# Quality checking and expression analysis of high-throughput small RNA sequencing data

Matthew Lloyd Beckers
School of Computing Sciences
University of East Anglia

A thesis submitted for the degree of
*Doctor of Philosophy*

September 2015

# Abstract

The advent of high-throughput RNA sequencing (RNA-seq) methods have made it possible to sequence transcriptomes for the cell-wide identification of small non-coding RNAs (sRNAs) and to assess their regulation using differential expression analysis by comparing two or more different conditions. During an analysis of a typical set of sRNA sequencing (sRNA-seq) libraries, a large variety of tools and methods are used on the dataset in order to understand the data's quality, content, and to summarise the knowledge gained from the entire analysis. Many of the tools available to do this were created for mRNA sequencing (mRNA-seq) datasets. In this thesis, we present and implement a processing pipeline that can be used to assess the quality and the differential expression of sRNA-seq datasets over two or more different conditions. We then utilise aspects of this pipeline in various sRNA-seq experiments. Firstly, we combine our pipeline with current tools for miRNA identification to assess the regulation of miRNAs during larval caste differentiation in a novel genome; the European bumblebee (*Bombus terrestris*). Secondly, we explore the differential expression during cell stress of all classes of sRNAs using two cell lines in humans. We also find that a specific protein, Ro60, is required for the expression of mRNA-derived sRNAs during stress, similar to the way in which sRNAs derived from Y RNAs are regulated. Finally, we utilise our understanding of sRNA mapping patterns, alongside current tools for miRNA identification, to search for functional miRNAs and other sRNAs in the novel genomes of two diatoms. The lack of canonical miRNA predictions in this study has repercussions for the evolutionary theory behind miRNAs. The implementation of our pipeline for sRNA-seq data provides an interactive and quality controlled workflow that can be used to process a dataset from raw sequences to the results of several differential expression experiments for all identified sRNA classes within a sequenced transcriptome.

# Acknowledgements

Firstly, I would like to thank my supervisor, Vincent Moulton, and my co-supervisors Tamas Dalmay and Thomas Mock for their support and guidance. Additionally, I would like to thank Irina Mohorianu, Matt Stocks and members of Tamas Dalmay's lab group for their help and collaboration throughout this PhD. Finally, I would like to thank my partner, Kate Kerry, and my parents, Sue and Martin, for their love and support over the last four years and beyond.

# List of publications

- Lopez-Gomollon S, **Beckers M**, Rathjen T, Moxon S, Maumus F, Mohorianu I, Moulton V, Dalmay T, Mock T, *Global discovery and characterization of small non-coding RNAs in marine microalgae* BMC Genomics, 15:697, 2014.

- The Bumblebee Genome Consortium, *The genomes of two key bumblebee species with primitive eusocial organization*, Genome Biology, 16:76, 2015.

- **Beckers M**, Mohorianu I, Stocks M, Applegate A, Dalmay T, Moulton V, *An interactive pipeline for quality checking, normalisation, and differential expression analysis of high-throughput small RNA sequence data*, in preparation.

- Collins DH, **Beckers M**, Mohorianu I, Moulton V, Dalmay T, Bourke AFG, *A MicroRNA Associated With Caste Determination in a Bumblebee is Expressed from a Mirtron Within a Homologue of Vitellogenin*, in preparation.

# Contents

# List of Figures

# List of Tables

# Glossary

**FASTQ** A human-readable file format that stores nucleotide sequences and their associated quality scores. 14

**HTS** The name given to the generation of sequencing technologies that parallelized the sequencing process to greatly increase the rate at which reads can be sequenced. 15

**miRNA** A class of sRNAs that are between 21-23nt in length. They are produced from a ncRNA transcript that forms a distinct hairpin-like secondary structure. 8, 9

**mRNA** RNA that is transcribed from DNA before being transported to the cytoplasm where it is used to encode proteins.. 11

**mRNA-seq** An alternative name for RNA-seq to clearly differentiate it from sRNA-seq. 2

**piRNA** A class of sRNAs that are 26-31nt in length. They form an RNA-protein complex with piwi proteins and are involved in epigenetic silencing. 9

**RNA-seq** A next generation sequencing tehcnology that measures the quantity and presense of RNA in an organism's transcriptome at a given moment in time. 1

**RNAi** The name given to the process and pathways whereby RNA sequences are used to silence translation of protein from mRNA. 1

**snoRNA** A class of sRNA that guides chemical modifications of other RNAs. 11

**sRNA** A class of RNAs that are 20-25 nucleotides in length. Many types of sRNAs act as sequence specificity guides in the RNAi pathway. 1

**sRNA-seq** A form of RNA-seq that focusses on an organism's sRNA transcript content. 2

**tRNA** An RNA molecule used to transport amino acids to ribosomes for protein synthesis. 11

**tsRNA** A sRNA that is derived from the longer tRNA transcript through RNA cleavage. 11

**Y RNA** A class of ncRNAs that are between 84-113nt in length. Its secondary structure bares similarity to that of a miRNA. 12, 80

**YsRNA** A sRNA that is 22-36nt in length and is produced from the longer Y RNA transcript. 80

# Chapter 1

# Introduction

An organism's cellular processes are based upon the flow of biological information from its genes to its products. This information is encoded and exchanged by a number of different biopolymers in a set of processes termed the Central Dogma of molecular biology (figure 1.1). Deoxyribonucleic acid (DNA) is responsible for storing the information within an organism's genome. The information is utilised by first transcribing it to ribonucleic acid (RNA), where it becomes functional within an organism's transcriptome. From here, the information can be translated to protein, an organism's fundamental building blocks, from messenger RNA (mRNA). However, a gene may also be transcribed to other classes of RNA, called non-coding RNAs (ncRNAs) [Eddy, 2001], that use their encoded information to both process and regulate the flow of information from the genome to the proteome. The regulation of genetic expression allows genes to have a more complex relationship with their products.

In particular, small non-coding RNAs (sRNAs) have been found to regulate the expression of mRNA through a process known as RNA interference (RNAi) [Mello and Conte, 2004]. With the continued use of next generation sequencing, and in particular RNA sequencing (RNA-seq) technologies, research on sRNAs has broadened in several directions. There is now a heavy focus on projects that sequence whole sRNA transcriptomes with the aim to expand the known sRNAs family as well as our knowledge of the kind of organisms that utilise these sRNAs. This has lead to a further expanded understanding of a wide range of other ncRNAs that can be processed to produce sRNAs under certain circumstances [Tuck and Tollervey, 2011]. In Chapter 2 we review the biology of sRNAs. This includes briefings on the sRNA classes that use the RNAi pathway and also the more recent insights into the extended group of non-RNAi sRNAs.

Although the use of RNA-seq for mRNA experiments is very common, sRNAs

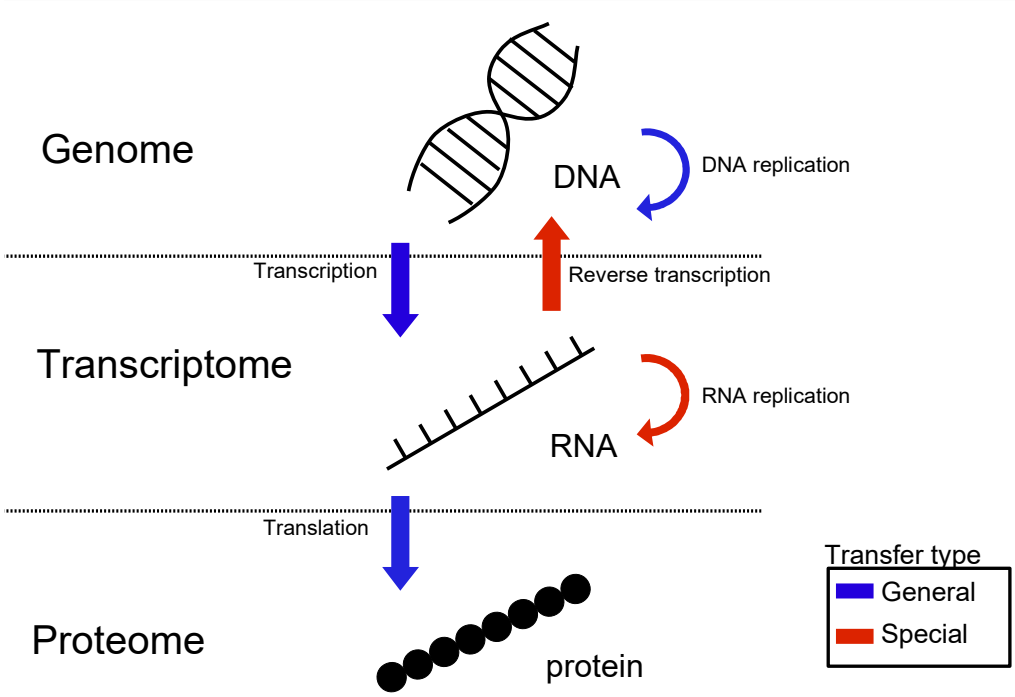



Figure 1.1: The Central Dogma of molecular biology [Crick, 1970] states that information can be exchanged between DNA, RNA, and proteins in a number of different ways. Three general transfers are known to occur in all organisms. A further three special transfers occur under special circumstances or in only certain classes of life (e.g. reverse transcription in viruses).

can also be studied by using a variation of this sequencing method (sRNA-seq). To process sRNA-seq data, it is common to utilise tools that were originally created for use on mRNA-seq data. These include quality check tools such as FASTQC [Andrews, 2010], and normalisation and differential expression packages such as edgeR [Robinson et al., 2010] and DESeq2 [Love et al., 2014]. FastQC can be used to aid the quality control of any data in a FASTQC format, whereas edgeR and DESeq2 focus on the analysis of preprocessed RNAseq experiments. However, sRNA sequencing (sRNA-seq) data is different to the data found from mRNA sequencing (mRNA-seq) and, as we find in this thesis, it is extremely beneficial to take the particular characteristics of sRNAs into account when preprocessing and calculating differential expression in sRNA experiments. Chapter 3 reviews the computational methods and tools for the preprocessing, manipulation, and analysis of sRNA-seq datasets.

With the advent of next generation sequencing in the early 2000s, the amount of data that can be retrieved for an experiment has been dramatically increasing owing to the reduction in costs per nucleotide [Stein, 2010]. Although more

in-depth data strengthens the results and conclusions of biological experiments, the rate of increase has now eclipsed that of Moore's Law, which describes the rate at which computer processing power has been increasing since the 1960s. This has repercussions for the ease in which researchers can complete their experiments, often requiring the use of efficient tools, specialised data processing centres, or cloud computing [Stein, 2010]. In Chapter 4 we describe a sRNA-seq processing pipeline that we developed over the course of this thesis that contains novel approaches to the quality checking, normalisation, and differential expression of sRNA-seq data. We also provide a comparison of stages of this pipeline to those of other widely used tools, and detail our own final implementation of this tool, aimed at both bioinformaticians and biologists for use on their own personal computers. Dr. Matthew Stocks of the University of East Anglia (UEA) implemented the workflow and much of the interfaces for our tool. In addition, some of the ideas and concepts for the pipeline were provided by Dr. Irina Mohorianu (UEA) with further input by the author. The author's contributions included all data analysis, comparisons to other approaches, and the implementation of specific methods and visualisations into the software.

In Chapter 5 we describe an analysis of sRNA-seq data from a novel organism, the European bumblebee (*Bombus terrestris*), to identify new and conserved miRNAs and understand the regulation of these miRNAs during larval caste development. Dr. David Collins (UEA) undertook the biological experiments and sample preparations described in this chapter.

In Chapter 6 we utilise two sRNA-seq datasets to understand the regulation of mammalian sRNA transcriptomes during cell stress. Biological experiments and sample preparations were carried out by Dr. Adam Hall (University of Leicester) and Dr. Carly Turnbull (UEA) with follow up northern blot validations by Martina Billmeier and Prince Panicker.

In Chapter 7 we carry out an investigation on sRNA transcriptomes of two diatom species, which are members of the eukaryotic supergroup Chromalveolata, in which sRNAs have been less studied [Cerutti and Casas-Mollano, 2006]. We use mapping characteristics of single sRNA-seq libraries for both organisms to identify and predict the presence of various types of sRNAs in these two organisms. RNA sample preparations and northern blot validations were carried out by Dr. Sara-Lopez Gomollon (CSIC) and Dr. Tina Rathjen. Growth experiments were conducted by Dr. Thomas Mock.

The bioinformatics work undertaken in Chapters 5, 6, and 7 was carried out by the author in consultation with Prof. Vincent Moulton (UEA), Prof. Tamas

Dalmay (UEA), and Dr. Irina Mohorianu (UEA).

Finally, Chapter 8 discusses the work completed in this thesis as a whole, and suggests proposals for future work.

# Chapter 2

# The biology of small RNAs

## 2.1 Summary

This chapter details the biological background surrounding sRNA research. This is primarily split into two fields of research: that of sRNAs known to be involved in the RNAi pathway, and secondly the more recent identification of an extended group of sRNAs derived from other non-coding RNAs.

## 2.2 The discovery of RNA interference

The discovery of the RNAi pathway was the result of an attempt to develop a new gene knockdown technique that could be used to downregulate genes. Fire et al. [1991] had previously shown that injecting antisense RNA into the worm *Caenorhabditis elegans* inhibited the expression of complimentary mRNA. However, Guo and Kemphues [1995] later found that gene expression was inhibited by both antisense and sense RNA, which they used as a control. Fire et al. [1998] went on to show that, in addition to this surprising result, double stranded RNA (dsRNA) was able to suppress gene expression at a higher level than the use of sense or antisense RNA alone. In fact, the silencing effect had only ever been caused by small amounts of dsRNA within purified assays of the single-stranded RNA.

The studies in *C. elegans* defined the new silencing mechanism as RNA interference. Instead of the oringal hypothesis that antisense RNA was able to compliment mRNA and block translation on its own, dsRNA was involved in a more complex and efficient silencing pathway that is catalytic at its core. Studies in plants had previously uncovered a very similar phenomenon under the name of post-transcriptional gene silencing (PTGS) [Jorgensen et al., 1996; Que and

Jorgensen, 1998]. A third eukaryotic lineage, fungi, have also been shown to contain a similar mechanism [Fulci and Macino, 2007; Nicolas et al., 2010], which has been termed "quelling". These biological pathways all contain related proteins that are involved in a regulatory mechanism shared by many lineages, suggesting an origin in the eukaryotic ancestor [Mello and Conte, 2004].

### 2.2.1 The canonical RNAi pathway

Generally, RNAi describes a pathway that is triggered by a number of different forms of dsRNA. The dsRNA is processed by RNase-III-type endonucleases. These are a family of proteins that include Dicer and Drosha (see section 2.2.2). These endonucleases cleave the long dsRNAs into much smaller RNAs of a specific length at specific points along the RNA. The small RNAs produced by Dicer-like proteins are referred to as the mature sequence and the transcript from which it is cut from is its precursor. One strand of the dsRNA is then incorporated into a protein complex called the RNAi silencing complex (RISC), along with the Argonaute (AGO) protein. AGO serves to recruit the short strand of RNA, which itself is able to guide RISC to a mRNA target. Once bound, the RISC prevents the mRNA from being translated.

Whilst this pathway is common to all sRNAs, the specific details differ between both the class of sRNA being processed and the organism that produces them [Carthew and Sontheimer, 2009]. The following sections summarise the protein machinery involved in sRNA processing and the different classes of RNA that are processed by them.

### 2.2.2 RNA interference machinery

#### Dicer

Dicer-like proteins are members of the RNase III family, a group of proteins that catalyse the cleavage of double-stranded RNA [Carthew and Sontheimer, 2009]. Table 2.1 lists the domains found in a typical Dicer in the order that they commonly appear along the protein (N terminus to C terminus).

The PAZ domain, shared with Argonaute proteins, binds RNA duplex ends, utilising the characteristic short overhangs of the dsRNA. Two RNase III domains perform the actual excision of the sRNA from its precursor. The sRNA's length is dictated by the distance between the PAZ domain and the protein's processing centre [MacRae et al., 2007]. Different categories of sRNAs are processed by

different variants of Dicer (usually termed Dicer-like) but most are cut into lengths ranging from 21 to 24 nucleotides [Meister and Tuschl, 2004].

Dicer domains are conserved across many eukaryotic lineages even if their domain structure and organisation is subject to variation [Cerutti and Casas-Mollano, 2006]. The greatest variability is the absence of the dsRBD and PAZ domain. *Giardia intestinalis* contains a Dicer composed of the PAZ and RNase III domains only and has been shown to function and process 25-27nt sRNAs [MacRae et al., 2006]. Other species, such as *T. brucei*, encode only RNaseIII domains [Shi et al., 2006]. Therefore the only domain that is continuous across RNAi-functional organisms are the two RNase III domains [Cerutti and Casas-Mollano, 2006].

### Argonaute

Once the mature dsRNA has been excised from its precursor, one of its strands is incorporated into an RNAi Silencing Complex (RISC) that includes the Argonaute protein. The other strand, in most cases, is no longer needed and is degraded. The Argonaute's choice of strand is dependent on the stability of the 5' ends [Khvorova et al., 2003; Schwarz et al., 2003]. A helicase enzyme is responsible for unwinding the duplex and will do so from the easier, or less stable, end. Another function of the 5' end is determining the type of Argonaute that the sRNA binds to [Kim, 2008]. The large array of *Arabidopsis* Argonautes select their sRNAs depending on their bias towards a particular base at the most 5' position. Many studies have also found that sRNAs hold a particular preference towards pyrimidine bases, particularly uridine [Aravin et al., 2006, 2003; Chen et al., 2005], and this has become a prominent feature for the presence of sRNAs in sequencing analysis.

The dual PAZ - PIWI domain structure of Argonaute is well-conserved amongst eukaryotic lineages [Cerutti and Casas-Mollano, 2006]. One exception is the highly divergent *Giardia intestinalis*, which contains a divergent PIWI domain.

Table 2.1: Common domains within the DICER group of RNase III proteins

| Domain | Function |
|---|---|
| DEX D/H | Missing in some; varying function |
| DUF 283 | Might bind to Dicer cofactors |
| PAZ | RNA duplex binding; measurement of mature sRNA. Missing in some |
| Platform | Present in some; unknown function |
| RNase III x2 | Cleavage of dsRNA |
| dsRBD | Binds other end of dicer. Numbers between 0 - 2 |

## 2.3 The evolution of RNA interference

This section describes what is currently understood about the many different mechanisms, functions, and pathways of RNA interference, and how they relate to the eukaryotic lineages that use them.

Although sRNAs involved in silencing pathways are produced at highly specific lengths, the size range varies between different organisms and even different silencing pathways within the same organism. Baulcombe [2004] describe plant siRNAs as between 21 to 26 nucleotides in length and micro-RNAs (miRNAs) as having a slightly narrower range of 21 to 24 nucleotides. A broader range of 19 to 28 nucleotides is described by Kim [2005] for all siRNAs and 18 to 24 nucleotides for miRNAs.

Classification of sRNAs is predominantly defined by the differences in biogenesis, since all sRNAs mediate silencing through the same mechanisms [Kim, 2005]. RNAi has emerged as a gene regulatory mechanism that is evolutionary conserved between eukaryote lineages but with a large number of class-specific, and even species-specific, variations [Meister and Tuschl, 2004].

Since there is conservation of certain RNAi pathways over many different lineages, it is likely that the last common ancestor of eukaryotes had a working set of RNAi machinery and pathways. However, as shown by a few unicellular eukaryotes, RNAi is also not essential for eukaryotic organisms [Cerutti and Casas-Mollano, 2006]. A handful of organisms analysed so far show no signs of an RNAi pathway as we know it. RNAi machinery are absent from some small genomes such as those of *Saccharomyces cerevisiae*, the excavates *Trypanosoma cruzi* and *Leishmania major*, the Archaeplastida *Cyanidioschyzon merolae*, and the malaria-causing *Plasmodium falciparum* [Cerutti and Casas-Mollano, 2006]. Although these organisms are all single-celled, other similar organisms have been shown to utilise a familiar RNAi pathway, particularly *Chlamydomonas reinhartdii* [Molnr et al., 2007]. It is not clear whether RNAi is linked with the advent of multicellularity or more simply a product of genome complexity [Casas-Mollano et al., 2008].

The following sections detail the important classes of sRNAs that have been identified so far and that appear in the majority of organisms, suggesting an ancient origin.

### 2.3.1 Small interfering RNAs

Small interfering RNAs (siRNAs) were originally discovered in plants, where they were involved in post-transcriptional gene silencing and predominantly in response to viral infection. The early discoveries suggested sRNAs originally evolved in eukaryotes to defend against invading viruses [Baulcombe, 2004]. An origin in the ancient eukaryotic ancestor is likely due to the presence of some kind of related RNA silencing mechanism in plants, animals, and fungi.

siRNAs have also been found that are derived from repetitive elements, particularly transposons. These are stretches of DNA that often encode the machinery necessary to replicate themselves across the genome. To control these "genome parasites", many organisms utilise the silencing activity of sRNAs [Malone and Hannon, 2009], such as piwi-interacting small RNAs (piRNAs). These sRNAs are 24-31nt in length and are found in animal gametes where they associate with the piwi protein [Brennecke et al., 2007] to guide the silencing of transposons [Malone and Hannon, 2009]. Although other sRNAs outside of gametes have been found to be derived from transposable elements in plants and animals, the evidence suggests that these have a limited abundance [Ghildiyal and Zamore, 2009].

### 2.3.2 Micro-RNAs

Micro-RNAs (miRNAs) were originally discovered as RNAs between 21 to 26 nucleotides in length that derive from a precursor hairpin structure transcribed from an organism's own genes [Baulcombe, 2004]. Although some types of siRNAs are endogenous to the organism, the primary distinction between the two classes are that siRNAs are processed from a length of dsRNA, whilst the miRNA's precursor is always a hairpin structure [Kim, 2005].

The miRNA pathways have diverged most notably between plant and animal lineages. This variation is apparent even in the final mechanisms of action of the silencing effector complex that catalyses the silencing of targets.

In animals, the processing of miRNA from dsRNA is achieved by two RNA-III-type endonucleases called Drosha and Dicer [Kim, 2005]. The processing is dependent on the transcribed RNA folding into a stem-loop structure, termed the primary miRNA (pri-miRNA) [Carthew and Sontheimer, 2009] (see figure 2.1 for an example). In the nucleus, Drosha, as part of the microprocessor complex [Carthew and Sontheimer, 2009], cuts out the precursor stem-loop (pre-miRNA) from specific positions on the transcript so that a 2-nucleotide long 3' overhang is left at the end of the hairpin secondary structure [Meister and Tuschl, 2004]. The

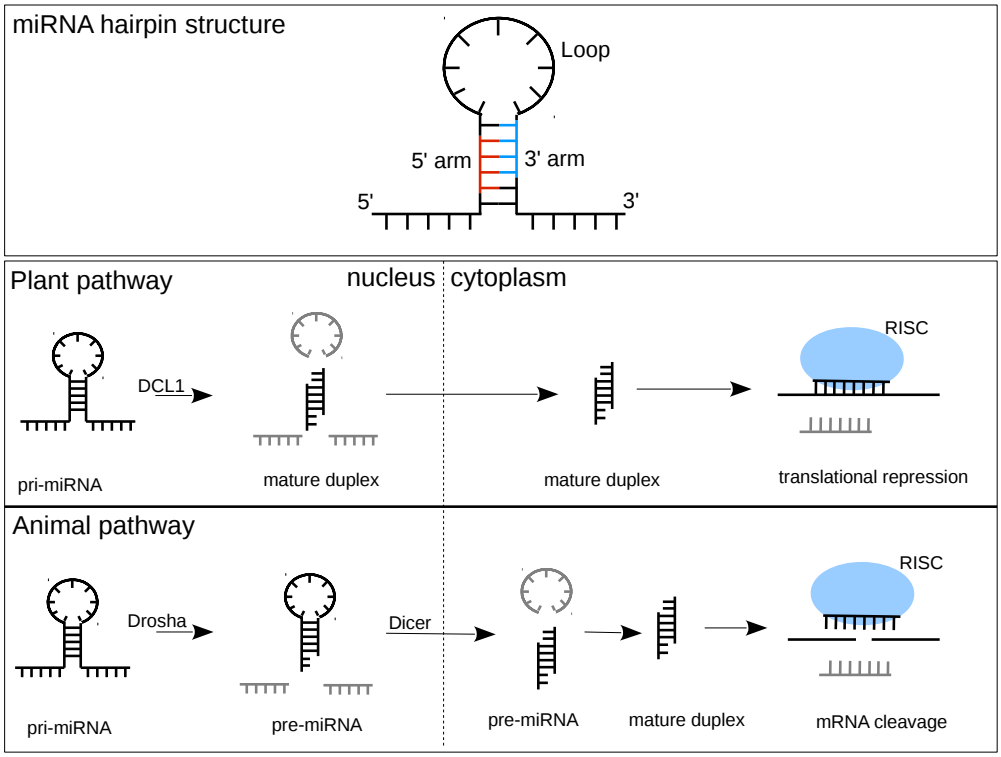



Figure 2.1: Differences between the two miRNA biogenesis pathways in plants and animals.

pre-miRNA is exported from the nucleus where a Dicer excises a short double-stranded RNA with 3' 2-nucleotide overhangs. One of these strands is incorporated into the effector complex as the mature miRNA. The other is degraded.

The maturation of plant miRNA is simpler. The pri-miRNA transcript is processed in the nucleus into the miRNA duplexes by Dicer alone [Baulcombe, 2004], usually DCL1 in *Arabidopsis* [Carthew and Sontheimer, 2009]. Drosha has only been identified within metazoan organisms, and so is thought to be a diverged form of the animal dicer protein [Cerutti and Casas-Mollano, 2006]. Other differences in plant and animal miRNAs include the length of pre-miRNA, which is more variable in plants at 100 to 900 nt. In animals the pre-miRNA is usually restricted to below 100 nt in length [Cuperus et al., 2011].

Since miRNAs have only been identified so far in two of the six eukaryotic supergroups, and the biogenesis of the miRNAs differ between these supergroups, miRNAs are expected to have evolved independently in these lineages and were not present in the last common eukaryotic ancestor [Cerutti and Casas-Mollano, 2006].

## 2.4 The extended small RNA pathway

The several classes of sRNAs summarised above are well-defined in the literature. The biogenesis of these sRNAs are longer precursor RNA transcripts whose sole purpose is to generate sRNAs to regulate gene expression. However, it is becoming apparent that both novel classes and current classes of sRNAs are derived from transcripts that have other primary purposes, representing a layering of information on the genome [Tuck and Tollervey, 2011]. The following section details the current literature on new and novel sRNAs that are thought to be associated with at least parts of the RNAi pathway.

### 2.4.1 Small RNAs derived from other RNA transcripts

As our understanding of genomes progresses, its complexity is becoming increasingly realised. Genes were initially seen as having a one-to-one relationship with the transcripts and proteins that they code for. However, this understanding is becoming increasingly challenged on every level - from transcription to epigenetics. For example, the introns in genes may be spliced out in varying combinations, resulting in different proteins after translation [Matlin et al., 2005].

Non-coding RNA is now beginning to reveal alternative functions where the original transcript is further processed, leaving behind stable fragments of specific length that are likely functional [Tuck and Tollervey, 2011]. Non-coding RNAs that show this activity include transfer RNA (tRNAs) [Cole et al., 2009; Haussecker et al., 2010; Lee et al., 2009], messenger RNA (mRNA), and small nucleolar RNAs (snoRNAs).

#### tRNA-derived small RNAs (tsRNAs)

Several studies in human and mouse organisms [Cole et al., 2009; Haussecker et al., 2010; Lee et al., 2009] have identified distinct classes of sRNAs that all derive from tRNAs, termed tRNA-derived sRNAs (tsRNAs). These classes, summarised in table 2.2 are generally defined by where the sRNA derives from the tRNA. Lee et al. [2009] defines three classes: tRF-3 fragments from matching exactly to the 3' end, tRF-5 fragments mapping exactly to the 5' end of the tRNA, and tRF-1 fragments that map to the 3' end of the pre-tRNA sequence. tRF-3 fragments are found with the CCA motif that is post-transcriptionally appended to the 3' end of tRNAs, confirming that tRF-3 fragments are at least processed from mature tRNAs. Haussecker et al. [2010] categorizes their findings into type

Table 2.2: The classifications of various tRNA-derived RNAs

| Category | lengths | tRNA Position | Cleavage mechanism |
|---|---|---|---|
| tRF-5 (type I) | 18-22 | 5' end of tRNA | Dicer |
| tRF-3 (type I) | 18-22 | 3' end of tRNA | Dicer |
| tRF-1 (type II) | 17-25 | 3' end of pre-tRNA | RNA-Z |
| tRNA halves | 30-35 | Either end | Stress-activated nucleases |

I fragments (from the 3' end of the pre-tRNA) and type II (from either end of the mature tRNA). Both findings show fundamentally similar products.

tsRNAs are usually found as 13-22nt RNAs [Tuck and Tollervey, 2011]. tRF-3 and tRF-5 tsRNAs are generally not found in high abundance from the same tRNA, suggesting a functional model similar to the selection of single-stranded miRNA from its duplexes by Argonaute.

A final type of tRNA-derived sequences are tRNA halves, which are produced when the organism is under stress [Thompson and Parker, 2009]. As the name suggests, these fragments are produced by cleavage at the anticodon loop. The cleavage is caused by stress-activated nucleases: RNY1 in *Saccharomyces cerevisiae* (yeast), part of the RNAse T2 family, and angiogenin in mammals, part of the RNAse A family. tRNA halves have been shown to inhibit translation activity in eukaryotes [Zhang et al., 2009]. The function of tsRNAs is more unclear, but there is evidence to suggest that they compete with other sRNAs for Argonaute proteins [Haussecker et al., 2010].

### Y RNAs

Y RNAs are ncRNAs that that vary in length between 84 and 113nt with a secondary structure similar to that of a precursor miRNA [Nicolas et al., 2012]. Although their sequence conservation is very low, the Y RNA's secondary structure is conserved across different genes [Chen and Wolin, 2004]. Y RNAs are transcribed from their respective genes by RNA polymerase III and go on to form the Ro-Ribonucleoprotein (roRNP) complex with two proteins: La and Ro60. Functionally, Y RNAs have been shown to promote chromsomal DNA replication [Christov et al., 2006]. However this function is independent from forming the roRNP complex. Little is known about the function of the Ro-RNP complex itself, however, much like that of tRNA halves, the RNA component of the complex is specifically cleaved into small RNAs that are between 22-36nt in length when the cell is under stress [Rutjes et al., 1999]. Although this pathway appears to be highly similar to miRNA biogenesis, it has since been shown that Y5 RNAs at least are not dependent on Dicer and are not involved with the canonical RNAi

pathway [Nicolas et al., 2012].

## 2.5 Discussion

We have provided an overview of sRNA biology and a brief understanding of the variety of classes and functions that exist for silencing gene expression using other short RNA molecules. Such diversity of sRNAs, and the variation within classes across organisms, is a challenge for bioinformatic analysis of the sRNA content of genomes. Many methods and tools have been proposed and used for sRNA discovery and functional annotation of genomes. These are described in detail in the next chapter.

# Chapter 3

# Preprocessing and analysis sRNA-seq data

## 3.1 Summary

There are a large variety of ways to process and analyse sRNA data. However, the majority of these methods share similarities along one common pipeline. First raw FASTQ files are processed by removing adapter sequences, filtering reads that are unsuitable for analysis, and converting the clean set of reads to a more useful non-redundant format. Secondly, a variety of quality checks may be performed on the data to assess its fitness for further analysis, whether certain samples should be removed, or even if any samples are of appropriate quality. Based on these quality checks, normalisation techniques can be applied to the sequence counts in order to make the expression estimates comparable across samples. The sequences are then annotated using various external sources or prediction programs. Lastly a differential expression analysis may be performed to understand the various changes within the transcriptome that was sequenced across samples.

What follows is a detailed description of the characteristics of sRNA sequencing, the type of sequencing data that we are specifically interested in, and a review of the methods used at each stage in this generic pipeline. The review is restricted to Illumina sequencing data, a platform that has proved extremely popular for sRNA sequencing due to its ability to record sequences at a high resolution of abundance (termed "sequencing depth").

## 3.2 sRNA data sources

Datasets for bioinformatic analysis on sRNAs can be derived using a number of molecular biology techniques. These techniques aim to quantify a population snapshot of sRNAs within an organism's transcriptome. Widely used techniques include microarray analysis and RNA sequencing (RNA-seq) [Mortazavi et al., 2008]. Although microarray analysis was the preferred method of quantifying RNA transcriptomes up to 2008, it suffers from technical limitations that prevent it from accurately quantifying RNAs with lower expression estimates [Casneuf et al., 2007], and has recently been eclipsed by the capabilities of RNA-seq using high throughput sequencing (HTS) platforms.

Each HTS platform is most useful for specific applications, in part due to the trade-offs between sequence coverage (the number of reads a sequencing run can produce) and individual read length (table 3.1). sRNA analysis is specifically interested in transcriptomic sequences that are no more than 30 to 40 nucleotides in length. Due to the low read length requirements, Illumina sequencing is generally desired because this will give the most coverage.

The depth of a sequencing run is extremely important [Tarazona et al., 2011], and there is a balance to be struck between good coverage and restricting potential sources for noise such as loci with low well-represented with higher depth.

Table 3.1: A summary of the capabilities of current high-throughput sequencing technologies.

| Platform | Reads per run | Read length |
|---|---|---|
| Pacific Biosciences | 50,000 per cell | Up to 15,000 bp |
| Ion Torrent sequencing | 400 bp | up to 90 million |
| 454 | 1 Million | 700 bp |
| Illumina G2 | Up to 6 billion | 50-300 bp |

Downsides to HTS platforms include recently assessed inaccuracies in the reported abundances of sequences, particularly for Illumina [Hafner et al., 2011; Jayaprakash et al., 2011; Sorefan et al., 2012; Sun et al., 2011]. This is caused by particular sequences forming highly stable structures with the adapters that are ligated to the sequences during the sequencing process. The more stable sequences are more likely to be sequenced and end up being over-represented in relation to their actual expression in the transcriptome. Several modifications to the sequencing preparations have been proposed to mitigate this bias, including 'HD adapters' [Sorefan et al., 2012] that contain a randomized subsequence that alters the structure's stability. Nevertheless there is a trade-off between efficiency

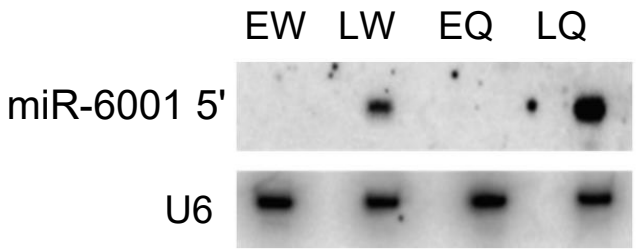



Figure 3.1: An example of a northern blot used to validate and quantify a specific mature miRNA transcript within four different treatments of the bumblebee *B. terrestris*.

of computational methods and the reliability of such methods.

Because of this decreased reliability, validation of bioinformatic results by other, usually low-throughput, means is neccessary. One such method is northern blot analysis, which is frequently used to validate RNA-seq and microarray experiments [Koscianska et al., 2011; Taniguchi et al., 2001]. The northern blot technique involves separating purified RNA using electrophoresis and then hybridizing it to the complementary strand of the sequence of interest on a membrane. A blot indicates the abundance of the sequence of interest in the form of dark bands on the membrane that vary in intensity, and the band's position indicates their length. An example blot is shown in figure 3.1. Due to their comparitively lower throughput, northern blots are best suited as a method to independently validate expression profiles of a handful of sequences found through computational analysis [Koscianska et al., 2011].

### 3.2.1 Replicates and experiments

There are two types of replicates. Technical replicates are carried out to adjust for differences in sequencing runs, whereas biological replicates adjust for differences in biological samples. When sequencing using Illumina machines, the technical replicates show a negligible difference and so are generally not needed [Marioni et al., 2008].

Biological replicates, however, are important in sequencing experiments because they represent independent experiments using the same treatment or conditions. This type of replicate can increase the statistical signficance of an analysis [McCormick et al., 2011]. Many differential expression techniques now require that there is more than one replicate and this requirement is becoming more achievable as sequencing costs fall [McCormick et al., 2011].

## 3.3 Preprocessing and quality checking of RNA expression data

In this section we describe the main stages commonly used in sRNA-seq projects. This starts with the initial input of FASTQ files provided by the sequencer and ends with a summary of the annotation and differential expression of sRNAs within the experiment.

### 3.3.1 Adapter removal

For Illumina data, sRNA sequences are usually presented for analysis in FASTQ files with the 5' adapter removed but with the trailing 3' adapter still attached. Finding and removing this adapter is crucial to processing an accurate sRNA sequence that will map back to the reference genome. This is commonly done by simply matching the first 6-8nt of the 3' adapter, trimming the adapter away, and returning the rest of the sequence for later processing. Sequences for which no adapter can be found are often simply discarded, since there is not enough information to discern the end of the sRNA sequence correctly.

If HD adapters are used (see section 3.2), the additional randomized nucleotides can be removed by trimming the final set of sequences by 4 nucleotides on either end.

### 3.3.2 Sequence filtering

After adapter removal, other undesirable characteristics of sequenced data can be filtered out in order to find a clean enough set of reads to map to the genome. Characteristics to filter on include:

- **Base calling errors** If any unknown bases (usually denoted as an 'N' nucleotide) are found in the processed sequences, it is recommended that the sequence is discarded to reduce ambiguity as long as sequencing depth is sufficient [McCormick et al., 2011].

- **Sequence length** Since the majority of sequences of interest are around 19-25nt long, it is generally recommended that sequences less than 18nt long should be discarded due to the higher chances of these mapping to erroneous places on the reference sequence [McCormick et al., 2011]. A maximum length threshold is often also imposed but this is dependent on whether longer ncRNAs are of interest.

### 3.3.3 Genome and annotation alignment

The preprocessed set of sequences are commonly annotated, first by aligning reads to the reference genome, if any, and secondly by using other annotation databases to further classify sequences as deriving from particular annotations.

There are many different alignment tools available, a review of which was conducted by [Fonseca et al., 2012]. For sRNAseq data, commonly used alignment tools include PatMaN [Prüfer et al., 2008], Bowtie [Langmead et al., 2009], and MicroRazerS [Emde et al., 2010]. These tools generally work efficiently to align reads but the speed and specificity can be affected by the addition of mismatches, gaps, and repetitive alignments. Due to the sRNA's small lengths, mismatches and gaps are not usually used for alignment because further ambiguity can greatly increase the chances of an incorrect alignment [McCormick et al., 2011]. Repetitive alignments, however, can be computationally expensive to process. Bowtie, however, allows the user to specify that only the first match of each read is returned, preventing the additional run time cost of mapping any further reads. This is useful for knowing if a read is genome-matching or not, but will obscure potentially important alignments of the read elsewhere on the genome. PatMaN, on the other hand, has no such parameter and may take a long time to map all repetitive sequences if there are a lot of them. A further issue is how to deal with repetitive sequences. Solutions range from discarding sequences that map more than once or a certain predefined threshold to more sophisticated "probability mapping" techniques, where it is assumed that reads are more likely to be derived from the loci that contain the most unique reads [McCormick et al., 2011].

Aside from reference genomes, a number of other sequences corresponding to annotations such as genes and ncRNAs may be used to assign reads to particular annotations. This is done either by aligning the reads to a further set of sequences or by finding overlapping alignments between reads and annotations aligned to a common reference, which can be accomplished using a tool such as BEDtools [Quinlan and Hall, 2010] or a suite of packages for Bioconductor: IRanges, GenomicRanges, and GenomicFeatures [Lawrence et al., 2013]. Annotating reads at this stage can aid in the quality assessment of the mapped libraries.

### 3.3.4 Quality measures and tools for RNA-seq data

Although RNA-seq is a very quick and convenient method of analysing RNA expression levels, it can often be prone to biases and errors that make a quality-

checking step an essential part of any analysis once the initial data has been retrieved from the sequencing machine. Quality of RNA-seq data can be affected by many different processes during both the preparation of samples in the lab and actual sequencing.

For very small multicellular organisms such as insects, extraction of enough RNA to use for RNA-seq is difficult due to contamination of surrounding organs, and the need to pool samples from multiple organisms together. This introduces various biases caused by contamination and variation of RNA expression across individuals, which can cause incomparable samples in the worst case [Amaral et al., 2014]. Systematic biases, such as ligation biases [Hansen et al., 2010; Sorefan et al., 2012], can also artificially affect the depth and resolution of gene expressions across an experiment.

Many standalone tools exist for assessing the initial quality of RNA-seq datasets. FASTQ files derived from the output of Illumina sequencers contain a line that encodes Phred quality scores [Ewing and Green, 1998]. The quality scores indicate the confidence of the nucleotide found in the same position of the sequence and can be read and used by assessment tools such as FASTQC [Andrews, 2010].

After preprocessing the dataset for low-quality base calls, errors, and other nucleotide-level biases, the sequence-level expression of datasets can be compared between libraries. Examining the correlation of abundances between libraries is often an effective way of examining the consistency of samples. Amaral et al. [2014] effectively used this method to reveal samples that were heavily effected by biases in Drosophila RNA-seq libraries.

## 3.4 Methods of normalisation

The goal of normalisation is to minimize the variability in sequence abundances between samples that arises from technical accuracy of the sequencing method and other noise in the data. The simplest type of variation is a systematic scaling of all counts with respect to another sequencing run.

Many methods and types of normalisation exist for all types of data from microarray to RNA-seq. There is no one method that suits all situations, and each method tends to have its own advantages and disadvantages. These methods can be conveniently split up into several categories that describe how they work. The following is a brief description of some of the most commonly used normalisation methods used and reviewed for RNA-seq datasets.

### 3.4.1 Scaling methods

Scaling normalisation adjusts the counts of all sequences by a global factor for each sample. Its main assumption is that sRNA expression is proportional to total library size, and the main differences that need to be normalised is the scaling of each sample distribution [Smyth et al., 2003].

#### Total count

A simple form of scaling by library size is to divide each sequence count by the sum of all counts in the library. We can find the normalised value of an expression value $x_{gk}$ of sequence $g$ in library $k$ (and where $G$ is the total number of sequences in the experiment) by calculating [Robinson and Oshlack, 2010]

$$N(x_{gk}) = \frac{x_{gk}}{\sum_{g=1}^{G} x_{gk}} \times C. \tag{3.1}$$

Here, the normalised value is multiplied by a constant $C$, which is an upscaling constant relative to the magnitude of sRNA data to prevent the counts from being very small fractions. Generally, this can be set to the mean of all library sizes.

This makes sense under the assumption that libraries are often sequenced at varying depths, causing all reads to be sequenced more by a set factor compared to another library. However, in practice, other biases between libraries, especially the biological differences, cause the relationship between two library distributions to be non-linear.

One particular issue arises when a group of reads are only highly expressed in some of the samples being normalised. If there are a large number of novel sequences present in one sample, scaling by a statistic based on the total count will result in under-represented sequences in the sample with extra reads [Robinson and Oshlack, 2010]. This is often the case when comparing across samples from an experiment in which Dicer or Argonaute has been knocked-out, where a whole group of sRNAs may not be expressed in the knock-out sample.

When used in mRNA-seq data, this method may also include a factor to weight the counts by gene length to account for biases in the number of sequences that longer genes can acquire [Mortazavi et al., 2008]. For sRNA-seq data, single reads are generally not aggregated into much longer genes and so this additional weighting is not needed.

### Percentile count

Due to the nature of RNA-seq datasets, using the total count for each library as a linear normalisation factor tends to reflect the expression of just a few high count genes. In their analysis of mRNA data, Bullard et al. [2010] noted that 5% of genes made up 50% of the total counts. This effectively scales the libraries towards the differences of these highly abundant sequences between samples.

One strategy to prevent such a bias is to use sum of counts that are at or under a percentile of the count distribution in one library. By normalising by total counts up to the median quartile, the assumption is that the scaling factor will capture the population of gene counts that should be a more steady state. In practice, the median is generally found to be very low due to the nature of the distribution and the upper quartile is used instead [Bullard et al., 2010].

### Trimmed mean of means

Trimmed mean of means (TMM) is a scaling normalisation method by Robinson and Oshlack [2010] that aims to improve on total count normalisation by applying a correction factor to library sizes used to normalise the samples. The method requires a reference sample to be chosen. Then, a correction factor is obtained from the weighted mean of the log ratios between sequence abundances in each observed sample and the reference sample after excluding differentially expressed genes. The filtering step ensures that the factors obtained are based on the core set of genes that are not differentially expressed and so should be generally equal between the two samples.

Robinson and Oshlack [2010] recommend to use scaling factors found by TMM as factors within the statistical tests for differential expression rather than modifying the original count data. This has the advantage of preserving the original nature of the data as counts, which certain statistical tests such as Fischer's exact, binomial, Poisson and $\chi^2$ tests rely on [Zhou et al., 2013].

An extension to this type of normalisation is *TMM-baySeq-TMM* (TbT) [Kadota et al., 2012]. This uses TMM in a 3-step process where an empirical Bayesian method, baySeq [Hardcastle and Kelly, 2010], is also used to determine differentially expressed genes after an initial normalisation. Differentially expressed genes are then removed before a final normalisation to make sure that only non-differentially expressed genes are being normalised.

TMM is implemented in the edgeR package [Robinson et al., 2010] by using the function `calcNormFactors()`.

#### DESeq normalisation

DESeq is a method used to find differentially expressed sequences that focusses on the assumption that the majority of the population of genes are not differentially expressed [Anders and Huber, 2010]. The method uses size factors $F_j$ for each library $j$ to adjust its counts so that they are comparable. The size factors are obtained by finding the the ratio between each gene count in a library and the same gene in a pseudoreference sample, which in turn is found by taking the geometric mean across samples for each gene. The size factor for a library, termed $F_k$ here for $K$ number of libraries, is then the median of the resulting ratios:

$$F_k = median(\frac{x_{gk}}{(\prod_{v=1}^{K} s_v)^{1/K}}). \tag{3.2}$$

If most genes are not differentially expressed, this statistic should be identifying a median from a series of fold changes that are mostly not differentially expressed, leaving a normalised fold change that represents the amount of scaling needed for each library.

DESeq normalisation is available in the R packages DESeq [Anders and Huber, 2010] and DESeq2 [Love et al., 2014] by using the function `estimateSizeFactors()`.

#### Spike-in sequences

Spike-in normalisation relies on the accuracy of spiking samples at the wet lab stage with a synthesized RNA that does not normally appear in the evaluated RNA samples. When added to each sample at the same concentration, the abundance of each sequence in the resulting RNA-seq data can be assessed and used as a scaling factor for the rest of the libraries. The assumption is that all other RNA quantities are scaled in the same way that the synthetic RNA is. The efficacy of such a method was assessed in Fahlgren et al. [2009]. A significant disadvantage, however, is the need to plan, prepare and add the spike-in sequences during the biological preparation of the samples.

### 3.4.2 Quantile normalisation

Quantile normalisation is a method used in microarray data to make the distributions of the probe intensities the same across samples [Irizarry et al., 2003]. It assumes that the overall distributions of gene probe intensities remains the same across samples and that differentially expressed genes are simply found at a different rank in compared samples, displacing the expression of other genes.

To normalise by quantile normalisation, sequences in each library are separately ordered from the most abundant to the least abundant. Then, each sequence in each library is assigned a summary - usually the mean - of the counts of all sequences at the same rank in all compared libraries.

The method has been adopted for RNA-seq data. [Bullard et al., 2010] used a variant of quantile normalisation in their assessment of normalisation procedures for mRNA-seq data. In this variant, normalised counts for each sorted row are calculated from the median rather than the mean, and the final counts are rounded to integer values to preserve the nature of the count data.

Because quantile normalisation was originally developed for probe intensity data, it can cause some unwanted effects on count data. Firstly, a gene may end up with a normalised expression value in a sample where it was never found originally. Secondly, ties are much more likely to occur in count data and must be dealt with correctly so that sequences that were found at the same rank in a sample do not end up with varying expression levels after normalisation.

### 3.4.3 Evaluation of normalisations on sRNA-seq data

Although there have been a large number of review papers detailing differences in known normalisation techniques for RNA-seq datasets, only a handful of these have focussed on normalisation strategies for sRNA-seq datasets specifically [Dillies et al., 2013; Garmire and Subramaniam, 2012; McCormick et al., 2011]. It is always important to evaluate the results of normalisation methods during the analysis of specific datasets, and these reviews indicate potential ways to do this.

Garmire and Subramaniam [2012] used mean square error (MSE) and Kolmogorov-Smirnov (K-S) statistics as data-driven measures to assess differences between 7 normalisation methods, including TMM, quantile, and scaling. They noted that both these statistics measured differences in the distribution of abundances between samples, and as such would generally favour quantile normalisation, which also aims to equalise the abundance distributions. To this end, they included correlations against QPCR runs as a different dimension of measurement. They found quantile normalisation to be superior to the other methods for small numbers of unique sRNAs. TMM behaved "abnormally", although this was criticised by Zhou et al. [2013] as being attributable to an issue with implementing the method.

A second area of criticism was targeted at the application of K-S and MSE statistics between different conditions. This gives no consideration to true differential expression between samples, which would adversely affect the statistics

[Zhou et al., 2013]. A more suitable metric is the coefficient of variation between replicates, which is used in Dillies et al. [2013] on miRNA-seq datasets to assess 7 normalisations including total count, upper quartile, median, DESeq, and quantile. They found that TMM and DESeq performed equally the best, with little or no difference between their use on mRNA-seq data or miRNA-seq data. Conversely, Quantile normalisation was found to particularly increase intra-replicate variance in miRNA-seq data, which should be avoided.

Other metrics for evaluating normalisation methods include comparisons of the normalisation outcome to a different type of expression data, such as qRT-PCR or QPCR used in Anders and Huber [2010]; Bullard et al. [2010]; Garmire and Subramaniam [2012]; Sun and Zhu [2012]; and Rapaport et al. [2013]. The largest barrier to using this method is that it requires a second, usually costly, experiment to be carried out and analysed on the same samples using different equipment. This may be fine when the goal is solely to evaluate general properties of normalisation methods on RNA-seq datasets, but is not feasible to evaluate normalisation results during specific experiments.

Importantly, all reviews cited here attempt to assess normalisation methods on either miRNA-seq data or mRNA-seq data, but there are no known analyses of how these methods perform on a total sRNA-seq dataset. If these methods perform differently between miRNA-seq data and mRNA-seq data, where the number of unique sequences differ somewhat, then they are likely to differ again when using sRNA-seq datasets (see Chapter 4). In addition, Bullard et al. [2010] note that the biggest difference between the methods they tested was the ability to handle low counts and zero-count data; a property that sRNA-seq datasets have much more of due to its increased sequenced diversity and sparsity of counts.

Simulated datasets were used in Kadota et al. [2012]; Reeb and Steibel [2013]; Robinson and Oshlack [2010]; and Dillies et al. [2013]. This method of evaluating normalisations provides a way of identifying the sensitivity and specificity of each method without the need for further biological intervention. However, the nature of simulated data means that they are likely to only show that a normalisation can work under ideal circumstances.

## 3.5 Methods for assessing differential expression

The past half of the decade has seen a sharp increase in the number of different tools and packages that can be used to fully analyse an Illumina RNA-seq dataset from beginning to end. Many of these packages maintain a heavy focus on RNA-

seq data, assuming that sRNA-seq data is then a simplified special case. This section aims to review some of these tools in terms of their ability to analyse specifically sRNA-seq datasets, taking care to highlight their differences.

The use of the R statistical package [Ihaka and Gentleman, 1996] with Bioconductor [Gentleman et al., 2004] remains a popular choice for analysing RNA-seq data. The packages available to use include EdgeR [Robinson and Smyth, 2007], and DESeq/DESeq2 [Anders and Huber, 2010; Love et al., 2014]. Both of these packages start by modelling the distribution of count data as a negative binomial distribution with a mean $\mu$ to variance $v$ relationship of $v = \mu + \alpha\mu^2$. $\alpha$, in this instance, is a dispersion factor that can be used to model the overdispersion that RNA-seq datasets contain [Rapaport et al., 2013]. The way this is done differs between edgeR and DESeq, but both tools effectively seek to share information on the dispersion of the data between genes to account for the likely low numbers of replicate samples. edgeR first estimates a common dispersion for genes that is related to the mean of replicate counts. It then employs an empirical Bayes strategy that squeezes per-gene dispersions towards these common dispersions. DESeq, on the other hand, uses the direct relationship between the variance and the mean given above, but computed along with a library size factor found by DESeq's normalisation method (see section 3.4). Both tools then use these dispersions in an exact test to find significantly differentially expressed genes. The general effect is that the noise found at low abundances tends to push the dispersion estimates up in genes with lower counts, resulting in a smaller proportion of genes being called differentially expressed at low abundances. DEseq2 seeks to further improve on the ranking of differentially expressed genes by fold changes. It does so by first finding a dispersion estimate for each gene using a maximum liklihood estimation (MLE) approach on the gene's replicate abundances [Love et al., 2014]. GLM regression is then used to fit a curve that regresses the estimators on to the normalised abundances. The curve provides new dispersion estimates that are dependent on the average count, and the per-gene estimates are then shrunk towards these fitted values using an empirical Bayes approach. Because the dispersion estimates now depend on the average abundance, dispersions for low count genes are likely to be higher, which will reduce the resulting fold changes. However, Seyednasrollah et al. [2015] found that these changes appeared to increase the number of incorrect differentially expressed gene calls (false positives) along with the number of total genes found significantly differentially expressed, rather than further controlling for them. In addition, it remains to be seen how well any of these tools work on full sRNA-seq datasets.

## 3.6 Computational tools and methods for discovering sRNAs

The use of high-throughput sequencing approaches has aided the increase in the number of sRNAs and the quantity of sRNA data available for analysis. In reaction to these discoveries, the number and quality of methods and tools dedicated to identifying sRNAs and their functions has also increased. This section is dedicated to a summary of the approaches used to identify sRNAs from high-throughput sequencing data and a more in-depth comparison of tools that are important to this project.

### 3.6.1 A summary of approaches for sRNA discovery

Computational identification of sRNAs generally relies on a set of characteristics that sRNAs exhibit within RNA-seq datasets. What follows is a summary of characteristics that are commonly analysed in order to pick out the most interesting loci from a sRNA-seq dataset. Many of these characteristics are inherited from characteristics that all RNAs exhibit within datasets and can be used to annotate many other types of RNAs, both coding and non-coding, as well.

**Size class distributions.** Because classes of sRNAs are generally of specific lengths, highly expressed sRNAs will contribute their size to the overall size class distribution of a dataset. Simply viewing the overall distribution of a dataset can give an estimate of the type of sRNA populations that are expressed within it, and this is often the first type of analysis that is carried out. In addition to this, size class distributions for localised regions of a genome can be a good indicator of sRNAs that have a lower expression.

**Mapping characteristics.** sRNAs typically produce "stacked" or "blocky" patterns when mapped to a genome, which distinguish themselves from longer RNA degradation [Cole et al., 2009; Langenberger et al., 2010] (figure 3.2). The patterns of the blocks will vary in length and proximity depending on the type of sRNA that produces them. For example, miRNAs have a characteristic dual block pattern of two 21-23nt blocks where one block (the mature sequence) is typically more expressed than the other block and tasiRNAs will produce phased blocks that are adjacent to each other.

**Read Count Complexity** is related to the mapping patterns, where the complexity of read counts is the non-redundant count divided by the redundant count for a population of sRNAs [Xu et al., 2014]. A read population that is rich in sRNAs usually has a lower complexity when compared to a population that is

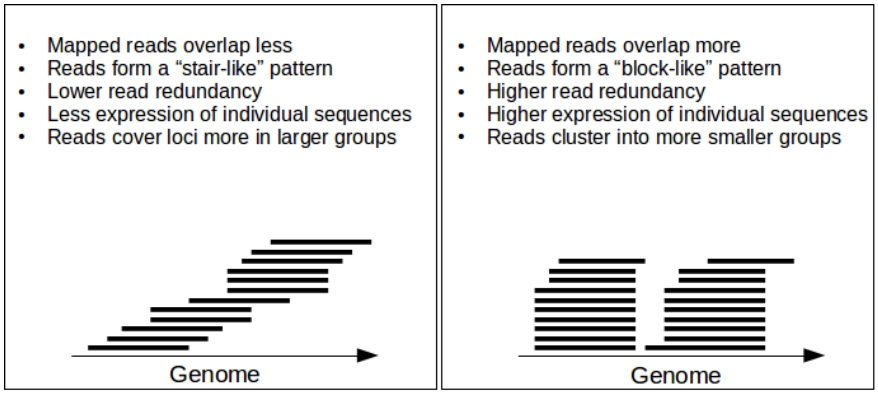



Figure 3.2: A schematic visualising mapping patterns of sequenced reads that are characteristic of mRNA degradation products, an example of noise, (left) and a hypothetical sRNA loci (right).

rich in degradation products.

**Differential expression** of reads between multiple samples provide a strong indicator that sRNAs exist that are expressed under certain conditions. The analysis of datasets produced under differing biological conditions is a robust method of identifying specialised sRNAs and their functions. This is discussed in much greater detail in the previous section, 3.3.

sRNAs such as miRNAs are produced from genomic loci that can fold into tight **localised secondary structures** once transcribed. The use of secondary structure prediction tools such as RNAfold [Hofacker, 2003] and Mfold [Zuker, 2003] can provide additional evidence for sRNA loci. Sequences are folded using a dynamic programming algorithm [Eddy, 2004] that efficiently computes the energy required to base pair combinations of nucleotides in the sequence. The candidate structure is the structure with the minimum free energy (MfE) out of all possible combinations. This is the secondary structure of the sequence that is most stable within the cell. Because these programs only consider the sequence as determining the structure, the influence of external factors on the secondary structure can not be ruled out, and often the most likely structure is one with a suboptimal free energy [Gardner and Giegerich, 2004]. miRNAs are identified this way in particular since their precursor hairpins are distinctive from other RNA secondary structures across organisms [Loong and Mishra, 2007]. However, the precision can be improved by the use of multiple alignments, using the idea that bases that pair together are more likely to be conserved [Washietl et al., 2005]. Lastly, significance of the MfE of a particular secondary structure can be computed by determining its $Z$-score when comparing it to the MfE of random sequences of the same length that conserve the di-nucleotide frequency of the

original [Clote et al., 2005].

**Sequence homology**. Identifying known sRNAs within sRNA libraries is commonly accomplished by matching the short reads with sRNAs that have already been identified in the target, or related, organisms that are stored in a public database such as miRBase [Griffiths-Jones et al., 2006].

**Source transcript**. A sRNAs precursor transcript can be derived from a multitude of other coding and non-coding RNAs, as described previously. Successfully identifying the precursor can help to categorise the sRNA.

**Sequence Motifs**. A number of classes of RNAs show a clear bias towards particular subsequences. For example, miRNAs tend to have an affinity for a U nucleotide at the 5' end of their sequences, and piRNAs tend to have both a U at their 5' end and the complimentary nucleotide at the 10th position [Malone and Hannon, 2009]. Sequence motifs are commonly visualised using graphics such as sequence logos [Schneider and Stephens, 1990]. The frequency of a nucleotide at each position in a set of alignments is represented by the height of a nucleotide's letter, where the total height of the stack of nucleotides at each position usually indicates the information entropy of that position. Tall nucleotides can indicate a consensus at that position. An alternative to this plot can be found in Berry et al. [2006]. Instead of information content, the y axis represents the log relative frequency of a nucleotide with respect to the background frequency, for which the background GC content is adjustable. In chapter 6 these graphs are referred to as "Berry Logos".

Table 3.2: A summary of tools developed to identify and characterise sRNAs from sRNA transcriptome data

| Tool | Description | References |
|---|---|---|
| RNAz | Prediction of ncRNA secondary structures | Washietl et al. [2005] |
| miRCat | miRNA prediction | Moxon et al. [2008]; Stocks et al. [2012] |
| miRDeep | miRNA prediction | Friedlnder et al. [2012] |
| MIReNA | miRNA prediction | Mathelier and Carbone [2010] |
| MapMi | miRNA homolog identification | Guerra-Assuno and Enright [2010] |
| BlockBuster | ncRNA loci detection | Langenberger et al. [2009] |
| DeepBlockAlign | Pattern similarity of ncRNA loci | Langenberger et al. [2012] |
| NiBLS | sRNA loci identification | MacLean et al. [2010] |
| SiLoCo | sRNA loci identification | Moxon et al. [2008] |
| CoLIde | sRNA loci identification | Mohorianu [2012] |

### 3.6.2 Methods for the identification of sRNA-producing loci

As HTS technologies have improved to allow more sensitive and deeper coverage of transcriptome data-sets, our understanding of the pathways of RNA regulations has broadened. Functional sRNAs can be found deriving from many different areas of the genome, including other RNA transcripts [Tuck and Tollervey, 2011]. These sRNAs can be identified by the patterns that the sequenced data makes when mapped to the genome - the sRNAs are usually of a specific length and spliced from precursor transcripts at the same sites. To assist in the identification of novel sRNAs, tools have been developed that do not require any other prior knowledge of the characteristics of particular sRNAs. These "general" ncRNA discovery tools allow researchers to hone in on loci of interest and focus on highly expressed sequences that may have new functions.

The clustering of sRNAs can be applied on two levels: clustering of individual reads into sRNAs, and the clustering of sRNAs into sRNA loci.

#### Local clustering

A defining characteristic of sRNA reads is their specific length and position when compared to varying lengths and positions of degradation. This is an artefact of the specific processing that sRNAs undergo during their biogenesis. Simple metrics, such as the ratio between the number of reads mapped to a loci and the length of the loci used by Cole et al. [2009], indicate that processing patterns can differ significantly between types of ncRNA. However, sRNA loci often contain a large amount of alternative reads that show a different processing pattern but with lower expressions. The sum of these expressions are often important in determining the overall expression of a particular sRNA loci and it may be misleading to use, for example, the abundance level of the most expressed read. It is therefore advantageous to group highly overlapping reads as derived from one particular sRNA. This problem is non-trivial, but several tools exist, described in later sections, that alleviate the issue.

#### sRNA loci generation

The second level of clustering is premised by the observation that sRNAs are usually found in clusters of similarly functioning sRNAs on the genome. An example of this is the mature and star arms of a miRNA, where both arms may contain sRNAs that are expressed at high levels. Clustering reads on this level can help

reveal sRNA hotspots on the genome that require further examination. However, clustering loci inherits the issue of when to stop clustering. Overclustering can cause the loci to become too fragmented, whereas the opposite approach can lead to clusters that are too large and meaningless to the analysis.

The following sections summarise tools that attempt to identify sRNAs and other ncRNAs by identification of ncRNA loci.

### SiLoco

SiLoco [Moxon et al., 2008] defines sRNA loci by grouping reads that are closer than a maximum allowed gap when aligned to the genome. Statistics, such as the normalised abundance of these loci, weighted by the individual read's repetitiveness, are then calculated and presented in a table.

### Blockbuster - identification of ncRNA blocks

Blockbuster [Langenberger et al., 2009] groups a set of mapped genomic reads into first blocks and then clusters of blocks based on the amount of similarity in their mapped loci. Reads are modelled as Gaussian densities using their start and stop positions and a standard deviation that is weighted by a tunable parameter. ncRNA loci are composed of reads that are not separated by more than 39nt. An iterative greedy algorithm is then used to group blocks starting with the location of the highest density peak in a locus. The block's expression is then the sum of the grouped Gaussian peaks, allowing the expression of a particular locale on the genome to be composed of the individual read expressions.

The blocks are converted to smoothed curves with areas that equal the number of reads in the block. The height of the peak is then a value affected by both the expression level of the reads in the block and its "coherence". Higher peaks therefore represent highly expressed reads that have more specific processing patterns, allowing them to be computationally categorised apart from degradation blocks.

In addition to simplifying loci identification, Langenberger et al. [2010] use BlockBuster to attempt to classify ncRNA based on the block patterns that their sRNAs leave on the genome. BlockBuster's output is used to quantify characteristics such as the number of blocks in a sRNA cluster, the length of a block, the block overlap, and the block height in order to train a random forest model with some success. The idea shows that there is a great deal of difference between the expression patterns of ncRNAs such as miRNAs, tRNAs, and snoRNAs.

#### NiBLS

NiBLS [MacLean et al., 2010] groups reads into sRNA loci by finding reads grouped by close proximity that together give a threshold clustering coefficient. The set of reads is first modelled as vertices on a graph $G$ where edges connect two reads if they are on the same chromosome and the difference between the start of one and the end of another is less than a parameter $M$. A loci is then created if the clustering coefficient $\gamma$ is larger than a second parameter $C$. This effectively models how 'spread out' a group of reads are. Stacked reads indicative of sRNAs will produce a higher clustering coefficient than less overlapping reads which are more indicative of degradation products.

#### CoLide

CoLide [Mohorianu et al., 2013] defines its sRNA loci based on the similarity in variation in the differential expression and sizes of neighbouring reads. The algorithm defines each sRNA as being upregulated, downregulated or not regulated using an offset fold change analysis on confidence intervals created over replicate expression levels. Reads are then merged into a loci if they are close enough to each other and their expression patterns are the same.

### 3.6.3 miRNA identification tools

miRNAs have been a prime candidate for the use of automatic identification tools because they can be relatively well defined in terms of their structure and expression pattern on the genome. This is in contrast to other sRNAs, such as piRNAs, that do not fold and can be harder to pick out above the noise. This section compares and describes tools that take different approaches to identifying miRNAs.

MapMi [Guerra-Assuno and Enright, 2010] focuses on validating the conservation of known miRNAs in novel genomes. It achieves this by aligning mature miRNA sequences to the supplied genome and then re-assessing the secondary structure of potential precursors by extending and folding the resulting alignments and assigning a score based on the alignment quality and precursor structure. Results are then filtered by a predefined score threshold and output back to the user.

Both miRCat [Stocks et al., 2012] and miRDeep2 [Friedlnder et al., 2012] attempt to predict new miRNAs using the genome mapping patterns and abundance levels of sRNA-seq data. miRCat first clusters sRNA reads using a prox-

imity based on a maximum gap proximity method. Clusters that pass a set of characteristics, such as minimum number of reads and non-overlapping reads, are folded with flanking regions and base pairing is assessed against a set of base-pairing rules. The minimum free energy of the structure is also calculated and checked against a randomly shuffled version of the sequence to obtain a $p$-value. Clusters that pass all rules are written to file as potential miRNAs.

miRDeep2 attempts to compare a set of aligned reads to the model of Dicer processing. Generally speaking, this model defines a miRNA as distinct "read stacks" in close proximity that can indicate the presence of the abundant mature sequence, the less abundant star sequence, and a left-over loop region that can be viewed as background degradation on the precursor transcript. In this way, the aim is similar to how algorithms such as BlockBuster attempt to identify ncRNAs. miRDeep2 selects miRNA loci by searching for highly abundant 20-24nt reads. Potential precursors are selected using these reads as guides. A Bayesian algorithm is used to score the final list of possible miRNAs indicating how likely they are to be a true miRNA based on prior known miRNAs. Prior probabilities are estimated from animal data, making miRDeep best suited for datasets within this lineage.

An important part of miRNA prediction programs is their application to a wide range of organisms. This is made particularly difficult by the stark differences of miRNAs found in animals versus plants. As a result, prediction tools are often developed with the characteristics of miRNAs from one lineage to start with. To extend the use of miRCat to animal lineages, a second set of parameters were estimated based on the characteristics of animal miRNAs. In the case of miRDeep, Yang and Li [2011] extended the miRDeep algorithm in a new tool, miRDeep-P, which altered some parameters, for example the size of the excised potential precursors, and also re-estimated prior probabilities for the scoring algorithm. miRCat has sets of alternative default parameters for both animal and plant organisms 3.3. This shows the amount of dependency that these methods have on previously identified miRNAs, and identification of miRNAs within unrelated species is still a challenge when there are no miRNAs to calibrate the tools with.

## 3.7 Conclusions

We have presented and discussed the main methods for analysing sRNA-seq datasets for each stage of an analysis. Such methods have been adapted from

Table 3.3: A comparison of miRCat parameters for plant and animal data. The main differences are in the size of the hairpin, where plant hairpins are known to be significantly longer on average.

| Parameter | Plant | Animal | Description |
|---|---|---|---|
| extend | 100 | 40 | nucleotides either side of mature to make fold |
| min_hairpin_len | 60 | 55 | minimum length of final hairpin |
| max_unpaired | 50 | 60 | maximum percentage unpaired nucleotides in hairpin |
| min_paired | 17 | 17 | minimum number of paired bases of mature |
| max_gaps | 3 | 3 | maximum number of consecutive unpaired bases in mature |
| max_genome_hits | 16 | 10 | maximum number of hits that a read can have to be considered |
| min_hits | 2 | 2 | |
| min_length | 20 | 21 | minimum length of mature miRNA |
| max_length | 23 | 23 | maximum length of mature miRNA |
| max_overlap_percentage | 80 | 80 | |

those used on the related mRNA-seq datasets. However, it is clear that sRNA-seq data contains its own unique chatacteristics and properties, such as distinct size classes and and a sparser count matrix, that should not be overlooked when adapting these methods, especially when assessing the data's quality and normalising for systematic biases. In the next chapter, we extend the methods used in each stage of a standard differential expression pipeline (preprocessing, quality checking, normalisation, and differential expression) to further take into account these unique properties. In doing so, we create a sRNA-specific processing pipeline, with an emphasis on quality checking and correct normalisation, to be used in subsequent chapters.

# Chapter 4

# An interactive pipeline for the analysis of high-throughput small RNA sequence data

This chapter is adapted from Beckers M, Mohorianu I, Stocks M, Applegate A, Dalmay T, Moulton V, "*An interactive pipeline for quality checking, normalisation, and differential expression analysis of high-throughput small RNA sequence data*", in preparation.

## 4.1 Summary

There currently exists a myriad of tools and software pipelines available to process, analyse and test the differential expression of RNA-seq datasets. However, few of these are properly tailored towards sRNA-seq datasets. In addition, the increase in sequencing depth and decrease in cost per sequencing run has produced larger datasets that are more memory intensive to run, usually requiring specialist hardware such as dedicated servers.

In this chapter, we present a processing pipeline that includes steps for quality checking, normalisation, and differential expression that are all correctly tailored towards the analysis of sRNA-seq datasets. The pipeline was developed over the course of completing several sRNA-seq differential expression experiments, the results of which are presented in later chapters. As such, this chapter serves as a detailed explanation of methods that will be used later on in this thesis.

We also describe an implementation of this pipeline, as part of the software package "The UEA sRNA Workbench", that was designed to utilise hard drive resources more than RAM resources so that it can be used to process large datasets

on non-specialist personal computing devices.

## 4.2 Background

The sequencing of sRNA technologies such as Illumina follows a similar protocol to that of mRNA-seq (see chapter 3). Numerous tools have been developed to handle large mRNA-seq datasets, many of which can also be used for sRNA-seq data, especially at the preprocessing step. These tools are commonly presented as a pipeline of subsequent operations on the data. First, sequenced reads are presented in FASTQ-formatted files. The 3' adapters are trimmed by matching the first 7-8nt of the adapter sequence using tools such as FASTX [Gordon and Hannon, 2010], or cutadapt [Martin, 2011]. The trimmed reads are then mapped to a reference genome and corresponding annotations using one of a variety of mapping tools [Fonseca et al., 2012]. The mapped sequences are subject to normalisation of their abundances across the samples and, finally, differentially expressed reads are identified using one of several differential expression approaches [Rapaport et al., 2013; Seyednasrollah et al., 2015].

An extremely important aspect to all steps of a differential expression pipeline is assessing the quality of the processed data in order to maintain the accuracy and efficacy of downstream bioinformatics analyses [Watson, 2014]. However, at many stages of the pipeline, the quality assessments are often overlooked, leading to potentially misleading results. For RNA-seq data, quality checking tends to focus on two aspects: (1) the quality and composition of sequences from the raw FASTQ files and (2) the quality of the comparison between libraries based on the processed sequence abundances.

Bioinformatics methods developed for differential expression analysis of RNA-seq data have thus far largely focused on analysing mRNA datasets. However, the difference in the way in which mRNA-seq and sRNA-seq datasets are processed after alignment causes a difference in the properties of the resulting datasets. For sequence data output by Illumina machines, the final expression values of mRNAs are found by summing together all reads that map to a particular mRNA because these are much longer than the resulting reads produced by the sequencing machines [Mortazavi et al., 2008]. sRNAs, however, are shorter than the reads produced, which are simply tabulated in to unique reads with an associated abundance in order to represent each sRNA. As a result, sRNA datasets contain many more unique entries than mRNA datasets but the mean and median of the expression levels are within the range of the dataset's noise (figure 4.1 (a)) and the

resulting count matrix is more sparse. The differing properties of the two types of data can have repercussions when attempting to use established mRNA-seq methods on sRNA-seq data, especially when attempting to fit a standard distribution. For sRNA-seq data, quality checks should also additionally focus on the specific characteristics of sRNAs, but these are often overlooked. For example, in a review of sRNA sequencing experiments [McCormick et al., 2011], the quality control discussion was limited to handling the quality of base calls and assessing size distribution graphs.

Differences in the two datasets can be mitigated by extracting only the known sRNAs of interest - usually miRNAs - out of a sRNA-seq dataset. We have termed this type of dataset as "miRNA-seq" and it is commonly done where the interest is solely on the abundance of known miRNAs (see Camps et al. [2014] for an example of such an analysis). This reduces the diversity of the data points and brings the count distribution more in line with mRNA-seq data (figure 4.1 (a) and (b)). However, this limits the analysis to a single class of known sRNAs. A more informative analysis can be done on the entirety of the transcriptome and may also incorporate further prediction tools to identify new types of sRNAs.

A further issue that becomes more apparent in sRNA-seq data is the significance of low abundance fold changes when assessing the differential expression of sequences. In RNA-seq data, low transcript abundance appears to be correlated with a lack of preference to aligning to genes rather than intergenic regions Ramskld et al. [2009], suggesting a detectable background level of noise. In sRNA-seq data, this background level is likely where the vast majority of transcripts are expressed, resulting in the vast majority of large fold changes found at the level of background abundance. Differential expression tools such as DESeq2 [Love et al., 2014] and edgeR [Robinson et al., 2010] deal with this issue by estimating the relative dispersion of sequences and incorporating this into later statistical tests for differential expression significance. An alternative more stringent solution is to apply an offset directly to fold change estimations that directly downweights fold changes from abdundance levels that are near the level of the offset [Mohorianu et al., 2011].

In addition to the technical aspects of differential expression analysis on sRNA data, the logistics of tying together multiple tools into one analysis can make processing RNA-seq data more complicated. A common solution is therefore to group tools into a software pipeline to allow the end user to more easily run a complete analysis. After the setup is complete, the likely lengthy procedure can be executed and left to run without the need for much further input from the

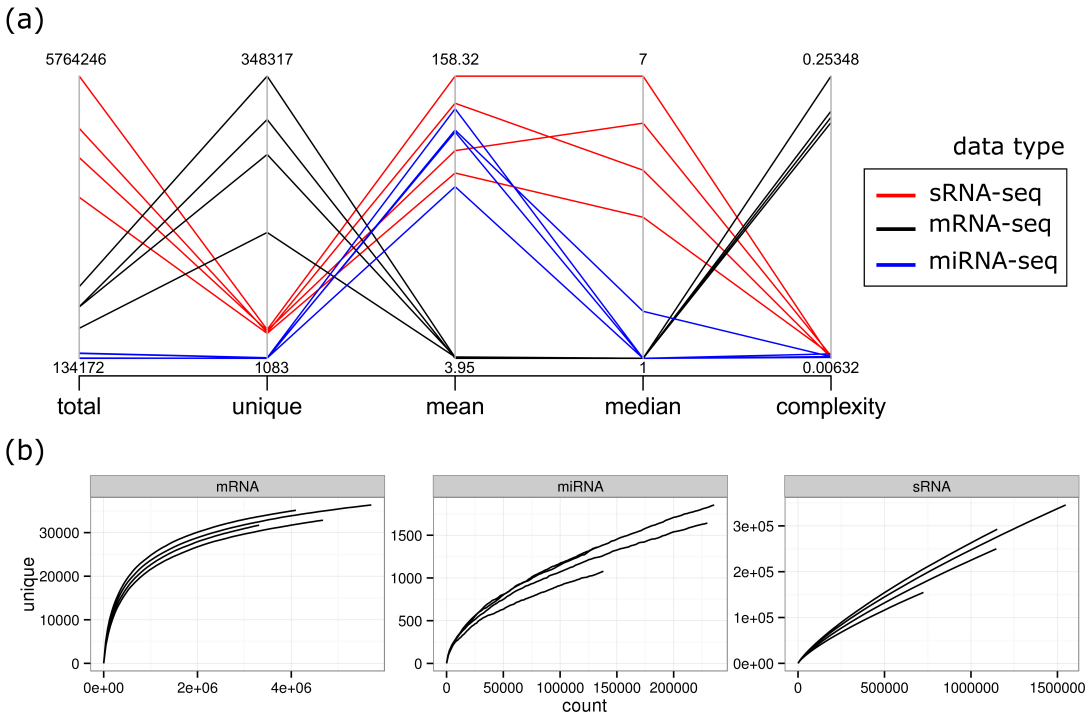



Figure 4.1: The statistical differences between preprocessed mRNA-seq, miRNA-seq, and sRNA-seq datasets. (a) Parallel coordinates for several statistical summaries of datasets from the same experiment (b) Rarefaction curves indicating the number of unique sequences found when the data is resampled to certain depths.

user. Currently there are several mRNA-seq tools available that can be configured to handle to some extent the various stages of a sRNA differential expression pipeline (see table 4.1). However, none of these cover the entire analysis of a sRNA dataset.

In this chapter, we present a processing pipeline specifically for the analysis of sRNA-seq datasets. The pipeline includes novel approaches to quality checking, normalisation, and differential expression analysis for use on a set of complete transcriptome samples derived from a sRNA-seq dataset. Rather than annotating and extracting just the known miRNAs and other sRNAs to analyse, processing the complete dataset has the advantage of identifying further novel differentially expressed sRNAs when combined with current sRNA sequence prediction programs.

Table 4.1: A summary of current RNA-seq and sRNA-seq packages and tools available. Most columns indicate whether a certain feature is available (Y) or not (N).

| Tool | Format | Data type | FASTQ quality | Nucleotide freq | Adapter trimming | nucleotide complexity | abundance filter | size filter | replicate filter | Size class | Annotation | MA/scatter | Normalisation | DE | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DEseq | R package | RNA-seq | N | N | N | - | - | - | - | N | N | Y | DEseq2 | Y | Love et al. [2014] |
| EdgeR | R package | RNA-seq | N | N | N | - | - | - | - | N | N | Y | multiple | Y | Robinson et al. [2010] |
| baySeq | R package | RNA-seq | N | N | N | - | - | - | - | N | N | N | quantile | Y | Hardcastle and Kelly [2010] |
| RSEQtools | Software | mRNA-seq | N | N | N | - | - | - | - | N | Y | N | RPKM | Y | Habegger et al. [2011] |
| DARIO | Web | ncRNA-seq | N | N | N | - | - | - | - | Y | Y | N | - | N | Fasold et al. [2011] |
| Cyber-T | Web | RNA-seq | N | N | N | - | - | - | - | N | N | N | VSN | Y | Kayala and Baldi [2012] |
| ncPRO-Seq | Software | sRNA-seq | Y | Y | N | - | - | - | - | Y | Y | N | - | N | Chen et al. [2012] |
| shortran | Software | sRNA-seq | N | N | Y | Y | Y | N | N | Y | Y | Y | total count | Y | Gupta et al. [2012] |
| RobiNA | Software | RNA-seq | Y | Y | Y | Y | N | N | N | N | N | Y | RPKM | multiple | Lohse et al. [2012] |
| omiRas | Web | miRNA-seq | Y | N | Y | Y | Y | N | N | Y | Y | N | DESeq | DESeq | Mller et al. [2013] |
| Kraken | Software | RNA-seq | Y | Y | Y | Y | N | Y | N | Y | Y | Y | - | N | Davis et al. [2013] |
| TCC | R package | RNA-seq | N | N | N | - | - | - | - | N | Y | N | DEGES/TbT | multiple | Sun et al. [2013] |

## 4.3 Datasets

This section summarises the datasets used to illustrate our methods.

RNA-seq data from Vidal et al. [2013] is used to examine the differences between mRNA-seq and sRNA-seq data in plants. The data is from the organism *Arabidopsis thaliana* and consists of two conditions each with two replicates for both poly-A enriched mRNA and isolated sRNA fractions. The data was originally used to identify nitrate-responsive miRNAs and genes.

We used several datasets to demonstrate the uses of our pipeline. These were selected from the GEO database [Barrett et al., 2013] based on the criteria that only data with at least two replicates are used and the replicates are checked to ensure they were of a high enough quality to be used in a differential expression analysis in order to properly demonstrate the whole pipeline. We selected one good quality dataset from each of the plant and animal kingdoms.

The first, termed the "H" dataset, is an experiment on the effects of hypoxic conditions (in which cells are deprived of oxygen) on human MCF7 cells [Camps et al., 2014]. The experiment is split into a time series of four conditions: Normoxia (N00), Hypoxia at 16 hours (H16), Hypoxia at 32 hours (H32), and Hypoxia at 48 hours (H48). Each condition is replicated twice.

The second dataset, termed the "F" dataset, is an experiment in *Arabidopsis thaliana* to investigate the ability of the plant to avoid inappropriate silencing of its own coding genes by the silencing pathway used in defence of viral genes [Zhang et al., 2015]. In these experiments, three different mutants were sequenced that contained combinations of knocked out genes: rdr6-11 (*rdr*); ein5-1, ski2-3 (*es*); ein5-1, ski2-3, rdr6-11 (*esr*). A wild type (*col0*) was also sequenced. Each treatment was repeated three times, each with two technical replicates.

## 4.4 Methods and Results

In this section we describe the stages of the analysis pipeline, outlined in the schematic shown in figure 4.2, together with results found using our demonstration data. We also provide a comparison of the resulting differential expression analysis to two other differential expression tools.

### 4.4.1 Quality checking

The main quality check (QC) steps are undertaken after the reads have been tabulated, mapped to a reference genome, and annotated with other annotation

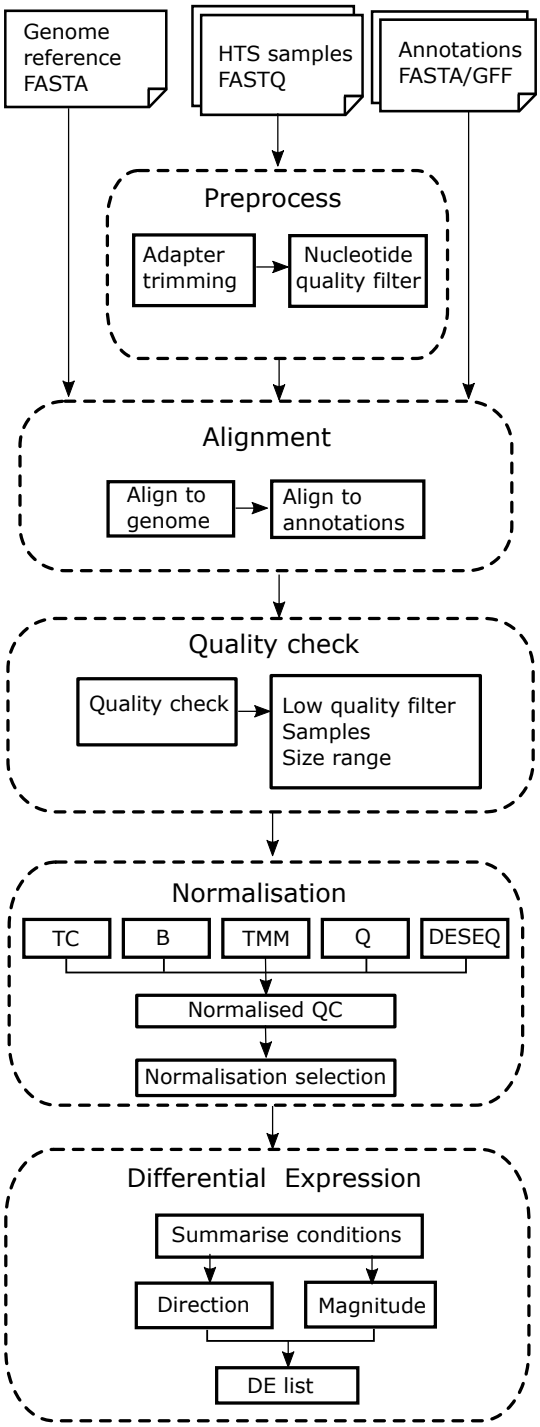



Figure 4.2: Schematic of the sRNA analysis pipeline.

sources. The quality checks assess the overall library similarities as well as deviations between sizes of sRNAs in individual libraries.

### Total library statistics

First, we assess the total library size (redundant count) and the number of unique sequences (non-redundant count) for each library. These values should be similar across libraries. The percentage of both redundant and non-redundant counts that mapped out of the total counts can indicate whether large numbers of sequences or highly abundant sequences were missed from the alignment. This can be due to contamination from other species or an incomplete reference sequence. An additional informative statistic is the count complexity of libraries and size classes. This is derived by dividing the non-redundant count by the redundant count. Complexity values that are close to 1 indicate a highly diverse set of low abundant sequences whereas lower complexity values are caused by a more homogeneous set of highly abundant sequences.

Quantitative assessment of the similarity between any two libraries is complicated by the dominance of low abundance sequences that appear in all libraries. Instead, we assess the similarity of libraries by considering the overlap between sets of the most abundant sequences because it is at the higher levels of abundance that discrepancies in sequence rankings are most important. The overlap is calculated using the Jaccard index, where for two sets of sequences $X$ and $Y$ we compute:

$$J_{XY} = \frac{X \bigcap Y}{X \bigcup Y} \tag{4.1}$$

This returns an index between 0 (no shared sequences) and 1 (all sequences are shared by both libraries). The Jaccard index is usually computed for all combinations of libraries and presented as a symmetrical matrix of indices. We expect to find that comparisons between replicates have a much higher Jaccard index than comparisons between samples from different conditions. However, all indices should be suitably high to maintain proper comparability between samples. The selection of a threshold for the number of ranked sequences to be compared can affect the resolution between library comparisons. This is illustrated by a Jaccard series for the H data shown in figure 4.3. However, most choices that only compare a small proportion of the dataset can produce an informative set of statistics.

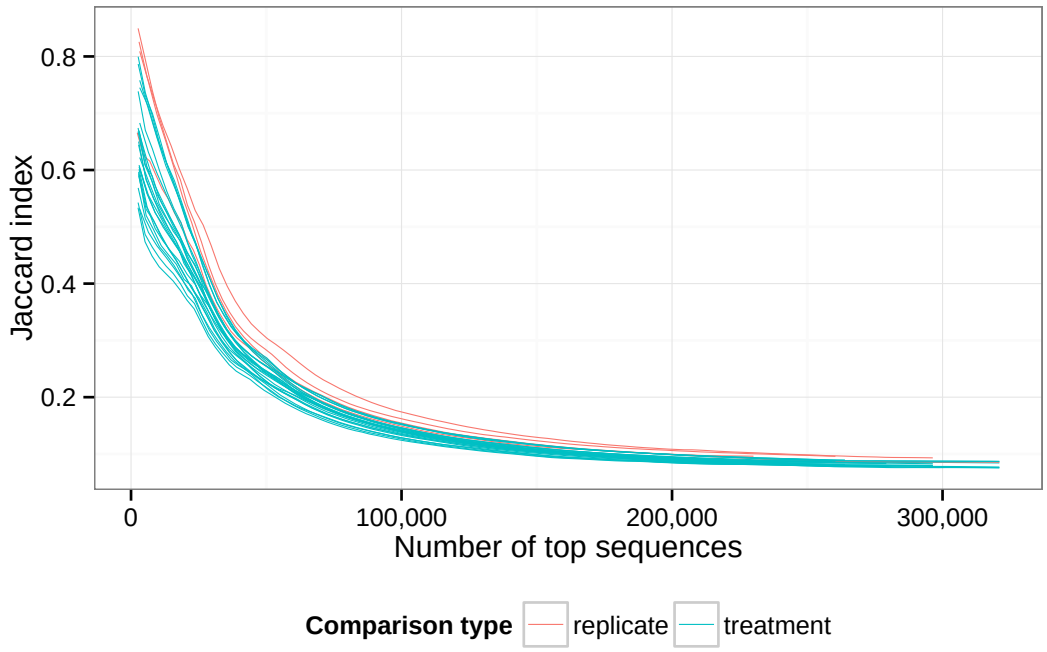



Figure 4.3: For all combinations of libraries in the H dataset, a series of Jaccard indices was calculated for varying magnitudes of sequence sets. These are plotted from the smallest to the largest sets with colour indicating the type of comparison: "replicate" is a comparison between two sample replicates of the same condition and "condition" is a comparison between samples in two different conditions.

### Size class distributions

Importantly for sRNA datasets, we also check the redundant, non-redundant, mapping percentage, and count complexity statistics for each sequence size class to identify important characteristics of each size class, or those that contain potential issues. We assess these statistics as a series of size class distribution plots shown in figure 4.4. The H dataset contains a peak in the redundant count distribution at 22-23nt where the count complexity is also very low. This indicates the presence of a few highly abundant sRNAs at these lengths. In contrast, the size class distribution of the F dataset indicates high numbers of unique 25nt sRNAs. If other annotations are used, this may also indicate the type of feature that certain discrepancies arise from. The differences between the H and F size class distributions are typical of the differences between the sRNA populations of plant and animal organisms [Mohorianu et al., 2012].

The size class distributions also reveal potential issues with both datasets, where the distribution of certain samples or conditions deviate from the rest at particular size classes. In the H data, one replicate of the H32 condition contains more unique reads than the other samples for sizes lower than 22nt, and there is a markedly higher complexity for an H16 replicate across the lower and higher range of size classes. In the F data, the mapped proportions reveal that much of the size classes for 22nt and 23nt could not be aligned to the reference genome. The *es* mutant is likely infected with viral siRNAs, which can be typical of plant sRNA-seq libraries. However, this type of contamination is unlikely to affect further analysis because these reads are not considered during normalisation of mapped expression levels and do not seem to impact the count of the mapped size classes in *es*.

### Replication comparability

The replication quality is checked by comparing replicates of each condition using MA plots, log scatter plots, Jaccard indices, and log fold change distributions by size classes.

MA plots and scatter plots are good visual indicators of the similarity of counts across the spectrum of abundance levels. Whilst it is easy to see strong correlations and deviations using the scatter plot, the MA plot directly displays log fold changes (M values) between replicates against their log average abundances (A values). This is important to ensure that the log fold changes of replicates are generally as close to 0 as possible.

The MA plots in figure 4.5 (a) show that the most dispersed fold changes are

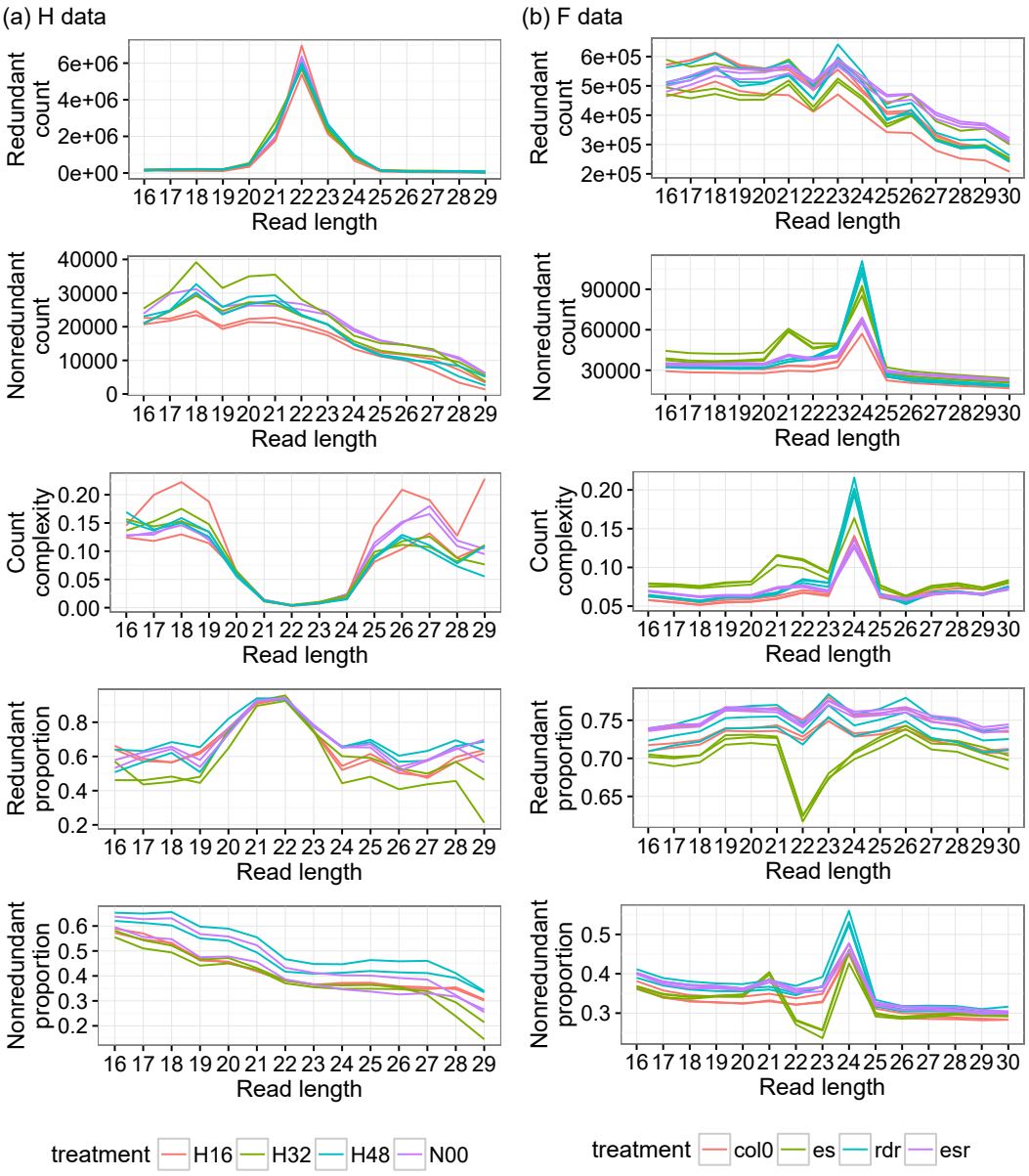



Figure 4.4: Size class distribution for all statistics produced for both demonstration datasets during the initial quality check stage. The type of statistic for each distribution is indicated by its y axis label.

those found between the replicates of H16. This appears to be a problem with the first replicate of H16, which results in lower jaccard indices against all other libraries (figure 4.6 (a)). In comparison to the H dataset, the MA plots between replicates of the F dataset show a very narrow distribution of M values (figure 4.5 (b)). These indicate good quality, comparable replicates, which is confirmed by the Jaccard index that rarely dips below 0.9 similarity (figure 4.6 (b)).

To assess the deviations between replicates of individual size classes, log fold changes are assessed for each size class using boxplots (figure 4.7). Both the range of the distributions and the deviations of fold changes from being symmetrically distributed around 0 can suggest issues that may affect further analysis. An alternative statistic is to assess within-group variance by calculating the coefficient of variance over all replicates for a condition [Dillies et al., 2013]. In the case of two replicates, the coefficient of variance is less informative than the fold change. Since sample sizes remain quite small for RNAseq studies, we opted to primarily use fold changes. However, the coefficient of variance would be a more efficient comparison with increasing number of replicates because it prevents combinatorial issues.

Replicate fold changes in the H dataset reveal a consistent deviating distribution of fold changes for the largest two size classes in all conditions except the control (N00) (figure 4.7 (a)). In comparison, the fold change distributions for replicate combinations of the F data have a tendency to deviate from 0 by the same amount. The latter discrepancy is less problematic and usually corrected by normalisation.

### Post-quality filter

The QC stage has two main purposes. Firstly, it allows us to understand the broad nature of our datasets. For example, the various size class distributions revealed the main size classes in both datasets (22-23nt in H and 24nt in F). Secondly, we are able to act on the identified causes of low quality within the data, usually by excluding outlying replicates, whole samples, size classes, and individual reads. In the worst case, this may show that the data is not viable for further analysis and certain samples should be re-created or re-sequenced.

The MA plots and Jaccard index for the H data showed that the H16 condition was composed of replicates with poor comparability. Since there are only two replicates per treatment, the H16 treatment was removed from further analysis. The F dataset also shows fold change distribution deviations between replicates, but these are similar for all size classes and can probably be compensated for by

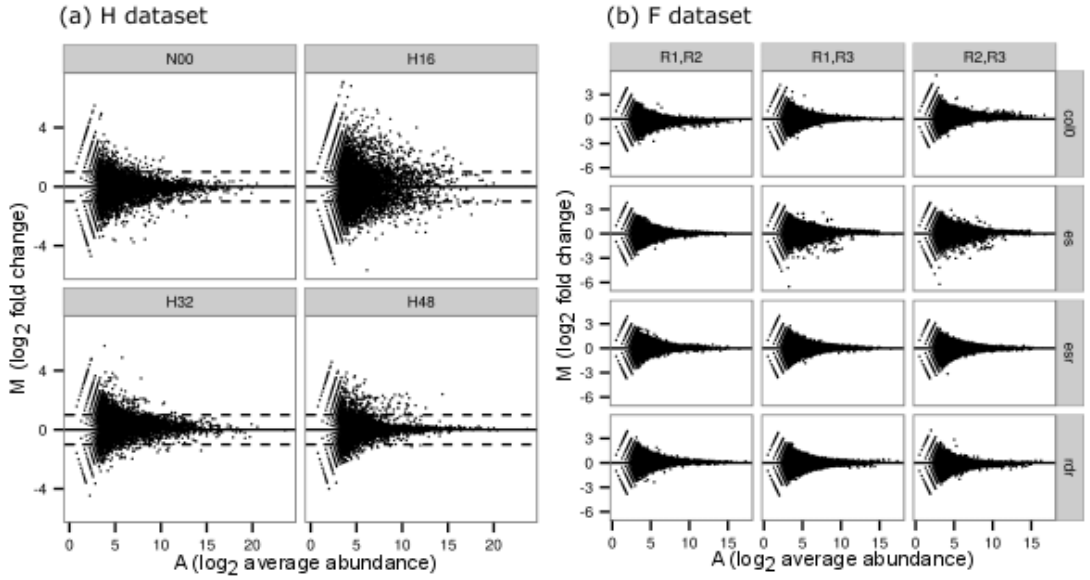



Figure 4.5: MA plots comparing combinations of replicates for both demonstration datasets.

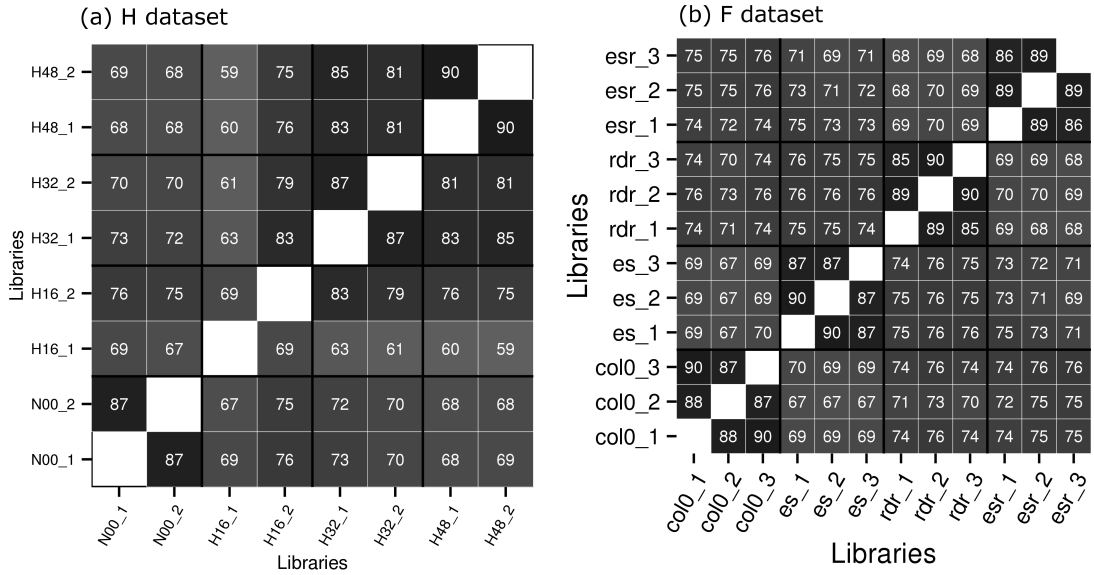



Figure 4.6: Jaccard index matrices for all library pairs of demonstration datasets. For the H data, 10,000 sequences were used for the index calculation. The F data calculation used 1,000 sequences.

using a scaling normalisation method.

The fold change distributions of size classes between replicates indicated high levels of deviation for the two longest size classes. Because we were less interested in these size classes (26nt and 27nt), we were able to remove these particular reads from further analysis to allow more accurate normalisation and differential expression analysis on the remaining size classes.

### 4.4.2 Normalisation

After the sRNA libraries have been assessed for quality, the expression levels are normalised across all libraries before differential expression between treatments can be calculated. Because there appears to be no single normalisation method that works best for all sRNA datasets [Dillies et al., 2013; Garmire and Subramaniam, 2012; Zhou et al., 2013], we incorporated several existing methods into our pipeline implementation and introduced a post-normalisation quality check step to select a normalisation method with the best outcome. The normalisation methods we used were Total Count (TC), Upper Quartile (UQ), Trimmed Mean of Means (TMM), DESeq, and Quantile normalisation modified for count data. These normalisations are reviewed in Chapter 3.

We also conducted a sampling based normalisation called Bootstrap normalisation. This is an adapted version of the Li and Tibshirani method [Li et al., 2012], where sampling with replacement is proposed. When two technical replicates are sequenced to different depths, the replicate with the larger library size may be normalised to the lower size by scaling all abundances down by the correct factor. However, figure 4.8 shows that it can not be assumed that all abundances require the same scaling factor, even between technical replicates sequenced at different depths. To alleviate this issue, the larger replicate may be resampled down towards the lower replicate using sampling-without-replacement on redundant sequences. This is implemented in our pipeline as a fifth normalisation method.

### 4.4.3 Post-normalisation quality check

Normalisations are evaluated for effectiveness in reducing unwanted variation in two ways: assessment of variation in count distribution over all samples, and the ability to minimize offset differences between the replicates for each size class.

Count distributions are shown as a boxplot of log abundances for each sample. Figures 4.9 (a) and 4.10 (a) show these count distributions for all available

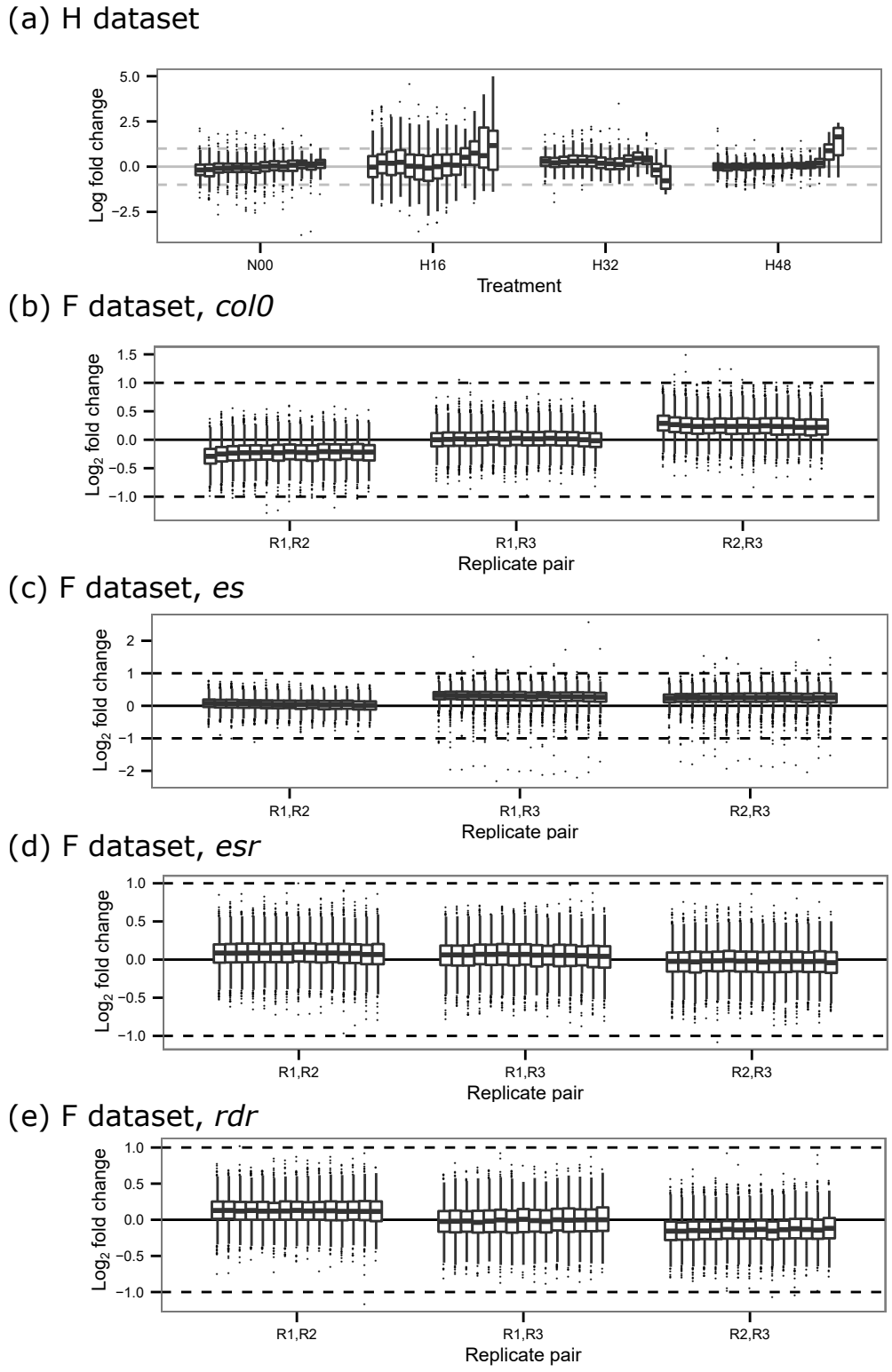



Figure 4.7: An example of log fold change assessment between replicates split by size classes.

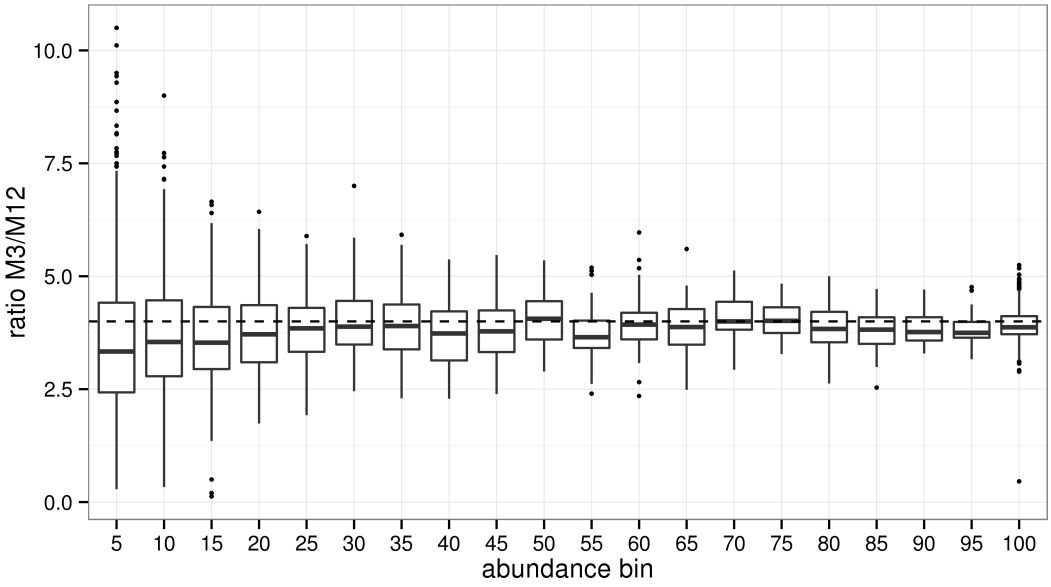



Figure 4.8: Demonstration of scaling between two technical replicates that have been sequenced under differing multiplexed conditions. M3 has been multiplexed with three other samples and M12 has been multiplexed with 12 other samples, leaving a four-fold difference between the two replicates. The y axis indicates the scaling factor required for each read to make the counts equal.

normalisations using the H dataset and F dataset respectively. Because the distribution of log abundances in sRNA-seq data are skewed heavily towards 0 (see figure 4.1 (a)), boxplots that depict the full distribution are indiscernible from one another. Instead, we visualise the abundance distributions of a subset of the most expressed sequences. These are found by summing the abundances of each sequence across all samples and taking the top $N$ sequences, where $N$ can be altered to view a variety of different abundance distribution windows. The graphic indicates the closeness of the distributions over all samples and will generally favour quantile normalisation.

To assess the differences between replicates, we calculate fold change distributions by size class between replicate pairs. Appropriate normalisations must minimize the interquartile range of all distributions whilst centering each distribution on the zero line. The assumption is that replicates should contain the minimal amount of variation between abundance levels and any normalisation that can lower this variation should be better than a normalisation that increases this variation.

Figure 4.9 (b) shows, for each size class, the comparability of abundances between replicates from H32 using fold-changes. These fold-changes should be minimised and centered on zero. Whilst the TMM, DESeq, and quantile methods

all appear to help centre the distributions of all size classes, the total count, bootstrap, and upper quartile methods do not improve on the distributions found by using the raw counts. This suggests that using TMM, DESeq, or quantile as the chosen normalisation for this analysis would be the best decision. However, this result is not the same for comparing normalisations on other datasets. The results of applying each normalisation method on the F dataset show that fold change distributions between replicates 2 and 3 of sample *esr* are correctly centered by all normalisations except TMM and DESeq, which appear to overcorrect their scaling by a larger difference than there was originally from the fold change distributions (figure 4.10 (b)). In this instance, quantile normalisation could be chosen because it also adjusts the abundance distributions to be equal. Dillies et al. [2013] found that quantile normalisation led to a significant increase in intra-group variation, but this is not seen in the assessment of quantile normalisation adapted for RNA-seq data when used on either datasets used here.

### 4.4.4 Calculating the differential expression of sRNA reads

In the following sections, we describe our method of calculating the log fold changes of sequences in such a way that they may be ranked without the introduction of uninteresting noisy candidates by low count sequences. We have termed this the Log Offset Fold Change method (LOFC).

The method first employs the conservative use of confidence intervals (CI) built on the distribution of a sequence's replicated abundance levels. For each sequence in each condition a CI was calculated using either Chebyshev's intervals [Singh et al., 2006] or the minimum and maximum abundance levels if only two replicates are used. For a selected comparison between a reference and observed condition, we then calculate both a direction of regulation and a magnitude as described below.

For each sequence, a directional descriptor from the set {up ($U$), down ($D$), straight ($S$)} is chosen in a similar way to the method applied in Lopez-Gomollon et al. [2012]. $S$ is used if the CIs overlap, otherwise $U$ indicates that the observed CI is higher than the reference, and $D$ indicates the opposite result.

The magnitude of a sequence between conditions is considered on proximate extrema of the reference and observed CIs. This is calculated using the log offset fold change on the extrema of confidence intervals, selected depending on the direction of differential expression. For a confidence interval $CI$ belonging to a read $i$ which has an upper limit $CI_{max}$ and lower limit $CI_{min}$, and comparing two

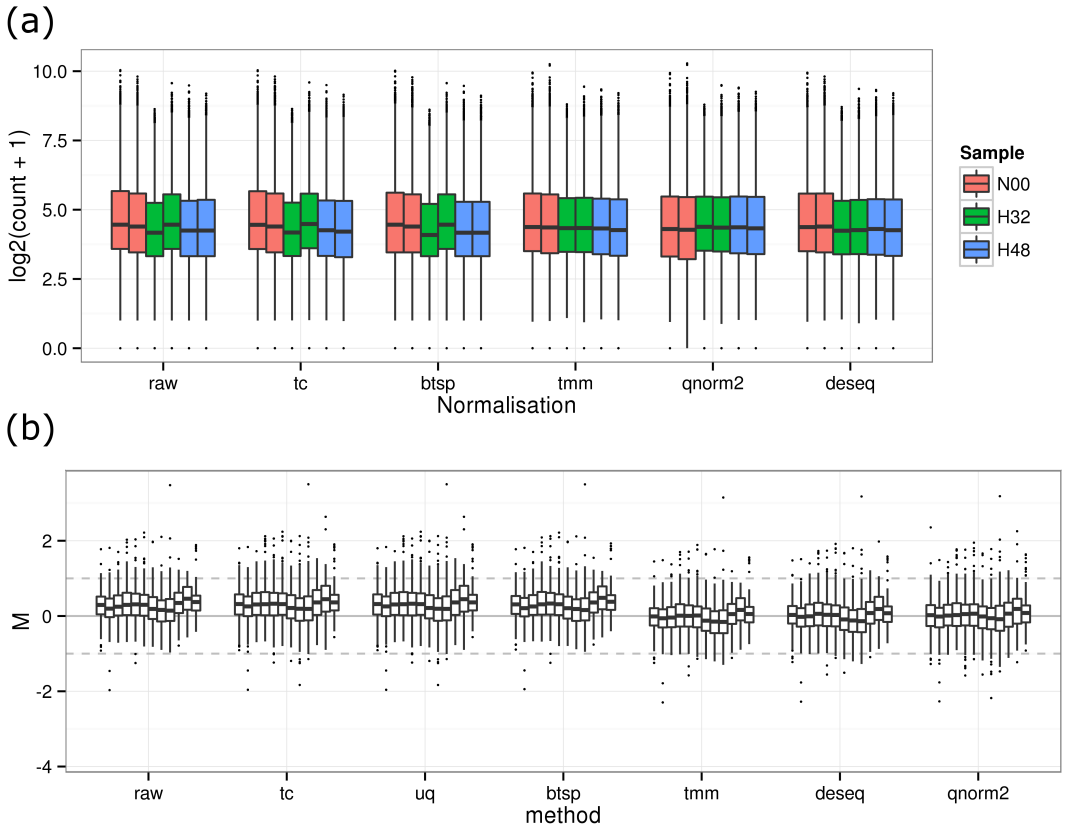



Figure 4.9: Graphics to aid the post-normalisation quality check step for the H dataset. (a) Abundance distribution of the top 20,000 abundance levels found by ranking sequences by their total abundance across libraries. (b) log2 fold change distributions for each size class between the two replicates of condition H32. Any fold change calculated from abundance levels below 20 were excluded. The normalisations listed a;along the x axis are unormalised (raw), total count (tc), bootstrap (btsp), trimmed mean of means (tmm), modified quantile normalisations (qnorm2), and DEseq normalisation (deseq).

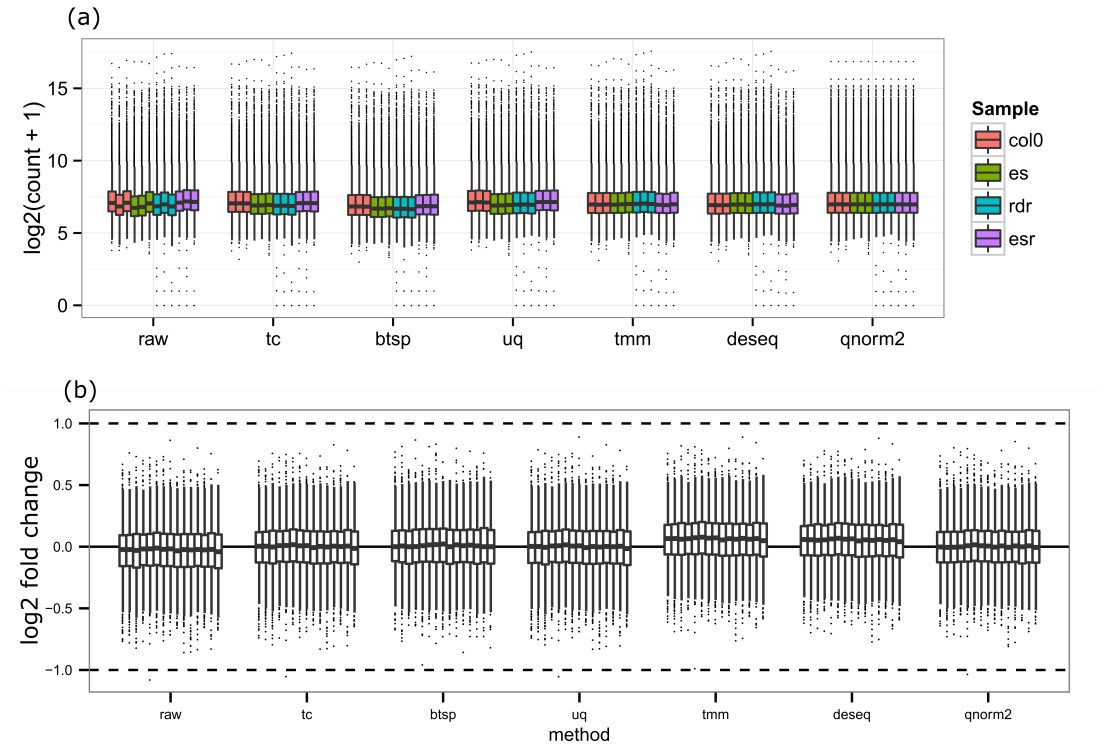



Figure 4.10: Graphics to aid the post-normalisation quality check step for the F dataset. (a) Abundance distribution of the top 20,000 abundance levels found by ranking sequences by their total abundance across libraries. (b) log2 fold change distributions for each size class for condition *esr* replicate 2 vs replicate 3. Any fold change calculated from abundance levels below 20 were excluded. The normalisations listed a;along the x axis are unormalised (raw), total count (tc), bootstrap (btsp), trimmed mean of means (tmm), modified quantile normalisations (qnorm2), and DEseq normalisation (deseq).

conditions $j$ to $j'$, the LOFC is computed using the formula

$$LOFC_{i,j\to j'} \begin{cases} log\left(\frac{CI_{max,j'}+o}{CI_{min,j}+o}\right), & \text{if U} \\ log\left(\frac{CI_{min,j'}+o}{CI_{max,j}+o}\right), & \text{if D,} \end{cases} \tag{4.2}$$

where $O$ is a set value termed the offset. The offset approach, initially described in Mohorianu et al. [2011], helps reduce the number of false positives from low abundance sequences and allow fold change values to be directly used when assessing the relative significance of differentially expressed sequences.

To determine an appropriate offset for a dataset, we estimate the abundance level where the majority of noise-related reads lie. We define strand bias as

$$SB = |0.5 - p| + |0.5 - n| \tag{4.3}$$

where $p$ and $n$ are the number of unique positive strand and negative strand reads respectively in a window. We found that low abundance loci tend to have a high strand bias and loci within the noise to signal range have no preferred strand bias. Based on this observation, we assigned sRNAs to windows of a set length along the reference genome and the total abundance and strand bias was calculated for each window. For all abundance levels $A$, the distribution of $N$ strand biases was compared to a random uniform distribution using the Kullback-Leibler divergence [Mohorianu et al., 2011]

$$KL_A = \sum_{i=1}^{N} log_2\left(P_i\right)\left(\frac{log2\left(P_i\right)}{log2\left(Q\right)}\right), \tag{4.4}$$

where $P_i$ is the proportion of strand biases that took the value $i$ and $Q$ is the uniform distribution $1/N$.

We define the signal to noise threshold (the offset) as the value for which the global minimum of KL divergences is reached. Abundance levels lower than this threshold tend to have a higher divergence from a uniform strand bias due to a low number of incident reads, and abundance levels that are higher than the threshold have an increasing divergence measure due to biologically relevant reads. The minimum is found after calculating smoothed values from the distribution using Loess smoothing [Cleveland and Devlin, 1988]. This is done to prevent local minimums from biasing the more general trend across differing abundance levels (see figure 4.11 (a)).

We assessed the dependence of this offset on the number of strand bias bins,

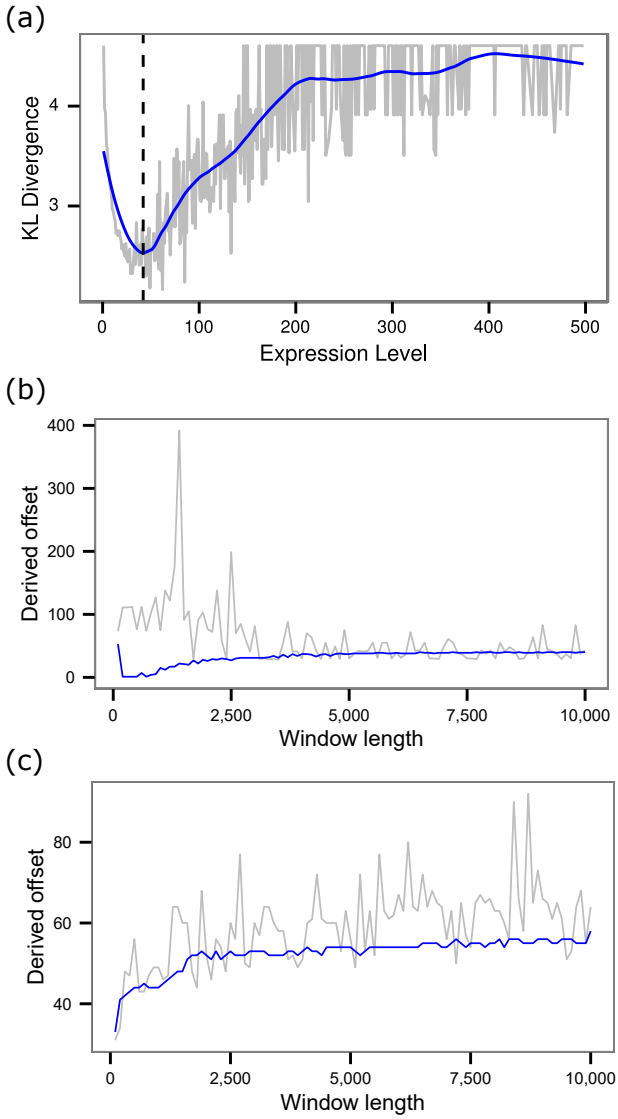



Figure 4.11: Derivation of the offset for a sample using the Kullback-Leibler (KL) divergence. (a) The result of calculating the KL divergence measure on strand bias bins for each level of abundance given on the $x$-axis. The length of window used was 4,000nt. The grey line indicates unsmoothed KL divergence values and the blue line is divergence values smoothed by Loess (span=0.3). The offset abundance level is identified by the minimum of smoothed divergence values as 42. (b) and (c) The results of calculating the offset in this way for varying window lengths. The grey line is the offset found at the minimum of the unsmoothed divergence curve and the blue line is the offset found at the minimum of the smoothed divergence curve. (b) is run on the N00_1 of the H dataset and (c) is on wt1 of the F dataset.

alignment window length, and the type of organism the data is sequenced from. In the H dataset, representing animal data, the number of strand bias bins heavily affected the resulting offset up to 100 bins, at which point no further difference can be seen on the KL curve (data not shown). The offset was also affected by alignment window length and can vary erratically when using the raw measures. The smoothed values, however, return a more consistent offset across differing window lengths. For the H dataset, the curve is generally less well defined at window lengths under 2,500nt, which returns an offset biased towards the lower end of abundance levels. However, longer windows than 2,500nt produce a stable offset when using the smoothed curve values (figure 4.11 (b) and A.1). In our plant dataset (F dataset), the smoothed minimum was similarly variable below a length of 2,500nt, but was able to stabilise thereafter for most of the samples. The *rdr* samples, however, contained a brief notable increase in the offset between around 1,250nt and 3,750nt (figures A.2, A.3). We therefore selected a window length of 4,000nt for deriving a suitable offset.

### 4.4.5 Comparison of the LOFC method to other tools

To assess the usefulness of our pipeline for identifying important differentially expressed sequences, we compared our method with two highly cited methods for differential expression: DESeq2 [Love et al., 2014] and edgeR [Robinson et al., 2010]. Both methods assume the data fits a standard binomial model and add dispersion estimators to account for deviations from this model. The LOFC method, however, does not assume the data fits any particular model, but only that fold changes are more important with increasing log-average abundance. Additionally, we more stringently filter for the requirement that the confidence distribution of replicate counts for a sequence is suitably different between the two samples to warrant calculation of differential expression.

Both DESeq2 and edgeR use adjustable $P$-values to indicate a threshold of significant differential expression, which is normally set at $P = 0.05$. Additionally, these tools also allow the user to set a $log_2$ fold change (LFC) threshold as part of the significance test, allowing sequences to be chosen that are significantly greater than a set fold change. We set a LFC of 1 for the significance tests in DESeq2 and edgeR, and use a LOFC of 1 for extracting selected differentially expressed sequences from our LOFC method. An LFC of 1 was selected based on empirical evidence that a sequence with a log2 fold change of 1 can be detectable on a northern blot or via qPCR [Morey et al., 2006].

We compared the three different differential expression approaches in several

different ways. To understand the differences between the significance cutoffs in each approach, we plotted the LOFC of sequences grouped by the tools in which they were found to be significant and analysed the overlaps of the resulting sets of significant calls. Figure 4.12 shows this comparison for the H dataset. Any sequences that were marked as Straight were not included in the plot, and none of these sequences were found to be significantly differentially expressed by the other two methods. However, a further 13 sequences in N00 vs H32 and 18 sequences in H32 vs H48 were above an absolute LOFC of 1 but were not called significant by DESeq2 or edgeR. Many of these sequences showed a high LFC under both tools but did not pass the tests for significance, despite several of the confidence intervals showing an appreciative proximate distance from each other (figure 4.13). A second important difference is demonstrated by the numerous sequences that are called significant by one or both tools whilst having an absolute LOFC of below 1. These sequences tend to have non-overlapping confidence intervals but their low average abundance means that they may pollute the fold change ranking with high but ultimately inconsequential fold changes. Both edgeR and DESeq2 rank their significant sequences by LFC values found using the mean of all replicates. A comparison of these values against the LOFC values is shown by figure 4.14. Low abundance values that have otherwise high LFC values are pushed towards 0 LOFC by using offsets. Whilst many of the affected values are sequences found differentially expressed by edgeR (shown in blue in the figure), DESeq2 appears to be able to better reject the low abundance LFC values that also have a low LOFC. However, this appears to be at the expense of missing some sequences that are just above 1 LOFC at an increased log average abundance. edgeR, on the other hand, is more comparable with LOFC at higher log average abundances but is far more sensitive at lower log average abundances.

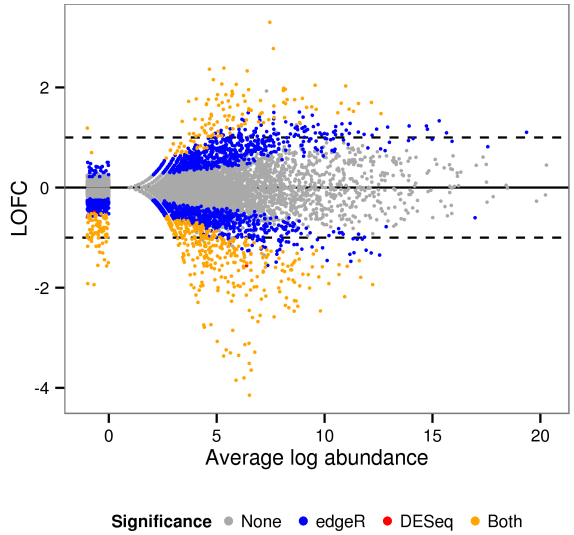
### 4.4.6 Software

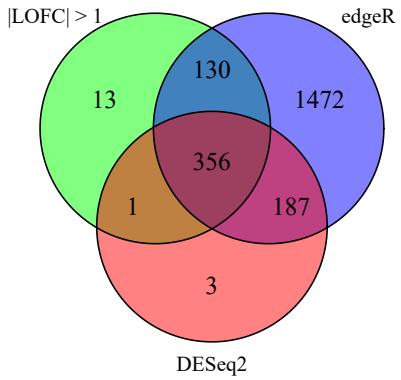
We now describe our implementation of the pipeline detailed in the previous sections.

#### Workflows

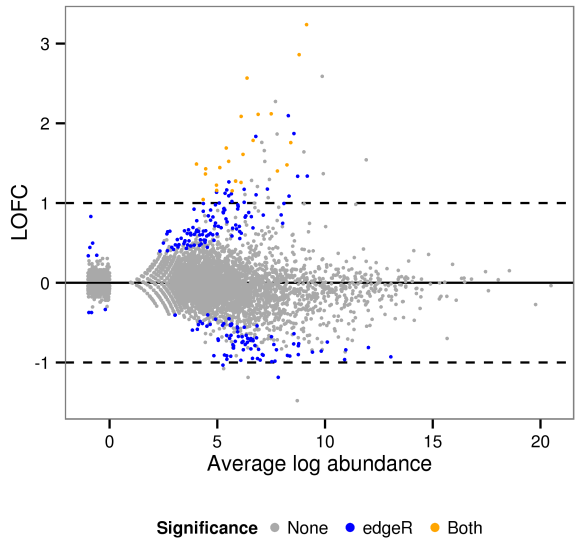
The pipeline is built into the UEA small RNA Workbench package [Stocks et al., 2012] but in contrast to the original multi document style of the workbench, it is presented to the user as a workflow diagram linking each distinct part of the pipeline together (See Figure 4.15). The workflow diagram consists of multiple

(a) N00 vs H32




(b) H32 vs H48



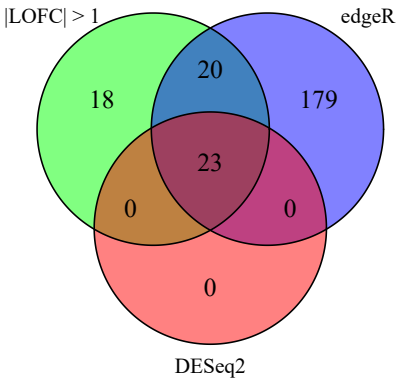




Figure 4.12: (left) MA plot of LOFC values against the average log abundance with sequences found significantly expressed by other tools highlight as described in the legend. (right) A Venn diagram depicting the amount of overlap between sequences called significantly differentially expressed in edgeR, DESeq2, and sequences greater than an absolute LOFC of 1 in the LOFC method.

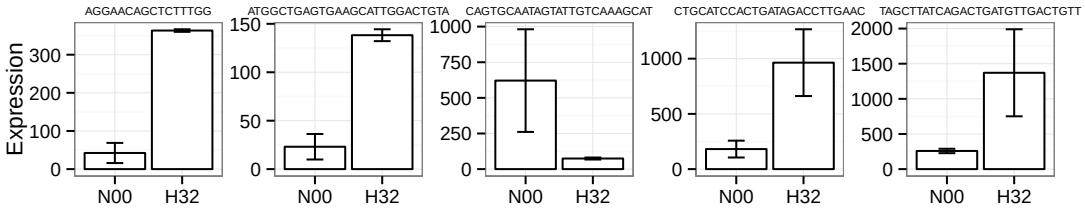



Figure 4.13: Normalised abundance levels and confidence intervals of the five most differentially expressed sequences under an LOFC analysis that are not called significant by other tools.

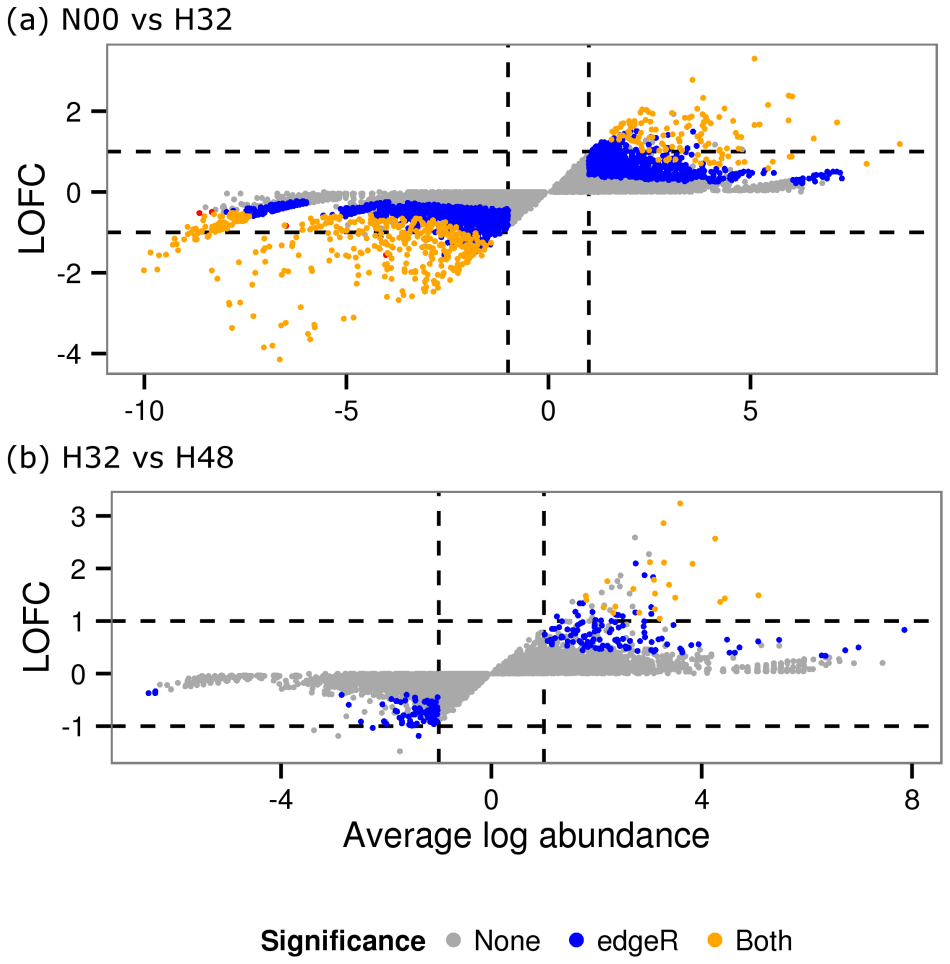



Figure 4.14: LOFC values plotted against LFC values for both comparisons in H dataset. Significance of LFC values are shown in colour depending on which tool found them significant. The LFC values were taken from DESeq2 and were calculated from average abundances over replicates.

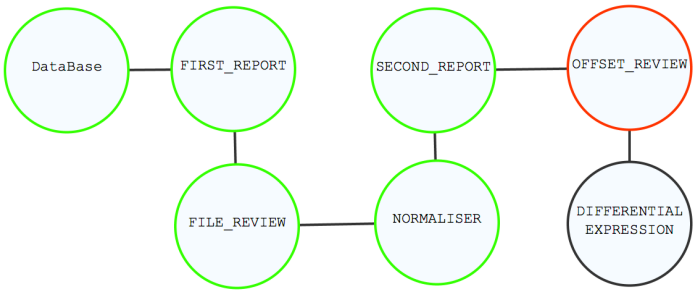



Figure 4.15: An example of the UI workflow diagram presented to the user in the Workbench implementation of our pipeline

user configurable nodes that represent the various stages in the analysis and the user can click on each part of the workflow to individually configure that area. The workflow is designed to be easy to use by both biologists and bioinformaticians, foregoing the need for many separate programs that require interlinked inputs/outputs.

Initially, the sequencing data can be processed from raw FASTQ formatted files by using an updated version of the adapter removal tool previously described in [Stocks et al., 2012]. The tool also provides the ability to process samples produced using the HD protocol described in [Sorefan et al., 2012]. However, this is currently not part of the differential expression pipeline and is instead currently available in the workbench as a separate standalone tool.

The first stage in setting up a differential expression workflow using our pipeline is to organise the data by creating a sample hierarchy that describes the original wet lab experiment. This is visualised as a tree structure where leaf nodes represent biological replicates and the parents of these nodes represent the samples. An example of a sample hierarchy is given in Figure 4.16. Users build the hierarchy by inputting their FASTA formatted samples into a setup wizard. They can then provide a reference genome, also in FASTA format, and an optional GFF file of annotations corresponding to the genome build that will be used for the annotation stage.

The tool currently accepts one or more GFF files for further annotation of genome-mapped sequences. The user is able to filter the features found in the GFF file down to only those of interest. Reads are then annotated by searching for overlaps between the features in the GFF file and the aligned sequence for each reference sequence. By sorting both annotation alignment set and the read alignment set, the search for overlaps between these two sets can be computed efficiently by advancing to the next alignment in one set if we have exceeded

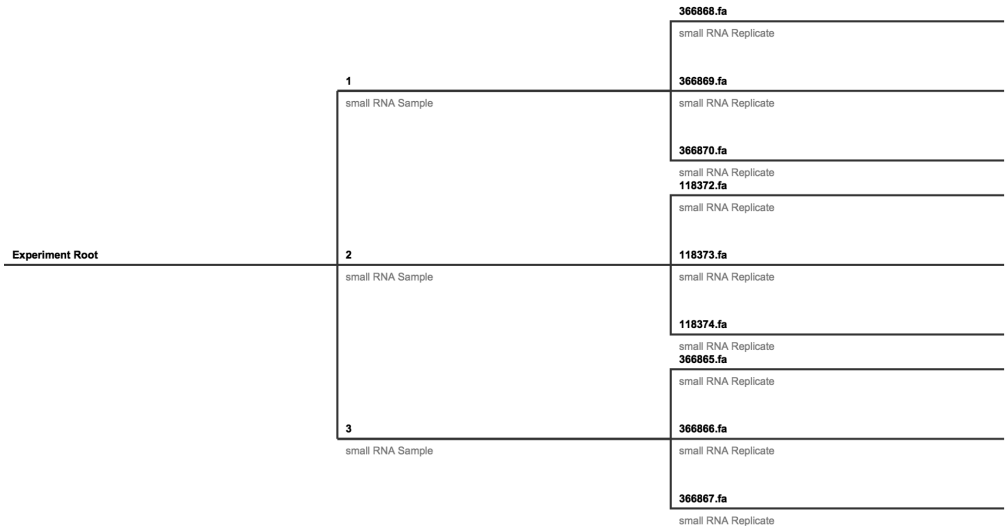



Figure 4.16: An example of the hierarchical visualisation used to depict the user's experimental setup as the input to our pipeline.

the position of its end coordinate in the other set. Alignments are also cached and re-checked only up to just before the start coordinate of the alignment being checked against. This ensures that the minimum number of checks are made between sets to find all possible overlaps.

The quality check stages are implemented as a report that pauses the workflow analysis and presents to the user the graphs described in section 4.4.1. These are implemented using D3.js; a data visualisation library for JavaScript [Bostock et al., 2011]. The user is able to dynamically configure the annotation classes and normalisation methods that are displayed for most of these graphs. Some examples of quality checking graphs produced in these reports are shown in figure 4.17. After assessing the quality of the data, the user may make adjustments to the set of samples, size classes, or select the normalisation method to be used before continuing. Finally, prior to beginning the differential expression stage of the analysis, the user can review the offset values and select the desired smoothing value for the KL divergence curve.

### Implementation

The quality check, normalisation, and differential expression steps are computationally intensive and pose significant demands on both processor and memory. Our aim is to facilitate its use by users with access to a wide range of comput-

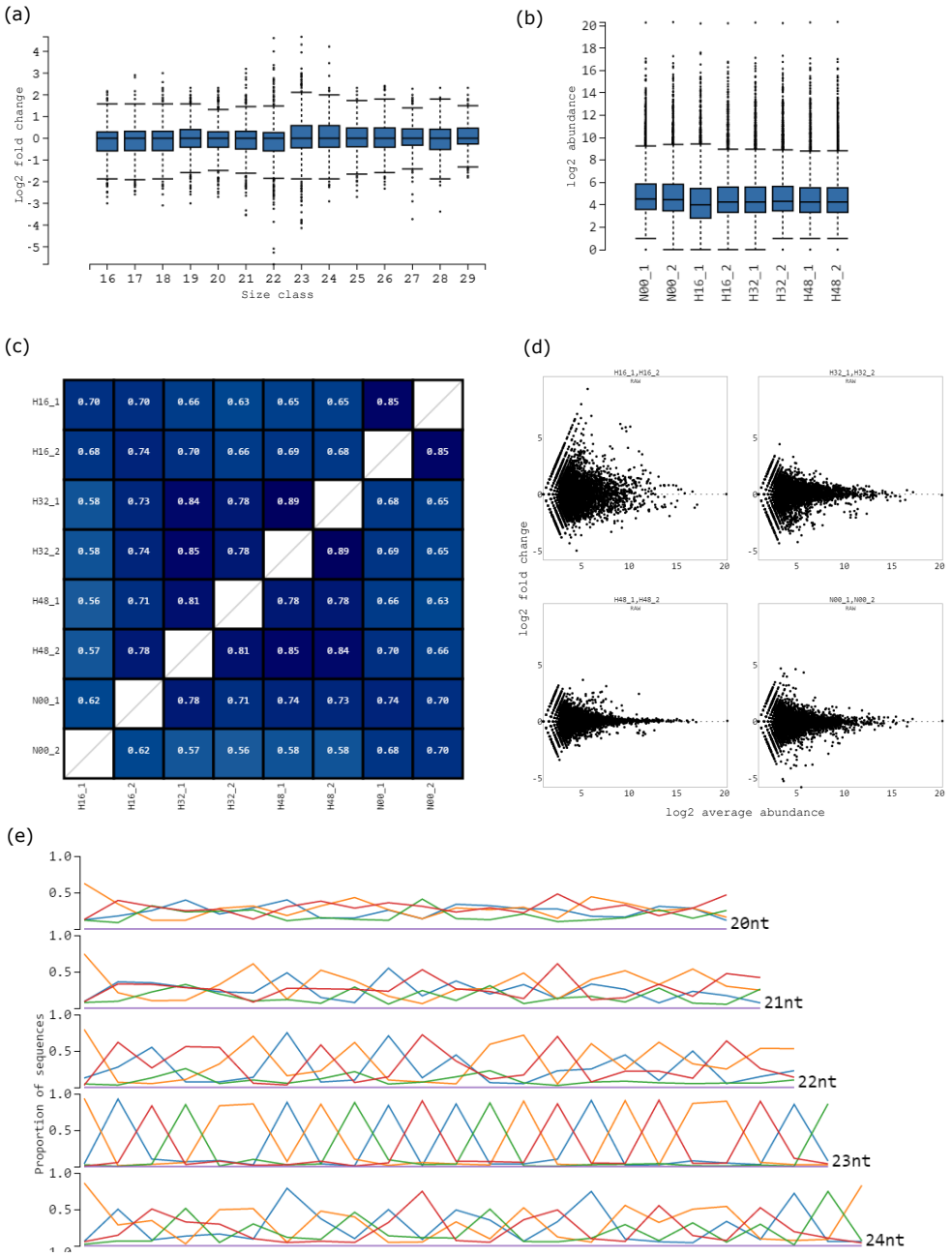



Figure 4.17: Examples of some of the plots produced as output during the first quality check step of the UEA sRNA Workbench: (a) Fold change boxplots between two replicates for raw data (b) abundance boxplots for raw data (c) a Jaccard matrix heatmap of the top 1,000 sequences between all libraries (d) MA plots comparing replicates in each condition (e) positional frequencies of nucleotides split by size class. The graphs were produced using a subset of the Hypoxia dataset.

ing hardware including standard desktop equipment. Moreover, the UEA sRNA Workbench has been designed using a virtual machine, Java, which comes with the caveat of having a fixed amount of RAM that it can call upon during runtime but allows multiple operating systems to be supported. To this end, an embedded database model was employed to assist in organising the large amounts of data structures required to conduct the analysis. We opted to use the H2 SQL database [Mueller, 2006] because of its speed when initially importing large tables and sequential operations. The use of the database allows greater control to be placed over the amount of memory that is required by the workbench during the analysis. In addition, large portions of data can be stored on disk and read into the pipeline using structured queries, then systematically handed back to the disk. The change from the RAM-only standard model to one that also employs disk can have negative effects on the runtime for any computational procedure. However, it also means that a larger number of datasets can be processed on lower specification hardware such as a desktop computer.

To address the increase in running time, certain strategies for caching and loading of data are employed during a run. The amount of data that can be cached for use in the analysis is finely tuned depending on how much RAM the virtual machine has available at any one point in time. Theoretically, the more RAM that is made available to the system, the faster the analysis will complete. A server version is also available that uses no disk caching and utilizes RAM only for maximum performance, which is similar to the way in which other RNA-seq software packages are run.

## 4.5 Discussion

In this chapter, we have described an analysis pipeline for the annotation and differential expression of sRNA-seq data that takes into account the unique characteristics of sRNA-seq datasets. This was implemented into a user friendly interactive workflow within the UEA sRNA Workbench.

### 4.5.1 Acting on thorough quality checks can improve downstream analysis

We propose to extend the usual quality checks made on RNA-seq data to further take into account the discrepancies of individual size classes between replicates and treatments. The information gained from the quality checks can be used

to filter from further analysis any low quality or outlying replicates, reads, size classes, or even whole conditions if this is not detrimental to the experiment's power. Interestingly, the removal of data has more of an effect on the outcome of DESeq2 and edgeR tools than the LOFC method due to the calculation of global dispersion estimates that incorporate the information from all libraries when comparing any two conditions. For example, leaving the H16 condition from the H dataset in the experiment has no effect on the two comparisons that LOFC was calculated on, but will produce larger dispersion estimates and different alpha values in the statistical tests run by DESeq2 and edgeR. It is therefore crucially important that the appropriate quality checks are made and acted on before providing data to these differential expression tools.

### 4.5.2 Normalisation quality checks are useful for selecting the most appropriate method

Due to the wide variety of normalisation methods, and the lack of consensus on a method that works for all sRNA data, we advocate testing several different methods and using a variety of normalisation measures to identify a normalisation that maintains a high degree of similarity for abundance distributions across all libraries and a low degree of difference between replicates. The inconsistency of normalisation results is demonstrated here by the ability of TMM and DESeq to minimize differences between some replicates in the H dataset but can not prevent deviations between replicate pairs in the *esr* treatment of the F dataset. We also introduce a sampling normalisation method, Bootstrapping, intended to reduce some of the linear scaling issues that total count introduces. In our demonstration data, bootstrapping appears to perform no better than total count data. Part of the difficulty in properly assessing the effects of normalisation lies in the need to select demonstration datasets with enough replicates that are both comparable enough yet have enough issues to resolve the differences between normalisations when correctly measured. While these datasets are not able to show the differences between total count and bootstrapping, others may reveal some important differences. Additionally, if the quality or consistency of libraries are too low, they may be unsalvageable by any normalisation, save for assessing each annotation class individually. This is a scenario that is tackled in chapter 6 of this thesis.

### 4.5.3 Offset fold change is a reasonable alternative to dispersion estimates

Whereas DESeq2 and edgeR utilise a dispersion estimator coupled with a variety of tests for significance, we find that the use of an offset to downweight low-abundance fold changes coupled with a test for overlapping confidence intervals is enough to equate the effect of the dispersion estimators. The chief differences appear to be that, depending on the offset chosen, low log average abundance reads are likely to be more penalised in the LOFC method whilst reads with a higher log average abundance are more likely to be higher up the rank of differentially expressed reads.

To identify a suitable offset, we assessed the low end of the abundance distribution of sRNA loci for their divergence from a uniform strand bias. For each library, the offset was chosen to be around the maximum abundance level that favoured the most uniform strand bias. Although this produces an offset that is non-arbitrary, it introduces further parameter considerations such as the size of the loci and the span that is used to smoothen the resulting divergence estimates. The span is easily tuned by the use of an interface and slider in our resulting software. However, consideration of a suitable loci length is difficult due to fluctuation in divergence minimums for small loci, which necessitates trying a variety of loci lengths to identify those that are more stable. This is a computationally expensive operation, and further work is needed to understand exactly why small changes in loci can produce large changes in the resulting divergence minimum.

## 4.6 Conclusions

With the introduction of this sRNA processing pipeline, we hope to provide new methods and approaches to ensure that sRNA datasets are properly assessed for their quality and correct normalisation before differential expression analysis takes place. By implementing this into a software package that is simple to use and with low computer memory requirements, we also hope to make these new methods for quality checking and detetecting differential expressed sequences as convenient to use as possible wihout the need for a high performance computing environment.

# Chapter 5

# Identification of miRNAs involved in caste differentiation of bumblebees

This chapter is adapted from Collins DH, Beckers M, Mohorianu I, Moulton V, Dalmay T, Bourke AFG, "*A MicroRNA Associated With Caste Determination in a Bumblebee is Expressed from a Mirtron Within a Homologue of Vitellogenin*", in preparation.

The miRNAs identified in this chapter have been published as part of "The Bumblebee Genome Consortium, *The genomes of two key bumblebee species with primitive eusocial organization*, Genome Biology, 16:76, 2015".

## 5.1 Summary

In this chapter, we focus on the use of sRNA-seq datasets for the identification of novel and conserved miRNAs in a novel genome model. To this end, we analyse an experiment on the regulation of miRNAs during caste differentiation of the bumblebee *Bombus terrestris*, and use data mining strategies outlined in the previous chapter to understand the miRNA population in this novel genome model.

## 5.2 Background

Many animals have the ability to conform to one of several different phenotypes throughout the stages of its life. This ability is termed "Phenotypic Plasticity" [Pfennig et al., 2010; West-Eberhard, 1989], and is a highly interesting topic of

study for genetics. Extreme cases of phenotypic plasticity occur in many eusocial insects, where larvae develop into specific castes that contribute different skill sets to the colony in an altruistic manner. In this case, the larvae are totipotent, meaning they have the ability to develop into more than one phenotype, before developing into a fixed phenotype where the cells are specialised to perform particular tasks. Such an event is termed Caste Differentiation. The evolutionary causes of phenotypic plasticity in social insects are mostly understood [Bourke, 2011]. However, the mechanisms behind Caste Differentiation are less clear [Smith et al., 2008].

*Bombus terrestris* is a species of eusocial bee in which the individuals of a hive specialise to reproducers (the Queen caste) and non-reproducers (the worker caste) [Goulson, 2003], which differ in both reproductive capability and morphology. Just after hatching, the larvae are totipotent and able to develop into either of the two castes, losing their totipotency through a series of endogenous changes. This chain of events is thought to be triggered by a pheromone produced by the Queen within 3-5 days of egg hatching [Cnaani et al., 2000]. To date, two studies have identified a number of genes found differentially expressed between castes in *B. terrestris* [Colgan et al., 2011; Pereboom et al., 2005]. These studies concluded that the regulation of gene expression was highly important during development of fixed castes and that these genes were not necessarily the same as those found in *A. melifera*, the closely related eusocial honeybee. In addition, miRNAs have a key role in the development of plastic traits in other insects, such as the development of wings in pea aphids in response to the population size [Legeai et al., 2010], however no such RNAi research has been conducted on the development of castes in any species of bee to date.

The aim of this study is to use sRNA-seq data to first identify conserved and novel miRNAs in an organism where no miRNAs have yet been identified. We will use the recently assembled genome of *B. terrestris* together with miRNA information from related species and miRNA prediction tools to assess the presence of miRNAs in the bumble bee. Secondly, we will assess the differential expression of miRNAs and related ncRNAs from larval stage to adult stage bees developing into both the queen and worker castes. We will also compare larval and adult stages for further differential expression, expecting the adult stage to have increased differential expression due to their loss of totipotency.

## 5.3 Methods

### 5.3.1 Biological methods

This experiment used 40 *Bombus terrestris* colonies after raising them for 28-93 days, depending on when the first males eclosed. Seven colonies were excluded from the experiment due to either the loss of the queen or contamination of worker bees. From the remaining 33 colonies, we retained the queen in 13 of them to allow them to generate worker-destined larvae (queenright), and removed the queen in 20 colonies to generate queen-destined larvae (queenless). In the queenright colonies, we removed up to half of the 1st or 2nd instar larvae (1-3 days old) every 2-3 days until approximately 10-14 days after first worker eclosion. In the queenless colonies, we removed up to half of the 1st or 2nd instar larvae every 2-3 days for 6 days after the queen was removed. In both queenright and queenless colonies, we allowed all unsampled larvae to develop to the 4th (final) instar, which is beyond the point in larval development when caste fate has been irreversibly determined. We then sampled approximately half of the 4th instar larvae from both sets of colonies. All 1st and 2nd instar larvae where treated as "early instar" and 4th instar larvae were called "late instar".

The sampled larvae were used to determine the colonies that produced the highest proportion of expected castes. Four colonies were selected for sampling each instar and caste type, creating 16 samples of 4 conditions with 4 replicates each. The conditions are Early Worker (EW), Late Worker (LW), Early Queen (EQ), and Late Queen (LQ).

We used total RNA extracted from the queen- and worker-destined larvae to construct 16 cDNA libraries. To make the cDNA libraries, we first enriched the total RNA for small RNAs (sRNA) (i.e. enriching the fraction of total RNA that was less than 200 bp in length) using a mirVana miRNA isolation kit (Ambion, Foster City, California, USA) according to the manufacturer's instructions. We then prepared the libraries using the TruSeq small RNA library preparation kit v.1.5 (Epicentre Technologies, Madison, Wisconsin, USA) with HD modifications to the 3' adapter to reduce sequencing bias [Sorefan et al., 2012]. To ligate the adapters to the sRNA sequences, we followed the protocol provided with the TruSeq 1.5 library preparation kit with some modifications. Following preparation of the cDNA, we amplified each library with a unique index sequence using Illumina index primers (1-16) before using PCR. We separated the PCR products on an 8% polyacrylamide gel, to identify the 21-23mer miRNA band on the gel, and cut out the gel section that contained it. Finally, we packed

the 16 prepared cDNA libraries in dry ice and sent them for miRNA-seq on the Illumina HiSeq2000 platform, which was conducted by BaseClear B.V, Leiden, The Netherlands.

### 5.3.2 Bioinformatic analysis

#### Preprocessing

We removed 3' adapters from sequences by matching the first 8 nucleotides of the adapter sequence and trimming the 3' end of each sequence from 4 nucleotides upstream of the adapter start coordinate to take in to account the multiplexed nucleotides of the HD adapter. Any sequence that did not contain an adapter was excluded from the rest of the analysis.

We then filtered the trimmed sequences by keeping only those with read sizes between 16nt and 30nt and that also contained at least 3 different nucleotides. The filtered set of sequences were then mapped to the *Bombus terrestris* genome Version 1.0 using PatMaN [Prüfer et al., 2008] with no mismatches or gaps.

We subjected the mapped sequences to several quality checks to ensure that the libraries were comparable. Bootstrap normalisation (chapter 4) and quantile normalisation (chapter 3) were both attempted on the data. We chose to keep the Bootstrap normalised data because it minimized the coefficient of variation between replicates whilst adequately also minimizing the difference between the count distributions at the top end of abundances over all samples.

#### miRNA gene prediction

Because the *B. terrestris* genome was not yet released, miRNA annotations were not available and had to be predicted for a novel genome. We used a combination of two different prediction programs, miRCat [Stocks et al., 2012] and miRDeep2 [Friedlnder et al., 2012], as well as miRNA alignments from related species and other animals to identify the broadest possible population of both new and conserved miRNAs. We used MapMi [Guerra-Assuno and Enright, 2010] to find potential miRNA precursor sites based on mature miRNA sequences from miRBase that were conserved in other species of the Hexapoda sub-phylum.

miRCat and miRDeep2 were both executed on all available samples after they had been mapped to the genome. miRDeep2 was run using default settings and miRCat was run using the default animal parameters.

We used a custom script, supplied online at `https://github.com/mattlbeck/collins_et_al_MCDB`, to first merge all runs for the two prediction tools into one

set of miRNAs, each containing a set of "predictions" that could differ between samples. Different miRNAs in this sense were defined by their exact location. These miRNAs were assigned an "arm" (either 3' or 5') based on the side of the precursor they were mostly on and were grouped into distinct miRNA precursor entries. Precursors were then labelled as having been identified by a subset of the three tools. miRNA names were based on their miRBase ID if MapMi had predicted them, and were assigned a unique ID otherwise.

To allow the most number of reads derived from miRNAs to be annotated, the reads were aligned to the resulting set of miRNA precursors.

### Differential Expression

To calculate differential expression of normalised reads, we used the methods outlined in chapter 4. Briefly, replicates for each sample were converted to confidence intervals and the magnitude of differential expression for each read between two treatments was found using the Log-offset fold change method based on the proximity of confidence intervals. If the confidence intervals overlapped, the sequence was not regarded as differentially expressed and eventually filtered from the final set of results. The offset used was found using the methods described previously for each library and the median offset of all involved replicates was used for each comparison. This allowed us to rank sequences without concern for low-abundance reads disturbing the ranking.

To find important differentially expressed reads, we used a LOFC cut-off of 1, meaning that read counts needed to have either doubled or halved after accounting for the effects of the offset. This left us with a manageable set of reads to investigate further.

### Summarisation of differentially expressed sRNAs

Our method of differential expression allows us to robustly rank and group individual reads by their differential expression between different conditions. However, assessing differential expression of individual reads presents two problems. Firstly, several reads may derive from the same location of the genome or are simply slight nucleotide variants of one another. This can complicate the final list of differentially expressed reads, since the same sRNA can have various levels of differential expression. Secondly, validation by northern blot does not discriminate single read variants because its intensity is based on the sum intensity of all reads that match the probe.

To mitigate both of these issues, we grouped differentially expressed reads into single, merged sRNAs if they overlapped each other. A second round of differential expression analysis was conducted on counts derived from the sum of reads pertaining to each merged sRNA. We validated several differentially expressed miRNAs using northern blots.

## 5.4 Results

### 5.4.1 Quality check results

The general characteristics of each library was assessed using mapping quality scores and the distribution of various abundance statistics over the different size classes. These are shown in figure 5.1. The size class distribution of redundant counts reveals a peak at 22nt that corresponds to the presence of abundant miRNA reads for most of the samples. However, some of the replicates for LW and LQ conditions have a much lower 22nt peak in relation to the other size classes, which may indicate a problem with these replicates or a true downregulation of many miRNA sequences. The size class with the lowest count complexity is 23nt, especially for the EW condition. This does not coincide with the 22nt peak, which suggests there are also important, but less abundant, sRNA classes at the 23nt size class.

Although the number of mapped sequences vary between 2,000,000 and 6,000,000 (figure 5.1 (a)), the proportions of sequences mapping to certain annotations remain similar.

We used the Jaccard index to assess the similarity of composition of the top 10,000 sequences between libraries (figure 5.3). This provides an indication of how related or comparable one library is to another. The Jaccard index is described further in chapter 4. Because replicates are generally more alike to one another than libraries from different experiments, we expected the index between blocks of replicates (along the diagonal in figure 5) to be closer to 1 than away from the diagonal. This is the case for all treatments except LW samples, which appear to have a lower similarity between its top sequences. The replicate LW3 had a particularly poor similarity index for its top sequences compared to any other library, which agrees with the MAplot comparisons of the LW replicates seen in figure 5.2 (a).

The results of the quality check were used to remove samples and size classes that showed poor comparability. We removed replicates LW4 and LQ1 as well as reads with lengths above 27nt in order to help the assessment of accurate

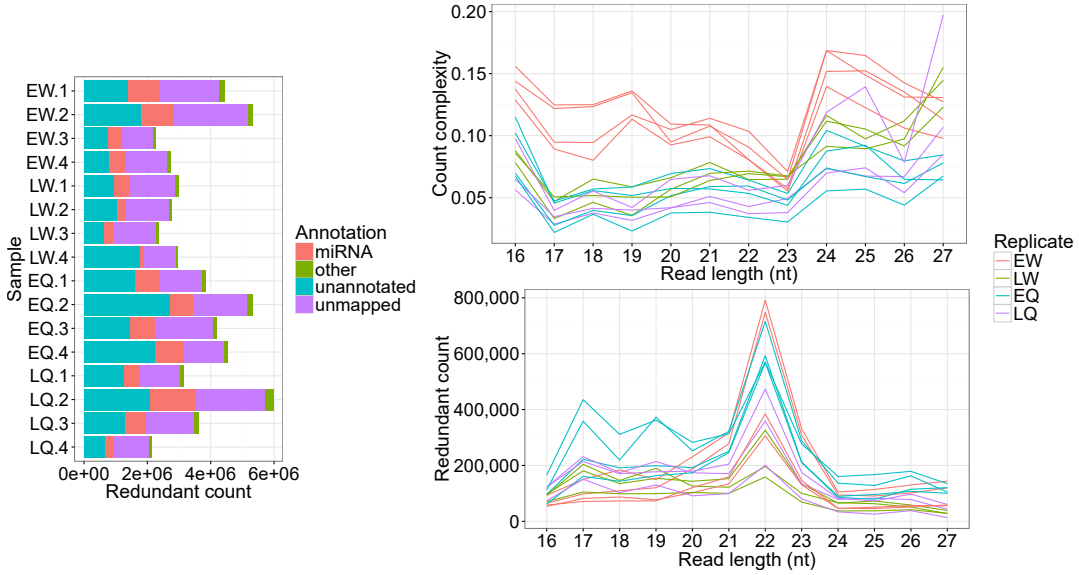



Figure 5.1: Characteristics of the sRNA-seq libraries. (a) Proportion of redundant reads that mapped to the genome (unannotated) and to tRNA and miRNA annotations. (b) Redundant counts and (c) count complexities of reads over size classes. Complexity is defined as the number of non-redundant reads divided by redundant reads. Both (b) and (c) only show replicates that were not removed at the quality checking step.

differential expression for the other size classes, which importantly includes the miRNA class.

### 5.4.2 miRNA identification

A total of 2,048 miRNA precursors were identified using a combination of the three tools. Figure 5.4 (a) shows the distribution of predictions when shared between the various tools. miRCat, using the default animal settings, predicted numerous miRNAs that were not found by the other tools. MapMi also identified 429 miRNA precursors not found by the two prediction tools but conserved in miRBase. However, the precursor lengths of these MapMi-only predictions suggest that these are not identified by the other animal-specific tools because their precursors far exceed the length assumed for an animal miRNA by both prediction tools. The miRDeep only precursors are predominantly very small, suggesting a tendency for miRDeep to find shorter precursors than are usually found in both miRBase or identified by miRCat. The precursors that multiple tools identify have precursors that are within the range of 60nt and 80nt.

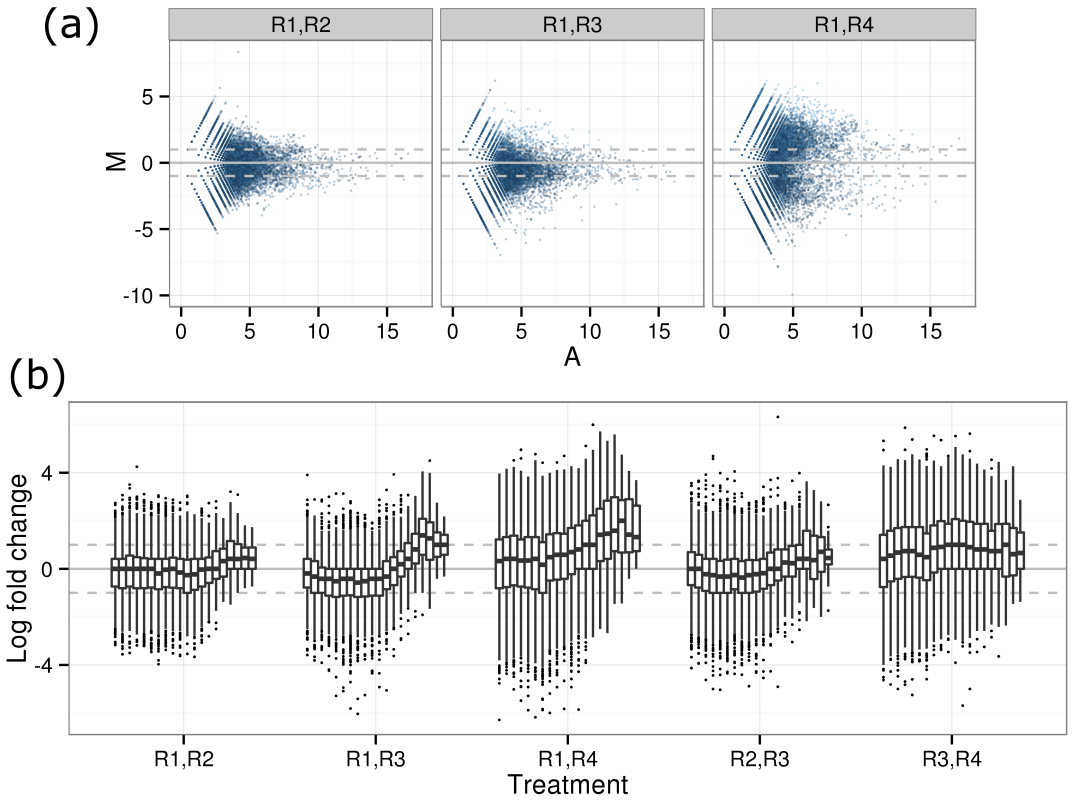



Figure 5.2: Replicate comparisons for the Late Worker treatment. (a) shows alpha-blended MA plots that indicate a skewed $log_2$ fold change distribution between replicates 2 and 3 and a highly dispersed distribution between replicates 1 and 4. (b) separates the distribution in to individual size classes, revealing the that the source of the issues are mostly from the largest size classes.

Figure 5.3: A symmetrical table showing Jaccard similarity indices for all library pairs between the top 500 sequences for each library. The similarity is measured by the Jaccard index where an index of 100 indicates that the two libraries share the same top sequences and an index of 0 indicates that none of the top sequences are shared between libraries.

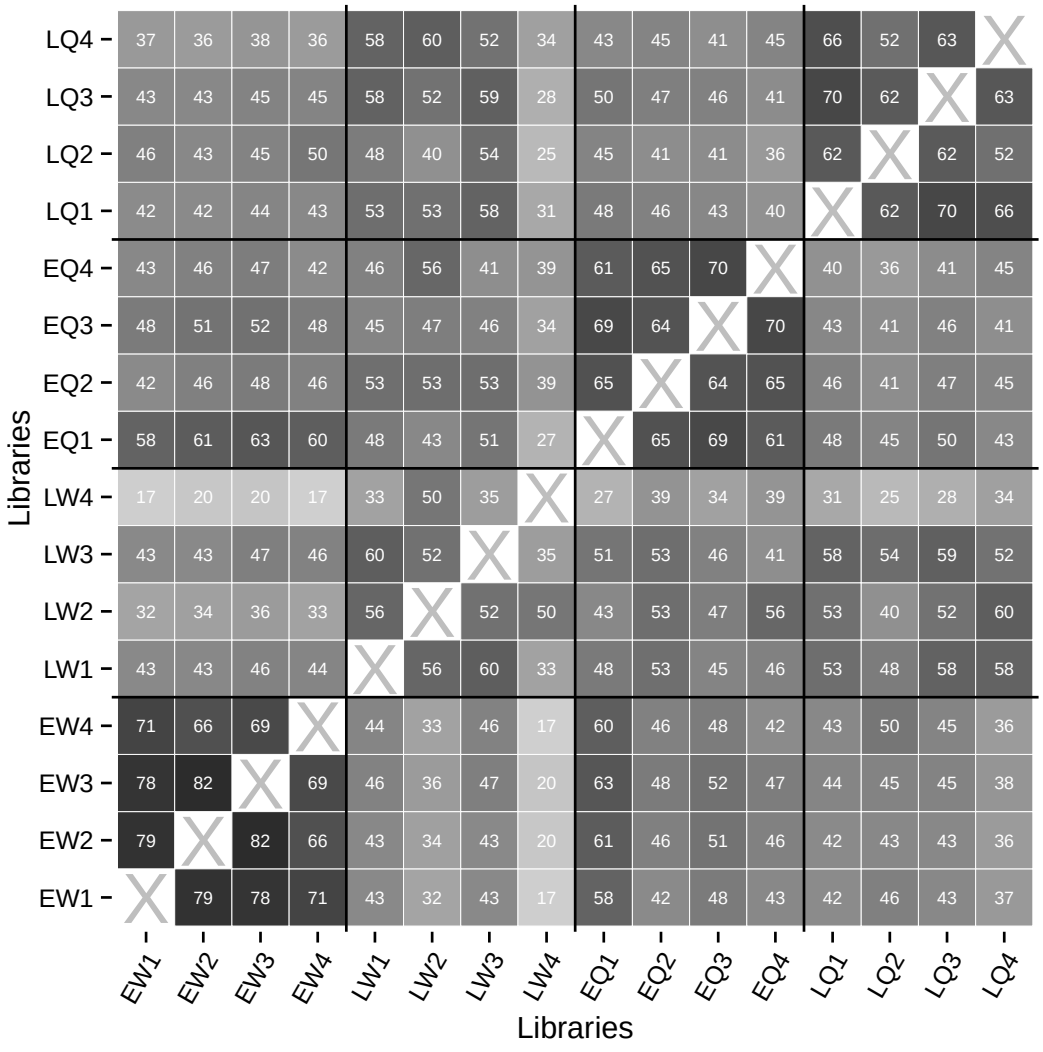
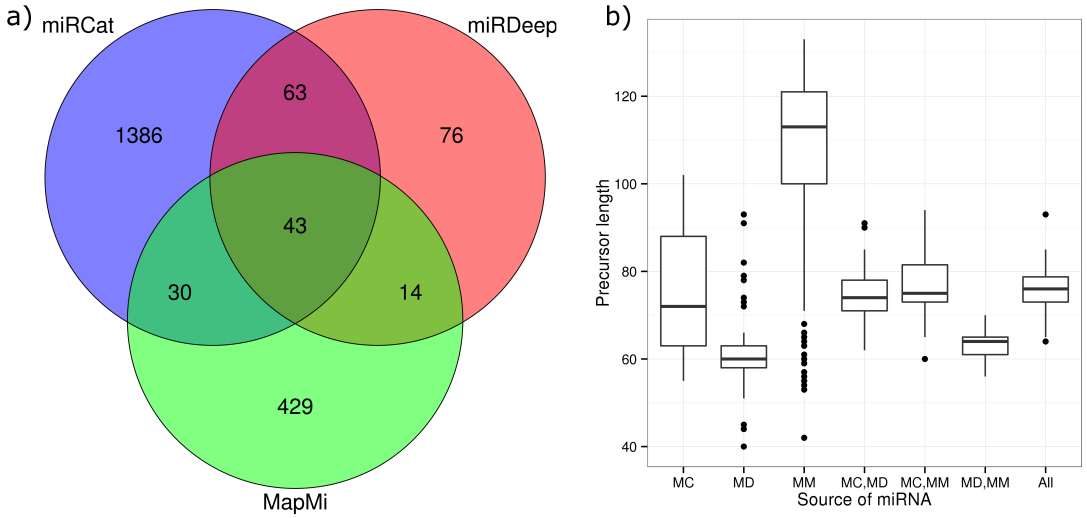


Figure 5.4: A summary of miRNA predictions. a) indicates the number of predicted precursors that were found by miRCat, miRDeep or conserved from miRBase using MapMi and the number of predictions shared by the results of these tools. b) shows the distribution of precursor sizes found by each tool or combinations of the tools. The x axis indicates which tools a particular distribution is for using the abbreviations MC (miRCat), MD (miRDeep), MM (MapMi), and "All" indicating that all tools identified these precursors.




### 5.4.3 Differential expression

After calculating LOFC values for all reads, any reads that were expressed by more than absolute 1 LOFC were summarised and evaluated to identify interesting sRNAs. Throughout the remainder of these results we call sRNAs expressed by more than absolute 1 LOFC as "differentially expressed".

The use of four different treatment comparisons allowed us to categorise differentially expressed reads based on their expression patterns over several variables. We looked at the correlation of related comparisons using the LOFC of all reads (figure 5.5 (a)). This showed a stronger correlation (Pearson coefficient of 0.56) between EW/LW and EQ/LQ comparisons compared to the correlation between EW/EQ and LW/LQ, which was not significant (Pearson coefficient of -0.05). Part of the reason for this difference is the difference in the number and amplitude of differentially expressed reads when going from a Worker sample to a Queen sample. However, several miRNAs are notably differentially expressed, both upregulated and downregulated, between the castes in the Late developmental stage.

A total of 47 miRNAs were differentially expressed in at least one comparison. 4 tRNA/rRNA reads and 11 other ncRNAs were also differentially expressed above a threshold of absolute 1 LOFC. The remaining 258 sRNAs were unanno-

tated. These reads corresponded to 26 upregulated sRNAs 42 downregulated sRNAs, which included 9 distinct downregulated sRNAs and two distinct upregulated sRNAs.

### 5.4.4 Identification of differentially expressed miRNAs

Ten differentially expressed sRNAs corresponded to miRNA annotations. All sRNAs were differentially expressed in at least one of the development comparisons. Five of these miRNAs were validated as being differentially expressed for these comparisons.

Out of the ten differentially expressed miRNAs, two miRNAs were also differentially expressed between Late Worker and Late Queen. These miRNAs corresponded to both arms of the miR-6001 precursor (figure 5.6 (a)), and have little to no abundance during Early stages but increase in abundance significantly more in Late Queen than Late Worker. Validation by northern blot confirms this pattern of differential expression (figure 5.6 (b)).

miR-6001 is a miRNA previously only identified in honeybees [Chen et al., 2010]. The precursor sequence is found within the fourth intron of predicted a predicted vitellogenin-6-like protein coding gene (protein accession number $XP_0$03400264.1). which is also conserved between bumblebees and honeybees.

## 5.5 Discussion

Although the evolutionary causes of eusociality in insects is generally understood [Bourke, 2011], the mechanisms used by larvae to develop towards specific roles or castes within a colony have been found to vary significantly, even between species of eusocial bee [Cardinal and Danforth, 2011]. In addition, the specific regulatory pathways behind caste determination is unclear [Smith et al., 2008]. In this sRNA-seq analysis we assessed libraries taken from an experiment on the regulation of miRNAs when larval and young adult Bumblebees undergo caste differentiation. The main aim was to identify conserved and new miRNAs that are involved in pathways for the development of these organisms into their distinct castes. However, since the reference genome was new and in a draft stage with very little annotation, a secondary aim was to identify as many miRNAs as possible that were either conserved from related species or otherwise unique to this organism. This was achieved through three different miRNA prediction tools. The resulting sets of miRNAs found by these tools suggested a disparity between the types of miRNAs found by each tool, especially when analysing the length

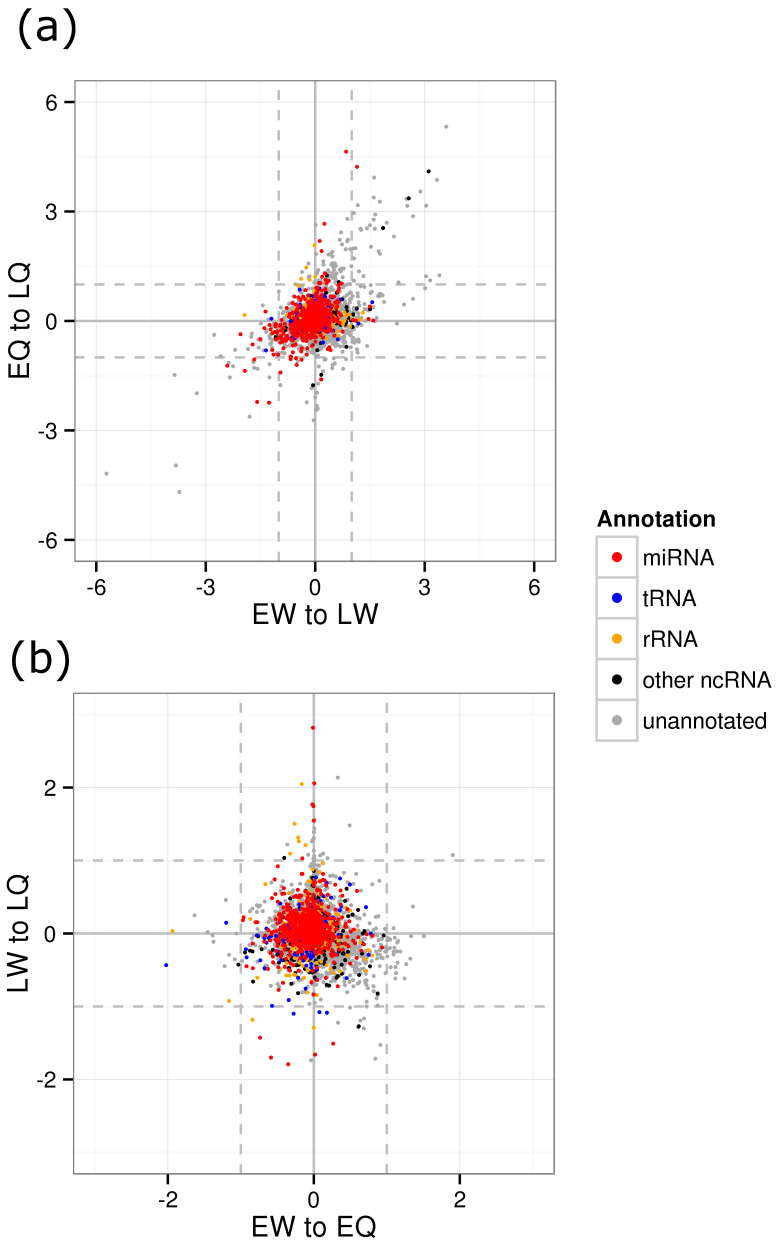
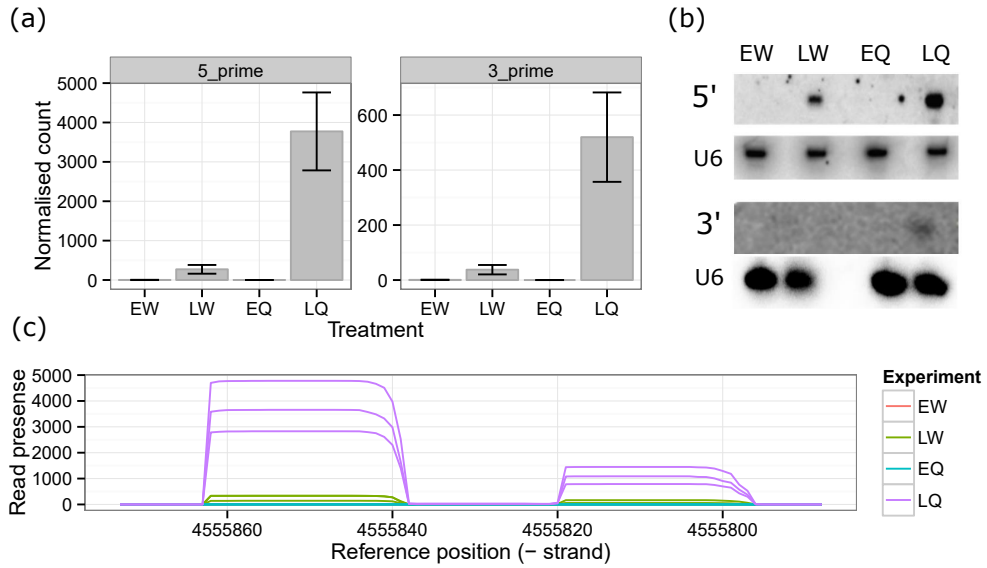



Figure 5.5: Cross plots of offset fold change results. The plots show the amount of LOFC between -1 and 1 in a 2D space created by plotting related comparisons against each other. (a) shows results in the space of Early conditions compared against Late conditions and (b) shows results in the space of Worker conditions compared to Queen conditions. The LOFC values are based on proximity comparisons, and any overlapping confidence intervals were assigned an LOFC of 0 for the purposes of visualisation. Note that miRNAs (in red) are plotted on top of all other annotation classes.

Figure 5.6: Differential expression and validation of both arms of miR-6001 over all four condition. (a) shows the total expression of reads associated with each arm of the miRNA, including confidence intervals. The results of a northern blot validating each arm are shown in (b). (c) is a presence plot of the miRNA precursor, indicating the total number of reads that cover each nucleotide for each condition.




of the precursor sequences. Conserved miRNAs tended towards having a longer precursor, and future work may involve tuning the parameter set of miRCat or miRDeep to take into account the ability for *B. terrestris* to have miRNAs with a longer precursor than assumed by the default parameter sets.

The large number of different treatments and replicates in this study facilitated the flexibility of the analysis. We were able to fully utilise the quality check stages outlined in chapter 4 after preprocessing and genome alignment of sRNA sequences to identify any replicates that were anomalous and decide on the best normalisation and library comparison strategy.

Based on a differential expression analysis that took into account all possible changes between conditions of the experiment, we found that differential expression of reads tended to be highly correlated between Workers and Queens throughout their developmental stages. When looking at differential expression between the castes, however, fewer reads were differentially expressed and the findings were not shared between development stages. Only the two arms of miR-6001 were significantly differentially expressed between Late development stages of Workers and Queens. The northern blot validations also showed only this miRNA as differentially expressed between castes, a finding which also val-

idates our differential expression methodology. Genome scanning revealed that the precursor sequence of mir-6001 could be found within the fourth intron of the predicted gene coding for a vitellogenin-6-like protein. Such miRNAs that derive from the introns of other genes have been termed mirtrons, and have mostly been found in mammals such as *Drosophola melanogaster* and *Caenorhabditis elegans* [Jan et al., 2011; Ruby et al., 2007]. The trans-regulatory functions of mirtrons do not differ from those of regular (non-mirtron) miRNAs; both are incorporated into the RNA-induced silencing complex (RISC) and target mRNA transcripts for silencing in the same way, but it is unclear whether their different modes of biogenesis have any cis-regulatory consequences related to their host gene [Westholm and Lai, 2011]. One interesting possibility is that the miRNA might be regulated by the upstream regulatory sequences of the host protein-coding gene itself. Therefore, the miRNA and the host gene would be co-expressed and affect the same pathways or phenotypes, in this case larval caste determination. Such a process has special relevance in the case of miR-6001 because it suggests a novel link between miRNA regulation of caste determination and vitellogenin. Vitellogenins are an important class of nutritive proteins induced by juvenile hormone and linked to reproduction in numerous insects [Sappington and S. Raikhel, 1998] and a storage protein in Hymenoptera including ants [Wheeler and Buck, 1995]. Storage proteins play key roles in insects that undergo metamorphosis, since they accumulate in late-instar larvae and are used in the rapid synthesis of amino acids prior to metamorphosis [Hunt et al., 2003]. This suggests that vitellogenin is a candidate for a caste-associated gene in eusocial Hymenoptera and that further investigations should focus on the potential link between the miR-6001 duplex expression and vitellogenin.

# Chapter 6

# Differential expression of small non-coding RNAs under cell stress

## 6.1 Summary

In the last chapter, we utilised our pipeline in conjunction with miRNA prediction tools in order to identify functions of conserved and new miRNAs in a novel organism. In that study, the ability to analyse a complete sRNA transcriptome allowed us to identify and calculate the differential expression of novel miRNAs. Here, we use the full transcriptome to understand the changes in expression of sRNAs that are derived from a large diversity of other ncRNAs during cell stress in organisms with a more robust set of annotations. Additionally, this new study highlights the technical challenges that can be faced when assessing datasets that represent highly divergent conditions with high rates of differential expression.

## 6.2 Background

As well as miRNAs, siRNAs, and piRNAs, sRNA sequences have been found to be produced from longer ncRNAs that have other primary functions. Such ncRNAs include Y RNAs [Hall et al., 2013], reviewed in chapter 2. Y RNAs are known to produce Y RNA-derived sRNA (YsRNA) sequences that are 22-32nt long following stress stimuli on cells and in the presence of two auto-immune proteins Ro60 and La. A similar response happens with tRNAs, where they are cleaved into smaller RNA fragments following cell stress in the presence of certain endonucleases [Thompson and Parker, 2009]. Cells respond to stress through

large changes in gene and RNA expression [Holcik and Sonenberg, 2005] and it is possible that other functional sRNAs may be produced and expressed as a result.

In this chapter, we analyse two sRNA sequencing experiments on the differential expression of sRNAs when mammalian cells are placed under stress using the immunostimulant Poly(I:C). This chemical stimulates a viral infection in cells; an environment where they undergo cellular stress and begin the process of apoptosis (cell death). The first experiment includes a mouse Ro60-/- mutant in order identify other sRNAs that are potentially dependent on Ro60 for expression. The second experiment, sequenced using HD adapters, attempts to understand the changes in sRNA transcriptome expression under cell stress of two human cell lines.

## 6.3 Materials

The first cell stress experiment was conducted on mouse cells with and without a Ro60 knockout background. Wild-type and Ro60-/- mouse embryonic stem (mES) cell lines were grown at 37°C in a 5% $CO_2$ humidified incubator.

Cells from both cell lines were exposed to Poly(I:C) treatment, creating four different conditions: wildtype (*wt*), Ro60 mutant (*ro60*), wildtype with Poly(I:C) treatment (*wt_pic*), and Ro60 knockdown with Poly(I:C) treatment (*ro60_pic*). Libraries were then pooled and sequenced using the HiSeq 2000 system (Illumina) with a 50 cycle read length. The sequencing was done using two different lanes for each experiment and three biological replicates. The dataset resulting from this experiment will be referred to as the Ro60 dataset.

The second experiment was conducted on two human cell lines: MCF7 and SW1353. Both conditions were alternatively treated with Poly(I:C), creating four conditions, a wild-type and poly(I:C) condition for each cell line, where each condition was biologically replicated three times. The library preparation protocol was identical to the previous study. For sequencing, the HD adapters were used. This dataset will be referred to as the cell line dataset.

## 6.4 Methods

### 6.4.1 Preprocessing and alignment

All libraries were trimmed for adapters by matching the first 8 nucleotides of the adapter sequence perfectly, where sequences with HD adapters were addi-

tionally trimmed by four nucleotides before the start and after the end of the adapter trimmed sequence. The processed reads were then aligned to their reference genomes (mouse genome GRCm38 available at `http://Dec2015.archive.ensembl.org/Mus_musculus/Info/Annotation` and human genome GRCh37 available at `http://grch37.ensembl.org/index.html`) using PaTMaN [Prüfer et al., 2008] with no mismatches or gaps.

We assessed the quality of mapped reads using quality checking methods detailed in chapter 4, including the use of MA plots and size class fold changes to assess within-condition replicate similarity.

Differential expression was carried out similar to the methods in Chapter 4 with a fixed offset of 20.

### 6.4.2 Annotation

To ensure we were able to annotate as many reads as possible, annotations were downloaded for the mouse and human genomes from several different databases. tRNA sequences were downloaded from tRNAdb [Jhling et al., 2009]. These were post-transcriptionally modified by removing any introns in the tRNAs, given by tRNAdb, and appending the CCA motif to the 3' ends of all sequences. Rfam11 [Burge et al., 2012] and the mature sequences from miRBase [Griffiths-Jones et al., 2006] were retrieved in FASTA format and all genome matching sRNA reads were subsequently re-mapped to these annotations using PaTMaN.

We also downloaded coding gene annotations sets for both human and mouse genomes in GFF format. This format was used in order to identify the specific feature that sRNAs may be mapping to within the gene models. To do this, we used BEDtools [Quinlan and Hall, 2010] to find overlaps between our set of mapped sRNAs and the set of gene models. We then categorised the matches as having derived from coding sequences (CDS), untranslated regions (UTR), or introns if the overlap was only found against the gene feature itself.

Coordinates for the two mouse Y RNAs Rny1 and Rny3 were taken from the MGI website by querying the two Y RNA identifiers. Any reads that mapped to these regions using PaTMaN were identified as Y RNAs.

### 6.4.3 Normalisation and differential expression

After assessing the results of total count normalisation and quantile normalisation, we proceeded with quantile normalisation for all datasets. To correctly normalise the cell line data, we separated the sequences by both the annotation

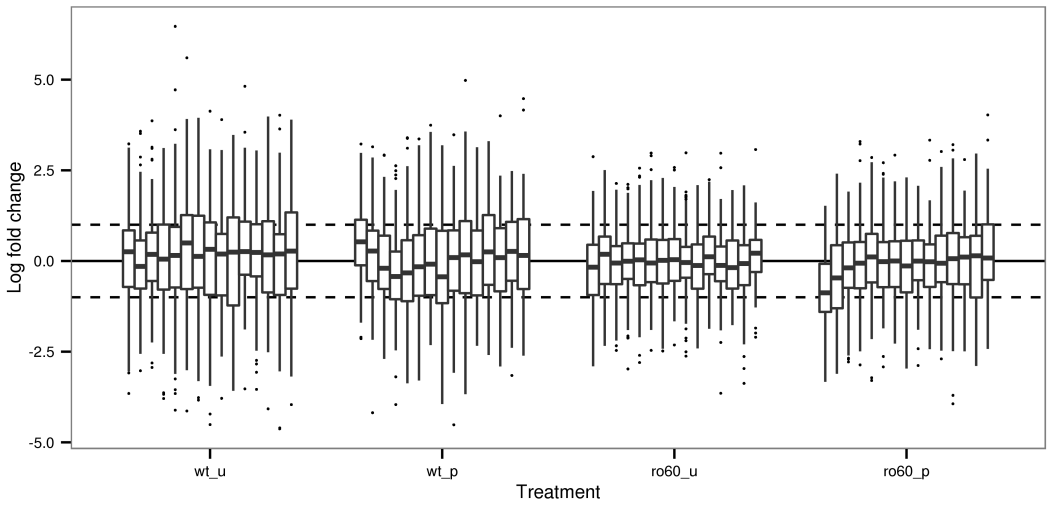



Figure 6.1: Log fold change distributions between the two remaining replicates for all conditions. The distributions are shown for each individual size class.

group and one of the two condition variables, depending on the comparison being investigated. Differential expression for all datasets was then calculated using the LOFC method (Chapter 4) with an offset of 20 for all samples. We applied LOFC to all comparisons between conditions that involved a change in only one of the condition variables e.g. cell line, Poly(I:C) treatment, or mutant phenotype.

## 6.5 Results

### 6.5.1 Quality checking and normalisation

To check the consistency of replicates in the Ro60 dataset, we visually assessed the within-sample fold changes using MA plots and the distribution of log offset fold changes for each size class. This showed large deviations from 0 fold-change for size classes less than 25nt. In order to allow accurate assessment of the larger size classes, where the ncRNAs of interest such as the Y RNAs are likely to be, we removed the lower deviating size classes (24nt or less) from the remaining analysis as well as the least similar replicate from each condition. This allowed quantile normalisation approach to normalise the remaining expression levels more accurately (figure 6.1).

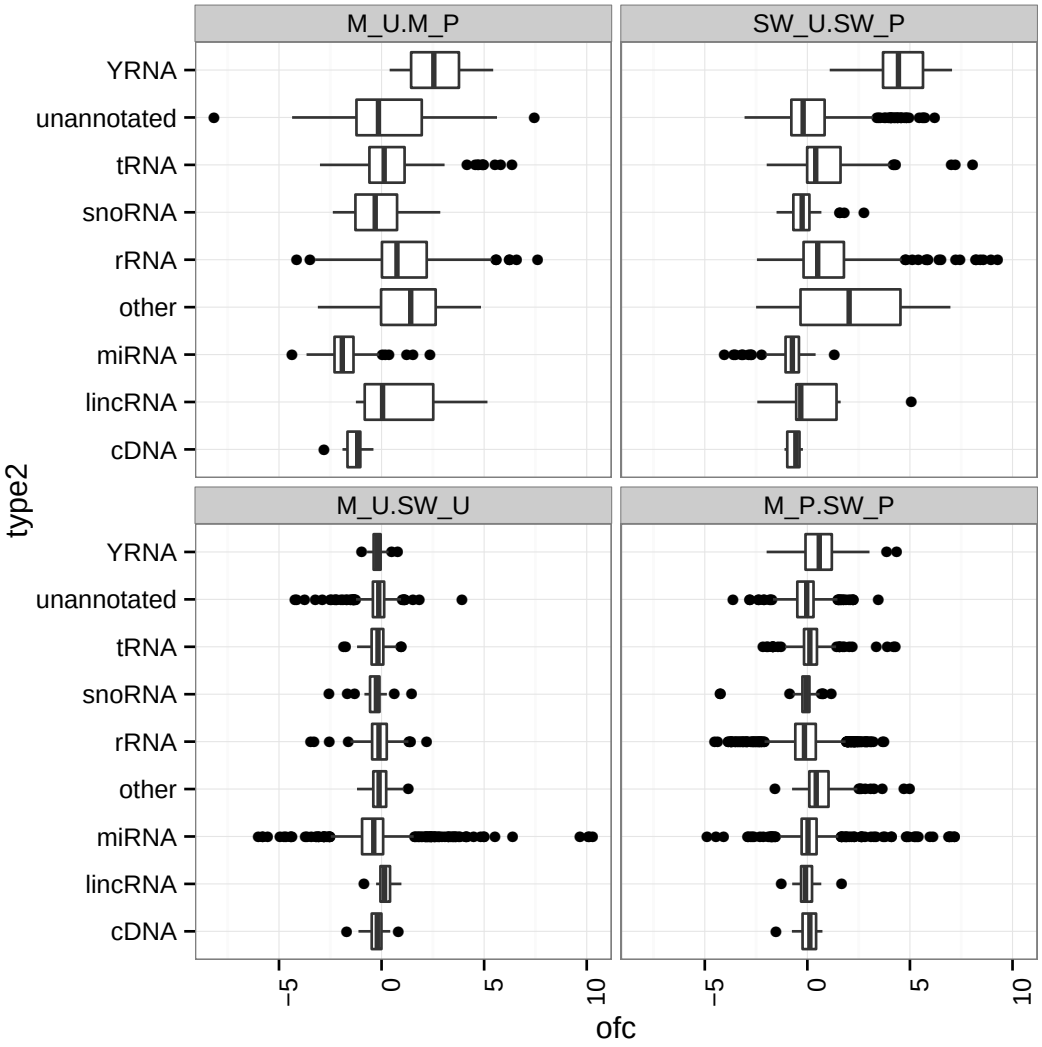
In contrast to the Ro60 datasets, the replicates of the cell line dataset showed a close agreement to one another for all conditions. However, acceptable normalisation of these libraries was prevented by a high disagreement of expression levels between both cell line conditions and stress phenotypes. To further un-

derstand this variation, we separated the libraries into their annotation sets as shown in figures 6.2 and 6.3. These show the distribution of LOFC values using the un-normalised data and reveal that the ncRNA annotations showed distinct differential expression distributions. For example, rRNAs were a highly numerous and upregulated class of ncRNA found in both cell lines whereas miRNAs appeared to be consistently downregulated but with a much smaller inter-quartile range for their distribution of fold changes. The pattern of fold changes found in the miRNA MA plot (figure 6.3) was particularly interesting because it is highly similar to the cone-shaped pattern produced in correctly normalised MA plots by non-differentially expressed sequences. We reasoned that, because so many rRNA sequences in the Poly(I:C) libraries are upregulated compared to the untreated libraries, they have taken up much of the sequencing "real estate" from the truly non-differentially expressed miRNAs. This suggested that a path to correctly normalise and identify truly differentially expressed miRNAs between untreated and Poly(I:C) conditions would be to separate these sequences from the rest of the library and normalise them on their own. This was attempted with both total count and quantile normalisation, selecting quantile as the method that produced the most centered fold change pattern of the miRNAs on 0 LFC.

### 6.5.2 YsRNAs are produced under stress only in the presence of Ro60

As YsRNA biogenesis is known to be dependent on Ro60, sequences derived from mY1 and mY3 RNAs were first assessed as a positive control to see if the data set could reliably be used to find other Ro60 dependent sRNAs. A presence plot was generated for both Y RNAs which plots the appearance of each nucleotide in all sequencing reads against its position in the genome (Figure 6.4). For the Y RNA gene Rny1, this plot shows a clear upregulation of YsRNAs at the 3' end of the gene between control and poly(I:C) treated wildtype cells, and further expression of a YsRNA at the 5' end in wildtype cells. The Ro60-/- samples, however, show very little or no expression of any reads along the length of the gene. The expression of Rny3 sRNAs is much lower and more variable with no convincing upregulation. Earlier Northern blot data showed that Rny3 and its YsRNA expression is generally much lower compared to the other Y RNAs and does not seem to be representative of YsRNA biogenesis. However, the mY1 presence plot did correlate with Northern blot analysis from Ro60 knockout experiments confirming that YsRNAs are dependent on Ro60. This in turn demonstrated that the sequencing data set could reliably be used to

Figure 6.2: Boxplots showing the distribution of LOFC values on the unnormalised cell line data for all assessed differential expression comparisons.



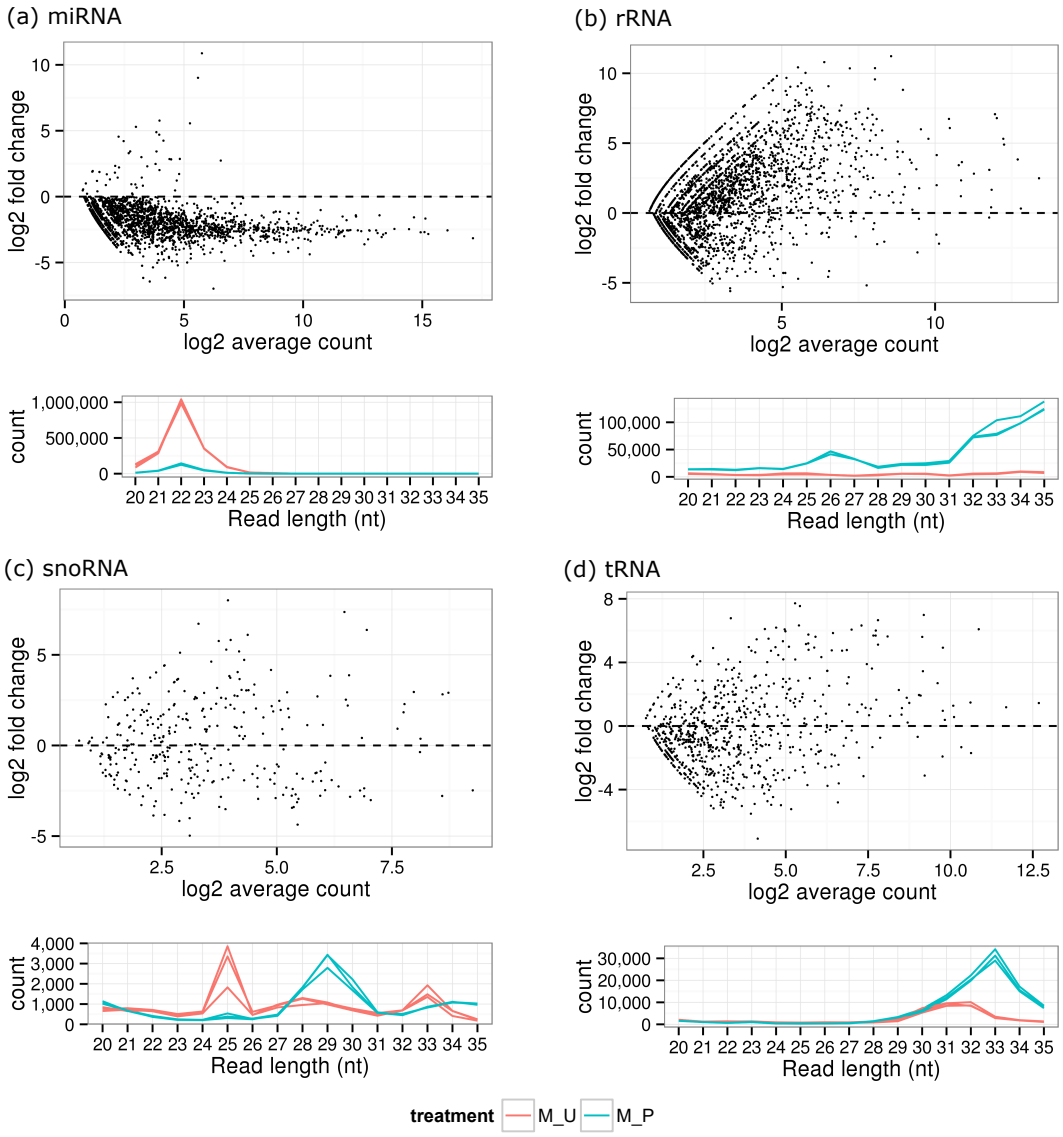
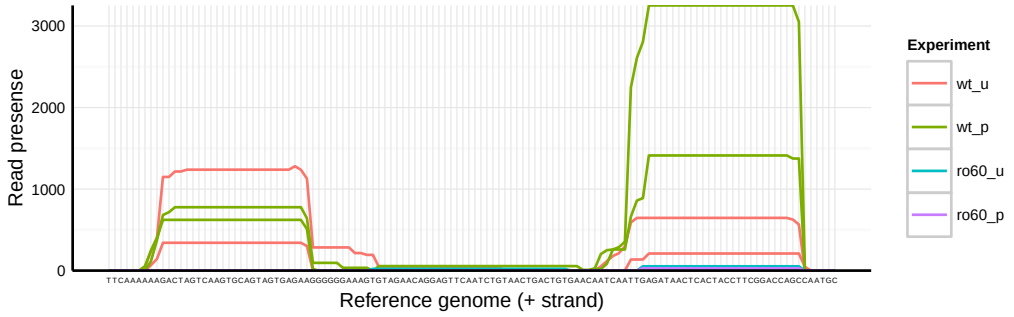
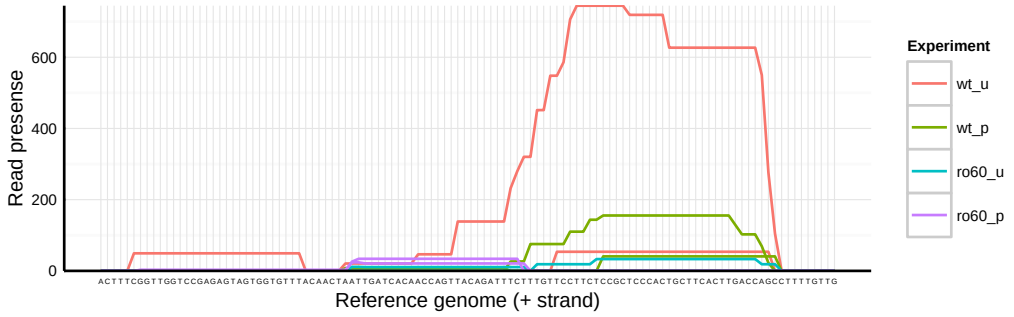



Figure 6.3: MA plots and size class distributions for selected individual annotation categories comparing untreated to Poly(I:C) conditions in the MCF7 dataset.




(a)




(b)

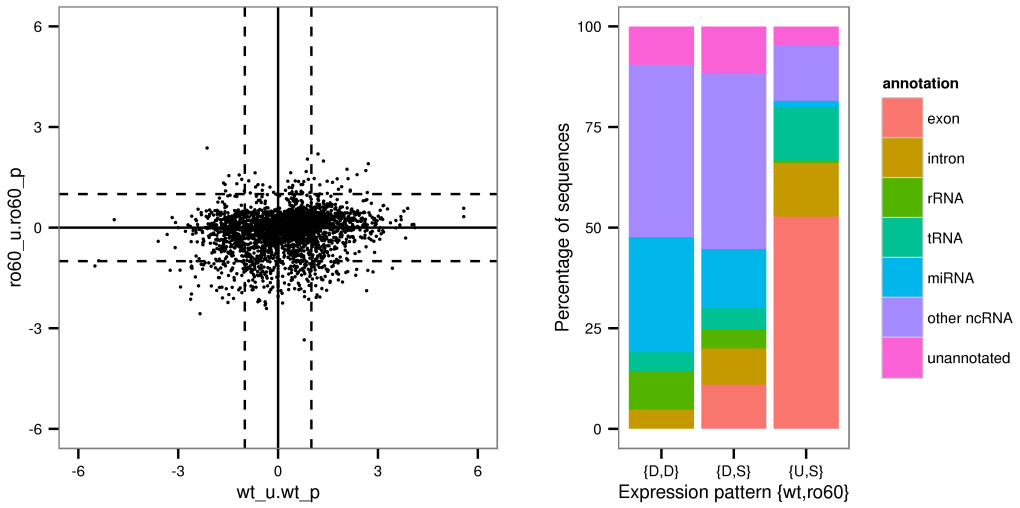
Figure 6.4: Presence plots for the coverage of (a) Rny1 and (b) Rny3 genes in the Ro60 dataset. Presence is calculated by summing the normalised expression levels of all reads that cover each nucleotide.

find other potential Ro60-dependent sRNAs.

### 6.5.3 Various ncRNAs are highly differentially expressed under stress

Calculating LOFC values for ncRNAs between the wild-type conditions indicates that there exists many ncRNAs that are differentially expressed in both directions (figure 6.5 (a)). This is very different to the distribution of LOFC values between the Ro60-/- conditions, which contains much less differential expression, leaving a large number of sequences that are only differentially expressed between the wild-type conditions. This indicates that many ncRNA derived sRNAs are only produced with the assistance of the Ro60 protein, although some ncRNAs are still found to be differentially expressed without it.

Sequences can be grouped into particular expression patterns according to how they were regulated in both comparisons between the unstressed and stressed conditions. These are denoted as $\{wt,ro60\}$ using (U)p, (D)own, and (S)traight




(a) (b)

Figure 6.5: Differential expression of annotated ncRNAs under stress. (a) Cross plot of the LOFC values of wt vs wt_pic against ro60 vs ro60_pic for all sequences that were regulated in at least one of the comparisons. (b) In the three largest expression patterns, the percentage of sequences that belong to each annotation group.

symbols, where *wt* indicates the unstressed to stressed comparison in the wild type cells and *ro60* indicates the unstressed to stressed comparisons in the Ro60-/- mutant. Figure 6.5 (b) shows a large difference in the proportion of sequences belonging to each annotation when split up by these expression patterns. The largest difference between the expression patterns is that exon-derived reads make up around 50% of the sequences upregulated in wildtype but unregulated in Ro60-/- (US). In contrast, only 7% of reads are derived from exons in DS and no exon-derived reads appear in the DD expression pattern.

### 6.5.4 miRNA regulation is more variable between cell lines than during cell stress

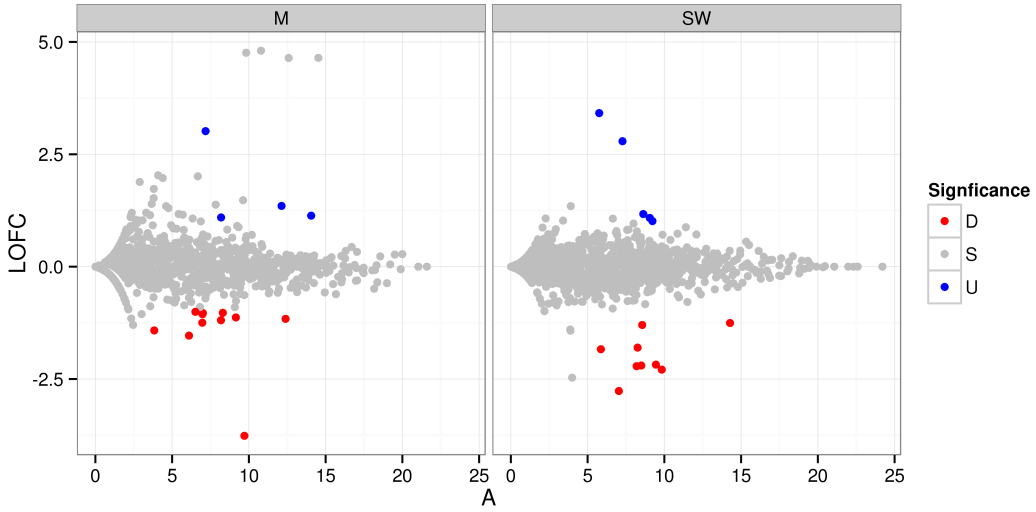
To most accurately assess any miRNAs that may be highly differentially expressed between untreated and Poly(I:C) treated conditions, we separated the dataset into libraries by cell line and normalised only the miRNA matching reads so that the majority of highly abundant miRNAs were found at 0 LOFC. This was achieved using quantile normalisation. We used the same approach to assess miRNA expression between cell lines for both unstressed and Poly(I:C) treated conditions, but in this case the normalisation appeared to have little effect because

the dispersion of fold changes was already very high. The final MA plots, after calculating differential expression using the LOFC method, are shown in figure 6.6. When comparing untreated to Poly(I:C) treated samples, we identified 23 miRNAs that were regulated above absolute 1 LOFC in at least one of the two comparisons, 6 of which were regulated as such in both conditions (table 6.1). Many more miRNAs were differentially expressed above absolute 1 LOFC when comparing cell lines in either treatment (31% for untreated and 24% for treated) than when comparing untreated to Poly(I:C) for either cell lines (about 1% for both cell lines). However, the high levels of differential expression that were apparent when comparing cell lines made these comparisons difficult to normalise.

Table 6.1: Mature miRNA LOFC levels between untreated and Poly(I:C) conditions for sequences found above absolute 1 LOFC in either MCF7 or SW1353. If a sequence was only found above this level in one cell line, the expression level is shown for the other cell line but it is designated as being (S)traight regulated.

| miRNA | MCF7 | SW1353 | Pattern |
|---|---|---|---|
| miR-1246 | 3.02 | 2.79 | UU |
| miR-1260b | -3.76 | -2.29 | DD |
| miR-1268a | -1.06 | -0.46 | DS |
| miR-1268b | -1.03 | -0.48 | DS |
| miR-145-3p | -0.80 | -1.80 | SD |
| miR-149-3p | -1.42 | -0.78 | DS |
| miR-181b | 1.13 | 0.88 | US |
| miR-181b-5p | 1.35 | 1.08 | UU |
| miR-221-5p | -0.76 | -1.25 | SD |
| miR-222-5p | -0.12 | -1.84 | SD |
| miR-23a-5p | -1.19 | -2.18 | DD |
| miR-23b-5p | -1.25 | -2.77 | DD |
| miR-2478 | 1.10 | 1.17 | UU |
| miR-27b-5p | -1.03 | -0.75 | DS |
| miR-29b-1-5p | -0.41 | -2.20 | SD |
| miR-29b-5p | -0.39 | -2.21 | SD |
| miR-3184-3p | -1.13 | -0.55 | DS |
| miR-365-5p | -1.00 | -0.60 | DS |
| miR-371-5p | 0.29 | 1.01 | SU |
| miR-423-5p | -1.16 | -0.54 | DS |
| miR-423a | -1.13 | -0.55 | DS |
| miR-4286 | -1.53 | -0.48 | DS |
| miR-4485 | -0.16 | 3.42 | SU |

(a)




(b)

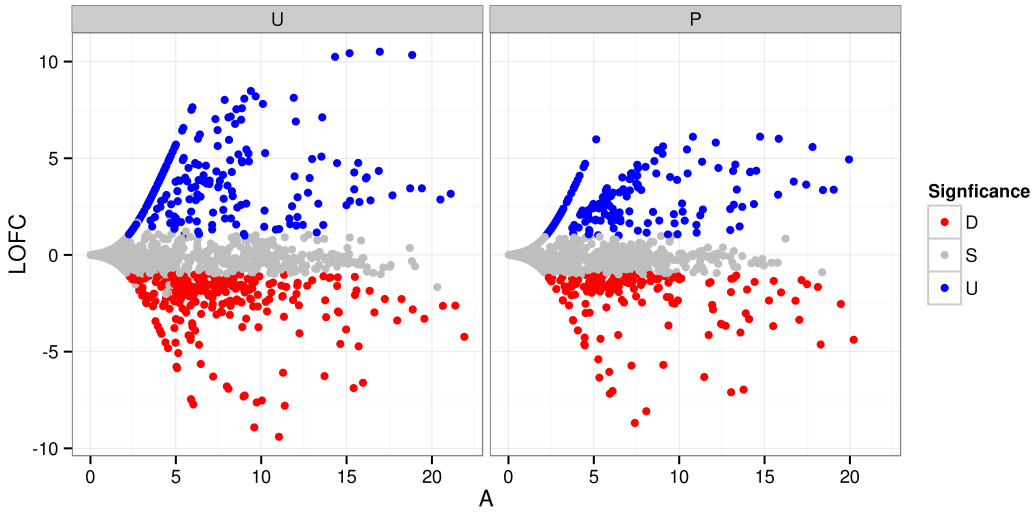



Figure 6.6: LOFC analysis of normalised miRNAs comparing treatments Unstressed to Stressed in MCF7 datasets (M), Unstressed to Stressed in SW1353 (SW) datasets, MCF7 to SW1353 cell lines in Unstressed datasets (U) and MCF7 to SW1353 cell lines in Poly(I:C) treated datasets (P). Colours indicate miRNAs that do not have overlapping confidence intervals and have a proximate LOFC above and below 1.

### 6.5.5 Differentially expressed mRNA fragments reveal a notable splice site motif

Owing to the drastically different annotation proportions assigned to gene features from USD expressed sRNAs, we further investigated the location, mapping characteristics, and sequence motifs of the gene-derived sRNAs found in this expression pattern group. After further grouping the reads in to sRNAs made up of closely neighbouring and overlapping differentially expressed reads, we further incorporated any remaining non-differentially expressed reads and recalculated the LOFC of these sRNAs. In doing so, we selected 24% of the sRNAs created that still maintained the desired expression pattern after combining the read expression levels, and selected 32 of the sRNAs for further examination that showed an overall differential expression of greater than 1 LOFC for each comparison. The median length of these sRNAs was 35nt. Each was almost always composed of a single length of closely overlapping sRNAs, distinctive of sRNA processing from a larger transcript. These mapping patterns are shown in the presence plots of figure 6.7. Each sRNA also has a distinctive slope in expression at the 5' end made by the production of variable length sRNAs, and a sharp drop in expression at the 3' end where all sRNAs end at the same location. This is similar to the mapping characteristics that many miRNAs have, where the conserved end of the mature miRNA is more stable than the other end. After examining the sequences, we also noticed a common motif to many of the sRNAs at the processed 5' end, where there exists a span of three to four T nucleotides. This can be clearly seen in logo plots when the sequences are aligned based on the largest increase in expression between two nucleotides at the 5' end of the presence plots (figure 6.8).

## 6.6 Discussion

In light of the increasing understanding of the expanded sRNA transcriptome, we carried out several experiments to understand the regulation of sRNAs, including those derived from ncRNAs with other functions, when cells are placed under stress. The first experiment, a study into the importance of the Ro60 protein in sRNA biogenesis, revealed a diverse set of sRNAs which rely on the interaction with Ro60 to be expressed. Interestingly these included a large number of 30nt sequences derived from within exons, which are predominantly found with a repeating T motif at their 5' end, suggesting a sequence recognition mode of splicing as part of their biogenesis. Interestingly, the La protein, which forms the

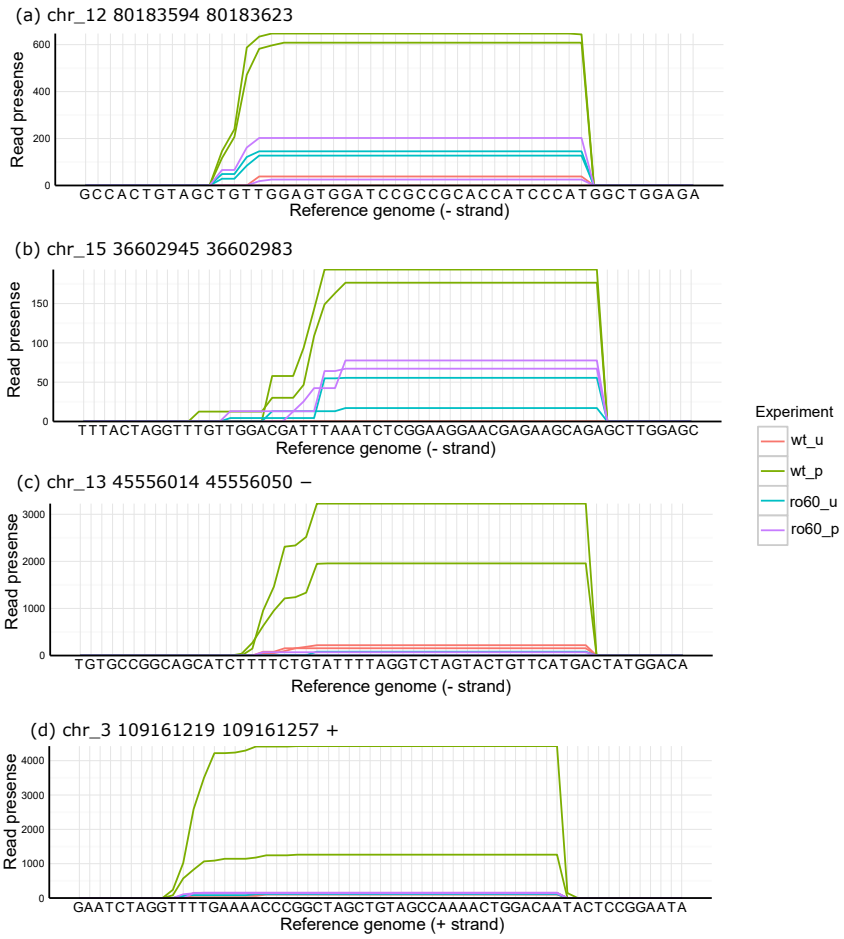



Figure 6.7: Presence plots showing the coverage of each sRNA that showed a USD pattern for differential expression. Each line represents the coverage for a specific sample. Plots are always shown 5' to 3'.

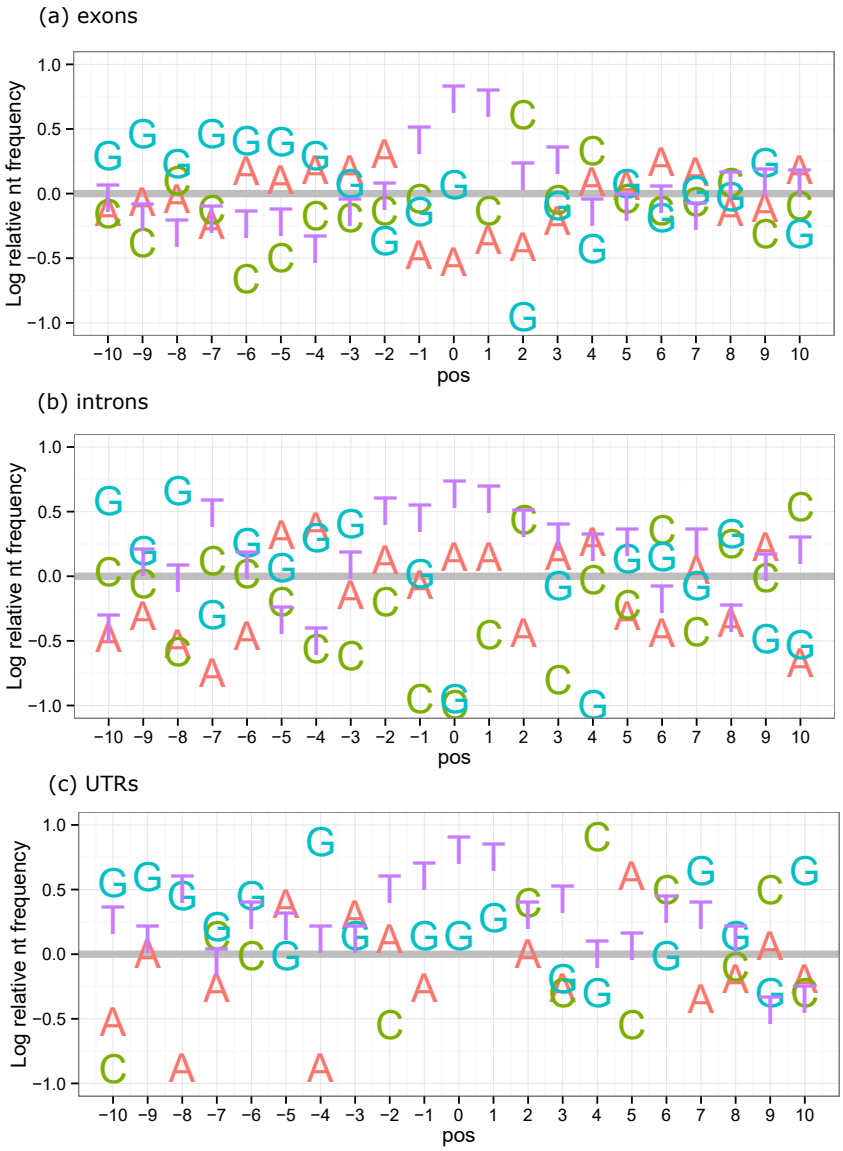



Figure 6.8: "Berry logos" showing sequence motifs for sequences that were aligned based on the most likely splice site location (at position 0) of gene-derived sRNAs.

roRNP complex with Ro60 and Y RNAs, primarily binds to poly(U) recognition sites on RNA molecules [Gottlieb and Steitz, 1989; Rinke and Steitz, 1982; Wolin and Cedervall, 2002]. Future experiments may be carried out to ascertain why such sRNAs are recruited by the roRNP complex in the same fashion as Y RNAs. One possible function is that, when cells undergo stress, Ro60 might be recruited to sRNAs which arise due to aberrant splicing errors because they may resemble misfolded RNAs which Ro60 has been shown to regulate [O'Brien and Wolin, 1994]. Alternatively, because these exon-derived sRNAs are essentially mRNA fragments and could therefore be remnants of transcript splicing, Ro60 might enter the nucleus following stress and modulate splicing of a subset of genes.

The second study, sequenced at a higher depth using the more accurate HD adapters, revealed a large amount of differential expression when two human cell lines are placed under cellular stress. This also demonstrated a potential difficulty when attempting to accurately analyse the differential expression of a sample where few high abundance read counts are not differentially expressed. In such a scenario, there is no baseline with which to normalise the samples to and it may not be possible to identify sequences with true differential expression versus differential expression as a result of losing sequencing space to a large population of differentially expressed sequences. The low variance of fold change that miRNAs showed at high log average abundance did however suggest that the majority of highly abundant miRNAs were not differentially expressed between unstressed and Poly(I:C) treated cells in either cell line. Whilst we did identify some miRNAs with a lower abundance that were differentially expressed as a result of this analysis, wet lab experiments to validate these miRNAs have yet to be finalised. Future studies between conditions with very different transcriptomes should include a quality check stage to determine an accurate zero baseline, potentially using sequences that are known to be not differentially expressed between conditions, or otherwise artificially spiking the data.

# Chapter 7

# Identification of small RNAs in microalagae

This chapter is adapted from Lopez-Gomollon S, Beckers M, Rathjen T, Moxon S, Maumus F, Mohorianu I, Moulton V, Dalmay T, Mock T "*Global discovery and characterization of small non-coding RNAs in marine microalgae*", BMC Genomics, 15:697, 2014.

## 7.1 Summary

The aim of this study is to identify sRNAs within the transcriptomes of two diatom species by methodical analysis of sRNA high-throughput sequencing datasets. Very little is known about sRNAs in these species, so this study was a data mining exercise that identified patterns within the sRNA libraries and attempted to explain these patterns.

## 7.2 Background

Diatoms are unicellular, photosynthetic phytoplankton that are dominant within both freshwater and seawater ecosystems where they form the basis of many food webs [Armbrust et al., 2004]. They are currently classified within the Chromalveolata supergroup of eukaryotes as a group of heterokonts. Model diatoms include *Phaeodactylum tricornutum*, *Thalassiosira pseudonana*, and *Fragilariopsis pseudonana*.

The evolutionary ancestor of chromalveolates is thought to have formed from a secondary endosymbiotic event between a photosynthetic eukaryote, the ancestor of land-plants, and a heterotrophic eukaryote. As a result, members of this lineage

contain plastids surrounded by four membranes [Falciatore and Bowler, 2002], as opposed to three in Plantae, and possess a mosaic genome consisting of genes that are othologous to both animal and plant lineages. Furthermore, approximately 5% of the genome of the diatom *P. tricornutum* consists of genes that have orthologs in bacteria, suggesting that diatoms take part in substantial horizontal gene transfer with bacteria. This 'soup' of different genes has resulted in its fast divergence from other eukaryotic lineages, namely the Archeaplastida (plants, green algae, and red algae) and the Opisthokonta (animals and fungi) [Armbrust et al., 2004].

There is also evidence of fast divergence between the heterokonts themselves that is faster than within plant, animal, or fungi lineages. Bowler et al. [2008] found that the pennate diatom *P. tricornutum* shared just 57% of its genes with *T. pseudonana*, a centric diatom. This relatedness is similar to the degree of divergence between fish and mammals, which started around 550 million years ago.

### 7.2.1 Current sRNA research in micro-algae

The rate of evolution of diatoms and other unicellular chromalveolates opens up an interesting question. sRNAs have been identified in most major eukaryotic lineages, but how conserved might the silencing mechanism be, if it is at all present, in a rapidly diverging lineage such as those of diatoms? Studies within the relatively newly sequenced genomes of several diatoms are beginning to uncover evidence for a possible silencing mechanism. However, the extreme differences between diatoms and other organisms present a challenge to sRNA sequencing technology, since methods used on plants or animals are not guaranteed to work with the genomes of micro-algae.

#### RNAi machinery in algae

Searches for homologs of key RNAi proteins in some chromoveolates have shown that there are highly divergent Dicer and Argonaute proteins within these organisms. These are summarised in table 7.1 Some were so divergent that they could not be assigned with confidence [Cerutti et al., 2011]. RdRP proteins were also present but had a very limited distribution amongst algae.

An analysis of the conservation of RNAi machinery in all eukaryotes by Cerutti and Casas-Mollano [2006] included the *T. pseudonana* genome in its draft stage. A homolog of Argonaute was identified, but no homologs of Dicer or RdRP were

Table 7.1: A summary of identified homologues to components from the RNAi pathway in the diatoms *T. pseudonana* and *P. tricornutum*.
[1]Norden-Krichmar et al. [2011], [2]Riso et al. [2009]

| | *T. pseudonana* | *P. tricornutum* |
|---|---|---|
| Dicer-like | PAZ, RNaseIII, RNaseIII [1]<br>DEAD, Hel-C, PAZ, DSRM [1]<br>RNaseIIIa, RNaseIIIb [2] | dsRBD, RNaseIIIa, RNaseIIIb [2] |
| Argo-Piwi | PAZ, PIWI [1,2] | PAZ, PIWI [1,2] |
| RdRP | RdRP [1,2] | RdRP [2] |

found. Argonaute or Piwi proteins were the most conserved elements of the RNAi pathway in eukaryotes, with presence noted in every eukaryote that has shown RNAi abilities.

Norden-Krichmar et al. [2011] identified several proteins that contained Dicer domains but no protein that had a complete set of Dicer domains. However, one homolog contained a PAZ and two RNaseIII domains, similar to the functional Dicer of *Giardia intestinalis*. No evidence of Drosha was found, which has only been identified in mammalian lineages. One Argonaute homolog was identified containing both PAZ and PIWI domains [Cerutti and Casas-Mollano, 2006; Cerutti et al., 2011; Riso et al., 2009], but no protein was found that included all domains from plant Argonaute homologs.

Evidence for RNAi machinery thus remains allusive in diatom genomes. Whilst RNAi proteins have been identified that function despite the lack of certain domains, it is also true that these proteins are used for other functions in the cell.

### Small RNA identification in algae

There have been very few studies on sRNAs within algae to date. Three of these studies have been on diatoms *P. tricornutum* and *T. pseudonana* and the rest on other algae, both highly related and less related to diatoms (see table 7.2). No sRNA analysis has been completed on *F. cylindrus* or the coccolithophore *Emiliania huxleyi*.

Table 7.2: Summary of the six papers that identified putative miRNAs in diatoms and related species.

| **Paper** | **Species** | **Sequences identified** | **Mature Lengths** | **Hairpin Lengths** | **Comments** |
|---|---|---|---|---|---|
| Lu and Liu [2010] | *P. tricornutum*; *T. pseudonana* | 6 novel hairpin candidates (5 in PT, 1 in TP) | 20-21 | 45-57 | No conservation between diatoms. Precursor sequences radically different to homologs of mature miRNAs |
| Norden-Krichmar et al. [2011] | *T. pseudonana* | 29 novel hairpins | 18-24 | 70-132 | dataset enriched at 28-32nt. No conservation with miRBase |
| Huang et al. [2011] | *P. tricornutum* | 13 novel hairpins | 18-25 | 101-260 | dataset normalized enrichment around 22nt |
| Cock et al. [2010] | *E. siliculosus* | 26 novel hairpins | 21-23 | 78-152 | Targets recently evolved leucine-rich domains involved in regulating multicellualrity |
| Liang et al. [2010] | *P. yezoensis* | 15 homolog miRNAs with high read counts, 1 novel hairpin | 21-22 | 66-251 | novel miRNA not conserved |
| Molnr et al. [2007] | *C. reinhardtii* | 21 novel hairpins, 47 longer hairpins | 18-24 | less than 150; 150-729 | Long hairpins generated phased sRNAs |

Lu and Liu [2010] attempted computational prediction of miRNAs based on EST sequences from *P. tricornutum* and *T. pseudonana*. miRNA homologs in plants were identified within the ESTs and 6 novel miRNA candidates were identified based on their ability to fold into hairpin structures; 5 in *P. tricornutum* and one in *T. pseudonana*. There was no conservation of mature miRNAs between diatoms. In addition, although the mature sequences were highly conserved with other miRNAs in the family, the precursors of the candidates were radically different. It was suggested that, since the RNAi machinery in diatoms are poorly conserved, there may be a large difference in miRNAs of diatoms too.

Norden-Krichmar et al. [2011] used two sequencing technologies in order to compare and contrast the usefulness of both SOLiD deep-sequencing and the more accurate 454 sequencing analyses. The 454 dataset showed an unusual enrichment of 28-32nt unique reads. The SOLiD dataset had a flatter frequency with a bias at either end of the size class spectrum. miRNAs were only predicted for a characteristic size range of 18-24nt, and the 29 candidates found originated exclusively from the SOLiD library. The predicted precursors varied in length around 100nt and all candidates lacked conservation with other miRNAs in miRBase. The more highly represented miRNAs were not detected in controlled northern blot experiments.

Huang et al. [2011] used Illumina sequencing to identify sRNAs in *P. tricornutum* under both Nitrogen limiting and silicon limiting conditions with a third control dataset. In contrast to Norden-Krichmar et al. [2011], the unique size class distribution of reads was enriched at around 22nt. However, the entire distribution represents a bell curve, which may have been due to the distribution of reads across the portion of gel that was cut from the size fractioning analysis. Thirteen novel miRNAs were predicted from all datasets using Mfold [Zuker, 2003] with mean precursor lengths at 235nt, none of which were conserved in other organisms. When two of the candidates were analysed by northern blot, only longer precursors were detected, suggesting the reads may be part of a longer degraded transcript.

Norden-Krichmar et al. [2011] also assessed the affinity of sRNA reads to transposable element regions. 2% of the *T. pseudonana* genome contained repeats, and as many as 15% of the sRNA reads mapped to them, suggesting a possible silencing pathway for transposable elements. The sRNA reads also tended to cluster along the genome, creating possible sRNA hotspots. The majority of other sRNA reads were produced from just a few hot-spots.

### Small RNAs in related algae

Perhaps the closest related model organism to diatoms is the brown alga *Ectocarpus siliculosus*. The brown algae split from the diatom lineage relatively recently and evolved multicellularity, including related genes and pathways, independently. Cock et al. [2010] analysed a fully sequenced genome of the model organism and computationally identified ncRNAs, including snoRNAs and sRNAs. Twenty-six sRNA sequences were found to match the required parameters for miRNAs, but none of these candidates were experimentally validated. The sequences preferred to target leucine-rich repeat domains of recently evolved genes, suggesting that the miRNA candidates may have evolved to regulate processes involved with multicellularity. In addition, other sequenced sRNA reads were found to significantly map to transposons, suggesting silencing pathways targeted at controlling transposable element activity.

miRNAs have also been predicted for a second multicellular seaweed; the red algae *Porphyra yezoensis* [Liang et al., 2010]. In this study, a sRNA library was prepared using Illumina in an attempt to identify young, lesser expressed miRNAs as well as possible mature miRNAs. 33,324 miRNA orthologs were identified, 15 of which had relatively high read counts. Comparisons between other species showed 16 miRNAs that were conserved between *P. yezoensis* and *C. reinhardtii*. Novel miRNAs were also computationally predicted using available EST data. Reads that were not identified as other ncRNAs were used to predict hairpin structures with minimum free energies (MFE) from -86.2 to -22. Only one miRNA had a MFE level below -25. These predicted miRNAs had mature sequences that were 21 or 22nt. They were not conserved with any other species. No experimental validation was attempted on any of the candidate miRNAs.

*Chlamydomonas reinhardtii* is a green algae that is much less related to the diatom lineage. A complex set of miRNAs and small RNAs, including phased siRNAs, were identified in *C. reinhardtii* [Molnr et al., 2007]. Unlike *E. siliculosus*, *C. reinhardtii* is single-celled, which contradicts the idea that miRNAs are mostly required for regulating multicellular processes [Casas-Mollano et al., 2008]. 21 miRNAs were found that could form a stable precursor loop. The miRNAs were similar to higher plants and animals. 47 other longer miRNAs were found that gave rise to phased siRNAs, which was proposed to represent young miRNAs in the process of evolving. None of the miRNAs found were conserved between *C. reinhardtii* and plants. In 8 of the examples tested, the expression of the mature miRNA candidate could be validated by northern blots. *C. reinhardtii* so far represents the only unicellular organism that uses miRNAs to regulate pathways.

Table 7.3: Genome and gene data sources for the species used in this analysis.

| Species | Version | URL |
|---|---|---|
| *T. pseudonana* | v3.0 | http://genome.jgi-psf.org/Thaps3/ |
| *F. cylindrus* CCMP 1102 | v1.0 | http://genome.jgi-psf.org/Fracy1/ |
| *E. huxleyi* 1516 | v1.0 | http://genome.jgi-psf.org/Emihu1/ |

## 7.3 Methods

### 7.3.1 Library preparation and preprocessing

Total RNA was extracted from *T. pseudonana*, *F. cylindrus* and *E. huxleyi* cultures and sequenced using Illumina Genome Analyzer II at TGAC (Norwich, UK).

The libraries were received as FASTQ files containing reads of 50nt including the 5' adapters. These adapters were trimmed from the resulting sequences by matching the first 6 nucleotides of the 5' adapter exactly. Sequences where no adapter was found were discarded. This left a redundant set of sequences for each library with lengths between 16 and 44 nucleotides inclusive. We mapped the remaining sequences using PatMaN [Prüfer et al., 2008] with no mismatches to their respective genomes for *T. pseudonana* v3.0, and *F. cylindrus* CCMP 1102 v1.0. The 1217 strain of *E. huxleyi* was mapped to the current draft assembly of the 1516 strain v1.0. The data sources are given in table 7.3.

#### Search for miRNAs

To identify possible miRNAs within the sRNA libraries, we ran both miRCat [Stocks et al., 2012] and miRDeep2 [Friedlnder et al., 2012] on the mapped reads. Since diatoms are highly unrelated to the plant and animal lineages that have been most studied for miRNAs, we ran miRCat once with default plant parameters and a second time with the default animal parameters. The default parameters were also used for miRDeep. The resulting predictions were collated and manually curated. Mature sequences with an abundance of less than 100 were considered to be unreliable predictions due to their low level of expression, which is unlikely to show up upon validation. We assessed the remaining sequences by their expression pattern on the genome and folded structure.

miRNAs identified by [Norden-Krichmar et al., 2011] in *T. pseudonana* were also checked against the *T. pseudonana* library to identify similarities between these experiments.

### Search for locally significant expression patterns

It is possible to identify small RNAs based on their characteristic mapping patterns. Functionally related sRNAs tend to aggregate in clusters at genomic loci and it is often possible to cluster mapped sRNAs based on this feature (see chapter 3 section 3.6.2). Molnr et al. [2007] successfully utilised a proximity algorithm for clustering their dataset, however the depth of their libraries was far less. A similar approach with our data was found to produce significantly longer loci that may not be meaningful. For this analysis, a different approach was used, based on locally significant size class distributions on a sliding window.

To identify loci that contain locally significant expression patterns, reads were binned into 300nt windows across each genome. For each region, a chi-squared statistic was calculated, comparing the size class distribution, between 16 and 40 nucleotides in length, of a region against a uniform distribution. Regions were filtered if they contained less than 100 reads or if the chi-squared statistic was not significant ($P$ ¡ 0.05). The remaining windows were combined if they were adjacent.

### Analysis of sequences derived from other sources

To characterise the remaining sequences, an annotation pipeline was created that mapped sequences to different reference databases and combined the results, labelling each sRNA as derived from a particular annotation (figure 7.1).

tRNAscan-SE [Lowe and Eddy, 1997] was used with default parameters to find putative tRNAs across all three genomes. Because any reads derived from these tRNAs would reflect post-transcriptional modifications, we removed intron sequences from the predicted tRNAs and a 'CCA' motif was appended to the 3' end. We then mapped reads to the mature tRNA sequences using PatMaN with no mismatches.

sRNAs were also aligned to coding gene transcripts. Those that overlapped with annotation provided by JGI (see table 7.3) and labelled as derived from either the intron or exon.

sRNAs that did not map to tRNAs or genes were aligned to the Rfam database using BLAST [Altschul et al., 1990] in order to identify possible homologous ncRNA transcripts that were unannotated on diatom genomes. These included snoRNAs and rRNAs. To maximise the likelihood of identifying homologs based on extremely short queries, the word size was set to 7 and hits with an e-value of at most 10 were considered. Hits were accepted that had an identity of more than 80% along the full length of the read.

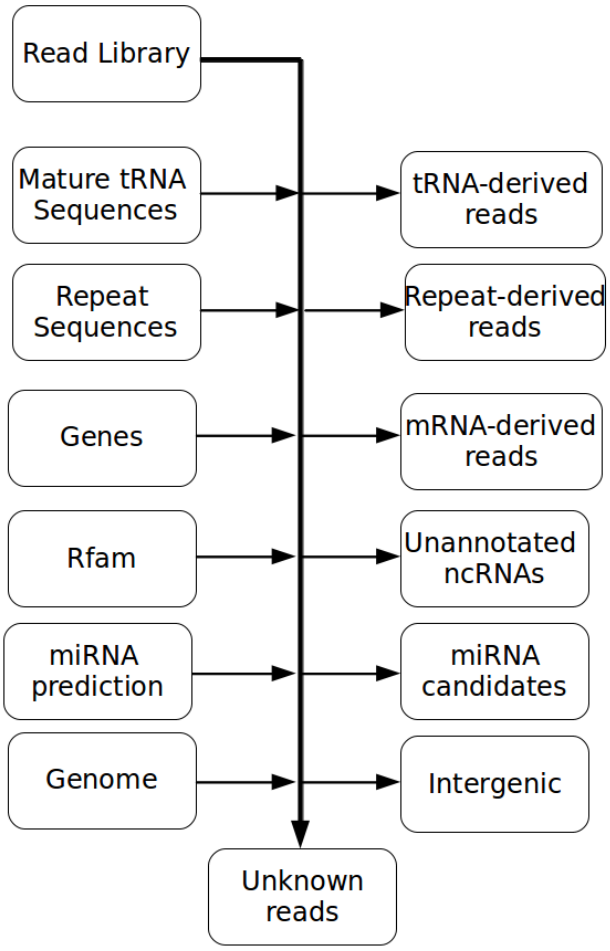



Figure 7.1: Schematic summarising the workflow used to annotate sequences

Table 7.4: The proportion of reads that could be mapped to the respective genomes for the three species.

| | *T. pseudonana* | | *F. cylindrus* | | *E. huxleyi* | |
|---|---|---|---|---|---|---|
| | Redundant | Unique | Redundant | Unique | Redundant | Unique |
| Total | | | | | | |
| Adapter trimmed | 63419714 | 5437751 | 40025978 | 1943536 | 83409723 | 577666 |
| After Mapping | 39755382 | 1473924 | 29006623 | 442026 | 12471322 | 80953 |
| Mapped | 62.7% | 27.1% | 72.5% | 22.7% | 14.9% | 14.0% |

Reads were classified as derived from at most one feature. Using this method, we ranked features in order of importance: tRNAs >repetitive elements >exons >introns >homologous ncRNAs.

## 7.4 Results

### 7.4.1 Library preprocessing

Table 7.4 gives the number of reads that could be processed from each library. Approximately two thirds of the reads could be mapped back to their respective genomes for *T. pseudonana* and *F. cylindrus*. However, when looking at unique reads this mapping percentage was much lower, suggesting that many low-abundance reads could not be mapped. Because very few *E. huxleyi* reads could be mapped to the 1516 strain's genome, the analysis of this library was not completed.

In all three libraries, sequences mapping to specific regions of the genome were removed from further analysis because their abundance totalled a large proportion of the genome, suggesting that these reads were subject to sequencing bias and did not represent real abundances.

After accounting for sequencing bias in each library, the length distributions we obtained are shown in figure 7.2. Canonical miRNAs in plants and animals are 21-23nt long and length distributions from such organisms will reflect this as peaks at these size classes. However, the distributions for both diatoms do not show any such peak. Other size classes are enriched, however. In particular, these include size classes around 27-30nt in both diatoms as well as some of the smaller size classes: 16nt and 19nt in *T. pseudonana*, and 16nt and 18nt in *F. cylindrus*. These enrichments may indicate other classes of sRNAs that are stable within the transcriptome and thus may have a function.

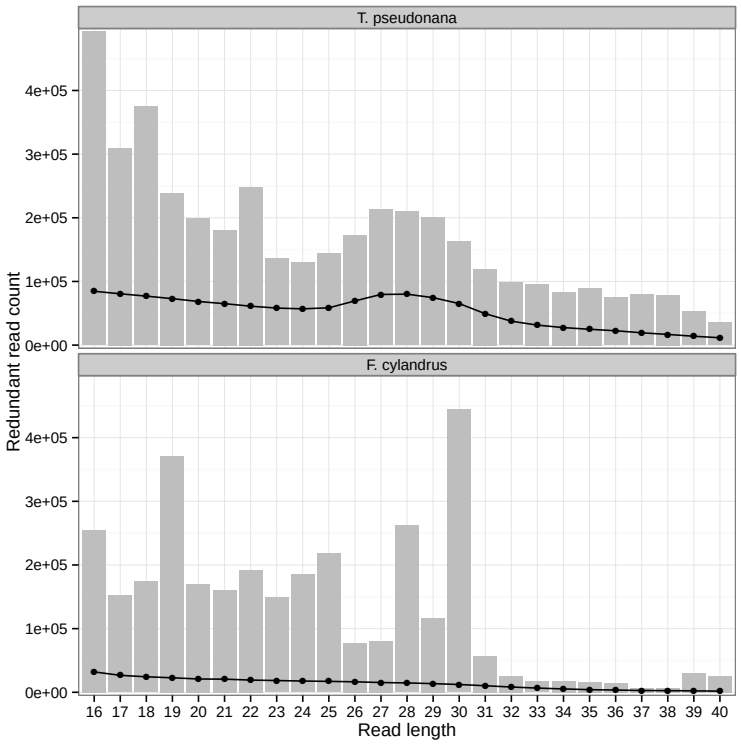
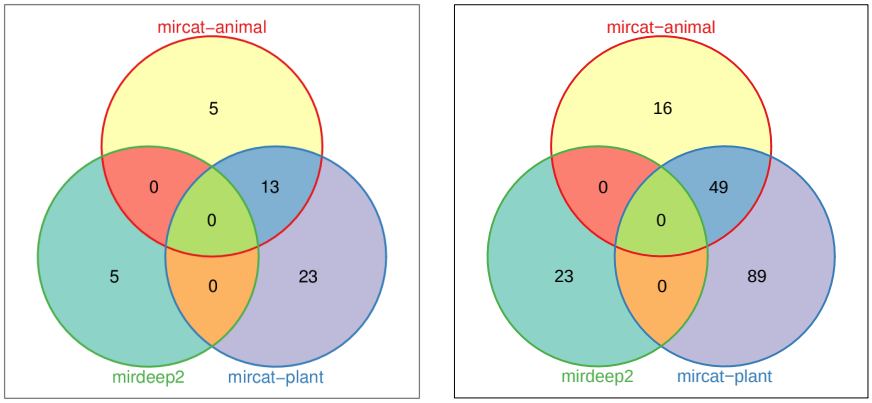



Figure 7.2: sRNA Length Distributions of all three microalgae species after mapping and filtering highly abundant regions that otherwise obscure the remainder of the distribution. Bars indicate redundant counts and lines indicate non-redundant counts.




(a) *T. pseudonana* (b) *F. cylindrus*

Figure 7.3: Venn Diagrams depicting the overlap of predictions between miRNA prediction tools for a) *T. pseudonana* and b) *F. cylindrus*.

### 7.4.2 miRNA predictions

In total, the three miRNA prediction runs produced 46 predictions for *T. pseudonana* and 177 predictions for *F. cylindrus*. The two miRCat runs agreed substantially on their predictions, whereas miRDeep2 never agreed with miRCat's predictions (figure 7.3). Most predictions were based on low mature read counts of between 1 and 20. Using a cutoff threshold of 100 read counts, we kept 7 sequences in *F. cylindrus*, and 1 sequence in *T. pseudonana* for further investigation (table D.1, and figure 7.4). Two of the sequences, Fc3 and Fc4, were validated by northern blots.

Out of the 29 miRNAs predicted by Norden-Krichmar et al. [2011], one miRNA, named '921_306_230_F3!AR2_G31013_21nts_x451' and found 524 times in their library, mapped to sRNAs in our library. There was no sRNA that matched this sequence's size exactly. However, the sequence does map to two of the tRNAs predicted by tRNAscan-SE; a duplicated tRNAHisGTG. The sequence mapped to the middle of the tRNA, where low read counts are found in our libraries. In addition, most of the predictions were aligned to other locations on the genome and overlapped other features such as tRNAs, rRNAs, and repetitive elements. These additional alignments are listed in table D.1 and further illustrates how predicted miRNAs can be mistaken for other types of ncRNA. The findings prompted us to focus on other sources of sRNAs in our diatom libraries.

### 7.4.3 Analysis of other potential sRNAs

Figure 7.5 summarises the proportions of sRNAs that map to particular genomic features. The complexity of each feature is also shown and varies between as little as 0.05 in tRNAs to 0.72 in exons. High complexity for exon sequences is expected, since these are most likely degradation products with low abundance. The low complexity in tRNAs is interesting and suggests that sequences derived from tRNAs are highly abundant and stable within the transcriptome. Furthermore, the tRNAs account for 40.1% of *F. cylindrus* sequences and 11.9% of *T. pseudonana* sequences.

The length distributions in figure 7.6 show different size class enrichments for different features. In *T. pseudonana*, the 26-30 peak is derived from exons and repetitive elements. Intriguingly, the reads derived from repetitive elements are almost exclusively made of these size classes. tRNAs appear to be composed of mostly 16nt reads, with an enrichment at 18-19nt as well. The picture is very different in *F. cylindrus*, with tRNA-derived reads now accounting for the

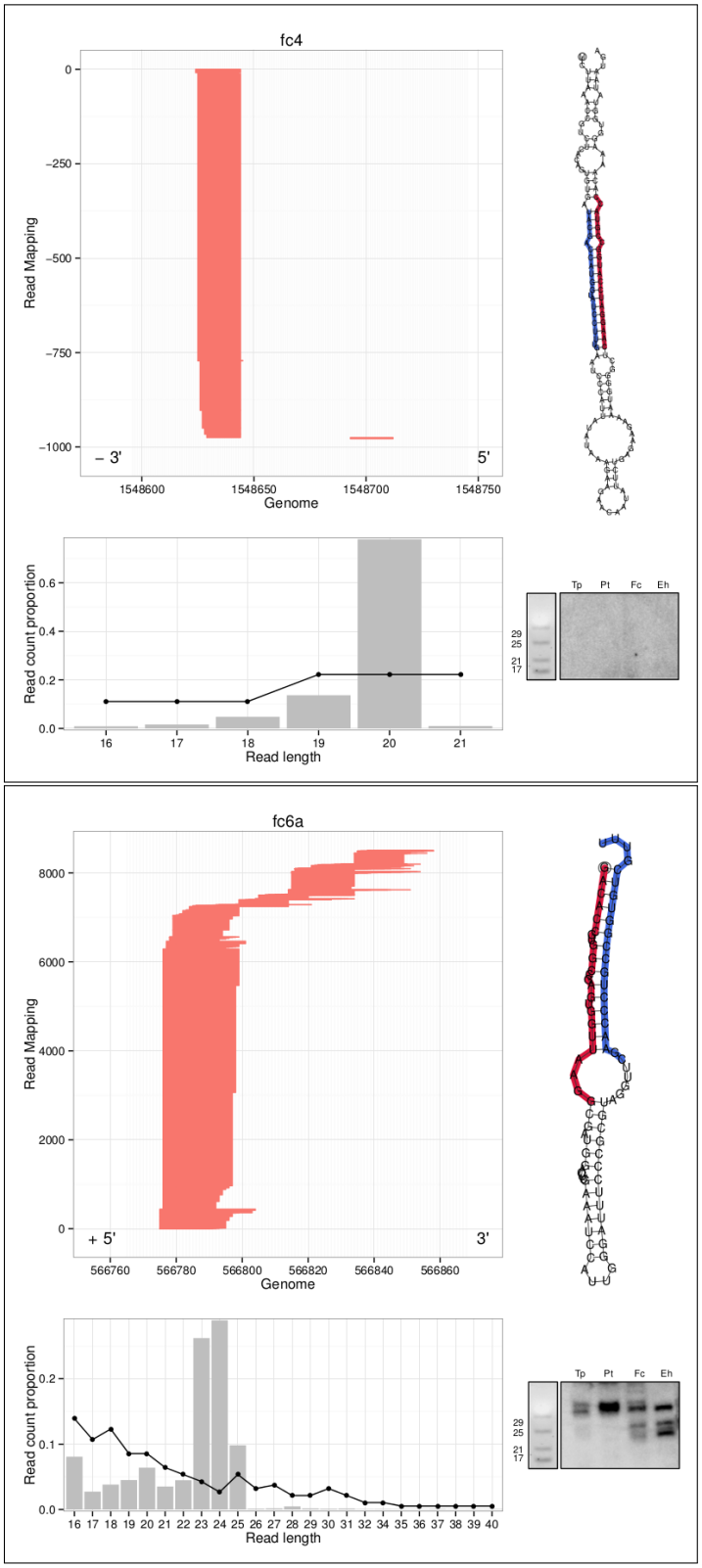



Figure 7.4: The mapping patterns, secondary structure, and northern blot validations of two predicted miRNAs. Mapping patterns are shown by representing each read as a red line along the reference genome (the x axis)

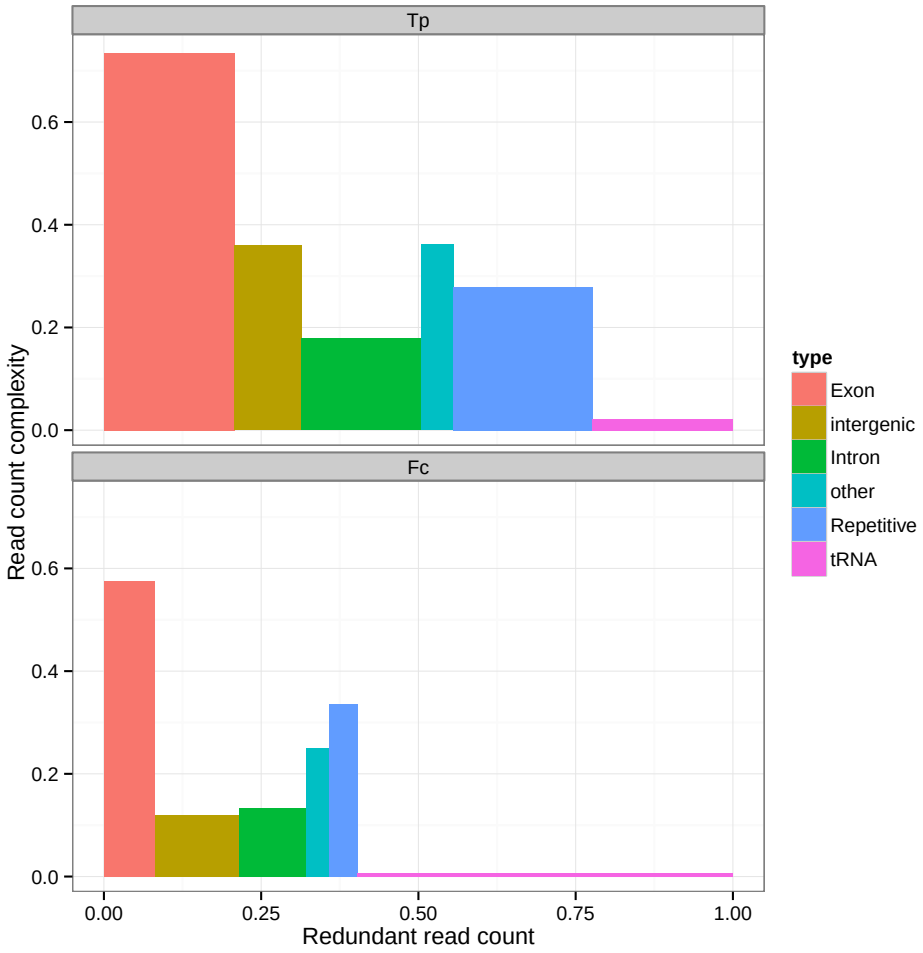


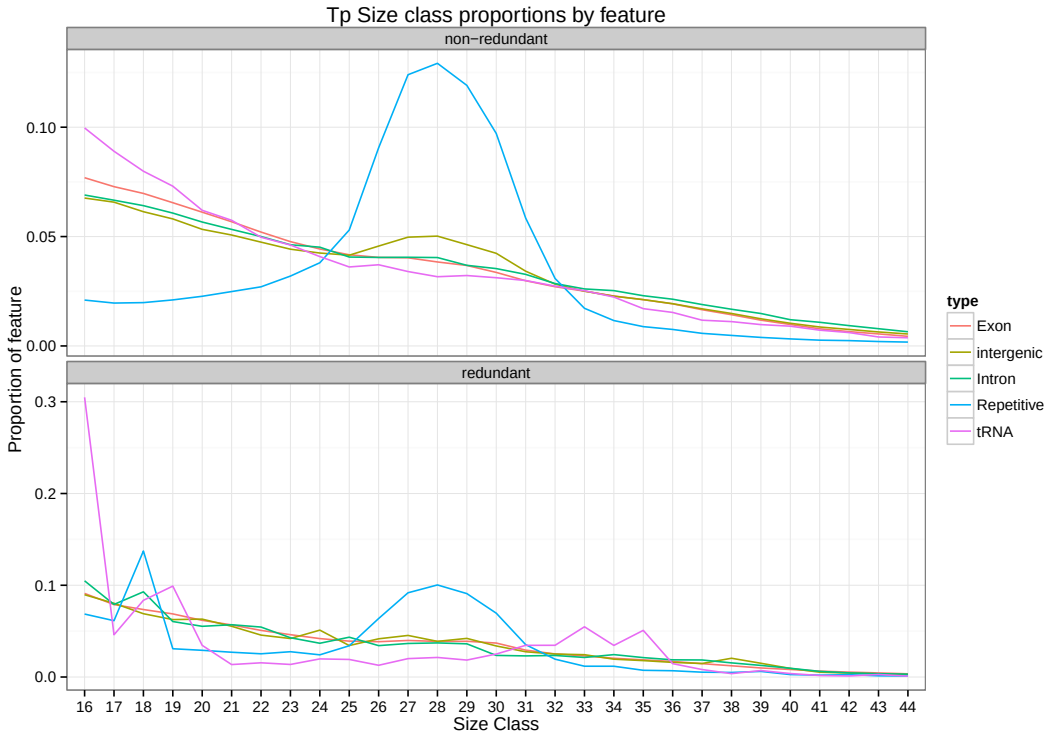

Figure 7.5: Proportional bar charts for *T.pseudonana* and *F. cylindrus* showing the proportion of features that all sRNAs map to. These features include exons, introns, tRNAs, Repetitive elements, and intergenic regions (where no features could be found that overlap the read). The y-axis indicates the complexity of each feature class, which is defined as the non-redundant count divided by the redundant count.

enrichment at 28-30nt. Sequences derived from repetitive elements and tRNAs were investigated further.

### sRNAs derived from tRNAs

Previously identified tsRNAs have been found to map almost exclusively to either end of tRNAs [Lee et al., 2009]. As shown in figure 7.8, reads that map to either end of the tRNA account for the majority of the size class enrichments, whereas internal reads generally do not have an enriched size class. In *T. pseudonana*, 3' matching reads are generally 22nt long, whereas 5' matching reads are 16nt long. In *F. cylindrus*, 5' matching reads tend to be 30nt long, whereas 3' matching

(a) *T. pseudonana*




(b) *F. cylindrus*

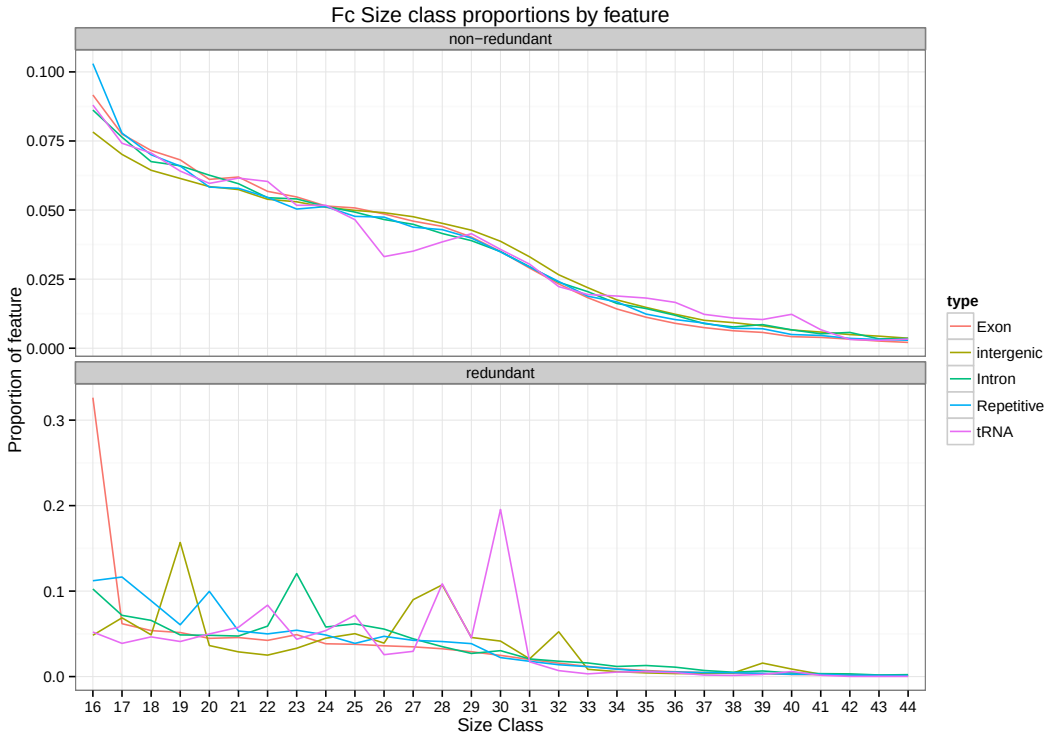
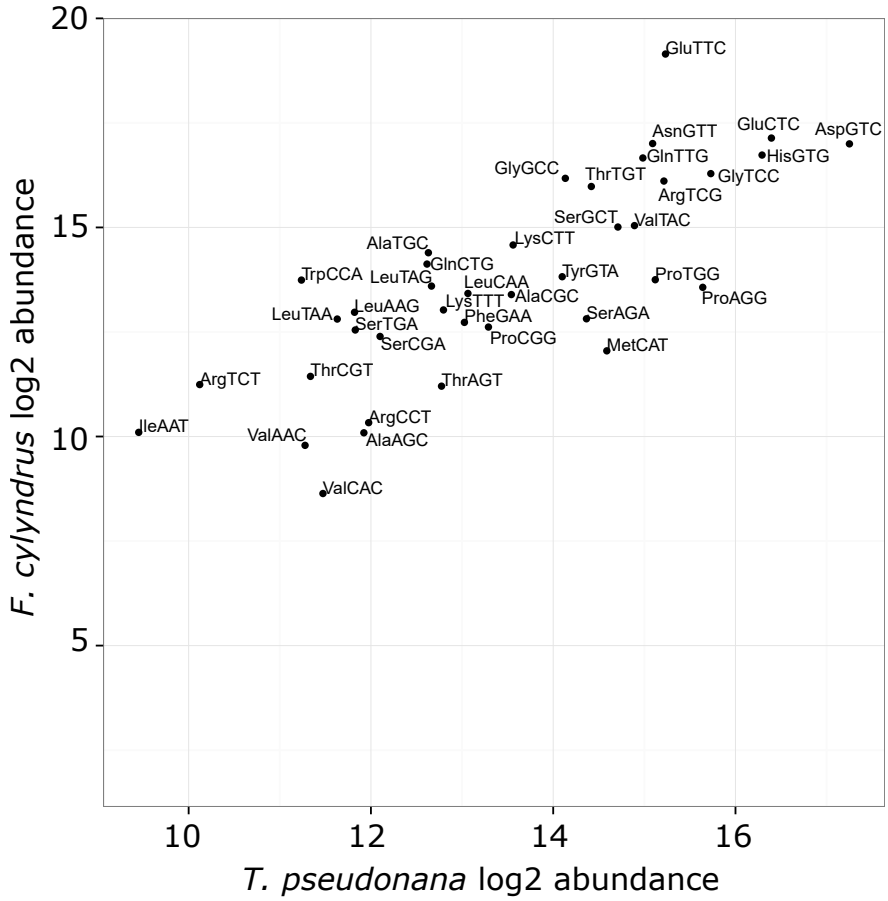



Figure 7.6: Length Distributions for *T.pseudonana* and *F. cylindrus*, showing the proportion of small RNA sizes that make up each feature class.

Figure 7.7: Analysis of total tRNA read abundances between the two diatoms.




reads are similar to *T. pseudonana* - around 22nt.

78 sequences are shared between tRNAs in *T. pseudonana* and *F. cylindrus* that have at most 3 mismatches between sequences. Of these 78, one sequence, separated by two mismatches, showed a consistently high read count between the two diatoms (47,000 and 74,507 read counts respectively), and is shown in the top half of figure 7.9. This maps to the AspGTC tRNA in both organisms and indicates a consistent expression across organisms, suggesting a possible conserved function for the tsRNA. Furthermore, a comparison of the total abundances of each tRNA type between the two diatoms reveals a strong correlation (figure 7.7).

A clear distinction is made in the literature between tRNA-halves, which are produced by angiogenin in plants and animals as a result of cell stress [Thompson and Parker, 2009], and smaller tsRNAs, generally around 18-24nt that are thought to be produced by Dicer-like proteins [Cole et al., 2009]. The results for these diatom libraries indicate that *F. cylindrus* appears to have more tRNA-halves

derived from the 5' end of its tRNAs, whereas *T. pseudonana* is producing more smaller tsRNAs from both ends. Northern blots were also able to confirm that the longer tRNA-halves were upregulated under stress of *T. pseudonana* (figure 7.9).

#### Small RNAs derived from repetitive elements

A larger proportion of sequences derive from repetitive elements in *T. pseudonana*, where a specificity in the size classes is also apparent at 28-30nt (Figure 7.6). Further analysis showed that the repeat-derived sRNAs are the only class to have a bias for a specific nucleotide, in this case U, at their five prime end (figure 7.11). Nucleotide biases at the five prime end have previously been shown to be associated with Argonaute selection of sRNAs in both plants and animals [Kim, 2008]. In particular, transposon-acting piRNAs in animals show a bias for Uridine [Malone and Hannon, 2009]. *F. cylindrus* shows no particular size class specificity for the sequences that map to repetitive elements and there is also no 5' bias.

## 7.5 Discussion

The low relatedness of diatoms to other organisms that have been well studied for sRNAs meant that this study focused on the identification of sRNAs without the aid of comparative genomics. The use of data from two different species facilitated comparisons across the diatom clade.

The presence of sequencing bias in the Illumina datasets mean that we are less confident that sequences with a high read count also have a high rate of expression within the transcriptome. However, subsequent validation of both tRNA-derived sRNAs and miRNA-like sequences show that at least some of these abundances represent functional expression.

This study helps to shed light onto the possible populations of sRNAs within diatoms. We identified very few sRNAs that could be characterised as miRNAs, which suggests that diatoms have either lost this pathway through subsequent evolution or that their ancestors never evolved a miRNA pathway similar to that found in plants and animals. Instead, data from *T. pseudonana* indicates the presence of a pathway that functions to regulate transposable elements, and both diatoms contain tRNA-derived sequences that are similar in characteristics to previously identified tsRNAs [Thompson and Parker, 2009], and thus may have similar proposed functions such as repressing translation in the cell. The

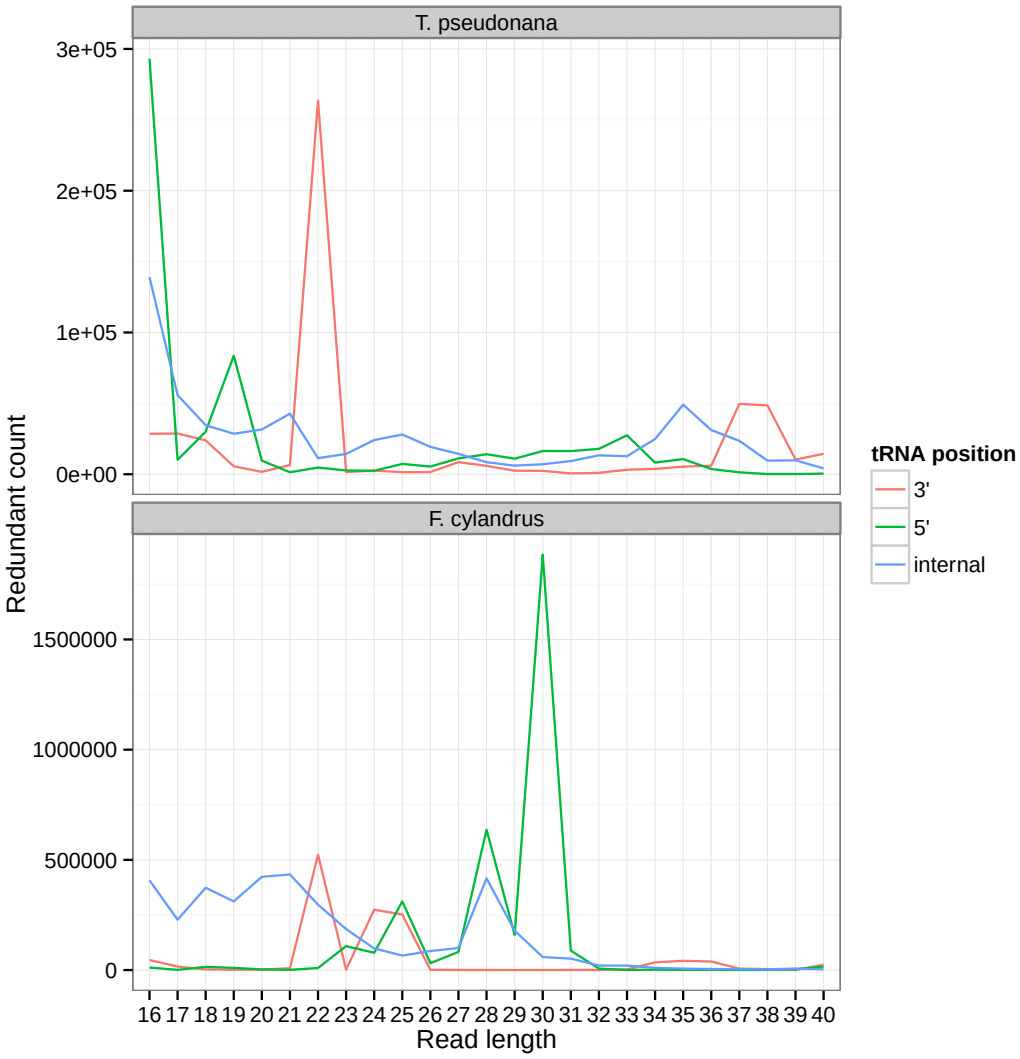



Figure 7.8: Length Distributions for *T.pseudonana* and *F. cylindrus* tRNA-derived reads grouped by the positions on the tRNAs that the reads map to. Reads that aligned precisely to 5' or 3' ends of the tRNAs were grouped as such, otherwise the read was classified as "internal".

(a) *T. pseudonana*

(b) *F. cylindrus*

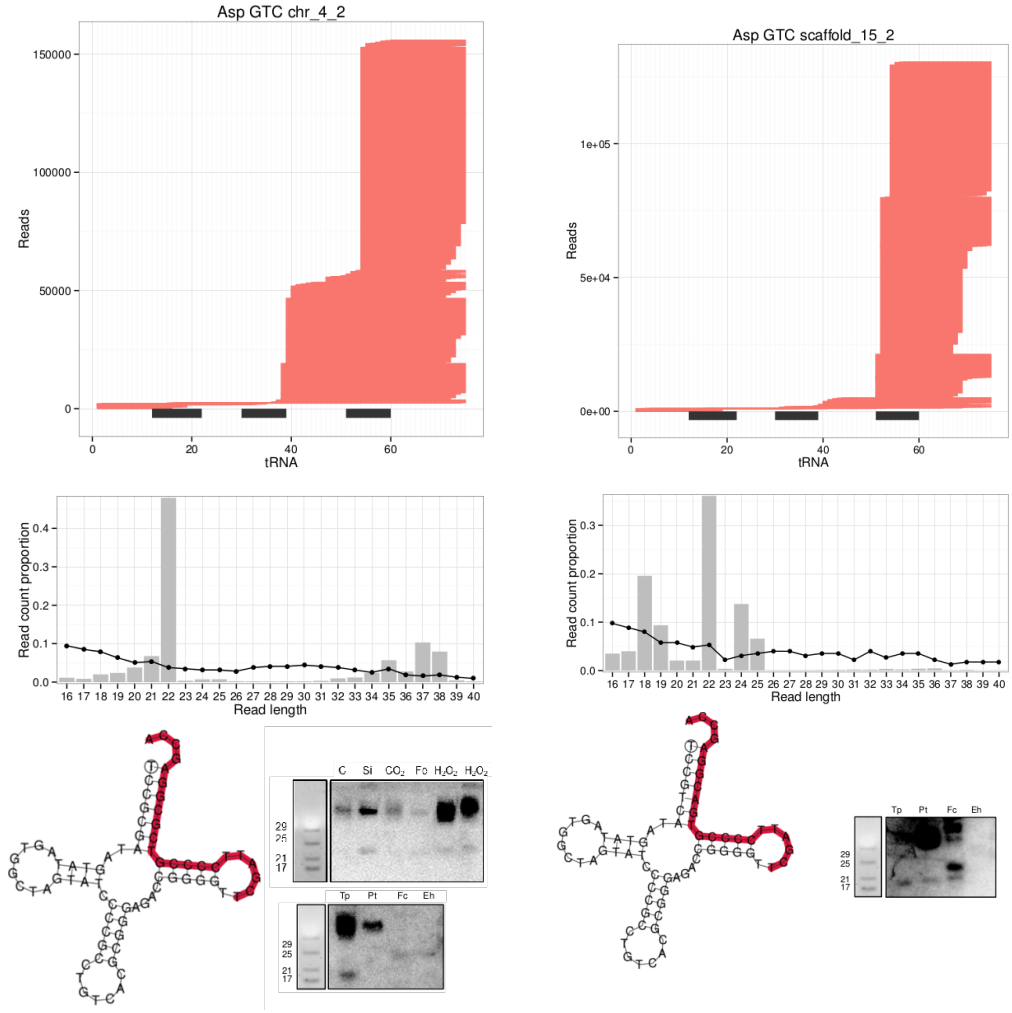
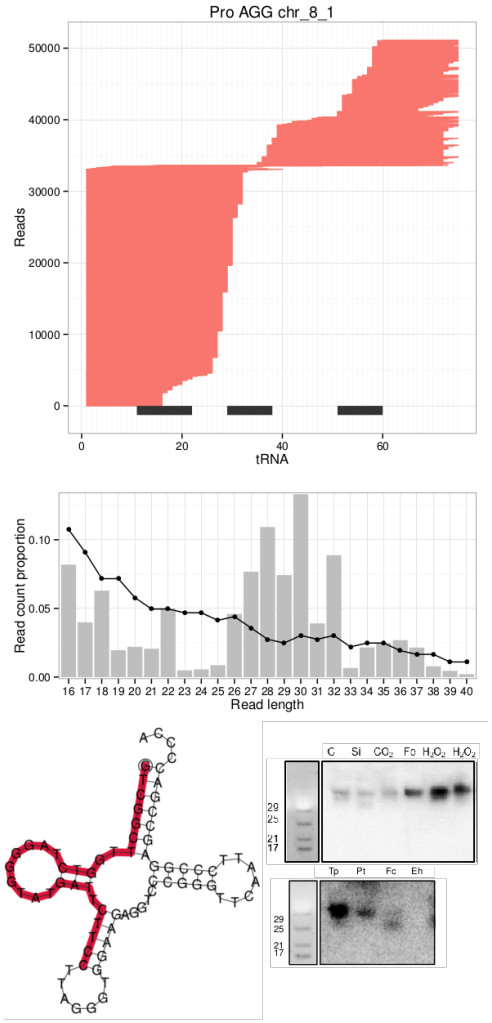



Figure 7.9: Examples of identified tsRNAs in *T. pseudonana* and *F.cylindrus*. The top-left plot is a map of all reads aligned to the tRNA. Black bars beneath the mapped reads indicate where the loop regions are on the tRNA. The tRNA secondary structures (top-right) show the positioning of the top most abundant read in red. Northern blots using probes from the most abundant sequence are shown in the bottom right. For *T. pseudonana*, blots were also done showing upregulation of 30nt+ sequences (tRNA-halves) under stress of the organism.

(a) *T. pseudonana*




(b) *F. cylindrus*

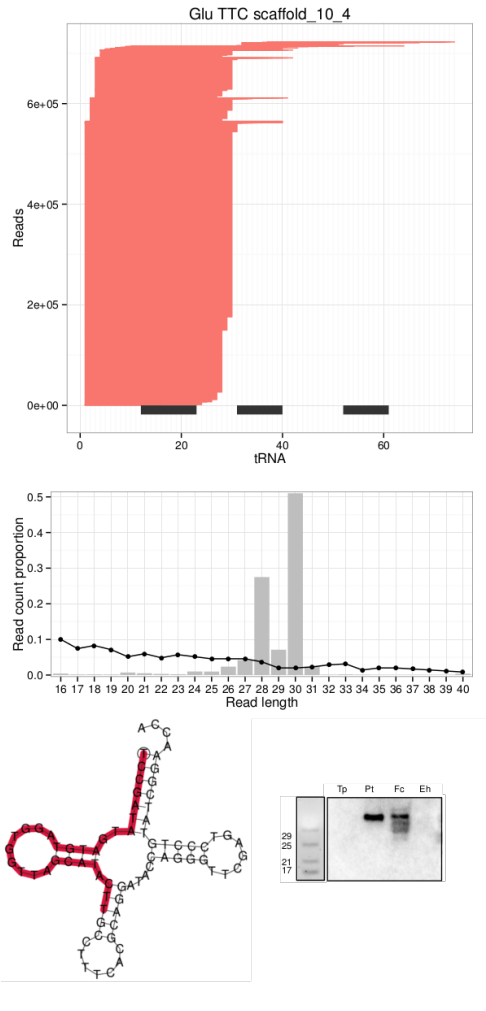



Figure 7.10: Further examples of identified tsRNAs in *T. pseudonana* and *F.cylindrus*. See the caption in figure 7.9 for figure details.

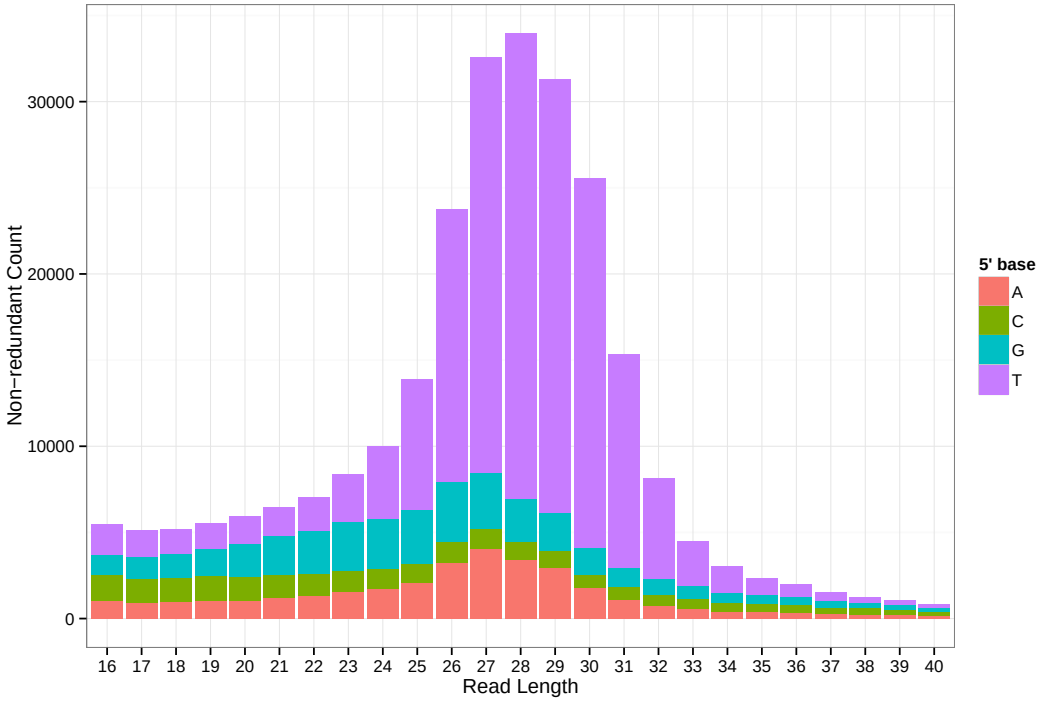



Figure 7.11: 5'-most base distribution across size classes in *T. pseudonana* repetitive elements.

similarities between the relative tsRNA abundances of each tRNA type between the two diatoms suggests also provides further evidence of a shared function for these tsRNAs between the two diatoms.

Importantly, many of the sequences would have been missed if post-transcriptional modifications for the tRNA sequences had been overlooked. This underlines a potential source of lost sRNA data, where sRNAs are derived from transcripts that are modified after translation, causing the sRNAs to be discarded because they do not map to the genome. Many other sRNA transcripts undergo post-transcriptional modifications, which likely affect the results of sRNA sequencing projects [Ebhardt et al., 2009; Findei et al., 2011; Kim et al., 2010].

Future work into these datasets may help to understand the nature of the miRNA-like sequences that have been identified. Advances in the genome assembly of the 1217 *E. huxleyi* strain will also allow analysis on a third microalgae dataset, strengthening the comparative analysis.

# Chapter 8

# Conclusions and future work

## 8.1 Summary

In the previous chapters we have described and utilised a processing pipeline that we have specifically tailored towards sRNA-seq differential expression experiments. The pipeline relies on the use of at least two replicates on each condition to closely assess the quality of the data and consequently to find interesting differentially expressed sequences by filtering and penalising noisy expression levels and subsequent fold changes. In chapter 4 we demonstrated the use of this pipeline as it is implemented in the UEA sRNA Workbench using an animal and a plant sRNA-seq dataset. We then compared the results of differential expression using our LOFC approach to two other currently available methods.

In Chapters 5 and 6 we used our pipeline to process and analyse several interesting experiments. In Chapter 5 we identified novel and closely conserved miRNAs in the European bumblebee and used these annotations to identify an important miRNA for the caste differentiation of bumblebee larvae. In Chapter 6 we focussed on a wider set of ncRNA and cDNA annotations to understand their regulatory changes when human and mouse cell lines are placed under stress. We discovered that many RNA classes are highly differentially expressed when cells are placed under stress but that miRNAs are the least regulated of all RNA classes that we looked at. We also identified particular RNA classes, such as exon-derived RNAs, that are only processed under stress with the help of the Ro60 protein that is already known to associate with the stress regulated YRNAs.

Finally, in Chapter 7 we conducted a search for sRNAs in two novel diatom genomes. In this case, in the absense of comparable conditions, we relied on mapping patterns of our reads to identify sRNA locus. Whilst a search for miRNAs appeared to be inconclusive, we did discover that tRNA-derived sRNAs appear

to be highly abundant in both organisms, and that these are also upregulated under stress of these cells. A futher class, associated with repetetive elements, appeared to be less size specific but contained a preference for T at their 5' end. We concluded that diatoms did contain some forms of sRNAs, but that these may well not be like the miRNAs found in plant and animal linneages, although they tend to be routinely mistaken for such classes.

## 8.2 Future work

### 8.2.1 Integrating sRNA loci aggregation strategies

A large limitation to our pipeline is the absense of any loci aggregation. When sRNAs are aggregated into sRNA loci, the small variations of each sRNA can be grouped to represent a single expression level. This makes the interpretation of differentially expressed sRNA results easier because there can be fewer or no alternative differential expression values for each annotation.

However, the solution for sRNA loci aggregation is not clearly defined; it is often difficult to tell what the borders of sRNA loci should be. In addition, the change in the depth and amount of RNA-seq data has caused some current approaches to finding loci to lose their relevancy. Whereas tools such as SegmentSeq [Hardcastle et al., 2012] and SiLoco [Moxon et al., 2008] have a tendency to create loci that are too large, grouping many different sRNAs together, other tools such as NiBLs [MacLean et al., 2010] are slow to use on large datasets with multiple replicates [Mohorianu et al., 2013].

In chapters 5 and 6 we experimented with a simple method of aggregating closely overlapping reads into sRNAs that used highly differentially expressed reads as the starting reads to group other overlapping reads to. The final differential expression values were then based on a recalculation of the differential expression of the summed abundances for each sRNA. Unfortunately this has the tendency to miss out sRNA loci that may only be differentially expressed when the aggregated value is taken into account.

Recently, a tool called CoLIde was inclduded as a standalone tool in the UEA sRNA Workbench [Mohorianu et al., 2013]. This finds loci based on the similarity of differential expression over neighbouring sRNAs. The approach for differential expression analysis is similar to that of the LOFC method described here, and in the future this could be integrated into the pipeline to provide information on potential sRNA loci rather than single sequences.

### 8.2.2 Integration of sRNA prediction tools

In this thesis we have focussed primarily on the differential expression of sequences. However, in Chapter 5 we also combined the results of two miRNA prediction programs to enhance our set of annotations with novel miRNA predictions. This worked well, but required a great deal of processing power to utilise both miRCat and miRDeep to run on each sample, since they were only used for single sample analysis. Prediction tools like these could be a great asset to an annotation and differential expression pipeline such as the one we implemented. However, changes will be required to make them run more efficiently over multi-sample datasets, with rules such as the ones that we used in Chapter 5 to summarise findings over multiple samples.

### 8.2.3 Normalisation of highly differentially expressed datasets needs work

The accurate differential expression analysis of sRNA-seq datasets requires that the two compared conditions are normalised to eliminate variations that were not originally caused by the biological experiments. Such added variations include the difference in sequencing depth, outlying reads caused by a variety of technical biases, and the requirement for sequences to share sequencing space with the rest of the sequences in the sample. These issues are particularly difficult to equate in experiments with large amounts of differential expression, such as the cell line dataset in chapter 6. TMM normalisation is able to estimate more accurate normalisation factors by identifying outliers that are using up the sequencing space and causing opposite differential expression for true unregulated sequences [Robinson and Oshlack, 2010]. However, it often does not go far enough. In Chapter 6 we resorted to seperately normalising the miRNAs, which clearly appeared to be artificially downregulated by a loss of sequencing space. However, the increased variation of other sRNA classes meant that it was not possible to isolate biases due to differences in depth from biases due to the sharing of sequencing space with highly differentially expressed sequences. Many normalisations work on the assumption that most sequences in a sample are truly not differentially expressed [Maza et al., 2013]. When this is not the case, it can be impossible to approximate the differences in sequencing depth between two or more samples. Other types of normalisation include uses of synthetic spike-in sequences, the expression of which can be measured to estimate differences in depth. However, these still don't address all biases - mainly the downregulation

caused by an upregulation of a large and abundant group of sequences [Locati et al., 2015].

### 8.2.4 Improvements to the detection of noisy expression levels

In Chapter 4 we employed an offset approach to downweight low abundance fold changes which are highly likely to be uninteresting due to the expression levels being within the range of noise. Since the noise range was likely to differ between experiments and even samples, we proposed a method of estimating the limit of likely noisy expression levels by finding the abundance at which sRNA reads may be biased to one strand. In practice, however, the abundance limit is affected by the choice of sRNA loci, which, as described in previous future work sections, can be hard to properly define. Our simple method of splitting the genome into windows of a certain length produced stable abundance limits at fairly high window lengths but these were longer than sRNA loci are likely to be. This is due to the artifical nature by which true sRNA loci are split up by fixed length windows. Other more complicated loci detection methods, while using more processing power and runtime, could be used to more accurately estimate strand biases of particular loci. In addition, other characteristics of sRNA loci may better predict noisy loci, such as the size class distribution of the aligned reads.

## 8.3 Conclusions

The ever-decreasing costs for sequencing sRNA libraries means that the field of bioinformatics must find ways of dealing with larger datasets, the benefits of which can be truly maximised by using the most appropriate tools for the job. In this thesis we have described a variety of approaches to checking the quality of large datasets with several replicates, many of which focus on the specifc characteristics of sRNAs, such as the importance of their size class seperation. Our approach to implementing a pipeline through to differential expression analysis also offers an alternative way to process large datasets that is more hard drive intensive and thus easier to run on standard machines.

We have also explored the versatility of this pipeline by incorporating the results of sRNA differential expression analyses in several different experiments in order to both initially discover new sRNAs and also further identify their

functions. Importantly, the use of completely analysing the whole sRNA-seq dataset has facilitated the discovery of knowledge for many different known and novel sRNA classes. We hope that the methods and tools introduced here enable researchers to better understand and produce higher quality results from their sRNA-seq datasets.

# References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410. 102

Amaral, A. J., Brito, F. F., Chobanyan, T., Yoshikawa, S., Yokokura, T., Van Vactor, D., and Gama-Carvalho, M. (2014). Quality assessment and control of tissue specific RNA-seq libraries of Drosophila transgenic RNAi models. *Frontiers in Genetics*, 5. 19

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106. 22, 24, 25

Andrews, S. (2010). FASTQC: A quality control tool for high throughput sequence data. 2, 19

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M. J., Kuramochi-Miyagawa, S., Nakano, T., Chien, M., Russo, J. J., Ju, J., Sheridan, R., Sander, C., Zavolan, M., and Tuschl, T. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, 442(7099):203–207. 7

Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The Small RNA Profile during Drosophila melanogaster Development. *Developmental Cell*, 5(2):337–350. 7

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., and Bechner, M. (2004). The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, 306(5693):79–86. 95, 96

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data setsupdate. *Nucleic Acids Research*, 41(D1):D991–D995. 40

Baulcombe, D. (2004). RNA silencing in plants. *Nature*, 431(7006):356–363. 8, 9, 10

Berry, C., Hannenhalli, S., Leipzig, J., and Bushman, F. D. (2006). Selection of Target Sites for Mobile DNA Integration in the Human Genome. *PLoS Computational Biology*, 2(11). 28

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309. 61

Bourke, A. F. G. (2011). *Principles of Social Evolution*. OUP Oxford. 67, 76

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R. P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J. A., Brownlee, C., Cadoret, J.-P., Chiovitti, A., Choi, C. J., Coesel, S., Martino, A. D., Detter, J. C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M. J. J., Jenkins, B. D., Jiroutova, K., Jorgensen, R. E., Joubert, Y., Kaplan, A., Krger, N., Kroth, P. G., Roche, J. L., Lindquist, E., Lommer, M., MartinJzquel, V., Lopez, P. J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L. K., Montsant, A., Secq, M.-P. O., Napoli, C., Obornik, M., Parker, M. S., Petit, J.-L., Porcel, B. M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T. A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M. R., Taylor, A. R., Vardi, A., Dassow, P. v., Vyverman, W., Willis, A., Wyrwicz, L. S., Rokhsar, D. S., Weissenbach, J., Armbrust, E. V., Green, B. R., Peer, Y. V. d., and Grigoriev, I. V. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219):239–244. 96

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell*, 128(6):1089–1103. 9

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94. 21, 23, 24

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., and Bateman, A. (2012). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*. 82

Camps, C., Saini, H. K., Mole, D. R., Choudhry, H., Reczko, M., Guerra-Assuncao, J. A., Tian, Y.-M., Buffa, F. M., Harris, A. L., Hatzigeorgiou, A. G., Enright, A. J., and Ragoussis, J. (2014). Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia. *Molecular Cancer*, 13:28. 36, 40

Cardinal, S. and Danforth, B. N. (2011). The Antiquity and Evolutionary History of Social Behavior in Bees. *PLoS ONE*, 6(6):e21086. 76

Carthew, R. W. and Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–655. 6, 9, 10

Casas-Mollano, J. A., Rohr, J., Kim, E.-J., Balassa, E., Dijk, K. v., and Cerutti, H. (2008). Diversification of the Core RNA Interference Machinery in Chlamydomonas reinhardtii and the Role of DCL1 in Transposon Silencing. *Genetics*, 179(1):69–81. 8, 100

Casneuf, T., Peer, Y. V. d., and Huber, W. (2007). In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics*, 8(1):461. 15

Cerutti, H. and Casas-Mollano, J. A. (2006). On the origin and functions of RNA-mediated silencing: from protists to man. *Current genetics*, 50(2):81–99. 3, 7, 8, 10, 96, 97

Cerutti, H., Ma, X., Msanne, J., and Repas, T. (2011). RNA-mediated silencing in algae: biological roles and tools for analysis of gene function. *Eukaryotic cell*, 10(9):1164–1172. 96, 97

Chen, C.-J., Servant, N., Toedling, J., Sarazin, A., Marchais, A., Duvernois-Berthet, E., Cognat, V., Colot, V., Voinnet, O., Heard, E., Ciaudo, C., and Barillot, E. (2012). ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. *Bioinformatics*, 28(23):3147–3149. 39

Chen, P. Y., Manninga, H., Slanchev, K., Chien, M., Russo, J. J., Ju, J., Sheridan, R., John, B., Marks, D. S., Gaidatzis, D., Sander, C., Zavolan, M., and Tuschl, T. (2005). The developmental miRNA profiles of zebrafish as determined by small RNA cloning. *Genes & Development*, 19(11):1288–1293. 7

Chen, X. and Wolin, S. L. (2004). The Ro 60 kDa autoantigen: insights into cellular function and role in autoimmunity. *Journal of molecular medicine*, 82(4):232–239. 12

Chen, X., Yu, X., Cai, Y., Zheng, H., Yu, D., Liu, G., Zhou, Q., Hu, S., and Hu, F. (2010). Next-generation small RNA sequencing for microRNAs profiling in the honey bee Apis mellifera. *Insect Molecular Biology*, 19(6):799–805. 76

Christov, C. P., Gardiner, T. J., Szts, D., and Krude, T. (2006). Functional Requirement of Noncoding Y RNAs for Human Chromosomal DNA Replication. *Molecular and Cellular Biology*, 26(18):6993–7004. 12

Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610. 54

Clote, P., Ferr, F., Kranakis, E., and Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591. 28

Cnaani, J., Robinson, G. E., and Hefetz, A. (2000). The critical period for caste determination in Bombus terrestris and its juvenile hormone correlates. *Journal of Comparative Physiology A*, 186(11):1089–1094. 67

Cock, J. M., Sterck, L., Rouz, P., Scornet, D., Allen, A. E., Amoutzias, G., Anthouard, V., Artiguenave, F., Aury, J.-M., and Badger, J. H. (2010). The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature*, 465(7298):617–621. 98, 100

Cole, C., Sobala, A., Lu, C., Thatcher, S. R., Bowman, A., Brown, J. W. S., Green, P. J., Barton, G. J., and Hutvagner, G. (2009). Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *Rna*, 15(12):2147–2160. 11, 26, 29, 110

Colgan, T. J., Carolan, J. C., Bridgett, S. J., Sumner, S., Blaxter, M. L., and Brown, M. J. (2011). Polyphenism in social insects: insights from a transcriptome-wide analysis of gene expression in the life stages of the key pollinator, Bombus terrestris. *BMC Genomics*, 12(1):623. 67

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563. viii, 2

Cuperus, J. T., Fahlgren, N., and Carrington, J. C. (2011). Evolution and functional diversification of MIRNA genes. *The Plant Cell Online*, 23(2):431–442. 10

Davis, M. P. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A. J. (2013). Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63(1):41–49. 39

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Lalo, D., Gall, C. L., Schaffer, B., Crom, S. L., Guedj, M., and Jaffrzic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683. 23, 24, 46, 48, 51

Ebhardt, H. A., Tsang, H. H., Dai, D. C., Liu, Y., Bostan, B., and Fahlman, R. P. (2009). Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Research*, 37(8):2461–2470. 115

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929. 1

Eddy, S. R. (2004). What is dynamic programming? *Nature Biotechnology*, 22(7):909–910. 27

Emde, A.-K., Grunert, M., Weese, D., Reinert, K., and Sperling, S. R. (2010). MicroRazerS: rapid alignment of small RNA reads. *Bioinformatics*, 26(1):123–124. 18

Ewing, B. and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, 8(3):186–194. 19

Fahlgren, N., Sullivan, C. M., Kasschau, K. D., Chapman, E. J., Cumbie, J. S., Montgomery, T. A., Gilbert, S. D., Dasenko, M., Backman, T. W., and Givan, S. A. (2009). Computational and analytical framework for small RNA profiling by high-throughput sequencing. *Rna*, 15(5):992–1002. 22

Falciatore, A. and Bowler, C. (2002). Revealing the molecular secrets of marine diatoms. *Annual review of plant biology*, 53(1):109–130. 96

Fasold, M., Langenberger, D., Binder, H., Stadler, P. F., and Hoffmann, S. (2011). DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Research*, 39(suppl):W112–W117. 39

Findei, S., Langenberger, D., Stadler, P. F., and Hoffmann, S. (2011). Traces of post-transcriptional RNA modifications in deep sequencing data. *Biological Chemistry*, 392(4). 115

Fire, A., Albertson, D., Harrison, S. W., and Moerman, D. G. (1991). Production of antisense RNA leads to effective and specific inhibition of gene expression in C. elegans muscle. *Development*, 113(2):503–514. 5

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *nature*, 391(6669):806–811. 5

Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177. 18, 35

Friedlnder, M. R., Mackowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52. 28, 31, 69, 101

Fulci, V. and Macino, G. (2007). Quelling: post-transcriptional gene silencing guided by small RNAs in Neurospora crassa. *Current Opinion in Microbiology*, 10(2):199–203. 6

Gardner, P. P. and Giegerich, R. (2004). A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5(1):140. 27

Garmire, L. X. and Subramaniam, S. (2012). Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA*, 18(6):1279–1288. 23, 24, 48

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J.

(2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80. 25

Ghildiyal, M. and Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nature Reviews Genetics*, 10(2):94–108. 9

Gordon, A. and Hannon, G. (2010). FASTX-Toolkit. 35

Gottlieb, E. and Steitz, J. A. (1989). Function of the mammalian La protein: evidence for its action in transcription termination by RNA polymerase III. *The EMBO Journal*, 8(3):851–861. 94

Goulson, D. (2003). *Bumblebees: Their Behaviour and Ecology*. Oxford University Press. 67

Griffiths-Jones, S., Grocock, R. J., Dongen, S. v., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl 1):D140–D144. 28, 82

Guerra-Assuno, J. A. and Enright, A. J. (2010). MapMi: automated mapping of microRNA loci. *BMC Bioinformatics*, 11(1):133. 28, 31, 69

Guo, S. and Kemphues, K. J. (1995). par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell*, 81(4):611–620. 5

Gupta, V., Markmann, K., Pedersen, C. N. S., Stougaard, J., and Andersen, S. U. (2012). shortran: a pipeline for small RNA-seq data analysis. *Bioinformatics*, 28(20):2698–2700. 39

Habegger, L., Sboner, A., Gianoulis, T. A., Rozowsky, J., Agarwal, A., Snyder, M., and Gerstein, M. (2011). RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*, 27(2):281–283. 39

Hafner, M., Renwick, N., Brown, M., Mihailovi, A., Holoch, D., Lin, C., Pena, J. T. G., Nusbaum, J. D., Morozov, P., Ludwig, J., Ojo, T., Luo, S., Schroth, G., and Tuschl, T. (2011). RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA*, 17(9):1697–1712. 15

Hall, A. E., Turnbull, C., and Dalmay, T. (2013). Y RNAs: recent developments. *BioMolecular Concepts*, NA(NA):NA. 80

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12):e131–e131. 19

Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422. 21, 39

Hardcastle, T. J., Kelly, K. A., and Baulcombe, D. C. (2012). Identifying small interfering RNA loci from high-throughput sequencing data. *Bioinformatics*, 28(4):457–463. 117

Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A. Z., and Kay, M. A. (2010). Human tRNA-derived small RNAs in the global regulation of RNA silencing. *Rna*, 16(4):673–695. 11, 12

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431. 27

Holcik, M. and Sonenberg, N. (2005). Translational control in stress and apoptosis. *Nature Reviews Molecular Cell Biology*, 6(4):318–327. 81

Huang, A., He, L., and Wang, G. (2011). Identification and characterization of microRNAs from Phaeodactylum tricornutum by high-throughput sequencing and bioinformatics analysis. *BMC genomics*, 12(1):337. 98, 99

Hunt, J. H., Buck, N. A., and Wheeler, D. E. (2003). Storage proteins in vespid wasps: characterization, developmental pattern, and occurrence in adults. *Journal of Insect Physiology*, 49(8):785–794. 79

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314. 25

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–264. 22

Jan, C. H., Friedman, R. C., Ruby, J. G., and Bartel, D. P. (2011). Formation, Regulation and Evolution of Caenorhabditis elegans 3UTRs. *Nature*, 469(7328):97–101. 79

Jayaprakash, A. D., Jabado, O., Brown, B. D., and Sachidanandam, R. (2011). Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Research*, 39(21):e141–e141. 15

Jhling, F., Mrl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., and Ptz, J. (2009). tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research*, 37(suppl 1):D159–D162. 82

Jorgensen, R. A., Cluster, P. D., English, J., Que, Q., and Napoli, C. A. (1996). Chalcone synthase cosuppression phenotypes in petunia flowers: comparison of sense vs. antisense constructs and single-copy vs. complex T-DNA sequences. *Plant Molecular Biology*, 31(5):957–973. 5

Kadota, K., Nishiyama, T., and Shimizu, K. (2012). A normalization strategy for comparing tag count data. *Algorithms for Molecular Biology*, 7(1):5. 21, 24

Kayala, M. A. and Baldi, P. (2012). Cyber-T web server: differential analysis of high-throughput data. *Nucleic Acids Research*, 40(W1):W553–W559. 39

Khvorova, A., Reynolds, A., and Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–216. 7

Kim, V. N. (2005). Small RNAs: classification, biogenesis, and function. *Mol Cells*, 19(1):1–15. 8, 9

Kim, V. N. (2008). Sorting out small RNAs. *Cell*, 133(1):25–26. 7, 111

Kim, Y.-K., Heo, I., and Kim, V. N. (2010). Modifications of small RNAs and their associated proteins. *Cell*, 143(5):703–709. 115

Koscianska, E., Starega-Roslan, J., Sznajder, L. J., Olejniczak, M., Galka-Marciniak, P., and Krzyzosiak, W. J. (2011). Northern blotting analysis of microRNAs, their precursors and RNA interference triggers. *BMC Molecular Biology*, 12:14. 16

Langenberger, D., Bermudez-Santana, C., Hertel, J., Hoffmann, S., Khaitovich, P., and Stadler, P. F. (2009). Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, 25(18):2298–2301. 28, 30

Langenberger, D., Bermudez-Santana, C. I., Stadler, P. F., and Hoffmann, S. (2010). Identification and classification of small RNAs in transcriptome sequence data. In *Pac Symp Biocomput*, volume 15, pages 80–87. 26, 30

Langenberger, D., Pundhir, S., Ekstrm, C. T., Stadler, P. F., Hoffmann, S., and Gorodkin, J. (2012). deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns. *Bioinformatics*, 28(1):17–24. 28

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25. 18

Lawrence, M., Huber, W., Pags, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol*, 9(8):e1003118. 18

Lee, Y. S., Shibata, Y., Malhotra, A., and Dutta, A. (2009). A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes & development*, 23(22):2639–2649. 11, 108

Legeai, F., Rizk, G., Walsh, T., Edwards, O., Gordon, K., Lavenier, D., Leterme, N., Mreau, A., Nicolas, J., Tagu, D., and Jaubert-Possamai, S. (2010). Bioinformatic prediction, deep sequencing of microRNAs and expression analysis during phenotypic plasticity in the pea aphid, Acyrthosiphon pisum. *BMC Genomics*, 11(1):281. 67

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3):523–538. 48

Liang, C., Zhang, X., Zou, J., Xu, D., Su, F., and Ye, N. (2010). Identification of miRNA from Porphyra yezoensis by high-throughput sequencing and bioinformatics analysis. *PLoS One*, 5(5):e10698. 98, 100

Locati, M. D., Terpstra, I., deLeeuw, W. C., Kuzak, M., Rauwerda, H., Ensink, W. A., vanLeeuwen, S., Nehrdich, U., Spaink, H. P., Jonker, M. J., Breit, T. M., and Dekker, R. J. (2015). Improving small RNA-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization. *Nucleic Acids Research*, page gkv303. 119

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., and Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40(W1):W622–W627. 39

Loong, S. N. K. and Mishra, S. K. (2007). Unique folding of precursor microRNAs: Quantitative evidence and implications for de novo identification. *Rna*, 13(2):170–187. 27

Lopez-Gomollon, S., Beckers, M., Rathjen, T., Moxon, S., Maumus, F., Mohorianu, I., Moulton, V., Dalmay, T., and Mock, T. (2014). Global discovery and characterization of small non-coding RNAs in marine microalgae. *BMC Genomics*, 15(1):697.

Lopez-Gomollon, S., Mohorianu, I., Szittya, G., Moulton, V., and Dalmay, T. (2012). Diverse correlation patterns between microRNAs and their targets during tomato fruit development indicates different modes of microRNA actions. *Planta*, 236(6):1875–1887. 51

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550. 2, 22, 25, 36, 39, 56

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research*, 25(5):0955–964. 102

Lu, Y. Z. and Liu, J. (2010). In silico identification of MicroRNAs and their targets in diatoms. *African Journal of Microbiology Research*, 4(13):1433–1439. 98, 99

MacLean, D., Moulton, V., and Studholme, D. J. (2010). Finding sRNA generative locales from high-throughput sequencing data with NiBLS. *BMC bioinformatics*, 11(1):93. 28, 31, 117

MacRae, I. J., Zhou, K., and Doudna, J. A. (2007). Structural determinants of RNA recognition and cleavage by Dicer. *Nature Structural & Molecular Biology*, 14(10):934–940. 6

MacRae, I. J., Zhou, K., Li, F., Repic, A., Brooks, A. N., Cande, W. Z., Adams, P. D., and Doudna, J. A. (2006). Structural Basis for Double-Stranded RNA Processing by Dicer. *Science*, 311(5758):195–198. 7

Malone, C. D. and Hannon, G. J. (2009). Small RNAs as guardians of the genome. *Cell*, 136(4):656–668. 9, 28, 111

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517. 16

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):pp. 10–12. 35

Mathelier, A. and Carbone, A. (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, 26(18):2226–2234. 28

Matlin, A. J., Clark, F., and Smith, C. W. J. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews. Molecular Cell Biology*, 6(5):386–398. 11

Maza, E., Frasse, P., Senin, P., Bouzayen, M., and Zouine, M. (2013). Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. *Communicative & Integrative Biology*, 6(6):e25849. 118

McCormick, K. P., Willmann, M. R., and Meyers, B. C. (2011). Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence*, 2(1):2. 16, 17, 18, 23, 36

Meister, G. and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature*, 431(7006):343–349. 7, 8, 9

Mello, C. C. and Conte, D. (2004). Revealing the world of RNA interference. *Nature*, 431(7006):338–342. 1, 6

Mller, S., Rycak, L., Winter, P., Kahl, G., Koch, I., and Rotter, B. (2013). omiRas: a Web server for differential expression analysis of miRNAs derived from small RNA-Seq data. *Bioinformatics*, 29(20):2651–2652. 39

Mohorianu, I. (2012). *Deciphering the regulatory mechanisms of small RNAs in plants*. PhD thesis, University of East Anglia, Norwich. 28

Mohorianu, I., Lopez-Gomollon, S., Schwach, F., Dalmay, T., and Moulton, V. (2012). FiRePatFinding Regulatory Patterns between sRNAs and Genes. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(3):273–284. 44

Mohorianu, I., Schwach, F., Jing, R., Lopez-Gomollon, S., Moxon, S., Szittya, G., Sorefan, K., Moulton, V., and Dalmay, T. (2011). Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *The Plant Journal*, 67(2):232–246. 36, 54

Mohorianu, I., Stocks, M. B., Wood, J., Dalmay, T., and Moulton, V. (2013). CoLIde: A bioinformatics tool for CO-expression based small RNA L oci Ide ntification using high-throughput sequencing data. *RNA Biology*, 10(7):1221–1230. 31, 117

Molnr, A., Schwach, F., Studholme, D. J., Thuenemann, E. C., and Baulcombe, D. C. (2007). miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*, 447(7148):1126–1129. 8, 98, 100, 102

Morey, J. S., Ryan, J. C., and Dolah, F. M. V. (2006). Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biological Procedures Online*, 8(1):175–193. 56

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628. 15, 20, 35

Moxon, S., Schwach, F., Dalmay, T., MacLean, D., Studholme, D. J., and Moulton, V. (2008). A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, 24(19):2252–2253. 28, 30, 117

Mueller, T. (2006). H2 Database Engine. `http://www.h2database.com/html/main.html`. [Online; accessed 2016-01-28]. 63

Nicolas, F. E., Hall, A. E., Csorba, T., Turnbull, C., and Dalmay, T. (2012). Biogenesis of Y RNA-derived small RNAs is independent of the microRNA pathway. *FEBS Letters*, 586(8):1226–1230. 12, 13

Nicolas, F. E., Moxon, S., Haro, J. P. d., Calo, S., Grigoriev, I. V., Torres-Martnez, S., Moulton, V., Ruiz-Vzquez, R. M., and Dalmay, T. (2010). Endogenous short RNAs generated by Dicer 2 and RNA-dependent RNA polymerase 1 regulate mRNAs in the basal fungus Mucor circinelloides. *Nucleic Acids Research*, 38(16):5535–5541. 6

Norden-Krichmar, T. M., Allen, A. E., Gaasterland, T., and Hildebrand, M. (2011). Characterization of the Small RNA Transcriptome of the Diatom, Thalassiosira pseudonana. *PLoS One*, 6(8):e22870. xv, 97, 98, 99, 101, 106

O'Brien, C. A. and Wolin, S. L. (1994). A possible role for the 60-kD Ro autoantigen in a discard pathway for defective 5s rRNA precursors. *Genes & Development*, 8(23):2891–2903. 94

Pereboom, J. J. M., Jordan, W. C., Sumner, S., Hammond, R. L., and Bourke, A. F. G. (2005). Differential gene expression in queenworker caste determination in bumble-bees. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1568):1145–1152. 67

Pfennig, D. W., Wund, M. A., Snell-Rood, E. C., Cruickshank, T., Schlichting, C. D., and Moczek, A. P. (2010). Phenotypic plasticity's impacts on diversification and speciation. *Trends in Ecology & Evolution*, 25(8):459–467. 66

Prüfer, K., Stenzel, U., Dannemann, M., Green, R. E., Lachmann, M., and Kelso, J. (2008). PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, 24(13):1530–1531. 18, 69, 82, 101

Que, Q. and Jorgensen, R. A. (1998). Homology-based control of gene expression patterns in transgenic petunia flowers. *Developmental Genetics*, 22(1):100–109. 5

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842. 18, 82

Ramskld, D., Wang, E. T., Burge, C. B., and Sandberg, R. (2009). An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLoS Comput Biol*, 5(12):e1000598. 36

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):R95. 24, 25, 35

Reeb, P. D. and Steibel, J. P. (2013). Evaluating statistical analysis models for RNA sequencing experiments. *Frontiers in Genetics*, 4. 24

Rinke, J. and Steitz, J. A. (1982). Precursor molecules of both human 5s ribosomal RNA and transfer RNAs are bound by a cellular protein reactive with anti-La lupus antibodies. *Cell*, 29(1):149–159. 94

Riso, V. D., Raniello, R., Maumus, F., Rogato, A., Bowler, C., and Falciatore, A. (2009). Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Research*, 37(14):e96–e96. xv, 97

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140. 2, 21, 36, 39, 56

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25. 20, 21, 24, 118

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887. 25

Ruby, J. G., Jan, C. H., and Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–86. 79

Rutjes, S. A., Heijden, A. v. d., Utz, P. J., Venrooij, W. J. v., and Pruijn, G. J. M. (1999). Rapid Nucleolytic Degradation of the Small Cytoplasmic Y RNAs during Apoptosis. *Journal of Biological Chemistry*, 274(35):24799–24807. 12

Sappington, T. W. and S. Raikhel, A. (1998). Molecular characteristics of insect vitellogenins and vitellogenin receptors. *Insect Biochemistry and Molecular Biology*, 28(56):277–300. 79

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100. 28

Schwarz, D. S., Hutvgner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P. D. (2003). Asymmetry in the Assembly of the RNAi Enzyme Complex. *Cell*, 115(2):199–208. 7

Seyednasrollah, F., Laiho, A., and Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, 16(1):59–70. 25, 35

Shi, H., Tschudi, C., and Ullu, E. (2006). An unusual Dicer-like1 protein fuels the RNA interference pathway in Trypanosoma brucei. *Rna*, 12(12):2063–2072. 7

Singh, A., Maichle, R., and Lee, S. (2006). On the computation of a 95% upper confidence limit of the unknown population mean based upon data sets with below detection limit observations. Technical report, US Environmental Protection Agency, Office of Research and Development. 51

Smith, C., Anderson, K., Tillberg, C., Gadau, J., Suarez, A., Sherratt, A. E. T. N., and Whitlock, E. M. C. (2008). Caste Determination in a Polymorphic Social Insect: Nutritional, Social, and Genetic Factors. *The American Naturalist*, 172(4):497–507. 67, 76

Smyth, G. K., Yang, Y. H., and Speed, T. (2003). Statistical Issues in cDNA Microarray Data Analysis. In Brownstein, M. J. and Khodursky, A. B., editors, *Functional Genomics*, number 224 in Methods in Molecular Biology, pages 111–136. Humana Press. 20

Sorefan, K., Pais, H., Hall, A. E., Kozomara, A., Griffiths-Jones, S., Moulton, V., and Dalmay, T. (2012). Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence*, 3(1):4. 15, 19, 60, 68

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biology*, 11(5):207. 2, 3

Stocks, M. B., Moxon, S., Mapleson, D., Woolfenden, H. C., Mohorianu, I., Folkes, L., Schwach, F., Dalmay, T., and Moulton, V. (2012). The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics*, 28(15):2059–2061. 28, 31, 57, 60, 69, 101

Sun, G., Wu, X., Wang, J., Li, H., Li, X., Gao, H., Rossi, J., and Yen, Y. (2011). A bias-reducing strategy in profiling small RNAs using Solexa. *RNA*, 17(12):2256–2262. 15

Sun, J., Nishiyama, T., Shimizu, K., and Kadota, K. (2013). TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics*, 14(1):219. 39

Sun, Z. and Zhu, Y. (2012). Systematic comparison of RNA-Seq normalization methods using measurement error models. *Bioinformatics*, 28(20):2584–2591. 24

Taniguchi, M., Miura, K., Iwao, H., and Yamanaka, S. (2001). Quantitative Assessment of DNA Microarrays - Comparison with Northern Blot Analyses. *Genomics*, 71(1):34–39. 16

Tarazona, S., Garca-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21(12):2213–2223. 15

Thompson, D. M. and Parker, R. (2009). Stressing out over tRNA cleavage. *Cell*, 138(2):215–219. 12, 80, 110, 111

Tuck, A. C. and Tollervey, D. (2011). RNA in pieces. *Trends in Genetics*, 27(10):422–432. 1, 11, 12, 29

Vidal, E. A., Moyano, T. C., Krouk, G., Katari, M. S., Tanurdzic, M., McCombie, W. R., Coruzzi, G. M., and Gutirrez, R. A. (2013). Integrated RNA-seq and sRNA-seq analysis identifies novel nitrate-responsive genes in Arabidopsis thaliana roots. *BMC Genomics*, 14(1):701. 40

Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2454–2459. 27, 28

Watson, M. (2014). Quality assessment and control of high-throughput sequencing data. *Frontiers in Genetics*, 5. 35

West-Eberhard, M. J. (1989). Phenotypic Plasticity and the Origins of Diversity. *Annual Review of Ecology and Systematics*, 20:249–278. 66

Westholm, J. O. and Lai, E. C. (2011). Mirtrons: microRNA biogenesis via splicing. *Biochimie*, 93(11):1897–1904. 79

Wheeler, D. E. and Buck, N. A. (1995). Storage proteins in ants during development and colony founding. *Journal of Insect Physiology*, 41(10):885–894. 79

Wolin, S. L. and Cedervall, T. (2002). The LA protein. *Annual review of biochemistry*, 71(1):375–403. 94

Xu, P., Mohorianu, I., Yang, L., Zhao, H., Gao, Z., and Dalmay, T. (2014). Small RNA Profile in Moso Bamboo Root and Leaf Obtained by High Definition Adapters. *PLoS ONE*, 9(7):e103590. 26

Yang, X. and Li, L. (2011). miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics*, 27(18):2614–2615. 32

Zhang, S., Sun, L., and Kragler, F. (2009). The Phloem-Delivered RNA Pool Contains Small Noncoding RNAs and Interferes with Translation. *Plant Physiology*, 150(1):378–387. 12

Zhang, X., Zhu, Y., Liu, X., Hong, X., Xu, Y., Zhu, P., Shen, Y., Wu, H., Ji, Y., Wen, X., Zhang, C., Zhao, Q., Wang, Y., Lu, J., and Guo, H. (2015). Suppression of endogenous gene silencing by bidirectional cytoplasmic RNA decay in Arabidopsis. *Science*, 348(6230):120–123. 40

Zhou, X., Oshlack, A., and Robinson, M. D. (2013). miRNA-Seq normalization comparisons need improvement. *RNA*, 19(6):733–734. 21, 23, 24, 48

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415. 27, 99

# Appendices

# Appendix A

# Kulback-Leibler divergence analysis for all libraries

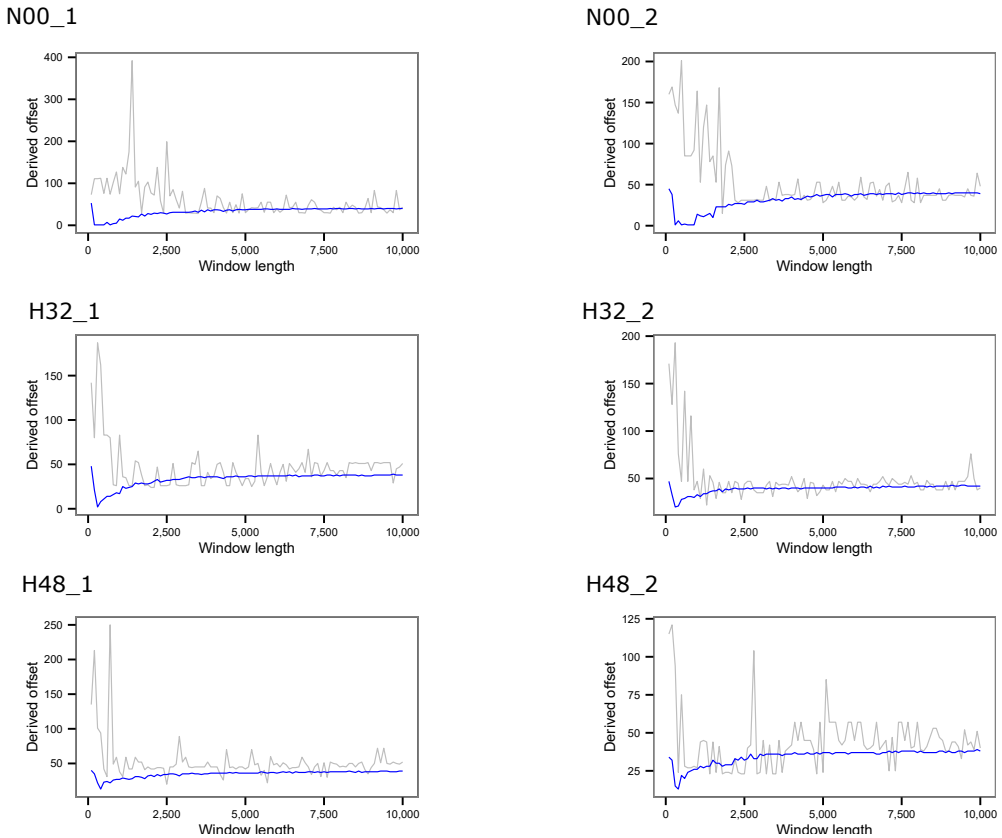



Figure A.1: H (Human) data: The effect of varying the alignment window length when derving an offset based on the Kullback-Leibler divergence on strand bias.

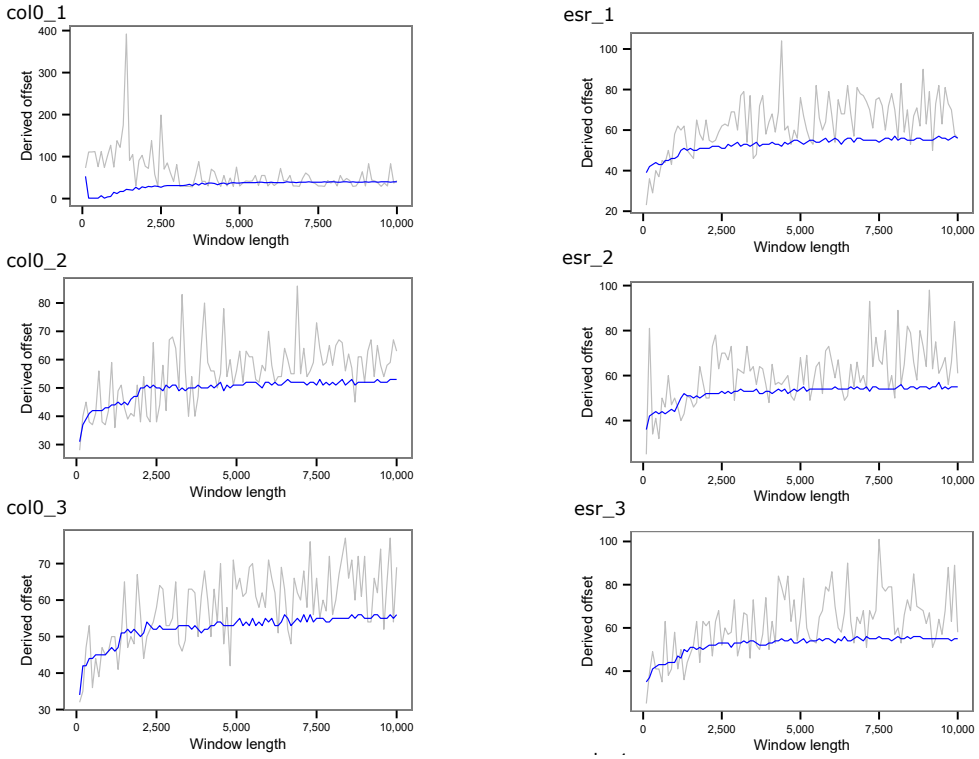



Figure A.2: F (Arabidopsis) data conditions col0 and es: The effect of varying the alignment window length when derving an offset based on the Kullback-Leibler divergence on strand bias.

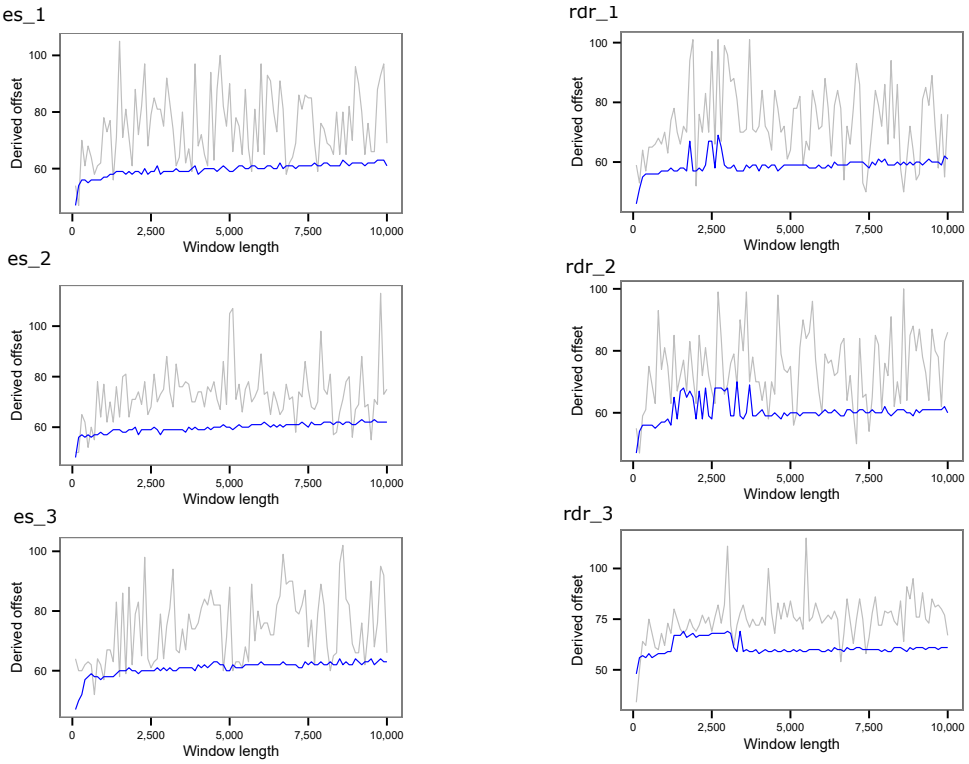



Figure A.3: F (Arabidopsis) data conditions esr and rdr: The effect of varying the alignment window length when derving an offset based on the Kullback-Leibler divergence on strand bias.

# Appendix B

# Additional preprocessing results

Table B.1: Preprocessing results for the mouse cell line data, showing the starting number of sequences, proportions that were lost during preprocessing, and the number of final cleaned sequences

| sample | replicate | lane | total | % no adapter | % filtered | output | % output | unique |
|---|---|---|---|---|---|---|---|---|
| wt_u | 1 | 1 | 14,359,941 | 51.5 | 1.9 | 6,687,427 | 46.6 | 113,928 |
| wt_u | 1 | 2 | 14,251,620 | 51.8 | 1.9 | 6,599,063 | 46.3 | 119,071 |
| wt_u | 2 | 1 | 8,392,267 | 44.9 | 3.0 | 4,365,698 | 52.0 | 78,368 |
| wt_u | 2 | 2 | 8,335,702 | 45.4 | 3.0 | 4,306,044 | 51.7 | 83,381 |
| wt_u | 3 | 1 | 7,558,548 | 32.0 | 0.5 | 5,101,448 | 67.5 | 39,840 |
| wt_u | 3 | 2 | 7,527,737 | 32.3 | 0.4 | 5,063,453 | 67.3 | 42,918 |
| wt_p | 1 | 1 | 7,794,914 | 41.1 | 1.7 | 4,455,718 | 57.2 | 82,892 |
| wt_p | 1 | 2 | 7,775,480 | 41.5 | 1.6 | 4,419,283 | 56.8 | 86,496 |
| wt_p | 2 | 1 | 8,959,157 | 47.0 | 2.6 | 4,518,555 | 50.4 | 62,275 |
| wt_p | 2 | 2 | 8,927,489 | 47.3 | 2.5 | 4,475,774 | 50.1 | 67,026 |
| wt_p | 3 | 1 | 6,400,707 | 36.9 | 0.6 | 4,003,031 | 62.5 | 58,393 |
| wt_p | 3 | 2 | 6,321,411 | 37.3 | 0.5 | 3,931,809 | 62.2 | 61,358 |
| ro60_u | 1 | 1 | 8,198,347 | 37.9 | 1.7 | 4,957,281 | 60.5 | 46,852 |
| ro60_u | 1 | 2 | 8,158,817 | 38.3 | 1.6 | 4,905,531 | 60.1 | 49,885 |
| ro60_u | 2 | 1 | 10,419,734 | 41.2 | 0.7 | 6,047,093 | 58.0 | 47,460 |
| ro60_u | 2 | 2 | 10,289,584 | 41.6 | 0.7 | 5,935,869 | 57.7 | 50,465 |
| ro60_u | 3 | 1 | 7,161,695 | 34.8 | 0.4 | 4,639,910 | 64.8 | 31,772 |
| ro60_u | 3 | 2 | 7,085,932 | 35.2 | 0.4 | 4,565,940 | 64.4 | 34,649 |
| ro60_p | 1 | 1 | 8,272,143 | 36.9 | 2.0 | 5,060,575 | 61.2 | 110,750 |
| ro60_p | 1 | 2 | 8,275,024 | 37.3 | 1.9 | 5,034,841 | 60.8 | 117,325 |
| ro60_p | 2 | 1 | 11,185,537 | 32.9 | 0.7 | 7,426,679 | 66.4 | 75,310 |
| ro60_p | 2 | 2 | 11,050,350 | 33.3 | 0.7 | 7,291,977 | 66.0 | 79,750 |
| ro60_p | 3 | 1 | 8,510,520 | 49.5 | 1.1 | 4,207,117 | 49.4 | 65,582 |
| ro60_p | 3 | 2 | 8,410,467 | 49.8 | 1.0 | 4,133,677 | 49.1 | 69,288 |

Table B.2: Preprocessing results for the datasets for human treated cell lines, showing the starting number of sequences, proportions that were lost during preprocessing, and the number of final cleaned sequences

| sample | replicate | total | % no adapter | % filtered | output | % output | unique |
|---|---|---|---|---|---|---|---|
| m_u | 1 | 5,042,731 | 29.8 | 2.9 | 3,390,169 | 67.2 | 254,723 |
| m_u | 2 | 3,563,313 | 29.9 | 2.1 | 2,421,824 | 68.0 | 195,454 |
| m_u | 3 | 4,900,612 | 32.0 | 2.2 | 3,221,816 | 65.7 | 231,258 |
| m_p | 1 | 4,892,506 | 25.7 | 4.4 | 3,418,702 | 69.9 | 162,059 |
| m_p | 2 | 5,439,093 | 30.9 | 3.2 | 3,588,094 | 66.0 | 149,129 |
| m_p | 3 | 3,143,996 | 18.0 | 2.6 | 2,494,310 | 79.3 | 127,617 |
| sw_u | 1 | 5,408,835 | 28.1 | 1.3 | 3,816,975 | 70.6 | 197,521 |
| sw_u | 2 | 6,172,065 | 25.9 | 1.0 | 4,514,048 | 73.1 | 212,334 |
| sw_u | 3 | 6,065,547 | 28.8 | 2.4 | 4,172,491 | 68.8 | 189,084 |
| sw_p | 1 | 5,894,300 | 25.2 | 2.8 | 4,244,293 | 72.0 | 127,225 |
| sw_p | 2 | 3,533,262 | 15.5 | 5.3 | 2,800,603 | 79.3 | 81,298 |
| sw_p | 3 | 4,822,206 | 20.4 | 3.4 | 3,671,581 | 76.1 | 108,967 |

# Appendix C

# Additional annotation information

Table C.1: Number of sequences belonging to each annotation type split by a sequence's expression pattern for the ro60 experiment

| type | DD | DS | SD | SS | SU | UD | US |
|---|---|---|---|---|---|---|---|
| SRP | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| UTR | 0 | 1 | 2 | 55 | 0 | 0 | 17 |
| YRNA | 0 | 0 | 0 | 5 | 0 | 0 | 3 |
| exon | 0 | 25 | 12 | 648 | 0 | 0 | 289 |
| intron | 1 | 27 | 17 | 330 | 0 | 1 | 78 |
| lincRNA | 1 | 15 | 9 | 126 | 0 | 0 | 21 |
| miRNA | 4 | 18 | 9 | 52 | 0 | 0 | 4 |
| misc | 0 | 0 | 0 | 7 | 0 | 0 | 0 |
| rRNA | 5 | 55 | 32 | 383 | 0 | 0 | 20 |
| snRNA | 0 | 6 | 2 | 35 | 0 | 0 | 3 |
| snoRNA | 7 | 98 | 32 | 396 | 0 | 1 | 43 |
| tRNA | 1 | 23 | 14 | 227 | 1 | 2 | 82 |
| unannotated | 2 | 32 | 23 | 268 | 0 | 1 | 24 |

Table C.2: Non-redundant count of annotations for the cell type dataset.

| annotation | M_U1 | M_U2 | M_U3 | M_P1 | M_P2 | M_P3 | SW_U1 | SW_U2 | SW_U3 | SW_P1 | SW_P2 | SW_P3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| YRNA | 171 | 170 | 142 | 533 | 469 | 468 | 142 | 93 | 151 | 592 | 549 | 568 |
| cDNA | 14,057 | 12,650 | 11,041 | 14,979 | 10,584 | 13,905 | 8,627 | 5,084 | 7,906 | 9,424 | 9,954 | 8,449 |
| lincRNA | 1,375 | 1,210 | 1,002 | 1,282 | 905 | 1,174 | 1,095 | 689 | 1,023 | 1,006 | 1,021 | 848 |
| miRNA | 4,450 | 4,607 | 3,999 | 2,960 | 2,625 | 2,741 | 4,739 | 3,559 | 4,296 | 3,701 | 3,776 | 3,363 |
| other | 3,861 | 3,572 | 3,146 | 4,847 | 3,745 | 4,378 | 2,980 | 1,831 | 2,612 | 3,745 | 3,846 | 3,442 |
| rRNA | 8,116 | 7,321 | 6,350 | 13,131 | 10,945 | 11,824 | 5,137 | 3,304 | 4,755 | 11,637 | 11,942 | 10,813 |
| snoRNA | 986 | 978 | 791 | 1,846 | 1,500 | 1,725 | 877 | 544 | 735 | 1,841 | 1,914 | 1,674 |
| tRNA | 2,498 | 2,599 | 2,335 | 2,520 | 2,155 | 2,307 | 1,617 | 1,165 | 1,388 | 2,451 | 2,552 | 2,309 |
| unannotated | 24,910 | 21,842 | 19,597 | 16,077 | 11,727 | 14,284 | 21,596 | 12,894 | 19,399 | 14,861 | 15,504 | 13,614 |

# Appendix D

# Predicted miRNAs in diatoms

Table D.1: Summary of predicted miRNAs, their alignments on the reference genomes, and the conflicting annotation features that they overlap.

| ID | sequence | count | seqid | start | end | strand | feature | tool |
|---|---|---|---|---|---|---|---|---|
| fc1 | CAAGGATCCATGGCCGTACC | 759 | scaffold_12 | 1548625 | 1548644 | - | rRNA | miRDeep |
| fc2 | CTTGTTAATGTTACCGTTAG | 148 | scaffold_18 | 446404 | 446423 | - | intergenic | miRDeep |
| fc3 | GACACCGTGGCCGAGTGGTTAAGG | 2406 | scaffold_10 | 566775 | 566798 | + | tRNA | miRDeep |
| fc3 | GACACCGTGGCCGAGTGGTTAAGG | 2406 | scaffold_4 | 1479318 | 1479341 | + | tRNA | miRDeep |
| fc3 | GACACCGTGGCCGAGTGGTTAAGG | 2406 | scaffold_149 | 1278 | 1301 | + | tRNA | miRDeep |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_2 | 2472025 | 2472042 | - | tRNA | miRCat-plant |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_10 | 1706149 | 1706166 | - | intergenic | miRCat-plant |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_10 | 1706193 | 1706210 | + | intergenic | miRCat-plant |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_4 | 2738184 | 2738201 | + | intergenic | miRCat-plant |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_4 | 2741614 | 2741631 | - | intergenic | miRCat-plant |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_4 | 2741554 | 2741571 | - | intergenic | miRCat-plant |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_4 | 2721156 | 2721173 | + | tRNA | miRCat-plant |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_10 | 1709408 | 1709425 | - | intergenic | miRCat-plant |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_2 | 2457673 | 2457690 | + | tRNA | miRCat-plant |
| fc4 | GTTCGAGTCCCTGTATCG | 163 | scaffold_10 | 1695233 | 1695250 | + | tRNA | miRCat-plant |
| fc5 | CTAGTTGGTTATGACGCGG | 18 | scaffold_34 | 305847 | 305865 | - | tRNA | miRDeep |
| fc5 | CTAGTTGGTTATGACGCGG | 18 | scaffold_39 | 465945 | 465963 | + | tRNA | miRDeep |
| fc5 | CTAGTTGGTTATGACGCGG | 18 | scaffold_24 | 882056 | 882074 | + | gene | miRDeep |
| fc5 | CTAGTTGGTTATGACGCGG | 18 | scaffold_6 | 1062570 | 1062588 | + | tRNA | miRDeep |
| fc6 | GTGATAGTCTTGTAGTT | 5875 | scaffold_3 | 3185397 | 3185413 | - | match | miRDeep |
| tp1 | GTTTGATGTTGCGAGGA | 6701 | chr_7 | 164398 | 164414 | - | intergenic | miRDeep |
| tp1 | GTTTGATGTTGCGAGGA | 6701 | chr_17 | 259936 | 259952 | + | intergenic | miRDeep |
| tp1 | GTTTGATGTTGCGAGGA | 6701 | chr_17 | 655509 | 655525 | + | intergenic | miRDeep |