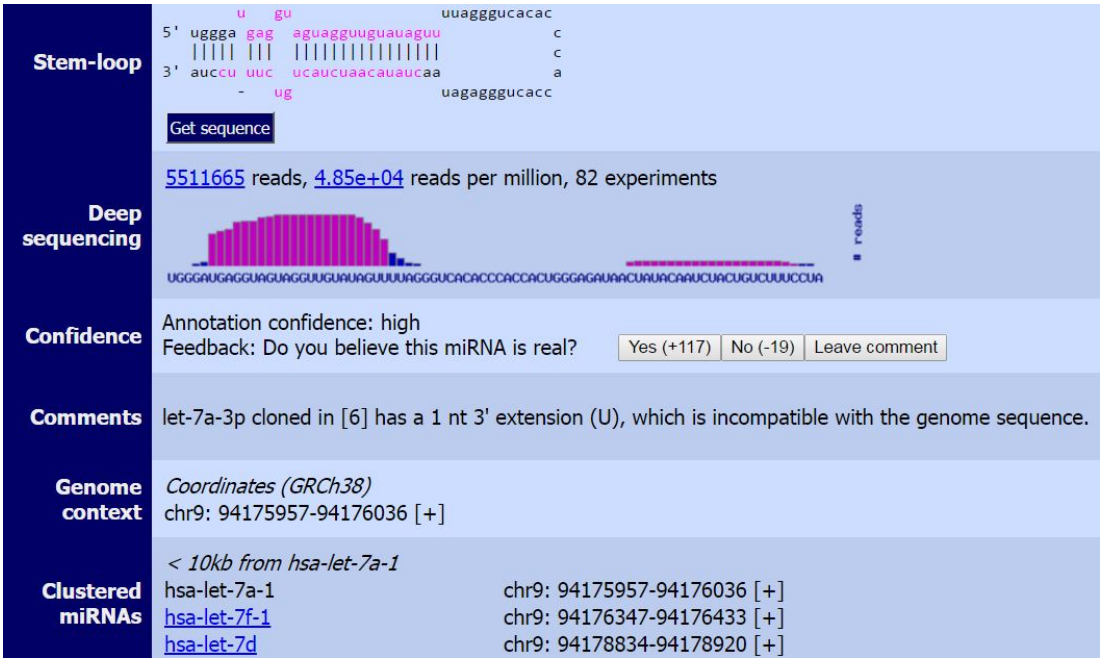


Figure 2.10: **Example of information displayed by miRBase for a selected miRNA.** miRBase entry for *Homo sapiens* let-7a-1 stem-loop, showing information about hairpin sequence and structure, deep sequencing alignment, genome context and clustered miRNAs.

arm of the precursor must pair in the mature microRNA duplex; (iii) the hairpin structure must have a folding free energy of <-0.2 kcal/mol/nt [5]. By applying these criteria, miRBase has created a set of high-confidence miRNAs, representing 22% of the miRNAs in 38 investigated species. In human, less than 20% passed all the criteria [5].

The user can browse or download data after filtering based on high/low confidence (see Figure 2.11). miRBase is commonly used as a reference dataset when predicting novel miRNAs, tools assessing their performance based on the number of predictions from miRBase that they detect or miss.

(b) RFAM [1] [160] - is a collection of ncRNA families represented by manually curated sequence alignments, consensus secondary structures and annotation gathered from corresponding Wikipedia, taxonomy and ontology resources [160]. The primary aim of RFAM is to annotate new members of known RNA families on nucleotide sequences, particularly complete genomes (including

[1] http://rfam.xfam.org/

Homo sapiens miRNAs (1881 sequences) [GRCh38]

ID	Accession	RPM	Chromosome	Start	End	Strand	Confidence	Fetch
hsa-let-7a-1	MI0000060	4.85e+04	chr9	94175957	94176036	+	✓	☑
hsa-let-7a-2	MI0000061	3.92e+04	chr11	122146522	122146593	-	✓	☐
hsa-let-7a-3	MI0000062	3.89e+04	chr22	46112749	46112822	+	✓	☐
hsa-let-7b	MI0000063	2.65e+04	chr22	46113686	46113768	+	✓	☐
hsa-let-7c	MI0000064	3.3e+04	chr21	16539828	16539911	+	-	☐

Figure 2.11: **Browsing miRNA annotations in miRBase.** The user can filter the entries based on weather they are high confidence annotations.

miRNAs, tRNA, rRNA, siRNA, snoRNA, lncRNA and other ncRNA) [161], the current release, RFAM 12.0, containing 2,450 entries [160]. For each RNA family, RFAM provides sequences, alignments, covariance model, trees and secondary structure images. RFAM can be used to identify new family members in other sequence databases and for annotating ncRNAs in genomes or metagenomes [160].

Regarding miRNA prediction, RFAM can be used as a negative control dataset in performance assessment experiments. Users can download a dataset of small ncRNAs (excluding miRNAs), then check that no such sequences are identified as miRNAs. Some miRNA prediction tools (miRDeep2 [2], miRAuto [162]) use RFAM to filter out tRNA, rRNA, snRNA and snoRNA sequences from the input dataset as a preprocessing step before identifying miRNAs.

2.4.4 Commonly used file formats

There are a series of standard file formats that are generally used in sRNA data processing, explained below:

(a) FASTQ - is the standard format for storing the output of high-throughput sequencing instruments such as Illumina [163]. It is a text-based format for storing both the biological sequence (nucleotides) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity [164].

A FASTQ file normally uses four sections per sequence (see Figure 2.12).

Section 1 begins with a '@' character and is followed by a sequence identifier and an optional description, section 2 contains the raw sequence (1 or multiple lines), section 3 begins with a '+' character and is optionally followed by the same sequence identifier, and section 4 encodes the quality values for the sequence in section 2, and must contain the same number of symbols as letters in the sequence (1 or multiple lines).

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Figure 2.12: **Example of entry in a FASTQ file.**

(b) FASTA - is a text-based format for representing nucleotide or peptide sequences. The FASTA format consists of 2 sections for each entry. The first section is a description line, which must begin with the greater-than (>) symbol in the first column, containing a code and optionally a description. The second section contains the sequence (see Figure 2.13) (can span over multiple lines). It is recommended that all lines of text be shorter than 80 characters in length, to fit on terminal windows. The format originates from the FASTA software package [165], but has now become a standard in the field of bioinformatics. The simplicity of the format facilitates easy processing of the sequences in any programming language.

```
>hsa-let-7a-1 MI0000060
UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUG
UCUUUCCUA
```

Figure 2.13: **Entry for human miRNA precursor hsa-let-7a-1 in fasta format.**

(c) GFF - stands for General Feature Format and consists of one line per feature, each containing 9 columns of data, plus optional track definition lines, columns being tab-delimited (see Figure 2.14). Also, all but the final field in each feature line must contain a value;"empty" columns should be denoted

with a '.'. The features are: seqname (chromosome or scaffold); source (name of the program that generated this feature); feature (e.g. Gene, Variation, Similarity); start position of the feature; end position; score (a floating point value); strand (+ (forward) or - (reverse)); frame ('0', '1' or '2'); attribute (list of tag-value pairs, providing additional information). Many annotation databases, including miRBase, keep their information in GFF format.

```
chr1  .  miRNA_primary_transcript  17369  17436  .  -  .  ID=MI0022705;Alias=MI0022705;Name=hsa-mir-6859-1
chr1  .  miRNA  17409  17431  .  -  .  ID=MIMAT0027618;Alias=MIMAT0027618;Name=hsa-miR-6859-5p;Derives_from=MI0022705
chr1  .  miRNA  17369  17391  .  -  .  ID=MIMAT0027619;Alias=MIMAT0027619;Name=hsa-miR-6859-3p;Derives_from=MI0022705
```

Figure 2.14: **miRBase entry for human miRNA precursor hsa-mir-6859-1 and its mature sequences in GFF format.**

(d) PATMAN - is the format of the output file generated by the sequence alignment tool PatMaN [153]. The file uses one line for each entry, containing tab-separated fields that represent the target and query sequence identifier, the start and end position of the alignment in the target sequence, the strand and the number of edits per match (see Figure 2.15).

```
chr1    AGAACTCAAGAAGGTGGACTTCA(17)    3042706  3042728  +    0
chr1    ATGCAAATCAAAACAACCCTGAG(31)    3042851  3042873  -    2
```

Figure 2.15: **Example of alignment output in PatMaN format.**

(e) SAM/BAM - SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section (see Figure 2.16). If present, the header must be prior to the alignments. Header lines start with '@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information [166].
BAM is the compressed binary version of the SAM format. SAM/BAM file formats are used by many alignment algorithms, including Bowtie 2.

(f) BED - is a format used to store annotated data, and has three required fields and nine additional optional fields. The number of fields per line must

Figure 2.16: **Example of alignment output in SAM/BAM format.**

be consistent throughout any single set of data. The first three required BED fields are: chromosome, start and end of the sequence on that chromosome. This file format is used by BEDTools [167], a tool kit for genomics analysis, that enable genome arithmetic operations (intersect, merge, count, complement, and shuffle on genomic intervals from multiple files).

(g) SRA - is a compressed file format of raw data, that supports files such as FASTQ and BAM. It is used for efficient data storage by public databases (NCBI, EBI, and DDBJ). Users can download published raw data in SRA format, then decompress it using the SRA toolkit [168].

2.5 Discussion

In this chapter we gave an introduction for animal and plant miRNAs, together with their biogenesis, functioning mechanisms and roles in organisms. We then presented high throughput sequencing technologies, helper tools and file formats used for miRNA prediction, their existence facilitating the development of miRNA prediction and analysis algorithms.

Because the field of miRNA and small non-coding RNA research is so complex, many tools have been developed to aid in processing, analysing and visualising of such data. Continuous efforts are aimed towards understanding various aspects of miRNAs, and, as progress is achieved in this field, the need to expand and adapt helper tools remains constant.

Although a substantial amount of research was focused on miRNAs in the last couple of years, there are still aspects that we do not yet fully understand. In animals, miRNA targeting can be described as a complex network of mutually interacting elements. The complexity of interactions between miRNAs, they targets and multiple other elements during miRNA targeting, has created challenges

for understanding the exact roles and functioning of each element involved. This is also a difficult task for the tools offering animal target prediction and miRNA-target interaction visualisation. Therefore, further improvements in this area could be achieved to improve the accuracy of such tools.

In plants, miRNAs from different species have developed divergent biogenesis pathways (being produced by homolog Dicers). The different lines of evolution suggest that they are continuously evolving and adapting, which also means that plant miRNAs have developed a more complex pattern for their biogenesis. This has caused the tools trying to predict plant miRNAs to achieve lower accuracy.

miRNAs have proved to be essential in all eukaryotes, because of their crucial roles in organ development and implication in disease. Their importance can be also justified by the fact that they evolved independently in the plant and animal kingdoms, in an example of convergent evolution. They have different modes of action in the two kingdoms and no seeming correspondence between their sequences. However, the presence of miRNAs in all plant and animal species suggests that life as multicellular organisms could not have been sustained without the miRNAs facilitating the developmental patterning needed.

Recently, miRNAs are widely used for understanding, diagnosing and treating various diseases, both in animals and plants. Abnormal expression levels of certain miRNA (biomarkers) in different tissues or organs can indicate the presence of disease or cancer, helping in early diagnosis, which can be crucial for an efficient treatment. By having a clear image of which miRNAs are involved in certain processes, researchers can develop treatments, by controlling the up- or downregulation of these miRNAs. Improving the current knowledge in this area could become the basis for future medical research, which could make use of more accurate information about organism-wide miRNA and sRNA interactions, or individual specific details (for achieving personalised treatments).

Therefore, it is essential to continue researching miRNA discovery and analysis methods. In the next chapter we will focus on giving a description of existing miRNA detection software, together with a review of their performance. This is relevant to our research, as we developed and tested our new algorithm, miRCat2, which is the focus of this thesis, taking into consideration these previous tools and their performance.

Chapter 3

miRNA detection methods

3.1 Summary

This chapter gives a review of the more commonly used tools for novel miRNA detection from HTS datasets (both for plants and animals). First, we describe the most important features of each algorithm, focusing on miRCat and miRDeep2, to understand how they perform miRNA prediction. Second, we give a comparison of the tools performance, by analysing their results and assessing their sensitivity and specificity rates, run time and memory consumption. We then present a critique of these tools.

3.2 Overview

As miRNAs have been an important area of research over the last decade, a growing need to discover miRNA sequences and analyse their functions has arisen. When HTS was introduced [138], it produced large amounts of data, which became too much for manual processing. Biologists have asked for the help of bioinformaticians, who started developing tools to analyse such data.

Several computational tools for identifying animal and/or plant miRNAs from HTS data have been developed. Some of the more commonly used tools, in order of appearance, are: miRDeep [149], miRCat [1, 4], miReap (http://mireap.sourceforge.net/), MIReNA [169], miRAnalyzer [170], miRDP (formerly known as miRDeep-

P) [171], miREvo [172], deepBlockAlign [173], miRDeep2 [2], MaturePred [174], miRDeep* [175], miRAuto [162], miRPlant [3], miR-PREFeR [176], Mirinho [177] and miRA [178]. Links for their official download page, number of citations (verified on Friday 23rd September, 2016), type of interface and preferred organism for each of these tools are presented in Table 3.1. There are tools that predict miRNAs from other kinds of input data. For example, miRNA Digger [179] uses degradome data. These kinds of tools are not directly comparable to the ones mentioned above and we will not review them here.

Tool name	Link	Cited	Interface	Organism
[149] miRDeep	*https://www.mdcberlin.de/8551903/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep	678	CLI	Animal
[1] miRCat	http://srna-workbench.cmp.uea.ac.uk/tools/analysis-tools/mircat/	281	Both	Both
miReap	http://sourceforge.net/projects/mireap/	N/A	CLI	Both
[169] MIReNA	http://www.lgm.upmc.fr/mirena/index.html	87	CLI	Both
[170] miRanalyzer	http://bioinfo5.ugr.es/miRanalyzer/miRanalyzer.php	225	Both	Both
[171] miRDP	http://faculty.virginia.edu/lilab/miRDP/	92	CLI	Plant
[172] miREvo	https://github.com/akahanaton/miREvo	28	Both	Both
[173] deepBlockAlign	http://rth.dk/resources/dba/	13	CLI	Both
[2] miRDeep2	https://www.mdcberlin.de/8551903/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep	382	CLI	Animal
[174] maturePred	*http://nclab.hit.edu.cn/maturepred/	33	CLI	Plant
[175] miRDeep*	http://www.australianprostatecentre.org/research/software/mirdeep-star	59	GUI	Animal
[162] miRAuto	*http://nature.snu.ac.kr/software/miRAuto.htm	2	GUI	Plant
[3] miRPlant	http://www.australianprostatecentre.org/research/software/mirplant	9	CLI	Plant
[176] miR-Prefer	http://www.cse.msu.edu/ leijikai/mir-prefer/	10	CLI	Plant
[177] Mirinho	http://mirinho.gforge.inria.fr/	1	CLI	Both
[178] miRA	https://github.com/mhuttner/miRA	0	CLI	Plant
[179] miRNADigger	http://www.bioinfolab.cn/	0	CLI	Both

Table 3.1: **miRNA detection tools, links to their official page as declared in their publication papers, number of citations as taken from Google Scholar on Friday 23rd September, 2016 and suitable organism to run on.** * - link was not accessible on checked date; interface type: CLI = Command Line Interface, GUI = Graphical User Interface.

To assess the performance of software tools, together with the strengths and weaknesses of distinct algorithms, a series of metrics are generally used. When predicting miRNAs, it is important to establish the number of real miRNAs detected, often referred to as true positives (TP), the number of false predictions, called false positives (FP), the number of miRNAs that were present in the input dataset but not detected - false negatives (FN) and the number of sRNAs not predicted that are not miRNAs - true negatives (TN).

Using these concepts, we define the following metrics for assessing the performance of software tools:

- Sensitivity (sometimes also called recall) - is the number of predicted miRNA

from reference dataset (eg. miRBase) divided by the total number of miRNA reads from miRBase, present in the file.
The formula for sensitivity is: TP / (TP + FN).

- Specificity (sometimes also called precision) - is the number of predicted miRNA in miRBase out of the total number of predicted miRNAs.
  The formula for specificity is: TP / (TP + FP).
- Accuracy - is the number of total sequences that are well classified (predicted true miRNAs and unpredicted true negatives) divided by the total number of sequences in the input dataset.
  The formula for accuracy is: TP + TN / (TP + FP + TN + FN).

The miRNA detection tools mentioned above have different approaches in selecting their miRNA candidates, and also tend to generate distinct results. Reviews show that many suffer from high false positives and negatives, lack of consistency across species and high runtime and memory consumption [180–182]. These are indicators that improvements to such software are required or new software needs to be developed.

Consequently, we have developed a new piece of software, miRCat2, as a complete redesign of the miRCat algorithm and also including useful features inspired from the miRDeep2 algorithm; therefore it is important to understand how these two pieces of software work. By analysing their algorithms in more detail, we were able to identify their strengths and weaknesses, which enabled us to improve on their results and performances.

In this chapter we give a brief overview for each of the above existing miRNA detection methods, focusing more deeply on miRCat and miRDeep2. We then give a short review for the most relevant tools for this thesis.

3.3 Algorithm description of miRNA detection tools

3.3.1 miRCat

Here we give a review of the miRCat tool, as found in UEA small RNA Workbench v3.2 [4]. miRCat [1, 4] is a sRNA analysis tool that predicts miRNAs from HTS datasets, both in animals and plants. It is included in the UEA small RNA Workbench [4], which is a collection of tools designed for the processing and analysis of sRNA data. The Workbench includes helper tools (Adapter Removal, Filter, Sequence Alignment), analysis tools (miRCat [1], miRProf, SiLoCo [1], ta-siRNA prediction [1], PAREsnip [183], CoLide [184]) and visualisation tools (RNA/Folding Annotation, VisSR). The UEA Small RNA Workbench can be run on any operating system running Java (Windows, Linux and Mac OSX), and has a user-friendly graphical interface or can be run from the command line.

miRCat receives two files as input: the reference genome of the studied organism and the sRNA sequence file in FASTA or PatMaN format. If the file is in FASTA format, before processing, miRCat maps the sRNA sequences full length to the genome. To do this, it uses PatMaN [153] (PatMaN is provided in the dependencies archive for the tool-kit).

miRCat has two sets of default values for the parameters, one for animals and one for plants, the values being customisable from the GUI or provided as a configuration file in command-line mode.

The workflow of the algorithm of miRCat is summarised in Figure 3.1. Now we present a more detailed description of its key features, as this is important later on.

1. Candidate selection

After the input files are processed and the sRNA reads are mapped to the genome, miRCat looks for genomic regions that have sRNAs aligned to them (sRNA loci), containing at least one read with abundance (or read count) equal to or greater than *min_abundance* parameter (default five).

These loci need to meet the following criteria to be considered to contain

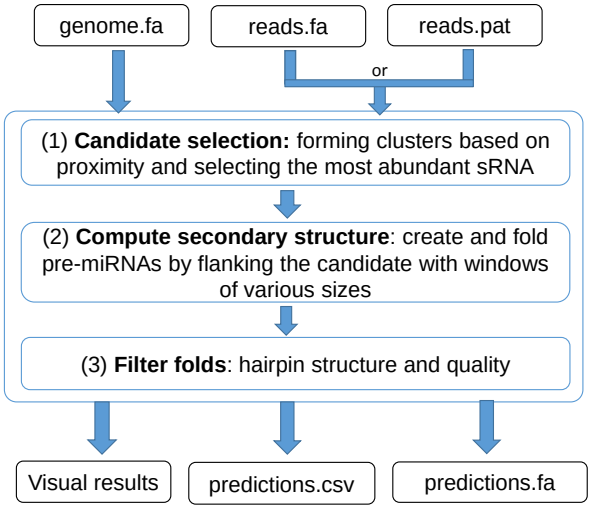


Figure 3.1: **Flowchart diagram representing the miRCat algorithm.**

a potential miRNA precursors. Firstly, each sRNA in a loci must be no more than 50 (animals) or 200 (plants) nts away from its closest neighbour (*hit_dist* parameter). This way, adjacent loci are always separated by at least *hit_dist* nts, while sRNAs inside a cluster are situated closer on the genomic location. Secondly, at least 90% of sRNAs in a cluster must have the same genomic orientation (*percent_orientation*). This is necessary because only on rare cases there are equal amount of sense and antisense matches on a real miRNA precursor, as usually the miRNA and miRNA* come from the same strand and should be highly expressed [185].

Once a list of loci has been produced, these are further analysed to find likely miRNA candidates. For each locus, the most abundant sRNA read is selected as the likely miRNA. If the selected sRNA has the standard miRNA-like features, such as the length between accepted values (*min_length*, *max_length*), a minimum percent of G and C nts in its composition (*min_gc*), and does not match the genome more then *genome_hits* times, then the software proceeds to check if it is a true miRNA.

2. Computing the secondary structure

Fourteen flanking sequences of different sizes (from 10 to 200 nts, empirically determined), surrounding the sRNA on both sides are extracted from the genome to create pre-miRNA candidates (see Figure 3.2). The flanking sequences together with the miRNA candidate form windows of sequences of various lengths that are later checked for the optimal secondary structures.

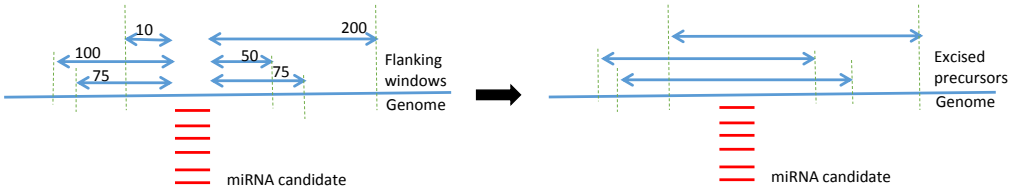


Figure 3.2: **Determining the secondary structure of the candidate miRNA in miRCat.** Multiple flanking sequences of varying lengths are used to obtain the potential precursors, which are then folded and further processed.

Each sequence window is then folded using the RNAFold tool from the ViennaRNA package [154], producing a MFE secondary structure for the putative miRNA. Each folded hairpin is being divided in three parts: 5' flank (before miRNA - could contain miRNA* and loop), miRNA, 3' flank (after miRNA - could contain miRNA* and loop), on which further filters are applied.

3. Folds filtering

miRCat computes discriminative features on the trimmed secondary structure that are useful for classifying miRNAs. The most important features are:

- The number of consecutive gaps and/or mismatches between miRNA and miRNA* must be no more than *max_gaps* (default 3): miRNA and miRNA* should not contain bulges larger than 3 nts [15, 65];
- The number of paired nucleotides between the miRNA and the miRNA* must be at least *min_paired* (17) of the 25 nucleotides centred around the miRNA [114];
- The length of the hairpin must be at least *min_hairpin_len* (75nt for plants and 50nt for animals), to ensure there is enough space for all miRNA biogenesis

products (miRNA, miRNA*, loop, 3' and 5' overhang) [15, 114];

- The percentage of paired bases in the hairpin must be at least $max\_percent\_unpaired$ (50%) of base-pairs in the hairpin [114];
- The miRNA and miRNA* should never basepair with itself and should not contain loops [15];
- If the hairpin structure is not perfect, then it checks to see if it should allow complex loops (more than one bulge, pairing nts inside loop) [53, 111–113].

The adjusted minimum free energy per 100 nts ($AMFE$, $AMFE = \frac{MFE}{length\_of\_hairpin} * 100$) is then computed for the potential precursors that pass the above criteria, to get a comparable measurement of a hairpins MFE. If the AMFE is below the mfe_param, the hairpin information is stored. Because miRNA precursors are very stable [35], miRCat selects then the hairpin with the lowest AMFE as the pre-miRNA candidate [65], which guarantees is the most stable secondary structure for the candidate. On this hairpin it validates the miRNA*: it finds the possible locations of the miRNA* in the hairpin, checks the properties (no pairs inside the sequence, minimum number of gaps, minimum no of nts paired) and looks for its sequence among the input reads.

The total abundance for the miRNA and miRNA* from overlapping sequences are added and their percent out of the sum of all sequences on the hairpin should be less than the *overlap_parameter*. This is done to ensure that the miRNA and miRNA* locus is clearly defined and the reads do not map randomly over the precursor.

The pre-miRNA candidate is then tested using RANDFold [155], checking if its p-value provides statistical evidence that it is a miRNA precursor.

miRCat then outputs the information about the sequences and precursors that passed all filters in the GUI, the user having the option of saving it to a file (csv, FASTA).

3.3.2 miRDeep2

miRDeep2 is a sRNA analysis software tool that predicts miRNAs for animal organisms from high-throughput sequencing datasets [2]. It is built on the original

miRDeep algorithm [149] (which is presented in Figure 3.3), adding additional features and packages. There are several versions of miRDeep/miRDeep2, developed and adapted by various research groups: miRDeep* [175] has a graphical user interface and is used for animals, while miRDP [171] and miRPlant [3] adapted the miRDeep algorithm for plants. These tools are explained in more detail in the following subsections.

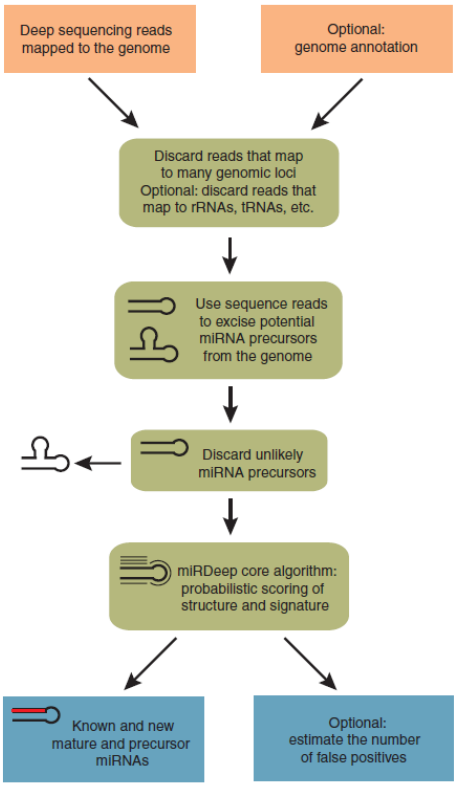


Figure 3.3: Flowchart diagram representing the miRDeep algorithm. Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology [149], copyright 2008.

As input, miRDeep2 receives a sRNA sequence file in redundant format (FASTA), the genome of the species to be analysed (FASTA) and a file of the reads mapped to the genome (arf format). The arf mapped file is obtained by mapping the reads to the genome using the alignment algorithm Bowtie [186] and the mapper tool provided in the miRDeep2 package. Optionally, the software also takes files containing miRNA precursors, mature miRNAs for that species or for related species (FASTA).

Now we present the miRDeep2 algorithm in more detail.

1. Excising precursors

After the input files are read, the algorithm first parses the reads on the genome and discards the "false reads": it keeps only the sRNAs with length >18 nts that map less than or equal to 5 times to the genome.

In the second step, for every sRNA, it looks 70 nts downstream on the genome and if it encounters a read with higher counts, this sRNA is chosen. This is done iteratively until no higher read stack is found, this way selecting the sRNA with highest local abundance. For that read it obtains the precursor sequence by excising twice: once including 70 nts upstream and 20 nts downstream flanking sequences, and once 20 nts upstream and 70 nts downstream (see Figure 3.4).

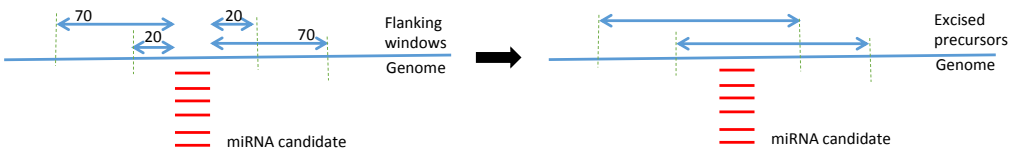


Figure 3.4: **Selecting the potential precursor sequences in miRDeep2.**

Thirdly, it maps all the reads to the previously excised precursors using Bowtie [186]. If the optional files are given, miRDeep2 also maps the referenced mature miRNAs to the precursors.

In the fourth step, the precursors are then folded using the RNAFold tool [154] and the software analyses that the folds are consistent with miRNA biogenesis, by applying a filtering step that discards potential precursors that do not have a hairpin structure. For the remaining hairpins, optionally, randfold p-values are calculated.

Further computations are performed on these precursors to select which might correspond to *bona-fide* miRNAs (true miRNAs).

2. Core algorithm

For the selected potential precursors, miRDeep2 then probabilistically integrates HTS information based on a simple model for miRNA precursor processing by Dicer. If a sequence is an actual miRNA precursor that is expressed in the HTS sample, then one expects that one or more HTS reads correspond to one or more of the three Dicer products: the mature miRNA sequence, the miRNA* sequence and the loop [149].

If a file with already known miRNAs for the species is given, it defines a set of seeds by selecting the first 7 positions of each known miRNA.

The precursors are processed using the following routine.

For each precursor it identifies the sequence corresponding to a mature miRNA, miRNA* and loop and records its start and end positions, strand and frequency. Then, the precursors need to pass a series of filters: all parts of the hairpin (miRNA, miRNA* and loop) need to be identifiable; it must present no bifurcations (additional hairpins on the structure); there should be a minimum of 60% base pairing in the duplex; there should be not more than 6 nts between the mature miRNA and the miRNA* in length; the reads mapped to the miRNA location should be not longer than 25; the reads corresponding to the miRNA* should be aligned where expected ($\pm$ 1 nt) and 90% of the reads must map in consistence with the Dicer processing (the three Dicer products should be clearly delimited).

The precursors that pass the filtering step are then tested further, by computing a probabilistic score for the combined compatibility of MFE, frequencies of reads and positioning in correspondence with Dicer processing. A number of features contribute to the score.

The basic score [149] is calculated by fitting the values from the MFE of the hairpin and the total frequency of the miRNA, miRNA* and loop into a Gumbel distribution [187]. The Gumbel distribution is used in statistics to predict the chance that an event will occur, based on a number of previous samples of various distributions. The parameters to create the Gumbel distribution describing miRNA precursors were generated using data from *C. elegans* real miRNA precursors. The value given by checking the candidate hairpin against the Gumbel distribution represents the probability of it being a real miRNA precursor. If however, the miRNA* sequence is missing, miRDeep2 then reduces the basic score

to 0, because a strong miRNA candidate would also have the miRNA* sequence expressed in the sample.

A series of predefined values are then added to the basic score in each of the following cases: if the seed is preserved (the seed is made of the first 7 characters of a miRNA; the presence of a seed will greatly contribute to the score, but its absence will not be decisive), if the miRNA* is present and begins where expected and if the RANDFold p-value is significant.

If the score is greater than a threshold (e.g. score >50), then the sequence is considered a miRNA.

3.3.3 miRDP

miRDP [171], formerly known as miRDeep-P, is based on the core algorithm of miRDeep, but has been adapted to work on plants. The flanking sequence for excising precursors was extended to 250, to accommodate the longer plant secondary structures. These are then processed by the miRDeep core algorithm with a plant-specific scoring system. Additional filters for plant-specific criteria based on known characteristics of plant miRNA genes are then applied.

3.3.4 miREvo

miREvo [172] is built on the miRDeep2 predictor and is suitable for both plant and animal data. SmiREvo extends the miRDeep2 algorithm for evolutionary analyses. Specifically, it uses whole-genome alignments to identify miRNA homologues in related species. It also includes tools to compare expression of miRNA homologues across species, if sRNA sequencing data are available for both species. It uses modified prediction parameters for plant analyses.

3.3.5 miRDeep*

miRDeep* [175] is a tool with user-friendly graphic interface modelled on miRDeep, with improved precision of detecting novel miRNAs [175]. It introduces new strategies for preprocessing and identifying precursor miRNAs (improved

precursor excision), Bowtie mapping and target prediction for known and novel miRNAs. The tool is implemented entirely in Java without requiring any pre-dependent computational tools, making it portable and easy to install. The computational efficiency allows it run on a desktop computer.

3.3.6 miRPlant

miRPlant [3] is a plant miRNA detection tool built on the miRDeep*, miRDeep and miRDP tools, providing a user-friendly interface. It has customisable parameter set-up, an improved method of pre-miRNA selection and allows for internal loops (multiple loops between miRNA and miRNA*), which contributed to at least a 10% improvement in specificity compared to miRDP [3].

3.3.7 miReap

miReap (http://sourceforge.net/projects/mireap/) combines sRNA position and depth with a model of microRNA biogenesis to discover miRNAs from HTS sRNA libraries. miReap is a command line tool and takes as input a sRNA file (FASTA), a file containing sRNA mapping information and the reference genome (FASTA).

The format of the sRNA mapping file should be read_ID, chr_ID, start, end, strand(+/-)(delimited by tab or space), although it does not provide a tool for obtaining the mapped file in the required format. miReap does not have extensive documentation, being unclear for which Kingdom it was designed, but has been previously used both on animal [180] and plant data [188–190].

miReap classifies a stem-loop hairpin as a typical pre-miRNA only when it satisfied the following criteria [191]: mature miRNAs were present in one arm of the hairpin precursors, which lacked large internal loops or bulges, and the secondary structures of the hairpins were steady, with the free energy of hybridization less than 20 kcal/mol.

3.3.8 MIReNA

MIReNA [169] can find miRNAs at the genome scale and from deep sequencing data. It uses a rule-based scheme with sharp cut-offs with only five parameters to

identify pre-miRNA/miRNA pairs. The rules are based on the following features: the lack of base pairing in the mature miRNA, the difference in length between the two candidate miRNA strands, the fraction of base-paired nucleotides in the hairpin, and two measures of energetic stability. As a second filtering step, it considers only hairpins where the sequenced RNAs map in consistence with Drosha/Dicer processing and it can consider several potential miRNA duplexes within one precursor structure.

It can handle four kinds of data (known miRNAs, deep sequencing data, potential miRNAs occurring in long sequences, and putative pre-miRNAs containing potential miRNAs).

3.3.9 miRanalyzer

miRanalyzer [170] first removes reads that map to known miRNAs or other transcripts, considering only the remaining reads as new miRNAs. miRanalyzer uses Bowtie [186] to map input reads to the target genome. It implements a machine learning algorithm based on the random forest classifier that is initially trained on a set of known miRNAs from human, rat or nematode. It considers features like: energetics, structure, bulges and number of mapped reads. The tool has fitted parameters for each species analysed (31 commonly used species, including 6 plants). It can also perform differential expression analysis and predict targets using the TargetSpy tool [192].

miRanalyzer is available through a web server and also as a stand-alone version that can be run on local machines. Apart from specifying the number of allowable mismatches, and the acceptable P level for a credible prediction, the user, however, is restricted from making any other parameter changes in the algorithm.

3.3.10 deepBlockAlign

deepBlockAlign [173] provides a scoring of the read signature (read alignment and counts), but does not evaluate the RNA structure. It uses a variant of Needleman-Wunsch algorithm (an alignment algorithm for sequence data) [193] to identify blocks of mapped reads that have similar features, including read begin positions

and block height (block read count). In a second step, similar groups of blocks are identified using a variant of the Sankoff algorithm (an alignment algorithm for blocks of sequences) [194]. These groups of blocks should correspond to gene loci. To predict novel miRNAs, the method finds loci that have block features similar to known miRNAs.

While the profiles might be different for plants and animals, or specific to particular tissues or pathological conditions, the method can compare to all known profiles from the entire miRBase database of miRNAs, giving it good coverage. Since this method does not evaluate the RNA structure, it can predict miRNAs that do not have canonical structure, or whose conformation is not easily predicted by computational methods.

3.3.11 MaturePred

MaturePred [174] is a plant specific miRNA prediction algorithm, which regards the miRNA/miRNA* duplexes as a whole to capture more of its characteristics and constructs a model based on SVM (support vector machines, a supervised learning approach) to predict the position of miRNAs inside their precursors. The proposed model considers in a total of 160 features, ranging from position-specific features of a single nucleotide to structure-related, energy-related and stability-related features.

3.3.12 miRAuto

miRAuto [162] is a tool that can be used to predict plant miRNA, providing a user-friendly interface and integrated analysis. miRAuto uses database information and predicted/statistical approaches to make its predictions, which provide reliable results in both model and non-model plant species for conserved and novel miRNAs [162]. miRAuto chooses its candidates based on expression analysis of the 5'-end position of mapped small RNAs in reference sequences, to prevent the possibility of mRNA fragments being included as candidate miRNAs.

3.3.13 miR-PREFeR

miR-PREFeR [176] takes a genome file of a species and one or multiple sRNA read alignment files (SAM format) of the same species as input.

The pipeline first generates candidate regions and candidate mature sequences of each candidate region based on the alignment depth. In the next step, these regions are folded using RNALfold. Regions with qualified stem-loop structures are then examined using published plant miRNA annotation criteria: the sRNA data should provide evidence of precise miRNA/miRNA* excision, criteria related to structure characteristics of the miRNA/miRNA* duplex. In addition, expression information from multiple sRNA data samples is also used to improve the accuracy of the prediction.

By default, the pipeline makes a checkpoint after each major step of a job, and makes checkpoints periodically within the time-consuming folding stage. This provides users an easy way to restart unfinished jobs. By restarting a job from the latest checkpoint other than starting it from the beginning, a lot of time/resources can be saved for long-running jobs on large plant genomes.

3.3.14 Mirinho

Mirinho [177] is a plant miRNA detection tool that detects pre-miRNA both with or without an input sRNA file. It offers a novel alternative to a classical MFE folder based on a thermodynamic Nearest-Neighbour (NN) model for computing the MFE and predicting the classical hairpin structure of a pre-miRNA. The free energies thus computed correlate well with those of RNAfold [154], but the method has quadratic instead of cubic complexity and is much more efficient in practice [177]. Mirinho uses only knowledge of the length of the loop and stem-arms and the MFE of the pre-miRNA hairpin to classify miRNAs.

3.3.15 miRA

miRA [178] can be used to identify miRNA precursors in plants. It requires an aligned sRNA file (SAM format) and a corresponding reference genome (FASTA format), and evaluates precursor secondary structures and precursor processing

accuracy. It does not require cross-species miRNA sequence conservation and it allows for a heterogeneous miRNA precursor population (non-characteristic secondary structures).

The miRNA detection is based first on identify genomic contigs based on sRNA sequencing data; secondly, analysing secondary structures for every cluster; lastly, verifying that RNA sequencing data-based read coverage of miRNA precursor candidates is consistent with miRNA precursor processing resulting in the expression of one or more mature/star miRNA duplexes.

3.4 Performance of existing miRNA detection tools

To chose the method of miRNA detection that is appropriate for a specific experiment, we need to better understand their performance. In some cases, the user might want to get a set of results in the fastest way, other times they would choose a method with higher accuracy even if it could run for days or weeks. Some users are constricted by low RAM and need a tool without many technical requirements, other times they can make full use of high performance computing facilities.

Here we review the performance of the tools described above to give a better idea which tool is most suitable for which cases. This will be important for the design of our new tool.

1) miRCat was originally tested on several high-throughput plant sRNA datasets and showed high levels of both sensitivity and specificity [1]. In an *Arabidopsis* leaf dataset, miRCat predicted 89 miRNA loci using default parameters (83 of these were known miRNA sequences and 6 were novel miRNA loci), while there were 91 known miRNA loci with a sRNA abundance of five or more (default threshold for miRCat) in the dataset. This showed 91.2% sensitivity and, even if all novel predictions are considered FP, this would give a specificity of 99.93%.

We have compared miRCat to miRDeep2 on zebrafish data, using default parameters, and produced a Venn diagram, showing the intersection between the

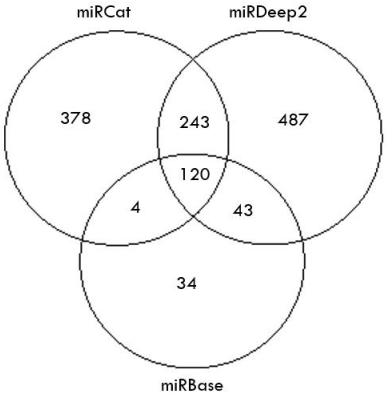


Figure 3.5: **Venn diagram of miRCat and miRDeep2 predictions on zebrafish data.** Figure shows all miRCat and miRDeep2 predictions and their overlap with miRBase miRNAs.

two software tools and the miRBase annotations (Figure 3.5). Both tools detect high numbers of miRBase miRNAs, miRDeep2 having 39 additional miRBase predictions. Although there is a considerable overlap between the two tools, both for known (120 miRNAs) and novel predictions (243 miNRAs), their results show different numbers for both cases.

2) In the paper publishing miRA [178], the tool is compared to two plant tools: miR-Prefer and miRDP. For each method and dataset they calculate the recall rate (i.e. sensitivity or true positive rate) RR = TP/(TP + FN). Some of their results are presented in Table 3.2 [178].

For both *Chlamydomonas reinhardtii* datasets, miRA recall rates are over 80%, the reference set being taken from a study publishing miRNAs in *C. reinhardtii* [195]. The larger number of novel miRNAs derived from the data in Set1 compared to those from Set2 is related to the larger sequencing depth of the former (see Table 3.2). Recall rates for miRDP and miR-PREFeR are significantly smaller, dropping below 50% in some cases; in a direct comparison of miRDP and miR-PREFeR, the former seems to perform better with low sequencing depth data, while miR-PREFeR outperforms miRDP with deeper sequencing data.

The *Arabidopsis thaliana* results are compared to miRBase, and the recall rates of miRA and miR-PREFeR are near identical, with miRA predicting more

novel miRNAs, indicating possibly more FP. This is believed to be related to miR-PREFeR's requirement of the existence of star-sequence associated reads, whereas miRA does not impose a minimum expression threshold on the star sequence. The performance of miRDP is significantly lower than that of miRA and miR-PREFeR, having low RR rate and high number of novel predictions.

Organism	Method	Nref	Nrecall	RR	Ntot
Chlamydomonas reinhardtii					
Set1	miRA	47	39	0.83	281
Set1	miRDP	47	14	0.3	964
Set1	miR-PREFeR	47	29	0.62	60
Set2	miRA	15	12	0.8	175
Set2	miRDP	15	7	0.47	51
Set2	miR-PREFeR	15	3	0.2	6
Arabidopsis thaliana					
Set1	miRA	246	122	0.5	517
Set1	miRDP	246	80	0.12	695
Set1	miR-PREFeR	246	119	0.48	138

Table 3.2: **Performance comparison of miRA, miRDP, and miR-PREFeR.** Nref is number of known miRNAs for the organism. Nrecall gives the number of identified known miRNAs. RR is the recall rate (sensitivity). Ntot gives the total number of identified miRNAs. This is reproduced from [178].

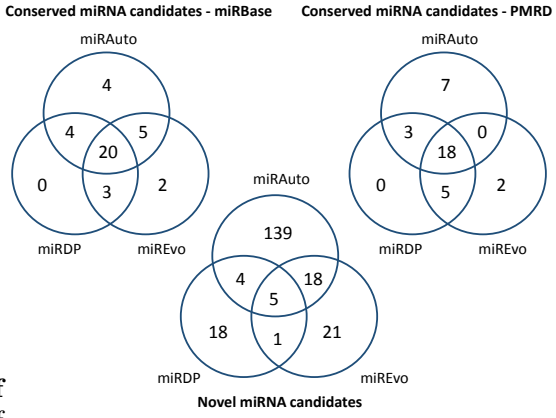


Figure 3.6: **Venn diagrams comparing the performance of miRAuto, miRDP and miREvo.** The results are shown for miRBase and PMRD miRNAs and novel predictions. This is reproduced from [162]

3) miRAuto was compared in its paper [162] to miRDP and miREvo, using sRNA sequencing data from hot pepper fruit tissue (*Capsicum annuum*).

In the comparison test, small RNAs with a count of 500 or higher and 18-25 nt in length were selectively chosen to determine the number of miRNA candidates predicted by all three programs (see Figure 3.6). The results were compared to annotated sequences from miRBase and PMRD [196] databases.

A total of 38 and 35 miRNAs from each database, respectively, were predicted by the three programs. Among these, 20 (52.6%, miRBase) and 18 (51.4%, PMRD) miRNAs were predicted by all three programs, suggesting that miRAuto performs comparably to the other two programs. For novel miRNA candidates, a total of 206 miRNAs were predicted by the three programs. Of these, 5 (2.4%) miRNAs were commonly predicted by all three programs, and 23 (11.2%) and 9 (4.4%) miRNAs from miRAuto/miREvo and miRAuto/miRDP, respectively.

miRAuto predicted a higher number of novel miRNA candidates (but possible false positives).

4) miRDeep* has benchmarked its sensitivity and specificity in detecting mature miRNAs against miRDeep, miRDeep2, miRanalyzer, and MIReNA [175], on two datasets, mock-treated and R1881-treated LNCaP cells (human prostate cancer cells). For the number of predictions for each of these tools see Table 3.3.

	miRDeep*	miRDeep2	miRDeep	miRanalyzer	MIReNA
Mock treated					
Known miRNA in raw RNAseq reads	240	240	240	240	240
No. predicted miRs	208	272	203	1063	162
No. predicted miRs in miRBase	173	208	164	215	77
No. predicted novel miRs	35	64	39	848	85
Precision	83.17%	76.47%	80.79%	20.23%	47.53%
Recall	72.08%	86.67%	68.33%	89.58%	32.08%
R1881 treated					
Known miRNA in raw RNAseq reads	285	285	285	285	285
No. predicted miRs	237	320	235	1321	190
No. predicted miRs in miRBase	192	229	180	261	88
No. predicted novel miRs	45	91	55	1060	102
Precision	81.01%	71.56%	76.60%	19.76%	46.32%
Recall	67.37%	80.35%	63.16%	91.58%	30.88%

Table 3.3: **Comparative analysis of the sensitivity and specificity of miRDeep*, against miRDeep2, miRDeep, miRanalyzer and MIReNA**. Precision = Number of predicted miRNA in miRBase/Number of predicted miRNA. Recall = Number of predicted miRNA in miRBase/RNAseq reads found in miRBase. This is reproduced from [175].

The majority of the highest scoring miRNAs from the results of miRDeep*, miRDeep2 and miRDeep, are already in miRBase, and thus more likely to be *bona-fide* miRNA. miRanalyzer output approximately five times more predictions compared with miRDeep*, miRDeep2 and miRDeep, although had a higher proportion of novel miRNA with higher number of reads. MIReNA predicted an overall small number of miRNAs, but many of them are not in miRBase, both the precision and recall of MIReNA being lower than that of miRDeep*.

In terms of precision, miRanalyzer had the lowest value with only 19.76% and 20.23%. This indicates that ∼80% of the miRNA predicted by miRanalyzer are novel and without any validation. The precision for miRDeep* and miRDeep was >70%, with miRDeep* having a 2.9% and 5.8% higher percentage for the mock-treated and R1881-treated datasets, respectively (see Table 3.3).

miRDeep* also slightly outperformed miRDeep in detecting validated miRNA with 5.5% and 6.7% higher recall. Interestingly, miRDeep2 had lower precision (76.47% and 71.56%) compared with the original miRDeep and miRDeep*, but had higher recall ratio (86.67% and 80.35%).

Some miRNA, such as miR-25 and miR-200a, were found to be highly expressed in the LNCaP dataset, but were not detected by miRDeep due to improper excision of the pre-miRNA region in those algorithms.

miRDeep* and miRDeep2 were also compared on data generated before and after inducing anti-dicer in MCF-7 cells to turn off the miRNA biogenesis pathway. The novel miRNAs predicted by miRDeep* have a lower average log(FC) (fold change) compared with miRDeep2, which demonstrates that these novel predictions are more likely to be generated from the miRNA biogenesis pathway. This is further supported as the percentage of miRNA with a negative fold change after Dicer knock-down, is greater than that of miRDeep2. However, miRDeep2 was able to predict more novel miRNA than miRDeep*.

5) MIReNA is compared with miRDeep in its paper [169], using *C. elegans* and *H. sapiens* data. MIReNA provides a slightly lower number of predictions with a lower sensitivity against a higher signal-to-noise ratio than miRDeep (see Table 3.4), but overall the two tools appear complementary: by running MIReNA one can recover a number of miRNAs in miRBase, which have been missed by miRDeep and vice versa. MIReNA predicts 5 (4 of which are MIReNA specific) new pre-miRNAs for *C. elegans* and 63 (29 of which are MIReNA specific) for *H. sapiens*; miRDeep predicts 1 (non-specific) novel pre-miRNA for *C. elegans* and 64 (30 of which are miRDeep specific) for *H. sapiens*.

This might appear because of the different ways of the potential miRNA selection between MIReNA and miRDeep at the beginning of the algorithm, that for MIReNA is less restrictive. MIReNA accepts pre-miRNA/miRNA pairs where each miRNA matches a read in the dataset, while miRDeep only considers those pairs where the miRNA is most represented by reads matching the pre-miRNA. This means that for the same pre-miRNA, MIReNA may consider several pre-miRNA/miRNA pairs, whereas miRDeep considers only one. Another main difference is in the second filtering step of MIReNA, which considers the same ideas

Homo sapiens

Method	Pred precs	Sens	Signal to noise	Specif in mirbase	new precs	
					All	Specif
miRDeep	284	70.55	08:01	31	64	30
MIReNA	266	64.42	09:01	11	63	29

Caenorhabditis elegans

Method	Pred precs	Sens	Signal to noise	Specif in mirbase	new precs	
					All	Specif
miRDeep	120	85.51	12:01	10	1	0
MIReNA	116	79.71	17:01	2	5	4

Table 3.4: **Comparison of MIReNA and miRDeep.** The table shows the number of predicted precursors (2nd column), sensitivity (3rd), signal-to-noise ratio (4th), number of specific (that is, captured by one method but missed by the other) miRNAs in miRbase (5th), total number of new predicted precursors (6th) and number of new specific predicted precursors (7th). An exact match with the miRNA in miRbase or with a read is required for the results in the last three columns. This table was reproduced from [169].

	Rice DS1		Rice DS2		*A. thaliana*		*M. truncatula*		*P. persica*	
Tool	miRDP	miRPlant	miRDP	miRPlant	miRDP	miRPlant	miRDP	miRPlant	miRDP	miRPlant
Precision	0.82 (31/38)	0.95 (36/38)	0.7 (44/63)	0.83 (52/63)	0.405	0.51	0.22	0.66	0.2	0.55
Recall	0.22 (31/144)	0.25(36/144)	0.24 (44/181)	0.29 (52/181)	0.35	0.65	0.1	0.325	0.29	0.65

Table 3.5: **Comparison of performance for miRPlant and miRDP tools.** Precision = known miRNAs/predicted miRNAs. Recall = known miRNAs/total known miRNAs. This table was reproduced from [3].

used in miRDeep but encodes them in a set of combinatorial rules instead of defining a probabilistic model.

6) miRPlant was benchmarked against miRDP in its publication [3]. They have been tested on rice (two datasets), *Arabidopsis thaliana*, *Medicago truncatula* and *Prunus persica* (see Table 3.5). In rice, miRPlant has better performance than miRDP (because miRPlant uses a flexible method to form the precursor candidates from the genomic region surrounding the sRNA reads).

For the next three organisms, the predicted miRNAs were ranked in descending order of score for each tool, and then the top 100 miRNAs from miRPlant and miRDP were chosen for comparison. Table 3.5 shows that miRPlant consistently outperforms the former tool in all samples.

7) miR-Prefer has been compared to mirDP, miRanalyzer, mirDeep2, mirDeep* and MIReNA [176], Table 3.6 shows the performance of each of these tools on two *A. thaliana* datasets.

miRanalyzer has the best sensitivity (explicable by the fact that it tries to detect annotated miRNAs that are saved in its own database), but the numbers of predictions are large on both datasets, most of these predictions are likely to be

Dataset	miR-PREFeR	miRDP	miRanalyzer	miRDeep2	miRDeep*	MIReNA
Athl-2 (two samples. Number of known miRNAs expressed 240)						
Number of predicted miRs	155	1263	2182	182	2018	152
Number of expressed miRs predicted	127	86	201	64	10	35
Number of novel predictions	28	1177	1981	118	2008	117
Sensitivity	0.53	0.36	0.84	0.27	0.04	0.15
Athl-6 (six samples. Number of known miRNAs expressed: 243)						
Number of predicted miRs	185	3021	13114	291	1472	411
Number of expressed miRs predicted	136	128	209	79	7	44
Number of novel predictions	49	2893	12306	212	1465	367
Sensitivity	0.56	0.53	0.86	0.33	0.03	0.18

Table 3.6: **Performance comparison of miR-PREFeR, miRDP, miRanalyzer, miRDeep2, miRDeep* and MIReNA.** A miRNA is considered to be expressed in the input dataset if at least 20 reads were mapped to the miRNA precursor region in the dataset.

FP. miRDeep2, miRDeep* and MIReNA have low sensitivity on both datasets, which indicates that they should not be used for annotating plant miRNA (they were originally designed for animal data).

miR-PREFeR has the second highest sensitivity and low FP rate, most of its predictions corresponding to previously annotated miRNAs. In all, 77.8% of the predicted miRNAs by miR-PREFeR have the same start positions and 81% of the predictions have the same lengths as the annotations. On the other hand, other tools show much lower consistency with previously annotated miRNAs.

In terms of time and memory resources, miR-PREFeR achieves the fastest runtime, while miRDP and MIReNA have long running time on both datasets. Without manually paralleling the jobs, it is even difficult to run the two tools on small genomes on a personal computer. miR-PREFeR uses less memory than the other tools on both datasets.

8) A software comparison between miRDeep, miRDeep2 and miRanalyzer [181] showed a >80% similarity of known miRNAs in each of the six biological datasets tested upon (see Figure 3.7). The datasets are: neuroblastoma cell line, blood cell line (PMBC), a chronic myelogenous leukemia cell line (K562), acute promyelogenous leukemia cell line (HL60), a breast cancer cell line and a simulated dataset (created using Flux Simulator [197], adding 100 known miRNAs that were 'spiked in' randomly at a prevalence of 0.1%). The tools are also compared to DSAP [198], which predicts only known miRNA signatures, so its performance will not be covered here.

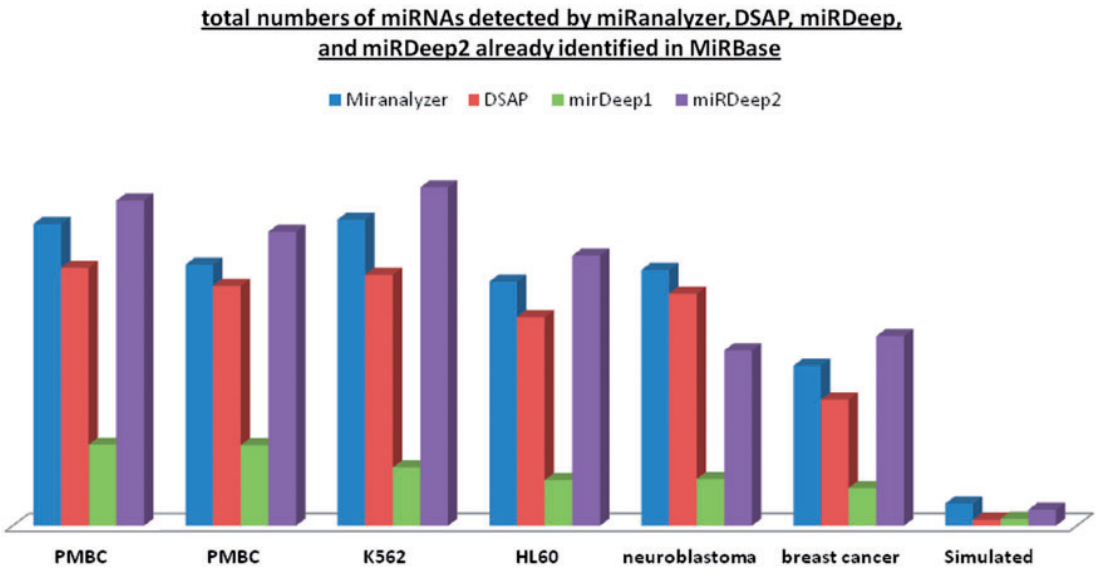


Figure 3.7: **Total numbers of miRNAs detected by miRanalyzer, DSAP, miRDeep and miRDeep2 already identified in MiRBase.** [181], by permission of Oxford University Press.

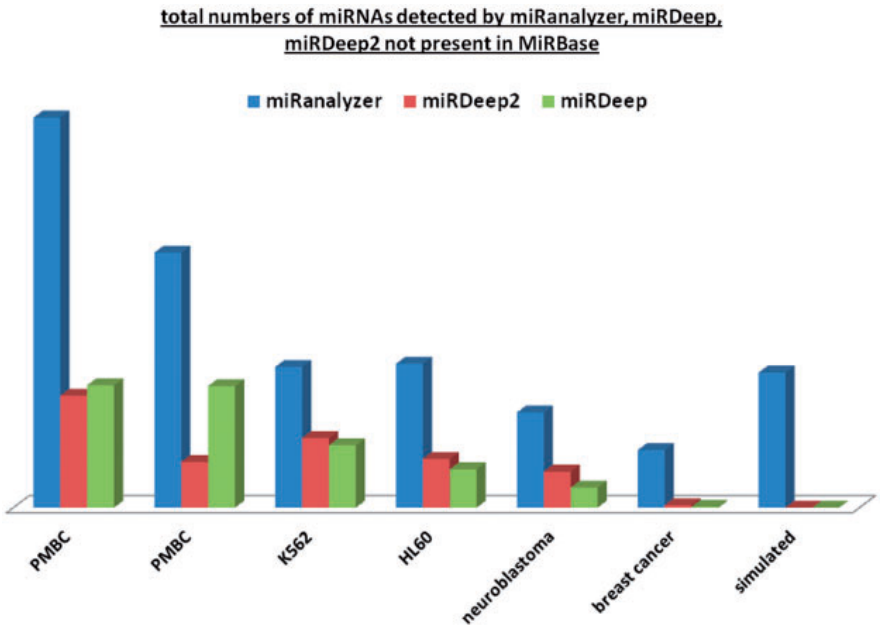


Figure 3.8: **Total numbers of novel miRNAs detected by miRanalyzer, miRDeep and miRDeep2.** [181], by permission of Oxford University Press.

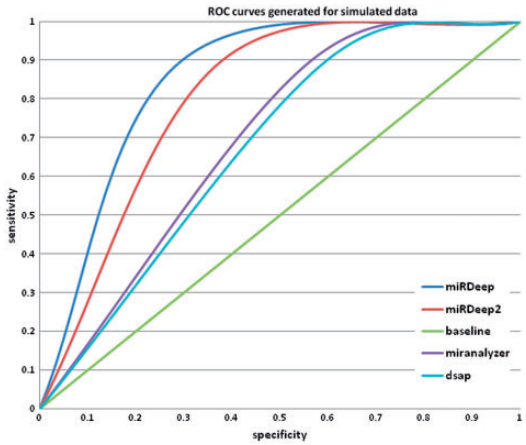


Figure 3.9: **ROC curve on performance of miRDeep/miRDeep2 and miRanalyzer, generated using simulated data.** [181], by permission of Oxford University Press.

Dataset	miRAnalyzer	miRDeep	miRDeep2
PMBC1	12	13	3
PMBC2	7	8	2.5
NB	7	9	4
K562	5	13	3
HL60	8	10	2.5
Breast Cancer	5	8	2
Simulated	13	12	3

Table 3.7: **Calculation time in hours taken by miRDeep, miRDeep2 and miRanalyzer to complete their analysis.** This was reproduced from [181].

In all cases, except the neuroblastoma dataset and the simulated dataset, miRDeep2 generated slightly higher numbers of known miRNAs and the additional miRNAs identified were most often a miRNA from the same family and/or precursor sequence. In the case of the novel miRNA candidates, however, there was a lower percent overlap in the predictions; particularly, between miRAnalyzer and miRDeep/miRDeep2 suggesting that perhaps in comparison to miRDeep, miRAnalyzer is better suited to detect low-expressed candidates (see Figure 3.8).

Another experiment was conducted on the simulated dataset. The tools were tested to have predicted the miRNAs at their correct mapping locations. ROC curves were generated for the simulated dataset. ROC curves [199] are created by plotting the TP against the FP at various threshold settings for the software parameters, thus presenting the sensitivity as a function of FP. ROC analysis can be used as a tool for selecting an optimal model for parameters, which would assure the best output performance.

The ROC curves in Figure 3.9 show that miRDeep/miRDeep2 have slightly better levels of specificity than miRanalyzer and DSAP. Based on the simulation data, accuracy levels for each test were calculated at 80.4% and 75.4% for miRDeep and miRDeep2, and 68.3% for miRanalyzer.

To determine how effective the programs were at identifying novel miRNAs,

the study chose predictions that overlapped in each of the four programs from the neuroblastoma dataset and validated them with Taqman RT-PCR (laboratory technique that monitors in real-time the amplification of targeted DNA molecule). Of the 16 overlapping predictions, 12 novel miRNAs were validated successfully. However, the hairpins predicted by miRDeep and miRanalyzer in many cases were discontinuous representations of each other. The predicted hairpin size varied when compared between miRanalyzer and miRDeep (the average hairpin length predicted by miRanalyzer was 20 bases longer).

To test time consumption of the tools, the study first broke down miRDeep and miRDeep2 into separate tasks, and timed them individually. The amount of time spent by miRDeep to map the reads to the target genome was ∼20% longer than that of miRDeep2. On average, miRDeep took three times as long to complete its analysis (10.5 h) compared with that of miRDeep2 (2.87 h) (see Table 3.7). For miRanalyzer, which at the time was web-based, was difficult to be sure of the result, as ones data is usually placed in a compute queue.

One area in which miRDeep and miRanalyzer both demonstrate apparent weakness is lack of specificity to detect the precursor sequence. When examining the novel miRNAs predicted by miRDeep and miRanalyzer, two instances were detected where precursors were predicted poorly in relation to the mature sequence.

9) In another study [180], miRDeep, miRanalyzer, MIReNA and miReap were compared, amongst others, on HTS datasets derived from three different genomes, i.e. *H. sapiens*, *G. gallus* and *C. elegans*. Other tools included in this study were miRExpress [200], miRTRAP [201], DSAP [198], mirTools [202], and miRNAkey [203], which classify miRNAs based on previous annotations, but do not predict novel ones. As they are not suitable for the scope of our analysis, we do not focus on them.

For a runtime analysis of the software, miReap took less computational time (10 min for *G. gallus* and 43 min for *H. sapiens*) compared with miRDeep (10 days for *C. elegans* and one month for *H. sapiens*) and MIReNA (10 days for *C. elegans* and more than one month for *H. sapiens*).

To evaluate the sensitivity of the tools (percentage of predicted miRNAs out of

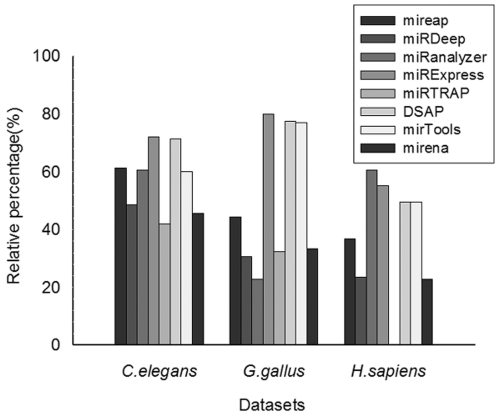


Figure 3.10: **Comparison of the sensitivity of various software tools, including miRDeep, miRanalyzer, MIReNA and miReap, when predicting known miRNAs.** The percentage of predicted miRNAs out of the total miRNAs in miRBase is shown. [180], by permission of Oxford University Press.

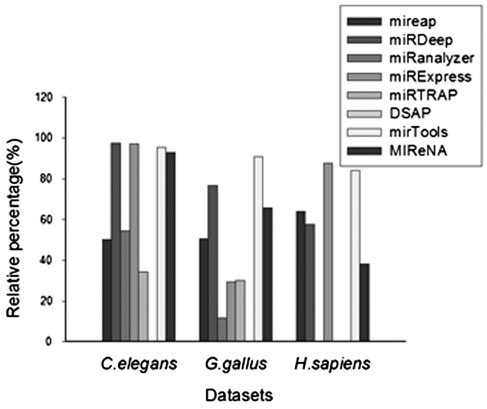


Figure 3.11: **Comparison of the accuracy of various software tools, including miRDeep, miRanalyzer, MIReNA and miReap, when predicting known miRNAs.** The percentage of predicted miRNAs in miRBase is compared with the total number of predicted miRNAs. [180], by permission of Oxford University Press.

the total number of miRNAs in the reference), the results of the tools on the three datasets were compared to a reference set comprised of either just the miRBase annotations or miRBase annotations together with sRNAs predicted using three or more software tools (extended reference dataset).

The number of predicted miRNAs compared to miRBase are presented in Figure 3.10, and compared to the extended reference dataset are presented in Table 3.8. When compared to miRBase, miRanalyzer had the highest success of 60.6% when predicting miRNAs for *H. sapiens* and a satisfactory high success with *C. elegans*. When compared to the extended reference dataset, miReap had the highest success of 59.85% for *H. sapiens*. These results suggest that different software tools were suited to predicting miRNAs in specific datasets.

The results regarding accuracy (percentage of predicted miRNAs out of the total number of predictions) were then calculated. When compared to miRBase (see Figure 3.11), miRDeep had the highest success of 97.41% when predicting *C. elegans*, the general ranking of the tools regarding accuracy being miRDeep, MIReNA, miRanalyzer and miReap. When compared to the extended reference dataset (see Table 3.9), miRanalyzer had the highest success of 100% in *H. sapi-*

	miRExpress	DSAP	*miRanalyzer*	mireap	mirTools	mirRDeep	MIReNA	miRTRAP
C.elegans	71.19	70.34	*59.32*	60.59	52.12	47.46	42.8	25
G.gallus	78.5	75.45	*76.17*	22.22	28.14	18.82	29.03	19.18
H.sapiens	54.52	48.87	*49.68*	59.85	23.26	16.72	18.82	0

Table 3.8: **Comparison of the sensitivity of various software tools, including miRDeep, miRanalyzer, MIReNA and miReap, when predicting known miRNAs, reported to an extended reference dataset.** Entries are shaded with black and white gradients, where black represents the highest percentage and white the lowest. [180], by permission of Oxford University Press.

	mirTools	miRExpress	MIReNA	miRanalyzer	mirRDeep	mireap	miRTRAP	DSAP
C.elegans	95.24	97.11	86.32	51.44	94.92	40.73	24.28	0.1
G.gallus	90.81	29.46	56.84	11.68	18.26	28.44	40.38	0.21
H.sapiens	84.36	87.66	31.28	100	39.06	35.82	NA	0.08

Table 3.9: **Comparison of the accuracy of various software tools, including miRDeep, miRanalyzer, MIReNA and miReap, when predicting known miRNAs, showing the percentage of miRNAs from an extended reference dataset compared with the total number of predictions.** Entries are shaded with black and white gradients, where black represents the highest percentage and white the lowest. [180], by permission of Oxford University Press.

ens, while miRDeep performs best on *C. elegans*. All tools seem to have high variance between datasets, thus, the performance accuracy when predicting miRNAs also depends on the dataset used.

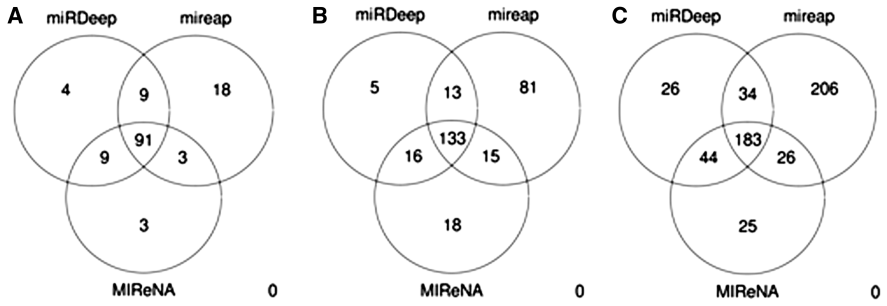


Figure 3.12: **Venn diagram of predicted known miRNAs by miRDeep, miReap and MIReNA.** (A) *C. elegans*, (B) *G. gallus* and (C) *H. sapiens*. [180], by permission of Oxford University Press.

An intersection of the predicted known miRNAs by miRDeep, miReap and MIReNA are shown as Venn diagrams in Figure 3.12 for the three organisms, respectively. The highest overlap was at the intersection of predicting *C. elegans*, whereas it was comparatively lower for *G. gallus* and *H. sapiens*. The prediction

of known miRNAs in *C. elegans* comprised more often of sequences that originated from genome locations close to each other (clusters of miRNAs) , whereas for *H. sapiens*, they originated from different locations on the genome (they were more discrete).

When predicting novel miRNAs, only the predictions that were detected by three or more tools were considered as true miRNAs. This can be however partially unreliable because of the differences in these tools results: one or two tools might predict a true miRNA, but it will not be considered genuine, because it was not predicted by a third method. Out of fourteen novel *G. gallus* miRNAs, miReap and miRDeep had the highest detection rate. For the fifteen novel *H. sapiens* predictions, MIReNA had the highest frequency and miRDeep was ranked second. Out of three novel *C. elegans* miRNAs predicted using three or more software tools, MIReNA had the highest percent, followed by miReap, miRDeep and miRanalyzer.

Based on this analysis, the study recommends MIReNA as the first choice for nematode and mammal datasets. Combinations of miReap, miRDeep and miRanalyzer can also be used with nematode, but miRDeep can also be used with mammals. In vertebrates, miReap was the first choice, while miRDeep can also be integrated in the analysis. miRDeep has better performance when predicting novel miRNAs for *C. elegans*, because it used *C. elegans* data for parameter estimation. However, MIReNA had better performance when predicting novel miRNAs with *C. elegans* and *H. sapiens* compared with miRDeep.

3.5 Discussion

We will now give a summary for the performance of each of the tools, based on the reviews above.

miRCat [4] has achieved a 91.2% sensitivity on its test dataset [1]. miRCat and miRDeep2 both show good percentages of known miRNAs detected, also having a considerable overlap of predictions. However, they both show high numbers of novel predictions (and possibly FP).

miRDeep2 [2] has generally a high sensitivity and specificity on a broad range

of animal datasets [2]. It has higher sensitivity than miRDeep* [175] and miRanalyzer [181], but has a slightly lower specificity than miRDeep* [175]. In a comparison with miRAnalyzer and miRDeep, miRDeep2 generated higher numbers of known miRNAs and the novel predictions were often from an annotated miRNA family [181].

miRPlant [3] has shown better rates of both sensitivity and specificity than miRDP in four organisms tested [3].

miReap needs less runtime to complete its analysis, requiring between 10 and 43 minutes, when compared to miRDeep and MIReNA, which took several days, in one case even more than one month [180]. It achieves the highest sensitivity in human datasets, compared to miRDeep and miRAnalyzer, but does not outperform the other tools constantly. miReap also presents the lowest specificity in all tested datasets, suggesting it might suffer from high rates of FP [180].

miRDeep [149] has high specificity, especially when run on *C. elegans* datasets (because it uses *C. elegans* as a model for parameter estimation) [180]. However, miRDeep was proved to miss some highly expressed miRNAs because of improper precursor excision [175, 181]. It has become inefficient in time consumption, taking ∼20% longer than miRDeep2 for mapping the reads to the target genome [181]. Moreover, it is 3 times slower overall, on average taking about 10.5 h to complete its analysis, compared to 2.87 h required by miRDeep2.

miRDeep* [175] has a better specificity than miRDeep2 and miRDeep, and also better sensitivity than miRDeep [175]. Analysing its results using Dicer mutant data, miRDeep* had a lower fold change average and a higher percent of negative fold change on novel predictions than miRDeep2, suggesting that its novel predictions are more likely to be TP [175].

miRDP [171] is more suitable for datasets with low sequencing depth, however it has a general low sensitivity and high number of novel predictions (FP) [178].

miRA [178] has a sensitivity of over 80% on its test dataset, performing better than miRDP and miR-Prefer. However, it has high number of novel predictions, lacking in specificity (FP) [178].

miR-Prefer [176] performs well in datasets with high sequencing depth, because it requires the star sequence to be present on the precursor as well [178]. It performs poorly in datasets with low sequencing depth [178]. It has a better sen-

sitivity than miRDP and low FP rate, most of the predictions corresponding to known miRNAs [176]. miR-Prefer achieves the fastest runtime and less memory when compared to miRDP, miRAnalyzer, miRDeep2, miRDeep* and MIReNA [176].

miRAuto [162] performs similarly to miRDP and miREvo for detecting known miRNAs, however it has the highest number of novel predictions, most of which are uncommon with the other tools [162]. This suggests it might allow FP amongst its predictions.

miRAnalyzer [170] has a high sensitivity ($\sim$ 90%) [175, 176]. This is achieved because it tries to detect annotated miRNAs that are saved in its own database [176]. It has 5 times more predictions than miRDeep*, having the lowest specificity rate ($\sim$ 20%) when compared to miRDeep*, miRDeep, miRDeep2 and MIReNA. Although a large proportion of the novel miRNAs have high number of reads, there is little overlap with the results of miRDeep and miRDeep2 [181], suggesting it is very likely that they are FP [175, 176]. It has also been reported that it might predict the hairpins poorly, varying in length and structure [181].

MIReNA [169] has shown low sensitivity and specificity rates [169, 175]. This might be explained by the fact that MIReNA is not very restrictive in its candidate selection at the beginning of the algorithm, but also because it considers several precursors for one candidate [169]. When compared to miRDeep, the results they generate are complementary, MIReNA detecting known miRNAs that miRDeep misses and vice versa [175]. In terms of resources used, it takes longer time and more memory to complete its tasks, being slower than miR-Prefer [176], miRDeep and miReap [180].

All the tools reviewed generate different results from the same input, showing preferences for certain datasets (possibly being over-trained and/or more suited for a specific organisms) [169, 180]. For example, miRDeep performs best on *C. elegans* data [180], miRDeep2 and miReap perform best on human data [180, 181]. In a study on plant datasets involving miRAuto, miRDP and miREvo, out of a total of 206 novel predictions detected by the three tools, only 5 sequences (2.4 %) were found in the intersection of all three of the results [162], suggesting that these tools look for distinct criteria that do not generate the same output.

In a study on miRDeep, miReap and MIReNA, there was a high percentage of overlap in *C. elegans*, which significantly decreased in *H. sapiens* and *G. gallus* [180], showing inconsistency of performance across different organisms.

This suggests that further improvements to such software are still required, to achieve a decrease of the false positive and false negatives rates and consistency across species. At the moment, when conducting a miRNA prediction project, it is recommended to use multiple miRNA prediction tools, then analyse their combined results, which requires more effort and resources. Also, these tools vary from plants to animals, but also from organism to organism, making it difficult for the user to select the most appropriate tool depending on the case. Therefore, there is the need for a single tool that can provide reliable results, for a wide range of organisms. For this reason we decided to develop miRCat2, presenting a new miRNA detection algorithm, which achieves the above mentioned goals.

Considering the performance of each tool reviewed above, but also based on popularity and frequency of usage, we selected the following tools to benchmark against our new algorithm: miRCat [4], miRDeep2 [2], miRPlant [3] and miReap. We have selected miRCat, because it presents good sensitivity and specificity rates, and it is suitable both for animals and plants, but also because we want to improve on its initial algorithm. miRDeep2 is at the moment one of the most popular tools for animal miRNA prediction, having high percentages for sensitivity and specificity. miRPlant is the equivalent of miRDeep2 for plant data, as it is developed on its algorithm, and shows better performance than miRDP, therefore miRPlant is more suitable for our testing. miReap is widely used both in animal and plants, having good results in some organisms, although it can lack in consistency. By developing miRCat2, we want to improve on their results, by achieving an increased accuracy over these tools.

3.6 Summary

In this chapter we gave an overview of the most commonly used miRNA detection tools from HTS datasets. We presented the key factors of their algorithms, then compared their performance, gathering information from a suite of reviews. Based on these factors, we evaluated the strengths and weaknesses of each algorithm

and we have chosen to benchmark our new tool, miRCat2, against miRCat [4], miRDeep2 [2], miRPlant [3] and miReap. We will give a complete analysis and details about this benchmarking in Chapter 5.

Chapter 4

Developing and testing the miRCat2 algorithm

Part of the work presented in this chapter is submitted as part of the manuscript "miRCat2: Accurate prediction of plant and animal microRNAs from next-generation sequencing datasets", Claudia Paicu, Irina Mohorianu, Matthew Stocks, Ping Xu, Aurore Coince, Martina Billmeier, Tamas Dalmay, Vincent Moulton and Simon Moxon.

4.1 Summary

As previously mentioned, many miRNA prediction methods, such as miRCat [4] and miRDeep2 [2], were designed when sequencing depth was typically orders of magnitude smaller than the output from the current generation of sequencers. As the size of HTS datasets are rapidly increasing [150, 151], the older algorithms struggle in terms of memory consumption and run time. The reviews we considered in the last chapter also show that many older methods suffer from high false positive and false negative rates and lack of consistency across species [180–182]. These are indicators that new algorithms for miRNA prediction are required.

In this chapter we begin by presenting a new algorithm, miRCat2, for identifying new miRNAs from HTS data in both plants and animals, which addresses

some of the issues mentioned above. We then give technical details about its implementation, as we have incorporated miRCat2 into the UEA sRNA Workbench [4]. Next we present the methods used for testing the miRCat2 algorithm and benchmarking its results against other commonly used software. We also describe several methods of computationally assessing the validity of novel miRNA predictions. The results for these methods are presented in the next chapter.

4.2 miRCat2 algorithm

We begin by describing the miRCat2 algorithm. In Figure 4.1 we give an overview of its key features. After mapping the reads to the genome, the algorithm first selects reads based on abundance significance (read counts), then filters based on read alignment patterns and secondary structure of the putative pre-miRNA hairpin are applied in the subsequent steps. We now give a detailed description of the algorithm.

4.2.1 Candidate selection

As the size of HTS datasets is rapidly increasing [151], previous methods, such as miRCat, can have systematic issues choosing miRNA candidates. For example, miRCat groups sequences on proximity on the reference genome sequence (all sequences within x nt from each other form a group), and selects one candidate from each group (see Chapter 3, Section 3.3.1). miRDeep2 has a similar approach to miRCat, looking 70 nts upstream and downstream the genome of every read and selecting the most abundant sRNA in the area (see Chapter 3, Section 3.3.2).

In larger datasets, these methods become inefficient, firstly, because higher coverage leads to a greater number of candidate groups. Many groups will not contain a *bona fide* miRNA, but will be tested anyway, greatly increasing computational runtime. Secondly, higher coverage leads to more low-abundance noise in sRNA sequencing datasets which may be a signal of random RNA degradation. As the number of genome mapping reads increases it has the effect of lengthening the miRCat group length and can join together two unrelated groups. More than one genuine miRNA can be located in a single miRCat group, but only one

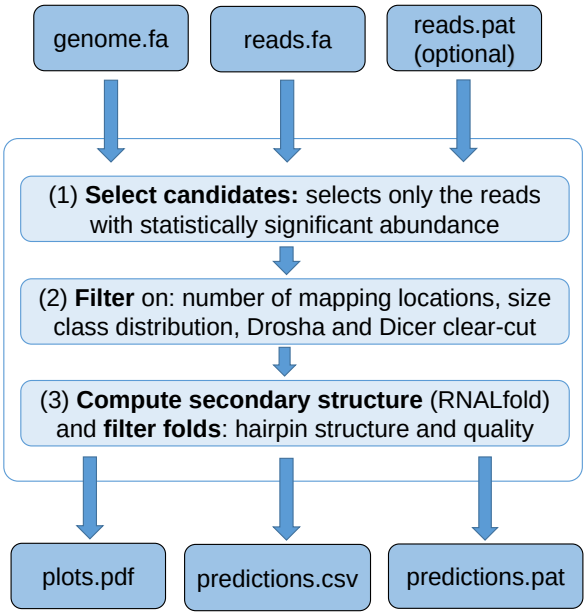


Figure 4.1: **Workflow of the miRCat2 algorithm.** The inner light-blue boxes represent processes, the outer dark-blue boxes are input and output files. The file formats are: .fa, fasta; .pat, PatMaN output; .csv, csv spreadsheet. These steps are explained in the following sections.

miRNA will be predicted from each group selected. In this way, genuine miRNAs can be overlooked. The same issue is found in the miRDeep2 upstream or downstream lookup for miRNA candidates.

miRCat2 implements a method of candidate selection specially designed to deal with high depth datasets. As sequencing depth increases, degradation products may obscure miRNA peaks (see Figure 4.2). To cope with this, we focus on selecting all the peaks at any given genomic location, while discounting reads that are at or below a background level that we compute from the data.

There are multiple peak calling algorithms which focus on detecting peaks in ChIP-seq data (protein interactions with DNA) [204–215]. These methods generally classify into three categories. The first type [204, 205] involves taking a moving average of sequence reads within a fixed or variable-width window and scanning the window through the entire genome, then a randomization scheme is used to determine the null distribution, to estimate the false discovery rate. The second class of algorithms [209, 216] use the same approach for finding peaks

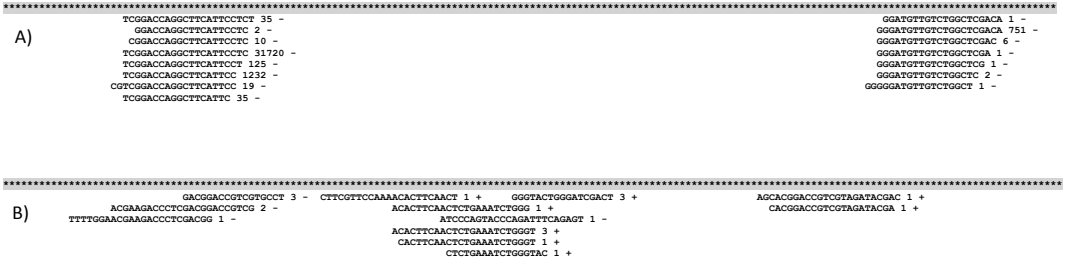


Figure 4.2: **Distribution of reads for a known miRNA locus A) and a random locus on the genome with incident degradation reads B)**. For each incident read we present, on the right, its abundance (read count),and the matching strand (+/-). A) Distribution of reads for sly-MIR166c (*S. lycopersicum*), on chromosome 1, positions 84381885 - 84382061. This shows the expected miRNA locus pattern, with a characteristic two-peak alignment corresponding to the 5'/3' miRNAs. B) Random distribution of reads for *S. lycopersicum*, on chromosome 1, positions 2076029 - 2076206. The lack of location, size class or abundance specificity, corroborated with the lack of a hairpin-like secondary structure, indicates that this alignment doesn't correspond to a miRNA locus.

but then make inferences based on a probabilistic model in order to assess the significance of the peaks, usually using a Poisson probability model. The third type [207] uses fitting Hidden Markov models, a more complicated approach then the previous ones. For more generic peak detection algorithms, the following reviews can be very helpful [217–220]. However, these algorithms are not specific for sRNA data and do not take into consideration any of the miRNA-specific features.

Therefore we decided to implement our own way of detecting abundance peaks in genome-aligned sRNA data. It is known that miRNAs and their complementary miRNA* sequence generally have significantly higher abundances in HTS datasets than non-miRNAs [35]. When aligning miRNA reads from an HTS experiment back to the pre-miRNA locus we see characteristic peaks forming, corresponding to the 5 and 3 miRNA sequences (Figure 4.2, A)). We can use this information to select a restricted group of sequences as candidates, on which further computation is performed.

By implementing a method to detect "peaks" of reads, we have improved the accuracy of the results and also eliminated the need for an additional parameter which sets a minimum required abundance for the predictions. Results showing

the accuracy of miRCat2 are presented in Chapter 5 in more detail.

Selecting the candidate miRNA loci

To identify putative miRNA loci based on "peaks" of reads (sequences with read counts above the background level), we use the following procedure (Figure 4.3).

a) The genome is split into windows of size l_w nts (default 300 nts for animals, 500 for plants), consecutive windows having an overlap of l_o nts (100 nts) (Figure 4.3.(A));

We have chosen these default values using empirical observation, applying the following reasoning. By having a window of 300/500 nts we are certain that the whole miRNA precursor can be contained on that window, considering the miRBase precursors have a mean length of 70 nts in animals and 200 nts in plants [5]. However, we need to make sure that we capture the context of the reads where the sRNA is located, therefore we need the window to be larger than just the precursor. By having an overlap of 100 nts we make sure that adjacent windows are not isolated, but they influence each other, to better define the context of sRNAs. The windows without any coverage are discarded.

b) Each window is split into subwindows of size l_{sw} (20 nt) and the mapped reads are assigned to subwindows based on location (to which subwindow they have most nts aligned, see Figure 4.3.(B)).

Each subwindow should have the length $13 < l_{sw} < 25$, so it can cover at least a half of the longer miRNAs, but not exceed the length of a miRNA. As sequences tend to overlap, by empirical observation, we have found that the most effective value for this parameter is 20.

c) Each window is compared with a simulated perfect random uniform distribution (RUD) on genome location, using the Kullback-Leibler divergence (KLD) [221] (Figure 4.3.(C)).

Generally, for any two probability distributions Q and P, the KLD of Q from P, denoted $D_{\mathrm{KL}}(P\|Q)$, is a measure of the information lost when Q is used to

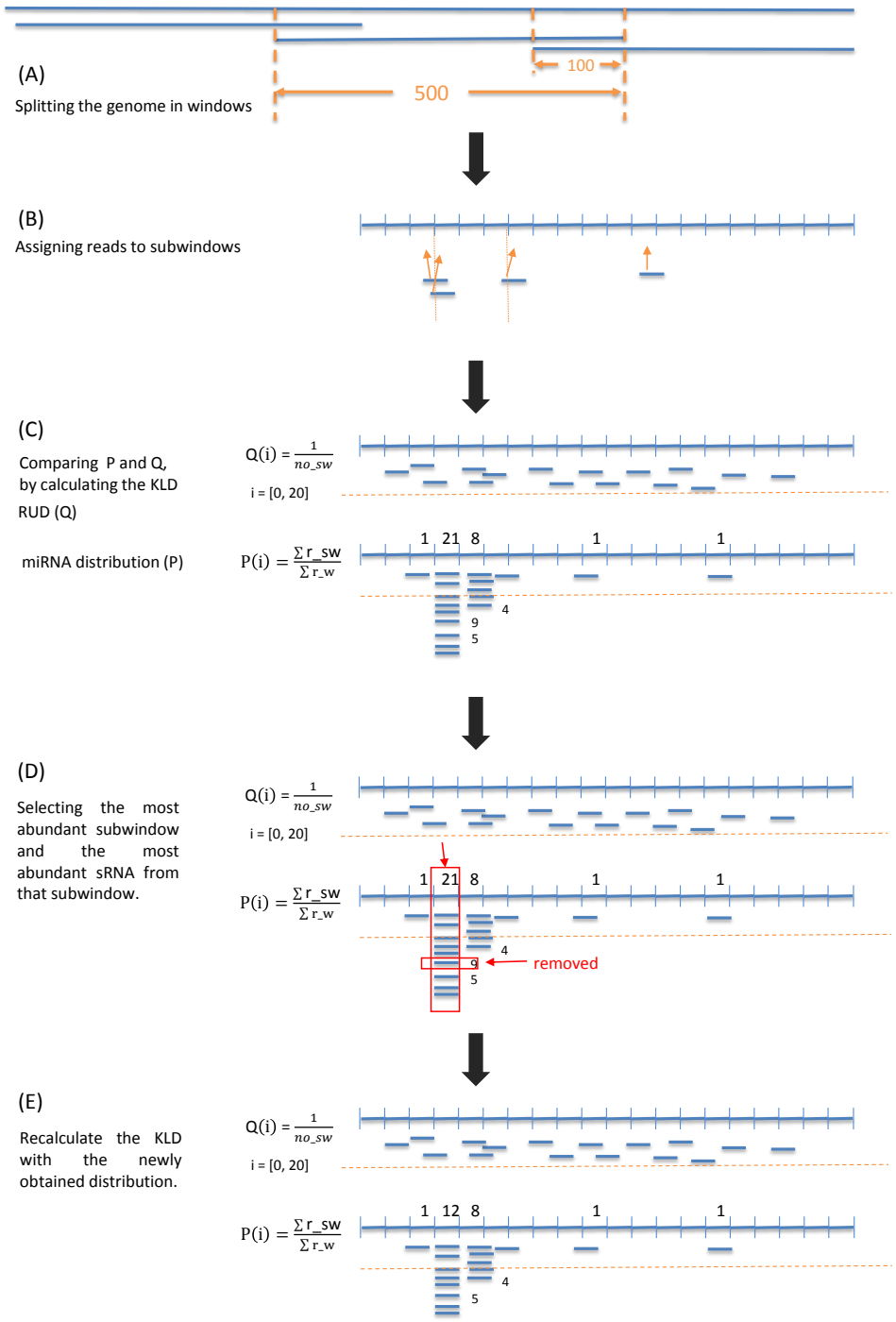


Figure 4.3: **Selection of candidate miRNA loci step by step in miRCat2.** (A) Splitting the genome in windows; (B) Assigning reads to subwindows based on location; (C) Comparing the distribution of reads, P, with and a RUD, Q; (D) selecting peak as miRNA candidate, removing it; (E) Recalculating the KLD on newly obtained distribution.

approximate P. It is computed using:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i |\ln\left(\frac{P(i)}{Q(i)}\right)|P(i) \tag{4.1}$$

where i is the index of an observation. To use it for sRNA data, this distance can provide statistical evidence concerning whether a distribution of reads on a genome window has a random uniform distribution or not. If not, then the tested distribution of reads could contain peaks. This entropy has been successfully used before for sRNA analysis, for determining a threshold above which the sRNA dataset does not have a strand bias [222].

In the algorithm of miRCat2, Q represents the simulated RUD, P represents the distribution of reads on the current window and each subwindow i is an observation. The probabilities for each subwindow are calculated from the read abundances: $P_{(i)} = \frac{\sum r_{sw}}{\sum r_w}$, where r_{sw} represents the abundance of the reads mapping to the subwindow and r_w represents the abundance of the reads mapping to the window, after a default offset of 1 has been added to each subwindow (empirically determined), to avoid allowing lowly expressed reads to look like they are peaks. The probability for the RUD is calculated using the following equation: $Q_{(i)} = \frac{1}{no_{sw}}$, where no_{sw} represents the total number of subwindows contained in a window;
d) The closer the score of the KLD is to 0, the closer the distribution of reads on the tested window is to a uniform distribution, which implies that it does not contain a peak. If the distribution is a RUD, then it is unlikely that a miRNA has evidence of expression at that genomic location and the window is discarded.

If the KLD is greater than a threshold (*rud_val*), then the current window contains at least one peak (the method can detect multiple peaks). In this case, the subwindow with the greatest probability is identified and the most abundant sRNA is selected (Figure 4.3.(D)). The KLD is applied again on a restricted area around this sRNA (*plateau_range*) to avoid detecting a peak that is actually a plateau (multiple neighbouring subwindows that are all highly expressed, see Figure 4.4). If this filter is passed, the sRNA is removed from the distribution and saved as a miRNA candidate for further investigation;

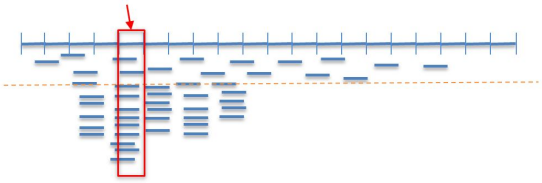


Figure 4.4: **Distribution of sRNA reads that would cause a peak detection that is actually a plateau.**

e) The KLD is recalculated with this newly obtained distribution (Figure 4.3.(E)). If the new KLD is still greater than the threshold, steps (c) to (e) are repeated until we reach an RUD (no more peaks). All removed sRNAs are miRNA candidates and are analysed in the subsequent filtering steps of the algorithm, which we describe in the next section.

Example of calculating steps (c) to (e).

To calculate the KLD on the reads distribution from Figure 4.3.(C), consider Q to be the RUD, P to be the distribution of reads and i the index of the subwindow. For simplicity, all the reads that do not have a number next to them have an abundance of 1. Each subwindow has its total abundance above it (sum of reads abundances incident to the subwindow).

We consider Q to be a perfect RUD, thus, the probability for each subwindow for Q will be the same: $Q(i) = \frac{1}{20} = 0.05$ (because we have 20 subwindows).

For P, the total abundance on the window is 32, and after we add an offset of 1 for each subwindow, we have a total abundance of 52. For the 4th subwindow, the probability is $P(4) = \frac{21}{52} = 0.403$.

After calculating each $P(i)$, we apply the KLD formula (equation 4.1) and we obtain $D_{\mathrm{KL}}(P\|Q) = 1.6912 > rud\_val$ (default value of $rud\_val = 1.23$, empirically observed), therefore P must contain a peak. In Figure 4.3.(D), the peak's location is identified (subwindow $i = 4$) and the most abundant sRNA from that subwindow is selected and removed (sRNA with abundance 9). This sRNA is saved separately for further computations.

After the sRNA removal, a new read distribution is obtained, as seen in Figure 4.3.(E). Each $P(i)$ is recalculated ($P(4) = \frac{12}{43} = 0.27$) and the KLD is computed

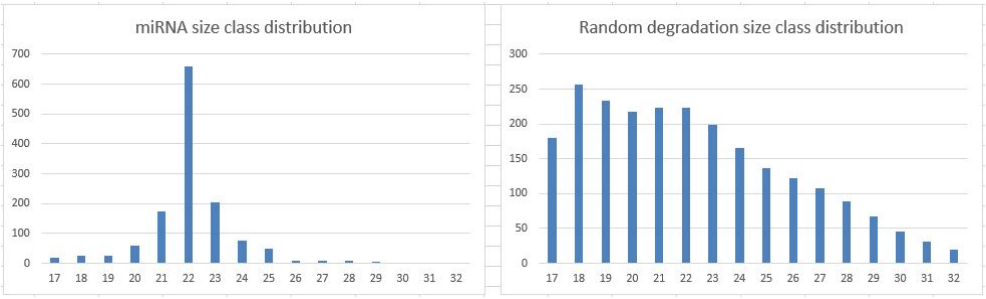


Figure 4.5: **Example of a miRNA size class distribution vs. a random degradation size class distribution.** Y-axis represents total counts, X-axis represents the size of the sRNA plotted.

again, the new value being $D_{\mathrm{KL}}(P\|Q) = 1.2115 < rud\_val$. In this case, there are no more peaks and the algorithm would start processing the next window on the genome. If the $D_{\mathrm{KL}}(P\|Q)$ would have been $> rud\_val$, steps (c) to (e) would have been reapplied for this distribution.

4.2.2 Filtering the sequences

After miRNA candidates are selected, potential false-positive predictions are excluded from down-stream analysis using a rule-based approach.

First, we discard the sequences that map to the genome more than *repeats* times as high-confidence miRNAs [5] are unlikely to be derived from repetitive regions of the genome (this parameter is user configurable, with a default value of 25).

Second, a **size class distribution filter** is applied, allowing us to focus on reads between 21 to 23 nt, which is the expected miRNA range. To check if the most enriched size class on the window containing the miRNA candidate is within the range, we compute the KLD on size classes, comparing the sRNA size class distribution (P) to a RUD on all size classes (Q) [184]. This is biologically appropriate because miRNAs tend to have significantly greater abundance than near-by sRNAs in their region [35]. For example, in Figure 4.5 the size class distribution of a miRNA is compared to a random degradation size class distribution (data simulated by assigning equal probabilities to all subwindows in the RUD).

The sequences contributing to the sRNA size class distribution are all the reads incident to the potential putative miRNA precursor (located $\pm max\_fold\_len$ on each side of the miRNA candidate). A size class can take values from min_size (default 16) to max_size (35), as sequences originating from sRNA data should have lengths between these bounds. Here also, the KLD is calculated after adding a default offset of 1 to each size class (empirically determined), to deal with lowly expressed reads.

If the KLD result is $> rud\_val$, then the size class distribution is different from random and contains a peak. We investigate whether the most abundant size class falls between 21 to 23 nts (the peak consists of sequences from the miRNA size range), otherwise the sRNA locus is discarded. Since a small set of annotated miRNAs in miRBase fall outside of this size range these values are configurable (min_len, max_len). If this criteria is met, the sRNA is saved for further analysis, otherwise it is discarded.

Third, to check whether the candidates have a miRNA-like alignment of incident reads, we also apply a filter that selects only sequences with **evidence of precise processing of the pre-miRNAs** by the miRNA biogenesis machinery, Drosha (animals) and Dicer (plants and animals) [15, 16, 21], as described in Chapter 2. Drosha processing excises the pre-miRNA hairpin from the primary transcript with high precision, Dicer then cleaves the hairpin loop giving rise the mature miRNA duplex. This should reflect in the alignment of sequences as the presence of one or two peaks corresponding to the miRNA/miRNA*. This filtering step ensures that the majority of reads aligned to the miRNA/miRNA* location have a high overlap (are variants of each other), and have the same genomic orientation. The distribution of reads of a genuine miRNA should have a similar shape to that shown in Figure 4.2,A) compared to a locus generated from random RNA degradation (Figure 4.2, B)).

To implement this filter, we define a cluster as all sequences that map to the same genomic location, having the start and the end of the mapping position within $clear\_cut$ nts of each other. We chose $clear\_cut$ to be 3 nts to account for the isomiRs that may also be generated during the miRNA biogenesis (sequences that have variations of a few nts with respect to the reference miRNA sequence)

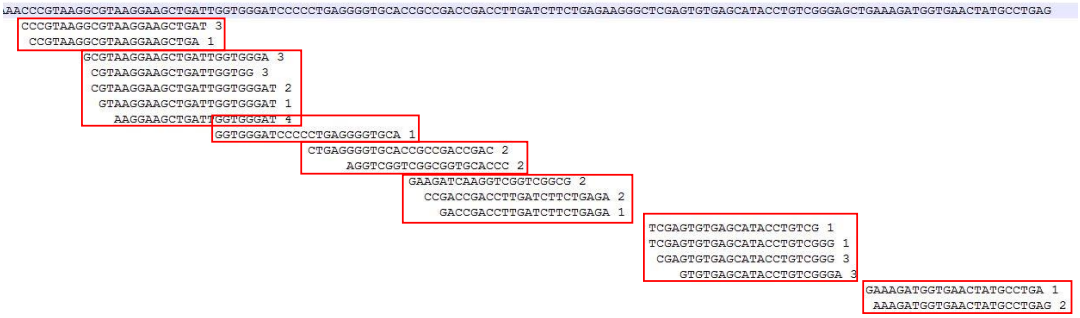


Figure 4.6: **Example of alignment of reads grouped on clusters computed on *S. lycopersicum* data.** Each red square represents a separate cluster.

[223]. isomiRs should contribute to the expression level of the miRNA, as they are often active in the cell, having the same roles as the putative miRNA [223].

The rules we use for a sRNA sequence s (with sRNA beginning position b_s and ending position e_s) to be considered to belong to a cluster C (with cluster beginning position b_C and ending position e_C) are:

- if $b_s \geq b_C$ and $e_s \leq e_C$, then the sRNA alignment is completely inside the cluster boundaries, therefore $s \in C$;
- if $b_C \geq b_s$ and $e_C \leq e_s$, then the sRNA alignment is completely covering the cluster location, therefore $s \in C$;
- if $(b_s - b_C \leq clear\_cut$ and $b_s - b_C \geq 0)$ or $(e_s - e_C \leq clear\_cut$ and $e_s - e_C \geq 0)$, then the sRNA alignment is with $clear\_cut$ nts to the right or left from the cluster location, therefore $s \in C$.
- we define mid_s to be $mid_s = \frac{(e_s - b_s)}{2} + b_s$ and mid_C to be $mid_C = \frac{(e_C - b_C)}{2} + b_C$. If $|mid_C - mid_s| \leq clear\_cut$, then the middle of the sRNA alignment is with $clear\_cut$ nts to the right or left from the middle of the cluster location, therefore $s \in C$;
- otherwise $s \notin C$.

Using the rules stated above, we identify all clusters on the window corresponding to each selected miRNA candidate, s. An example of such clustering is presented in Figure 4.6. Next, to evaluate the existence of a precise excision (e.g. resulting from Drosha and/or Dicer cleavage), we use the following criteria:

- on the cluster containing s, if the sum of the abundances of all sequences with same start and end positions ($\pm clear\_cut$ nts) as s represent $clear\_cut\_percent\%$

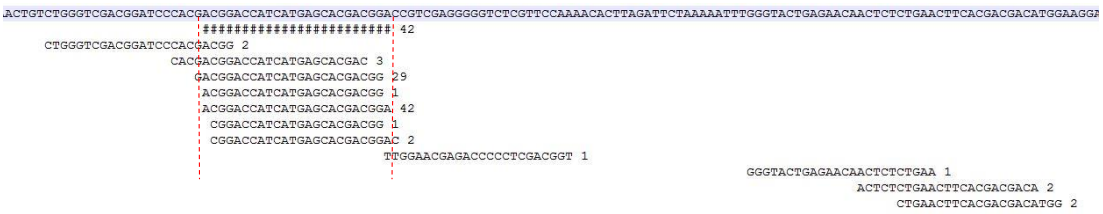


Figure 4.7: **Distribution of sequences with a miRNA-like structure on *S. lycopersicum* data**. The first sequence, encoded with #, is the candidate sRNA, (s). The red dotted lines delimit the start and end position of the candidate sRNA. The numbers on the right of each sequence represent their read abundance.

of the total abundance of the cluster (95 % in animals, 92 % in plants, chosen by empirical observation), then s is kept for subsequent analysis; otherwise, it is discarded;

- on the adjacent clusters to the cluster containing s, if the sum of the abundances of all sRNAs from adjacent clusters that overlap with s with more than *clear_cut* nts represents less than *overlap_percent*% of the total abundance of the s cluster (5%), then s is kept for further analysis; otherwise, it is discarded.

Example of deciding if a sRNA allignemnt respects the precise processing of Dicer and/or Drosha.

Consider the distribution of reads for the candidate from Figure 4.7, that has been grouped into clusters, such that the second cluster C contains the miRNA candidate sequence s. The total abundance of cluster C is 79, and there is one sequence that does not respect the precise Dicer and/or Drosha processing: the first top sequence from C (ACGACGGACCATCATGAGCACGACG , abundance 3), because its start is 4 nts smaller than the start of s. All the other sequences in C respect the criteria. We now calculate the percentage of the sequences that respect the criteria to be $\frac{76}{79} = 96.2\% \geq 95\%$, so s respects the first part of the criteria.

Then, we calculate the sum of all sRNAs from adjacent clusters that overlap with s with more than 3 nts. In the cluster on the left of C, we identify the sequence that fits this rule (CTGGGTCGACGGATCCCACGACGG , abundance 2). The cluster on the right does not have a significant overlap with s. We now

calculate the percentage of sequences from overlapping adjacent clusters to be $\frac{2}{79} = 2.53\% \leq 5\%$, so s respects the second part of the criteria. Therefore, s is saved for further analysis.

4.2.3 Computing the secondary structure

Most existing methods for miRNA prediction extract a fixed, arbitrary flanking region containing the miRNA candidate and fold it using RNA secondary structure prediction tools (such as RNAFold [154]) to identify a suitable hairpin-like precursor [1–3]. However, this approach is highly dependent on the length of the flanking region; therefore choosing an optimal length is a critical step to predict the correct secondary structure.

To address this, we employ RNALfold [154], previously used by miR-PREFeR [176] and miRA [178] (a modified version of RNALfold), which folds a large window efficiently, giving all possible structures contained within that region.

RNALfold is more efficient than RNAFold on larger sequences, as its algorithm has the time complexity of $O(n * l^2)$ compared to $O(n^3)$ for RNAFold, where n is the length of the sequence and l is the maximum length of the folded subsequence.

RNALfold outputs as result a list containing all possible local secondary structures within the selected region, in dot-bracket notation,and their corresponding minimum free energies (MFE, in kcal/mol). To be able to compare the stability of two folds of differing lengths, we calculate the adjusted minimum free energy per 100 nts (aMFE) for each secondary structure, as follows: $aMFE = \frac{MFE}{fold\_len} * 100$.

To detect the most appropriate secondary structure in miRCat2, we consider a window of max_fold_len nts (100 nts for animals, 250 nts for plants) on each side of the miRNA candidate and use it as the input, ensuring that the window is of sufficient length to capture the pre-miRNA structure on either side of the miRNA candidate.

The secondary structures that contain the miRNA candidate (see Figure 4.8) are kept for subsequent filtering based on the properties of miRNA precursors (see Figure 2.9 for miRNA precursor features).

The filters on the secondary structure include checking for:

- the minimum fold length: the folds should not be shorter than min_fold_len,

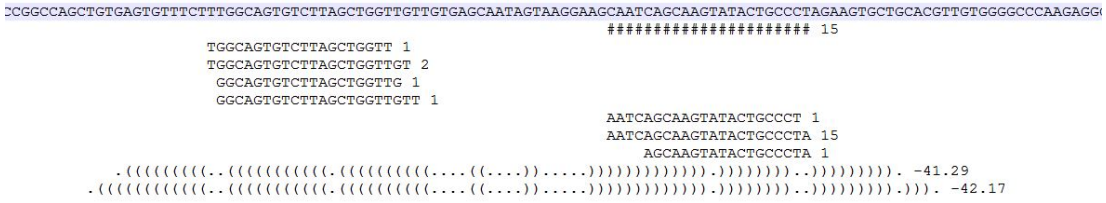


Figure 4.8: **RNALFold results that overlap with the position of miRNA hsa-mir-34a and passed all precursors filters (*H. sapiens* data)**. The folds are represented in the dot-bracket notation, together with their calculated aMFE.

to ensure there is enough space for the miRNA/miRNA* duplex and the loop [15, 114];

- maximum aMFE: miRNA precursors are highly stable, therefore it is important that their secondary structure has a low aMFE. A valid fold should have an aMFE lower than max_amfe [65].
- a hairpin-like structure: the fold should present a duplex, with the miRNA position aligned to the miRNA* (with a 3' overhang of 2 nts) and loop [15]. If there are sequences corresponding to the miRNA* position, it is checked that they align where expected.
- number of consecutive unpaired nts in the miRNA position: there should not be more than $gaps\_miRNA$ consecutively unpaired nucleotides on the location of the miRNA [15, 65] .
- the existence of multiple bulges in the loop area: in animals, the loop should be represented as a single bulge. In plants, however, as the hairpins are longer and the structures are more complex, we allow that loops contain up to *no_loops* bulges [53, 111–113].
- the existence of loops inside of miRNA and miRNA*: the miRNA and miRNA* sequence should never fold onto itself. If there is a loop inside one of these areas, the hairpin is not valid [15].
- percentage of nts paired on fold: there should be a minimum of $min\_pared\_perc\%$ nts that are paired on the hairpin. This is particularly important for longer precursors (e.g. in plants), where areas outside the miRNA/miRNA* duplex should also pair, to insure the stability of the secondary structure [114].
- number of nts paired on miRNA and miRNA*: the duplex should have a minimum of min_pared_nucl nts paired [15, 32].

- orientation of reads on fold: a minimum of $min\_orientation\%$ from all the reads that align to the hairpin location should originate from the same strand (sense or antisense) [47, 48].
- fuzzy alignment of reads on fold: checks that reads correspond to a clear product of miRNA biogenesis. The hairpin can be divided into the following regions: start of hairpin to miRNA, miRNA, miRNA to loop, loop, loop to miRNA*, miRNA*, miRNA* to end of hairpin. Although we checked for precise cutting by Dicer and/or Drosha around the miRNA/miRNA* position, it does not ensure that the other areas are clearly delimited. Each read is tested to have the start and the end within one of the above categories ($\pm clear\_cut$ nts if limits are exceeded). This filter ensures that at least $fuzzy\%$ of all reads on the hairpin fall clearly into one category.
- randfold p-value: optionally, randfold [155] can be applied, and the hairpin should have a maximum p-value of $pVal$ (default 0.05; a smaller p-value means that the hairpin has pre-miRNA properties with more confidence). Because randfold takes a significant amount of time to run and the miRCat2 algorithm classifies miRNA precursors with high accuracy disregarding this filter, we decided to let this be an optional step, set by the user.

Full details about each of these parameters (whether they are configurable or not, default values for animals and plants and the justification for the chosen values) are listed in Appendix A.

If there is more than one subwindow that passes all filters, the one with the lowest aMFE is considered to be the true precursor. This guarantees that the most stable secondary structure is chosen for the candidate, as true miRNA precursors are very stable [35].

Scoring the predicted precursors

miRCat2 computes a score for the output precursor, which is calculated based on the miRDeep2 model [149] (see Chapter 3, Section 3.3.2 for details). miRDeep2 calculates the basic score by fitting the values from the MFE of the hairpin and the total abundance of the miRNA, miRNA* and loop into a Gumbel distribution [187], using parameters generated from *C. elegans* real miRNA precursors. If the miRNA* sequence is missing, miRDeep2 then reduces the basic score to 0. To

this value, miRDeep2 next adds a series of hard-coded values, based on a series of criteria.

We have used the same parameters as miRDeep2 to characterise the Gumbel distribution describing a miRNA precursor, and we calculate the score based on it. If the miRNA* is missing, however, instead of using a value of 0, we reduce it to a tenth of the score's initial value. This is necessary because a prediction without a complementary precursor sequence is less likely to be a true miRNA; nevertheless, there are cases where only one miRNA is sequenced, while the miRNA* is not present in the dataset. Therefore we do not reduce the score to 0, but only decrease its value. This way, the score remains proportional to the MFE and total abundance of the hairpin. We do not add any further fixed, predefined values.

The miRCat2 score represents the probabilistic confidence that the prediction is a true miRNA precursor, but it does not influence the output of the method. When filtering the results based on this score as a post-processing step (at a default cut-off value of 5, empirically determined), we observed that miRCat2 performs well irrespective of this filtering (for more details see Chapter 5). This suggests that the core algorithm is robust and therefore we feel that it is unnecessary to implement this as a post-processing filtering field. It can rather be used as a ranking criteria for the results, a higher score meaning the prediction has a higher probability of being a true miRNA.

4.3 Implementation

The miRCat2 algorithm has been incorporated into the UEA small RNA Workbench [4] (with the help of Dr. Matthew Stocks) and is written in Java, version 1.8+; for optimal results, we recommend using the latest, stable, Java version. The Workbench also includes helper tools (Adapter Removal, Filter, Sequence Alignment), analysis tools (miRCat [1], miRProf, SiLoCo [1], ta-siRNA prediction [1], PAREsnip [183], CoLide [184]) and visualisation tools (RNA/Folding Annotation, VisSR). The Workbench and miRCat2 can be run on any operating system running Java (Windows, Linux, Mac OSX).

miRCat2 can be executed either through the user-friendly interface or from

the command line. Two sets of default parameters are provided, one for animals and one for plants, although the user can adjust these parameters (see Appendix A). The default parameters were set according to rules generally applicable to the annotated miRNAs from miRBase [5], for the specific kingdom. A list of all parameters and their default values is given in Appendix A.

miRCat2 requires as input a reference genome and a set of sRNA sequencing data (FASTA format, non redundant, adaptors trimmed; the files can be processed from FASTQ to the necessary format using the UEA sRNA Workbench [4]). It automatically maps the sequences to the reference genome using PatMaN [153], full length, with 0 gaps and no mismatches to create a mapped file, which can be used in later runs. The sequences that do not map are discarded. The user can optionally give the mapped sequences in PatMaN format as input, to speed up the processing: if a PatMaN file is given as input, the mapping step is skipped.

The output of miRCat2 is automatically saved as:

- a PatMaN file, containing the miRNA coordinates predicted
- a csv file, containing additional information about the miRNA*, hairpin and existing annotation of the sequence. The columns displayed are: "Precursor Score, Chromosome, Sequence, Abundance, Start, End, Strand, Mismatches, Hairpin Sequence, Hairpin Dot-Bracket, Hairpin Start, Hairpin End, Hairpin MFE, Hairpin aMFE, p-Value, Star Sequence, Star Abundance, Star Start, Star End, miRBase Precursor".
- a PDF file including, for every prediction, coverage plots of mapped abundances (see Figure 4.9, A)).
- a text file containing, for every prediction, the read alignments on the precursor (see Figure 4.9, C)).

Additionally, the user can also export parts of the results to different file formats: mature miRNA sequences to FASTA, precursor sequences to FASTA, precursor secondary structure to png (see Figure 4.9, B)).

miRCat2 uses RNALfold from the ViennaRNA package for folding of the secondary structure [154] and randfold for optionally calculating the statistical significance of the precursor structure [155]. All dependencies are included in the download package and no extra installation is required. The code can be

downloaded from http://srna-workbench.cmp.uea.ac.uk/downloadspage/, where users can also find the documentation and example files.

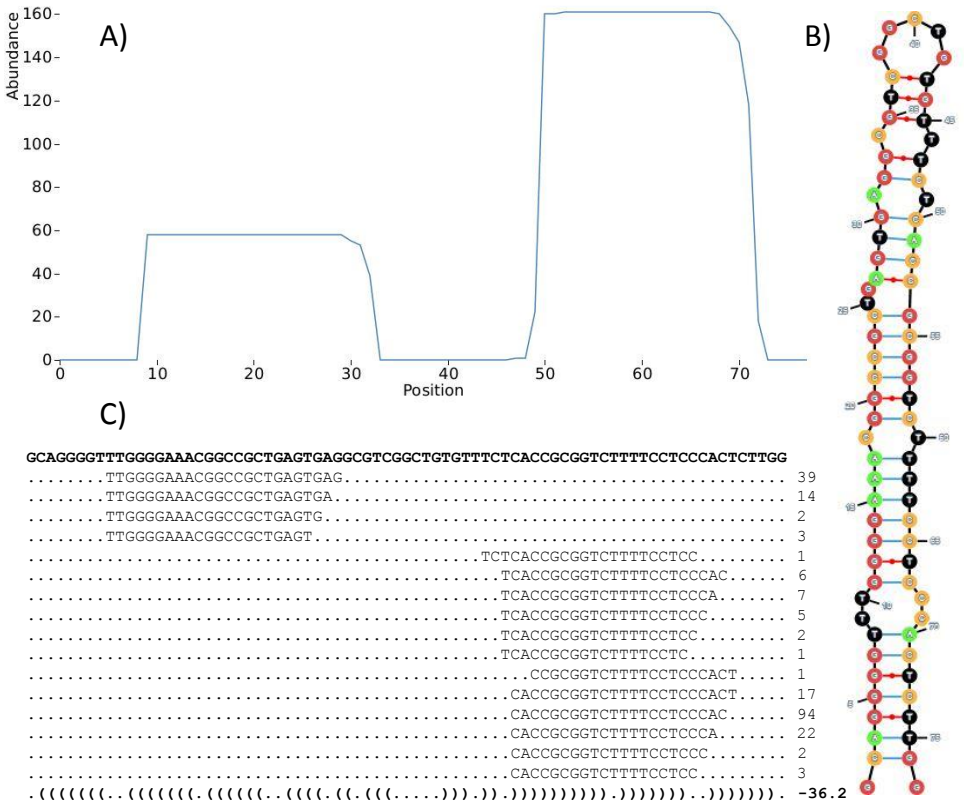


Figure 4.9: **Output of miRCat2 for a predicted sequence corresponding to hsa-mir-2110 (chromosome 10), depicting A) precursor coverage plots, B) precursor secondary structure and C) alignment of incident reads.** A) Precursor coverage plots, showing the total abundance of mapped reads for each nucleotide. B) Precursor secondary structure, colour-coded for each nucleotide type (A - green, C - orange, G - red, T - black). C) Alignment of incident reads on the precursor; the numbers of the right represent the read abundance. The last line presents the secondary structure in dot-bracket notation, together with its MFE.

4.4 Performance assessment methods

In this section we describe the methods we used to test miRCat2 and also to assess its performance. We have applied the below explained methods on miRCat2 and compared it to miRCat [1], miRDeep2 [2], miRPlant [3] and miReap (http://mireap.sourceforge.net/). Results of the tests conducted are described in the next chapter.

4.4.1 Data

To assess miRCat2, we ran it on multiple organisms and benchmarked the results against other commonly used miRNA detection tools, miRCat (version srna-workbenchV3.2), miRDeep2 (version miRDeep2.0.0.7), miRPlant (version miRPlant_V5) and miReap (version mireap_0.2). The datasets for each organism considered were taken from the following sources: *Danio rerio* [224], *Homo sapiens* [91, 225–229], *Mus musculus* [230–235], *Caenorhabditis elegans* [236], *Drosophila melanogaster* [237], *Heliconius melpomene* [238], *Xenopus laevis* [239] (animal datasets), *Solanum lycopersicum* [240, 241], *Glycine max* [242] and *Arabidopsis thaliana* [243] (plant datasets). We have downloaded these datasets from GEO [158] or SRA [168] databases. We also generated an *A. thaliana* dataset as described below. Information about the genomes used, accession numbers of small RNA datasets, trimmed adapter sequences and number of reads in each dataset can be found in supplementary file Supplementary_DataSources.xlsx.

One set of *A. thaliana* wildtype and DCL1 mutant data, each condition containing three replicates, was also created in our lab. The plants were grown, harvested and then sequenced by biologists in the Dalmay laboratory. The raw FASTQ files and processed csv files are publicly available on Gene Expression Omnibus (GEO) [158] under accession number GSE90771 (GSM2412286 to GSM2412288 are the wild type samples and GSM2412289 to GSM2412291 are the DCL1 mutant samples).

4.4.2 Data processing

All samples downloaded from GEO [158] were processed as follows: files were transformed to FASTA format (using the sratoolkit-2.4.2 [168]). 3' adapters were trimmed and sequences longer than 16 nt were kept for the subsequent steps. All *A. thaliana* samples that we sequenced were transformed from FASTQ to FASTA format, then the 3' adapters and the HD tags were trimmed (using the UEA small RNA Workbench) and sequences longer than 16 nt were kept for the subsequent steps. Next, all files were collapsed into non-redundant format (for each sRNA we kept the sequence and its abundance). Then the files were aligned full length, with 0 gaps and 0 mismatches to the respective reference genome using PatMaN [153]. For miRDeep2, the reads were mapped using mapper.pl; to make the results comparable with the other methods used, the sequences which mapped with mismatches were discarded. All software was run on the processed datasets with their default parameters.

4.4.3 Specificity and sensitivity assessment

To assess the specificity and sensitivity of the miRCat2 method (metrics defined in Section 3.2), we chose miRBase [5] as the reference miRNA annotation database. Although miRBase itself is not perfect (it might miss real miRNAs and contain false entries) [244], it is the standard miRNA annotation database and therefore, the most commonly used.

We downloaded from miRBase v21 the files containing mature sequences for each organism and mapped the test datasets against them (using PatMaN, full length, with 0 gaps and 0 mismatches), to determine which miRNAs are actually present in the file (which miRNAs were expressed in the sequenced sample). We consider all the miRNAs that have at least one sequence mapped to it, regardless of their read count (might contain reads with a low counts), to be expressed in the sample. In this way, we created the reference miRNA dataset for each test file.

Using this reference dataset, we computed the software specificity and sensitivity. The specificity was calculated using the formula TP / (TP + FP), and the sensitivity using the formula TP / (TP + FN) (see Section 3.2).

For the organisms for which miRBase provides a high-confidence miRNA annotation dataset, we determine the number of high-confidence and low confidence miRNA precursors from miRBase, along with the number of novel miRNA predictions. For calculating the specificity and sensitivity in these cases, both high-confidence and low confidence miRNA constitute the TP dataset together.

4.4.4 Fold change computation

The fold change of a sequence represents the difference in expression level of the same sequence in two samples. It is useful to calculate the fold change between wild type and mutant samples, to determine which sequences are differentially expressed (DE) in the mutant data. The differential expression is a strong indicator that the respective sequences are interacting with the mutated genes (are regulated by their activity). Because we compare to mutants in the miRNA biogenesis pathway, we can be quite confident that the sequences that are down-regulated in the mutant datasets represent *bona-fide* miRNAs.

Therefore, to validate miRNA predictions, we estimate fold changes between wild type and mutants of the miRNA biogenesis pathway. To do this, we use the following procedure.

Each experiment is processed individually, because their sequencing of data was designed separately. By one experiment we mean all the sRNA libraries (or samples) that were sequenced together for a specific organism, and it includes all conditions sequenced (the wildtype samples and the mutant samples, where present).

For estimating fold changes between wild type and mutants of the miRNA biogenesis pathway, we consider only the genome mapping reads. To compare datasets with different sequencing depths, we normalize all abundances using the RPM method (reads per million) [245] to the median total count (MTC) of each experiment [246, 247]. Briefly, we sum the abundances of genome mapping reads in each sample to obtain the total for each library; next we calculate the median of total counts for each comparison (the MTC value). We normalize the abundance of each read using: $normalized\_count = \frac{count}{total\_library\_count} * MTC$ [246]. We use MTC to keep the values close to their biological level, while a flat million

value could reduce or inflate them artificially. We repeat this procedure for all experiments.

To create a control dataset, as suggested in [226], containing reads whose abundances are unlikely to be affected by the mutations on the miRNA pathway, the reads in each experiment were mapped to a file containing tRNAs and snoRNAs of the respective species, using PatMaN, full length, with 0 gaps and 0 mismatches. The control file was created downloading tRNAs and snoRNAs from the RFAM database [160] through the RNAcentral web service [248] (http://rnacentral.org/). For each RFAM transcript, we compute its abundance as the algebraic sum of the normalized abundances of mapped reads, for each condition.

We then calculate the log_2 fold change using the normalized abundances of all predicted miRNAs, from each tool. The log_2 fold change for each miRNA is calculated for each set of replicates as the ratio between the median value of normalized abundances from the mutant samples to the median value of normalized abundances from the wildtype (control) samples. We use an offset approach, adding a count of 10 to both numerator and denominator, to avoid divisions by zero and cases where lowly expressed sequences appear to be differentially expressed [222]. This value was chosen empirically, although experiments have shown that most frequently the optimal offset resides between a count of 8 and 15 [249]. Next, we compare the percentage of reads that are significantly down regulated in the mutant samples ($\geq$ 2-fold downregulated).

Because we compare wildtype samples with miRNA biogenesis pathway mutant samples, we expect the miRNAs in the mutant datasets to be significantly downregulated, and therefore, DE. In a previous novel miRNA annotation study a cut-off of 30% difference was considered sufficient to classify a sequence as downregulated [226]. We thus take a strict cut-off of a 2-fold change (100% difference) to be the threshold for considering a sequence DE, to be confident that the sequence is indeed downregulated.

We then plot the cumulative percent of the log_2 fold change for each of the tools' results and control sequences, and compare the percentage of reads that are significantly down regulated in the mutant files ($\geq$ 2-fold downregulated). An example of such plot is presented in Figure 4.10. The percentage of miRNA

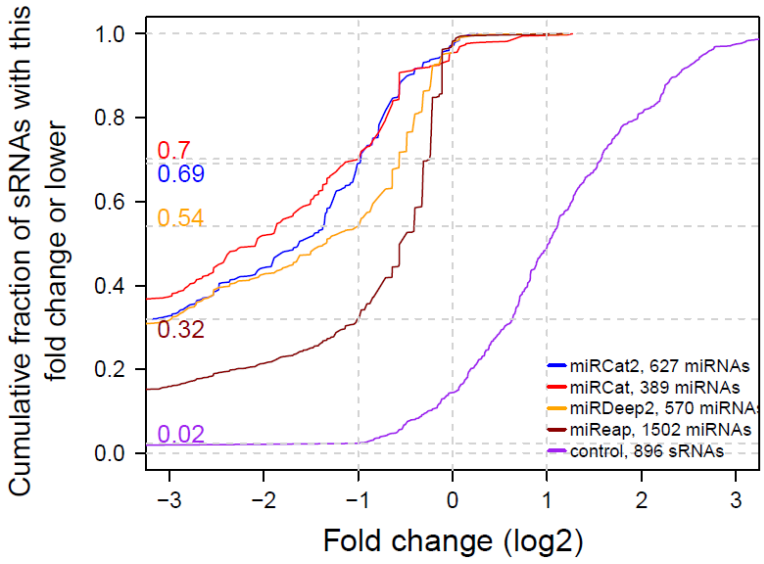


Figure 4.10: **Example of cumulative plot on the log_2 fold change between wildtype and mutant data.** Results are shown for a comparison of wildtype *H. sapiens* data [225] to a Dicer mutant.

predictions that are DE can be found by intersecting the line in -1 on the x-axis with the curve for the cumulative distribution of the log_2 fold changes for the respective tool (in -1 on the x-axis will be the percentage of sequences with a 2-fold change or lower).

For tools with high prediction accuracy we expect to see a significant differential expression (downregulation in the mutant samples i.e. have a fold change lower than or equal to -1) for the majority of the predicted miRNAs. As a control dataset containing reads independent of the miRNA biogenesis pathway, we use RFAM tRNA and snoRNA transcripts (see Section 4.4.4 for details on how the control dataset was computed). As expected, their expression level is not decreased in the mutant samples; moreover, in the animal datasets the expression of these transcripts is upregulated, due to the stochasticity of the sequencing technology. In plant samples we observe little differential expression for the control sequences, as the biogenesis of plant sRNAs is more complex. We should observe that all tools lead to a significantly different curve compared to the control dataset.

We produce cumulative plots on the results for each tool, considering: all the results (to have an overall comparison of the output of each method), only the

novel predictions (to validate novel miRNAs) and miRNAs present in the file but not detected by each tool (to check if the missed miRNAs presented or not miRNA features). Full results are presented in Chapter 5.

4.4.5 Validating novel predictions

To validate new predictions, we considered the following methods:
Comparing to known miRNAs of other species - Many miRNAs are highly conserved among species [38, 39], therefore, if a sequence is annotated as a miRNA in one species, it is very likely that there are orthologues of the miRNA in a related species (genes in different species that evolved from a common ancestral gene by speciation).

To check if novel predictions have already been annotated in other species, we mapped them to all known mature miRNAs from the same Kingdom, using PatMaN [153], full length and allowing 0 gaps and 1 mismatch (there can be a one nt disparity between the sequences, to allow for isomiRs and small variations). All novel predictions that correspond to at least one annotated miRNA in another species are considered to present strong evidence of being a *bona-fide* miRNA.

Finding the miRNA gene source by comparing to all genome annotations - because miRNA genes are transcribed from introns or intergenic regions [15, 16, 36, 47, 48], we can use this information to check whether or not the novel predictions have the same origin. To check the source of these sequences, we downloaded all available annotations on the respective genome (GFF file containing the loci for protein coding genes, exons, introns, sRNAs) and produces the intersection with the results of miRCat2, using bedtools (intersect) [167]. The predictions that have the same source as the miRNA genes (introns or intergenic regions) are more likely to be genuine miRNAs.

Detection of the same novel miRNA in multiple samples - If the experiment has replicates for the same condition, classifying the same sequence as a miRNA multiple times (in the replicates) presents evidence of it being a miRNA. It is highly unlikely that a FP should present miRNA-like features in multiple

samples (if a sequence is a true FP, then it must presented miRNA-like features for a specific dataset by change). This method has been successfully used before in a study of novel miRNA annotation [250]. In this way, the new detections are not restricted only to conserved genes, but novel miRNAs that are species-specific or tissue-specific can be validated.

Pooling of multiple samples - In another study [226], the replicates have been pooled together, to create a richer context for the sequences. For example, the miRNA sequence could be expressed in one sample and the miRNA* sequence could be expressed in another sample, but never together. By pooling the replicates both the miRNA and the miRNA* would appear on the precursor, giving strong evidence of miRNA biogenesis processing. If a sequence is present in both samples, their counts are added. The novel miRNAs predicted from pooled samples would present a higher confidence that they originate from a true miRNA. This method is efficient especially for validating lowly expressed sequences, that could look like FP because of their low read counts. However, the increased depth of the pooled datasets could obscure the miRNA signal in some cases, by increasing the read counts for all sequences in a miRNA locus.

4.5 Summary

In this chapter we presented miRCat2, a new miRNA prediction algorithm, which is suitable both for animal and plant data. The miRCat2 algorithm was designed to handle datasets with high sequencing depth, detecting "peaks" of highly abundant reads in the datasets. These reads are then checked to have a miRNA-like size class distribution and present an alignment in accordance with Dicer/Drosha processing. The secondary structure is then computed, on which further filters are applied, to verify that it folds into a hairpin. The results are then output with a ranking score together with visual plots for easy analysis.

In the next chapter we shall use the methods that we just described to assess the performance of miRCat2 and to computationally verify its novel predictions.

Chapter 5

miRCat2 results

Part of the work presented in this chapter is submitted as part of the manuscript "miRCat2: Accurate prediction of plant and animal microRNAs from next-generation sequencing datasets", Claudia Paicu, Irina Mohorianu, Matthew Stocks, Ping Xu, Aurore Coince, Martina Billmeier, Tamas Dalmay, Vincent Moulton and Simon Moxon.

5.1 Summary

In this chapter we evaluate the performance of our new algorithm, miRCat2, comparing it with miRCat, miRDeep2, miRPlant and miReap, which are the most commonly used tools that we found to have the best performance amongst current software. To benchmark these tools, we have measured their specificity and sensitivity. We then conducted cumulative plots on the fold changes between wild type and mutants in the miRNA biogenesis pathway, calculated on the results of each tool. Next, we investigated the new predictions of miRCat2 in more detail, to assert whether they are likely novel miRNAs or false positive predictions.

5.2 Specificity and sensitivity assessment

To assess the specificity and sensitivity of miRCat2, miRCat [4], miRDeep2 [2], miRPlant [3] and miReap (http://mireap.source-forge.net/), we generated results by running all software with their default parameters on multiple organisms. The data used is described in Section 4.4.1. We filtered the output of each tool as recommended by their authors (miRCat2: no filtering, miRCat: no filtering, miRDeep: filter by score cut-off of 0, miRPlant: filter by score cut-off of 4, miReap: no filtering).

For each method and input dataset we determined the number of high-confidence and low confidence miRNA precursors from miRBase v21 [5], the number of novel miRNA predictions, sensitivity (percentage of miRBase annotated miRNAs within the output) and specificity rates (percentage of miRNAs detected out of the total number of miRNAs expressed in the sample file). The specificity and sensitivity were calculated using the formulas defined in Section 3.2. The averages for each organism are presented in Table 5.1, for animal data and in Table 5.2, for plant data; full results for each individual dataset for the organism presented in the Table 5.2 and for the other organisms mentioned in Section 4.4.1 can be found in supplementary file Supplementary_Results.xlsx. The supplementary file consists of raw numbers of predictions and percentages for the two metrics per individual file, and averages on each tool. We used miRBase as a reference of accepted/studied miRNAs, although we acknowledge its caveats [244].

To calculate the sensitivity and specificity, any miRNA precursor from miRBase that has at least one sequence mapped to it, is considered to be expressed in the sample. This includes very lowly expressed miRNAs, which are difficult to predict, resulting in overall low sensitivity rates.

Comparing the prediction accuracy of miRCat2 with miRCat and miRDeep2/ miRPlant, we observe that miRCat2 has comparable specificity to other methods, whilst achieving an improved sensitivity. In particular, we predict a higher number of known miRNAs, whilst avoiding an increase in the number of false positives. For example, in *M. musculus*, miRCat2 detects 41 more miRNAs than miRDeep2, which has the highest specificity, while predicting only 21 additional (potentially novel) miRNAs. Analysing the results for animal data (Table 5.1), we

Animals

Organism	Tool	High-conf. miRNAs	Low-conf. miRNAs	Novel predictions	Specificity (%)	Sensitivity (%)
H. sapiens (23 datasets)	miRCat2	159	83	72	78.6 (±9.1)	30.6 (±3.3)
	miRCat	122	67	27	87.9 (±5.8)	23.9 (±2.5)
	miRDeep2	149	61	14	94 (±2.7)	26.5 (±4.5)
	miReap	148	108	227	52.3 (±14.3)	32.5 (±7.4)
M. musculus (21 datasets)	miRCat2	147	25	23	90.5 (±7.5)	39.8 (±3.2)
	miRCat	124	20	20	88.5 (±8.3)	33.5 (±1.9)
	miRDeep2	117	14	2	98.6 (±2)	29.7 (±7.2)
	miReap	114	21	134	48.7 (±12.3)	31.6 (±8.5)
D. rerio (2 datasets)	miRCat2	141	145	42	93.6 (±2.4)	88.6 (±2.3)
	miRCat	101	88	26	87.9 (±0.3)	58.2 (±2.5)
	miRDeep2	120	111	27	89.7 (±1.3)	71.5 (±3.0)
	miReap	137	132	43	86.2 (±0.2)	82.9 (±0.2)

Table 5.1: **Performance comparison of benchmarked tools on animal data (on average)**. miRCat2 performs well consistently, having a good specificity and sensitivity trade-off, while miRCat and miReap struggle in terms of specificity. miRDeep2 has good specificity, but lacks in sensitivity.

observe the miRCat2 has the best sensitivity in *M. musculus* and *D. rerio*, while on *H. sapiens* data it ranks second to miReap. However, we observe that miReap achieves this improved sensitivity for *H. sapiens* at a cost to specificity (with 26.3% lower than miRCat2), since it makes a large number of new predictions (155 more than miRCat2), which may be false positives.

For plant data (Table 5.2), miRCat2 offers the second best sensitivity in all three cases, with close percentages to the first predictor (which is miRPlant in *A. thaliana*, and miRCat in *S. lycopersicum* and *G. max*).

Another important fact is that miRCat2 predicts the greatest number of high confidence miRBase miRNA annotations in all of the tests conducted. This is an indicator that miRCat2 is more likely to predict true miRNAs, that the other tools might miss.

In terms of specificity, miRCat2 presents the highest rates in *D. rerio*, and miRDeep2 performs best in *H. sapiens* and *M. musculus*. miRCat2 has the second best rate in *M. musculus*, being close to miRDeep2 (∼8%) and to miRCat, the next highest tool for *H. sapiens* data (∼9%). In plants, miRCat2 has the highest specificity for *A. thaliana* and second best for *S. lycopersicum* and *G. max*, where miRPlant performs better. Both miRDeep2 and miRPlant achieve an improved specificity because of the post-filtering of the results based on the miRNA score

Plants

Organism	Tool	High-conf. miRNAs	Low-conf. miRNAs	Novel predictions	Specificity (%)	Sensitivity (%)
A. thaliana (7 datasets)	miRCat2	66	44	8	93.6 (±2.7)	38.3 (±2.7)
	miRCat	51	57	167	40.9 (±9)	37.9 (±1.8)
	miRPlant	62	52	7	93.3 (±5.4)	39.3 (±14.9)
	miReap	6	8	121	14.5 (±8.5)	4.9 (±0.6)
S. lycopersicum (14 datasets)	miRCat2	15	13	233	11.6 (±5)	44.2 (±12.8)
	miRCat	14	16	1204	2.7 (±1.1)	48 (±4.8)
	miRPlant	11	7	45	30.3 (±7)	28.9 (±13.1)
	miReap	4	5	1619	0.7 (±0.3)	13.6 (±3.2)
G. max (2 datasets)	miRCat2	N/A	129	269	32.7 (±3.8)	34.9 (±1.1)
	miRCat	N/A	149	865	15.4 (±4.5)	40.2 (±0.8)
	miRPlant	N/A	80	74	52 (±0.7)	21.6 (±4.9)
	miReap	N/A	25	2243	1.2 (±0.3)	6.8 (±0.8)

Table 5.2: **Performance comparison of benchmarked tools on plant data (on average)**. miRCat2 performs well consistently, having a good specificity and sensitivity trade-off, while miRCat and miReap struggle in terms of specificity. miRPlant has good specificity, but lacks in sensitivity.

values, as this removes from the analysis the predictions with lower confidence.

In all organisms other than *H. sapiens*, miReap performs poorly, especially in plants, where both sensitivity and specificity are very low. However, miReap has been used for plant miRNA detection in studies such as [188–190], but they only published conserved miRNAs and lab validated predictions, giving no information about the performance of the tool during the conducted experiments.

The low standard deviations (for more detailed plots see Figure 5.1) for miRCat2 present overall favourable numbers, which for sensitivity are second lowest in animals, and for specificity are lowest in *A. thaliana* and second lowest of all tools for *M. musculus* datasets, which means its predictions are more reliable and stable from dataset to dataset of the same organism (little fluctuations in results between datasets of same origin). The other tools generally tend to have the standard deviation high for one metric and low for the other. For example, miRDeep has a low standard deviation for specificity, but high for sensitivity in *M. musculus* data, miRCat has a high value for specificity, but a low value for sensitivity in *A. thaliana*, while miReap has the highest values for both metrics in *H. sapiens* and *M. musculus*, but the lowest in *D. rerio*, *S. lycopersicum* and *G. max*.

The results for specificity and sensitivity, together with the standard deviation

values are a solid indicator that miRCat2 performs overall more consistently than miRCat, miRDeep2, miRPlant and miReap. This suggests that the method used by miRCat2 is more robust. We used miRBase as a reference of accepted/studied miRNAs, although it is a collection of predicted miRNAs and not all sequences have been validated [5]. Therefore, we next use an objective, biologically meaningful test, by computing the differential expression of the results of the tools between wildtype and mutant in the miRNA biogenesis pathway samples.

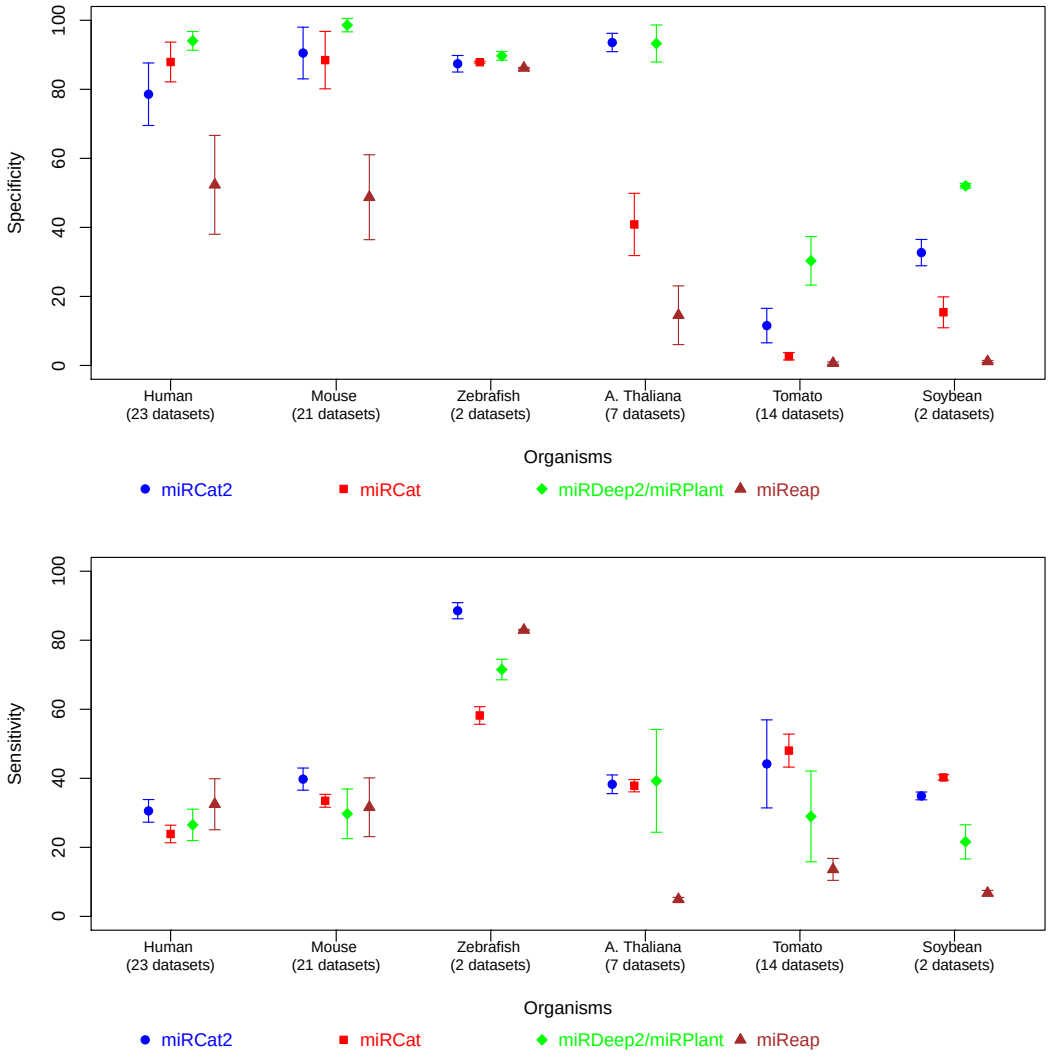


Figure 5.1: Standard deviation for specificity and sensitivity.

5.3 Performance assessment using fold change computation between wildtype and miRNA biogenesis mutant data

Because miRBase is likely to contain false positives and to compensate for the lack of in-depth miRNA annotations for some model organisms [244], we investigated whether or not the mature miRNAs predicted by each tool were dependent on Dicer/DCL1, Drosha and DGCR8 processing, which are key factors in miRNA biogenesis. This is a more robust method of validating the tools predictions. The reasoning behind this was that we would expect *bona-fide* miRNAs to have reduced expression in Dicer, Drosha, DGCR8 mutants versus wildtype samples. We consider a sequence as being down-regulated in the mutant dataset if the normalized expression is at least two fold lower in the mutant. We have calculated the fold change and conducted cumulative plots on it as explained in Section 4.4.4.

To check the quality of the datasets, we produced correlation plots between the expression levels in wild type and mutants for miRBase miRNAs (see Figure 5.2). If the experiment was created successfully, we expect to see higher counts in the wildtype dataset than in the mutant, therefore the plots should show a shift of the sequences above and parallel to the diagonal, and this pattern can be observed in the majority of cases (for example, the pattern can be certainly distinguished in the plot for *S. lycopersicum*). However, for *D. rerio* the pattern is not very clear, in *G. max* and *M. musculus* the sequences seem to group around the diagonal, rather than being shifted above, but in all cases more than a half of the sequences are plotted above the diagonal. This suggests that these datasets contain overall lower percentages of differential expression amongst miRNAs, therefore the tools results might also have lower percentages when plotted. We noticed that in the *H. sapiens* wildtype vs. Drosha mutant, there are some miRNAs that appear below the diagonal (more highly expressed in the mutant). This is likely because they might be mirtrons or have a Drosha-independent biogenesis pathway and therefore appear to be more highly expressed in a miRNA depleted background [225, 251, 252].

In Figure 5.3 we compare the performance of miRCat2, miRCat, miRDeep,

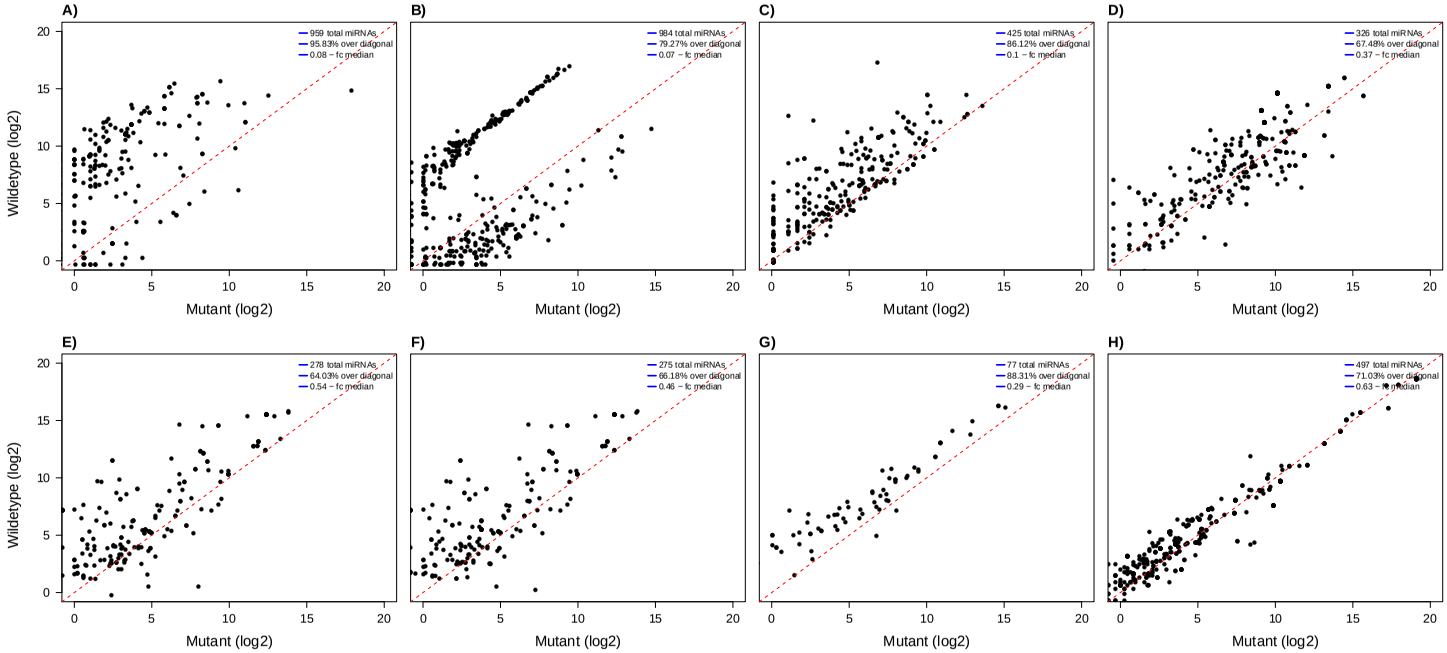


Figure 5.2: **Correlation plots of normalized abundances for expressed miRBase miRNAs in the wildtype, compared to mutant samples.** We present results for *H. sapiens* (subplots (A) Dicer and (B) Drosha knock-out), *M. musculus* (subplot (C)), *D. rerio* (subplot (D)), *A. thaliana* (subplots (E) and (F)), *S. lycopersicum* (subplot (G)) and *G. max* (subplot (H)). The plots give information about the percentage of miRNAs that are more abundant in the wildtype (above diagonal) and the median fold change, where a fold change of 0.5 means the sequence is down-regulated in the mutant. (A) *H. sapiens* wildtype vs. Dicer knock-out. (B) *H. sapiens* wildtype vs. DROSHA knock-out. (C) *M. musculus* wildtype vs. DGCR8 knock-out. (D) *D. rerio* wildtype vs. Dicer knock-out. (E-F) *A. thaliana* wildtype vs. Dicer knock-down. (G) *S. lycopersicum* wildtype vs. DCL1 knock-down. (H) *G. max* wildtype vs. DCL1 knock-down.

miReap and miRPlant with and without filtering. For miRCat2, we have used a score cut-off of five (empirically observed to separate most new predictions from already annotated miRNAs). The filtering has some impact both on miRCat2 and miRDeep2 in *H. sapiens*. In plants however we observe that miRCat2 performs well irrespective of this filtering, with a particularly large impact for miRPlant. For comparability purposes, we computed the cumulative plots of log_2 fold changes only on unfiltered outputs (see Figure 5.4).

For tools with high prediction accuracy we expect to see a significant differential expression (downregulation in the mutant samples i.e. have a fold change lower than or equal to -1) for the majority of the predicted miRNAs. As a control dataset containing reads independent of the miRNA biogenesis pathway, we use RFAM tRNA and snoRNA transcripts (see Section 4.4.4 for details on how the control dataset was computed). As expected, their expression level is not decreased in the mutant samples; moreover, in the animal datasets the expression of these transcripts is upregulated, due to the stochasticity of the sequencing technology. In plant samples we observe little differential expression for the control sequences, as the biogenesis of plant sRNAs is more complex. All tools produce a substantially different cumulative differential expression curve compared to the control dataset; miRCat2 performs better than other tools in all but one of the experiments, because it presents the largest fraction of predictions that are downregulated in the mutants. In *H. sapiens* vs. Dicer knock-out sample (see Figure 5.4,(A)), we observe that miRCat2 is a close second to miRCat, while in plant data there is a substantial gap between miRCat2 and the other tools.

The plots confirm the validity of the predictions, especially for *S. lycopersicum*, where miRCat2 shows a low specificity when detecting annotated miRBase miRNAs. This could be explained by the fact that miRBase has not been updated recently and currently only contains 77 annotated precursors in *S. lycopersicum* (miRBase v21), so these novel predictions might be real miRNAs which have not yet been annotated.

We next produced cumulative plots only for those sequences that were not previously annotated in miRBase and therefore are potential newly predicted miRNAs (see Figure 5.5). Here we still see a significant downregulation of predicted miRNAs in the mutant samples, although to a lesser extent than the plots

including all predictions. We observe no change in the ranking of the tools, miRCat2 performing better than the other tools in each of the experiments. In *M. musculus* we notice a drop for all tools in the percentage of sequences with at least a 2 fold change, which can be explained by the low number of novel predictions, where any sequence predicted has a large influence on the result. The high percentage of differentially expressed sequences among novel predictions, especially in plants, indicate that these sequences are likely to be *bona-fide* miRNAs.

To examine the low overall sensitivity rates, we have conducted cumulative plots on the miRNAs present in the dataset, but not detected by each tool. We expect these to have low counts, suggesting that they are very lowly expressed in the sample. Alternatively these sequences could be misannotations in miRBase and therefore do not show features consistent with canonical miRNA structure and biogenesis and therefore their expression would not be affected in the mutant datasets. Consequently, we expect to see a less significant change between the wildtype and mutant samples in the cumulative plot. In this case, a curve closer to the control line would indicate that the miRNAs missed by the tool were not changed in the mutants and therefore potentially should not have been predicted.

Looking at these plots (see Figure 5.6), we can observe a clear change in the shape of the lines for each tool (especially for miRCat2), suggesting that these miRNAs might not present the canonical miRNA features or were lowly expressed in the datasets analysed. Also, it is notable that miRCat2 consistently performs well, having the lowest value of DE amongst the miRBase miRNAs that were not identified in seven out of eight cases. This suggests that miRCat2 is less prone to false positives and false negatives than other methods, because it detects more miRNAs that are DE in the mutant data and because it successfully identifies the miRNAs that are DE, that other tools miss.

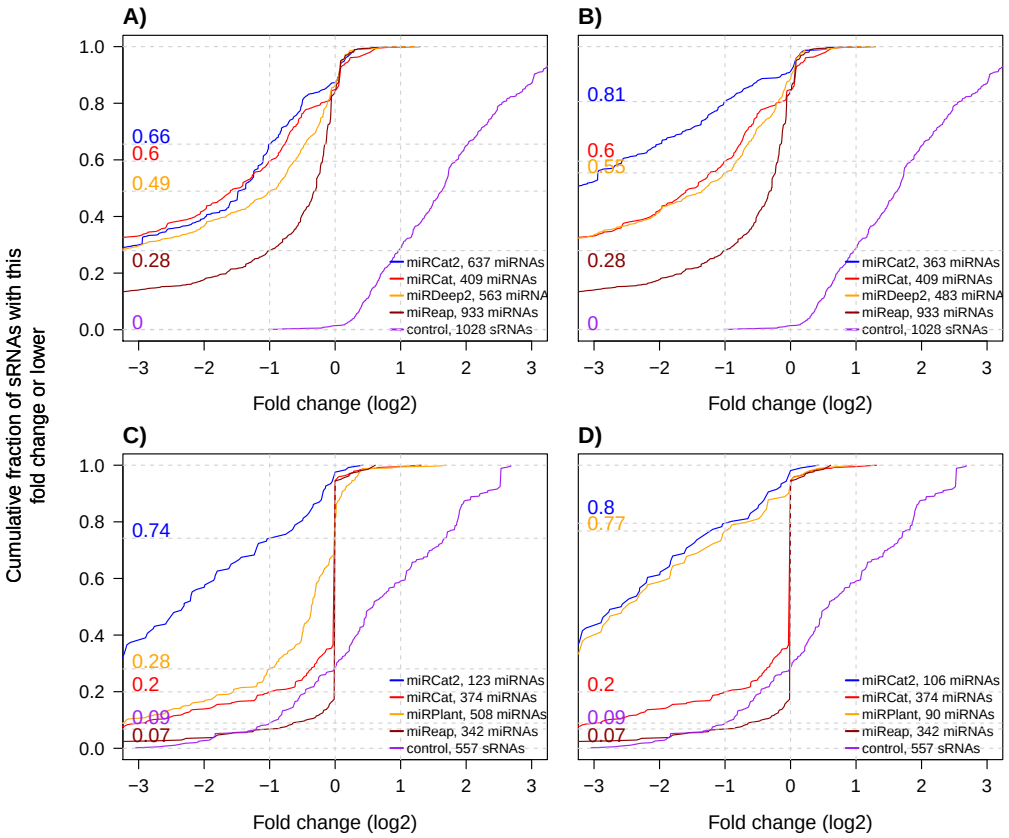


Figure 5.3: Comparison of filtered vs not filtered results for *H. sapiens* (subplots (A) and (B)) and *A. thaliana* (subplots (C) and (D)) data. In each plot we represent the cumulative distribution of differential expression for predictions conducted with miRCat2, miRCat, miRDeep2/miRPlant and miReap. The results were filtered based on the recommended cut-off of the score for miRDeep2 (0) and miRPlant (4) and a value of 5 for miRCat2, empirically determined. We observe that for both plant and animal data, the filtering has an effect on the performance of the tools. (A) *H. sapiens* wildtype vs. DROSHA knock-out, before filtering. (B) *H. sapiens* wildtype vs. DROSHA knock-out, after filtering. (C) *A. thaliana* wildtype vs. DCL1 knock-down, before filtering. (D) *A. thaliana* wildtype vs. DCL1 knock-down, after filtering.

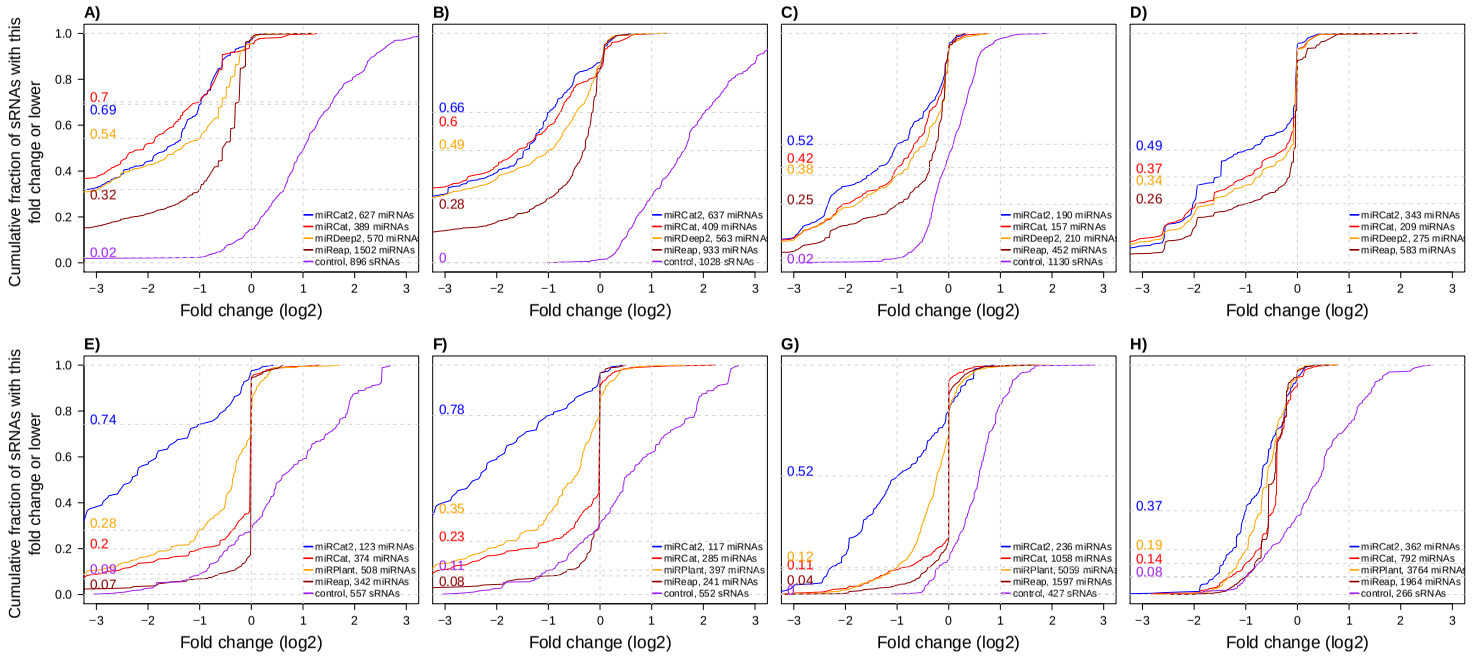


Figure 5.4: **Cumulative plots of log_2 fold changes of control vs. mutant datasets, calculated on the output of miRCat2, miRCat, miRDeep2/miRPlant and miReap and a control dataset formed of tRNAs and snoRNAs**. We present results for *H. sapiens* (subplots (A) Dicer and (B) Drosha knock-out), *M. musculus* (subplot (C)), *D. rerio* (subplot (D)), *A. thaliana* (subplots (E) and (F)), *S. lycopersicum* (subplot (G)) and *G. max* (subplot (H)). miRCat2 has the highest percentage of DE miRNAs in all but one of the experiments, were it classifies as a close second to miRCat. (A) *H. sapiens* wildtype vs. Dicer knock-out. (B) *H. sapiens* wildtype vs. DROSHA knock-out. (C) *M. musculus* wildtype vs. DGCR8 knock-out. (D) *D. rerio* wildtype vs. Dicer knock-out. (E-F) *A. thaliana* wildtype vs. Dicer knock-down. (G) *S. lycopersicum* wildtype vs. DCL1 knock-down. (H) *G. max* wildtype vs. DCL1 knock-down.

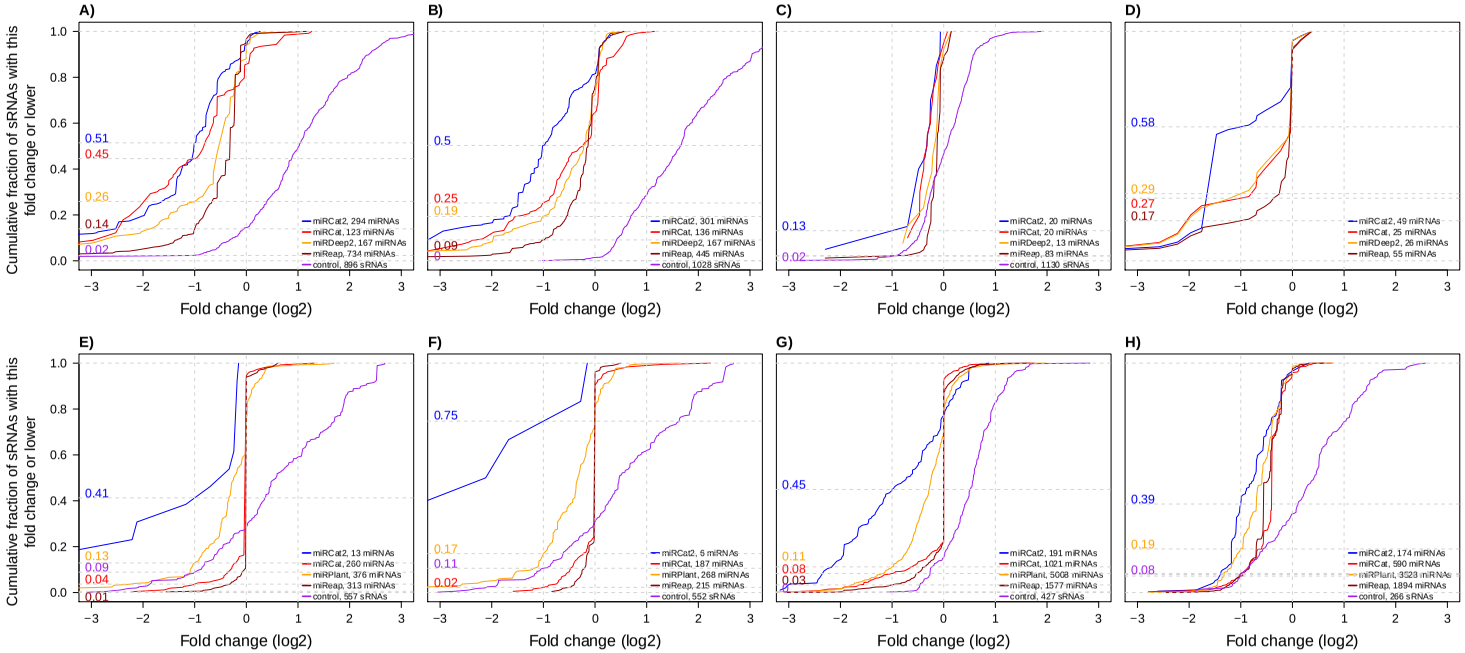


Figure 5.5: **Cumulative plots of log_2 fold changes of control vs. mutant datasets, calculated on the new predictions of miRCat2, miRCat, miRDeep2/miRPlant and miReap and a control dataset formed of tRNAs and snoRNAs**. We present results for *H. sapiens* (subplots (A) Dicer and (B) Drosha knock-out), *M. musculus* (subplot (C)), *D. rerio* (subplot (D)), *A. thaliana* (subplots (E) and (F)), *S. lycopersicum* (subplot (G)) and *G. max* (subplot (H)). miRCat2 has the highest percentage of DE miRNAs in all of the experiments. (A) *H. sapiens* wildtype vs. Dicer knock-out. (B) *H. sapiens* wildtype vs. DROSHA knock-out. (C) *M. musculus* wildtype vs. DGCR8 knock-out. (D) *D. rerio* wildtype vs. Dicer knock-out. (E-F) *A. thaliana* wildtype vs. Dicer knock-down. (G) *S. lycopersicum* wildtype vs. DCL1 knock-down. (H) *G. max* wildtype vs. DCL1 knock-down.

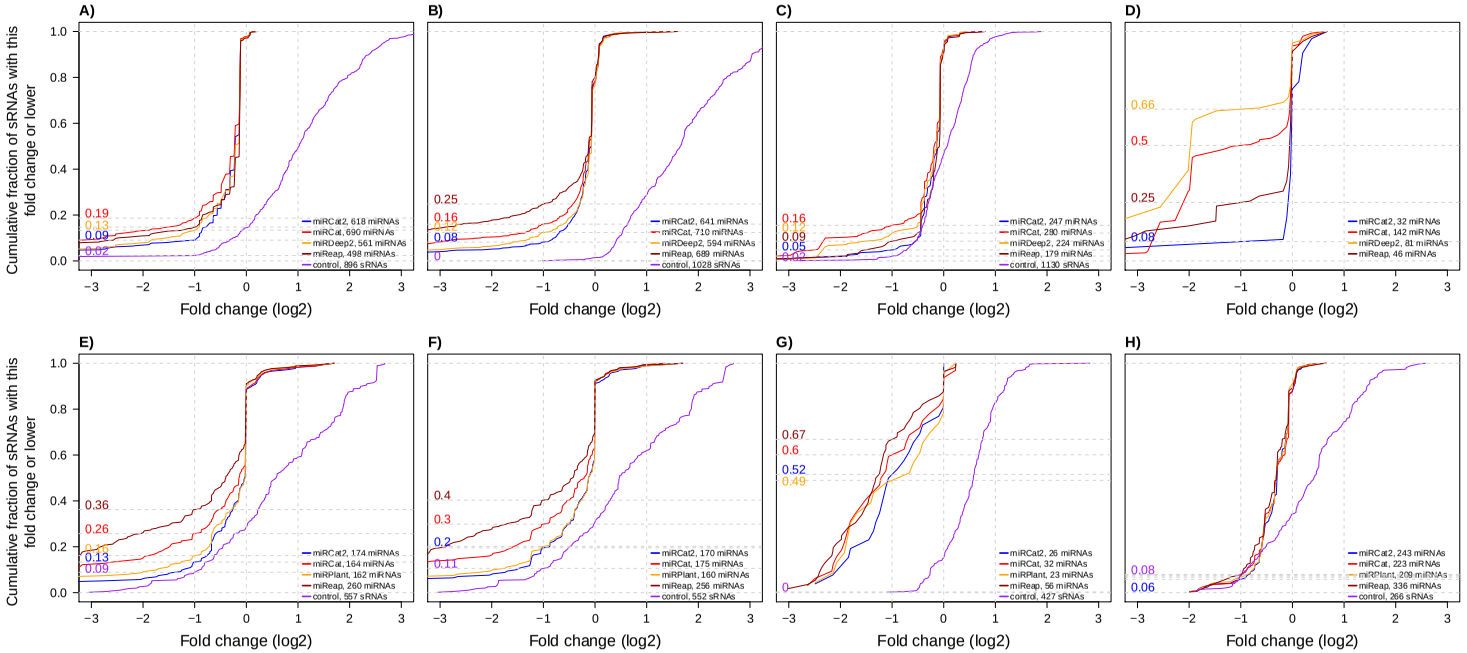


Figure 5.6: **Cumulative plots of log_2 fold changes of control vs. mutant datasets, calculated on miRBase miRNAs present in the datasets, but not detected by the predictions of miRCat2, miRCat, miRDeep2/miRPlant and miReap and on a control dataset formed of tRNAs and snoRNAs.** We present results for *H. sapiens* (subplots (A) Dicer and (B) Drosha knock-out), *M. musculus* (subplot (C)), *D. rerio* (subplot (D)), *A. thaliana* (subplots (E) and (F)), *S. lycopersicum* (subplot (G)) and *G. max* (subplot (H)). We expect to see a smaller differential expression between the wildtype and mutant samples in the cumulative plot i.e. a curve closer to the control line. miRCat2 presents the lowest differential expression in all experiments, suggesting that it is less prone to false positives than other methods. (A) *H. sapiens* wildtype vs. Dicer knock-out. (B) *H. sapiens* wildtype vs. DROSHA knock-out. (C) *M. musculus* wildtype vs. DGCR8 knock-out. (D) *D. rerio* wildtype vs. Dicer knock-out. (E-F) *A. thaliana* wildtype vs. Dicer knock-down. (G) *S. lycopersicum* wildtype vs. DCL1 knock-down. (H) *G. max* wildtype vs. DCL1 knock-down.

5.4 Run time and memory requirements

To test the memory requirements and run time of miRCat2, miRCat, miRDeep2, miRPlant and miReap, we monitored the execution of the tools on two datasets (*H. sapiens* and *A. thaliana*). We ran the tools on a Linux server with CentOS 5.11 operating system, 144GB of memory and 2 Intel Xeon X5550 processors. The memory consumption and run time for each tool are presented in Table 5.3. For this test we monitor the tools only after the initial read mapping step (to measure precisely only the miRNA prediction algorithm), using JConsole [253] for Java based tools (miRCat2, miRCat and miRPlant) and the Linux top command, for the others (miRDeep2, miReap).

	H. sapiens (Control2, 34.450.792 seq.)		*A.thaliana* (Arab_WTA 6.698.043 seq.)	
Tool	Memory (GB)	Time (h:m:s)	Memory (GB)	Time (h:m:s)
miRCat2	12.89	3:50:37	16.3	2:30:39
miRCat	19.94	00:28:31	6.992	00:04:07
miRDeep2	0.37	5:15:32s	N/A	N/A
miRPlant	N/A	N/A	1.2	00:55:00
miReap	61	00:45:43	2.1	00:03:22

Table 5.3: **Performance comparison of run time and memory consumption between miRCat2, miRCat, miRDeep2, miRPlant and miReap.** The number of sequences represent genome mapped sequences in each file.

It is notable that for miRCat and miRCat2, the user can define the memory constraints it wishes the tool to be run with. That is, it is possible for these tools to be run on a desktop with only 4GB of RAM (in the detriment of runtime). For the purpose of this test, we allowed up to a maximum of 64GB and monitored how much memory each software actually consumes. For miRPlant, the memory used is predefined to 1.2GB and cannot be modified by the user, which forces the Java Virtual Machine (JVM) to swap the RAM to be able to do its processing, which substantially increases the run time of miRPlant. Therefore, larger datasets can take a long time to run for miRPlant (for example, running miRPlant on the *G. max* datasets took more than 1 week each).

Comparing the performance of the tools on the *H. sapiens* dataset, which is a fairly large input, containing approximately 34.5 million reads, we observe

that miRDeep2 consumes the least amount of memory (0.37GB), followed by miRCat2 in second place (12.89GB). It is notable that the low RAM needed is however achieved by sacrificing run time, the two also having the highest run time. We also see that miReap has the highest memory consumption (61GB), using over 40GB more than miRCat (19.94GB), while also consuming almost twice the time to complete its processing, compared to miRCat.

Looking at the performance of the tools on the *A. thaliana*, which is a rather small input, containing approximately 6.7 million reads, we observe that miRCat2 has both the highest run time and memory consumption. miRCat has the second highest RAM requirement, but it is achieved substantially faster, having the second lowest run time. While miReap has both low RAM and time consumption, miRPlant achieves the lowest memory consumption by increasing its processing time.

Next we compared the results between the runs on the two datasets, to see the scaling of the algorithms from a small input (*A. thaliana*) to a larger input (*H. sapiens*), containing 5.2 times more sequences. We observe that both miRCat and miReap suffer a considerable increase both in memory and in run time consumption in the larger dataset. miRCat required approximately 3 times more memory and 7 times more runtime, while miReap needed 30 times more RAM and 15 times longer processing time. miRDeep and miRPlant are suitable only for one of the datasets and they are not directly comparable. We notice that for miRCat2, the memory requirement does not change, while the run time is increased only by 1.4 times. This suggests that miRCat2 is scaling better and is more suited for larger inputs (which lead to a larger number of mapped unique genomic locations), without needing proportional amounts of resources.

	H. sapiens (Control2, 34.450.792 seq.)		*A.thaliana* (Arab_WTA 6.698.043 seq.)	
miRCat2	Memory (GB)	Time (h:m:s)	Memory (GB)	Time (h:m:s)
DB on disk	12.89	3:50:37	16.3	2:30:39
DB in memory	18.1	0:52:52	15.5	0:32:37

Table 5.4: Performance comparison of run time and memory consumption for miRCat2, when constructing the database in memory or on disk. The number of sequences represent genome mapped sequences in each file.

The high processing time for both datasets for miRCat2 is partially due to the way the UEA small RNA Workbench is implemented: to decrease memory requirements, the data is stored in a local database (DB) (created at each run). This decreases the RAM consumed (especially for large datasets), but it adds to the run time, because it need to access the disk every time a query is sent to the database. However, this is not a substantial issue in practice, since it performs similarly to other current software and the user might choose to let it run overnight.

To avoid long run times, the user has the option to run the Workbench constructing the database in memory, which decreases the runtime. This is however recommended only if the user can make use of large amounts of RAM. In table 5.4 we compare the performance of miRCat2 when constructing the database on the disk with constructing it in memory. We observe that its runtime is substantially smaller in the second case. While the memory consumption is only slightly increased during the processing of miRCat2, we must mention that a larger amount of RAM was required for the database construction for the *H. sapiens* dataset (35.1GB).

5.5 Validation of novel miRNAs for miRCat2

To investigate the new predictions made by miRCat2, besides the fold change computation, we conducted a series of tests on the *S. lycopersicum* data to provide arguments that these new predictions are real miRNAs. The methods used for these tests are described in Section 4.4.5. We have chosen the *S. lycopersicum* dataset because it has the lowest specificity of all datasets upon which we tested miRCat2 when compared to miRBase, the low specificity suggesting that miRCat2 allows for FP. These tests confirm that there are many novel miRNAs amongst these predictions, which have not yet been annotated. All the tests below were conducted on *S. lycopersicum* sample WT1 (for which we also generated cumulative plots), to have a better understanding of the data.

Plots describing secondary structure and alignment information

miRCat2 outputs pdf files containing graphical representation of the hairpin

structure and plots with the alignment of sequences on the secondary structure. Using these plots, the user can check if the prediction presents the required miRNA features (hairpin structure and alignment depth and distribution - see Section 2.4.2). An example of a successful such prediction and of a more questionable prediction are presented in Figure 5.7 and Figure 5.8, respectively.

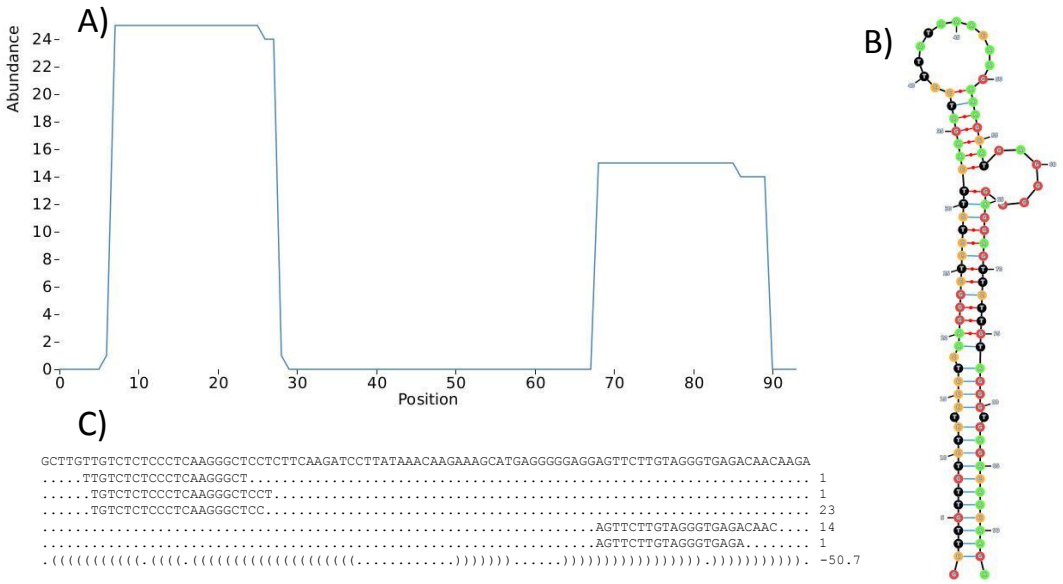


Figure 5.7: **Output of miRCat2 for a successful prediction (chromosome 10).** The information shown contains A) precursor coverage plots, B) precursor secondary structure and C) alignment of incident reads.

Comparing to known miRNAs of other species

Comparing the new *S. lycopersicum* predictions to other plant miRNAs, we have found that 54 out of the 190 new predictions for the *S. lycopersicum* dataset are likely homologs of miRNAs that are annotated in other plant species (28.4% of the new predictions). Table 5.6 gives information about these sequences, together with one example of a homologue annotated miRNA (most predictions have more than one corresponding annotated miRNA). This is a strong indicator that these loci represent *bone-fide* miRNA genes. A full list of homologs for each new miRNA prediction can be found in Supplementary_Homologs.xlsx

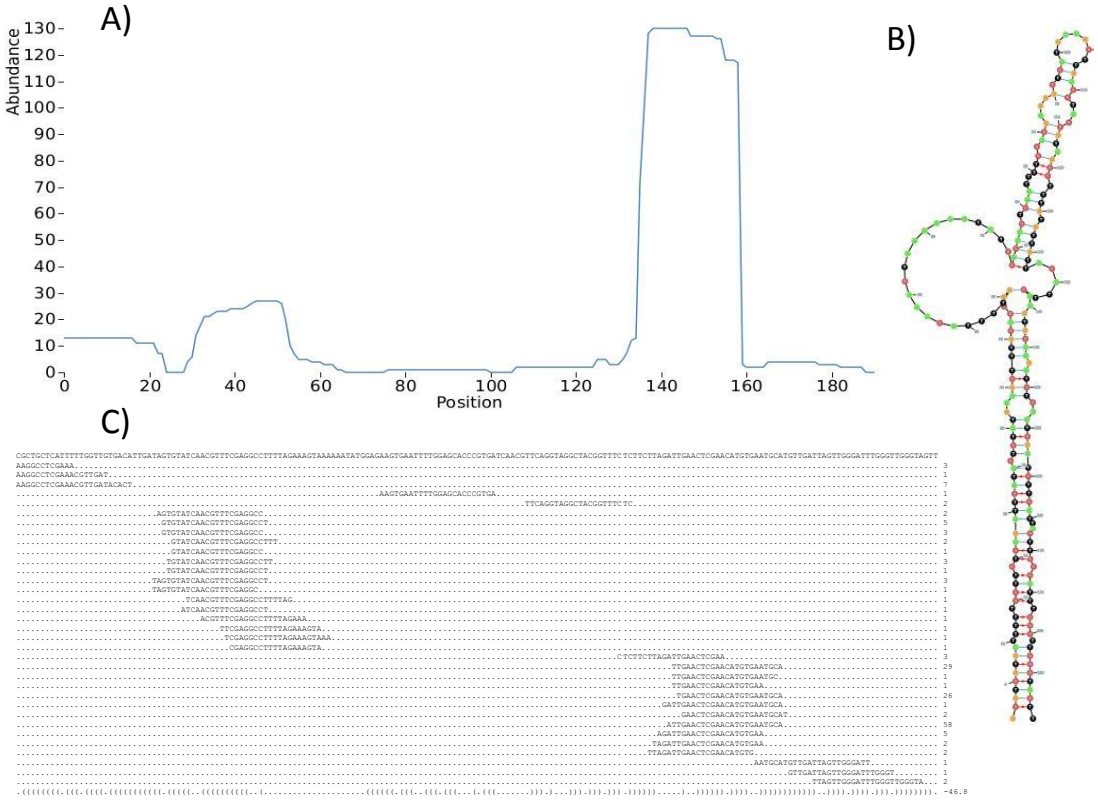


Figure 5.8: **Output of miRCat2 for a more questionable prediction (chromosome 10).** The information shown contains A) precursor coverage plots, B) precursor secondary structure and C) alignment of incident reads.

Detection of the same novel miRNA in multiple samples

We have also checked to see if the same sequence was predicted in other *S. lycopersicum* samples. We have run miRCat2 on all 4 samples of the same *S. lycopersicum* experiment and intersected the results. We then counted the new predictions from sample WT1 that have been predicted in more than one dataset. 170 out of the 191 new predictions (89%) are detected in more than one sample, 79 (41.3%) of these predictions having low counts (counts under 50). This presents evidence of them being true miRNAs [250], because it is highly unlikely that a FP should present miRNA-like features in multiple samples (if a sequence is a true FP, then it must present miRNA-like features for a specific dataset by change). Details about these novel predictions that were detected in multiple samples, including information about their abundances in the other samples

where they have been predicted can be found in supplementary file Supplementary_Multiple_Samples.xlsx.

Pooling of multiple samples

We pooled together 4 samples of *S. lycopersicum* and created a single FASTA file containing all sequences, adding the counts of the same sequence if found in multiple datasets. Then we run miRCat2 on the newly obtained data. We obtained similar results to the previous method, where we checked the predictions in all samples after individual runs. After intersecting the WT1 with the pooled sample results, we found that 166 out of the 191 new predictions (86.9%) in WT1 sample are detected in the pooled sample as well, 77 (40.3%) of which having low counts in the original file (counts under 50). This is strong evidence that the overlapping predictions, especially the lowly expressed reads, are true miRNAs. We show the predictions that have been predicted both in WT1 and in the pooled samples in supplementary file Supplementary_Pooled_Samples.xlsx.

Finding the miRNA gene source by comparing to all genome annotations

We have downloaded available *S. lycopersicum* annotated genome regions from the Sol Genomics website [1] [254], containing protein coding regions and sRNAs. The annotated genomic regions were produced by Infernal [2], adding to the miRBase entries and containing a total of 391 miRNAs. These annotations complement miRBase and give a wider view of existing *S. lycopersicum* miRNAs. We produced the intersection of the novel predictions of miRCat2 with the GFF annotation file, bedtools (intersect) [167]. The number of predicted sequences that overlap with annotated regions are shown in Table 5.5:

miRNA	21	intron	15	intergenic	137
gene	13	3' and 5' UTR	3	tRNA	1

Table 5.5: Intersection of novel predictions with annotated genes of the *S. lycopersicum* genome.

The first line in Table 5.5 shows sequences with an origin that could cor-

[1] ftp://ftp.solgenomics.net/tomato_genome/annotation/ITAG2.3_release/
[2] http://infernal.janelia.org/

respond to a miRNA genes, while on the second line, the sequences have an origin that does not correspond to miRNA biogenesis. There are 21 novel predictions that correspond to miRNAs annotated by Infernal, but are not present in miRBase, which confirms that improvements could be made to the current version of miRBase [5, 244]. Moreover, 15 loci were predicted from intronic regions and 137 from intergenic regions, from where miRNAs are usually generated [15, 16, 36, 47, 48]. This information, together with their incident read alignment and secondary structure, gives us good reason to believe they could be real miRNAs that the miRCat2 algorithm detected. Therefore, we conclude that out of 190 total novel predictions, 173 predictions (91.05%) have a miRNA-like origin for the *S. lycopersicum* dataset analysed.

Chromozome	Sequence	Count	Start	End	Strand	Homolog miRNA in other species
SL2.40ch10	TCGGACCAGGCTTCATTCCCC	79714	3321909	3321929	+	ath-miR166a-3p
SL2.40ch10	ATGGGTAGCACAAGGATTAATG	1070	62182533	62182554	+	sly-miR6027-5p
SL2.40ch03	TCTCGGACCAGGCTTCATTCC	119585	292999	293019	-	gma-miR166h-3p
SL2.40ch03	TCCAAAGGGATCGCATTGATC	8	7461884	7461904	+	gma-miR393h
SL2.40ch03	TAGCCAAGGATGACTTGCCT	13	9591894	9591913	+	tcc-miR169g
SL2.40ch03	CTGAAGTGTTTGGGGGAACTC	53	30608183	30608203	+	aly-miR395i
SL2.40ch03	CTGAAGTGTTTGGGGGAACTC	53	30608715	30608735	+	aly-miR395i
SL2.40ch03	TCGGACCAGGCTTCATTCCCC	79714	46023561	46023581	-	ath-miR166a-3p
SL2.40ch03	TCGATAAACCTCTGCATCCAG	1695	58491605	58491625	+	ath-miR162a-3p
SL2.40ch03	TTGACAGAAGATAGAGAGCAC	8015	61720039	61720059	+	smo-miR156c
SL2.40ch03	TTTGGATTGAAGGGAGCTCTA	64876	61786222	61786242	+	osa-miR159a.1
SL2.40ch03	TTGGACTGAAGGGTTTCCTTC	1352	61794716	61794736	+	stu-miR319-3p
SL2.40ch02	TGCCTGGCTCCCTGTATGCCA	210	16516665	16516685	-	mes-miR160g
SL2.40ch02	CTGAAGTGTTTGGGGGAACTC	53	27508523	27508543	-	aly-miR395i
SL2.40ch02	CTGAAGTGTTTGGGGGAACTC	53	27508705	27508725	-	aly-miR395i
SL2.40ch02	CTGAAGTGTTTGGGGGAACTC	53	27526675	27526695	-	aly-miR395i
SL2.40ch02	CTGAAGTGTTTGGGGGAACTC	53	27533088	27533108	-	aly-miR395i
SL2.40ch02	TATTGGCCTGGTTCACTCAGA	78	27921894	27921914	-	ath-miR170-5p
SL2.40ch02	TTGACAGAAGATAGAGAGCAC	8015	29897683	29897703	+	smo-miR156c
SL2.40ch02	TTGACAGAAGATAGAGAGCAC	8015	47055512	47055532	+	smo-miR156c
SL2.40ch01	TCGGACCAGGCTTCATTCCCC	79714	79167141	79167161	+	ath-miR166a-3p
SL2.40ch01	TGCACTGCCTCTTCCCTGGCT	85	82801320	82801340	-	smo-miR408
SL2.40ch01	TTGGCATTCTGTCCACCTCC	269	84522987	84523006	+	vvi-miR394a
SL2.40ch12	TTCCACAGCTTTCTTGAACTT	1957	2899110	2899130	+	vvi-miR396b
SL2.40ch12	TGTCGCAGATGACTTTCGCCC	828	6988948	6988968	-	sly-miR1919c-5p
SL2.40ch12	TTGGACTGAAGGGTTTCCTTC	1352	39442963	39442983	+	stu-miR319-3p
SL2.40ch12	TTGGACTGAAGGGAGCTCCCT	13907	47456665	47456685	-	ppt-miR319a
SL2.40ch00	TTCCACAGCTTTCTTGAACTG	20754	12537520	12537540	+	vvi-miR396b
SL2.40ch00	CTGAAGTGTTTGGGGGAACTC	53	17038754	17038774	-	aly-miR395i
SL2.40ch11	TATGTTCTCAGGTCGCCCCTG	607	47382919	47382939	-	stu-miR398a-3p
SL2.40ch07	TGACAGAAGAGAGTGAGCAC	832	324087	324106	+	osa-miR156k
SL2.40ch07	TAGCCAAGGATGACTTGCCT	13	2174513	2174532	+	tcc-miR169g
SL2.40ch07	TAGCCAAGGATGACTTGCCT	13	2180718	2180737	+	tcc-miR169g
SL2.40ch07	TGACAGAAGAGAGTGAGCAC	832	48503782	48503801	+	osa-miR156k
SL2.40ch06	AAGCTCAGGAGGGATAGCGCC	45	1372252	1372272	-	cca-miR390
SL2.40ch06	TCGGACCAGGCTTCATTCCCC	79714	33130363	33130383	+	ath-miR166a-3p
SL2.40ch06	TCGATAAACCTCTGCATCCAG	1695	39463195	39463215	+	ath-miR162a-3p
SL2.40ch06	TGATTGAGCCGTGCCAATATC	30	40810831	40810851	+	bna-miR171g
SL2.40ch05	CTGAAGTGTTTGGGGGAACTC	53	1703003	1703023	+	aly-miR395i
SL2.40ch05	CTGAAGTGTTTGGGGGAACTC	53	1705348	1705368	+	aly-miR395i
SL2.40ch09	TGAAGCTGCCAGCATGATCTA	66	59575910	59575930	+	osa-miR167d-5p
SL2.40ch09	TGAAGCTGCCAGCATGATCTA	66	59584688	59584708	+	osa-miR167d-5p
SL2.40ch09	TGAAGCTGCCAGCATGATCTA	66	63883415	63883435	+	osa-miR167d-5p
SL2.40ch09	TCGGACCAGGCTTCATTCCCC	79714	64446634	64446654	-	ath-miR166a-3p
SL2.40ch08	TCGGACCAGGCTTCATTCCCC	79714	2978492	2978512	+	ath-miR166a-3p
SL2.40ch08	TGACAGAAGAGAGTGAGCAC	832	49143120	49143139	-	osa-miR156k
SL2.40ch08	GCTCACTGCTCTATCTGTCACC	53	49143294	49143315	-	zma-miR156l-3p
SL2.40ch08	TGACAGAAGAGAGTGAGCGC	10	49143362	49143381	-	ath-miR156a-5p
SL2.40ch08	TGACAGAAGAGAGTGAGCAC	832	49143604	49143623	-	osa-miR156k
SL2.40ch08	TGACAGAAGAGAGTGAGCAC	832	49143859	49143878	-	osa-miR156k
SL2.40ch08	TGACAGAAGAGAGTGAGCAC	832	49276200	49276219	+	osa-miR156k
SL2.40ch08	TAGCCAAGGATGACTTGCCT	13	52708635	52708654	+	tcc-miR169g
SL2.40ch08	TTGGACTGAAGGGAGCTCCCT	13907	61949534	61949554	-	ppt-miR319a
SL2.40ch08	TTGCTGCCGACTCATTCATCCA	78	61949641	61949662	-	smo-miR319

Table 5.6: New predictions in *S. lycopersicum* that have homologs in other plant species (only one example shown). Homologoues sequences were obtained by matching miRCat2 new predictions to all mature miRNAs from miRBase with one missmatch.

5.6 Conclusions

We have presented a new tool for miRNA prediction, miRCat2, applicable on both plants and animals, which can be run both from the UEA small RNA Workbench graphical interface and from the command line. The miRCat2 output offers useful information about its predictions, implementing new features for users to better visualise and analyse its results. It produces descriptive plots depicting the secondary structure and the alignment of sequences on the hairpin, which constitute valuable information for easy manual processing and validation of the predictions. Another feature is that miRCat2 can be easily integrated into bioinformatics workflows available from the Workbench for a more complex analysis of the data.

We tested miRCat2 on ten model organisms and compared its results with four commonly used tools for miRNA discovery (miRCat, miRDeep2, miRPlant and miReap). miRCat2 shows a good trade-off between sensitivity and specificity (relative to miRBase annotation), performing well in both metrics, while other tools generally performed well only for one of these measures. More specifically, miRDeep2 and miRPlant had good specificity rates, but lacked in sensitivity (annotated miRNAs are not predicted). miReap had a good sensitivity in animals, but lacked in specificity, allowing a high number of new predictions, which could potentially contain false positives.

To evaluate the accuracy of the predictions we used the miRBase annotations and the objective and biologically meaningful mutant test (using Dicer/DCL1, Drosha, DGCR8 mutants). This approach alleviated the lack of in-depth miRNA annotations for some model organisms [244]. We have shown using the comparison of wildtype and mutant datasets, in the cumulative plots, that miRCat2 generally performs better than all other tools tested, both overall and when confirming novel annotations. The tool also remains consistent in its predictions across all animal and plant data whilst the other tools tend to perform better only on some of the organisms: miRCat and miRDeep2 perform well in *H. sapiens* and *D. rerio*, while miRPlant performs well in *A. thaliana*.

Advantages of other tools also include, for example, high specificity rates for miRDeep2 and miRPlant and high sensitivity in plants for miRCat and in *H.*

sapiens for miReap. Based on the cumulative plots, miRCat and miRDeep2 perform well in animal datasets tested (*H. sapiens*, *M. musculus*, *D. rerio*), and miRPlant in the *A. thaliana* datasets. In terms of resource usage, miRCat and miReap perform best at run times, while miRDeep2 and miRPlant require the least memory for their processing.

miRCat2 is based on a new peak selection and feature-filtering algorithm. This means that it can only detect miRNAs with conservative secondary structures and miRNA-specific features. In animals, the pre-miRNAs have a well-defined structure with little fluctuations, making the detection of miRNAs easier. In plants, however, there is a higher degree of variability in miRNA hairpin length [114] and hairpins can contain multiple loops and additional smaller hairpins [21, 36]. These features make the plant miRNA detection challenging. Therefore, rule-based tools, such as miRCat2, miRCat, miRDeep2, miRPlant and miReap, may perform poorly on plant data, missing miRNAs with uncharacteristic features or allowing a large number of false positives. The results for plant data show that miReap performs poorly, displaying low sensitivity and specificity and also the poorest performance on the comparison with mutant datasets. This indicates high false positive and false negative rates and, although it performs better on animal data, miReap should probably not be used for plant miRNA prediction.

Another criterion that influences the outcome of miRCat2 is the read abundance of a miRNA locus: miRCat2 may miss miRNAs that are lowly expressed in the input samples due to the calculations used to test against a random uniform distribution, for the identification of peaks. Nevertheless the detection of low abundance miRNAs is a common issue for all miRNA prediction tools. This is not necessarily a disadvantage, as low read counts would suggest that the miRNA may not be expressed in that particular sample. In another sample where the miRNA is more highly expressed it is more likely that it would be predicted.

The quality of the input datasets can also have a great impact over the results of miRCat2. For aligning the sRNA sequences to the reference genome, miRCat2, as well as miRCat, use PatMaN [153], because of its efficiency in aligning short sequences to large databases. However, PatMaN does not compute any quality checks over the alignments, i.e. it finds all the possible matches for a sequence, irrespective of how many times or in which region of the genome it was matched.

Therefore, short or poor quality sequences might be aligned multiple times, likely even to loci they did not originate from, increasing the level of noise and making the processing more prone to error. Other miRNA prediction tools, such as miRDeep and miRPlant, use Bowtie2 [152] for sequence alignment, which allows the user to customise the maximum number of times a sequence can be matched to the reference genome. Moreover, if the user wishes to align with gaps, it uses a seed approach to ensure the correctness of the method, and outputs the gaped alignment locations only if no full alignments were found. This can offer further insurance for the quality of the mappings. For these reasons, miRCat and miRCat2 can be affected stronger by the poor quality of the input datasets. While this does not affect their ability to classify real miRNAs, it can result in lower specificity, allowing more predictions. This may be one of the reasons, amongst others, for their poor specificity in the soybean and tomato datasets.

miRCat2 generates a score as a mean of ranking its predictions and performs well irrespective of a filtering based on this score. This suggests that the core algorithm is robust.

In terms of run time, miRCat2 compares favourably with miRDeep2, although miReap was faster. For example, on a *H. sapiens* dataset, containing approximately 34.5 million reads, miRCat2 generated the results in 3h50m, while miRDeep2 generated the results in 5h15m (all tests performed on a Linux server with CentOS 5.11 operating system, 144GB of memory and 2 Intel Xeon X5550 processors). In terms of memory usage, the amount allocated for one miRCat2 run is user-defined making it versatile to run on a wide range of specifications.

In conclusion, miRCat2 provides improved identification and characterization of new miRNAs over a range of organisms, that are not predicted by other tools. It should therefore contribute to a better, more in depth understanding of miRNAs, both in plants and animals.

5.7 Summary

In this chapter we have presented the results of miRCat2, compared to miRCat, miRDeep2, miRPlant and miReap. We have presented the sensitivity and specificity of the tools when compared to miRBase, then we have compared their

performance by calculating the fold change in mutants in the miRNA biogenesis pathway. After all the tests conducted, we conclude that miRCat2 performs the most consistently of all tools, having higher statistics in both in animal and plant data.

After discovering new miRNAs and adding them to miRBase, biologists can make good use of them, by studying their roles in the organisms. If these sequences were not annotated, real miRNAs would be overlooked, therefore it is very important to have an image of existing miRNAs as close to reality as possible. In the next chapter we give an example of how these predicted miRNAs are analysed and studied, and present the results for a study of miRNAs in colorectal cancer. The chapter provides a wider view on the area of research on miRNAs and emphasises the importance of having accurate miRNAs annotations, and therefore the need for miRCat2.

Chapter 6

sRNA and miRNA differential expression analysis to study the effects of sulphoraphane treatment on human colorectal cancer

This analysis is submitted as part of the manuscript "Sulforaphane modulates microRNA expression in colorectal cancer cells to potentially implicate the regulation of the CDC25A, HMGA2 and MYC oncogenes", C A Dacosta, C Paicu, I Mohorianu, W Wang, P Xu, T Dalmay, Y Bao"

I have conducted the bioinformatics analysis of the sRNA sequencing data under the supervision of Dr. Irina Mohorianu

6.1 Summary

In this chapter we describe an applied method of conducting analysis of HTS sRNA datasets (generated in Dr. Tamas Dalmay's laboratory). We describe the evaluation of the quality of the sequencing data and how to identify and assess the effect of technical issues occurred either during the library preparation or the sequencing. Next, we give an overview several normalization approaches and select the most appropriate one for this data. Lastly, we explain a procedure for conducting differential expression and identifying the sRNAs that have major roles in the studied conditions. This analysis framework provides an objective overview of the genome-wise expression study. In particular for miRNA studies, it facilitates the identification, with reasonably high accuracy, of a small number of candidate sRNAs which may be linked to the gene regulation. This is of relevance to this thesis as it shows how miRNA predictions can be used in the study of diseases and cancer.

6.2 Introduction

The analysis of sRNA datasets can facilitate the identification of sRNAs which play a role in certain biological processes or diseases. A clear image of the sequences involved provides researchers with a better understanding of the functioning of the organism. The conclusions resulting from such studies, i.e. the set of regulatory sRNAs and the understanding of their mode of action, can then be used, for example, to develop new treatments for diseases (for animals), or to improve quality of products and resistance to pathogens (for plants).

The experimental design of sRNA data consists of multiple conditions: control, treatment or mutant. Usually, experiments have control samples, which are the made from the unaltered cell line/tissue/organism to be studied (wild-type) and can be used as reference, to objectively identify differential expression between the treatment and the control.

Each condition in the experiment can consist of one or multiple libraries (technical or biological replicates). As the HTS sequencing cost has decreased over time, it is more common nowadays to have multiple technical replicates for each

condition, as this gives a better overview of the biological reality, it offers statistical significance for the analysis conducted and it can reduce both FP and FN rates [246, 255].

Because a series of errors can be introduced both during library preparation and HTS sequencing [148, 246, 256–258], it is highly important to check the quality of the libraries produced. The correctness of the results is dependant on the quality of the datasets, otherwise they could be misleading. It is assumed that only 81% of sequences would be free of errors for a successful sequencing experiment [256], thus, we need to make sure we only include the reads that are likely to be correct. A description of the most common possible errors is given in Section 6.3.1.

There are a series of tools that perform HTS sRNA dataset processing, quality assurance and analysis of sRNA experiments. Amongst the most commonly used are: RNA-SeQC [259] (quality check), RSeQC [260] (quality check), Kraken [261] (quality check and datasets processing), UEA sRNA Workbench [4](datasets processing, data visualisation tools, expression profiling [184]), SeqAssist [262] (quality check), sRNAtoolbox [263] (expression profiling and differential expression), FastQC [264] (FASTQ files quality check).

These tools use different methods and provide help when conducting sRNA analysis, however, they lack functionalities (e.g. different tools offer quality check of different criteria). In addition, these tools do not offer a full framework. In the next section, we describe our methods for performing quality check on sRNA datasets, normalization for expression levels and differential expression, which have been successfully used before for HTS sRNA data, on an extended time-course analysis of sRNAs during tomato development and many others [222, 265–267]. The methods we used have been recently implemented as a pipeline of tools for sRNA data analysis, which can be found in the UEA sRNA Workbench [4, 249], providing an end to end processing and sRNA analysis.

6.2.1 Datasets

We applied the methods described below in a study on the effects of sulforaphane (SFN) on the expression of microRNAs in human (*H. sapiens*) colorectal ade-

(A) Lane 1 (Caco-2 Libraries)

Library Name	Index Primer	Treatment	New file Name
LIB 1	1	8 h SFN	B1
LIB 2	2	8 h SFN	B2
LIB 3	3	8 h SFN	B3
LIB 4	4	Control	A1
LIB 5	5	Control	A2
LIB 6	6	Control	A3
LIB 7	7	24 h SFN	C1
LIB 8	8	24 h SFN	C2
LIB 9	9	24 h SFN	C3

(B) Lane 2 (CCD-841 Libraries)

Library Name	Index Primer	Treatment	New file Name
LIB 10	1	8 h SFN	Y1
LIB 11	2	8 h SFN	Y2
LIB 12	3	8 h SFN	Y3
LIB 13	4	Control	X1
LIB 14	5	Control	X2
LIB 15	6	Control	X3
LIB 16	7	24 h SFN	Z1
LIB 17	8	24 h SFN	Z2
LIB 18	9	24 h SFN	Z3

Table 6.1: Library names and information for two sequencing experiments datasets (Caco-2 cell line, CCD-841 cell line). SFN = sulforaphane.

nocarcinoma Caco-2 cells and non-cancerous colorectal CCD-841 cells, to help ascertain the roles of microRNAs in the anti-cancer effects of sulforaphane. Data was generated by Christopher Dacosta (who is part of Dr. Tamas Dalmay's group) for Caco-2 cell line and CCD-841 cell line, to determine miRNA differential expression induced by the effects of sulforaphane in colorectal cancer.

Caco-2 and CCD-841 cells were treated for 8 or 24 h with sulforaphane or DMSO (dimethyl sulfoxide) alone (control). Final DMSO concentrations were 0.05%. Total RNA was then isolated, and microRNAs were cloned as cDNA-based libraries. The libraries were then subject to deep sequencing. Two experiments were produced, one for each cell line (one for Caco-2, one for CCD-841 cell line), the library names, conditions and corresponding file names can be seen in Table 6.1. Each experiment has three replicates for each of the three conditions: control (unchanged cells samples), 8 hours after treatment with sulforaphane and 24 hours after treatment with sulforaphane. The data can be accessed online and downloaded from GEO database [158] under accession number GSE89363 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89363).

To describe this method, we will explain it step by step by performing the analysis on the Caco-2 libraries (for simplicity, we will refer only to this one experiment for quality check, normalization and differential expression analysis, then give an overview and results for the CCD-841 Libraries, as well).

6.2.2 Statistical concepts

Boxplots are used in descriptive statistics as a simple, convenient way of representing data in a plot, giving information about the range, the median and the quartiles of the data. It is produced by drawing a rectangle between the second and third quartiles, usually with a vertical line inside to indicate the median value. The lowest and highest values are shown as vertical lines either side of the rectangle (whiskers), indicating variability outside the quartiles. During this work, we generate standard box plots, where the whiskers extend to 5% and 95% of the data, respectively. If there are outliers in the data, they may be represented as individual points. Box plots are non-parametric, displaying variation of data in samples without making any assumptions of the statistical distribution. The spacings between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data, and show outliers [268].

MA plots [269] is a visual representation of two datasets which have been transformed into the M (log ratio) and A (mean average) scale. It is used to create a visual representation of differential expression between two samples (by plotting the log fold change (M) against the average read count(A)). This plot is used to determine if a normalization method can correct technical biases [270] (after the normalization, there should be little to no DE between replicates in the MA plots).

The Jaccard similarity index [271] is a statistical method of comparing two samples by observing the proportion of members that are common and the proportion of members that distinct. It is defined as the size of the intersection divided by the size of the union of two sample sets, A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6.1}$$

The Jaccard similarity index can take values in the interval $[0, 1]$ where 0 means that there is no similarity between the samples, and 1 means that the samples are identical.

6.3 sRNA datasets processing and quality check

6.3.1 Errors and biases when constructing sRNA libraries

Library preparation is extremely important, because the method used can significantly affect the diversity and abundance of the sRNA that are sequenced [258]. If biases were introduced during this step, relative read counts for different sequences within the same library can be affected. Some sequences present in the biological samples may even be absent in the libraries because of preparation bias [246].

Errors and biases can also occur in the sequencing step. The most commonly used HTS technologies for sRNA datasets, the Illumina [144] and SOLiD [272] platforms, create their libraries by ligating RNA adapters of known sequence to the 5' and 3' ends of single molecules in a purified sRNA population [258]. The adapter-ligated sequences are reverse-transcribed, amplified by PCR to increase the depth of the library, applied to the platform and amplified again to form millions of clusters of DNA of the same sequence, which are then sequenced in parallel [246].

The steps of adapter ligation, reverse transcription and PCR amplification have the potential to induce errors. Adapter ligation is the most important one, as the ligation efficiency is very sensitive to nucleotide base composition at the ligation site and to sRNA modifications. The identity of at least the three 3'-most nucleotides of the sRNA sequence affects ligation efficiency, with a different base preference at each position (5'-nucleotide: A > G $\sim$ C > U; middle nucleotide: A > C > U > G; 3'-nucleotide: A > C > G > U) [273]. There are many studies that focus to solve the adapter bias problem, developing new adapters [265, 274–276] or developing new treatments for improving ligation with one of the adapters (the 3'- or 5'-adapter) [276]. Another method to avoid this issue is to use a ligation-independent library preparation [277], but this method is not perfect, either [246]. Because of this preference, some sRNAs are more likely to be ligated than others, resulting in having higher probability of being sequenced.

Errors also occur during reverse transcription and amplification. The 2'-O-methylation of sRNA (sRNA have a methyl group added to the 2' hydroxyl of the

nucleotide) reduces the efficiency of reverse transcription [274]. The step of PCR amplification can be a problem with sequences that have very low or very high GC content, reducing the likelihood that these sequences will be represented in the final population. Techniques that do not require the initial library amplification have been developed for DNA sequencing and RNA sequencing, providing a less biased library preparation for low GC sequences [278, 279].

Another type of error that HTS technologies suffer from, is introducing substitutions, additions or deletions of nucleotides in the sequenced reads [148]. Although this types of error have low impact (11.5% to 0.1% error rate), it is important to be aware of such errors and make sure they do not affect the overall quality of the samples.

Because of these errors and biasses that can occur during library preparation and sequencing, it is important to check the quality of the datasets, before performing further analysis. The quality check also ensures there have been no errors and biases introduced throughout the processing of the datasets (by programming errors).

6.3.2 Quality check on FASTQ files

The sequencing companies provide the data as files in FASTQ format (for details on the file format see Section). Each nucleotide in a sequence has a Phred quality score associated with it (encoded with a single ASCII character), which represents the probability for a given nucleotide of being one of the four bases. If an ambiguous base exists (low confidence in any particular nucleotide), it is usually denoted with an "N" in the sequence.

To check if the sequencing quality of the libraries (Caco-2 cell lines, for file names and information for each library see Table 6.1) were up to a high standard, for each file, we conducted the following procedure: we calculated the FASTQ quality score at each nucleotide for all sequences, then generated standard box plots using R [280] (as described in Section 6.2.2). The box plots can be seen in Figure 6.1.

We expect to see the boxplots representing high Phred quality scores, in this case all positions having the lower quartile at a high value of at least 64%, which

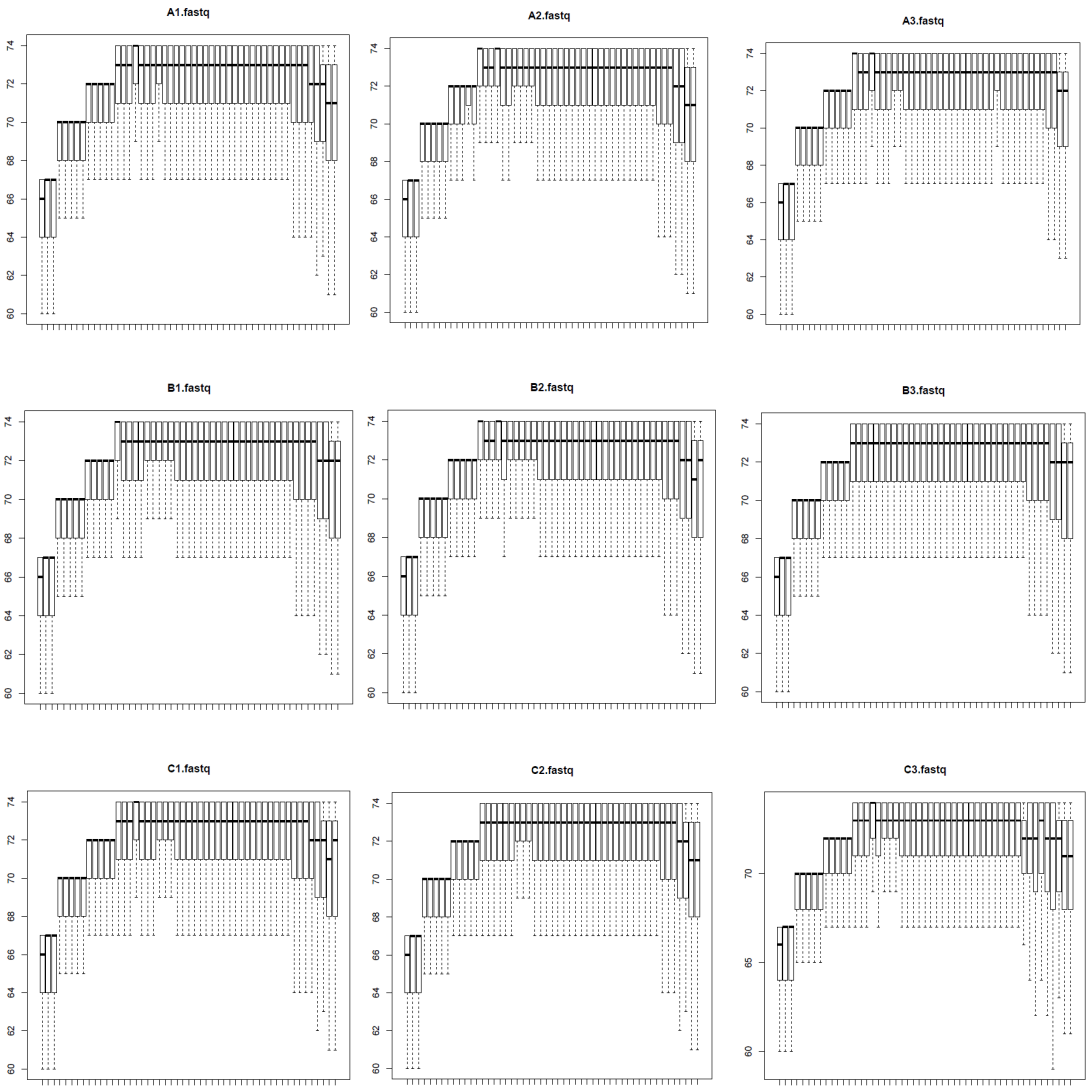


Figure 6.1: Boxplots for the Phred score per nucleotide, for each library. Replicates are based on the same line and can easily be compared. The boxplots show good quality score per nucleotide for all files.

		FASTQ ->FASTA			Proportions (%)	
Condition	File	total	accepted	rejected	accepted	rejected
Control	A1	19,985,254	19,973,288	11,966	99.94	0.06
	A2	26,152,561	26,136,918	15,643	99.94	0.06
	A3	16,176,855	16,167,508	9,347	99.94	0.06
8h SFN treatment	B1	16,002,804	15,993,315	9,489	99.94	0.06
	B2	14,603,739	14,594,340	9,399	99.94	0.06
	B3	13,554,322	13,545,464	8,858	99.93	0.07
24h SFN treatment	C1	23,029,947	23,016,183	13,764	99.94	0.06
	C2	22,223,346	22,210,160	13,186	99.94	0.06
	C3	26,035,997	26,020,325	15,672	99.94	0.06

Table 6.2: **Statistics for transforming files from FASTQ to FASTA format.** After transforming from FASTQ to FASTA format, the proportions of accepted/rejected reads were calculated.

means the data was sequenced with high accuracy. It is normal that there are slightly lower values at the beginning and at the end of the reads, mainly because the HD adapters [265, 266] are found in those specific areas, whereas in the area of the actual sequence (20-25 nts in the middle) there are the higher confidence score interval: 70-75%.

Another sign of good quality is a small degree of dispersion on the boxplots (distance between the quartiles), which represents the variation in confidence between reads. If there was a large variation, it would mean some reads (or nts within some reads) were sequenced with high confidence, while other were sequenced with low confidence. The datasets present low variance 2-4%, assuring consistency and overall high quality.

To transform the FASTQ files to FASTA files, we used a custom made perl script, that selects the id and sequence for each read and outputs it in FASTA format. If the sequence of the read contains an "N" (ambiguous base), the read is discarded. At the end, the script returns the total number of reads in the file, the accepted reads and the rejected reads (the ones containing "N"s). These numbers are helpful to make sure there is a low percentage of sequences with ambiguous bases, assuring the libraries were built with accuracy of sequencing. The numbers for transforming the FASTQ files to FASTA for the libraries can be seen in Table 6.2, all samples having an acceptance percentage of over 99.9%, with only ∼0.06 rejection rate.

The total counts of each library assures the sequencing succeeded with very

high depth, all of them having tens of millions of reads. The replicates have close total counts amongst each other, which could give better chances of having a high degree of similarity, as expected. There is a significant gap in the total count between A2 and A3, however this can possibly be fixed using normalization methods.

Having analysed the boxplots and the statistics from transforming FASTQ files to FASTA, which present low percentages of sequences containing "N"s, we conclude that they give a good overview of the sequencing quality, assuring that all of the libraries from the investigated experiment were made with high confidence.

6.3.3 Adapter removal

The libraries were constructed using HD (high definition) adapters [265, 266, 275, 276] which can reduce the ligation bias, by synthesizing, besides the adapter sequence, an additional 4 random bases on each side of the RNA sequence, called the HD tag (the signatures are assigned all possible combinations of 4 nucleotides in equal amount). The HD adapters are more efficient because sRNAs can anneal to a pool of different sequences (represented by the HD tag-adapter combination) instead of a single adapter sequence. Libraries generated with HD adapters were found to recover more different sRNA sequences and the abundance of a sRNA sequence read correlated in quantity with the real expression level [265, 276].

The adapters are introduced artificially because they are necessary for sequencing, but they are not part of the studied organism, therefore they must be removed before mapping the sequences to the genome. To do this, the first 8 nts of the 3 adapters were identified (sequence TGGAATTC) and trimmed, then four nucleotides on the 5' and 3' ends of the reads were removed (which corresponded to the NNNN tags on the HD adapters), using the UEA small RNA Workbench [4].

The data presented in the Table 6.3 summarizes the proportion of reads at each step of adapter removal. After trimming the adapter sequence, all libraries have a high percentage of over 92% of the reads with lengths over 16 nts, which is an indicator of efficient adapter ligation [281]. Looking at sequences with length

		After adapter removal		Proportions (%)		Sum (%)	After HD tag removal	Proportions (%)
Condition	File	>16	≤16	>16	≤16		>16	
Control	A1	18,594,759	789,750	93.10	3.95	97.05	16,534,880	82.74
	A2	24,722,116	48,093	94.59	0.18	94.77	23,917,718	91.45
	A3	15,463,203	77,505	95.64	0.48	96.12	15,215,090	94.05
8h SFN treatment	B1	14,818,006	155,644	92.65	0.97	93.62	14,253,093	89.07
	B2	13,486,309	190,311	92.41	1.30	93.71	12,868,213	88.12
	B3	12,628,998	303,665	93.23	2.24	95.48	12,234,317	90.26
24h SFN treatment	C1	21,928,868	348,282	95.28	1.51	96.79	20,327,308	88.26
	C2	21,045,058	580,709	94.75	2.61	97.37	19,459,528	87.56
	C3	24,137,624	773,688	92.76	2.97	95.74	21,543,196	82.74

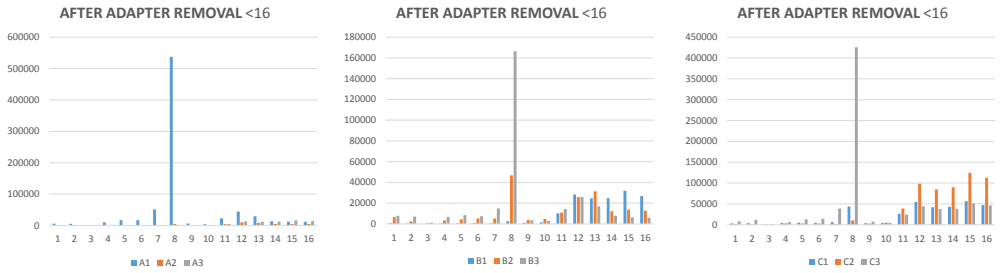
Table 6.3: **Statistics for trimming the HD adapters.** After the adapter removal step, the proportions of sequences with lengths smaller or greater than 16 were calculated. Fragments smaller that 16 nts are counted to verify a potential adapter-adapter contamination. The sum represents the total percentage of sequences that contained the adapter sequence. After the HD tag removal, the proportions of sequences with length greater than 16 were calculated. All percentages were calculated out of the total number of sequences in the FASTQ file.

below 16 nts, the proportions vary between 0.18% (A2) and 3.95% (A1). Column "Sum" in Table 6.3 represents the total number of sequences in the FASTA file that contained the adapter sequence, regardless of its length, concluding that only a small number of reads did not present it.

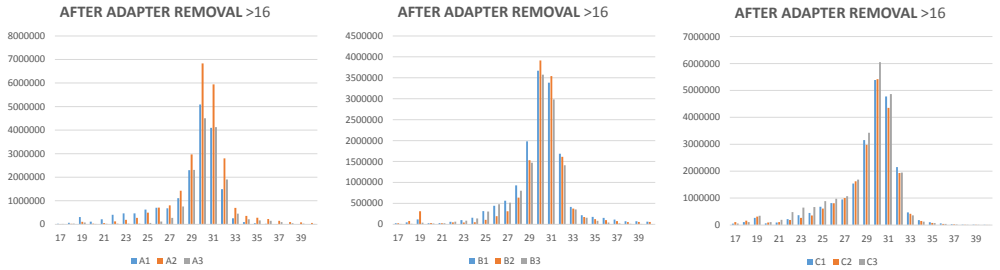
We then created size class distribution histograms (Figure 6.2) on all sequences after adapter removal, to better understand what kind of sRNAs the data contains.

In the plots for the short sequences, we can determine a clear peak at length 8 (see Figure 6.2,a). These sequences are most probably originating from an adapter-adapter ligation, considering the fact that the sequences should still present the HD tags (NNNN) on both 3' end and 5' end. In the two treatment conditions (B and C), we can also observe a higher number of sequences with lengths between 11 and 16, which most likely are adapter-adapter dimers. However, it is expected to have such sequences, and it does not affect the analysis, because their overall proportion out of the total number of reads is small (0.18% to 3.95%, see Table 6.3).

In the plots on the sRNA inserts, we can detect a clear peak between lengths 29 and 32 (see Figure 6.2,b). This is the expected distribution for a successful sRNA sequencing project, because between these lengths reside the miRNAs (which will



a: Size class distribution of sequences with lengths smaller than 16.



b: Size class distribution of sequences with lengths greater than 16.

Figure 6.2: **Size class distribution of sequences with lengths smaller (a) or grater (b) than 16, after adapter removal.** The plots depict the lengths of sequences on the x-axis against the total number of sequences with the specified length on the y-axis. Each condition is plotted separately, to facilitate the comparison of replicates.

be 21-24 nts after the HD tag removal: 21 + 4 (5' HD tag) + 4 (3' HD tag) = 29) and it is expected that the miRNA size class is the most enriched size class of sRNAs in the dataset.

After checking the adapter removal step, we proceeded with validating the trimming of the HD signatures. The proportions of sequences with length greater than 16 after the HD tag removal are also high, with values between 82.74% (A1 and C3) and 94.05% (A3). The proportion of accepted reads changes slightly after the HD tag trimming, suggesting that there were short sequences amongst the reads (which are discarded, because they are most likely degradation products) (see Table 6.3).

The files were then converted from redundant to non-redundant FASTA for-

mat, in which the reads with the same sequence are collapsed, each unique sequence being output only once, together with its total count. We then computed the complexity of each dataset, by dividing the number of non-redundant sequences to the number of redundant sequences [222]. The values for redundant, non-redundant number of sequences and the dataset complexity can be seen in Table 6.4.

The complexity can take values in the interval $(0, 1]$, a value closer to 1 meaning that the sequences have low abundances, with the majority having counts of 1. A value closer to 0 means that the sequences have high abundances, with many sequences being highly expressed (presumably miRNAs). All the datasets in the experiment present extremely low complexities, 0.03-0.05%, which validates the high depth of the sequencing (see Table 6.4).

We created size class distribution histograms after the HD tag removal, on redundant and non-redundant sequences, then plotted the complexity for each dataset (Figure 6.3).

The size class distribution plots on the redundant data confirm that the most enriched class is indeed the miRNA size-class (21-24 nts long). The highest peaks are at lengths 22-23 nts, which are the most frequent lengths for *H. sapiens* miRNAs. The size class distribution histograms on the non-redundant data should ideally look like the distribution for A1 and B2, with low level for size 17, rising until sizes 22-23, then decreasing again. While there are low levels for sizes greater than 25 for all libraries, some present high levels between sizes 17 and 20. These sequences might be degradation products (possibly resulting from miRNA degradation). Because they have low read counts they are not a concern at this stage in the analysis. The plots depicting the complexity of the datasets show a deep valley at sizes 21-24, suggesting the sequences of these sizes are the most abundant in the dataset, much more abundant than the other sizes (see Figure 6.3).

Following these tests at each step of the adapter removal, we observe that all libraries have high percentages of accepted reads with at least 16 nts in length, after trimming both the 3' adapter and the HD tag. Then we found the miRNA size class (21-24 nts) to be the most enriched size class in all datasets, and also having the lowest complexity (few sequences having high read counts). These are

Figure 6.3: **Size class distribution of sequences after HD adapter removal, considering redundant and non-redundant counts.** The plots depict the lengths of sequences on the x-axis against the total number of sequences with the specified length on the y-axis. The complexity represents the number of non-redundant sequences divided by the number of redundant sequences. Each condition is plotted separately, to facilitate the comparison of replicates.

strong indicators that all libraries have a good quality after this step.

6.3.4 Genome matching

The FASTA files in non-redundant format were matched full length against the human genome (version 38) using PatMaN [153] allowing 0 mismatches and 0 gaps. It is important to validate the proportion of genome matching sequences, to make sure there was no contamination (e.g. other organisms, bacteria) during the preparation of the libraries in the laboratory and we perform the analysis on

		After HD tag removal			Genome matching			Proportions (%)	
Condition	File	Red	NR	Compl	Red	NR	Compl	Red	NR
Control	A1	16,534,880	679,579	0.041	13,539,159	377,735	0.028	81.88	55.58
	A2	23,917,718	892,076	0.037	19,211,757	487,753	0.025	80.32	54.68
	A3	15,215,090	573,254	0.038	12,200,936	270,743	0.022	80.19	47.23
8h SFN treatment	B1	14,253,093	698,662	0.049	11,096,534	344,527	0.031	77.85	49.31
	B2	12,868,213	567,364	0.044	10,258,639	304,402	0.030	79.72	53.65
	B3	12,234,317	627,355	0.051	9,484,717	335,746	0.035	77.53	53.52
24h SFN treatment	C1	20,327,308	755,610	0.037	16,136,765	379,296	0.024	79.38	50.20
	C2	19,459,528	773,920	0.040	15,459,590	382,986	0.025	79.44	49.49
	C3	21,543,196	820,294	0.038	17,208,841	415,426	0.024	79.88	50.64

Table 6.4: **Number of sequences in redundant (Red) and non-redundant (NR) formats after HD tag removal and after genome matching.** The complexity (Compl) represents the number of non-redundant sequences divided by the number of redundant sequences, after each step. The proportions represent the percentage of sequences that mapped to the genome.

reads free of sequencing errors [256].

The number of redundant and non-redundant sequences, complexity and proportions of sequences that matched to the human genome are presented in Table 6.4. The percentages are calculated out of the total number of sequences in the FASTA file, after HD adapter trimming.

Compared to the previous step, we observe a major decrease in complexity for all datasets (for example, from 0.041% to 0.028% for A1), which means the majority of the sequences that matched to the genome are sequences with low complexity levels (and high read counts), and that a large number of sequences with low counts did not originate from the *H. sapiens* cell line. This logic is also confirmed by the proportion of non-redundant reads that matched to the genome: $\sim$ 50% in all libraries.

However, the redundant proportion of mapping reads is very high for all datasets: from 77.53% in B3 to 81.88% in A1. This is a good indicator that the majority of reads, and especially the high abundance reads originate from the desired organism.

In the next step, we eliminate reads with low sequence complexity from the non-redundant FASTA files [257]. A read has a low sequence complexity if it is represented in proportion of at least 75% only by one nucleotide (e.g. AAAAAAAAAACAATAAAAAAAAA) or a combination of only 2 nucleotides

		Genome matching			Proportions (%)	
Condition	File	Red	NR	Compl	Red	NR
Control	A1	12,249,604	83,644	0.007	83.12	53.15
	A2	17,264,912	149,208	0.009	81.48	54.46
	A3	11,123,437	62,451	0.006	81.44	46.01
8h SFN treatment	B1	9,954,438	88,639	0.009	79.47	50.00
	B2	9,329,583	69,063	0.007	80.85	52.57
	B3	8,370,321	91,143	0.011	78.89	53.78
24h SFN treatment	C1	14,503,754	89,972	0.006	80.65	47.86
	C2	13,727,533	93,375	0.007	80.71	47.40
	C3	15,457,968	93,581	0.006	81.18	47.11

Table 6.5: Number of sequences in redundant (Red) and non-redundant (NR) formats, complexity (Compl) and proportions of genome matching after eliminating reads with low sequence complexity.

(e.g. ACACACACACACACACACACACACAC). It is recommended to remove such sequences [257] and this should not affect the results of the analysis, as these sequences are generally low abundance, do not match to functional sRNAs, and most often do not originate from the biological sample, but are an artefact of sequencing.

After eliminating reads with low sequence complexity, we recalculated the number of redundant and non-redundant sequences, complexity and proportions of sequences that matched to the human genome (see Table 6.5). In this case, we notice the complexities of the datasets have dropped even lower (for example, from 0.028% to 0.007% for A1), while the proportion of redundant mapping reads has increased (for example, from 81.88% to 83.12% for A1). This suggests that the reads with low sequence complexity that we filtered out were lowly expressed and were not matching the human genome, therefore the decision of removing them from further analysis is beneficial.

To check that the size class distribution has not been affected by filtering out the reads with low sequence complexity and genome matching, we have replotted it and is presented in Figure 6.4. We observe there has been no change in the shape of the plots, keeping the same features for lengths 21-24: a high peak for the redundant sequences and a deep valley for the complexity charts. This proves the data has good quality after genome matching.

To sum up, the high proportion of sequences mapping to the human genome and low complexities, together with the high proportion of 21-24mers in the

Figure 6.4: **Size class distribution of sequences after eliminating reads with low sequence complexity and matching to the genome, considering redundant and non-redundant counts.** The plots depict the lengths of sequences on the x-axis against the total number of sequences with the specified length on the y-axis. The complexity represents the number of non-redundant sequences divided by the number of redundant sequences. Each condition is plotted separately, to facilitate the comparison of replicates.

datasets, are strong indicators that the libraries were constructed with good quality up to this point.

6.3.5 Replicates validation

Replicates are samples made from the same biological condition, therefore they should be very similar to each other. They should contain the same sRNA sequences and the same sequence should have fairly close abundances in different

	A1	A2	A3
A1	1	0.852	0.812
A2	0.852	1	0.820
A3	0.812	0.820	1

	B1	B2	B3
B1	1	0.828	0.764
B2	0.828	1	0.807
B3	0.764	0.807	1

	C1	C2	C3
C1	1	0.874	0.898
C2	0.874	1	0.857
C3	0.898	0.857	1

Table 6.6: The Jaccard similarity index in top 1000 non-redundant sequences from each library. The replicates from each condition are compared with each other.

	A1	A2	A3
A1	1	0.285	0.182
A2	0.163	1	0.139
A3	0.211	0.281	1

	B1	B2	B3
B1	1	0.218	0.228
B2	0.294	1	0.282
B3	0.238	0.219	1

	C1	C2	C3
C1	1	0.218	0.220
C2	0.208	1	0.212
C3	0.208	0.210	1

Table 6.7: The fraction of sequences in the intersection of each two samples. The fractions are calculated from the total non-redundant sequences of each library on each row. For example, the number of sequences found in the intersection between A1 and A2 represent 0.285 of the total sequences in A1 and 0.163 of the total sequences in A2.

	A1	A2	A3
A1	1	0.247	0.349
A2	0.664	1	0.406
A3	0.215	0.089	1

	B1	B2	B3
B1	1	0.282	0.272
B2	0.172	1	0.244
B3	0.287	0.412	1

	C1	C2	C3
C1	1	0.261	0.259
C2	0.279	1	0.262
C3	0.278	0.263	1

Table 6.8: The fraction of sequences in the specific difference of each two samples. The fractions are calculated from the total non-redundant sequences of each library on each column. For example, A1 has 0.664 specific sequences when compared to A2, while A2 has 0.247 specific sequences when compared to A1.

replicates. If this does not hold, we might need to remove one or multiple replicates from one condition, because they are not statistically similar.

To compare the similarity between the replicates, we first computed the Jaccard similarity index on top 1000 most abundant sequences. We computed the index in this way, because these are the most important in the samples (with highest counts and low possibility of being false entries).

Because we expect the replicates to be alike, we want their Jaccard similarity index to be as high as possible. Table 6.6 shows the Jaccard similarity index for the three conditions. For Condition A, there is a high similarity between A1 and A2 (0.852), but A3 has a slightly lower value when compared with both A1 and A2 ($\sim$0.81). In condition B, B3 seems to have a lower similarity to the others, with fairly low coefficient of only 0.764 when compared to B1. In condition C all three replicate comparisons have high similarity (0.857 - 0.898).

To check in more detail the replicates similarity, we then computed the frac-

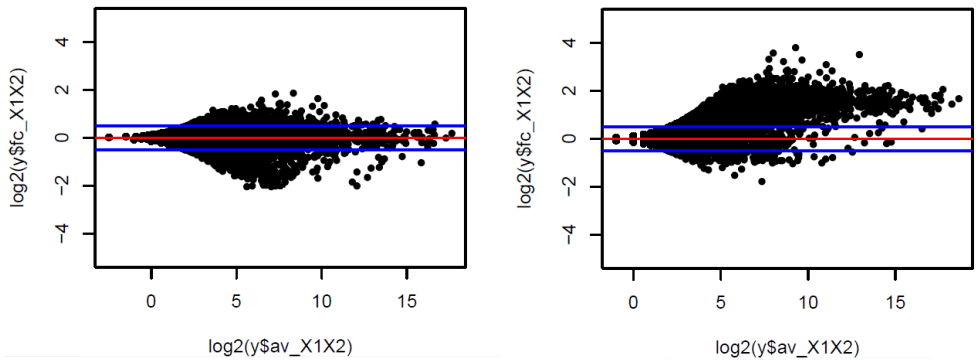


Figure 6.5: Examples of MA plots, on raw expression levels, after genome matching. The shape in panel (A) is a good distribution, while in panel (B) is a bad distribution when comparing replicates.

tions of sequences in the intersection of the datasets and in their specific differences (in one of the samples, but not the other). In the case for intersection, it is better if the fractions are larger (see Table 6.7). We notice that A2 has a low percentage both when compared to A1 and A3. Conditions B and C have good intersection fractions. In the case for specific difference, it is better if the fractions are smaller (see Table 6.8). A1 and A3 have very high percentages of specific sequences, and also B2 has a large percentage when compared to B3. Condition C has very good fractions for all samples.

In the next step, we produced MA plots (using R [280], as described in Section 6.2.2) to check if there are DE sequences amongst replicates. Because the same sequence should have fairly the same expression level in different replicates, we expect there are only few cases of DE. It is normal to have DE sequences for low counts, because any small difference would make the read seem differentially expressed in this case (e.g. a sequence with counts of 1 and 2 would seem DE). MA plots are also a good indicator if the samples are normalizable (if by applying a normalization method, the technical errors, such as DE, can be corrected).

In Figure 6.5 we present examples of MA plots that depict a high (A) and a low (B) similarity when comparing two replicates. If two replicates are similar, the points in the MA plot should be centred around the horizontal line in 0 (the red line) or around a line parallel to it. It is acceptable to have points outside the

blue lines (horizontal lines at 0.5 and -0.5), especially closer to the smaller values on the x-axis, but the points should converge to a tighter distribution towards the larger values, resulting in a fan-like overall shape (Figure 6.5, (A)). However, it is recommended that the fan is not too wide [282]. If the distribution is not centred around a horizontal line, being skewed below or above it (like in Figure 6.5, (B)), then one of the samples has DE sequences that cannot possibly be corrected through normalization: by trying to normalize the data so that the expression levels of DE sequences between two replicated are corrected, it is likely that other sequences would become DE as a results of the normalization. Therefore, samples that present such malformed MA plots should be avoided, because they do not present consistency in expression levels throughout their dataset.

We generated MA plots, comparing all replicates from each condition, dividing the data by size (corresponding to miRNA lengths, from 21 to 24 nts, as we are most interested in miRNAs). The MA plots comparing all replicates for all conditions are presented in Appendix B, Figures 1, 2 and 3.

For condition A, we can see the MA plot for A1 and A2 at size class 24 is suboptimal, however, because miRNAs with length of 24 represent only a small fraction of *H. sapiens* miRNAs, we decided it is not an issue. Looking at the MA plots between A2 and A3, and A1 and A3 we can observe in all cases a very wide fan, without a clear convergence of the points on the horizontal line. This indicates there might be a similarity issue between them, A3 possibly being the outlier (because it is common to both cases).

For condition B, the comparison between B1 and B3 has a very large fan, with a poor distribution of points, even towards larger numbers on the x-axis, suggesting these two datasets are not very similar. For condition C, all MA plots have a nice distribution, suggesting all three replicates are highly alike.

To further check the DE between the replicates, we produced box plots on the offset fold change (fc), grouped by size classes (see Figure 6.6). We expect the median lines for each box to be aligned inside each plot, ideally on the horizontal line in 0 (meaning that the majority of sequences are not DE), or parallel to it (meaning that the DE expression can be corrected through normalization). Another sign of good quality are small boxes (little variance in expression levels).

Looking at the box plots for condition A, we observe the median lines have

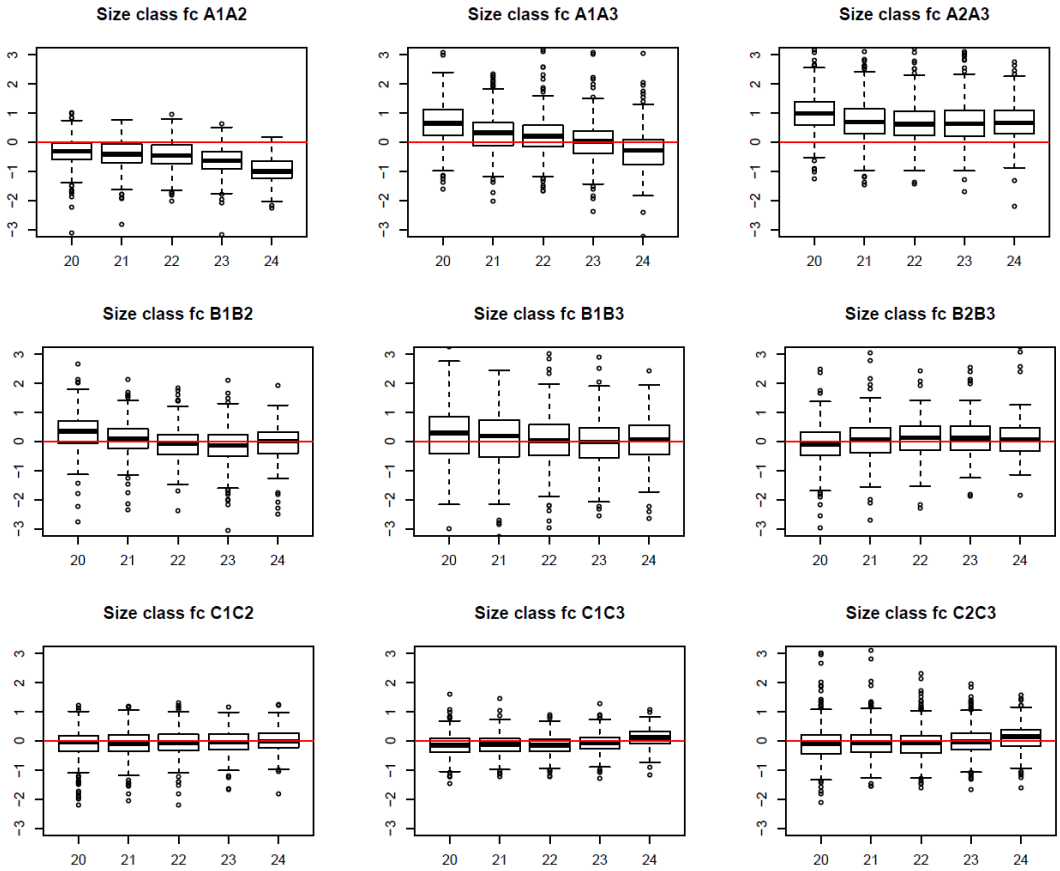


Figure 6.6: Box plots on the offset fold change between replicates. The box plots were created separately for each size class (20-24 nts).

a wavy shape in the comparison between A1 and A3, which is a sign of poor similarity, and also hard to correct through normalisation, suggesting that one of these two samples might have issues. The medians are aligned well for comparing A1 to A2 and A2 to A3, the fact that they are slightly below, respectively above the red line is not a problem. In condition B, the boxes are aligned nicely, however, they are very wide in the comparison between B1 and B3, suggesting variance from sequence to sequence. The boxplots in condition C present no faults.

After carefully analysing the Jaccard similarity indexes, the intersection and the specific difference fractions, the MA plots and the box plots on the offset

		miRNAs		rRNAs		snoRNAs		tRNAs	
Condition	File	Red %	Compl	Red %	Compl	Red %	Compl	Red %	Compl
Control	A1	69.86	0.0021	3.24	0.0309	0.46	0.0576	1.09	0.0203
	A2	70.64	0.0018	2.28	0.0344	0.37	0.0572	1.06	0.0299
8h SFN	B1	68.51	0.0023	2.49	0.0375	0.37	0.0657	1.35	0.0346
treatment	B2	71.82	0.0022	1.50	0.0321	0.21	0.0708	0.77	0.0461
	C1	71.35	0.0018	3.33	0.0266	0.41	0.0577	1.54	0.0195
24h SFN	C2	72.65	0.0019	3.38	0.0281	0.40	0.0611	1.56	0.0193
treatment	C3	71.06	0.0018	3.67	0.0274	0.47	0.0556	1.62	0.0183

Table 6.9: Proportions and complexity of redundant sequences that mapped to mature miRNAs, precursor miRNAs, rRNAs, snoRNAs and tRNAs. Mature and precursor miRNAs are shown only once, since their numbers are extremely close.

fold change between replicates, we decided to eliminate samples A3 and B3 from the analysis, as they appear to have the least similarities with the other samples from their respective conditions. Keeping them in the analysis could influence and distort the results, because they are not statistically consistent with the other replicates.

6.3.6 Datasets composition

Next, we checked for the composition of the data in the remaining samples, to ensure the reads correspond to functional sRNAs and not to other genome regions, which are not of interest to this analysis (e.g. coding regions). We expect that the majority of sequences would come from miRNAs and miRNA precursors, and only a smaller part should belong to other categories (rRNA, tRNA, snoRNA).

To find the exact proportion of reads that belong to each of the above mentioned sRNA types, we mapped the reads to *H. sapiens* mature miRNAs and their precursors (taken from miRBase) [5], and to rRNA, tRNA and snoRNA datasets taken from RFAM [160]. All libraries were aligned full length using PatMaN [153], allowing up to 2 mismatches and 0 gaps. Instead of requiring a strict match to miRBase sequences, we allowed up to 2 mismatches when mapping the read sequences to *H. sapiens* annotations from miRBase, to account for post-transcriptional modifications and different isomiRs (sequences that are 1-2 nts shorter/longer than the canonical miRNA, possibly with different 5 and 3 ends). One downside of this is that reads for different miRNAs that are very close in sequence (e.g. hsa-let-7a-5p and hsa-let-7f-5p) can be mixed up, but this can be

easily corrected by checking the individual sequences. In this way we make sure that different isomiRs (that could be more abundant than the miRBase forms [283]) were not missed, as well as to allow for any artefacts of library preparation and the possibility of some sequencing errors.

The proportion of redundant sequences that mapped to mature miRNAs, miRNA precursors, rRNAs, snoRNAs and tRNAs, together with their complexity, are presented in Table 6.9. While the percentage of miRNA mappings are very high (from 68.51% in A3 to 72.65% in C2), and their complexity is very low (∼0.002 in all samples), the other sRNAs have percentages of under 5% and much higher complexities (at least 10 times higher than miRNAs). The high levels of miRNAs, together with the similar proportions of miRNAs between replicates is a good indicator of datasets quality [246].

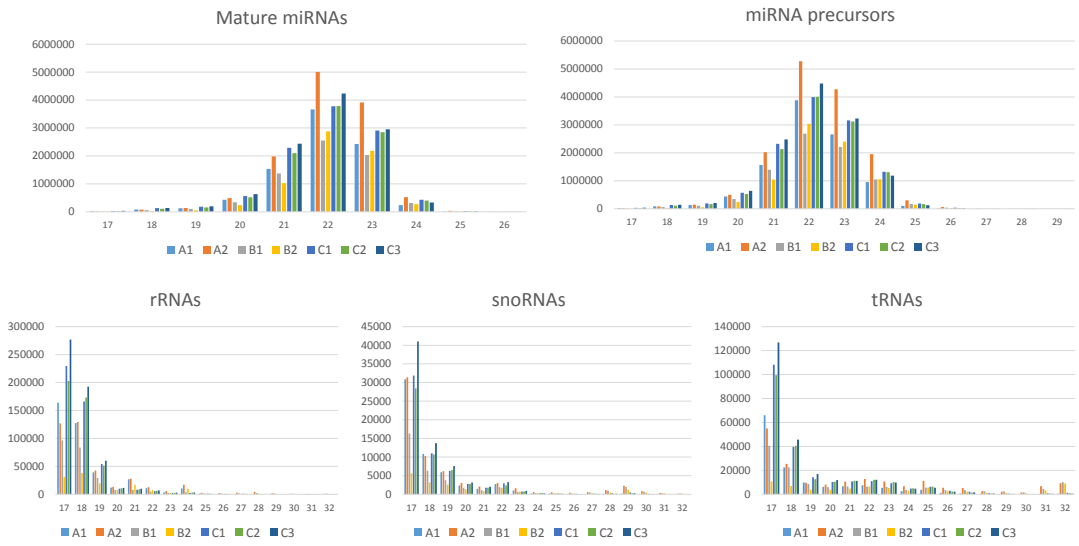


Figure 6.7: Size class distribution histograms for mature miRNAs, miRNA precursors, rRNAs, snoRNAs and tRNAs, on redundant sequences, after genome matching. The sequences were aligned full length to miRBase and RFAM annotations using PatMaN, allowing up to 2 mismatches and 0 gaps.

Generating the size class distribution histograms for the functional sRNA mappings (see Figure 6.7), we observe the clear peak at lengths 21-23 both for mature and precursor miRNAs. For rRNAs, snoRNAs and tRNAs we detect large amounts of sequences of 17-19 nts, which are most probably degradation

products. The plots suggest that the majority of 17-19 nt sequences in all datasets are derived from tRNA, rRNA and snoRNA loci (and not miRNA loci) [246]. However, there are also some small peaks at lengths 24, 29 and 31-32, respectively, which correspond to the respective functional sRNAs. This ensures the datasets have the required sRNA composition.

6.3.7 Quality check conclusions

We have checked the quality of the datasets after each processing step. First, we checked the quality of the FASTQ files, then all FASTQ files were converted to FASTA format, and sequences containing 'N's were discarded. Next, the adapter sequence was identified and trimmed, and the 'NNNN' sequences corresponding to the HD tag were removed. Then the files were converted from redundant to non-redundant format. The resulting reads were matched full length against the human genome, before and after eliminating reads with low sequence complexity. For each of these steps we confirmed that high proportions of the valid sequences have passed the processing. We produced size class distribution histograms and complexity line plots to check that the most enriched size class, with the lowest complexity is the miRNA size class. All samples passed these filters.

We then checked for the replicates similarity and consistency, by calculating the Jaccard similarity index, the fraction of sequences in the intersection and in the specific difference, and by producing MA plots and box plots on the offset fold change between replicates, grouped on size classes. After these steps, we decided it is beneficial for the analysis to eliminate samples A3 and B3 from further processing, as they do not present the expected features of similarity when compared to their replicates. We then checked that the datasets contain a high percentage of sequences corresponding to annotated miRNAs.

6.4 sRNA datasets normalisation methods

Sample variations can occur, including between-sample differences such as library size (i.e. sequencing depth) [245] as well as within-sample sequence-specific effects related to sequence length [284] and CG content [285]. To correct these er-

rors and differences among replicates and make samples from different conditions comparable, we need to first normalize the data. There are several normalization methods that are suitable for sRNA data [247], the most commonly used being RPM (reads per million) [245, 286], Quantile normalization [287, 288] and Bootstrapping [249, 282, 289, 290].

Because the normalisation is data-dependent (one method can be more suitable than the others for certain datasets), we need to check which method is best suited for these samples. RPM is inefficient in some cases, because it ignores the number of distinct reads within each sample [246]. Quantile normalisation is limited by the fact that it assumes a similar distribution of abundances per distinct reads among all libraries being normalized [246], which may lead to over-correction and increased within-condition variability [247]. Bootstrapping can lead to over- or underrepresentation of certain sequences [246]: by over-selecting particular reads, the proportions between sequences within the same sample can be altered. The method that corrects the most differences between replicates will be chosen as the most appropriate normalisation method.

6.4.1 RPM normalisation

The RPM method [245, 286] originally consists in dividing the read abundance by the total count of the library, and then multiplying it by an a priori defined normalization total (when the method was first proposed, the total was 1 million) (i.e. equation 6.2 and 6.3).

$$normalized\_read\_count = \frac{read\_count}{total\_library\_count} * 1000000 \tag{6.2}$$

$$normalized\_read\_count = \frac{read\_count}{total\_library\_count} * MTC \tag{6.3}$$

The total library count is the redundant count, computed by adding the counts of all sequences in that library. Equation 6.2 is applied on all sequences to get the normalised counts.

Because the total counts of the libraries are so large, each having tens of millions of reads (see Table 6.5), using a flat million value for normalisation would

artificially decrease the expression levels, possibly distorting them. Therefore, we use the median total count (MTC) instead (equation 6.3), which is defined as the median value for all library total counts in the experiment.

After normalising all samples, we replotted the MA plots and box plots on the offset fold changes between replicates, to evaluate its effect on the FC between replicates. The box plots are presented in Figure 4, showing all medians aligning on the horizontal line in 0: RPM corrected the issue for condition A, where comparisons A1 to A2 and A2 to A3 would be below, respectively above the line. The rest of the box plots present the same features as before normalisation. The MA plots are presented in Figures 5, 6 and 7. We notice a good distribution of the points for comparisons A1 to A2, B1 to B2 and all comparisons in condition C, the other ones having a very wide fan.

The box plots and the MA plots show that RPM is a suitable method for normalizing the datasets.

6.4.2 Quantile normalisation

Quantile normalization [287, 288] is used to make two distributions identical in statistical properties. The Quantile normalisation is performed by pooling together all sequences from all samples (even if they are not expressed in all of them), then sorting the samples, then overriding each value with the average (usually, arithmetic mean) of the values from all samples for a specific rank. The highest value in all cases becomes the mean of the highest values, the second highest value becomes the mean of the second highest values, and so on. At the end, all samples will have the same distribution.

We performed quantile normalization on all samples and produced box plots (see Figure 8) and MA plots (see Figures 9, 10 and 11). Although the MA plots show the required shapes, in the box plots we notice that the medians are not aligned on the horizontal line in 0, as expected, both for condition A and condition B. This occurs because of the presence of a much higher proportion of 0 counts in certain samples.

Therefore, the quantile normalisation method can be used for this data, but is suboptimal.

6.4.3 Bootstrapping normalisation

Bootstrapping [249, 289, 290] is used in statistics as a resampling technique and can be used for normalisation by randomly selecting sequences until a predefined number of reads (or a percentage) have been chosen. This method favours the more abundant reads as more likely to be picked, while the less abundant ones are more likely to be eliminated. This can be helpful, because low count reads are more likely to be false positives, are more often dataset specific and can appear as differentially expressed between replicates.

		Before	After bootstrapping							
Condition	File	% Red	95%	90%	85%	80%	70%	60%	50%	Min total
Control	A1	83.12	83.12	83.12	83.12	83.12	83.13	83.11	83.11	83.14
	A2	81.48	81.48	81.48	81.49	81.48	81.48	81.48	81.48	81.49
	A3	81.44	81.44	81.44	81.43	81.44	81.44	81.46	81.44	81.44
8h SFN treatment	B1	79.47	79.47	79.46	79.46	79.47	79.47	79.47	79.46	79.46
	B2	80.85	80.85	80.85	80.85	80.84	80.85	80.85	80.86	80.84
	B3	78.89	78.89	78.88	78.89	78.88	78.89	78.88	78.90	78.88
24h SFN treatment	C1	80.65	80.65	80.66	80.65	80.66	80.65	80.67	80.65	80.65
	C2	80.71	80.71	80.71	80.71	80.71	80.72	80.71	80.70	80.70
	C3	81.18	81.18	81.18	81.18	81.18	81.19	81.18	81.17	81.19

Table 6.10: Proportions of sequences that map to the genome after bootstrapping at different percentages.

We performed bootstrapping at the following percentages: 95%, 90%, 85%, 80%, 70%, 60%, 50%, to check whether the resampling distorted the proportions of reads within the same sample. If this is not the case, then the proportion of sequences that map to the genome after each step of bootstrapping should be consistent (the presence and proportion of the mapping sequences is not affected, only their counts might be decreased). In Table 6.10 we present the proportion of mapping sequences after each step of bootstrapping, which remain consistent. Box plots for offset fold changes between replicates after bootstrapping to 50% and 70% are presented in Figures 12 and 13.

However, we must choose the most appropriate percentage for the data, which should keep the read counts as close as possible to their original counts (the highest possible percentage). Therefore, we chose to apply bootstrapping to the minimum total (MT) of the libraries (total count of the B3 library). The MT

is the highest number that allows us to apply bootstrapping on all samples (we cannot select more than 100% for the MT sample), and because it represents more than 50% of the total in each of the samples, and it statistically correct to use it (see Table 6.10) [249]. Bootstrapping at the MT values means resampling all libraries, such that each of them will have a total of MT sequences.

Box plots on the data after bootstrapping to the MT value are presented in Figure 14 and MA plots are presented in Figures 15, 16 and 17. While the box plots look reasonable, all MA plots are skewed to a side, not respecting the fan shape. Because this happens in all sample comparisons, even in condition C, which presents very high similarities between replicates, we decided that bootstrapping is not a suitable method for normalizing the data.

6.4.4 Normalization methods conclusions

We have normalised the data using three techniques: RPM, quantile normalization and bootstrapping. For each normalisation method, we generated box plots and MA plots, to see which method is most suitable for the data in the libraries. By analysing the plots we found the RPM normalisation to correct the most errors for the replicates, and therefore we continued the data analysis based on the RPM normalised read counts. The box plots and MA plots have also confirmed that the decision of eliminating A3 and B3 from the analysis is appropriate, because no normalization method succeeded in correcting their errors.

6.5 sRNA differential expression analysis

To find DE miRNA sequences, we have selected all the reads mapping to miRNAs (by mapping the samples to miRBase mature miRNAs, as described above) and used the offset fold change method [291, 292] on their RPM normalised counts.

This technique of calculating the differential expression has been compared to two other methods: unusual ratio [293], and modified SAM [294] in a paper publishing FiRePat [295], a tool for studying expression patterns, and showed that there is a good agreement between the methods, with more than 90% overlap between the outputs of each these techniques [295]. Other papers comparing

methods for differential expression analysis are [296], [297] and [292]. Because of its simplicity, we used the offset fold change method to identify DE sequences. We now review this method.

First, we added an offset of 20 to all read counts (value was chosen by empirical observation), to avoid lowly expressed reads being taken as DE reads.

Second, we compared the expression levels from each two conditions (A to B, A to C and B to C). If there is a two fold change in the expression levels (one condition has at least double the number from the other condition), then the read is considered DE. To make the comparison, there are two methods of deciding the expression level of a read for a condition, based on one or multiple replicates:

- Median or average - if there are at least three replicates, the median of the read counts from the replicates can be chosen for each sequence as the most representative value for the condition. If there are at least two replicates, the average between them can be used as the value for the respective condition. Then the values of the medians/means are compared directly.
- Confidence intervals - is a more stringent way of calculating the DE. For each sequence, we determine an interval in which the condition can take value, by selecting the minimum and the maximum read count from the replicates of the respective condition, which will be the limits of the confidence interval for the respective condition. If there is only one sample per condition, confidence intervals can be derived by adding and subtracting 10% of each read count to create a maximum and minimum value. After we calculate the confidence intervals for both conditions that we want to compare, if the intervals overlap, then we compare their mean value. If not, then we compare the minimum value from the higher interval to the maximum value of the lower interval.

  Figure 6.8 shows the confidence intervals for comparing the expression levels of hsa-miR-27b-5p in conditions A and B. Because the confidence interval in condition A is higher and non-overlapping with the interval in condition B, the minimum value of A (1330) is compared to the maximum value of B (410), and because there is more than a two fc, it results that this miRNA is downregulated in condition B.

We used confidence intervals to make the comparisons and the DE sequences are presented in Table 6.11. It is noticeable that for each comparison, all miRNAs

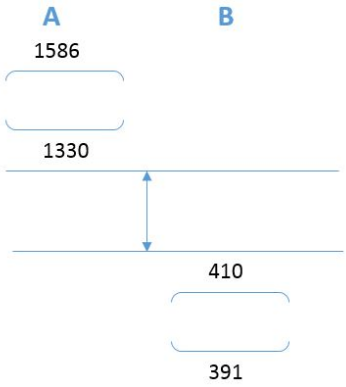


Figure 6.8: Confidence interval for a hsa-miR-27b-5p for conditions A and B.

DE type in second condition	A and B	A and C	B and C
Downregulated	hsa-miR-27b-5p hsa-miR-25-5p **hsa-miR-4286** hsa-miR-339-3p hsa-miR-33b-3p hsa-miR-27a-5p		
Upregulated		**hsa-miR-10a-5p** **hsa-miR-10b-5p** hsa-miR-182-5p hsa-miR-146a-5p hsa-let-7a-5p hsa-let-7f-5p	**hsa-miR-4286** **hsa-miR-10a-5p** **hsa-miR-10b-5p**

Table 6.11: Differentially expressed sequences corresponding to miRNAs. Sequences in bold represent miRNAs that were found DE in more than one comparison.

are either up- or downregulated, but not mixed. This suggests that all miRNAs are influenced in the same way by the treatment: after 8h SFN treatment (B), all miRNAs suffer a decrease in levels, but after 24h SFN treatment (C), all miRNAs experience a growth of expression. The fact that some miRNAs are found DE both in comparison A to C and B to C (hsa-miR-10a, hsa-miR-10b) is evidence that these miRNAs might have actual roles in colorectal cancer.

miRNA	Cond.	DE type	miRNA	Cond.	DE type	miRNA	Cond.	DE type
hsa-miR-26a-5p	A	UP	hsa-miR-10a-5p	B	DOWN	hsa-miR-26a-5p	B	DOWN
hsa-miR-141-3p	A	DOWN	hsa-miR-10b-5p	B	DOWN	hsa-miR-26a-5p	C	DOWN
hsa-miR-27a-3p	A	DOWN	hsa-miR-192-5p	B	DOWN	hsa-miR-26a-5p	C	UP

Table 6.12: miRNAs that are DE between replicates of the same condition.

We checked to see if there are any DE miRNAs between replicates from the same condition, to make sure the DE between conditions is correct (see Table 6.12). We notice that miRNAs hsa-miR-27a-3p, hsa-miR-10a-5p, hsa-miR-10b-5p are all DE between replicates of the same condition, which means there could probably be some issues with their sequences and cannot be very confident about their DE between conditions at this point. Full results are presented in Supplementary_Lane1_DE.xlsx.

However, we found many sequences that corresponded to one mature miRNA, with little variations from the original miRNA sequence (isomiRs). For example, in Table 6.13 are presented some of the hsa-miR-10a-5p and hsa-miR-10b-5p isomiRs that were found to be DE. When comparing conditions A and C, we found 15 distinct sequences that were DE and 22 distinct sequences that were not DE corresponding to hsa-miR-10a-5p, all presenting high abundances (over 100 reads). In this case, how do we chose which sequence is the most representative for each miRNA?

miRNA	isomiR
	TACCCTGTAGAACCGAAT
	TACCCTGTAGAACCGAA
	TACCCTGTAGAACCGAATTT
hsa-miR-10a-5p	TACCCTGTAGATCCGAATT
hsa-miR-10b-5p	TACCCTGTAGATCCGAATTT
	TACCCTGTAGATCCGAAT
	TACCCTGTAGAACCGAATT
	TACCCTGTAGATCCGAA

Table 6.13: Differentially expressed isomiRs corresponding to hsa-miR-10a-5p and hsa-miR-10b-5p.

IsomiRs may be 1 or 2 bases longer/shorter than the canonical miRBase sequence, with shifted 5 and 3 ends, because Drosha and Dicer may cut imprecisely, but usually 1 or 2 specific isoforms are dominant [283]. In most cases, about 60-70% of reads seem to be from sequences that exactly match the canonical miRBase sequence, however, interestingly, there are some cases where an isomiR with 1 or 2 bases missing/extra at each end is actually more dominant [283]. miRBase currently cannot distinguish between different isomiRs, and the single canonical sequence that it lists is what seemed to be the most dominant form at the time that it was added to miRBase. However, the isoform that is dominant can vary

between different tissue and cell types, perhaps due to the presence of different enzymes that post-transcriptionally modify the mature miRNAs [283].

DE type in second condition	A and B	A and C	B and C
Downregulated	**10 sequences:**	**21 sequences:**	**1 sequence:**
	hsa-miR-33b-3p	hsa-miR-625-3p	hsa-miR-4517
	hsa-miR-27b-5p	hsa-miR-3135b	
	hsa-miR-4286	hsa-miR-339-5p	
Upregulated	**0 sequences**	**9 sequences:**	**7 sequences:**
		hsa-miR-1827	hsa-miR-1827
		hsa-miR-6130	hsa-miR-6130
		hsa-miR-98-5p	hsa-miR-1297

Table 6.14: Top 3 miRNAs with largest offset fold change for each comparison and DE type.

Therefore, we decided to collapse all sequences corresponding to a miRNA (adding all isomiRs counts) and perform the DE analysis for non-redundant miRNAs. In this way, all isoforms account for the expression level of a miRNA. For this reason, this method is preferred to redundant miRNA reads in some studies [246]. After re-comparing all conditions, we found a different number of DE miRNAs (which are presented in Table 6.14). In comparing conditions A and B, most miRNAs have still appeared as DE (5 out of 6), but only two miRNAs in comparing conditions A and C and none in comparing B and C. However, other miRNAs have come up as DE, both up- and downregulated, proving that collapsing the isomiR count has a large influence on the miRNA expression levels (full results are presented in Supplementary_Lane1_DE.xlsx).

6.6 Analysis on the CCD-841 libraries

We applied the same methods for processing and analysing the data for the CCD-841 libraries. FASTQ files were converted to FASTA, and sequences containing 'N's were discarded. Next, the adapter sequence was identified and trimmed, and the 'NNNN' sequences corresponding to the HD tag were removed. Then the files were converted from redundant to non-redundant format. The resulting reads were matched full length against the human genome, before and after eliminating reads with low sequence complexity.

During quality check, we found low proportions of reads that presented the

adapter sequences for libraries X2 (70.5%) and Y3 (58.6 %), which suggests there might have been a problem during the adapter ligation step for these samples. After genome mapping, we found an incredibly low proportion of reads from Y1 (8.6%) and Y2 (5.9%) that originated from *H. sapiens* tissue, suggesting that errors were probably introduced while preparing these samples (possibly contamination from other organisms).

Looking at the size class distribution histograms after trimming the adapter sequence and the HD tag, we observed that conditions X and Z had peaks at read lengths 22 and 23, while the peak in condition Y was shifted towards lengths 20 and 21, suggesting the reads captured were more likely degradation products and sRNA fragments rather than functional sRNA. Because of the extreme mapping percentages and the most enriched size classes detected in the samples, which do not correspond to the expected results (the miRNA size classes), samples Y1 and Y2 were eliminated from further analysis.

To check the similarity between replicates, we computed the Jaccard similarity index, MA plots and box plots on fold change. Although overall the Jaccard similarity index was fairly consistent, it presented lower values than in the previous experiment ($\sim$ 0.60 - 0.75), the lowest value being between Z2 and Z3 (0.46). The composition of the datasets was as expected, containing high amounts of miRNAs.

The box plots have the correct alignment of the median lines both for condition X and Z, although the MA plots do not present the expected distribution, being skewed in condition X ad having very large fans in condition Z. After normalising all samples with RPM, quantile and bootstrapping techniques and generating MA plots and box plots for all normalisation methods, we observed that bootstrapping was the most appropriate option for this data, although no method succeeded in correcting all errors.

We then identified the differential expressed sequences by applying the offset fold change method (offset value 20), using confidence intervals, on redundant mature miRNA matches. We found 32 sequences downregulated between X and Y, 2 sequences down- and 37 sequences upregulated between X and Z and 55 sequences upregulated between YZ.

However, we cannot be very confident in the results of this analysis, because

the quality check proved that some of these libraries contain construction errors, the replicates showed very low similarity between them in conditions X and Z, while condition Y lacked in having multiple replicates, and the normalisation methods could not compensate for these errors.

Quality check tables, size class distribution histograms, the Jaccard similarity indexes are presented in Supplementary_Lane2_QC.xls; MA plots and box plots are presented in Appendix C; the results for the DE analysis are presented in Supplementary_Lane2_DE.xls.

6.7 Results

Without sRNA dataset analysis, biologists would have to look at possibly hundreds of miRNAs, which is impossible in practice, due to time and financial constraints. Therefore, bioinformatics dataset analysis has became a standard procedure for selecting the most important sequences to be then checked in the laboratory.

The results of the analysis for both experiments (Caco-2 libraries, CCD-841 libraries) were sent to Christopher Dacosta, who analysed the DE sequences and performed Northern Blots to confirm the activity of the miRNAs in the studied cell lines. The Northern Blot is a technique used in molecular biology research to study gene expression by detection of RNA in a biological sample [8].

For the Caco-2 experiment, he was able to confirm that let-7f-5p, let-7g-5p and miR-10a-5p were upregulated, whereas miR-193b-3p was downregulated by SFN treatment in Caco-2 cell line. He then proved by luciferase assays that let-7f-5p is able to bind to the 3'-UTRs of CDC25A and HMGA2, which are two oncogenes in colorectal cancer.

For the CCD-841 experiment, however, he was not able to validate the differentially expressed miRNAs in the laboratory, suggesting they might appear as DE due to technical errors and therefore, the results of the DE analysis might not reflect the real expression levels.

6.8 Discussion

Although this method of performing sRNA analysis proved its efficiency, it has its limitations too. The way of calculating the DE involves no statistical evidence that the difference in the expression levels is significant. Therefore, the procedure relies on the pre-processing and normalisation methods to correct the potential errors, and it becomes imperative that the proper methods for these steps are chosen, depending on the data. However, the pre-processing and normalisation methods cannot always ensure that the expression levels have been fully corrected to their biological values. This can provide a logical explanation for the fact that the DE sequences found in some bioinformatics analysis cannot be validated in the laboratory (e.g. the case of the CCD-841 experiment above).

A similar sRNA data analysis, using a slightly different approach, was conducted on a study on miRNA activity in non-tumorigenic MCF10A cell line, after inducing DNA damage at 4h, 24h and 48h [298]. This work was performed under the supervision of Dr. Simon Moxon, in collaboration with Dr. Adam Hall, who provided the small RNA sequencing data.

After the datasets were processed, the similarity between replicates was checked using correlation indexes and correlation plots. Sequences were normalised using the RPM method and then differentially expressed miRNAs were called using the edgeR package [299]. The edgeR package uses a Poisson model to estimate the dispersion between replicates and the model can separate biological variation from technical variation, giving a p-value as the statistical confidence that a sequence is DE [299]. For this study, a p-value of 0.05 was used as a cut-off to analyse microRNAs that were upregulated following DNA damage.

The results of this study were successfully published in the following paper [298], where I figure as co-author.

In this chapter we have presented a method for sRNA dataset processing, quality assurance and differential expression analysis. We have discussed about potential errors that might occur during library preparation, sequencing and through normalising the data. We have therefore proved that quality check, replicates validation and choosing the correct normalisation method are essential for conducting a correct differential expression analysis. We have then described

a method of calculating the differential expression, and presented the results for miRNAs in redundant and non-redundant format. These results were analysed and validated biologically through Northern Blots.

Chapter 7

Conclusions and future work

7.1 Summary

In this thesis we have presented a review of the most commonly used miRNA prediction tools, and then we developed a new miRNA prediction algorithm, miRCat2, which identifies miRNAs from HTS datasets in both plants and animals. We have tested miRCat2 on ten model organisms and benchmarked it against four similar tools (miRCat [1], miRDeep2 [2], miRPlant [3] and miReap (http://mireap.source-forge.net/)), showing that we achieve an improved performance. We have then presented a practical use for the annotated miRNAs, describing and applying a method of sRNA datasets quality checking and miRNA differential expression analysis.

We now present some possible improvements and future extensions to miRCat2, before summing up the key points of this research.

7.2 Future work

Future extensions might be implemented to improve the performance of miRCat2 or to add functionality and innovative features. We present these possible improvements below:

- **Decreasing the run time** - miRCat2 is currently implemented as a single-threaded process, but the miRNA candidates are processed individually, and

do not depend on one another. Therefore, it is possible to integrate multi-threading in the application, which could substantially decrease its run time.

- **Performing analysis on multiple datasets** - currently, miRCat2 can be run only on one dataset at a time. We could make use of the fact that the UEA small RNA Workbench provides an easy way of constructing a database using multiple replicates from the same organism, and run miRCat2 on multiple datasets. The results obtained would be more informative and the predictions would have better confidence that they are true miRNAs if they were predicted from multiple samples.
- **Auto-updating values for the parameter set** - the current parameter values were chosen based on the sequencing data and on features observed from miRBase entries (for the respective Kingdom). As these technologies are continuously evolving and changing, the values for the recommended default parameters might become obsolete, although the user can manually change them. A new feature could be implemented, that would analyse annotated miRNA sequences and extract their features, to auto-update and redefine the set of default parameters.
- **Integrating miRCat2 in predefined pipelines** - the UEA small RNA Workbench contains many useful analysis tools, with which the user can define custom pipelines (order of tools for data processing). It would be useful to have a set of predefined pipelines in which miRCat2 would be integrated. For example, after running miRCat2 on plant data, its results could be given as input to PAREsnip [183] (if degradome data is also available), to have their targets predicted. Alternatively, the results of miRCat2 can be included as reference miRNAs when running the differential analysis tool [249].

7.3 Conclusions

We have presented miRCat2, and showed that the predictions made by miRCat2 are more accurate than those made by similar software. Furthermore, miRCat2 performs consistently throughout all tested organisms, while the other tools tend to perform efficiently in only some of the datasets.

miRNAs have a very complex mechanism of functioning, to which new discoveries are still being added. The comparison of miRNA profiles (over different conditions and treatments) reveals differentially expressed loci, giving a better understanding of their function in biological processes. Therefore, the accurate classification of miRNAs can have a crucial impact in multiple areas of research of high importance, such as disease and cancer research or crop improvement. miRCat2 could successfully be used to classify miRNAs, that will be used for important future research projects. For example, in animals, there is a strong focus on using miRNAs expression for developing new treatments, but also as biomarkers in disease and cancer diagnosis, which is essential in many cases for treatment efficiency. In plants, researchers can use the discovered miRNAs to develop new strains of crops with pathogen-resistance or to improve plant resistance in unfavourable growing conditions, such as draught and soil nutrient deficiency, this way optimising the quantity and quality of food produced. These are important current and future issues that the sRNA and miRNA research can address.

Although most tools perform miRNA classification quite accurately, miRNA target prediction remains an open problem. Current tools predict hundreds of potential targets for each miRNA sequence, lacking in precision, which makes the validation of the results in the laboratory near impossible. Therefore, improvements to such software or new software is required, and miRNA target prediction is an important area on which miRNA research will focus on in the near future.

HTS technologies are continuing to evolve, requiring improvements for the tools used to analyse such data, to keep up with their progress. miRCat2 has an efficient method of dealing with increasing HTS sequencing dataset size, at the same time facilitating the expansion of sRNA knowledge and the exciting discovery of novel miRNAs, that are missed by other methods. In this way, miRCat2 contributes to the characterisation and understanding of miRNAs, both in plants and animals, expanding the broad field of miRNA and sRNA research.

Appendices

Appendix A

Here we present the parameters used by the miRCat2 algorithm. We show both the user-configurable and the predefined parameters, together with their default values in animal and in plant data and the justification for the proposed default value.

User-configurable parameters				
Parameter name	**Description**	**Default value animals**	**Default value plants**	**Justification for value**
min_len	Minimum length of a miRNA	20	20	Includes sequences that fall out of the regular miRNA size class
max_len	Maximum length of a miRNA	24	23	Includes sequences that fall out of the regular miRNA size class
min_fold_len	Minimum length of a hairpin	40	45	Lower value than minimum fold length for most organisms
max_fold_len	Maximum length of a hairpin	100	250	Higher value than maximum fold length for most organisms
max_amfe	Maximum value for the AMFE for a miRNA precursor	-22	-22	Empirically determined
complex	Complexity of sequence	0.90	0.90	Empirically determined
gaps_miRNA	Maximum number of consecutive gaps on the hairpin on the miRNA location	4	4	Empirically determined
repeats	Maximum number of times a sRNA can map to the genome (usually miRNAs map to a limited number of locations)	25	25	a miRNA sequence does not map repeated times to the reference genome
pVal	Threshold for the RANDfold output value	0.05	0.05	Statistically significant value
complex_loop	If a hairpin with multiple loops between the miRNA and miRNA* is allowed	false	true	Complex secondary structures have been previously seen in plants, but never in animals. If a complex loop is permitted, it should not contain more than 3 loops.
no_loop	Maximum number of bulks in the loop area of the precursor	0	3	Empirically determined
clear_cut_-percent	Percent of incident reads that should fall between the same start and end positions as the miRNA	0.95	0.92	Empirically determined, plant data is more variable
RANDfold	If RANDfold should be computed	false	false	Results are accurate without it and it slows the algorithm. Recomended if wanting to further restrict the results

Table 1: **Parameters involved in the algorithm of miRCat2, that are user-configurable.** The parameters are presented with their default values and the justification for using the respective value, for both animal and plant data.

Predefined parameters (user cannot change)				
Parameter name	**Description**	**Default value animals**	**Default value plants**	**Justification for value**
min_orientation	Percent of the reads that have the same strand on the hairpin	0.8	0.8	Empirically determined
min_pared_perc	Minimum percentage of nts that should be paired on a hairpin	0.5	0.5	Empirically determined
min_pared_nucl	Minimum number of nts that should be paired on a hairpin	15	15	Empirically determined
overlap_percent	Maximum percent of sRNAs in an adjacent cluster overlapping with the miRNA cluster in order to be considered a clear cut	0.05	0.05	Empirically determined
fuzzy	Percent of all reads aligned to the hairpin that should map in accordance to Dicer/DCL1 and Drosha products	0.9	0.9	Empirically determined
min_fold_len	Minimum length of a hairpin	40	45	Minimum length of the miRNA, miRNA* and loop added together
window	Number of nts a split of the genome has.	300	500	Large enough to contain a miRNA hairpin but small enough to represent a significant local context of reads
subwindow	Number of nts a split of the window has.	20	20	It can cover at least a half of the longer miRNAs (25nts), but not more than then maximum length
window_overlap	Number of nts two adjacent windows overlap	100	100	ensures that adjacent windows are not isolated, but they influence each other
depth	Number of iterations to perform the KLD on genome location if removing a sRNA does not bring it closer to a RUD	4	4	Empirically determined
rud_val	Threshold for the KLD below which we consider the distribution to be a RUD	1.23	1.23	Empirically determined
min_loop	Minimum number of nts that the loop should have	3	3	Empirically determined
clear_cut	Number of nts a sRNA can be shifted regarded to a miRNA in order to be considered to have the same start/end (isomir)	3	3	Empirically determined
under_clear_cut	Percent of sRNA in a cluster with the same cut in order to be considered a clear cut. This is considered if the clear_cut_percent fails on one of the sides of the miRNA	0.7	0.7	Empirically determined
min_size	Minimum length of a sRNA in the file	16	16	Sequences smaller than 16 nt are adapter-adapter sequences and should not exist in the dataset
max_size	Maximum length of a sRNA in the file	40	40	There are usually very few sequences with length over 40 in a sRNA dataset
offset	Value added to the reads distribution in order to avoid division by 0	1	1	Minimum read abundance
offset_low	Value added to the reads distribution in order to avoid division by 0 when read abundance is low	offset*0.9	offset*0.9	Offset becomes more significant than actual read abundances
plateau_range	Number of subwindow to be included in the local peak detection on both sides of the miRNA candidate	4	6	Empirically determined
3‘overhang	Number of nts the miRNA* is shifted compared to the miRNA	2	2	miRNA biogenesis

Table 2: **Predefined parameters involved in the algorithm of miRCat2, that cannot be changed by the user.** The parameters are presented with their default values and the justification for using the respective value, for both animal and plant data.

Appendix B

Here we present the MA plots and box plots produced for the analysis of sRNA sequencing libraries in a study on the roles of microRNAs in the anti-cancer effects of sulforaphane. We produced plots in order to check the quality and consistency of the replicates, and also for each of the normalisation methods, to assert which one is the most suitable for the analysed data. We present the results for Lane 1, conditions A, B and C.

Figure 1: MA plots comparing the offset fold change between replicates from condition A, before normalisation, grouped on size classes.

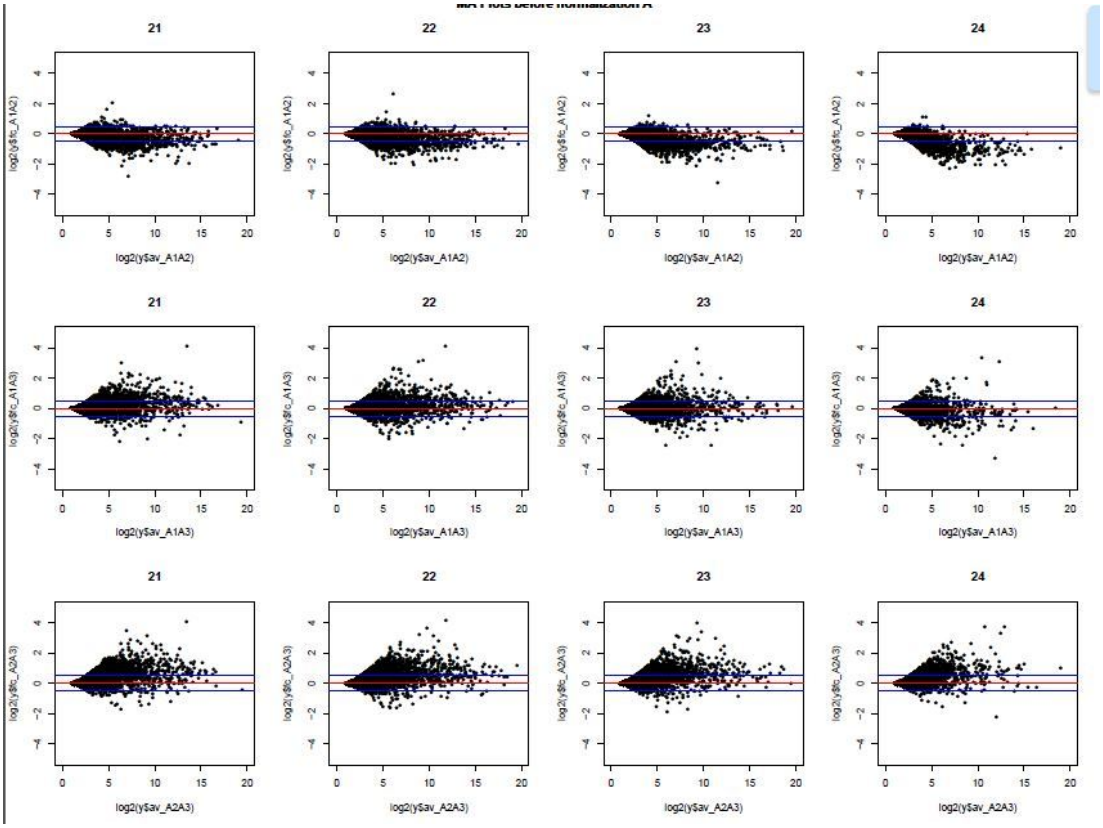


Figure 2: MA plots comparing the offset fold change between replicates from condition B, before normalisation, grouped on size classes.

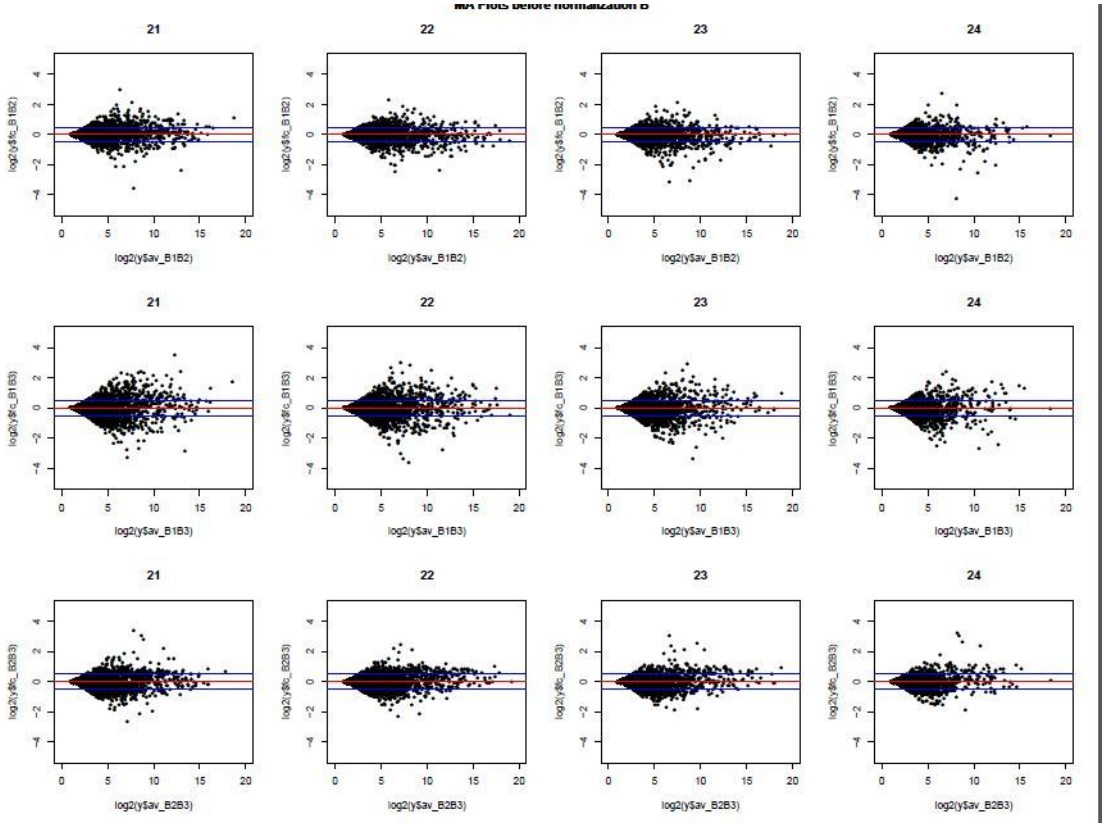


Figure 3: MA plots comparing the offset fold change between replicates from condition C, before normalisation, grouped on size classes.

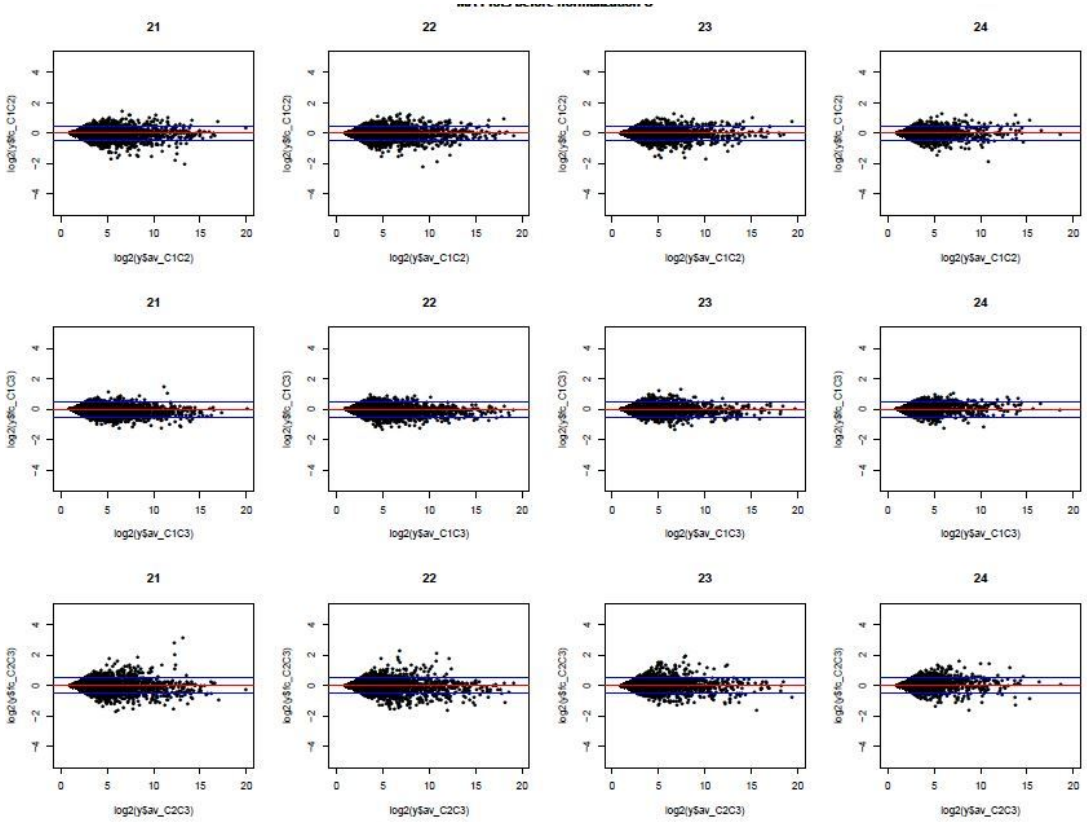


Figure 4: Box plots on the offset fold change between replicates, after normalisation using RPM, grouped on size classes.

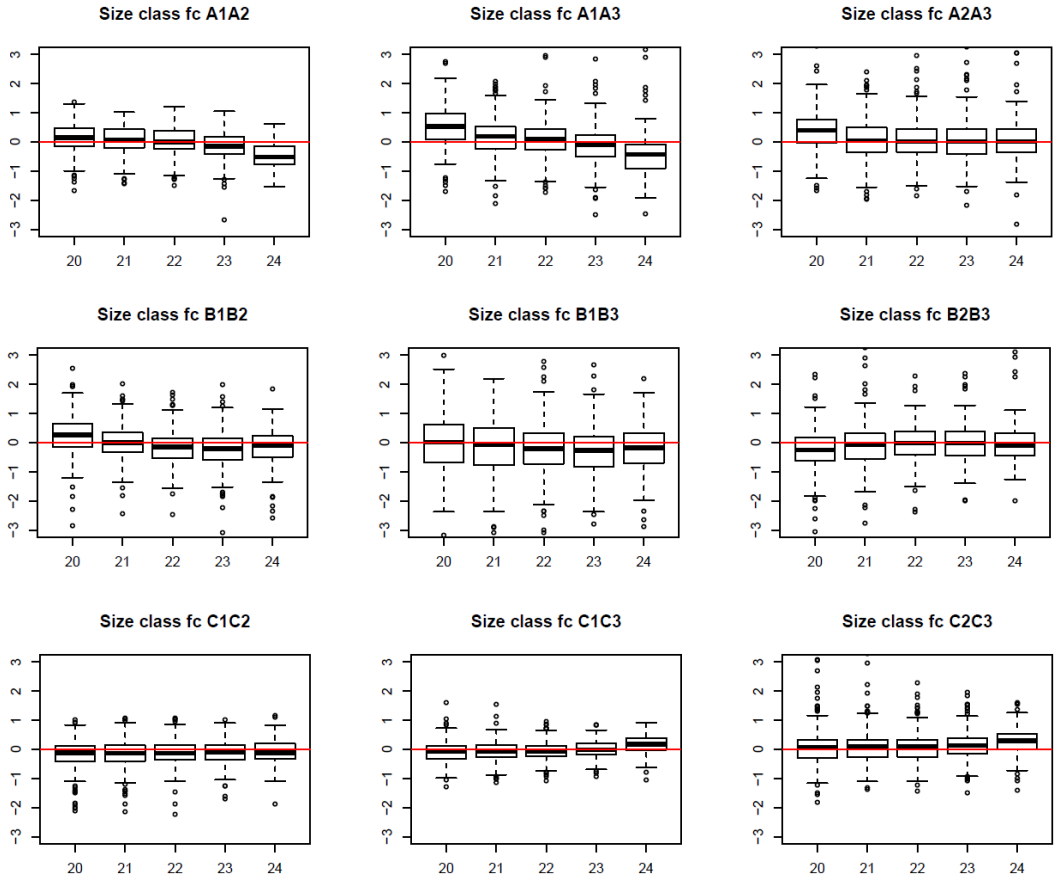


Figure 5: MA plots comparing the offset fold change between replicates from condition A, after normalisation using RPM, grouped on size classes.

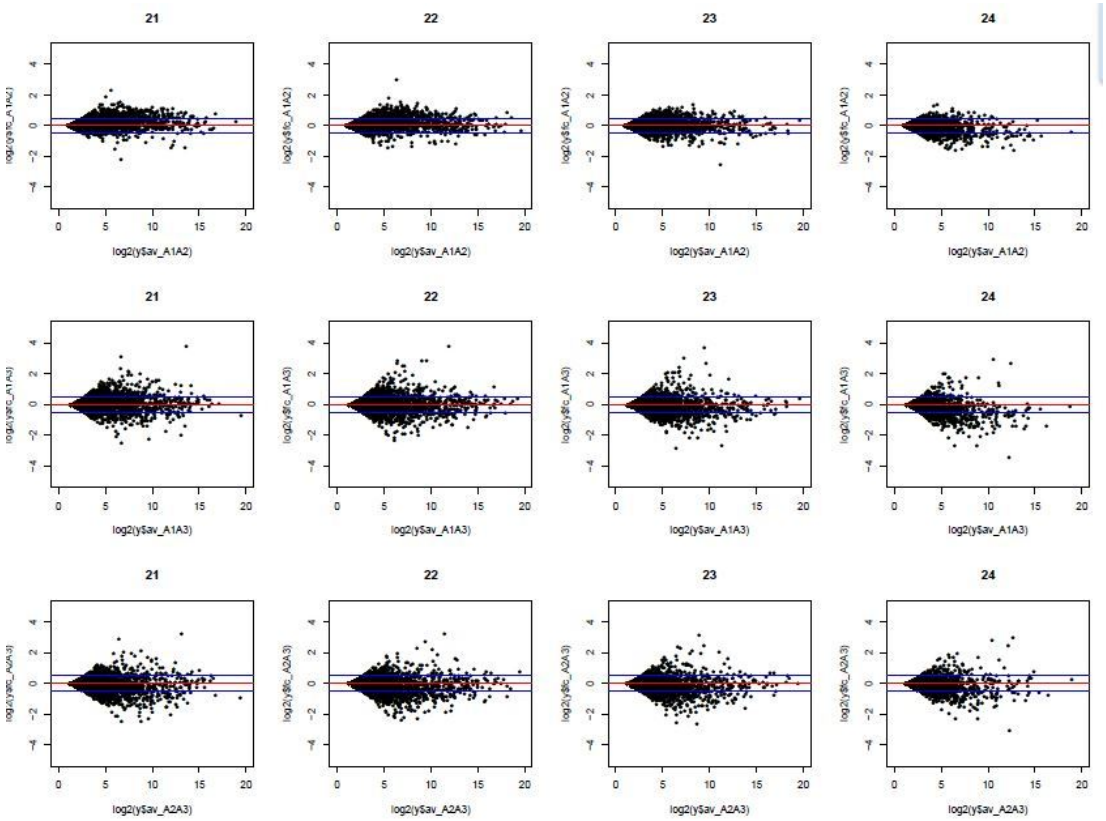


Figure 6: MA plots comparing the offset fold change between replicates from condition B, after normalisation using RPM, grouped on size classes.

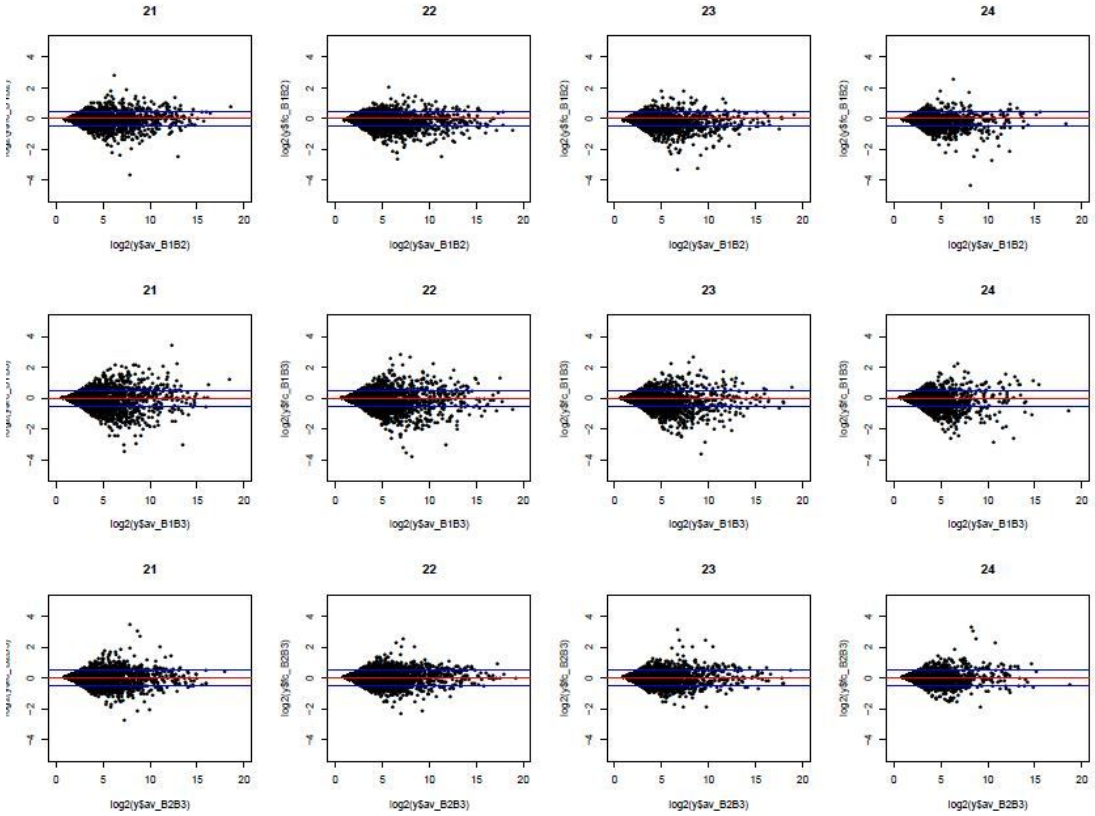


Figure 7: MA plots comparing the offset fold change between replicates from condition C, after normalisation using RPM, grouped on size classes.

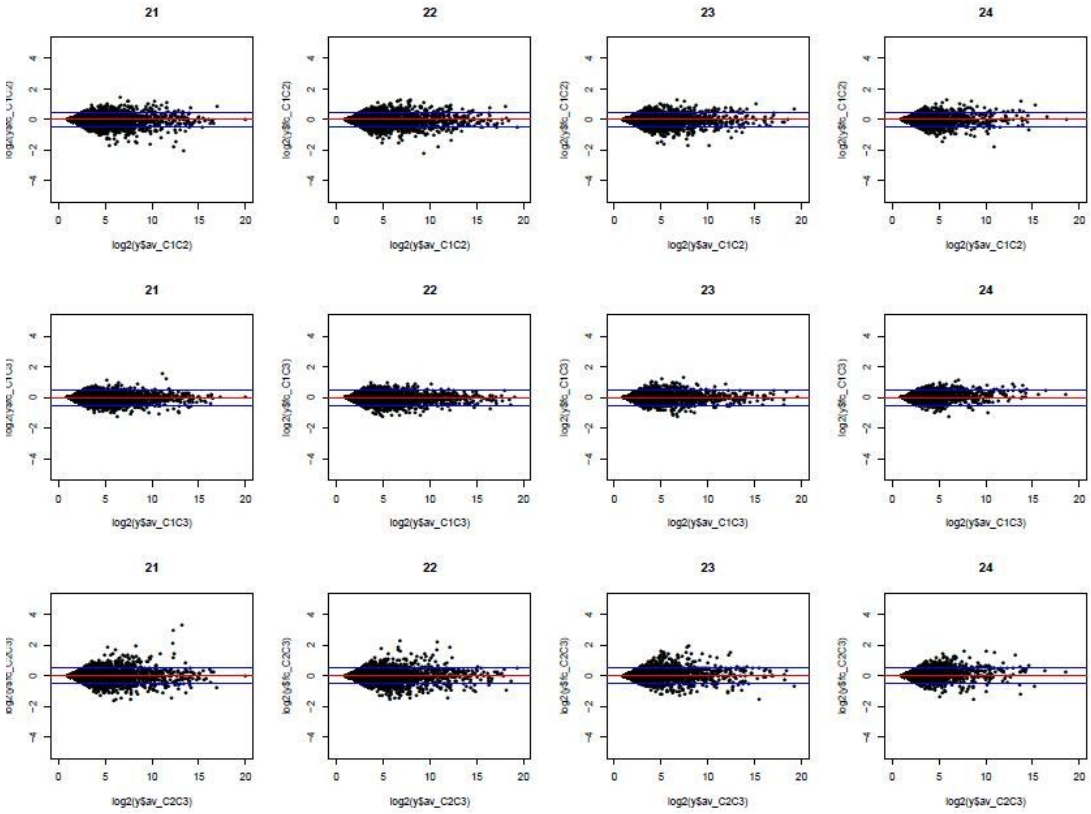


Figure 8: Box plots on the offset fold change between replicates, after normalisation using the quantile method, grouped on size classes.

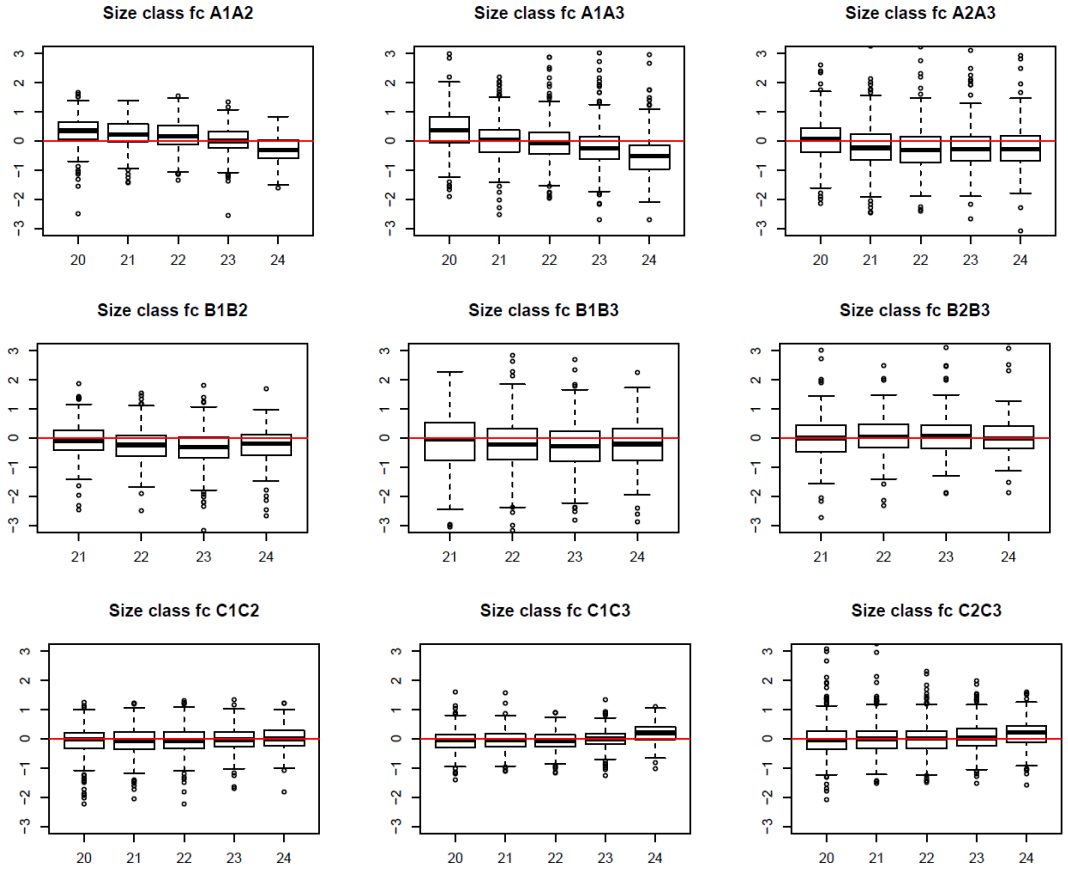


Figure 9: MA plots comparing the offset fold change between replicates from condition A, after normalisation using the quantile method, grouped on size classes.

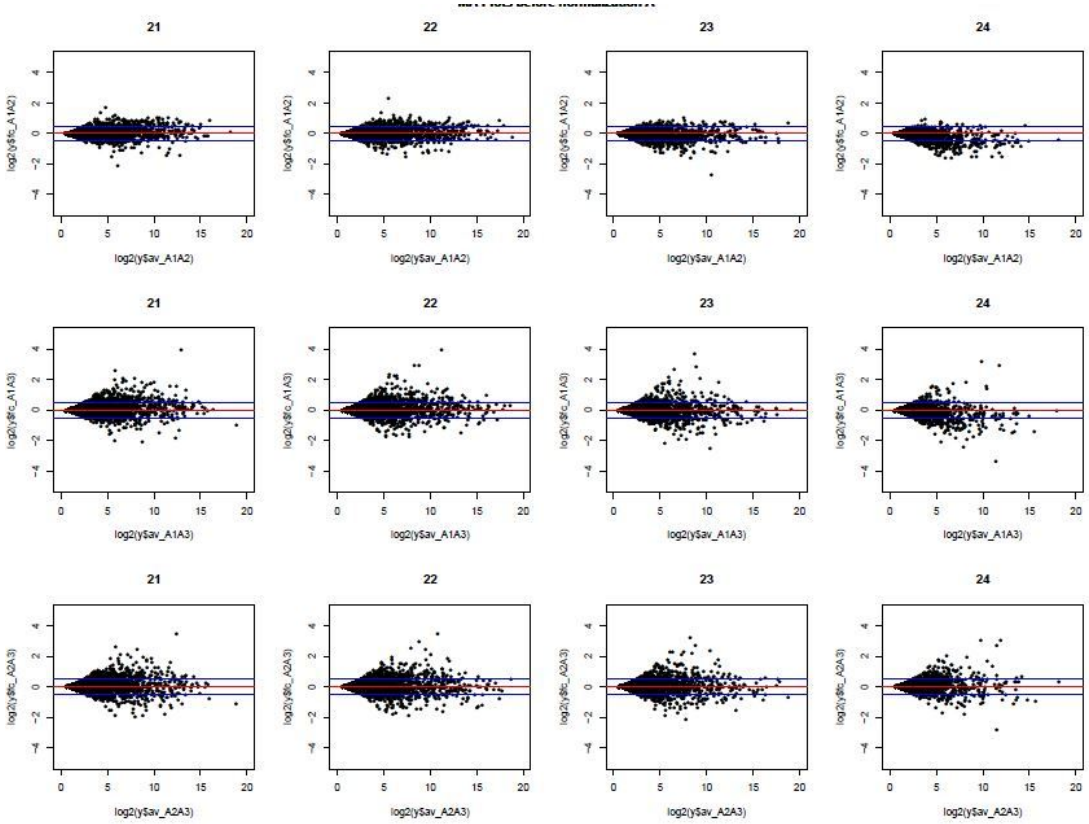


Figure 10: MA plots comparing the offset fold change between replicates from condition B, after normalisation using the quantile method, grouped on size classes.

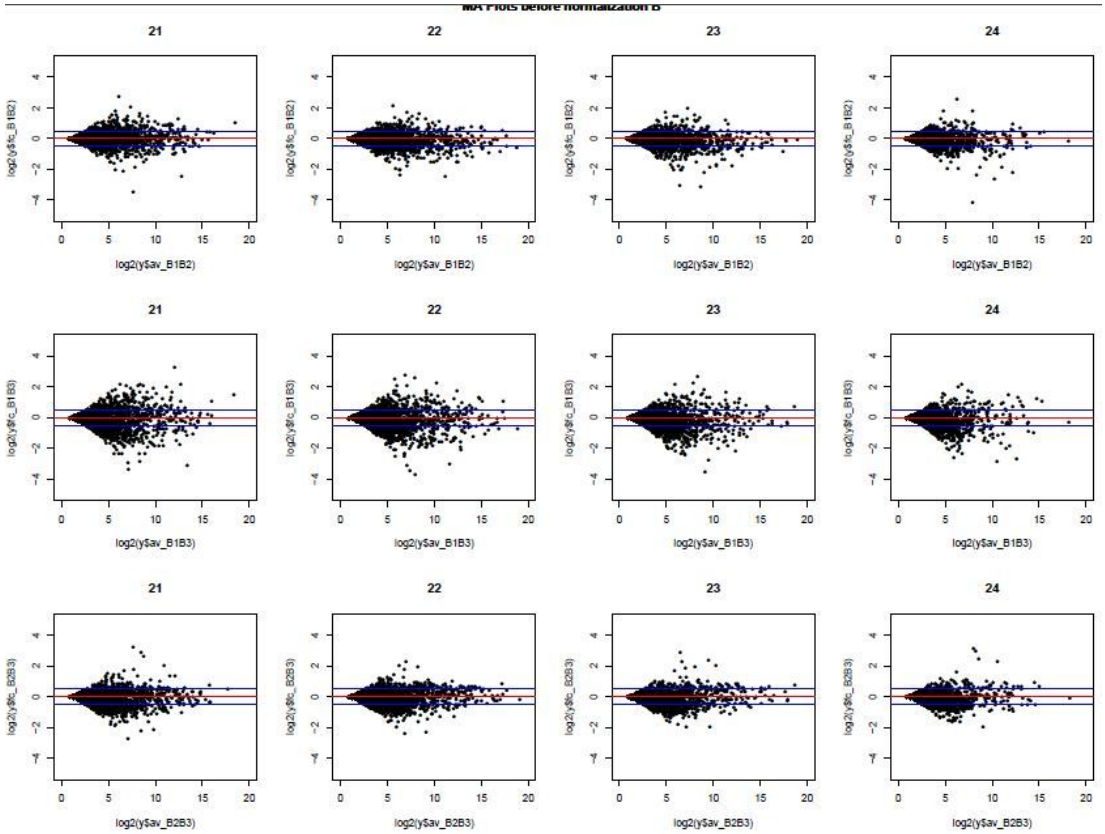


Figure 11: MA plots comparing the offset fold change between replicates from condition C, after normalisation using the quantile method, grouped on size classes.

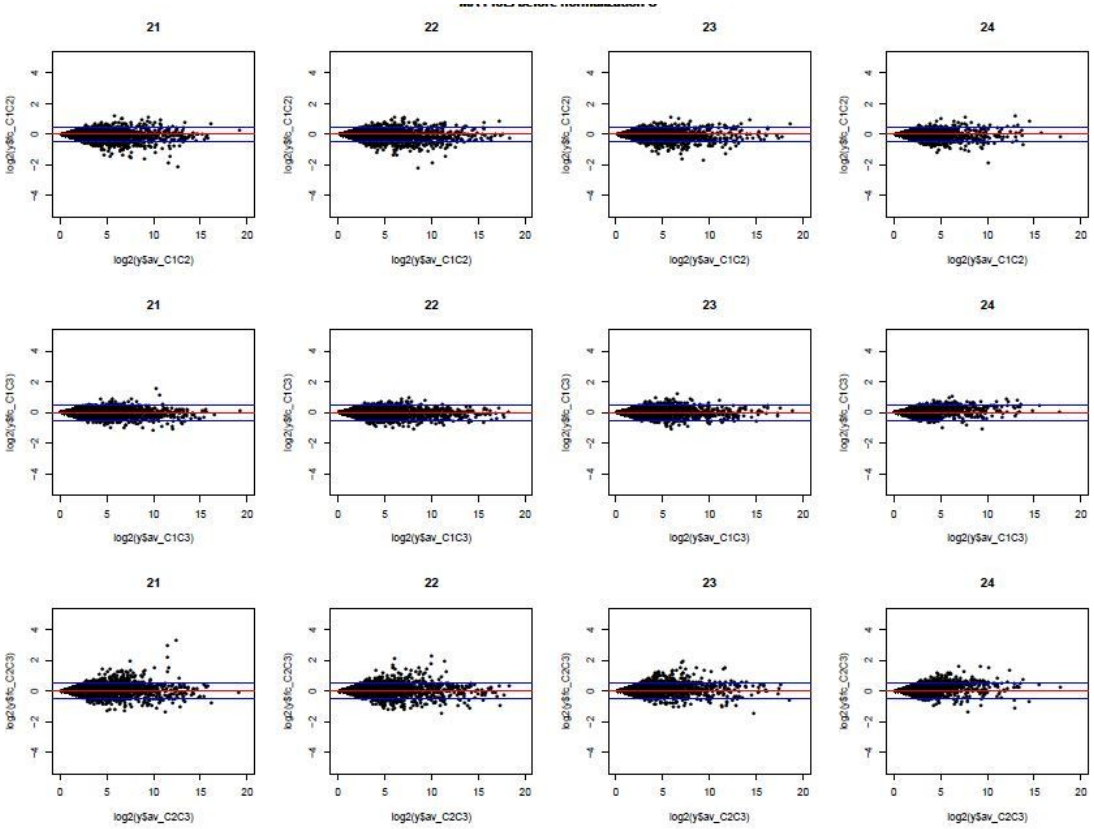


Figure 12: Box plots on the offset fold change between replicates, after normalisation using the bootstrapping method at 50%, grouped on size classes.

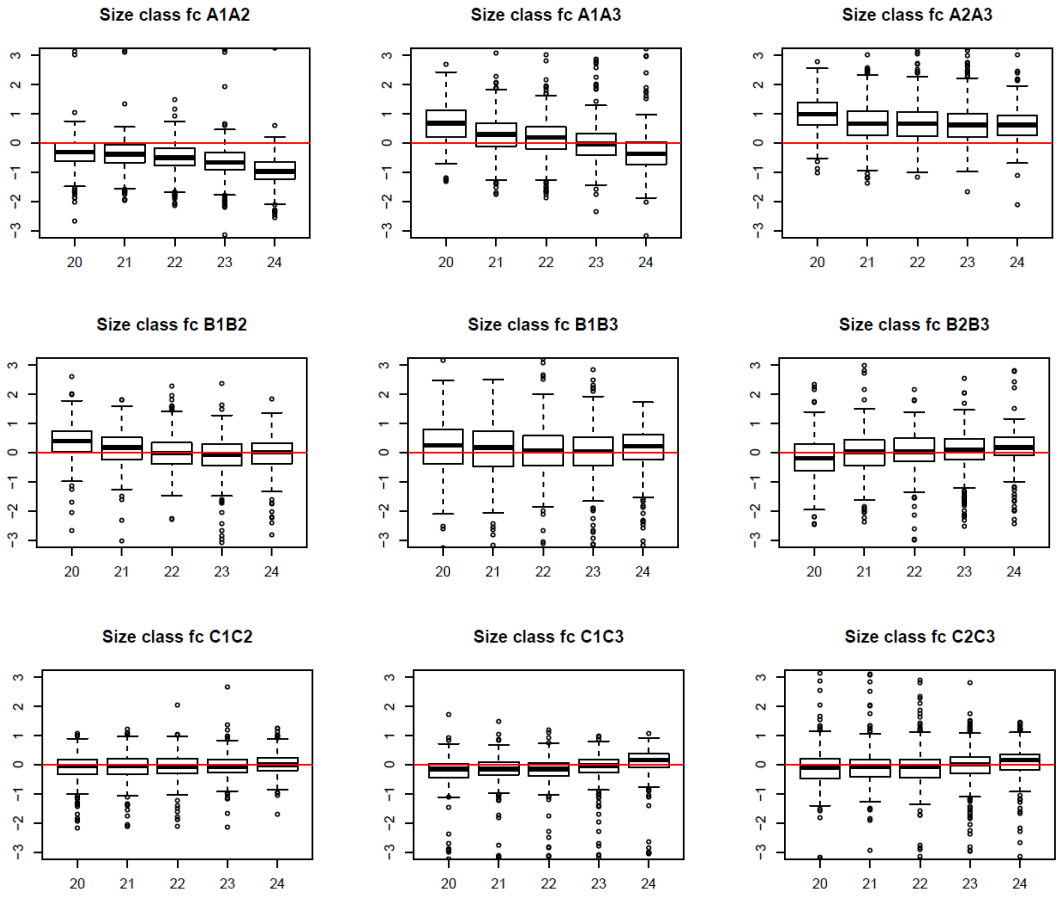


Figure 13: Box plots on the offset fold change between replicates, after normalisation using the bootstrapping method at 70%, grouped on size classes.

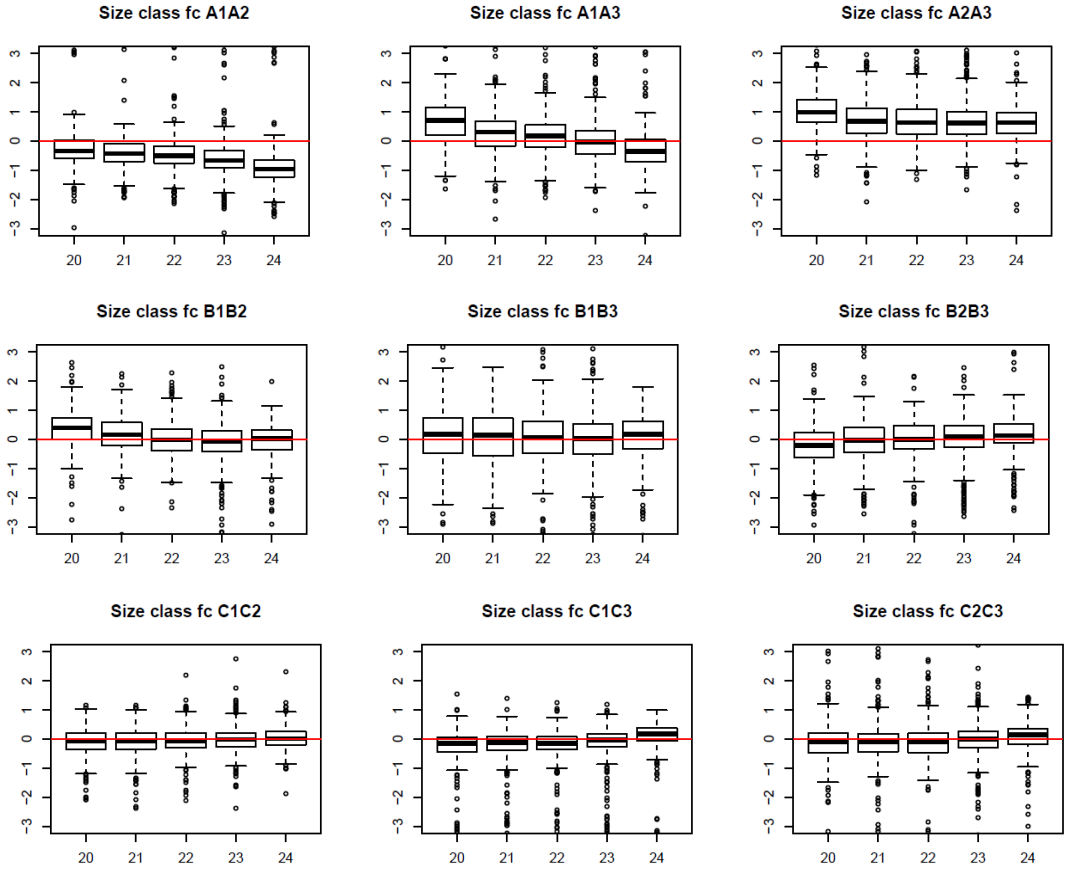


Figure 14: Box plots on the offset fold change between replicates, after normalisation using the bootstrapping method at minimum total, grouped on size classes.

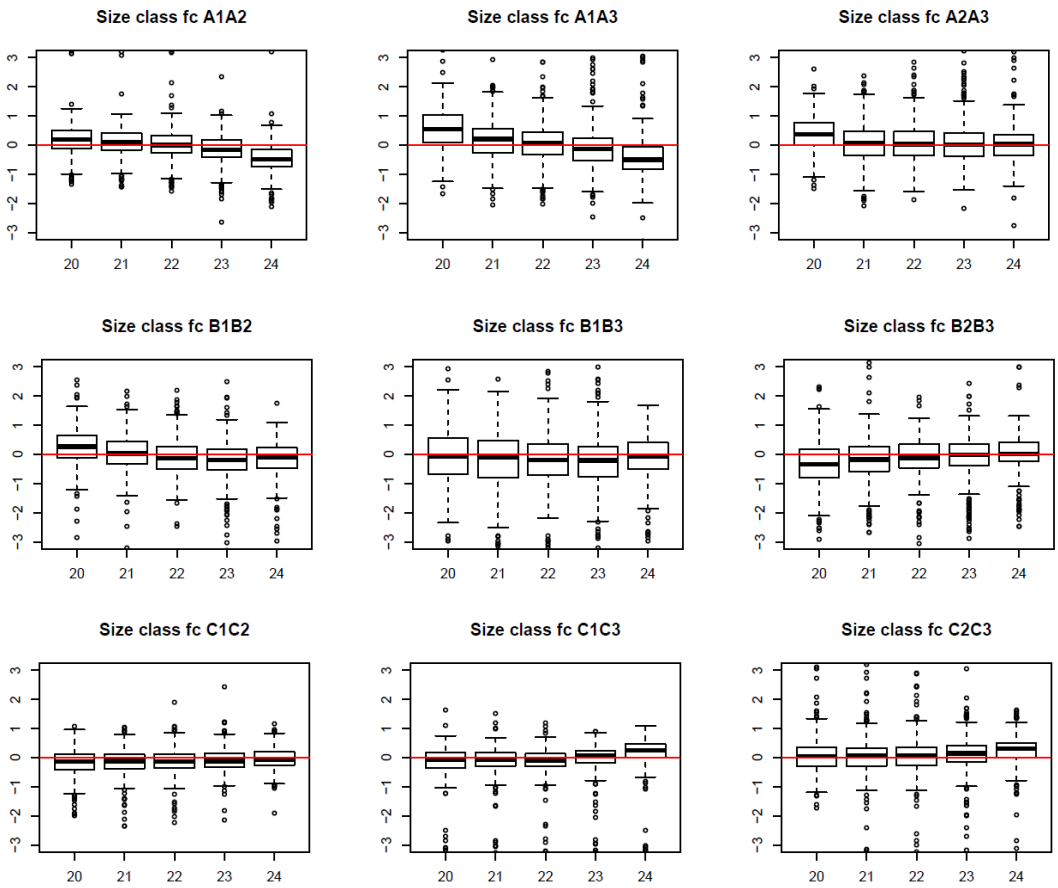


Figure 15: MA plots comparing the offset fold change between replicates from condition A, after normalisation using the bootstrapping method at minimum total, grouped on size classes.

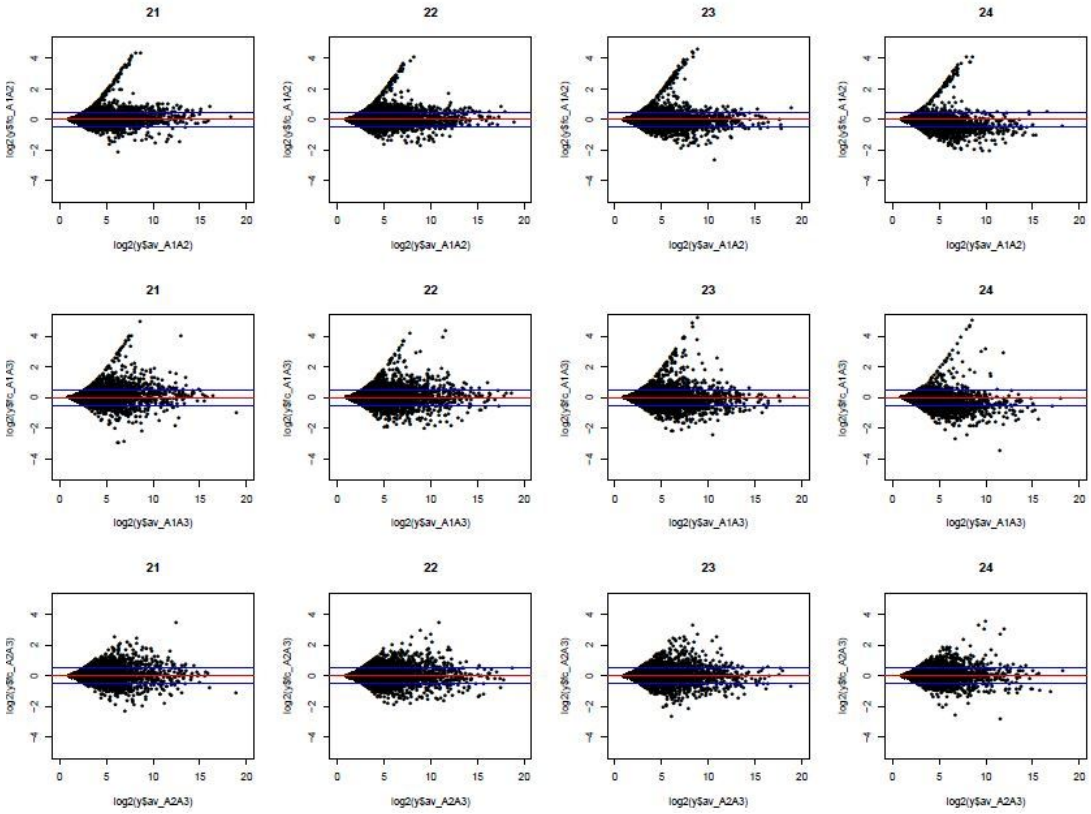


Figure 16: MA plots comparing the offset fold change between replicates from condition B, after normalisation using the bootstrapping method at minimum total, grouped on size classes.

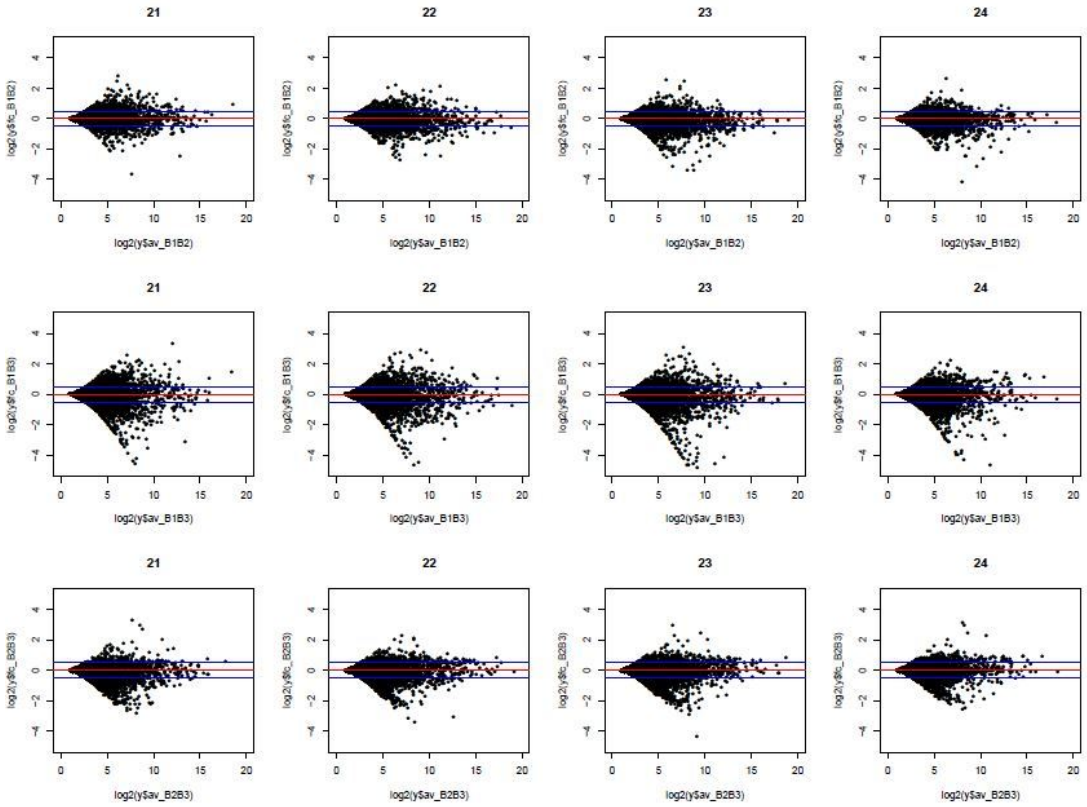
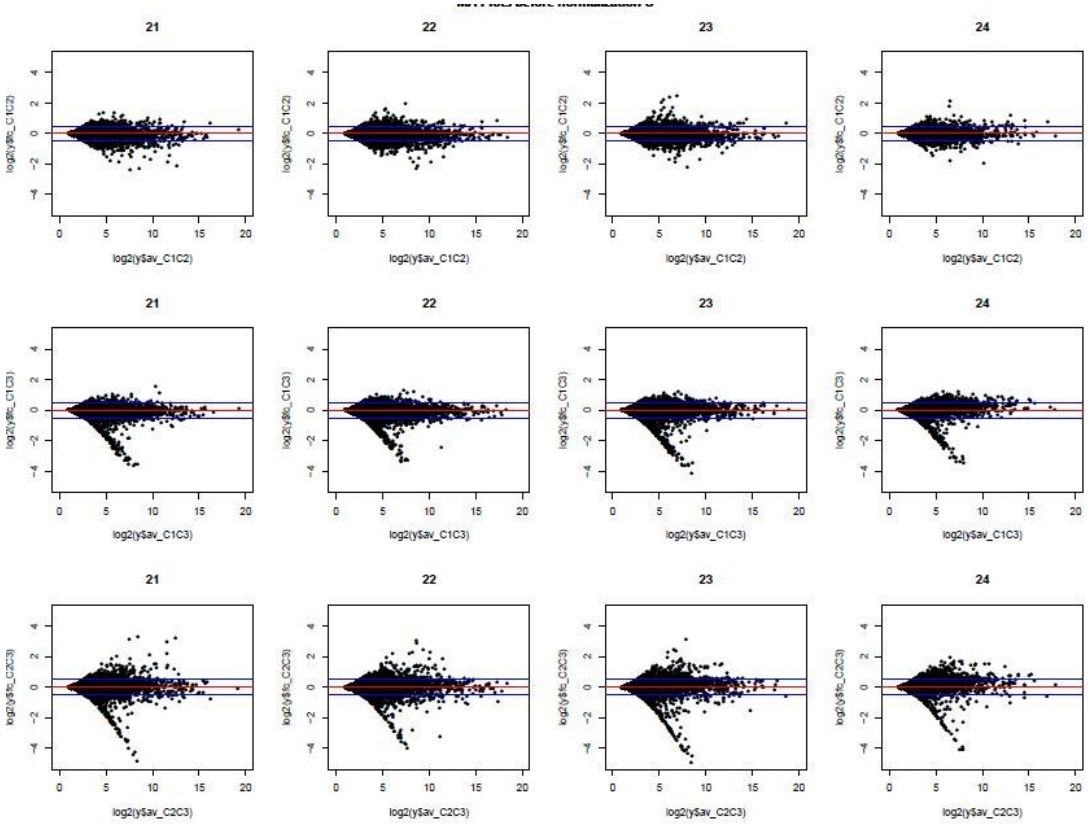


Figure 17: MA plots comparing the offset fold change between replicates from condition C, after normalisation using the bootstrapping method at minimum total, grouped on size classes.



Appendix C

Here we present the MA plots and box plots produced for the analysis of sRNA sequencing libraries in a study on the roles of microRNAs in the anti-cancer effects of sulforaphane. We produced plots in order to check the quality and consistency of the replicates, and also for each of the normalisation methods, to assert which one is the most suitable for the analysed data. We present the results for Lane 2, conditions X, Y (where Y has more than one valid replicate) and Z.

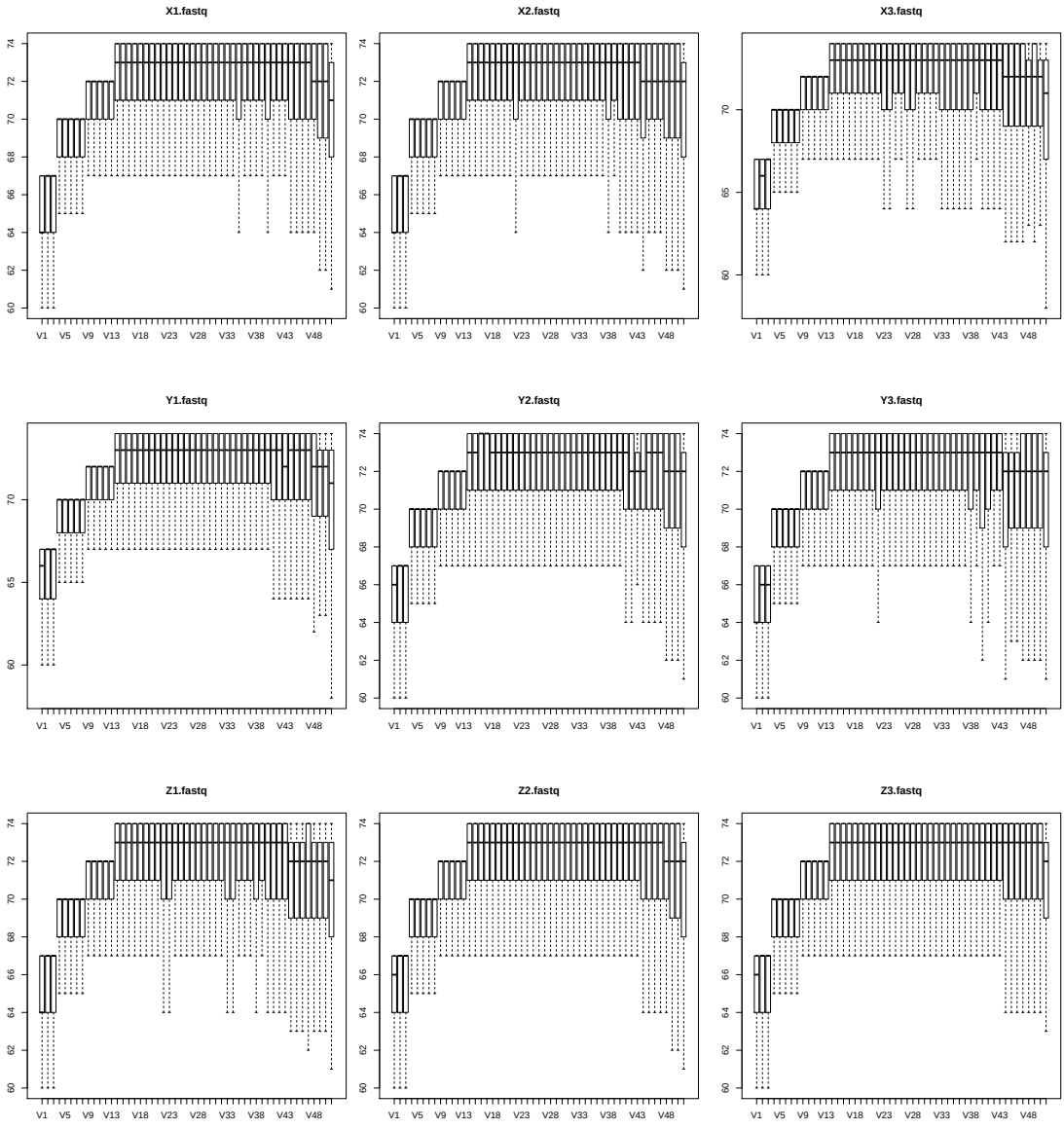
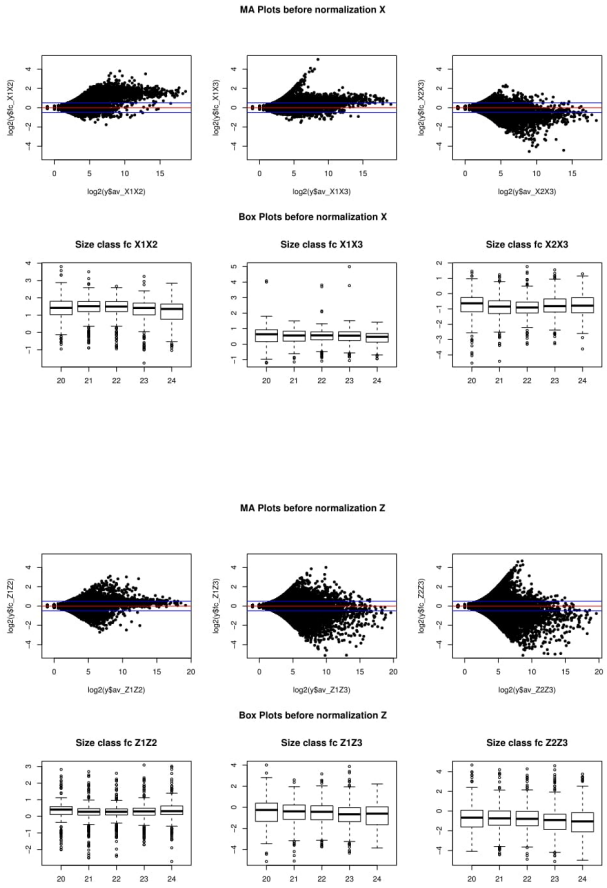
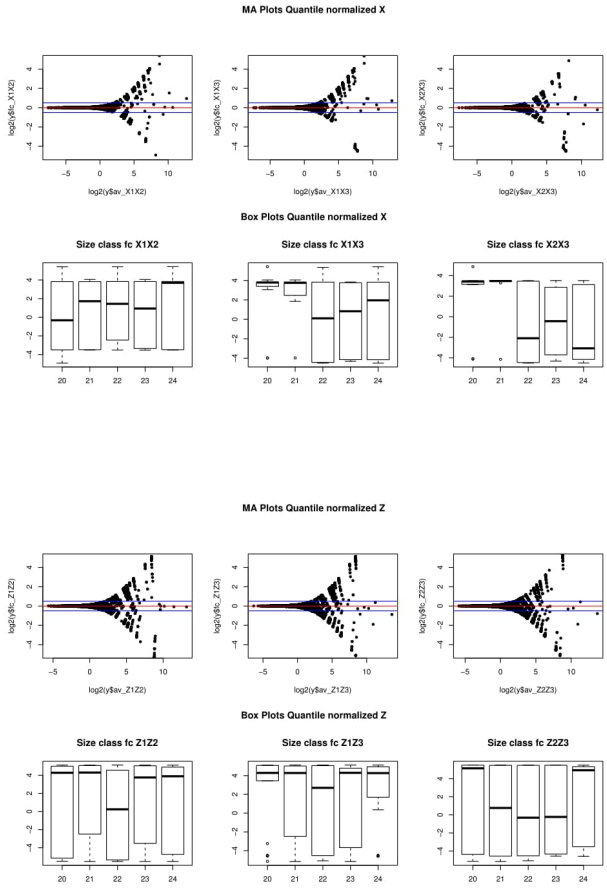
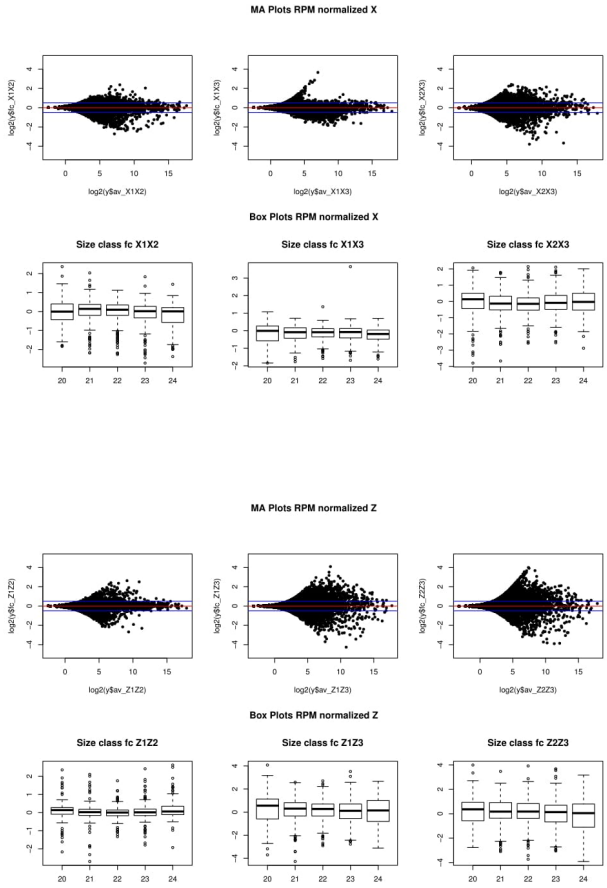


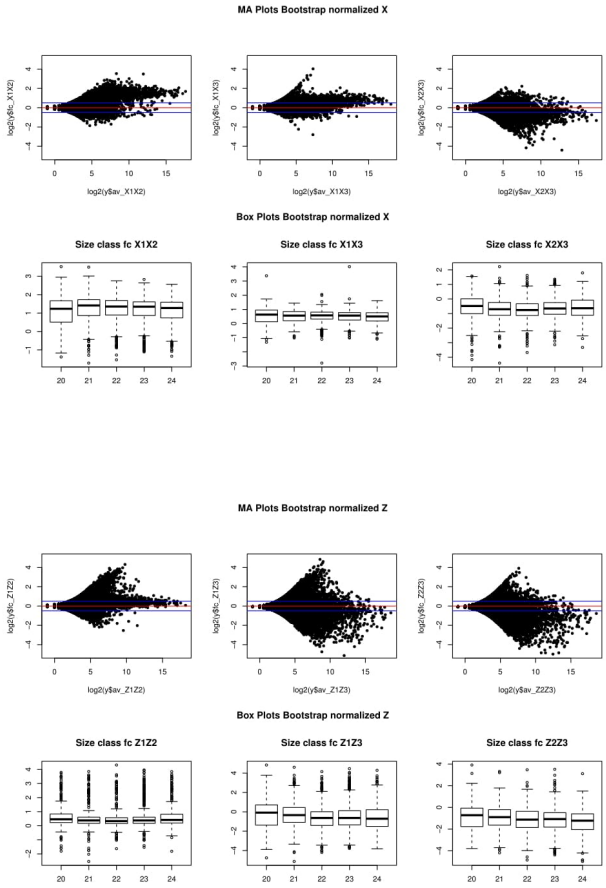
Figure 18: Boxplots for the FASTQ score per nucleotide, for each library. Replicates are based on the same line and can easily be compared. The boxplots show variable quality score per nucleotide in some files.

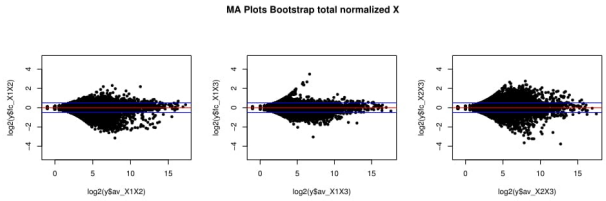

MA Plots before normalization X
log2(y$fc_X1X2)
log2(y$av_X1X2)
log2(y$fc_X1X3)
log2(y$av_X1X3)
log2(y$fc_X2X3)
log2(y$av_X2X3)
Box Plots before normalization X
Size class fc X1X2
Size class fc X1X3
Size class fc X2X3
MA Plots before normalization Z
log2(y$fc_Z1Z2)
log2(y$av_Z1Z2)
log2(y$fc_Z1Z3)
log2(y$av_Z1Z3)
log2(y$fc_Z2Z3)
log2(y$av_Z2Z3)
Box Plots before normalization Z
Size class fc Z1Z2
Size class fc Z1Z3
Size class fc Z2Z3

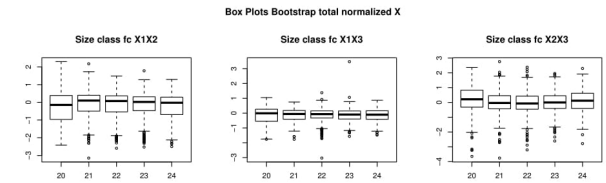
1

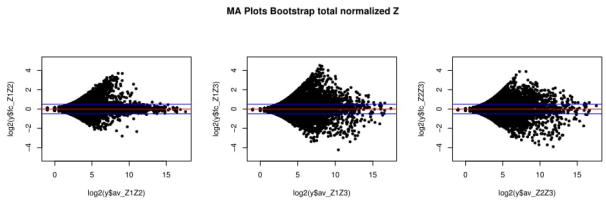

MA Plots Quantile normalized X
log2(y$fc_X1X2)
log2(y$av_X1X2)
log2(y$fc_X1X3)
log2(y$av_X1X3)
log2(y$fc_X2X3)
log2(y$av_X2X3)
Box Plots Quantile normalized X
Size class fc X1X2
Size class fc X1X3
Size class fc X2X3
MA Plots Quantile normalized Z
log2(y$fc_Z1Z2)
log2(y$av_Z1Z2)
log2(y$fc_Z1Z3)
log2(y$av_Z1Z3)
log2(y$fc_Z2Z3)
log2(y$av_Z2Z3)
Box Plots Quantile normalized Z
Size class fc Z1Z2
Size class fc Z1Z3
Size class fc Z2Z3

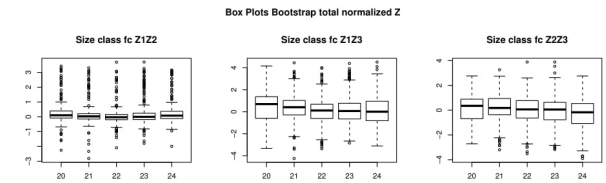

MA Plots RPM normalized X
log2(y$fc_X1X2)
log2(y$av_X1X2)
log2(y$fc_X1X3)
log2(y$av_X1X3)
log2(y$fc_X2X3)
log2(y$av_X2X3)
Box Plots RPM normalized X
Size class fc X1X2
Size class fc X1X3
Size class fc X2X3
MA Plots RPM normalized Z
log2(y$fc_Z1Z2)
log2(y$av_Z1Z2)
log2(y$fc_Z1Z3)
log2(y$av_Z1Z3)
log2(y$fc_Z2Z3)
log2(y$av_Z2Z3)
Box Plots RPM normalized Z
Size class fc Z1Z2
Size class fc Z1Z3
Size class fc Z2Z3

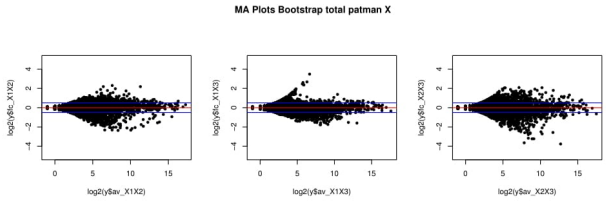

MA Plots Bootstrap normalized X
log2(y$fc_X1X2)
log2(y$av_X1X2)
log2(y$fc_X1X3)
log2(y$av_X1X3)
log2(y$fc_X2X3)
log2(y$av_X2X3)
Box Plots Bootstrap normalized X
Size class fc X1X2
Size class fc X1X3
Size class fc X2X3
MA Plots Bootstrap normalized Z
log2(y$fc_Z1Z2)
log2(y$av_Z1Z2)
log2(y$fc_Z1Z3)
log2(y$av_Z1Z3)
log2(y$fc_Z2Z3)
log2(y$av_Z2Z3)
Box Plots Bootstrap normalized Z
Size class fc Z1Z2
Size class fc Z1Z3
Size class fc Z2Z3

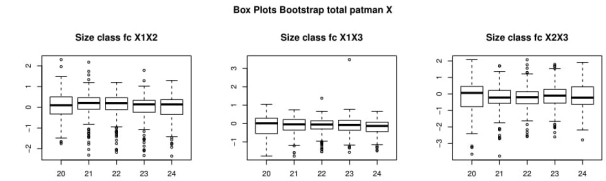

MA Plots Bootstrap total normalized X
log2(y$fc_X1X2)
log2(y$av_X1X2)
log2(y$fc_X1X3)
log2(y$av_X1X3)
log2(y$fc_X2X3)
log2(y$av_X2X3)

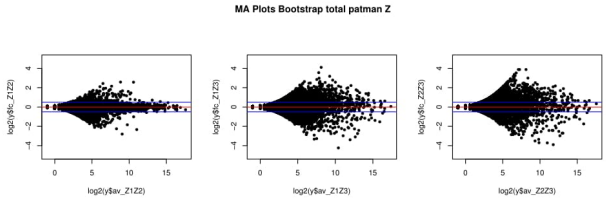

Box Plots Bootstrap total normalized X
Size class fc X1X2
Size class fc X1X3
Size class fc X2X3

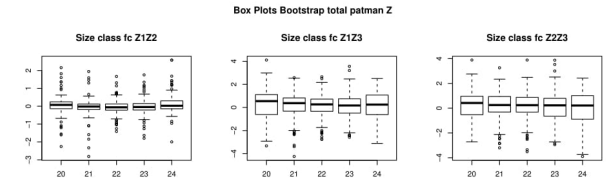

MA Plots Bootstrap total normalized Z
log2(y$fc_Z1Z2)
log2(y$av_Z1Z2)
log2(y$fc_Z1Z3)
log2(y$av_Z1Z3)
log2(y$fc_Z2Z3)
log2(y$av_Z2Z3)


Box Plots Bootstrap total normalized Z
Size class fc Z1Z2
Size class fc Z1Z3
Size class fc Z2Z3


MA Plots Bootstrap total patman X
log2(y$fc_X1X2)
log2(y$fc_X1X3)
log2(y$fc_X2X3)
log2(y$av_X1X2)
log2(y$av_X1X3)
log2(y$av_X2X3)


Box Plots Bootstrap total patman X
Size class fc X1X2
Size class fc X1X3
Size class fc X2X3


MA Plots Bootstrap total patman Z
log2(y$fc_Z1Z2)
log2(y$fc_Z1Z3)
log2(y$fc_Z2Z3)
log2(y$av_Z1Z2)
log2(y$av_Z1Z3)
log2(y$av_Z2Z3)


Box Plots Bootstrap total patman Z
Size class fc Z1Z2
Size class fc Z1Z3
Size class fc Z2Z3

List of Figures

List of Tables

References

[1] Simon Moxon, Frank Schwach, Tamas Dalmay, Dan MacLean, David J Studholme, and Vincent Moulton. A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, 24(19):2252–2253, 2008. 1, 2, 23, 34, 35, 37, 49, 61, 78, 81, 85, 154

[2] Marc R Friedländer, Sebastian D Mackowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research*, 40(1):37–52, 2012. 1, 2, 29, 35, 40, 61, 62, 64, 65, 66, 85, 93, 154

[3] Jiyuan An, John Lai, Atul Sajjanhar, Melanie L Lehman, and Colleen C Nelson. miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. *BMC bioinformatics*, 15(1):275, 2014. 2, 35, 41, 45, 54, 62, 64, 65, 78, 85, 93, 154, 200

[4] Matthew B Stocks, Simon Moxon, Daniel Mapleson, Hugh C Woolfenden, Irina Mohorianu, Leighton Folkes, Frank Schwach, Tamas Dalmay, and Vincent Moulton. 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, 2012. 2, 34, 37, 61, 64, 65, 66, 67, 81, 82, 93, 120, 127

[5] Ana Kozomara and Sam Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73, 2014. 2, 10, 27, 28, 70, 74, 82, 86, 93, 96, 112, 139

[6] Harvey Lodish, David Baltimore, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, and James Darnell. *Molecular cell biology*, volume 3. Scientific American Books New York, 1995. 4, 6, 8

[7] James D Watson and Francis HC Crick. The structure of DNA. In *Cold Spring Harbor symposia on quantitative biology*, volume 18, pages 123–131. Cold Spring Harbor Laboratory Press, 1953. 5, 9

[8] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D Watson, and AV Grimstone. Molecular Biology of the Cell (3rd edn). *Trends in Biochemical Sciences*, 20(5):210–210, 1995. 5, 151

[9] Daniel Horspool. File:Central Dogma of Molecular Biochemistry with Enzymes.jpg. https://commons.wikimedia.org/wiki/File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg, 2008. [Online; accessed 10-October-2016]. 5, 187

[10] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970. 6

[11] Rna structure (molecular biology). http://what-when-how.com/molecular-biology/rna-structure-molecular-biology/, 2016. [Online; accessed 30-September-2016]. 7, 187

[12] John S Mattick and Igor V Makunin. Non-coding RNA. *Human molecular genetics*, 15(suppl 1):R17–R29, 2006. 8

[13] V Narry Kim. Small RNAs: classification, biogenesis, and function. *Mol cells*, 19(1):1–15, 2005. 8

[14] Arne Weiberg, Marschal Bellinger, and Hailing Jin. Conversations between kingdoms: small rnas. *Current opinion in biotechnology*, 32:207–215, 2015. 8

[15] V Narry Kim. MicroRNA biogenesis: coordinated cropping and dicing. *Nature reviews Molecular cell biology*, 6(5):376–385, 2005. 8, 10, 11, 12, 39, 40, 75, 79, 90, 112, 187

[16] David P Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004. 8, 10, 12, 16, 18, 75, 90, 112

[17] Heng-Chi Lee, Liande Li, Weifeng Gu, Zhihong Xue, Susan K Crosthwaite, Alexander Pertsemlidis, Zachary A Lewis, Michael Freitag, Eric U Selker, Craig C Mello, et al. Diverse pathways generate microRNA-like RNAs and Dicer-independent small interfering RNAs in fungi. *Molecular cell*, 38(6):803–814, 2010. 8

[18] Jiahong Zhou, Yanping Fu, Jiatao Xie, Bo Li, Daohong Jiang, Guoqing Li, and Jiasen Cheng. Identification of microRNA-like RNAs in a plant pathogenic fungus Sclerotinia sclerotiorum by high-throughput sequencing. *Molecular genetics and genomics*, 287(4):275–282, 2012.

[19] Susanna KP Lau, Wang-Ngai Chow, Annette YP Wong, Julian MY Yeung, Jessie Bao, Na Zhang, Si Lok, Patrick CY Woo, and Kwok-Yung Yuen. Identification of microRNA-like RNAs in mycelial and yeast phases of the thermal dimorphic fungus Penicillium marneffei. *PLoS Negl Trop Dis*, 7(8):e2398, 2013.

[20] Kang Kang, Jiasheng Zhong, Liang Jiang, Gang Liu, Christine Yuan Gou, Qiong Wu, You Wang, Jun Luo, and Deming Gou. Identification of microRNA-Like RNAs in the filamentous fungus Trichoderma reesei by solexa sequencing. *PloS one*, 8(10):e76288, 2013. 8

[21] Xuemei Chen. MicroRNA biogenesis and function in plants. *FEBS letters*, 579(26):5923–5931, 2005. 8, 18, 75, 115

[22] Angie M Cheng, Mike W Byrom, Jeffrey Shelton, and Lance P Ford. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic acids research*, 33(4):1290–1297, 2005. 9, 16

[23] Marilena V Iorio, Manuela Ferracin, Chang-Gong Liu, Angelo Veronese, Riccardo Spizzo, Silvia Sabbioni, Eros Magri, Massimo Pedriali, Muller Fabbri, Manuela Campiglio, et al. MicroRNA gene expression deregulation in human breast cancer. *Cancer research*, 65(16):7065–7070, 2005. 15

[24] Aurora Esquela-Kerscher and Frank J Slack. Oncomirs - microRNAs with a role in cancer. *Nature Reviews Cancer*, 6(4):259–269, 2006. 15, 16

[25] Ming Lu, Qipeng Zhang, Min Deng, Jing Miao, Yanhong Guo, Wei Gao, and Qinghua Cui. An analysis of human microRNA and disease associations. *PloS one*, 3(10):e3420, 2008. 16

[26] Álvaro L Pérez-Quintero, Rafik Neme, Andrés Zapata, and Camilo López. Plant microRNAs and their role in defense against viruses: a bioinformatics approach. *BMC plant biology*, 10(1):1, 2010. 19, 20

[27] Matthew W Jones-Rhoades, David P Bartel, and Bonnie Bartel. MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, 57:19–53, 2006. 9, 18, 19

[28] PP Pashkovskiy and SS Ryazansky. Biogenesis, evolution, and functions of plant microRNAs. *Biochemistry (Moscow)*, 78(6):627–637, 2013. 9

[29] Fabrício F Costa. Non-coding RNAs, epigenetics and complexity. *Gene*, 410(1):9–17, 2008. 9

[30] Ryan J Taft, Michael Pheasant, and John S Mattick. The relationship between non-protein-coding dna and eukaryotic complexity. *Bioessays*, 29(3):288–299, 2007. 9

[31] Jana Hertel and Peter F Stadler. The expansion of animal microRNA families revisited. *Life*, 5(1):905–920, 2015. 9

[32] David P Bartel. MicroRNAs: target recognition and regulatory functions. *cell*, 136(2):215–233, 2009. 9, 12, 13, 14, 18, 79

[33] Benjamin P Lewis, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, 2003. 13

[34] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *cell*, 120(1):15–20, 2005. 9, 13

[35] Nelson C Lau, Lee P Lim, Earl G Weinstein, and David P Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–862, 2001. 10, 40, 69, 74, 80

[36] Meng Xie, Shuxin Zhang, and Bin Yu. microRNA biogenesis, degradation and activity in plants. *Cellular and Molecular Life Sciences*, 72(1):87–99, 2015. 10, 16, 17, 18, 19, 90, 112, 115

[37] Lisa A Urry, Michael L Cain, Steven A Wasserman, Peter V Minorsky, Robert B Jackson, and Jane B Reece. *Campbell biology in focus*. Pearson, 2014. 10

[38] Michel J Weber. New human and mouse microRNA genes found by homology search. *FEBs Journal*, 272(1):59–73, 2005. 10, 90

[39] Baohong Zhang, Xiaoping Pan, Charles H Cannon, George P Cobb, and Todd A Anderson. Conservation and divergence of plant microRNA genes. *The Plant Journal*, 46(2):243–259, 2006. 10, 90

[40] Mariana Lagos-Quintana, Reinhard Rauhut, Winfried Lendeckel, and Thomas Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–858, 2001. 10

[41] Rosalind C Lee and Victor Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–864, 2001. 15

[42] Amy E Pasquinelli, Brenda J Reinhart, Frank Slack, Mark Q Martindale, Mitzi I Kuroda, Betsy Maller, David C Hayward, Eldon E Ball, Bernard Degnan, Peter Müller, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, 2000. 11

[43] Mariana Lagos-Quintana, Reinhard Rauhut, Abdullah Yalcin, Jutta Meyer, Winfried Lendeckel, and Thomas Tuschl. Identification of tissue-specific microRNAs from mouse. *Current biology*, 12(9):735–739, 2002. 10, 15

[44] Rosalind C Lee, Rhonda L Feinbaum, and Victor Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993. 10

[45] Brenda J Reinhart, Frank J Slack, Michael Basson, Amy E Pasquinelli, Jill C Bettinger, Ann E Rougvie, H Robert Horvitz, and Gary Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *nature*, 403(6772):901–906, 2000. 10

[46] Brenda J Reinhart, Earl G Weinstein, Matthew W Rhoades, Bonnie Bartel, and David P Bartel. MicroRNAs in plants. *Genes & development*, 16(13):1616–1626, 2002. 11, 16, 17

[47] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20):4051–4060, 2004. 11, 80, 90, 112

[48] Xuezhong Cai, Curt H Hagedorn, and Bryan R Cullen. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *Rna*, 10(12):1957–1966, 2004. 11, 80, 90, 112

[49] Yoontae Lee, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, Patrick Provost, Olof Rådmark, Sunyoung Kim, et al. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956):415–419, 2003. 11

[50] Jinju Han, Yoontae Lee, Kyu-Hyun Yeom, Young-Kook Kim, Hua Jin, and V Narry Kim. The Drosha-DGCR8 complex in primary microRNA processing. *Genes & development*, 18(24):3016–3027, 2004. 11

[51] Ahmet M Denli, Bastiaan BJ Tops, Ronald HA Plasterk, René F Ketting, and Gregory J Hannon. Processing of primary microRNAs by the Microprocessor complex. *Nature*, 432(7014):231–235, 2004. 11

[52] Richard I Gregory, Kai-ping Yan, Govindasamy Amuthan, Thimmaiah Chendrimada, Behzad Doratotaj, Neil Cooch, and Ramin Shiekhattar.

The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014):235–240, 2004. 11

[53] Yan Zeng, Rui Yi, and Bryan R Cullen. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *The EMBO journal*, 24(1):138–148, 2005. 11, 40, 79

[54] Elsebet Lund, Stephan Güttinger, Angelo Calado, James E Dahlberg, and Ulrike Kutay. Nuclear export of microRNA precursors. *Science*, 303(5654):95–98, 2004. 11

[55] Rui Yi, Yi Qin, Ian G Macara, and Bryan R Cullen. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & development*, 17(24):3011–3016, 2003.

[56] Markus T Bohnsack, Kevin Czaplinski, and DIRK GÖRLICH. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *Rna*, 10(2):185–191, 2004. 11

[57] Emily Bernstein, Amy A Caudy, Scott M Hammond, and Gregory J Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366, 2001. 11

[58] Alla Grishok, Amy E Pasquinelli, Darryl Conte, Na Li, Susan Parrish, Ilho Ha, David L Baillie, Andrew Fire, Gary Ruvkun, and Craig C Mello. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106(1):23–34, 2001. 11

[59] György Hutvágner, Juanita McLachlan, Amy E Pasquinelli, Éva Bálint, Thomas Tuschl, and Phillip D Zamore. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293(5531):834–838, 2001.

[60] René F Ketting, Sylvia EJ Fischer, Emily Bernstein, Titia Sijen, Gregory J Hannon, and Ronald HA Plasterk. Dicer functions in RNA interference and

in synthesis of small RNA involved in developmental timing in C. elegans. *Genes & development*, 15(20):2654–2659, 2001. 11

[61] E Jean Finnegan and Marjori A Matzke. The small RNA world. *Journal of cell science*, 116(23):4689–4693, 2003. 12

[62] Young Sik Lee, Kenji Nakahara, John W Pham, Kevin Kim, Zhengying He, Erik J Sontheimer, and Richard W Carthew. Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*, 117(1):69–81, 2004. 12

[63] Caterina Catalanotto, Massimiliano Pallotta, Paul ReFalo, Matthew S Sachs, Laurence Vayssie, Giuseppe Macino, and Carlo Cogoni. Redundancy of the two dicer genes in transgene-induced posttranscriptional gene silencing in Neurospora crassa. *Molecular and cellular biology*, 24(6):2536–2545, 2004. 12

[64] Minju Ha and V Narry Kim. Regulation of microRNA biogenesis. *Nature reviews Molecular cell biology*, 15(8):509–524, 2014. 12, 13

[65] Anastasia Khvorova, Angela Reynolds, and Sumedha D Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–216, 2003. 13, 39, 40, 79

[66] Sam Griffiths-Jones, Jerome HL Hui, Antonio Marco, and Matthew Ronshaugen. MicroRNA evolution by arm switching. *EMBO reports*, 12(2):172–177, 2011. 13

[67] Ji-Joon Song, Stephanie K Smith, Gregory J Hannon, and Leemor Joshua-Tor. Crystal structure of Argonaute and its implications for RISC slicer activity. *science*, 305(5689):1434–1437, 2004. 13

[68] Jidong Liu, Michelle A Carmell, Fabiola V Rivas, Carolyn G Marsden, J Michael Thomson, Ji-Joon Song, Scott M Hammond, Leemor Joshua-Tor, and Gregory J Hannon. Argonaute2 is the catalytic engine of mammalian RNAi. *Science*, 305(5689):1437–1441, 2004.

[69] Gunter Meister, Markus Landthaler, Agnieszka Patkaniowska, Yair Dorsett, Grace Teng, and Thomas Tuschl. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular cell*, 15(2):185–197, 2004. 13, 14

[70] Marc R Fabian and Nahum Sonenberg. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nature structural & molecular biology*, 19(6):586–593, 2012. 13

[71] Andrew Grimson, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett-Engele, Lee P Lim, and David P Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, 27(1):91–105, 2007. 13

[72] Sergej Djuranovic, Ali Nahvi, and Rachel Green. miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science*, 336(6078):237–240, 2012. 13, 14

[73] Ariel A Bazzini, Miler T Lee, and Antonio J Giraldez. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science*, 336(6078):233–237, 2012. 13, 14

[74] Antonio J Giraldez, Yuichiro Mishima, Jason Rihel, Russell J Grocock, Stijn Van Dongen, Kunio Inoue, Anton J Enright, and Alexander F Schier. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *science*, 312(5770):75–79, 2006. 13, 14, 15

[75] Nadya Morozova, Andrei Zinovyev, Nora Nonne, Linda-Louise Pritchard, Alexander N Gorban, and Annick Harel-Bellan. Kinetic signatures of microRNA modes of action. *Rna*, 18(9):1635–1655, 2012. 13

[76] Clare A Beelman, Audrey Stevens, Giordano Caponigro, Thomas E LaGrandeur, Lianna Hatfield, David M Fortner, and Roy Parker. An essential component of the decapping enzyme required for normal rates of mRNA turnover. 1996. 13

[77] Hedda A Meijer, Martin Bushell, Kirsti Hill, Timothy W Gant, Anne E Willis, Peter Jones, and Cornelia H de Moor. A novel method for poly (A) fractionation reveals a large population of mRNAs with a short poly (A) tail in mammalian cells. *Nucleic acids research*, 35(19):e132, 2007. 13

[78] Soraya Yekta, I-hung Shih, and David P Bartel. MicroRNA-directed cleavage of HOXB8 mRNA. *Science*, 304(5670):594–596, 2004. 14

[79] Antonio Marco, Jamie I MacPherson, Matthew Ronshaugen, and Sam Griffiths-Jones. MicroRNAs from the same precursor have different targeting properties. *Silence*, 3(1):1, 2012. 14

[80] Stefan L Ameres and Phillip D Zamore. Diversifying microRNA sequence and function. *Nature reviews Molecular cell biology*, 14(8):475–488, 2013. 14, 18, 188

[81] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*, 19(1):92–105, 2009. 14

[82] Juliana Giusti, Danillo Pinhal, Simon Moxon, Camila Lovaglio Campos, Andrea Münsterberg, and Cesar Martins. MicroRNA-10 modulates Hox genes expression during Nile tilapia embryonic development. *Mechanisms of development*, 140:12–18, 2016. 14, 15

[83] Elizabeth E Caygill and Laura A Johnston. Temporal regulation of metamorphic processes in Drosophila by the let-7 and miR-125 heterochronic microRNAs. *Current Biology*, 18(13):943–950, 2008. 14, 15

[84] Eric A Miska, Ezequiel Alvarez-Saavedra, Matthew Townsend, Akira Yoshii, Nenad Šestan, Pasko Rakic, Martha Constantine-Paton, and H Robert Horvitz. Microarray analysis of microRNA expression in the developing mammalian brain. *Genome biology*, 5(9):1, 2004. 14, 15

[85] Po Yu Chen, Heiko Manninga, Krasimir Slanchev, Minchen Chien, James J Russo, Jingyue Ju, Robert Sheridan, Bino John, Debora S Marks, Dimos

Gaidatzis, et al. The developmental miRNA profiles of zebrafish as determined by small RNA cloning. *Genes & development*, 19(11):1288–1293, 2005. 15

[86] Xiangqian Guo, Songkun Su, Geir Skogerboe, Shuanjin Dai, Wenfeng Li, Zhiguo Li, Fang Liu, Ruifeng Ni, Yu Guo, Shenglu Chen, et al. Recipe for a busy bee: microRNAs in Honey Bee caste determination. *PLoS One*, 8(12):e81661, 2013. 15

[87] George Adrian Calin, Cinzia Sevignani, Calin Dan Dumitru, Terry Hyslop, Evan Noch, Sai Yendamuri, Masayoshi Shimizu, Sashi Rattan, Florencia Bullrich, Massimo Negrini, et al. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proceedings of the National academy of Sciences of the United States of America*, 101(9):2999–3004, 2004. 15

[88] Stefano Volinia, George A Calin, Chang-Gong Liu, Stefan Ambs, Amelia Cimmino, Fabio Petrocca, Rosa Visone, Marilena Iorio, Claudia Roldo, Manuela Ferracin, et al. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National academy of Sciences of the United States of America*, 103(7):2257–2261, 2006. 15

[89] Fadila Guessous, Ying Zhang, Alex Kofman, Alessia Catania, Yunqing Li, David Schiff, Benjamin Purow, and Roger Abounader. microRNA-34a is tumor suppressive in brain tumors and glioma stem cells. *Cell cycle*, 9(6):1031–1036, 2010. 15

[90] George Adrian Calin, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, Sashi Rattan, Michael Keating, Kanti Rai, et al. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*, 99(24):15524–15529, 2002. 15

[91] Mehmet Somel, Song Guo, Ning Fu, Zheng Yan, Hai Yang Hu, Ying Xu, Yuan Yuan, Zhibin Ning, Yuhui Hu, Corinna Menzel, et al. MicroRNA,

mRNA, and protein expression link development and aging in human and macaque brain. *Genome research*, 20(9):1207–1218, 2010. 16, 85

[92] Sébastien S Hébert, Katrien Horré, Laura Nicolaï, Aikaterini S Papadopoulou, Wim Mandemakers, Asli N Silahtaroglu, Sakari Kauppinen, André Delacourte, and Bart De Strooper. Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/β-secretase expression. *Proceedings of the National Academy of Sciences*, 105(17):6415–6420, 2008. 16

[93] Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic acids research*, 37(suppl 1):D98–D104, 2009. 16

[94] Xi Chen, Yi Ba, Lijia Ma, Xing Cai, Yuan Yin, Kehui Wang, Jigang Guo, Yujing Zhang, Jiangning Chen, Xing Guo, et al. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell research*, 18(10):997–1006, 2008. 16

[95] Douglas D Taylor and Cicek Gercel-Taylor. MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. *Gynecologic oncology*, 110(1):13–21, 2008.

[96] Carolina Salazar, Rahul Nagadia, Pratibala Pandit, Justin Cooper-White, Nilanjana Banerjee, Nevenka Dimitrova, William B Coman, and Chamindie Punyadeera. A novel saliva-based microRNA biomarker panel to detect head and neck cancers. *Cellular Oncology*, 37(5):331–338, 2014. 16

[97] GA Ganepola, John R Rutledge, Paritosh Suman, Anusak Yiengpruksawan, and David H Chang. Novel blood-based microRNA biomarker panel for early diagnosis of pancreatic cancer. *World J Gastrointest Oncol*, 6(1):22–33, 2014. 16

[98] Harry LA Janssen, Hendrik W Reesink, Eric J Lawitz, Stefan Zeuzem, Maribel Rodriguez-Torres, Keyur Patel, Adriaan J van der Meer, Amy K

Patick, Alice Chen, Yi Zhou, et al. Treatment of HCV infection by targeting microRNA. *New England Journal of Medicine*, 368(18):1685–1694, 2013. 16

[99] Shijun Hu, Mei Huang, Zongjin Li, Fangjun Jia, Zhumur Ghosh, Maarten A Lijkwan, Pasquale Fasanaro, Ning Sun, Xi Wang, Fabio Martelli, et al. MicroRNA-210 as a novel therapy for treatment of ischemic heart disease. *Circulation*, 122(11 suppl 1):S124–S131, 2010.

[100] Robert E Lanford, Elisabeth S Hildebrandt-Eriksen, Andreas Petri, Robert Persson, Morten Lindow, Martin E Munk, Sakari Kauppinen, and Henrik Ørum. Therapeutic silencing of microRNA-122 in primates with chronic hepatitis C virus infection. *Science*, 327(5962):198–201, 2010. 16

[101] Yukio Kurihara and Yuichiro Watanabe. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(34):12753–12758, 2004. 16, 17

[102] Zhixin Xie, Edwards Allen, Noah Fahlgren, Adam Calamar, Scott A Givan, and James C Carrington. Expression of Arabidopsis MIRNA genes. *Plant physiology*, 138(4):2145–2154, 2005. 16

[103] István Papp, M Florian Mette, Werner Aufsatz, Lucia Daxinger, Stephen E Schauer, Animesh Ray, Johannes Van Der Winden, Marjori Matzke, and Antonius JM Matzke. Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant physiology*, 132(3):1382–1390, 2003. 17

[104] Zhixin Xie, Lisa K Johansen, Adam M Gustafson, Kristin D Kasschau, Andrew D Lellis, Daniel Zilberman, Steven E Jacobsen, and James C Carrington. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol*, 2(5):e104, 2004. 17

[105] Wonkeun Park, Junjie Li, Rentao Song, Joachim Messing, and Xuemei Chen. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein,

act in microRNA metabolism in Arabidopsis thaliana. *Current Biology*, 12(17):1484–1495, 2002. 17

[106] E Jean Finnegan, Rogerio Margis, and Peter M Waterhouse. Posttranscriptional gene silencing is not compromised in the Arabidopsis CARPEL FACTORY (DICER-LIKE1) mutant, a homolog of Dicer-1 from Drosophila. *Current Biology*, 13(3):236–240, 2003. 17

[107] Virginie Gasciolli, Allison C Mallory, David P Bartel, and Hervé Vaucheret. Partially redundant functions of Arabidopsis DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. *Current Biology*, 15(16):1494–1500, 2005. 17

[108] Rogerio Margis, Adriana F Fusaro, Neil A Smith, Shaun J Curtin, John M Watson, E Jean Finnegan, and Peter M Waterhouse. The evolution and diversification of Dicers in plants. *FEBS letters*, 580(10):2442–2450, 2006. 17

[109] Ramya Rajagopalan, Hervé Vaucheret, Jerry Trejo, and David P Bartel. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes & development*, 20(24):3407–3425, 2006. 17

[110] Liang Wu, Huanyu Zhou, Qingqing Zhang, Jianguang Zhang, Fangrui Ni, Chang Liu, and Yijun Qi. DNA methylation mediated by a microRNA pathway. *Molecular cell*, 38(3):465–475, 2010. 17

[111] Liang Song, Michael J Axtell, and Nina V Fedoroff. RNA secondary structural determinants of miRNA precursor processing in Arabidopsis. *Current Biology*, 20(1):37–41, 2010. 17, 40, 79

[112] Schallum Werner, Heike Wollmann, Korbinian Schneeberger, and Detlef Weigel. Structure determinants for accurate processing of miR172a in Arabidopsis thaliana. *Current Biology*, 20(1):42–48, 2010.

[113] Julieta L Mateos, Nicolás G Bologna, Uciel Chorostecki, and Javier F Palatnik. Identification of microRNA processing determinants by random muta-

genesis of Arabidopsis MIR172a precursor. *Current Biology*, 20(1):49–54, 2010. 17, 40, 79

[114] Josh T Cuperus, Noah Fahlgren, and James C Carrington. Evolution and functional diversification of MIRNA genes. *The Plant Cell*, 23(2):431–442, 2011. 17, 39, 40, 79, 115

[115] Bin Yu, Zhiyong Yang, Junjie Li, Svetlana Minakhina, Maocheng Yang, Richard W Padgett, Ruth Steward, and Xuemei Chen. Methylation as a crucial step in plant microRNA biogenesis. *Science*, 307(5711):932–935, 2005. 18

[116] Mee Yeon Park, Gang Wu, Alfredo Gonzalez-Sulser, Hervé Vaucheret, and R Scott Poethig. Nuclear processing and export of microRNAs in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3691–3696, 2005. 18

[117] Krista M Bollman, Milo J Aukerman, Mee-Yeon Park, Christine Hunter, Tanya Z Berardini, and R Scott Poethig. HASTY, the Arabidopsis ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. *Development*, 130(8):1493–1504, 2003. 18

[118] N Baumberger and DC Baulcombe. Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11928–11933, 2005. 18

[119] Andrew L Eamens, Neil A Smith, Shaun J Curtin, Ming-Bo Wang, and Peter M Waterhouse. The Arabidopsis thaliana double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. *Rna*, 15(12):2219–2235, 2009.

[120] Liang Wu, Qingqing Zhang, Huanyu Zhou, Fangrui Ni, Xueying Wu, and Yijun Qi. Rice MicroRNA effector complexes and targets. *The Plant Cell*, 21(11):3421–3435, 2009. 18

[121] Wuli Bao, David M O'Malley, Ross Whetten, and Ronald R Sederoff. A laccase associated with lignification in loblolly pine xylem. *SCIENCE-NEW YORK THEN WASHINGTON-*, 260:672–672, 1993. 19

[122] Milo J Aukerman and Hajime Sakai. Regulation of flowering time and floral organ identity by a microRNA and its APETALA2-like target genes. *The Plant Cell*, 15(11):2730–2741, 2003. 19

[123] Javier F Palatnik, Edwards Allen, Xuelin Wu, Carla Schommer, Rebecca Schwab, James C Carrington, and Detlef Weigel. Control of leaf morphogenesis by microRNAs. *Nature*, 425(6955):257–263, 2003. 19

[124] Rebecca Schwab, Javier F Palatnik, Markus Riester, Carla Schommer, Markus Schmid, and Detlef Weigel. Specific effects of microRNAs on the plant transcriptome. *Developmental cell*, 8(4):517–527, 2005. 19

[125] Peter Huijser and Markus Schmid. The control of developmental phase transitions in plants. *Development*, 138(19):4117–4129, 2011. 19

[126] Stephen E Schauer, Steven E Jacobsen, David W Meinke, and Animesh Ray. DICER-LIKE1: blind men and elephants in Arabidopsis development. *Trends in plant science*, 7(11):487–491, 2002. 19

[127] Jo Ann Banks. MicroRNA, sex determination and floral meristem determinacy in maize. *Genome biology*, 9(1):1, 2008. 19

[128] Basel Khraiwesh, Jian-Kang Zhu, and Jianhua Zhu. Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(2):137–148, 2012. 19, 20

[129] Ramanjulu Sunkar, Yong-Fang Li, and Guru Jagadeeswaran. Functions of microRNAs in plant stress responses. *Trends in plant science*, 17(4):196–203, 2012. 19

[130] Xuefeng Zhou, Guandong Wang, and Weixiong Zhang. UV-B responsive microRNA genes in Arabidopsis thaliana. *Molecular systems biology*, 3(1):103, 2007. 20

[131] Shanfa Lu, Ying-Hsuan Sun, Rui Shi, Catherine Clark, Laigeng Li, and Vincent L Chiang. Novel and mechanical stress–responsive microRNAs in Populus trichocarpa that are absent from Arabidopsis. *The Plant Cell*, 17(8):2186–2203, 2005. 20

[132] Surekha Katiyar-Agarwal and Hailing Jin. Role of small RNAs in host-microbe interactions. *Annual review of phytopathology*, 48:225, 2010. 20

[133] Shanfa Lu, Ying-Hsuan Sun, Henry Amerson, and Vincent L Chiang. MicroRNAs in loblolly pine (Pinus taeda L.) and their association with fusiform rust gall development. *The Plant Journal*, 51(6):1077–1098, 2007. 20

[134] Xiang-Feng He, Yuan-Yuan Fang, Lei Feng, and Hui-Shan Guo. Characterization of conserved and novel microRNAs and their targets, including a TuMV-induced TIR–NBS–LRR class R gene-derived novel miRNA in Brassica. *FEBS letters*, 582(16):2445–2452, 2008. 20

[135] Mogens S Hovmøller, Amor H Yahyaoui, Eugene A Milus, and Annemarie F Justesen. Rapid global spread of two aggressive strains of a wheat rust fungus. *Molecular Ecology*, 17(17):3818–3826, 2008. 20

[136] J Graham Ruby, Calvin H Jan, and David P Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–86, 2007. 21

[137] Qian-Hao Zhu, Andrew Spriggs, Louisa Matthew, Longjiang Fan, Gavin Kennedy, Frank Gubler, and Chris Helliwell. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome research*, 18(9):1456–1465, 2008. 21

[138] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008. 21, 34

[139] Allan M Maxam and Walter Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2):560–564, 1977. 22

[140] Frederick Sanger, Steven Nicklen, and Alan R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977. 22

[141] F Sanger. Nucleotide sequence of bacteriophage (d x174 dna. 1977. 22

[142] Howard S Bilofsky, Christian Burks, James W Fickett, Walter B Goad, Frances I Lewitter, Wayne P Rindone, C David Swindell, and Chang-Shung Tung. The GenBank genetic sequence databank. *Nucleic acids research*, 14(1):1–4, 1986. 22

[143] Elaine R Mardis. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008. 22

[144] Pauline C Ng and Ewen F Kirkness. Whole genome sequencing. In *Genetic variation*, pages 215–226. Springer, 2010. 22, 123

[145] Anthony Rhoads and Kin Fai Au. PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015. 22, 23

[146] Hengyun Lu, Francesca Giordano, and Zemin Ning. Oxford nanopore minion sequencing and genome assembly. *Genomics, proteomics & bioinformatics*, 14(5):265–279, 2016. 22, 23

[147] Simon Bennett. Solexa ltd. *Pharmacogenomics*, 5(4):433–438, 2004. 22

[148] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008. 22, 23, 120, 124

[149] Marc R Friedländer, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel, and Nikolaus Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology*, 26(4):407–415, 2008. 23, 34, 35, 41, 43, 62, 80, 189

[150] Monya Baker. Next-generation sequencing: adjusting to data overload. *nature methods*, 7(7):495–499, 2010. 23, 66

[151] Tracy Tucker, Marco Marra, and Jan M Friedman. Massively parallel sequencing: the next big thing in genetic medicine. *The American Journal of Human Genetics*, 85(2):142–154, 2009. 23, 66, 67

[152] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, 2012. 25, 116

[153] Kay Prüfer, Udo Stenzel, Michael Dannemann, Richard E Green, Michael Lachmann, and Janet Kelso. PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, 24(13):1530–1531, 2008. 25, 31, 37, 82, 86, 90, 115, 131, 139

[154] Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):1, 2011. 26, 39, 42, 48, 78, 82

[155] Eric Bonnet, Jan Wuyts, Pierre Rouzé, and Yves Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–2917, 2004. 26, 40, 80, 82

[156] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. miRBase: tools for microRNA genomics. *Nucleic acids research*, 36(suppl 1):D154–D158, 2008. 27

[157] Ana Kozomara and Sam Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, page gkq1027, 2010. 27

[158] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. NCBI GEO: archive for functional genomics data setsupdate. *Nucleic acids research*, 41(D1):D991–D995, 2013. 27, 85, 86, 121

[159] National Institutes of Health et al. International nucleotide sequence database collaboration. 27

[160] Eric P Nawrocki, Sarah W Burge, Alex Bateman, Jennifer Daub, Ruth Y Eberhardt, Sean R Eddy, Evan W Floden, Paul P Gardner, Thomas A Jones, John Tate, et al. Rfam 12.0: updates to the RNA families database. *Nucleic acids research*, page gku1063, 2014. 28, 29, 88, 139

[161] Paul P Gardner, Jennifer Daub, John G Tate, Eric P Nawrocki, Diana L Kolbe, Stinus Lindgreen, Adam C Wilkinson, Robert D Finn, Sam Griffiths-Jones, Sean R Eddy, et al. Rfam: updates to the rna families database. *Nucleic acids research*, 37(suppl 1):D136–D140, 2009. 29

[162] Jeongsoo Lee, Dong-in Kim, June Hyun Park, Ik-Young Choi, and Chanseok Shin. miRAuto: An automated user-friendly MicroRNA prediction tool utilizing plant small RNA sequencing data. *Molecules and cells*, 35(4):342–347, 2013. 29, 35, 47, 51, 63, 189

[163] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–1771, 2010. 29

[164] Wikipedia. Fastq format — wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=FASTQ_format&oldid=740151340, 2016. [Online; accessed 23-September-2016]. 29

[165] David J Lipman and William R Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, 1985. 30

[166] SAM/BAM Format Specification Working Group et al. Sequence alignment/map format specification, 2014. 31

[167] Aaron R Quinlan. BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, pages 11–12, 2014. 32, 90, 111

[168] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic acids research*, page gkq1019, 2010. 32, 85, 86

[169] Anthony Mathelier and Alessandra Carbone. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, 26(18):2226–2234, 2010. 34, 35, 45, 53, 54, 63, 200

[170] Michael Hackenberg, Martin Sturm, David Langenberger, Juan Manuel Falcon-Perez, and Ana M Aransay. miranalyzer: a microrna detection and analysis tool for next-generation sequencing experiments. *Nucleic acids research*, 37(suppl 2):W68–W76, 2009. 34, 35, 46, 63

[171] Xiaozeng Yang and Lei Li. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics*, 27(18):2614–2615, 2011. 35, 41, 44, 62

[172] Ming Wen, Yang Shen, Suhua Shi, and Tian Tang. miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. *BMC bioinformatics*, 13(1):140, 2012. 35, 44

[173] David Langenberger, Sachin Pundhir, Claus T Ekstrøm, Peter F Stadler, Steve Hoffmann, and Jan Gorodkin. deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns. *Bioinformatics*, 28(1):17–24, 2012. 35, 46

[174] Ping Xuan, Maozu Guo, Yangchao Huang, Wenbin Li, and Yufei Huang. MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs. *PloS one*, 6(11):e27422, 2011. 35, 47

[175] Jiyuan An, John Lai, Melanie L Lehman, and Colleen C Nelson. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic acids research*, 41(2):727–737, 2013. 35, 41, 44, 52, 62, 63, 199

[176] Jikai Lei and Yanni Sun. miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. *Bioinformatics*, page btu380, 2014. 35, 48, 54, 62, 63, 78

[177] Susan Higashi, Cyril Fournier, Christian Gautier, Christine Gaspin, and Marie-France Sagot. Mirinho: An efficient and general plant and animal pre-miRNA predictor for genomic and deep sequencing data. *BMC bioinformatics*, 16(1):1, 2015. 35, 48

[178] Maurits Evers, Michael Huttner, Anne Dueck, Gunter Meister, and Julia C Engelmann. miRA: adaptable novel miRNA identification in plants using small RNA sequencing data. *BMC bioinformatics*, 16(1):1, 2015. 35, 48, 50, 51, 62, 78, 199

[179] Lan Yu, Chaogang Shao, Xinghuo Ye, Yijun Meng, Yincong Zhou, and Ming Chen. miRNA Digger: a comprehensive pipeline for genome-wide novel miRNA mining. *Scientific reports*, 6, 2016. 35

[180] Yue Li, Zhuo Zhang, Feng Liu, Wanwipa Vongsangnak, Qing Jing, and Bairong Shen. Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic acids research*, page gks043, 2012. 36, 45, 58, 59, 60, 62, 63, 64, 66, 190, 200

[181] Vernell Williamson, Albert Kim, Bin Xie, G Omari McMichael, Yuan Gao, and Vladimir Vladimirov. Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. *Briefings in bioinformatics*, page bbs010, 2012. 55, 56, 57, 62, 63, 190, 200

[182] Wenjing Kang and Marc R Friedländer. Computational prediction of miRNA genes from small RNA sequencing data. *Frontiers in bioengineering and biotechnology*, 3:7, 2015. 36, 66

[183] Leighton Folkes, Simon Moxon, Hugh C Woolfenden, Matthew B Stocks, Gyorgy Szittya, Tamas Dalmay, and Vincent Moulton. PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. *Nucleic acids research*, 40(13):e103–e103, 2012. 37, 81, 155

[184] Irina Mohorianu, Matthew Benedict Stocks, John Wood, Tamas Dalmay, and Vincent Moulton. CoLIde: a bioinformatics tool for CO-expression based small RNA Loci Identification using high-throughput sequencing data. *RNA biology*, 10(7):1221–1230, 2013. 37, 74, 81, 120

[185] Blake C Meyers, Michael J Axtell, Bonnie Bartel, David P Bartel, David Baulcombe, John L Bowman, Xiaofeng Cao, James C Carrington, Xuemei Chen, Pamela J Green, et al. Criteria for annotation of plant MicroRNAs. *The Plant Cell*, 20(12):3186–3190, 2008. 38

[186] Ben Langmead. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics*, pages 11–7, 2010. 41, 42, 46

[187] Emil J Gumbel. The return period of flood flows. *The annals of mathematical statistics*, 12(2):163–190, 1941. 43, 80

[188] Dongli He, Qiong Wang, Kun Wang, and Pingfang Yang. Genome-Wide Dissection of the MicroRNA Expression Profile in Rice Embryo during Early Stages of Seed Germination. *PloS one*, 10(12):e0145424, 2015. 45, 95

[189] Li-Chuan Wan, Feng Wang, Xiangqian Guo, Shanfa Lu, Zongbo Qiu, Yuanyuan Zhao, Haiyan Zhang, and Jinxing Lin. Identification and characterization of small non-coding RNAs from Chinese fir by high throughput sequencing. *BMC plant biology*, 12(1):1, 2012.

[190] Li-Chuan Wan, Haiyan Zhang, Shanfa Lu, Liang Zhang, Zongbo Qiu, Yuanyuan Zhao, Qing-Yin Zeng, and Jinxing Lin. Transcriptome-wide identification and characterization of miRNAs from Pinus densata. *BMC genomics*, 13(1):1, 2012. 45, 95

[191] Hui-juan Wang, Peng-jun Zhang, Wei-jun Chen, Deng Jie, Feng Dan, Yanhong Jia, and Li-xin Xie. Characterization and Identification of novel serum microRNAs in sepsis patients with different outcomes. *Shock*, 39(6):480–487, 2013. 45

[192] Martin Sturm, Michael Hackenberg, David Langenberger, and Dmitrij Frishman. TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC bioinformatics*, 11(1):1, 2010. 46

[193] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970. 46

[194] David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985. 47

[195] Attila Molnár, Frank Schwach, David J Studholme, Eva C Thuenemann, and David C Baulcombe. miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*, 447(7148):1126–1129, 2007. 50

[196] Zhenhai Zhang, Jingyin Yu, Daofeng Li, Zuyong Zhang, Fengxia Liu, Xin Zhou, Tao Wang, Yi Ling, and Zhen Su. PMRD: plant microRNA database. *Nucleic acids research*, 38(suppl 1):D806–D813, 2010. 51

[197] Brian E Howard and Steffen Heber. Towards reliable isoform quantification using RNA-SEQ data. *BMC bioinformatics*, 11(3):1, 2010. 55

[198] Po-Jung Huang, Yi-Chung Liu, Chi-Ching Lee, Wei-Chen Lin, Richie Ruei-Chi Gan, Ping-Chiang Lyu, and Petrus Tang. DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic acids research*, page gkq392, 2010. 55, 58

[199] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. 57

[200] Wei-Chi Wang, Feng-Mao Lin, Wen-Chi Chang, Kuan-Yu Lin, Hsien-Da Huang, and Na-Sheng Lin. miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC bioinformatics*, 10(1):1, 2009. 58

[201] David Hendrix, Michael Levine, and Weiyang Shi. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome biology*, 11(4):1, 2010. 58

[202] Erle Zhu, Fangqing Zhao, Gang Xu, Huabin Hou, LingLin Zhou, Xiaokun Li, Zhongsheng Sun, and Jinyu Wu. mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic acids research*, 38(suppl 2):W392–W397, 2010. 58

[203] Roy Ronen, Ido Gan, Shira Modai, Alona Sukacheov, Gideon Dror, Eran Halperin, and Noam Shomron. miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics*, 26(20):2615–2616, 2010. 58

[204] Alan P Boyle, Justin Guinney, Gregory E Crawford, and Terrence S Furey. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21):2537–2538, 2008. 68

[205] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, 5(9):829–834, 2008. 68

[206] Anthony P Fejes, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven JM Jones. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, 2008.

[207] Zhaohui S Qin, Jianjun Yu, Jincheng Shen, Christopher A Maher, Ming Hu, Shanker Kalyana-Sundaram, Jindan Yu, and Arul M Chinnaiyan. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC bioinformatics*, 11(1):369, 2010. 69

[208] David A Nix, Samir J Courdy, and Kenneth M Boucher. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC bioinformatics*, 9(1):523, 2008.

[209] Christiana Spyrou, Rory Stark, Andy G Lynch, and Simon Tavaré. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC bioinformatics*, 10(1):299, 2009. 68

[210] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75, 2009.

[211] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology*, 26(12):1351–1359, 2008.

[212] Teemu D Laajala, Sunil Raghav, Soile Tuomela, Riitta Lahesmaa, Tero Aittokallio, and Laura L Elo. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC genomics*, 10(1):618, 2009.

[213] Elizabeth G Wilbanks and Marc T Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PloS one*, 5(7):e11471, 2010.

[214] Debashis Ghosh and Zhaohui S Qin. Statistical issues in the analysis of ChIP-Seq and RNA-Seq data. *Genes*, 1(2):317–334, 2010.

[215] S Wilder. SWEMBL: a generic peak-calling program, 2010. 68

[216] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature biotechnology*, 26(11):1293–1300, 2008. 68

[217] G Palshikar et al. Simple algorithms for peak detection in time-series. In *Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence*, pages 1–13, 2009. 69

[218] Pan Du, Warren A Kibbe, and Simon M Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, 2006.

[219] RJ Urban, WS Evans, AD Rogol, DL Kaiser, ML Johnson, and Johannes D Veldhuis. Contemporary aspects of discrete peak-detection algorithms. i. the paradigm of the luteinizing hormone pulse signal in men. *Endocrine Reviews*, 9(1):3–37, 1988.

[220] Felix Scholkmann, Jens Boss, and Martin Wolf. An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals. *Algorithms*, 5(4):588–603, 2012. 69

[221] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 70

[222] Irina Mohorianu, Frank Schwach, Runchun Jing, Sara Lopez-Gomollon, Simon Moxon, Gyorgy Szittya, Karim Sorefan, Vincent Moulton, and Tamas Dalmay. Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *The Plant Journal*, 67(2):232–246, 2011. 72, 88, 120, 130

[223] Ryan D Morin, Michael D O'Connor, Malachi Griffith, Florian Kuchenbauer, Allen Delaney, Anna-Liisa Prabhu, Yongjun Zhao, Helen McDonald, Thomas Zeng, Martin Hirst, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome research*, 18(4):610–621, 2008. 76

[224] Daniel Cifuentes, Huiling Xue, David W Taylor, Heather Patnode, Yuichiro Mishima, Sihem Cheloufi, Enbo Ma, Shrikant Mane, Gregory J Hannon, Nathan D Lawson, et al. A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science*, 328(5986):1694–1698, 2010. 85

[225] Young-Kook Kim, Boseon Kim, and V Narry Kim. 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, 2016. 85, 89, 98, 192

[226] Marc R Friedländer, Esther Lizano, Anna JS Houben, Daniela Bezdan, Mónica Báñez-Coronel, Grzegorz Kudla, Elisabet Mateu-Huertas, Birgit

Kagerbauer, Justo González, Kevin C Chen, et al. Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol*, 15(4):R57, 2014. 88, 91

[227] Candida Vaz, Hafiz M Ahmad, Pratibha Sharma, Rashi Gupta, Lalit Kumar, Ritu Kulshreshtha, and Alok Bhattacharya. Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood. *BMC genomics*, 11(1):1, 2010.

[228] Jin Hou, Li Lin, Weiping Zhou, Zhengxin Wang, Guoshan Ding, Qiongzhu Dong, Lunxiu Qin, Xiaobing Wu, Yuanyuan Zheng, Yun Yang, et al. Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma. *Cancer cell*, 19(2):232–243, 2011.

[229] Chanseok Shin, Jin-Wu Nam, Kyle Kai-How Farh, H Rosaria Chiang, Alena Shkumatava, and David P Bartel. Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular cell*, 38(6):789–802, 2010. 85

[230] Andrew D Bosson, Jesse R Zamudio, and Phillip A Sharp. Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Molecular cell*, 56(3):347–359, 2014. 85

[231] Andrew J Modzelewski, Stephanie Hilz, Elizabeth A Crate, Caterina TH Schweidenback, Elizabeth A Fogarty, Jennifer K Grenier, Raimundo Freire, Paula E Cohen, and Andrew Grimson. Dgcr8 and Dicer are essential for sex chromosome integrity during meiosis in males. *J Cell Sci*, 128(12):2314–2327, 2015.

[232] Haneul Noh, Charny Park, Soojun Park, Young Seek Lee, Soo Young Cho, and Hyemyung Seo. Prediction of miRNA-mRNA associations in Alzheimers disease mice using network topology. *BMC genomics*, 15(1):1, 2014.

[233] J Groenendyk, C Hetz, L Kurgan, and M Michalak. P125Endoplasmic reticulum stress responses to disrupted endoplasmic reticulum ca2+ homeostasis. *Cardiovascular research*, 103(suppl 1):S22–S22, 2014.

[234] Jody Groenendyk, Xiao Fan, Zhenling Peng, Yaroslav Ilnytskyy, Lukasz Kurgan, and Marek Michalak. Genome-wide analysis of thapsigargin-induced microRNAs and their targets in NIH3T3 cells. *Genomics data*, 2:325–327, 2014.

[235] Xiangbing Meng, Shujie Yang, Yuping Zhang, Xinjun Wang, Renee X Goodfellow, Yichen Jia, Kristina W Thiel, Henry D Reyes, Baoli Yang, and Kimberly K Leslie. Genetic deficiency of Mtdh gene in mice causes male infertility via impaired spermatogenesis and alterations in the expression of small non-coding RNAs. *Journal of Biological Chemistry*, 290(19):11853–11864, 2015. 85

[236] Laura Garcia-Segura, Cei Abreu-Goodger, Armando Hernandez-Mendoza, Tzvetanka D Dimitrova Dinkova, Luis Padilla-Noriega, Martha Elva Perez-Andrade, and Juan Miranda-Rios. High-Throughput Profiling of Caenorhabditis elegans Starvation-Responsive microRNAs. *PloS one*, 10(11):e0142262, 2015. 85

[237] Mihye Lee, Yeon Choi, Kijun Kim, Hua Jin, Jaechul Lim, Tuan Anh Nguyen, Jihye Yang, Minsun Jeong, Antonio J Giraldez, Hui Yang, et al. Adenylation of maternally inherited microRNAs by Wispy. *Molecular cell*, 56(5):696–707, 2014. 85

[238] Alison K Surridge, Sara Lopez-Gomollon, Simon Moxon, Luana S Maroja, Tina Rathjen, Nicola J Nadeau, Tamas Dalmay, and Chris D Jiggins. Characterisation and expression of microRNAs in developing wings of the neotropical butterfly Heliconius melpomene. *BMC genomics*, 12(1):1, 2011. 85

[239] Ayisha Ahmed, Nicole J Ward, Simon Moxon, Sara Lopez-Gomollon, Camille Viaut, Matthew L Tomlinson, Ilya Patrushev, Michael J Gilchrist,

Tamas Dalmay, Dario Dotlic, et al. A Database of microRNA Expression Patterns in Xenopus laevis. *PloS one*, 10(10):e0138313, 2015. 85

[240] Sara Lopez-Gomollon, Irina Mohorianu, Gyorgy Szittya, Vincent Moulton, and Tamas Dalmay. Diverse correlation patterns between microRNAs and their targets during tomato fruit development indicates different modes of microRNA actions. *Planta*, 236(6):1875–1887, 2012. 85

[241] Michael Kravchik, Ramanjulu Sunkar, Subha Damodharan, Ran Stav, Matat Zohar, Tal Isaacson, and Tzahi Arazi. Global and local perturbation of the tomato microRNA pathway by a trans-activated DICER-LIKE 1 mutant. *Journal of experimental botany*, 65(2):725–739, 2014. 85

[242] Shaun J Curtin, Jean-Michel Michno, Benjamin W Campbell, Javier Gil-Humanes, Sandra M Mathioni, Reza Hammond, Juan J Gutierrez-Gonzalez, Ryan C Donohue, Michael B Kantar, Andrew L Eamens, et al. MicroRNA maturation and microRNA target gene expression regulation are severely disrupted in soybean dicer-like1 double mutants. *G3: Genes—Genomes— Genetics*, 6(2):423–433, 2016. 85

[243] Huan Wang, Xiuren Zhang, Jun Liu, Takatoshi Kiba, Jongchan Woo, Tolulope Ojo, Markus Hafner, Thomas Tuschl, Nam-Hai Chua, and Xiu-Jie Wang. Deep sequencing of small RNAs specifically associated with Arabidopsis AGO1 and AGO4 uncovers new AGO functions. *THe plant journal*, 67(2):292–304, 2011. 85

[244] Müşerref Duygu Saçar, Hamid Hamzeiy, and Jens Allmer. Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins. *Journal of integrative bioinformatics*, 10(2):215, 2013. 86, 93, 98, 112, 114

[245] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008. 87, 141, 142

[246] Kevin P McCormick, Matthew R Willmann, and Blake C Meyers. Experimental design, preprocessing, normalization and differential expression

analysis of small RNA sequencing experiments. *Silence*, 2(1):1, 2011. 87, 120, 123, 140, 141, 142, 149

[247] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013. 87, 142

[248] RNAcentral Consortium et al. RNAcentral: an international database of ncRNA sequences. *Nucleic acids research*, page gku991, 2014. 88

[249] Matthew Beckers, Irina Mohorianu, Matthew Stocks, Christopher Applegate, Tamas Dalmay, and Vincent Moulton. Comprehensive processing of high throughput small RNA sequencing data including quality checking, normalization and differential expression analysis using the UEA sRNA Workbench. *under revision*, 2016. 88, 120, 142, 144, 145, 155

[250] Eric Londin, Phillipe Loher, Aristeidis G Telonis, Kevin Quann, Peter Clark, Yi Jing, Eleftheria Hatzimichael, Yohei Kirino, Shozo Honda, Michelle Lally, et al. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate-and tissue-specific microRNAs. *Proceedings of the National Academy of Sciences*, 112(10):E1106–E1115, 2015. 91, 110

[251] Jr-Shiuan Yang and Eric C Lai. Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. *Molecular cell*, 43(6):892–903, 2011. 98

[252] Jakub O Westholm and Eric C Lai. Mirtrons: microRNA biogenesis via splicing. *Biochimie*, 93(11):1897–1904, 2011. 98

[253] Mandy Chung. Using JConsole to monitor applications. *Sun Developer Network*, 2004. 106

[254] Noe Fernandez-Pozo, Naama Menda, Jeremy D Edwards, Surya Saha, Isaak Y Tecle, Susan R Strickler, Aureliano Bombarely, Thomas Fisher-

York, Anuradha Pujar, Hartmut Foerster, et al. The Sol Genomics Network (SGN)from genotype to phenotype to breeding. *Nucleic acids research*, 43(D1):D1036–D1041, 2015. 111

[255] Kasper D Hansen, Zhijin Wu, Rafael A Irizarry, and Jeffrey T Leek. Sequencing technology does not eliminate biological variability. *Nature biotechnology*, 29(7):572–573, 2011. 120

[256] David J Studholme. Deep sequencing of small RNAs in plants: applied bioinformatics. *Briefings in functional genomics*, page elr039, 2011. 120, 132

[257] Margaret A Taub, Hector Corrada Bravo, and Rafael A Irizarry. Overcoming bias and systematic errors in next generation sequencing data. *Genome medicine*, 2(12):1, 2010. 132, 133

[258] Sam EV Linsen, Elzo de Wit, Georges Janssens, Sheila Heater, Laura Chapman, Rachael K Parkin, Brian Fritz, Stacia K Wyman, Ewart de Bruijn, Emile E Voest, et al. Limitations and possibilities of small RNA digital gene expression profiling. *Nature methods*, 6(7):474–476, 2009. 120, 123

[259] David S DeLuca, Joshua Z Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, 2012. 120

[260] Liguo Wang, Shengqin Wang, and Wei Li. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28(16):2184–2185, 2012. 120

[261] Matthew PA Davis, Stijn van Dongen, Cei Abreu-Goodger, Nenad Bartonicek, and Anton J Enright. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63(1):41–49, 2013. 120

[262] Yan Peng, Andrew S Maxwell, Natalie D Barker, Jennifer G Laird, Alan J Kennedy, Nan Wang, Chaoyang Zhang, and Ping Gong. SeqAssist: a novel toolkit for preliminary analysis of next-generation sequencing data. *BMC bioinformatics*, 15(11):1, 2014. 120

[263] Antonio Rueda, Guillermo Barturen, Ricardo Lebrón, Cristina Gómez-Martín, Ángel Alganza, José L Oliver, and Michael Hackenberg. sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic acids research*, 43(W1):W467–W473, 2015. 120

[264] Simon Andrews et al. FastQC: A quality control tool for high throughput sequence data. *Reference Source*, 2010. 120

[265] Ping Xu, Irina Mohorianu, Li Yang, Hansheng Zhao, Zhimin Gao, and Tamas Dalmay. Small RNA profile in moso bamboo root and leaf obtained by high definition adapters. *PloS one*, 9(7):e103590, 2014. 120, 123, 126, 127

[266] Ping Xu, Martina Billmeier, Irina Mohorianu, Darrell Green, William D Fraser, and Tamas Dalmay. An improved protocol for small RNA library construction using high definition adapters. *Methods Next Generation Seq*, 2:1–10, 2015. 126, 127

[267] Tina Rathjen, Helio Pais, Dylan Sweetman, Vincent Moulton, Andrea Munsterberg, and Tamas Dalmay. High throughput sequencing of microRNAs in chicken somites. *FEBS letters*, 583(9):1422–1426, 2009. 120

[268] Wikipedia. Box plot. https://en.wikipedia.org/wiki/Box_plot, 2016. [Online; accessed 25-October-2016]. 122

[269] Sandrine Dudoit, Yee Hwa Yang, Matthew J Callow, and Terence P Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, pages 111–139, 2002. 122

[270] Wikipedia. Ma plot. https://en.wikipedia.org/wiki/MA_plot, 2016. [Online; accessed 31-October-2016]. 122

[271] Raimundo Real and Juan M Vargas. The probabilistic basis of Jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996. 122

[272] Vicki Pandey, Robert C Nutter, and Ellen Prediger. Applied biosystems solid system: ligation-based sequencing. *Next Generation Genome Sequencing: Towards Personalized Medicine*, pages 29–42, 2008. 123

[273] Larry W McLAUGHLIN, Elcna ROMANIUK, Paul J ROMANIUK, and Thomas NEILSON. The effect of acceptor oligoribonucleotide sequence on the T4 RNA ligase reaction. *European Journal of Biochemistry*, 125(3):639–643, 1982. 123

[274] Daniela B Munafó and G Brett Robb. Optimization of enzymatic reaction conditions for generating representative pools of cDNA from small RNA. *Rna*, 16(12):2537–2552, 2010. 123, 124

[275] Anitha D Jayaprakash, Omar Jabado, Brian D Brown, and Ravi Sachidanandam. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic acids research*, page gkr693, 2011. 127

[276] Karim Sorefan, Helio Pais, Adam E Hall, Ana Kozomara, Sam Griffiths-Jones, Vincent Moulton, and Tamas Dalmay. Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence*, 3(1):1, 2012. 123, 127

[277] Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924):218–223, 2009. 123

[278] Lira Mamanova, Robert M Andrews, Keith D James, Elizabeth M Sheridan, Peter D Ellis, Cordelia F Langford, Tobias WB Ost, John E Collins, and Daniel J Turner. FRT-seq: Amplification-free, strand-specific, transcriptome sequencing. *Nature methods*, 7(2):130, 2010. 124

[279] Iwanka Kozarewa, Zemin Ning, Michael A Quail, Mandy J Sanders, Matthew Berriman, and Daniel J Turner. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes. *Nature methods*, 6(4):291–295, 2009. 124

[280] R Core Team et al. R: A language and environment for statistical computing. 2013. 124, 136

[281] Preethi H Gunaratne, Cristian Coarfa, Benjamin Soibam, and Arpit Tandon. miRNA data analysis: next-gen sequencing. *Next-Generation MicroRNA Expression Profiling Technology: Methods and Protocols*, pages 273–288, 2012. 127

[282] Irina Mohorianu, Amanda Bretman, Damian Smith, Emily Fowler, Tamas Dalmay, and Tracey Chapman. Novel approaches for analysing multi-level RNA-seq data: Sampling-based normalization and hierarchical differential expression. *under revision*, 2016. 137, 142

[283] Li Guo and Feng Chen. A challenge for miRNA: multiple isomiRs in miRNAomics. *Gene*, 544(1):1–7, 2014. 140, 148, 149

[284] Alicia Oshlack and Matthew J Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology direct*, 4(1):1, 2009. 141

[285] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, 2010. 141

[286] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11(1):1, 2010. 142

[287] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003. 142, 143

[288] Yee Hwa Yang and Natalie P Thorne. Normalization for two-color cDNA microarray data. *Lecture Notes-Monograph Series*, pages 403–418, 2003. 142, 143

[289] Jun Li and Robert Tibshirani. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical methods in medical research*, 22(5):519–536, 2013. 142, 144

[290] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994. 142, 144

[291] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):1, 2014. 145

[292] Xiaobei Zhou, Helen Lindsay, and Mark D Robinson. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic acids research*, 42(11):e91–e91, 2014. 145, 146

[293] Sorin Drăghici. *Data analysis tools for DNA microarrays*. CRC Press, 2003. 145

[294] Noah Fahlgren, Christopher M Sullivan, Kristin D Kasschau, Elisabeth J Chapman, Jason S Cumbie, Taiowa A Montgomery, Sunny D Gilbert, Mark Dasenko, Tyler WH Backman, Scott A Givan, et al. Computational and analytical framework for small RNA profiling by high-throughput sequencing. *Rna*, 15(5):992–1002, 2009. 145

[295] Irina Mohorianu, Sara Lopez-Gomollon, Frank Schwach, Tamas Dalmay, and Vincent Moulton. FiRePatfinding regulatory patterns between sRNAs and genes. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(3):273–284, 2012. 145

[296] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, 14(1):1, 2013. 146

[297] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology*, 14(9):1, 2013. 146

[298] AE Hall, WT Lu, JD Godfrey, AV Antonov, C Paicu, S Moxon, T Dalmay, A Wilczynska, PAJ Muller, and M Bushell. The cytoskeleton adaptor protein ankyrin-1 is upregulated by p53 following DNA damage and alters cell migration. *Cell death & disease*, 7(4):e2184, 2016. 152

[299] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010. 152