NTNU – Trondheim
Norwegian University of
Science and Technology

Identifying miRNA short reads as potential markers for biologically active miRNAs

Kristin Wahl

Master of Science in Computer Science
Submission date: June 2015
Supervisor: Pål Sætrom, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

Sammendrag

MikroRNA (miRNA) er en gruppe ∼22 nukleotider ikke-kodende RNA som har undergått omfattende forskning siden oppdagelsen i 2001. MiRNAs rolle i posttranskripsjonell genregulering i pattedyr og planter har blitt koblet til en rekke klinisk betydningsfulle sykdommer. MiRNA-sekvenser kortere enn 16 nukleotider antas å være kløyveprodukter av Ago2 eller degraderingsprodukt, og miRNA-studier har derfor kun inkludert segmenter av lengde 16-25 nt. I 2014 fant J. P. Mossin ∼10 nt korte miRNA-segmenter som hverken samsvarte med Ago2 kløyving eller kjente degraderingsprosesser, noe som utfordret de rådende antakelsene om korte sekvenser. I et forstudie utført høsten 2014 av samme forfatter som denne masteroppgaven, ble Mossins funn verifisert for flere datasett.

Denne rapporten presenterer et vellykket forsøk på å reprodusere funnene av både Mossins og forstudiet til denne masteroppgaven, ved å studere korte, 11-15 nt miRNA-sekvenser. Studiene ble utvidet til flere datasett fra både menneske og mus. Denne rapporten presenterer en rekke analyser som motbeviser de rådende antakelsene for korte sekvenser, og konkluderer med framfor å representere rester av degraderte miRNA passasjer-tråder, er majoriteten av korte miRNA-sekvenser i realiteten markører for biologisk aktive miRNA. Basert på resultatene i denne rapporten blir en modifisert modell av miRNA-aktivitet og degradering presentert.

Abstract

MicroRNAs (miRNAs) are a group of ∼22 nt non-coding RNAs that since their discovery in 2001 have been extensively studied, and their post-transcriptional gene-regulatory role in animals and plants have been linked to a number of clinically important diseases. Studies have only regarded miRNA segments of length 16-25 nt, assuming shorter reads to be either Ago2 cleavage or degradation products. The credibility of this assumption was questioned by J. P. Mossin in 2014, when short reads of length ∼10 nt was found not compatible with being products of cleavage or known degradation processes. A preliminary study from late 2014 by the author of this master's thesis verified Mossins findings.

This report presents a successful attempt at reproducing both the findings of Mossin and results from the preliminary study, by studying short reads of length 11-15nt. The experiments are extended onto multiple data sets of human and mouse genomes. The report presents a range of analyses discouraging the current assumption regarding short reads, concluding that rather than being remnants of passenger strands, the majority of short reads are actually markers for biologically active miRNAs. A modified model of miRNA activity and degradation is presented, substantiated by the findings of this study.

Preface

This master thesis is submitted as a completion of a Master's Degree in Computer Science at the Norwegian University of Science and Technology in Trondheim spring 2015. It is a continuation of a preliminary study of December 2014. Both the preliminary study and this master thesis have been completed at the Bioinformatics group of the Department of Cancer Research and Molecular Medicine at St. Olavs hospital, in management of the Department of Computer and Information Science, at NTNU.

For invaluable involvement, enthusiasm and guidance, I would like to give a special thanks to my supervisor Professor Pål Sætrom, whose accommodating feedback has been crucial for my work.

Trondheim, June 2015
Kristin Wahl

Contents

List of Figures

List of Tables

Abbreviations

A	Adenine
Ago	Argonaute
ANOVA	Analysis of Variance
Boxplot	Box-and-whiskers plot
BWM	Burrows-Wheeler matrix
BWT	Burrows-Wheeler transform
C	Cytosine
cDNA	Complementary DNA
D	Different class
DNA	Deoxyribonucleic aicd
E	Equal class
FC	Fold change
G	Guanine
IP	Immunoprecipitation
isomiR	miRNA isoform
KO	Knock out
LF	Least-First
miRNA	MicroRNA
moR	MicroRNA-offset RNA
mRNA	Messenger RNA
n	Set size
ncRNA	Non-coding RNA
nt	Nucleotide
p	Probability value
PCR	Polymerase chain reaction
pre-miRNA	Precursor-miRNA
pri-miRNA	Primary miRNA
r	Correlation coefficient
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNA-Seq	RNA Sequencing
RPM	Read-per-million
s	Slope
SD	Standard deviation
SR	Short read
T	Thymine
U	Uracil
UTR	Untranslated region
WT	Wild type

Chapter 1

Introduction

MicroRNAs (miRNAs) are short non-coding RNAs that play an important gene-regulatory role in animals and plants (Bartel, 2009). Derived from longer duplex precursor sequences, mature miRNAs are ∼22 nucleotides, and will typically be loaded into a RNA-induced silencing complex (RISC) to serve as a guide strand for targeting messenger RNAs (mRNA) by partially base-pairing with the mRNA. When paired to the target, the RISC may perform destabilization, translational repression or site-specific cleavage of the mRNA.

RISC is composed of an Argonaute (Ago) protein and a miRNA (Cenik & Zamore, 2011), and Ago proteins exist in multiple variants across genomes. Mammals encode four Ago proteins, Ago1-Ago4, of which all can be the active component in RISC. The activity of RISC depends on which Ago protein is active in the complex, and the only Ago protein known to be functionally distinct is Ago2, as only Ago2 retains the ability to cleave mRNA. The functionality of the other Ago proteins is still not fully understood.

Since the recognition of miRNAs in 2001, a still increasing amount of research has been conducted to attempt understanding the biogenesis of miRNAs, and abnormal expression levels of miRNAs have been linked to a number of clinically important diseases (Soifer, Rossi, & Saetrom, 2007). The miRNA mediated post-transcriptional gene regulation has been extensively studied, and the data experimented on has been short RNA segments of length 16-25 nt. In attempting to find ∼10 nt Ago2 cleavage products, Mossin (2014) found a number of such short reads aligning to mature miRNAs, however these reads were also found in samples where Ago2 were knocked out. Such short reads have been commonly assumed either Ago2 cleavage products or products of known degradation processes, however Mossins findings question the credibility of these assumptions.

The experiments of Mossin rose the question of what these short reads actually represent, and in Wahl (2014), I extended his experiments to reproduce his findings. I concluded that the prior assumptions regarding short reads cannot hold, and that short reads might be caused by an unknown biological function and/or unknown degradation processes. This is the outset for the study behind this report, where the objectives are to reproduce my findings, assess the credibility of prior assumptions regarding short reads, and investigate possible features and relations that can explain short read expression and correlation with miRNAs.

This report describes my successful attempt at reproducing my preliminary results on six independent data sets. The data experimented on are high-throughput sequencing data from independent experiments on mouse and human, and the data have been aligned to miRNA reference genomes and analysed. The short reads studied have been of length 11-15 nt, and different aspects of their existence have been investigated to enable a reliable discussion of the credibility of the current assumptions regarding short reads. Extensive analyses of features of short reads and short read associated miRNAs have been conducted to attempt explaining the underlying biological functionality of short reads, concluding in a modified model of miRNA activity and degradation.

The outline of this report is as follows. A theoretical introduction to all background information necessary for understanding the procedures and results is presented in Chapter 2, including previous work and the rationale for this project. In Chapter 3, a presentation of the data experimented on is given, before describing the methods used to produce all results. These results are presented and discussed separately in Chapter 4, and a discussion comparing and evaluating the results is presented in Chapter 5. Chapter 6 provides a conclusion of this report, and discusses possible directions for further studies.

Chapter 2

Background

This section provides the basic background information needed to understand the work presented in this paper. Bioinformatics is an inter-disciplinary field of science, and as such this section will contain a wide range of topics. The main audience of this paper has a background in computer science, so this section begins with covering some elementary topics from molecular biology, before explaining some background of methods and technology used. Finally, it elaborates the actual problem in question, related work and what the potential results may offer.

2.1 Cells, DNA, RNA and protein

The information presented in this section is based on Lewin (2006) except for where otherwise noted, and will cover the basic biology of cells, DNA and RNA, followed by the Central Dogma of molecular biology, and regulation of gene expression.

2.1.1 DNA and RNA

The smallest building block composing every organism is the cell. The human body consists of approximately 37 trillion cells (Bianconi et al., 2013), each responsible for chemical reactions necessary to maintain life, and for passing the information on how to do so to the next generation of cells (Sung, 2010). This information is crucial for constructing the organism and is defined in its complete set of genetic material, the genome. The information is stored in deoxyribonucleic acid (DNA) structures, consisting of chains of nucleotides, the basic unit of DNA. The genome is functionally divided into genes, which are sequences of DNA that encode for another nucleic acid, the ribonucleic acid (RNA).

Both DNA and RNA consist of polynucleotide chains, in which the building block nucleotide is composed of a nitrogenous base, a monosaccharide sugar and a phosphate. An alternating series of sugar and phosphate residues compose what is called the sugar-phosphate backbone, spanning from the 5' end to the 3' end. The nitrogenous bases protrude from the sugar-phosphate backbone by glycosidic bonds to the sugar. In DNA, the base can either be Adenine (A), Cytosine (C), Guanine (G) or Thymine (T). These nucleotides are connected in sequence to form a nucleotide chain, or DNA strand. Two opposing DNA strands are connected through hydrogen bindings between opposing bases, resulting in a DNA molecule shaped as a double helix with two sugar-phosphate backbones protecting its bases in between. An illustration of the DNA structure is given in Figure 2.1.

The sequences of nucleotides in the two chains can be represented as a string over the alphabet {A, C, G, T}. The only possible base bindings are Adenine-Thymine and Cytosine-Guanine, a constraint leaving the two strands reversely identical, and any of the two strands can identify the same DNA sequence.

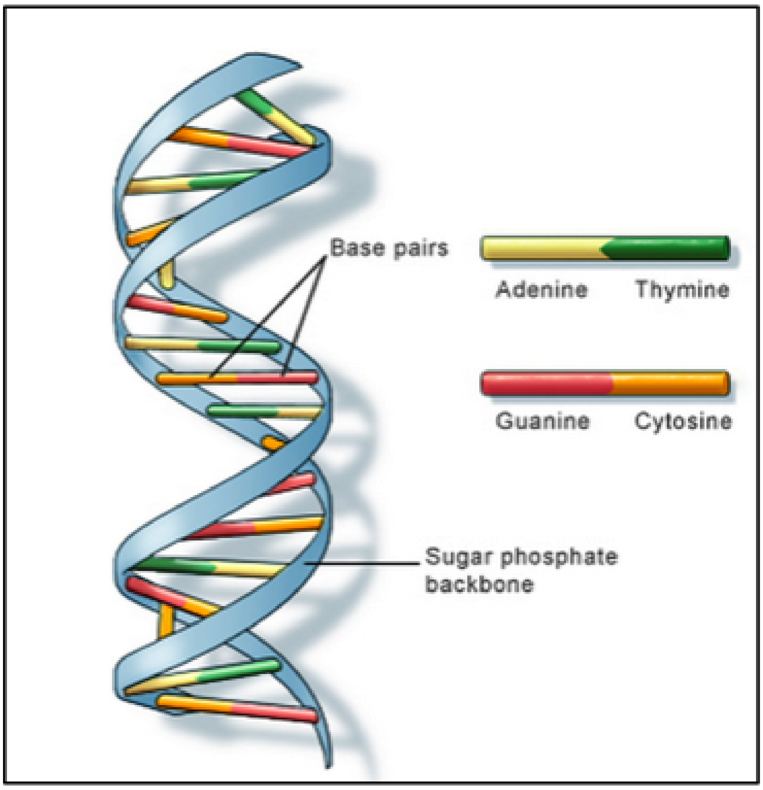


Figure 2.1: Simplified DNA structure. Image adapted from National Library of Medicine US, NLM (2015).

2.1.2 Central Dogma

The central dogma of genetics explains how functional proteins are produced from the information given in the DNA, a process called gene expression. Each cell has a copy of the genome in its nucleus. The full human genome is vast and contains approximately 3 billion base-pairs, however only about 1.5% of it actually encodes proteins (IHGSC, 2001). When the cell is to produce a certain protein, only a small part of the DNA encodes the process. As the cell only contains one replicate of the DNA, it is necessary to transcribe the relevant segment of the DNA, transport this out of the nucleus and to the production unit in the cell, the ribosome, and produce the protein by translating the transcribed DNA fragments. This gene expression process is illustrated in Figure 2.3.

The first step in this process is RNA transcription. In this process, the enzyme RNA polymerase and other transcription factors attach to a specific site in the DNA, uncoils the helix and synthesizes an antiparallel strand to the DNA template strand. This new strand is RNA, similar to DNA but differing by being single-stranded and containing the nucleotide Uracil (U) in replacement for Thymine. The RNA contains sections not required to produce the desired proteins, and some of these are removed to produce the messenger RNA (mRNA), of which a simplified structure is illustrated in Figure 2.2. A mRNA contains a coding region, which encodes for a specific protein, and two flanking untranslated regions (UTR) on each side, conventionally named 5' UTR and 3' UTR. The UTRs do not encode protein, but contain regulatory regions that may be utilized for post-transcriptional influence.

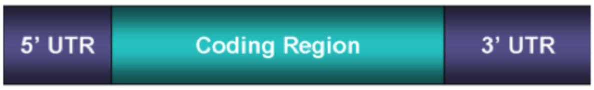


Figure 2.2: Simplified mRNA structure. Image adapted and modified from `http://en.wikipedia.org/wiki/File:Mature_mRNA.png`

The next step is RNA translation, which is the synthesis of proteins from mRNA. The mRNA is transported out of the nucleus and to the ribosome. It is then read using the genetic code, a language of 64 words consisting of all three-letter combinations of the four nucleotides known as codons. These codons encode for one start codon, three stop codons and a total of 20 amino acids, the building blocks of proteins. As the mRNA is read one codon at a time, an anticodon with the translated amino acid binds to it. When the next anticodon binds to the mRNA, the prior amino acid binds to the next one by peptide bonds, creating a growing chain of amino acids. When the stop codon is encountered the translation is terminated and the peptide chain is released into the cytoplasm, where it will fold itself into a three dimensional structure defining the function of the protein.

2.1.3 Gene expression regulation

Approximately 45 million base-pairs constitute the encoding part of the human genome, and encodes for an estimated number of 250,000 proteins. Not all proteins are produced in every cell; what proteins and the amount of them vary from different cell types, tissues and individuals, while the genome remains the same. Different genes in the genome can be translated to proteins in one cell while not in another, and produced proteins are degraded at different rates in different cells. The amount of a particular protein in a cell reflects the balance between the productive and degradative biochemical pathways for that protein, and to a certain degree this balance is self regulated by the the cell. This regulation is called gene expression regulation, and the major regulatory mechanisms affect the protein synthesis.

A gene can be in an ‘on’ or ‘off’ state regarding transcription, and must be ‘on’ for transcription to be initialized. Human genes are by default in an ‘off’ state (Hoopes, 2008), regulated by histone bindings. If in an ‘off’ state, a gene will not be transcribed and the corresponding proteins will not be synthesized, which is the major gene expression control point.

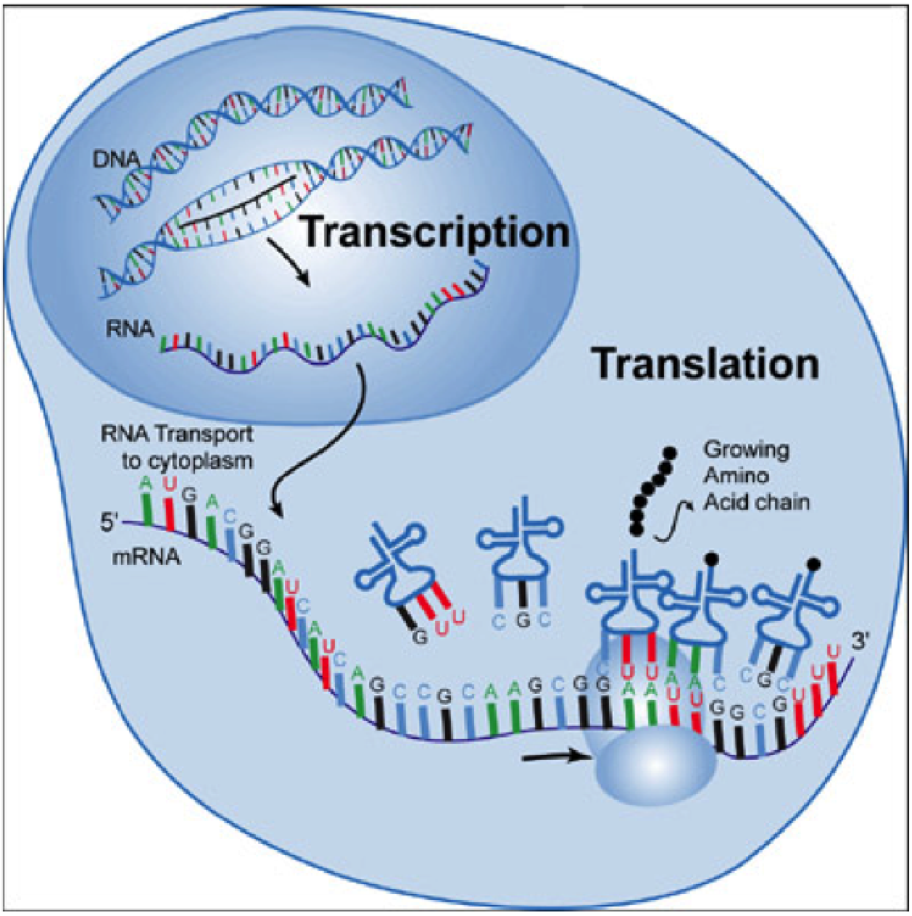


Figure 2.3: Central Dogma of molecular biology. Image adapted from Wiki Kids Ltd. (2014).

If the cell needs to alter the protein balance due to environmental changes, the particular protein might already be transcribed, thus altering the transcription is not sufficient. Post-transcriptional regulation is then necessary, and can be performed in different ways. The mRNA can be altered while in transit to the ribosome, the RNA translation initiation in the ribosome can be affected, and how the cell processes or degrades newly synthesized proteins can affect the protein levels in the cell.

A mechanism important for the work presented in this paper is the post-transcriptional regulation mediated by noncoding RNA. This has emerged as a critical mechanism for gene expression regulation (Carthew & Sontheimer, 2009), and will be described in more detail in the next section.

2.2 MicroRNA

There are two main classes of RNA, where one is translated into protein (mRNA) and the other not. The most common class is mRNA, whose function is presented in section 2.1.2. The other class of RNA are commonly known as non-coding RNA (ncRNA), and consists of a wide range of small functional RNAs that are active in different processes within the cell (Mattick & Makunin, 2006). Most are involved in protein synthesis, DNA replication or gene expression regulation. One such ncRNA important for the work presented in this paper is the microRNA, a ncRNA involved in gene regulation. This section presents the history and importance of miRNA recognition, their biogenesis and their function in the cell.

2.2.1 Biogenesis of miRNA

The information presented in this section is based on Ambros (2004), Bartel (2004), and Mattick and Makunin (2006), unless otherwise specified.

What has later been classified as microRNAs was first discovered in 1993 in a study of *C. elegans* (Lee, Feinbaum, & Ambros, 1993). They found the process of which the lin-4 gene negatively regulates the lin-14 protein was not due to a lin-4 protein, but rather a pair of short RNAs produced by lin-4. The pair consisted of one ∼22 nt RNA and one ∼61 nt RNA, the longest predicted to be the precursor of the shorter one, and their main interaction causing the down regulation was found to be multiple antisense complementarity with the 3' untranslated region of lin-14. This discovery launched a new branch of molecular biology research where gene regulation in animals and plants mediated by short RNAs became an important topic. In 2001 microRNAs (miRNAs) were recognized as a separate class of ncRNAs, and since their discovery, a large number of human miRNAs and miRNA target genes have been identified, and abnormal expression levels of some miRNAs and their targets have been linked to a number of clinically important diseases (Soifer et al., 2007).

Human mature microRNAs are small, single-stranded RNAs of approximately 20 – 23 nucleotides, with a primary function of identifying target mRNAs for destabilization or translation reduction. MiRNAs originate from the genome, and the miRNA biogenesis is illustrated in Figure 2.4. The precursor transcript is transcribed from the genome by normal RNA transcription in the nucleus and folds back on itself to form distinctive hairpin structures, each hairpin representing a potential miRNA precursor, and the whole transcript may serve as both an mRNA and a primary miRNA (pri-miRNA). Still in the nucleus, the enzymatic complex of Drosha and DGRC8 processes the pri-miRNA to produce precursor-miRNA (pre-miRNA) by cleaving the hairpin from the pri-miRNA. A pre-miRNA is an approximately 70 nucleotides long RNA duplex, folded by imperfect base-pairing into a stem-loop structure, containing two mature miRNA candidates. Figure 2.5 illustrates a miRNA hairpin, where purple nucleotides represent the candidate mature miRNA sequences.

The pre-miRNA is transported to the cytoplasm by Exportin-5 and Ran-GTP, where the pre-miRNA is recognized and the Dicer-TRBP complex cleaves off the loop of the pre-miRNA molecule, resulting in a 22 nucleotide miRNA:miRNA* duplex. The duplex represents two possible mature miRNAs, of which one will typically be incorporated into an Argonaute protein of the RNA-induced silencing complex to guide mRNA silencing.

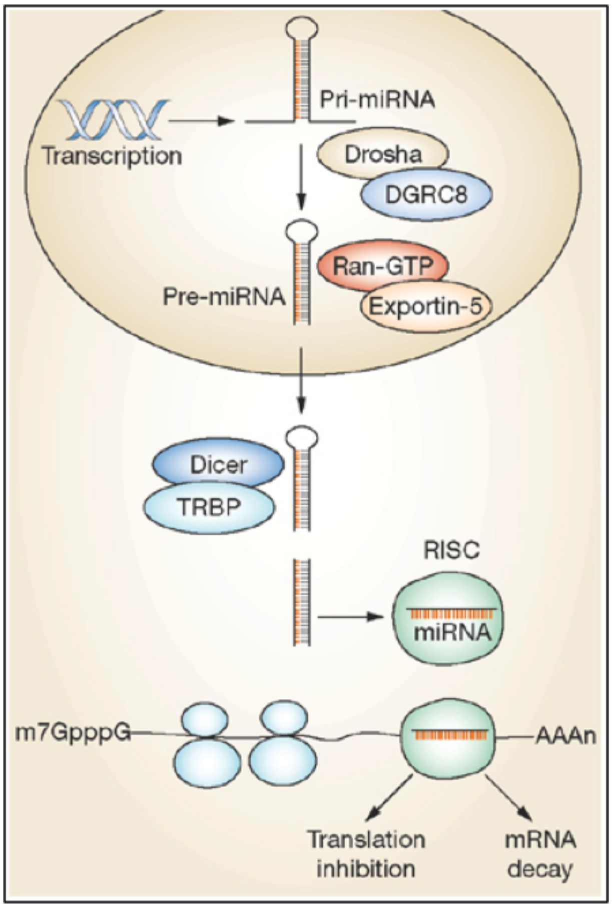
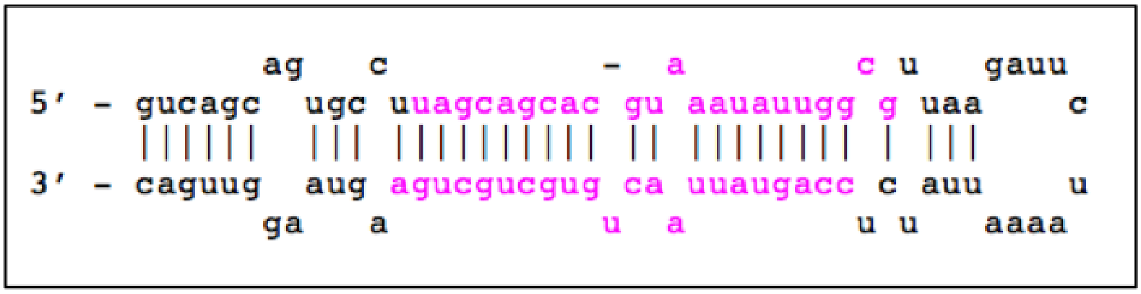


Figure 2.4: MicroRNA biogenesis. Image adapted from Tili et al. (2008).



```
           ag   c          -  a        c u   gauu
5' - gucagc  ugc uuagcagcac gu aauauugg g uaa    c
     ||||||  ||| |||||||||| || |||||||| | |||
3' - caguug  aug agucgucgug ca uuaugacc c auu    u
           ga   a          u  a        u u   aaaa
```

Figure 2.5: Example miRNA hairpin, hsa-mir-16-1. The purple nucleotides represent the two candidate mature sequences, one on each strand.

2.2.2 miRNA mediated gene regulation

The information given in this section is based on Cenik and Zamore (2011) and Saito and Saetrom (2010).

The RNA-induced silencing complex (RISC) is composed of an Argonaute protein and a small RNA guide, which in this context is a mature miRNA. Argonaute proteins are essential for development and differentiation in humans, and defend the cells against viral infections. Four different Argonaute (Ago) proteins are found in mouse and humans, Ago1, Ago2, Ago3 and Ago4, of which only Ago2 has the known ability to cleave RNA, while the unique functionality of the other proteins are still fully not understood.

The miRNA:miRNA* duplex produced by Dicer contains two mature miRNA candidates, of which only one will be incorporated into RISC. The process of incorporating the mature miRNA strand into an Argonaute protein to produce RISC is illustrated in Figure 2.6. The duplex is first loaded into the Argonaute protein, resulting in what is called pre-RISC. The Argonaute must then determine which miRNA strand to include as its target guide, a process described in the next section. When the mature miRNA strand is selected, the miRNA* strand is evicted from pre-RISC to be ultimately degraded and the resulting complex is mature RISC, ready to identify the target mRNA through the incorporated miRNA guide strand.

The Argonaute protein of RISC pre-organizes the 'seed region', nucleotides 2-7, of the miRNA so that their base edges are displayed and ready to at minimum partly pair with the corresponding nucleotides in the 3' UTR of the target mRNA. When paired to the target mRNA the activity performed by RISC depends on which Argonaute is active. All Argonautes may through a degree of complementarity perform destabilization or translational repression, while Ago2 has an additional slicer activity, of which a high degree of complementarity guides a cleavage of the target mRNA between its nucleotides across from the guide nucleotide 10 and 11.

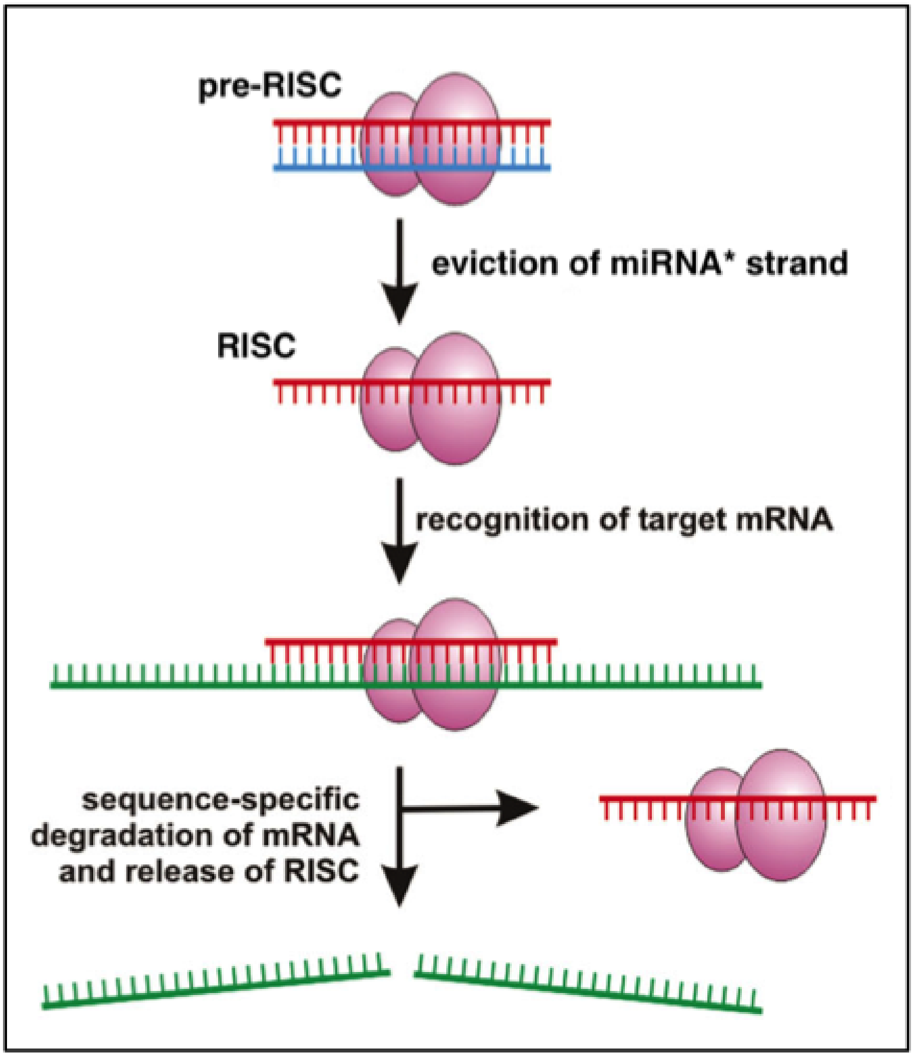


Figure 2.6: RNA-induced silencing complex (RISC). Image adapted and modified from Rutz and Scheffold (2004).

2.2.3 Argonaute strand preference and sorting mechanisms

The process of which pre-RISC selects one strand of the double stranded miRNA:miRNA* duplex for incorporation has been studied in multiple organisms. A strong bias is found toward the strand with the thermodynamically less stable 5' end; however, an Argonaute-dependent preference has also been reported. For *Drosophila* and *C. elegans*, miRNAs are sorted among the different Argonaute proteins based on the structural characteristics of their precursors (Tomari, Du, & Zamore, 2007). In *Arabidopsis*, the miRNAs are strictly sorted among Ago1-Ago5 by the 5' terminal nucleotide of each strand, independent of miRNA size or what biological pathway produced it (Mi et al., 2008). Attempts at verifying the existence of similar global sorting mechanisms on a handful of sequences of human cells have not been successful (Dueck, Ziegler, Eichner, Berezikov, & Meister, 2012; Meister et al., 2004).

A more recent study on human cells reported a possible existence of a unique sorting system operating on a small scale (Burroughs et al., 2011), and alternative cleavage sites for a pri-miRNA have been found to yield isomiRs with different strand preferences (Seong et al., 2014). Polikepahad and Corry (2013) reported an Ago1 preference for adenine as the 3' terminal nucleotide, and an Ago2 preference for 3' uracil in mice. A similar preference for 3' uracil has been suggested but not yet confirmed for human Ago2, however a bias against 3' adenine has been observed (Kandeel et al., 2014). Elkayam et al. (2012) further reports a preference for either A or U as the 5' terminal nucleotide of human miRNAs due to structural requirements of the Argonaute proteins.

Whatever strand is selected as the guide strand of RISC, it is strongly bound to the Argonaute protein. An Ago protein consists of multiple domains, and the 5' end of a miRNA is bound to the MID domain, and the 3' end to the PAZ domain of the Argonaute, leaving room for the stretch of miRNA to be contained in between . The 5' end of a miRNA is found to continously be bound to the MID domain, however there has been two different models for how the 3' end is bound. The *fixed-end* model states that the 3' end of miRNAs are also continuously bound, while the *two-state* model presents a more flexible, repeating binding and releasing of the 3' end dependent on the miRNA pairing with target mRNAs (Cenik & Zamore, 2011). Accordingly, the 3' end of miRNAs that are base paired with mRNAs will be released from Ago. The latter model has gained more support the last years, with multiple experiments revieling release of miRNA 3' end during mRNA basepairing (Sasaki & Tomari, 2012) (Y. Wang et al., 2009).

2.2.4 IsomiRs

As described in section 2.2.1, a pri-miRNA hairpin is processed into a ∼22 nt miRNA:miRNA* duplex representing two possible mature miRNA sequences. In reality, the hairpin of a pri-miRNA can give rise to a variation of expressed miRNA sequences, conventionally named isomiRs after Morin et al. (2008). Of all variations observed, the most abundant sequence is regarded the mature sequence of the hairpin and used for reference, while all others are regarded isomiRs of the same miRNA. All isomiRs are distinct sequences, and can be either templated or non-templated, depending on whether the sequence can be found within the pri-miRNA. Templated isomiRs can arise by a shift in the cleaving site of either Drosha-DGRC8 or Dicer-TRBP, resulting in variations in both ends of the miRNA:miRNA* duplex and thus trimming of the expressed sequences. Post-transcriptional removal of nucleotides in either end of the sequence also results in templated isomiRs. Post-transcriptional addition or substitution of nucleotides may however yield non-templated isomiRs, of which 3' additions of Adenine and Uracil are particularly common (Wyman et al., 2011). An example of possible isomiRs of the 5' strand of hsa-mir-16-1 are presented in Figure 2.7.

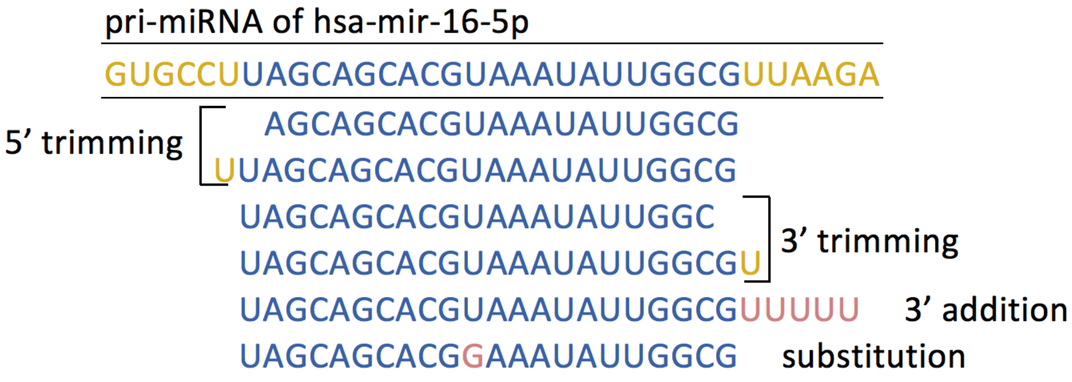


Figure 2.7: Example of possible isomiRs of the 5' strand of hsa-mir-16-1. The mature sequence is presented in blue, and its flanking yellow nucleotides represent the surrounding nucleotides from the pri-miRNA. The first four isomiRs illustrates trimming of the expressed sequence due to shift in cleavage positions, and results in templated variations. The last two isomiRs are non-templated, due to 3' additions and substitution not matching the pri-miRNA.

IsomiRs are found to interact with Argonaute proteins just as mature miRNAs (Cloonan et al., 2011), and recent studies have found them to be of functional and evolutionary importance (Tan et al., 2014).

2.3 Approaches for miRNA isolation

When studying miRNAs, the experiments often depend on Argonaute proteins, as these are the only known proteins to frequently interact with miRNAs to perform biologically important functions. Obtaining cell samples of miRNA and Argonaute interactions can be done by different approaches, two of which are relevant to this report. One approach is to genetically modify the DNA of an organism, effectively inactivating the gene, known as gene knockout. Another approach, which may be performed in combination with gene knockout, is to precipitate a desired protein from a cell sample by immunoprecipitation. The information presented in this chapter is based on information from Kaboord and Perr (2008) and Z. Wang (2009).

2.3.1 Gene knockout

Gene knockout (KO) is a procedure of which the DNA of an organism is modified to effectively inactivate a known gene. This approach is usually taken when the sequence of a specific gene is known while the true function of the gene is not. The function of the gene can then be studied by comparing the knockout organism with normal, wild-type individuals (WT), and any dissimilarities in behaviour or physicology can be investigated as possible effects of the inactivated gene. This procedure is most frequently used on mice due to the simplicity of modifying mouse DNA, and knockout mice are commonly used as animal models for human physiology and behaviour in experiments. When studying miRNAs and Argonaute proteins, the function of Argonaute proteins can be studied by performing Argonaute KO and comparing the KO individual with WT individuals. Investigating the dissimilarities between the two may provide insight into the function of Argonautes and miRNAs, and when performing KO on the different Ago proteins, the functions of them can be compared.

Gene knockout has some limitations. Naturally, mice do not share the same genome and physicology as humans, and observations in mice may not be transferable to humans. Also, altering the DNA of an organism may alter essential functions and processes in the organism and essentially turn lethal, resulting in studies of only the developmental stages of an organism and not the entire life cycle.

2.3.2 IP procedure

A cell sample may contain thousands of proteins and RNA segments, of which often only a specific subset is of interest to a study. When studying miRNAs and their association with Argonaute proteins, the desired sample will typically contain only Argonaute proteins and short RNA segments associated with them. To accommodate this need, a common procedure to use when studying miRNAs is immunoprecipitation. Immunoprecipitation (IP) is a procedure that uses high affinity antibodies to extract a specific protein from a sample. An appropriate antibody is chosen for the desired protein, which when incubated with the sample will bind to the protein. A solid substrate specifically designed to bind to the antibody, called beads, are added to the solution after incubation. When the protein-antibody complexes are bound to the beads, the solution is centrifuged, leaving the heavier beads and their bound complexes at the bottom and all lighter components of the sample on top. The supernatant is removed, and the beads are washed to remove non-specific binding. An illustration of the general procedure is given in Figure 2.8. Immunoprecipitation can be carried out in various ways, depending on the chosen antibody type, incubation conditions, bead type and washing procedure. The different approaches have different advantages, and might yield different results. Two approaches commonly used differ mainly in the choice of beads, where the traditional practice utilizes agarose beads and the more novel practice utilizes magnetic beads. Another approach is taken when known antibodies are unobtainable, of which specific tags are engineered onto the proteins of interest. These approaches are explained in the following sections.

2.3.2.1 Agarose and magnetic beads

Traditionally, the beads used in IP has been agarose beads. Agarose beads are highly porous sponge-like structures, with a large surface area resulting in a high binding capacity. Agarose beads will only bind with the desired protein if its surface is coated in antibodies, any region not coated will bind to whatever protein that sticks. The cost

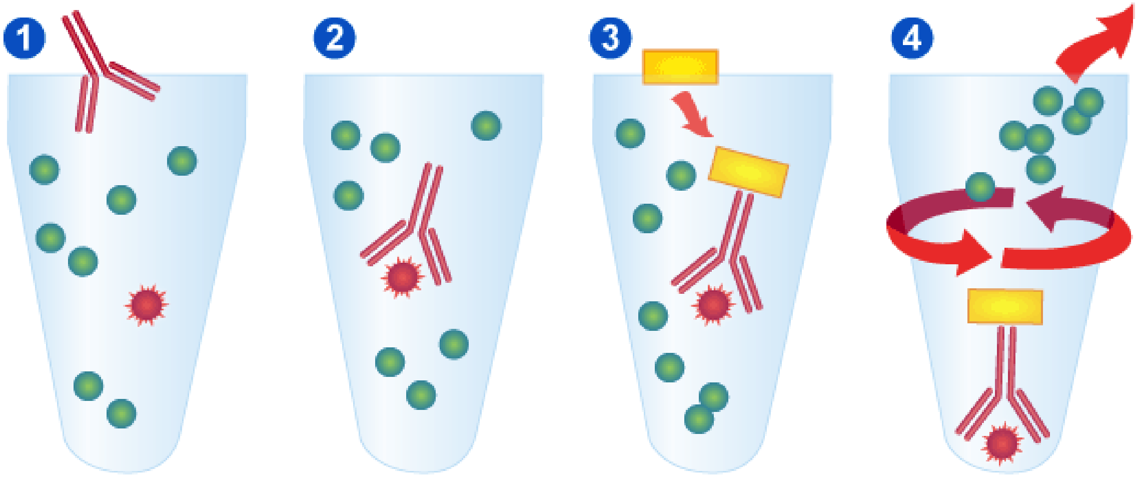


Figure 2.8: Stages of a simplified immunoprecipitation procedure. First, a suitable antibody is added to the sample. Second, the antibody binds to the protein of interest during the given incubation time. Third, the suitable beads are added to the sample and binds to the antibodies, making antibody-protein complexes insoluble. Last, the solution is centrigufed, the supernatant is removed and the beads washed. Image adapted from Leinco Technologies (2015).

of ensuring precision of the process might thus be high in terms of antibody amounts per bead. Another drawback of this approach is the delicate nature of the removal of supernatant. A perfect separation of beads and supernatant is very difficult, and the result often includes some supernatant or lacks some of the beads. The physical stress on the proteins from repeated centrifugation is also a disadvantage, especially if the protein complexes in question are fragile.

Some of the disadvantages of agarose beads have caused an increase in the use of the more novel magnetic beads. Compared to agarose beads, magnetic beads are smaller and with an even, spheric surface, resulting in a significantly reduced surface area, which in turn drastically reduces both the antibody cost and the binding capacity of each bead. However, the small and even size of the beads allow for a higher amount of beads per sample volume than with agarose beads, reducing the binding capacity loss. The main advantage of this approach is in the last stage of the procedure, where the beads are separated from the sample by using magnets as opposed to centrifugation and supernatant removal. This ensures minimum loss of precipitated proteins, reduces background noise, allows for more fragile complexes to be precipitated and is less time consuming than repeated centrifugations.

2.3.2.2 Tagged proteins

Immunoprecipitation is dependent on the availability of antibodies for the protein of interest, and proteins lacking available antibodies are unable to be immunoprecipitated. An alternative approach is to engineer specific tags onto the proteins of interest, of which known antibodies are available. The procedure is then similar, and can be implemented with either agarose or magnetic beads. In addition to enabling any protein to be immunoprecipitated, this approach is highly reproducible as the same tags and antibodies can be reused multiple times. However, engineering tags onto proteins might obscure the natural protein functions or introduce unnatural functions. Additionally, this approach precipitates overexpressed, tagged proteins and not the biologically natural proteins. The result might thus be of tagged proteins of obscured functionality, and the biological relevance of such results are questionable.

2.4 RNA Sequencing

The total amount of transcripts and their quantity in a cell at a specific point in time defines the transcriptome of the cell (Z. Wang, Gerstein, & Snyder, 2009). The transcriptome changes in time due to development and physiological changes in the cell, and cataloguing all transcripts and studying their expression levels in different cells, tissues or stages under the same or different conditions has proven powerful for understanding the functional elements of the genome. Different methods and technologies have been developed for analysing the transcriptome, and currently the most commonly used method is RNA-Seq (RNA sequencing) (Z. Wang et al., 2009), which the data used in this project is produced by.

RNA-Seq is an approach where the advantages from next generation sequencing are utilized, allowing high throughput and quantitative analysis of the entire transcriptome. As such, the whole transcriptome can be analysed, reads classified into different RNA types such as total RNA, mRNAs, and miRNAs, and their differential expressions determined in a limitless manner.

There is a still increasing number of different platforms for RNA-Seq, however they all follow the same basic principles, illustrated in Figure 2.9 (Farazi et al., 2012). The first step is to isolate the total RNA from the cell group in the sample. An adapter sequence is then ligated to the 3' end of each RNA segment, possibly including a sample specific barcode if more than one sample is to be processed simultaneously to reduce cost and overhead, allowing identification and subsampling later. Another adapter sequence containing a primer is then ligated to the 5' end of the RNA segment. The resulting sequences, consisting of the 5' adapter, actual RNA segment and 3' adapter, are then reversely transcribed into complementary DNA (cDNA) segments. These individual segments are amplified using a common amplification technique in molecular biology, polymerase chain reaction (PCR), resulting in a collection representing the relative expression levels of the original RNA transcripts. This collection is known as a cDNA library, and is used as input to a next generation sequencing platform for RNA sequencing.

The result generated by the high-throughput sequencing platform is typically in the form of FASTQ or FASTA, text-based file formats where every individual read is represented as an entry, accompanied by its quality score (only in FASTQ) and associated information such as alignment identifier, sequence read count and comments. Both formats are presented in Figure 2.10 for illustration.

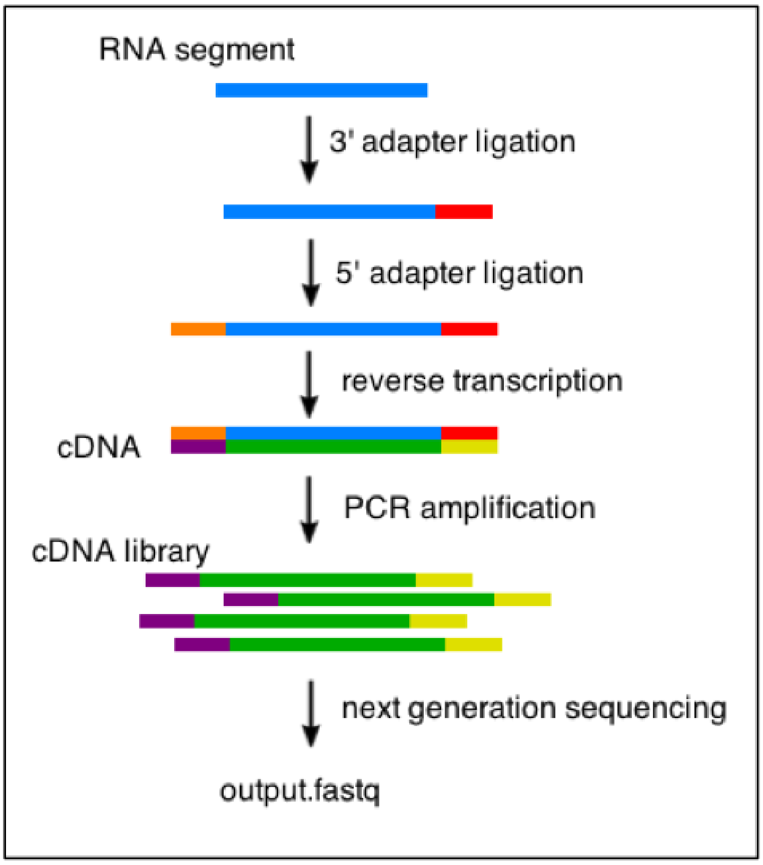
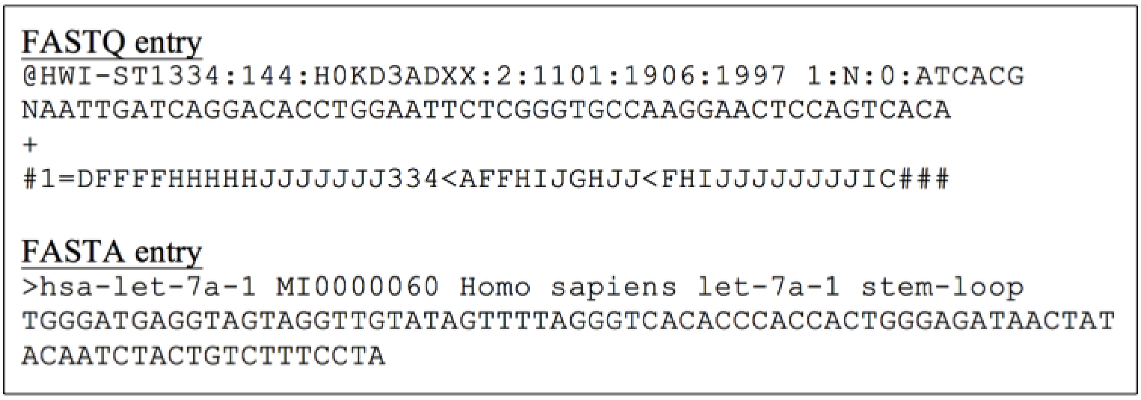


Figure 2.9: RNA-Seq process.



FASTQ entry

```
@HWI-ST1334:144:H0KD3ADXX:2:1101:1906:1997 1:N:0:ATCACG
NAATTGATCAGGACACCTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACA
+
#1=DFFFFHHHHHJJJJJJJ334<AFFHIJGHJJ<FHIJJJJJJJJIC###
```

FASTA entry

```
>hsa-let-7a-1 MI0000060 Homo sapiens let-7a-1 stem-loop
TGGGATGAGGTAGTAGGTTGTATAGTTTTAGGGTCACACCCACCACTGGGAGATAACTAT
ACAATCTACTGTCTTTCCTA
```

Figure 2.10: Example entries in FASTQ and FASTA format. The first line of every entry contains information about the entry, the following lines contains the nucleotide sequence of the read, and in FASTQ format, the quality score is given in a seperate line following a '+' sign.

The sequence read counts, or read frequencies, should be converted to relative frequencies for the sample by normalizing to the total sequence reads for the sample, a procedure described in detail in section 3.5. These frequencies are relative frequencies within the same sample; if absolute values for the sequences are preferred, this is obtainable by normalizing against known amounts of calibrator sequences that were added during cDNA library preparations.

2.5 NGS data processing

When the sequencing results are available, a series of processing steps must take place before the sequencing data is ready to be analysed. First, the reads may include adapter remnants, and these must be removed. To reduce the computational cost and improve accuracy when processing miRNA data, a typical second step is to merge duplicate reads to reduce the data size. Third, the reads must be aligned to a reference genome or a set of known sequences. This section elaborates these three pre-processing steps, briefly presents the technology available for the different tasks, and explains in detail the algorithm behind the most commonly used sequence alignment tools.

2.5.1 Adapter removal

The resulting reads from most sequencing platforms are usually longer than miRNAs, and may contain parts of or the full 3' adapter ligated to the sequence during cDNA library preparations. Identification of adapter-containing reads and removal of the adapter sequences is necessary for obtaining the original RNA sequences, and there is a range of tools that serve this purpose. These tools differ in accepted input file format, algorithms and functionality, such as whether reads containing adapter remnants should be discarded or trimmed, and tolerance of insertions or deletions.

For this work, the exact adapters are known and their position is at the 3' end of the reads, resulting in a simplified identification process. All miRNA reads used for this paper was found to contain at least part of a 3' adapter, however there are two different scenarios that come to play. As illustrated in Figure 2.11, the read may either run into the adapter, or the adapter is within the read. For both scenarios, all remnants of the adapter and eventual following characters (nucleotides) should be removed.

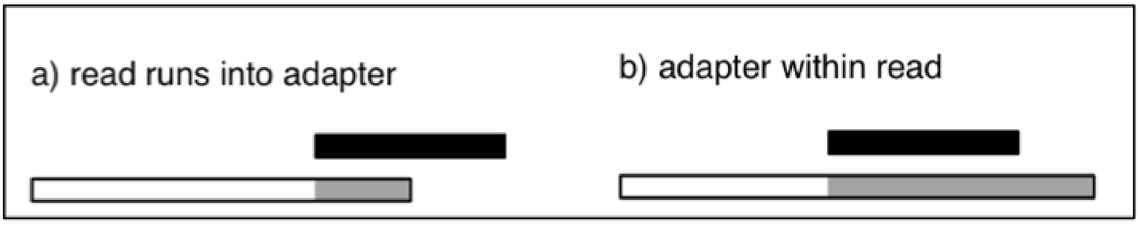


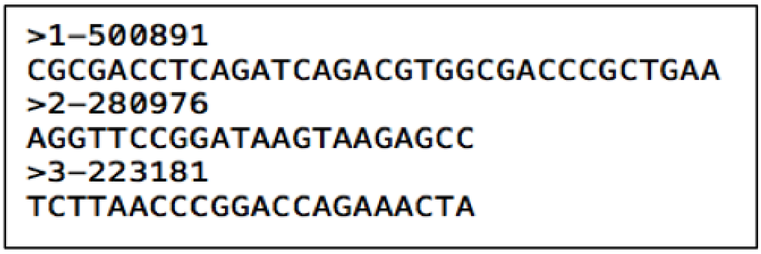
Figure 2.11: Adapter alignment scenarios. White rectangles are reads, black are adapters and grey are segments removed. Image adapted and modified from Martin (2011).

In theory, the adapter sequence remnants in the read should have a perfect match with the known adapter sequence ligated to the read in the cDNA library preparations. However, due to a non-negligible sequencing error rate in current technology, requiring an exact match might discard many valid reads. A more preferred approach is to use semi-global alignments (Gusfield, 1997), which when altered to penalize initial gaps in the read sequence, performs better. Cutadapt (Martin, 2011) is a much used command line tool that implements such a modified semi-global alignment algorithm, and this is the tool used in this project. Cutadapt computes an optimal alignment between the adapter and the read, calculates an alignment error rate and trims or discards the sequence based on whether the error rate is within defined limits.

2.5.2 Merging of duplicate reads

Duplicate reads are common, and they need to be handled, as processing of a substantial amount of identical reads in the following, computationally heavy procedure is highly unnecessary. It is often observed a high degree of duplicates in short RNA processing, so for this project duplicate handling is essential. The current practise is not to remove duplicates, but rather merge them and retain information on the amount of identical reads to use in later analysis.

A tool used for this project is the module *fastx_collapser*[1] from the *fastx* toolkit, which accepts either FASTQ or FASTA input formats and collapses identical reads into a single entry, retaining the total amount of the read in the original data. The entries are sorted by decreasing expression levels, and each entry is given an identifier, which is usually its position in the sorted list, followed by the total expression of the read in the sample, as exemplified in Figure 2.12.



```
>1-500891
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAA
>2-280976
AGGTTCCGGATAAGTAAGAGCC
>3-223181
TCTTAACCCGGACCAGAAACTA
```

Figure 2.12: Example of fastx_collapser output. The first line of every entry explains the expression rank of the read in the sample, followed by the total expression separated by a dash.

2.5.3 Sequence alignment to reference genome

To provide for expression analysis on sequencing data, the high throughput data must be aligned to a reference genome or a set of annotated sequences to identify where the sequences are transcribed from. This is a computationally heavy procedure as high throughput data is immense, and in attempting to overcome this challenge much research has been conducted and many different algorithms and platforms for the computation have been developed. The existing mapping tools have their individual specificities; however they share many similarities, and most utilize either hash table based algorithms or Burrows-Wheeler transform (BWT) based algorithms (Schbath et al., 2012). The main challenges of mapping algorithms are handling the genome size in an efficient manner, allowing for a

[1] Available from `http://hannonlab.cshl.edu/fastx_toolkit/index.html`

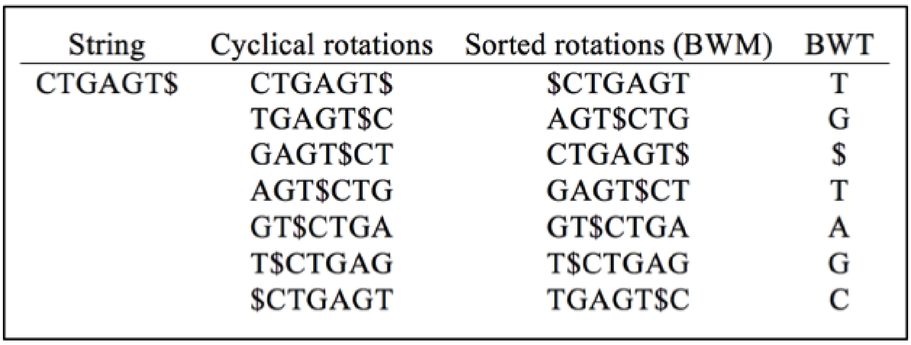
defined number of error/mismatches, locating multiple alignment matches and achieving acceptable performance regarding computational time and memory usage.

In general, hash table based algorithms split the genome into $k-mers$, find all locations of each $k-mer$ in the genome and hash the results in a list, essentially creating a hashed index of the genome. When searching for alignment matches for a read, the read is also split into $k-mers$, and a hashing procedure finds the corresponding $k-mers$ in the index, extracts the possible positions, and accepts a match if succeeding $k-mers$ have succeeding positions. This approach does not allow for mismatches, demands too much memory, and as the genome only has four characters and short combinations of these tend to appear multiple times, it will spend an unnecessary amount of time investigating identical sequences. Modifications are however available to overcome some of these challenges, such as only searching for the first 5' $k-mer$ of a read (*a seed*) and extending the search if a match is found (*seed and extend*), allowing mismatches in only n of the total $k-mers$ of the read (*pigeon hole principle*), and using seeds with “don't care” positions that are not evaluated in the search (*spaced seeds*). However, even if the algorithm first searches for a seed match, with RNA Seq data there is often many hits for a seed and all these must be extended, thus essentially many identical sequences are processed. The extend process is also time consuming, as it usually includes a form of dynamic programming for calculating the distance between the read and the genome segment following the matched seed. The range of approaches to error handling has proven efficient for finding partly identical matches; however if the errors are uniformly distributed among the $k-mers$ or seeds, it will be slow and insensitive. In summary, hash table based algorithms are not optimal regarding computational time as it scans identical sequences multiple times, while the error handling is efficient as long as the errors are not uniformly distributed.

While hash table based algorithms efficiently allows for mismatches, BWT based algorithms have a limited error handling ability with a heuristic that do not scale well on increasing number of errors. However, while hash table based algorithms have great memory requirements and slow computational times, BWT based algorithms are faster and more memory efficient, as BWT based algorithms only scan a repeating sequence once, and have a much more compact genome representation. Newer tools tend to prefer BWT based algorithms (Schbath et al., 2012), an approach taken by a common tool also used in this project, *Bowtie* (Langmead, Trapnell, & Pop, 2009). The next section presents in detail the BWT approach taken by Bowtie.

2.5.3.1 Burrows-Wheeler transform

The Burrows-Wheeler transform was discovered in the 1980s and early 1990s and published in Burrows and Wheeler (1994). It was intended for lossless text compression, and can as such be very efficient for searching in an input text. Burrow-Wheeler transformation of a string is a reversible permutation of the characters in the string. First, a character x not yet present and lexicographically smaller than any character in S is appended, and the resulting string is regarded as a cylinder. All cyclical rotations of $S+x$ are constructed and when lexicographically sorted, the BWT of the string is in the last column of the resulting Burrows-Wheeler matrix (BWM). Figure 2.13 illustrates this computation of the BWT of an input string S = ‘CTGAGT’ with x = ’$’.



String	Cyclical rotations	Sorted rotations (BWM)	BWT
CTGAGT$	CTGAGT$	$CTGAGT	T
	TGAGT$C	AGT$CTG	G
	GAGT$CT	CTGAGT$	$
	AGT$CTG	GAGT$CT	T
	GT$CTGA	GT$CTGA	A
	T$CTGAG	T$CTGAG	G
	$CTGAGT	TGAGT$C	C

Figure 2.13: Burrows-Wheeler transform of string S = 'CTGAGT' and x = '\$'. The second column holds all permutations of the string $S + x$, the third column represents the BWM and the last column the BWT of S.

The BWM has one important property called the Least-First (LF) Mapping, describing a correlation between the characters in the first and last column of the sorted rotations. Specifically, the i^{th} occurrence of any character x in the last column, L, corresponds to the i^{th} occurrence of x in first column, F, as well as in the input string. These LF-links are utilized when recovering the input string from the transform: the last character of the input string is the first character in L, so starting at $L[0]$, the LF-link $LF[0]$ returns the index i of $L[0]$ in F, and the preceding character of the input string is then $L[i]$. Recursively following these links results in the original string input. See Figure 2.14 for illustration of recovering the string 'acaacg' from the BWT 'gc\$aaac' using LF-mapping.

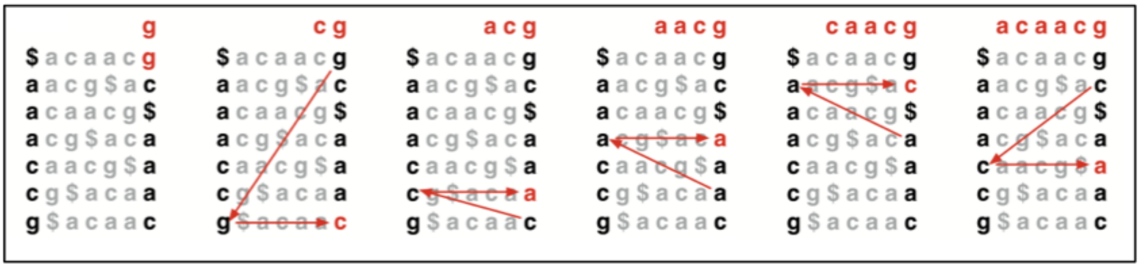


Figure 2.14: Recovering the input string of a Burrows-Wheeler transform using LF-Mapping. Image adapted and modified from Langmead et al. (2009).

When the characters are sorted by their right-context in this manner, the same characters tend to group together in the last column. When transforming RNA data, there are only four characters, and as the first column is sorted, it will contain many repetitions of As, Cs, Gs and Us, in that specific order. All instances need not be maintained, only the total number of the four characters. When recovering the original sequence or searching within it, the whole F is not required. Say there are 100 As, 80 Cs, 95 Gs and 120 U's, with $L[i]$ being the 60$^{\text{th}}$ occurrence of G in L, the LF-mapping is simply 100+80+60 = 240. The same compression of L is possible, although it will contain more entries, as it is not sorted. As F only has four entries and L is at most the same size of the input string, the memory footprint is relatively small compared to hash table based approaches.

The searching procedure by LF mapping does not allow for mismatches as only existing links are investigated, and different approaches have been made to overcome this constraint. *Bowtie* overcomes this by backtracking to a previously matched position, substitutes the nucleotide with a random one, and continuing the search from that point forward. To mitigate excessive backtracking, two indexes are used: a 'forward' index consisting of the BWT, and a 'mirror' index containing the reversed BWT, each index used for aligning opposing halves of the query. Using the two indexes and a limited number of allowed mismatches constrains the number of mismatches allowed in one half if already aligned with mismatches in the other, and reversely, reducing excessive search. Additionally, an upper limit on the number of allowed backtracks are set, and if reached the search is terminated.

2.6 Rationale

Since the recognition of miRNAs, a still increasing amount of research has been conducted to attempt to understand the miRNA biogenesis and pathways. Searching for articles regarding microRNAs in *Bibsys Ask*[1], *Pubmed*[2] or other scientific search engines return over 100 000 results, illustrating the focus miRNAs have gained the last decade. Most experiments and research have studied the miRNA biogenesis and miRNA mediated post-transcriptional regulation, focusing on the mature miRNA sequence, the guide strand, of the miRNA:miRNA* pair, disregarding the passenger strand miRNA* and assuming it to be rapidly degraded. The last few years, an increasing number of studies have focused on the passenger strand (Mah, Buske, Humphries, & Kuchenbauer, 2010) with results indicating that also these strands have a function other than degradation, and even coexistence of 5p and 3p guide sequences are being reported (Choo, Soon, Nguyen, Hiew, & Huang, 2014).

The degradation processes of miRNAs have recieved little attention, due to a perception of mature miRNAs being stable molecules protected by the Ago protein in the RISC complex (Zhang, Qin, Brewer, & Jing, 2012), and degradation of passenger strands have not been of interest. However, the need for a robust miRNA regulation is illustrated by the large amount of research that connects dysregulation of miRNAs with diseases, where decay is one of the expected regulatory processes (Ruegger & Grosshans, 2012). 3' tailing of guide strands are commonly observed and percieved as markers for miRNAs about to be degraded, a process which requires the 3' end of miRNAs to be accessible and not protected by an Argonaute protein, indicating that the *two-state* model might be more accurate than the *fixed-end* model. The review of Ruegger and Grosshans (2012) presents the current understanding of miRNA degradation, and based on the scarce research available on the subject, they conclude that miRNA degradation is a process probably of much higher significance than assumed, and that the degradation might be influenced by mRNA pairing which contradicts the current perception of miRNA/mRNA interactions being a one-way process.

The majority of research on miRNAs has studied RNA sequences filtered on read lengths of approximately 18-26 nts. This is due to an assumption that shorter reads are degradation products or Ago2 cleavage products, an assumption commonly agreed upon. The guideline for RNA-Seq by Farazi et al. (2012) recommends to evict reads not within the range

[1] http://www.ask.bibsys.no

[2] http://www.ncbi.nlm.nih.gov/pubmed

[16,25] nts due to the same assumption, which exemplifies a practice widely used. This has resulted in a void of research on miRNA sequences smaller than 16 nts.

J.P. Mossin conducted a study in 2014 (Mossin, 2014) with a primary focus on differential isomiRs, variants of the same miRNA, in high-troughput sequencing data. When searching for Ago2 cleavage products, he discovered that a group of small ∼10 nt reads aligned with the 3' end of mature sequences, which could not be cleavage products nor explained by previously reported degradation processes. The experiment was performed on high-throughput sequencing data from mouse, and Mossin proposed an extension of this particular part of his study to other and bigger datasets to decide whether the results were significant and reproducible. This was the starting point for my work in Wahl (2014), where I successfully reproduced Mossins results by studying reads of length 11-15 nts in the Meister (human) and Lundbæk (mouse) datasets, presented in section 3.2. I also performed a range of analyses discouraging the current short read assumptions, concluding that these short reads might be degradation products of unknown processes and/or results of an unknown biological function.

The work behind this report is an extension of my work in Wahl (2014), where the main intention is to reproduce my findings on a wider range of datasets. If so, the possibilities of short reads being either degradation products or results of an unknown biological function should be further explored. Possible Argonaute dependencies should be investigated, including any tendencies for a sorting mechanism influencing short read association. Features of short reads should be analysed, and an attempt at identifying miRNA features that may predict short read association of miRNAs should be performed. Any significant results of these analyses may provide better understanding of the role of miRNAs, the origin of short reads, and implicate an updated model of miRNA functionality and decay.

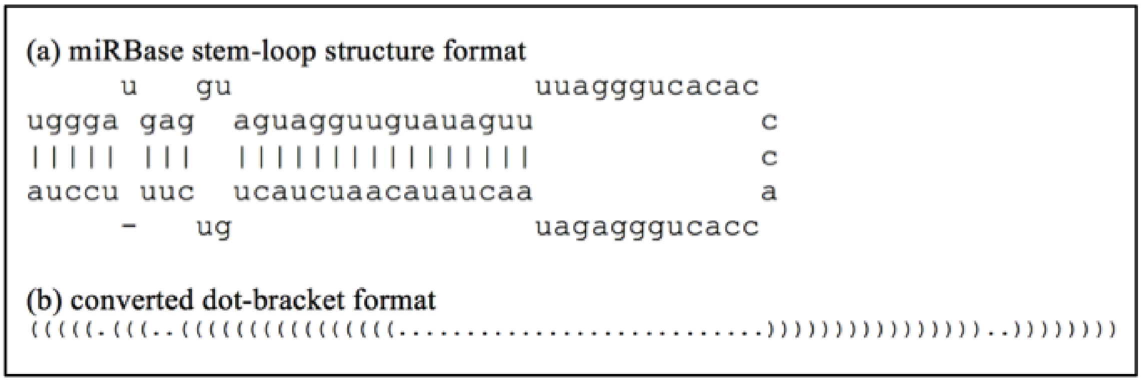
Chapter 3

Data and Methods

The datasets, methods and tools used in this project are briefly presented in this section.

3.1 Reference data

Conducting read mapping in this project has required existing reference data, which has been obtained from miRBase version 21 (Kozomara & Griffiths-Jones, 2014), the primary public microRNA sequence repository. The data obtained from miRBase is existing miRNA hairpins, mature miRNA sequences and miRNA stem-loop structures. The files available from miRBase contain annotated sequences across multiple genomes, and for the purpose of this project sequences representing human and mouse were filtered out, separately. The stem-loop structure data was in a format suitable for visualization only, and was converted to the more convenient dot-bracket format, where brackets and dots represent a base pairing or lack of it, respectively. An example of this conversion is presented in Figure 3.1.



(a) miRBase stem-loop structure format

```
     u   gu               uuaggucacac
uggga gag  aguagguuguauaguu           c
||||| |||  ||||||||||||||||           c
auccu uuc  ucaucuaacauaucaa           a
     -   ug               uagaggucacc
```

(b) converted dot-bracket format

```
(((((.(((..((((((((((((((((.........................))))))))))))))))..))))))))
```

Figure 3.1: Example of conversion from visual stem-loop structure format (a) to dot-bracket format (b).

3.2 Sample data

The analyses presented in this paper have been conducted on six different sample datasets, three from mouse and three from human. Five of these are produced by immunoprecipitation (section 2.3.2), while one present data from knock-out experiments on Ago2. Samples of immunoprecipitated Argonaute proteins are denoted *Ago IP*, while knock-out data are denoted *Ago2 KO*. Details on each dataset and where it can be accessed are presented in this section.

3.2.1 Human sample data

Three human datasets were processed, all of which provides immunoprecipitated Argonaute data. The first dataset is a published dataset by Dueck et al. (2012), available at the NCBI Gene Expression Omnibus[1] (GEO) as *GSE45506*. The experiment aimed to study miRNAs associating with different Argonaute proteins in human HeLa cells, and as such they performed IP of Ago 1 through 4 by using antibodies and agarose beads, as explained in section 2.3.2.1. The samples of interest are Ago1 IP, Ago2 IP and Ago3 IP, and these comprise the dataset denoted *Meister* throughout this report.

The second human dataset was published in Rybak-Wolf et al. (2014), produced to study Dicer targets and binding sites in human and *C. elegans*. The dataset is available at GEO as *GSE55333*, where the samples of interest are Ago2 IP and Ago3 IP. Immunoprecipitation was performed in human embryonic kidney 293 cells by using tagged Ago2 and Ago3 proteins and magnetic beads, as explained in section 2.3.2. The Ago2 IP and Ago3 IP samples of this dataset will be denoted *Rajewsky* throughout this report.

The third human dataset was publised in Burroughs et al. (2010), and is available at the DDBJ Sequence Read Archive[2] under accession number *DRA000205*. The original study aimed to better define the global contours of 3' miRNA additions in human THP-1 cells, and performed IP of Ago1 through 3 by using antibodies and non-magnetic, silica-based polymer beads. The samples of interest are Ago1 IP, Ago2 IP and Ago3 IP, and these comprise the dataset denoted *Daub* throughout this report.

3.2.2 Mouse sample data

The first mouse dataset is a published dataset by Polikepahad and Corry (2013), available at the NCBI Sequence Read Archives[3] by accession number *SRA056111*. The study aimed to determine the functional implications of antisense transcript binding to Argonaute proteins, and performed immunoprecipitation of Ago1 and Ago2 in mouse CD4+ T cells by using known antibodies. They immunoprecipitated Ago1 and Ago2 three times each, resulting in 6 individual samples denoted Ago1a IP, Ago1b IP, Ago1c IP, Ago2a IP, Ago2b IP and Ago2c IP, of which all are of interest and collectively denoted *Corry* throughout this report.

The second mouse dataset was published by D. Wang et al. (2012), in association with studying the functions of individual Argonaute proteins and microRNA activity in mammals. Ago1 through 3 were immunoprecipitated in mouse epidermal cells by using antibodies and agarose beads, and the result is available at the Yi Laboratory webpage[4]. The samples of interest are Ago1 IP, Ago2 IP and Ago3 IP, and throughout this report these will collectively be denoted *Rui*.

[1] http://www.ncbi.nlm.nih.gov/geo/

[2] http://trace.ddbj.nig.ac.jp/dra/index_e.html

[3] http://www.ncbi.nlm.nih.gov/sra

[4] http://yilab.colorado.edu/Data.html

The last mouse dataset was produced by M. Lundbæk, R. Mjelle and P. Sætrom in 2013 in association with yet unpublished work at Department of Cancer Research and Molecular Medicine at NTNU. Their experiment measured miRNA expression levels in Ago2 knockout cells and wild type cells, specifically three knockout samples and three wild type samples from two separate mouse cell lines. Studying these samples may provide further insight to the functionality of miRNA mediated Ago2 activity. The two cell lines will be denoted DOC and GH, and the combined data set of 12 samples will be denoted *Lundbæk* throughout this report.

3.3 Sample data processing pipeline

The main steps in the pipeline outlined in Farazi et al. (2012) have been followed in this project. cDNA libraries were already constructed, and the sample datasets used in this project are the NGS output data from their respective experiments. The data processing pipeline is illustrated in Figure 3.2, and has the sample data in FASTQ format as input to the process.

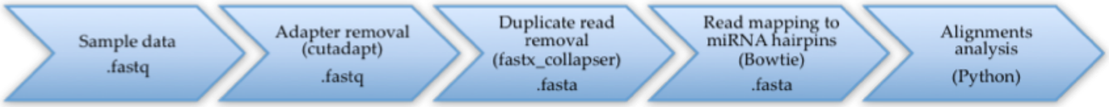


Figure 3.2: Data processing pipeline.

The sample data was already sorted by barcode, so the first step of the pipeline is to remove adapter remnants from the cDNA library preparations. The adapter sequence of the sample must be known prior to the operation, and with an adapter sequence A the trimming is done using *Cutadapt* with the command shown for step 1 in Table 3.1.

To reduce the computational cost in further steps, the second step removes duplicate reads and obtains the total count for each read. This is done by collapsing all identical reads into one entry and maintaining the total expression level for that read in the sample. No sequence length filtering was applied. The command used for performing this with *fastx_collapser* is illustrated in step 2 in Table 3.1, where "–Q 33" declares the correct format for the quality scores in the input .FASTQ file. The resulting .FASTA file contains an entry for every unique read with its associated total expression level and rank, and the entries are sorted on descending expression levels. See Figure 2.12 for an example. During this process the quality scores of the reads are evicted, as there is no reliable method for combining quality scores of multiple reads, however requiring exact alignment in the third step ensures reliability.

The third step is the most common NGS data processing step, in which the reads are mapped against a reference genome. For the human datasets the reference genome is all annotated human miRNA hairpins, for the mouse datasets the reference genome is all annotated mouse miRNA hairpins, both references obtained from miRBase. The mapping was performed using *Bowtie*, by first building bowtie indexes of the two sets of hairpins, and then aligning the sample dataset against its respective indexes. Bowtie provides fast and memory-efficient alignment due to, amongst many aspects, a sensitivity and accuracy compromise. By default, Bowtie reports only one alignment per read, which is not guaranteed to be the best one, and the default allowed number of mismatches is 2. Additionally, Bowtie tries to align both the read and its complementary sequence to the genome, resulting in possibly two alignments from different strands of miRNA hairpins. Bowtie provides for many user-defined options that modify the default settings, and for this study the maximum number of mismatches was set to 0 ("-v 0"), all alignments was to be reported ("-a") with a maximum number of 100 ("-m 100"), the reported alignments was to be the best alignments found ("--best") and reads should only be aligned to the forward strand ("--norc"). The command line for building the bowtie indexes and producing read mappings with bowtie is given in step 3.1 and 3.2 in Table 3.1, respectively.

Table 3.1: Tool commands for each step in the data processing pipeline

Step	**Tool command**
1	cutadapt –a A sample.fastq >trimmed.fastq
2	fastx_collapser -i trimmed.fastq -o collapsed.fasta -Q 33
3.1	bowtie-build hairpin_file.fa hsa_index
3.2	bowtie –f hsa_index collapsed.fasta –a –v 0 –m 100 --best --norc > alignments.txt

3.4 Alignments processing

The read alignments identified by Bowtie are not readily interpretable, and scripts were produced for performing a series of steps to enable flexible analysis of the data. The alignment data is first parsed and desired information stored in an object-oriented manner, before a range of aspects are investigated and visualised. The following section describes this process.

3.4.1 Data parsing

To enable sophisticated analysis, considering only the alignment data is not sufficient. The annotated miRNA hairpins, their stem-loop structures, and their annotated mature miRNA sequences are also required, and along with the alignment data this constitutes the input to the data parsing process. The underlying data structure used for the program is object oriented, and the program steps are illustrated in Table 3.2.

First, the annotated miRNA hairpins from the genome in question (mouse or human) are parsed, and each hairpin is represented as a Hairpin object, with its hairpin identity as object identifier, and the nucleotide sequence as an attribute. For every hairpin parsed, the corresponding dot-bracket format stem-loop structure is read, and stored as another attribute of the Hairpin object.

Table 3.2: Data parsing steps

Step	**Description**
1	Parse annotated miRNA hairpin and stem-loop sequences
2	Parse annotated mature miRNA sequences
3	Parse read alignments
4	Identify mature miRNAs and isomiRs
5	Identify short reads

Second, the annotated mature miRNA sequences are parsed. As there might be multiple mature sequences and unannotated isomiRs per miRNA hairpin, the identifiers from the annotated sequences are discarded. Each annotated mature sequence is given a new, unique identity, containing its corresponding hairpin ID and a suffix consisting of the total number of mature sequences for that particular hairpin at the time of initialization. The first mature sequence found for hairpin *hsa-mir-23a* would have the ID *hsa-mir-23a1*, the second *hsa-mir-23a2* and so on. Each annotated mature sequence is parsed, given a new identifier and represented as a MicroRNA object, with its corresponding Hairpin ID, start index relative to the hairpin, nucleotide sequence, and residential strand stored as attributes.

The third step is to load the read alignments. First, the read frequency of all alignments are read and their sum maintained for normalization (see section 3.5 for details). Then, every alignment entry is parsed, and their information maintained in different data structures. The length of every alignment is inspected, and if the length is in the range [11,15], the read is stored as a short read candidate. If the length is in the range [16,25], the read is investigated as an annotated mature sequence candidate, and if so, the expression level of the corresponding MicroRNA object is incremented by the normalized read frequency of the new alignment. If the miRNA candidate is not an annotated miRNA, the read is stored as an isomiR candidate.

The fourth step has two implementations, and evaluates the list of isomiR candidates. The alignments can either be run with regard to *all* templated expressed isomiRs, or it can regard only the *highest* isomiR if that isomiR is higher expressed than any annotated miRNA for the same strand. If all isomiRs are regarded, any sequence in the range [16,25] aligning to a hairpin with an expression level above 0.5 rpm (see section 3.5) is approved. If only one possible isomiR is regarded, the isomiR retained is the highest expressed one. If multiple isomiRs have the same expression level, the priority is an index offset closest to the annotated sequence, followed by a proximity to $sim22$ nt length of the sequence, closest end index to the annotated sequence and lastly, if all prior criterias are equal for two reads, the smallest one will be retained. In either way, the new isomiRs are saved as new MicroRNA objects with an ID declaring them unannotated. Following the example for hairpin *hsa-mir-23a* in the last paragraph, the ID of an unannotated isomiR would be *hsa-mir-23a5N* and *hsa-mir-23a3N* for 5' and 3', respectively. In either run mode, all annotated and retained isomiRs for the same strand are compared, and the highest expressed sequence is regarded the mature one, disregarding whether it is annotated or not, and the strand with the highest total expression level is regarded the guide strand of the duplex. Finally, if no expressed sequence is found for a hairpin, it is regarded unexpressed.

The fifth step is to investigate the short read candidates to identify true short reads and represent these as short read objects. First, as short reads might align to multiple hairpins, short reads aligned to hairpin strands not expressed in the sample are discarded. Secondly, when the pre-miRNA is cleaved from pri-miRNA, it is not perfectly cleaved at the mature sequences, and short segments from the pri-miRNA immediately adjacent to the mature sequences might be included in the pre-miRNA, called miRNA-offset RNAs (moRs) (Langenberger et al., 2009). Due to this, the short read candidates are evaluated as possible moRs, and if concluded as such, they are maintained as moRs and discarded as short reads. Short read candidates are classified as moRs if they have no more than 2 positions overlap with a mature miRNA sequence and resides adjacent to the 5' end of the 5' mature sequence, or in the 3' end of the 3' mature sequence. Figure 3.3 illustrates the different sections of a hairpin.

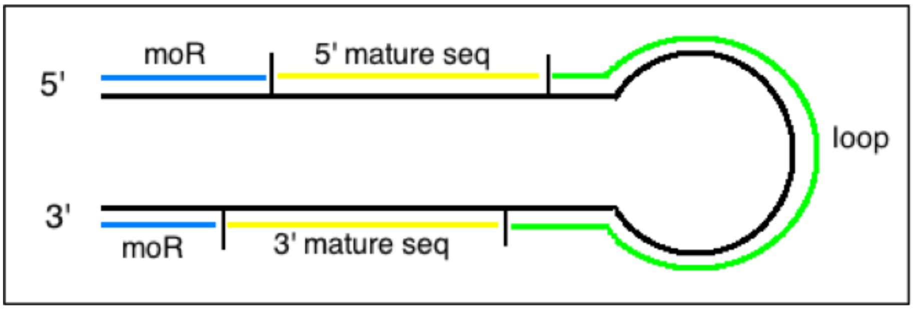


Figure 3.3: Different sections of a pre-miRNA hairpin, where blue is moRs, yellow is the pair of mature sequence candidates and green is the hairpin loop.

If the short read candidate is found to be an actual miRNA short read, different attributes must be identified. First, an alignment is performed to find if the short read aligns best with the 5' or 3' end of the corresponding miRNA sequence, the resulting position being either 'start' or 'end' to avoid confusion with prime definitions. Second, the alignment offset is found between the short read position and the corresponding boundary of the mature sequence: if the short read aligns to the start of the sequence, the offset is between the 5' end of the short read and the 5' end of the mature sequence, if it aligns to the end of the sequence, the offset is defined as the difference between the 3' end of the short read and the 3' end of the mature sequence. See Figure 3.4 for illustration of the alignment positions and offsets. This alignment is performed only against the mature sequence if analysis is to regard only the highest miRNA for each strand, however if all isomiRs are to be evaluated, this alignment must be done against all isomiRs of the corresponding hairpin strand. The isomiR to which the short read aligns best with is chosen, and if multiple isomiRs align equally, the highest expressed one is chosen. When the corresponding isomiR or mature sequence is chosen , all features of the short read is defined, and the alignment is presented as a Shortread object, containing attributes declaring the corresponding Hairpin ID, MicroRNA ID, nucleotide sequence, expression level, alignment position, alignment offset, corresponding prime and a Boolean value declaring whether the read origins from the highest expressed strand of the Hairpin.

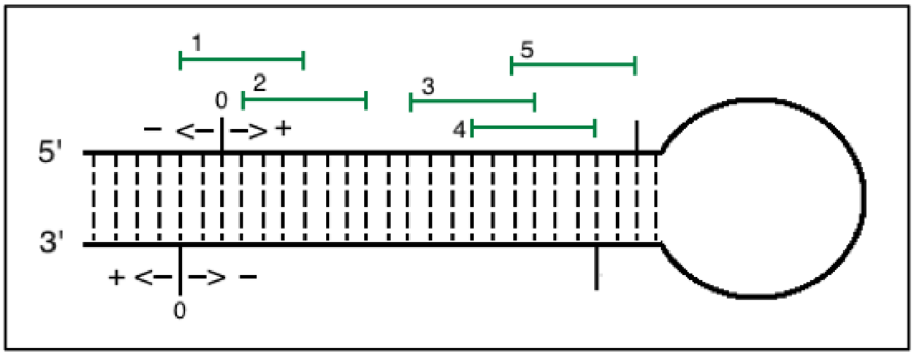


Figure 3.4: Short read alignment to mature sequences. If the shortread is closer to the 5' end than 3' of its corresponding mature sequence, it is positioned at the start, otherwise end. An alignment offset is calculated, and position shifts towards the 5' end are negative, towards the 3' end are positive, as shown for the 3' strand in the figure. Shortreads are shown in green, and shortread 1 and 2 will align to the start with a start offset of -2 and +1, respectively, while shortread 3, 4 and 5 will align to the end, with end offsets of -5, -2 and 0, respectively.

Finally, all MicroRNA objects are investigated, and annotated sequences not found in the sample are regarded unexpressed. The resulting data of interest are the Hairpin objects, MicroRNA objects, and Shortread objects, stored in three separate data structures easily accessible. These structures enable flexible analysis, which are presented in more detail in the next section.

3.4.2 Analysis and visualization

The object-oriented presentation of the processed data is accessible for flexible analysis and visualizations, and for every sample and dataset, a range of analyses are performed. First, a report is printed containing statistics about the sample, such as the number of expressed miRNAs, both annotated and unannotated, and the percentage of these associated with short reads, of which miRNAs residing in the guide strand and passenger strand are reported separately.

Further, most analyses in this project are done through visualizing correlations between different aspects of the data. All visualizations are performed using *Matplotlib* plots (see section 3.7), and three different plot types are used: bar chart, box-and-whiskers plot (boxplot), and scatter plot. For most plots the data visualized have been log^2 transformed to reduce variance, indicated by labels in the plots. Scatter plots and bar charts are relatively uncomplicated to interpret as they only represent two-dimensional data, however boxplots are more complex. Two example boxplot with explanation is given in Figure 3.5. The boxplot with notches, as presented in the right of the figure, show the 95% confidence interval around the median of the data sample. These are useful when comparing data sets: if the notches of two boxes do not overlap, the medians of the two boxes differ with 95% confidence.

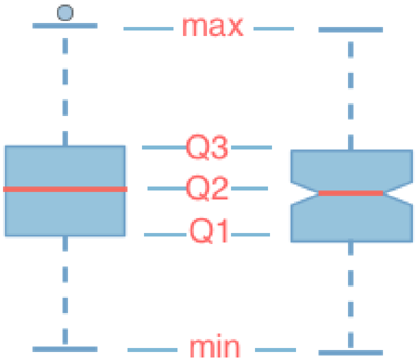


Figure 3.5: Two variants of box-and-whiskers plots. For both, the lower border of the box represents the first quartile ($Q1$) of the data sample, the higher border the third quartile ($Q3$). The median, $Q2$, is represented as a line within the box. The length of the box, IQR, is defined as $Q3 - Q1$, and is used for calculating the whiskers. The lower whisker (min) is given by $Q1 - 1.5 * IQR$, the higher (max) by $Q3 + 1.5 * IQR$. Any data points that resides outside this range may be included if desired, and are then represented as outliers, as illustrated with a circle in the boxplot to the let. Two other variant, represented by the boxplot to the right, includes notches. Notches represent the 95% confidence interval around the median.

Multiple statistical tests were performed on a subset of the data. Sample-specific tests were calculated directly on the Shortread objects, while external scripts calculated tests spanning across multiple samples. For external scripts, the object data needed was written as a matrix to file in a tabular delimited manner, with attribute factors represented by columns and reads represented by rows. More details on the statistical method are given in section 3.6.

3.5 Normalization

For the aligned reads loaded as described in the last section to be comparable with each other and across samples, a normalization of the read frequencies is necessary. After collapsing identical reads, each read is only represented once, and the identity of that read entry contains information on the total expression level from its original sample, as described in section 2.5.2. After aligning each read to the genome using Bowtie (see section 3.2), the same read might have aligned to multiple locations in the genome, and each of these alignments are represented as individual entries in the resulting alignment data. For each entry, information on how many additional locations the read aligned to is given.

For this project no sequence length filter is applied as short RNAs are desired, and as such many short sequences are included in the result. Short RNA sequences may very well align to multiple locations in the genome, which might introduce noise to the analysis, thus the frequency count should be corrected for additional matching sites. The corrected frequency for a single read is given as

$$C_r = \frac{F_r}{L_r} \tag{3.1}$$

where F_r is the collapsed read frequency, L_r is the total number of aligned locations for the read, and C_r is the corrected read count for read r.

To enable comparison across samples, the read counts must also be normalized against the total read count in the sample, which is done following a read-per-million (rpm) scheme given by

$$RPM(C_r) = \frac{C_r}{C_t} * 10^6 \tag{3.2}$$

where $RPM(C_r)$ is the read-per-million normalized count, C_r is the corrected read count, and C_t is the total read count within the sample. Throughout all computations and analyses, the expression level of every read or mature sequences is given as rpm normalized values.

3.6 Statistical methods

To identify statistically significantly differing features regarding miRNA short reads, different statistical methods have been used. The Wilcoxon signed-rank test are used to compare single values between data sets, while the ANOVA test is used for complex comparison of a range of features across data sets.

3.6.1 Wilcoxon signed-rank test

The *Wilcoxon signed-rank test* was published by Frank Wilcoxon in 1945 (Wilcoxon, 1945), and is a useful model for comparing repeated measurements of related samples, where the values cannot be assumed to be normally distributed. It compares n paired absolute values between two samples, and calculates the absolute difference and the sign of the difference. The resulting differences are sorted in an ascending order, and ranked from position 1 to n, discarding all values with a difference of zero. The sign are added to the rank, and the sum of the signed ranks are calculated, resulting in the test statistic value W. The null hypothesis of this test is that the median difference of the pairs is zero. For small n values, the W value is compared against a reference table of $W_{\alpha,n}$ values, and the null hypothesis is thrue if W is smaller than the corresponding $W_{\alpha,n}$. The reference table contain pre-defined W values for samples of specific sizes required to reject the null hypothesis with a certain probability. This probability value, p, denotes the probability of obtaining the observed signed ranks when the null hypothesis is true, and a significance level of α of $p = 0.05$ is used.

3.6.2 Statistical testing across multiple samples

To test for significantly differing features across samples, *Fisher's Analysis of Variance (ANOVA)* was calculated (Fisher, 1925). ANOVA is a collection of complex models for comparing group variances for a broad range of feature definitions, including possibilities for error strata. The basic operations of ANOVA is to perform statistical tests to determine if the means of different variables or samples are equal. ANOVA essentially calculates means and variances for every factor specified, dividing the variances of two factors in turn and calculates an *F-value* for the total data. As F increases, the evidence for evicting the null hypothesis increases, and this value can be used to calculate a corresponding p-value. The output from a computational approach to the ANOVA method is usually an ANOVA table containing the F-value, degrees of freedom, sum of squares, mean squares and p-value, and as with the Wilcoxon signed-rank test, the significance level is set to $p = 0.05$. The features analysed are short read position, residential strand, offset from miRNA sequence, the sample condition and associated Argonaute protein, and the conditional value evaluated is the expression level of short reads.

3.6.3 Implementation of statistical methods

The Wilcoxon signed-rank test was implemented using existing functionality in the SciPy statistics library, while the ANOVA test was implemented with existing functionality in the R language library. The data required for the ANOVA test were written to file in a tabular delimited manner for objects of each sample, to be used by external scripts when calculating the test. The specific function calls are presented in Table 3.3, and more information on the libraries is given in the next section.

Table 3.3: Function call for statistical methods.

Test	**Command**	**Tool**
Wilcoxon signed-rank	stats.wilcoxon(sample1, sample2)	SciPy
ANOVA	anova(lm(response ~ formula, data))	R

3.7 Tools and languages

The analysis of aligned sequence data has been implemented using mainly the *Python*[1] programming language, with the additional python based third party libraries *SciPy*[2], *Biopython*[3], *Matplotlib*[4], and *Numpy*[5]. SciPy is a scientific extension to Python, which includes the statistical library *stats* utilized for sample specific statistical test in this project. Biopython is a set of python based tools that simplifies bioinformatics associated tasks, such as parsing Bowtie output FASTA files (see section 3.2). Matplotlib is a two-dimensional plotting library used to produce publication quality plots presented in this

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

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

[3] http://www.biopython.org/

[4] http://www.matplotlib.org/

[5] http://www.numpy.org/

paper. Numpy provides support for multidimensional arrays and high-level mathematical functions, and has been utilized when calculating the data to be plotted by Matplotlib.

For statistical testing across multiple samples, specifically ANOVA tests (see last section for details), the *R*[1] programming language was briefly used.

All necessary source code for reproducing the results presented in the next chapter can be accessed at the online repository, where the input files accessible are the alignment output files from section 3.4.1 for the *Meister*, *Daub*, *Rajewsky*, *Corry*, *Rui* and *Lundbæk* data:
`https://bitbucket.org/kristwah/mirna-short-reads`

[1] `http://www.r-project.org/`

Chapter 4

Results

The most essential results from this project is presented in this chapter. First, statistics obtained after processing all data sets are presented. Second, the findings of Wahl (2014) are verified, and a general tendency of short reads aligning to miRNAs with offset 0 is found for all data sets, along with a coexpression of short reads and miRNAs. Third, the majority of short read associated miRNAs are found to be within the 20% most expressd miRNAs, indicating that short reads may be related to miRNA activity in the cell. Fourth, the terminal nucleotides of miRNAs are studied, revealing a 5' preference for Uracil regardless of any short read association, and no consistent preference for a 3' nucleotide. Fifth, the lengths of short reads and miRNAs are investigated as an attempt at explaining their origin. Sixth, an attempt at identifying statistically differentiating features across the data sets are presented. Seventh, a classification scheme for miRNA hairpins reveals that the vast majority of short read associated miRNAs originates from hairpins with a clear strand preference, and of these, the majority of short read associated miRNAs reside in the guide strand. Lastly, the results of repeating all prior analyses by regarding all expressed isomiRs are presented.

Throughout this chapter, short reads aligning to the 5' and 3' end of mature sequences are denoted *start reads* and *end reads*, respectively. Also, all presented analyses are based on the sample data sets *Daub*, *Corry*, *Lundbæk*, *Meister*, *Rajewsky* and *Rui*, presented in section 3.2, and will be denoted accordingly.

4.1 Sample data processing

When processing each sample data set as described in section 3.3 and 3.4.1, the resulting object oriented presentation of the data are readily available for analysis. For each data set, statistics from the results are retrieved for each sample, including the number of unique alignments, number of expressed hairpins and miRNAs, the number of expressed short reads and the share of miRNAs associated with short reads. These statistics will be presented in this section, to visualize the differences in quantities between the data sets and aid comparisons between the sets. The statistics for the human and mouse data sets are presented separately in the following sections. Statistics regarding unannotated miRNAs represent isomiRs that differ from the annotated mature sequences provided from *miRBase*, as described in Section 3.4.1. The statistics presented are the mean values from all subsamples of each data set; the sample-specific values, mean and standard deviations are presented in Appendix A.

4.1.1 Processing of human sample data

This section presents results from the processing of the three human sample data sets *Meister*, *Daub* and *Rajewsky*, described in Section 3.2.1. The number of annotated hairpins and mature miRNAs for the human genome is 1,865 and 2,562, respectively. A presentation of the average values from all samples of each data set is presented in Table 4.1.

The number of unique alignments differ drastically between the data sets, as well as the minimum absolute expression level of the top ten most abundant reads of each sample, with 99,834 in *Meister*, 56,001 in *Daub* and 511,611 in *Rajewsky*. All other values are quite similar for *Meister* and *Daub*, while *Rajewsky* show different trends. The number of expressed annotated hairpins and miRNAs, unannotated sequences, and short reads, are more than doubled for *Rajewsky* compared to *Meister* and *Daub*, probably due to higher read depth of its samples. However, the share of expressed sequences that are associated with short reads are similar for all data sets, indicating a correlation between expressed miRNAs and expressed short reads. For *Meister* and *Daub*, Ago2 IP show higher numbers of short reads than Ago1 IP and Ago3 IP, while for *Rajewsky*, Ago3 IP show the highest numbers. The *Rajewsky* data was produced by tagging Ago2 and Ago3 proteins in the sample cell, and as discussed in Section 2.3.2.2, immunoprecipitating tagged proteins might not provide true, biological reads. Additionally, the sequencing depth of this data set is deeper than the others, and in total this introdues some constraint as to the reliability of comparisons between *Rajewsky* and the other data sets. Detailed statistics from each subsample and the standard deviation of the means are given in tables A.1, A.2 and A.3 in Appendix A.

4.1.2 Processing of mouse sample data

This section contains the results from processing the three mouse sample data sets *Corry*, *Rui* and *Lundbæk*, described in Section 3.2.2. The number of annotated hairpins and mature miRNAs for the mouse genome is 1,185 and 2,112, respectively. A presentation of the average statistics from all samples of the *Corry* and *Rui* data sets are presented in Table 4.2. The number of unique alignments differs drastically between the two data sets, along with the top ten most abundant reads in the different samples. Even with lower unique alignments, a greater number of annotated and unannotated miRNA sequences are expressed in *Rui*, while short reads are clearly more frequently observed in *Corry*, with a total of 44.6% of all expressed hairpins in *Corry* being associated with short reads compared to only 3.4% in *Rui*. A closer look at the actual short reads in question for *Rui* show that the normalized expression levels are lower than the expression levels of short reads in *Corry*. The low number of unique short reads for *Rui*, as well as the low expression levels of these, renders comparisons between *Rui* and the other data sets rather restricted and unreliable. The detailed statistics for each subsample of both data sets are given in tables A.4, A.5 and A.6 in Appendix A. An interesting observation is that the standard deviation of the means from the *Corry* data set is quite high. As Corry contains triple samples of the same Ago IP, this is surprising, and might illustrate the level of noise present in the samples.

The *Lundbæk* data set differs from the others by knocking out Ago2 to enable comparison

Table 4.1: Statistics from processing the human sample data sets. All numbers presented are the mean values across all subsamples of each data set. The *Alignments* row presents the number of unique alignments from each data set. The top ten most abundant reads are represented in the *Reads* row, while the total expression level of mature sequences of each strand is represented in the *Expression* row. Short read-association (*SR-associated*) is calculated from the set of expressed annotated sequences and unannotated sequences. When considering mature sequences only, the number of unannotated mature sequences and corresponding SR-association is given in the *Matures* rows; when considering all observed isomiRs, the number of unannotated isomiRs and corresponding SR-association is found in the*IsomiRs* rows.

Data set		**Meister**	**Daub**	**Rajewsky**
Alignments	Unique	10,961	86,323	34,058
Reads	Top 10	99,834	56,001	511,611
Expression	5’	503,009	236,092	300,813
	3’	96,733	68,953	292,553
Hairpins	Expressed	583 (31.3%)	586 (31.4%)	1,308 (70.1%)
	SR-associated	193 (33.1%)	202 (34.4%)	439 (33.6%)
MiRNAs	Annotated	404 (15.8%)	474 (18.5%)	1,393 (54.4%)
Matures	Unannotated	486	482	1,375
	SR-associated	220 (24.7%)	238 (24.9%)	611 (22.1%)
IsomiRs	Unannotated	2,031	1,503	4,552
	SR-associated	499 (20.2%)	508 (25.7%)	1,100 (18.5%)
Short reads	Candidates	1,387	1,600	3,011
	Actual	1,246	1,412	2,731

Table 4.2: Statistics from the processing of *Corry* and *Rui* data sets. The *Alignments* row presents the number of unique alignments from each data set. The top ten most abundant reads are represented in the *Reads* row, while the total expression level of mature sequences of each strand is represented in the *Expression* row. Short read-association (*SR-associated*) is calculated from the set of expressed annotated sequences and unannotated sequences. When considering mature sequences only, the number of unannotated mature sequences and corresponding SR-association is given in the *Matures* rows; when considering all observed isomiRs, the number of unannotated isomiRs and corresponding SR-association is found in the*IsomiRs* rows.

Subset		**Corry**	**Rui**
Alignments	Unique	157,106	4,522
Reads	Top 10	83,082	57,108
Expression	5’	61,274	423,647
	3’	5,827	72,862
Hairpins	Expressed	439 (37.1%)	537 (45.3%)
	SR-associated	196 (44.6%)	18 (3.4%)
MiRNAs	Annotated	499 (23.6%)	530 (25.1%)
Matures	Unannotated	386	447
	SR-associated	238 (26.8%)	18 (1.8%)
IsomiRs	Unannotated	617	1,849
	SR-associated	280 (25.1%)	19 (0.8%)
Short reads	Candidates	984	21
	Actual	729	20

of knocked out cells and normal ones, rather than immunoprecipitate Ago2-associated RNA segments. The statistics obtained after processing this data set are presented in Table 4.3, where the samples have been grouped into *DOC KO, DOC WT, GH KO* and *GH WT* subsets, representing knockout and wild type of the two cell lines. The number of unique alignments, as seen in the second row of the table, are similar for all subsets, and the absolute expression levels of the top ten abundant reads of all subsets were above approximately 12,000 - 15,000. Generally, the share of expressed hairpins and miRNAs are quite similar in all subsets, where a slight increase is observed for WT compared to KO for all numbers except annotated miRNAs in the GH cells. The share of short read-associated sequences among the expressed sequences is also quite similar, and a slight increase in the WT cells compared to the KO cells is observed for both cell lines. The increase in numbers for isomiRs, short read candidates and actual short reads in WT cells is non-negligible, indicating that Ago2 activity influences the number of unique isomiRs and the short read frequency in the cell. However, the share of short read-associated mature miRNAs do not show a similar increase, thus Ago2 cannot be the only cause for short reads in the cell, and might mostly influence the short read association of lower expressed isomiRs. The detailed statistics for each subset is given in tables A.7, A.8, A.9 and A.10 in Appendix A. As with Corry, it is interesting to observe that also the triple samples of same conditions in the *Lundbæk* set show high SDs, indicating a noise level that cannot be overlooked.

Considering the differences in data input is important for a reliable fundament of further comparisons between the data sets. Generally, the human data sets yield higher numbers of expressed hairpins, expressed miRNAs, isomiRs and short reads than mouse datasets, regardless of the number of unique alignments. The share of expressed sequences associated with short reads is also higher in the human sets. The high standard deviations found for the repeated samples of *Corry* and *Lundbæk* might illustrate a non-negligible noise level present in the samples. However, the consistent share of short read associated miRNAs across all samples except *Rui* indicates a consistent, existing correlation between short reads and miRNAs.

Table 4.3: Statistics from processing the *Lundbæk* data set, divided into *DOC KO, DOC WT, GH KO* and *GH WT* subsets. All numbers presented are the average values across all samples of each subset. The *Alignments* row presents the number of unique alignments from each data set. The top ten most abundant reads are represented in the *Reads* row, while the total expression level of mature sequences of each strand is represented in the *Expression* row. Short read-association (*SR-associated*) is calculated from the set of expressed annotated sequences and unannotated sequences. When considering mature sequences only, the number of unannotated mature sequences and corresponding SR-association is given in the *Matures* rows; when considering all observed isomiRs, the number of unannotated isomiRs and corresponding SR-association is found in the*IsomiRs* rows.

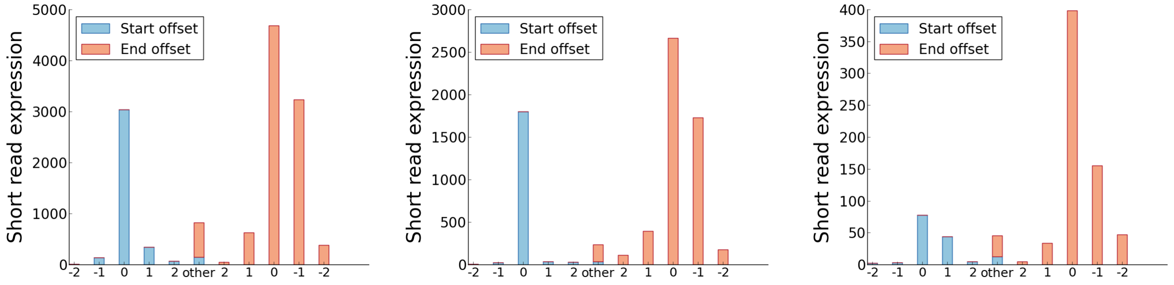
Data set		**DOC KO**	**DOC WT**	**GH KO**	**GH WT**
Alignments	Unique	29,895	32,310	28,115	32,626
Reads	Top 10	12,689	12,715	15,242	15,173
Expression	5’	268,898	176,531	285,002	166,034
	3’	74,340	117,001	125,970	127,211
Hairpins	Expressed	333 (28.1%)	393 (33.2%)	369 (31.1%)	376 (31.7%)
	SR-associated	96 (28.8%)	113 (28.8%)	99 (26.8%)	119 (31.6%)
MiRNAs	Annotated	366 (17.3%)	472 (22.3%)	420 (19.9%)	418 (19.8%)
Matures	Unannotated	238	286	277	286
	SR-associated	109 (18.0%)	128 (16.9%)	112 (16.1%)	133 (18.9%)
IsomiRs	Unannotated	1,696	1,813	1,751	1,840
	SR-associated	169 (8.2%)	235 (10.3%)	215 (9.9%)	272 (12.0%)
Short reads	Candidates	530	726	559	801
	Actual	296	448	387	553

4.2 Verification of findings in Wahl (2014)

The starting point for my work with miRNA short reads was the findings of Mossin (2014), which I essentially verified by successfully reproducing Mossins findings for the *Meister* and *Lundbæk* data sets in my work in Wahl (2014). In particular, I found that short reads align well with either the start or the end of mature miRNAs, discouraging a prior belief that short reads are either products of Ago2 cleavage or degradation. Additionally, I found that the correlation of short read and miRNA expression is not strictly linear, further discouraging the prior belief of short reads being degradation products, and indicating a more complex correlation between short reads and miRNAs. For this project, the first objective was to reproduce these findings on multiple data sets to further verify and ensure statistical reliability of earlier findings. This section presents the successful approach at verifying both short read alignments and coexpression of short reads and miRNAs. Although both alignments and coexpression results for *Meister* and *Lundbæk* were presented in Wahl (2014), I present them again in this report for a holistic view of the results and comparison between the data sets.

4.2.1 Short read alignments

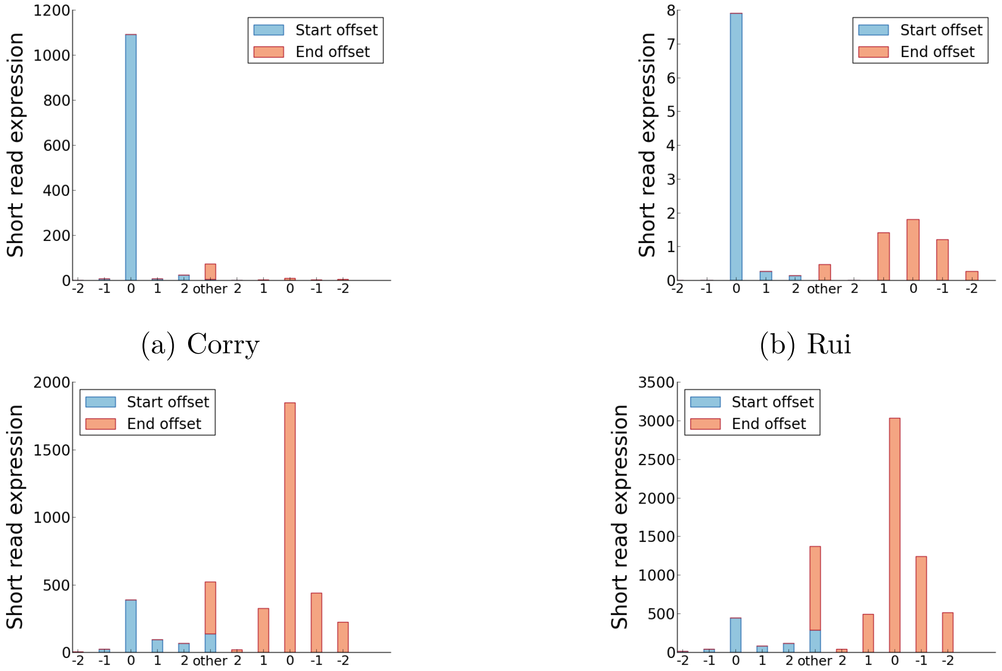
The first step of verifying the findings of Wahl (2014) was to align short reads to mature sequences of all sample data sets and compare the results. The alignments are performed following the scheme presented in Section 3.4.1. Alignments for the human data sets are presented in Figure 4.1, with 'Start offset' denoting the alignment offset for start reads, and 'End offset' denoting the alignment offset for end reads. All three human data sets show a tendency of short reads aligning well to either end of their corresponding mature sequences. Additionally, higher expression levels for end reads than start reads are observed in all data sets.



(a) Meister (b) Daub (c) Rajewsky

Figure 4.1: Total short read alignments for the human sample data sets.

Alignments for the mouse data sets are presented in Figure 4.2. The mouse IP data sets, *Corry* and *Rui*, show a high degree of perfect alignments for start reads, and *Rui* show well alignment also for end reads while *Corry* show bad alignments for end reads. The total expression level of short reads in *Rui* is drastically lower than any of the other data sets, in accordance with the quantities reported in Section 4.1, and comparing these results with the other data sets is imprecise. However, the alignment tendency is still similar, which is a good indication that short reads, independent of their expression levels, align well with mature miRNAs. An interesting observation is that start reads show higher expression levels than end reads, in contrast to the findings from the human data sets. The results from the *Lundbæk* data set show the same tendency for well alignments with mature miRNAs, and as with the human samples, show higher expression levels for end reads than start reads. An important observation is that except higher expression levels for Ago2 WT, there is a lack of significant alignment difference between the Ago2 KO and Ago2 WT samples, indicating that the short read existence and alignment with mature sequences are independent of Ago2 activity. Additionally, Ago2 cleavage would result in equal shares of start and end reads, which is not the case in either Ago2 KO or Ago2 WT samples.



(a) Corry

(b) Rui

(c) Lundbæk KO

(d) Lundbæk WT

Figure 4.2: Total short read alignments for the mouse sample data sets.

The results from both human and mouse data sets verify the reported tendency of short reads aligning well with their corresponding mature miRNAs. All human samples, as well as the mouse *Lundbæk* samples, show higher expression levels of end reads than start reads. The mouse IP samples, *Corry* and *Rui*, show higher expression levels for start reads than end reads, and *Corry* show overall bad alignments for end reads. The implication however remain the same: the alignment tendencies are not compatible with the prior assumptions that short reads are products of either Ago2 cleavage or degradation by known processes.

A closer investigation of possible differences between the 5' and 3' strand of the hairpins reveal no significant differences in alignments, however all samples show much higher expression levels for the 5' strand than the 3' strand. An exception is the *Rui* data set, however the overall short read expression level of this sample is too low to perform a reliable comparison. The alignment distribution for each strand of all data sets is presented in figures B.1, B.3, B.5, B.7, B.9 and B.11 in Appendix B.

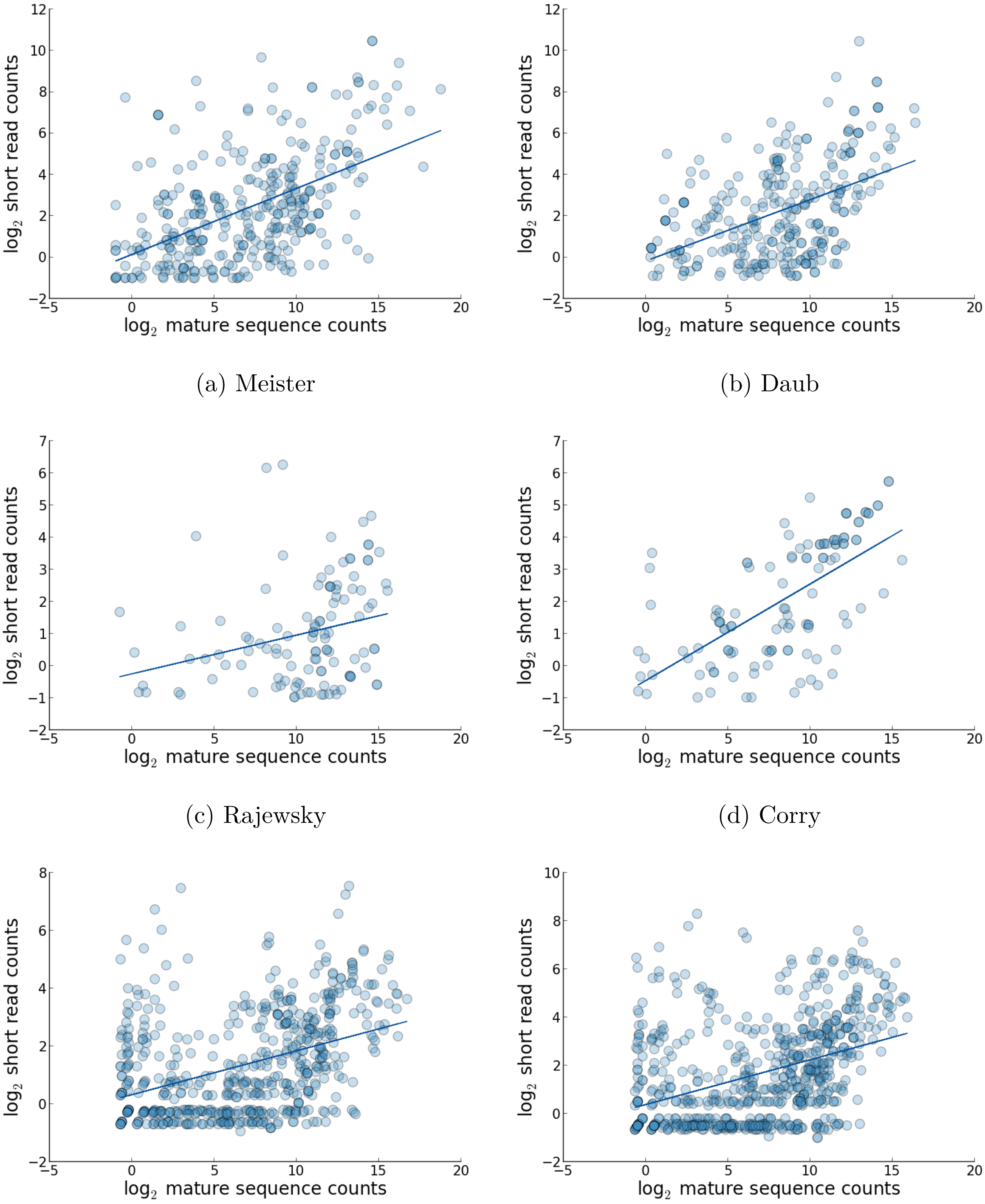
As discussed in Section 4.1, the data sets are of different size and contain different quantities and expression levels of miRNAs and short reads. To better visualize the dynamics of the alignment data, box-and-whiskers-plots were created, as described in Section 3.4.2. These show the same tendency of short reads aligning well to their corresponding mature sequences, and also that this alignment is not only true for a small set of highly expressed short reads, but rather a general tendency of all short reads. Alignment boxplots for all samples are given in figures B.2, B.4, B.6, B.8, B.10 and B.12 in Appendix B.

The boxplots present the results for all individual subsamples, and were analysed to investigate any possible Argonaute dependent alignments trends. For *Meister*, *Daub* and *Corry*, the expression levels were higher for Ago2 than the other Argonautes. *Rui* and *Rajewsky* showed higher levels for Ago1 and Ago3, respectively, however these samples showed generally lower expression levels and are not readily comparable. No general tendency for a preference for start or end reads among the Argonautes was observed.

4.2.2 Correlation of short read and miRNA expression

The second step of verifying my findings in Wahl (2014) was to investigate the correlation of short read and miRNA expression. If short reads are degradation products, the correlation should be linear and hold for all miRNAs. To investigate whether this is true, the expression levels of all short reads aligned to a mature sequence were summarized and plotted against the expression level of the corresponding mature sequence. To obtain a more reliable comparison across data sets, a threshold was set for the expression level of short reads and mature sequences, discarding all reads with an expression level below $t = 0.5rpm$. Both values were log_2 transformed before plotted against each other, and a regression line was calculated for the total set of data points. The results are presented in Figure 4.3, where *Rui* is omitted due to too few short reads above the expression threshold. All datasets show an only partly linear relationship, where some highly expressed miRNAs are associated with lowly expressed short reads and vice versa. This contradicts the assumption of short reads being degradation products, and verifies and supports the findings in Wahl (2014).

If short reads are cleavage products, the correlation is expected to be linear, and to only involve miRNAs from the passenger strand of a miRNA hairpin. When further analysing the results by comparing the results for the highest and lowest expressed strands, respectively regarded the guide and passenger strand of a miRNA hairpin, all data sets show that the vast majority of short read associated miRNAs reside in the guide strand. Most data sets show a poorly fitted regression line for the passenger strand, indicating a slight or no causal relationship between the expression of short reads and miRNAs residing in the passenger strand, further discouraging the assumption of short reads being cleavage products of Ago2. An interesting note is the observation of a more steep, linear correlation for end reads than start reads in all samples except *Corry* and *Rui*, which has too few short reads to be evaluated. Also, there seems to be a trend of higher numbers of end reads than start reads in the same samples. The correlation for the guide strand, passenger strand, start and end reads for the *Daub* data set are presented in Figure 4.4 as an example.



(a) Meister

(b) Daub

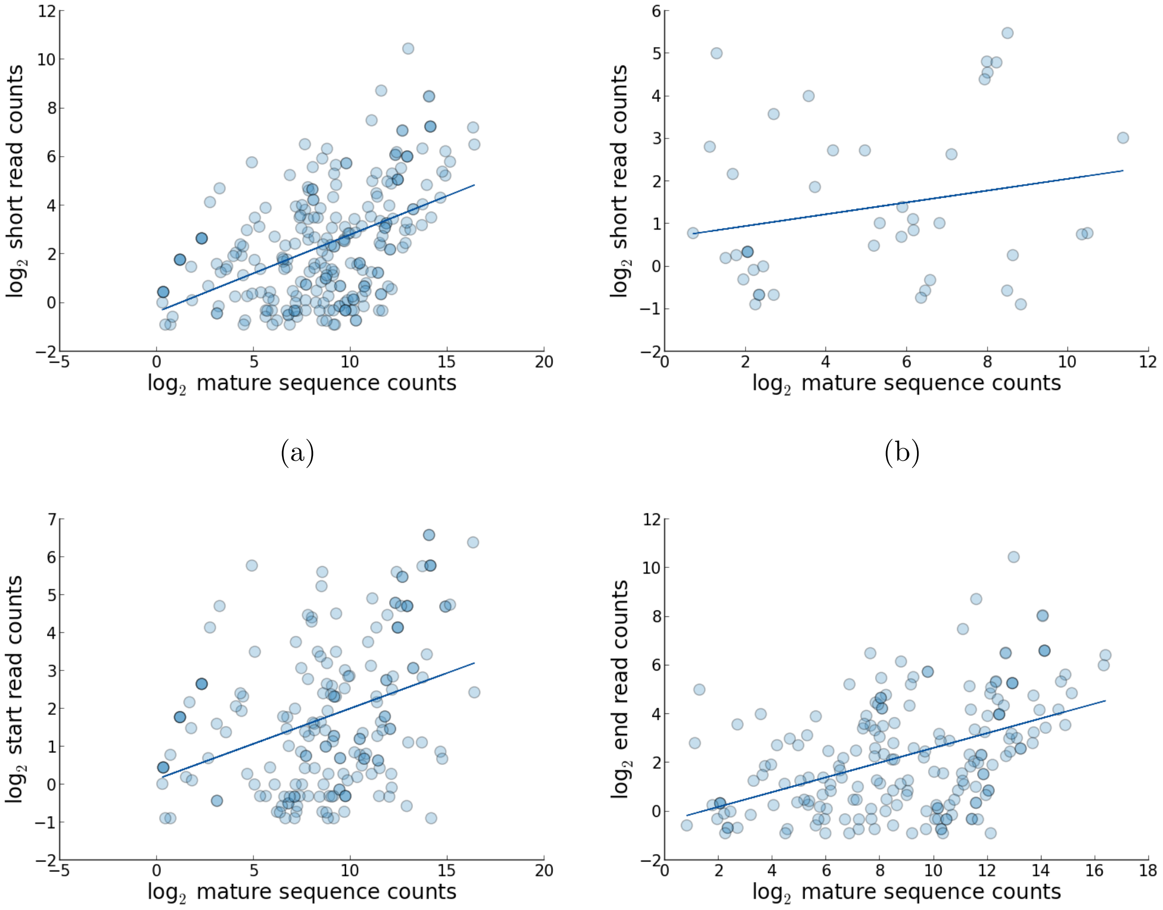
(c) Rajewsky

(d) Corry

(e) Lundbæk KO

(f) Lundbæk WT

Figure 4.3: Correlation between expression levels of short reads and their corresponding mature miRNAs for (a) Meister, (b) Daub, (c) Rajewsky, (d) Corry, (e) Lundbæk KO and (f) Lundbæk WT.



(a) (b)

(c) (d)

Figure 4.4: Correlation between the expression levels of short reads and their corresponding mature miRNAs for the (a) guide strand, (b) passenger strand, (c) start reads, and (d) end reads of the Daub data set.

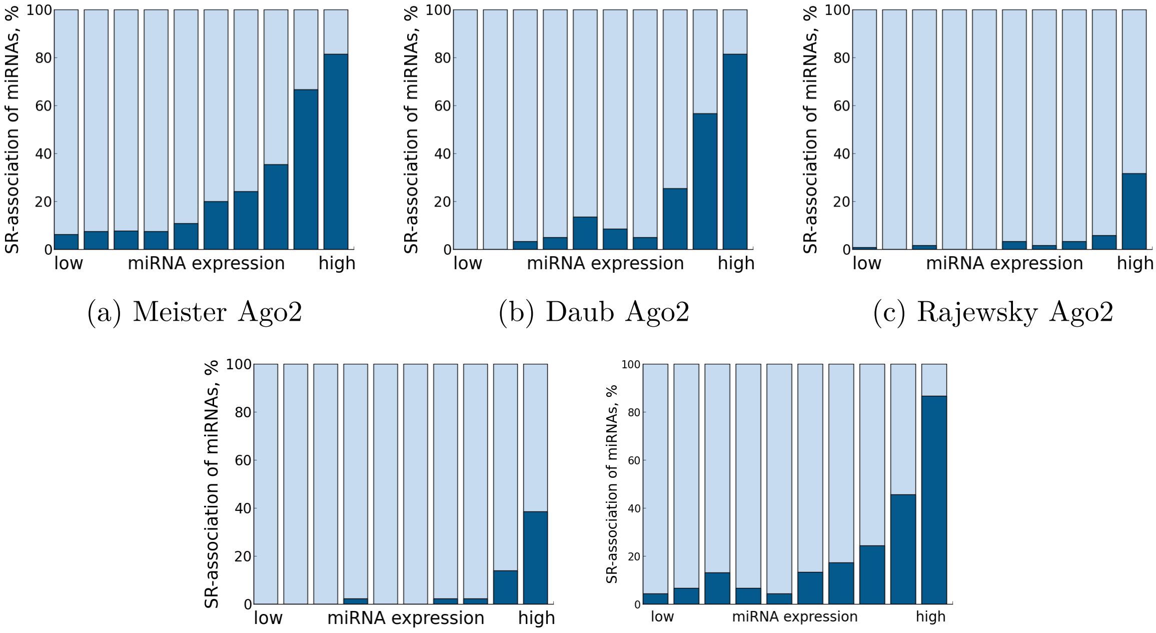
4.3 Short read association of miRNAs

The statistics presented in Section 4.1 revealed that for most data sets, the share of expressed mature sequences associated with short reads ranged from 16% to 26%, thus only these are represented in the correlation graphs presented in the last section. Further analysis is required to evaluate the total set of miRNAs, and investigate whether the short read association of miRNAs is dependent on miRNA expression levels. For each sample, the total set of expressed miRNAs was dynamically divided into ten equally sized bins, ranging from the 10% lowest expressed miRNAs to the 10% highest expressed miRNAs. For each bin, the percentage of its miRNAs associated with short reads was calculated and plotted in a bar chart. To obtain more reliable comparison across the data sets, a threshold of $t = 0.5rpm$ for both short read and miRNA expression was used, discarding all reads with lower expression levels.

As Ago2 is the only Argonaute present in all data sets, the results for Ago2 is presented in Figure 4.5. The *Rui* data set is omitted due to too few short reads with expression level above the threshold. For *Corry*, the Ago2c sample is representative for the three Ago2 samples and is presented, the same holds for the GH WT1 sample of *Lundbæk*.

If short reads are cleavage products, and thus remnants of discarded passenger strands, they should associate with the lower expressed miRNAs, which is clearly not supported in any of the samples. Rather, the results show a trend in all data sets where the percentage of miRNAs associated with short reads increases with the expression level of miRNAs, and the majority of short read associated miRNAs are found within the top 20% expressed miRNAs. The most expressed miRNAs in a cell are more likely to perform biological functions, thus the results indicate that short reads may be related to miRNA activity in the cell.

By closer investigation, an interesting observation is that the vast majority of short read associated miRNAs reside in the highest expressed strand, the guide strand, which is not compatible with short reads being passenger strand cleavage products. Additionally, the share of the 10% highest expressed miRNAs seem to be greater for the 5' strand than the 3' strand, and accordingly, the share of short read associated miRNAs. This trend is observed for all five data sets, however it is less clear in *Meister* and *Rajewsky*. The results for the *Daub* data set are presented in Figure 4.6 as an example.



(a) Meister Ago2

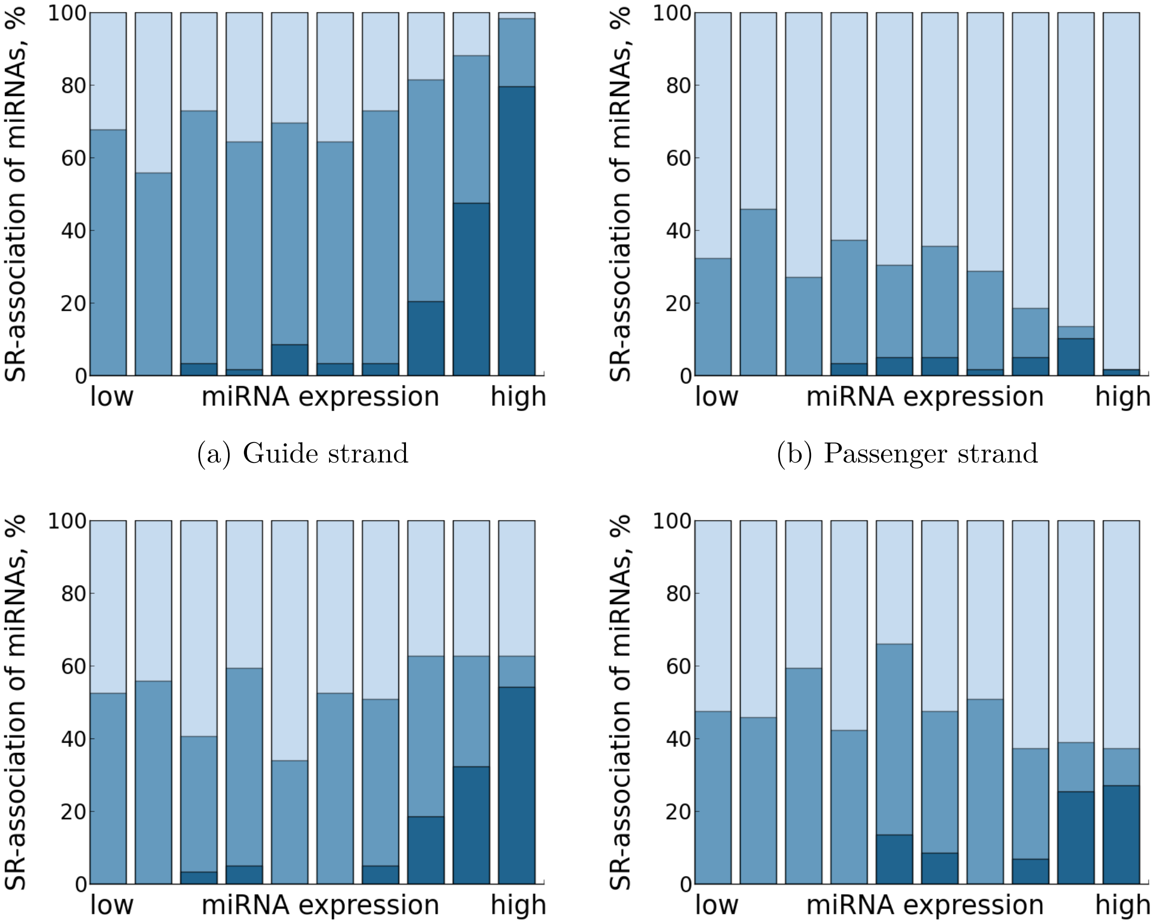
(b) Daub Ago2

(c) Rajewsky Ago2

(d) Corry Ago2c

(e) Lundbæk GH WT1

Figure 4.5: Short read association of miRNAs in relation to miRNA expression levels. All expressed miRNAs are divided into ten equally sized bins represented by ten bars in the figure. The share of miRNAs of each bin associated with short reads are represented by the darker colour of the within the bars.



(a) Guide strand

(b) Passenger strand

(c) 5’ strand

(d) 3’ strand

Figure 4.6: Short read association of miRNAs in relation to miRNA expression levels. The bins are created as in Figure 4.5, with the lighter blue shade within each bar representing the share of miRNAs and the darker blue shade representing the share of short read associated miRNAs in the bin belonging to the guide strand (a), passenger strand (b), 5’ strand (c) and 3’ strand (d) for the Daub data set.

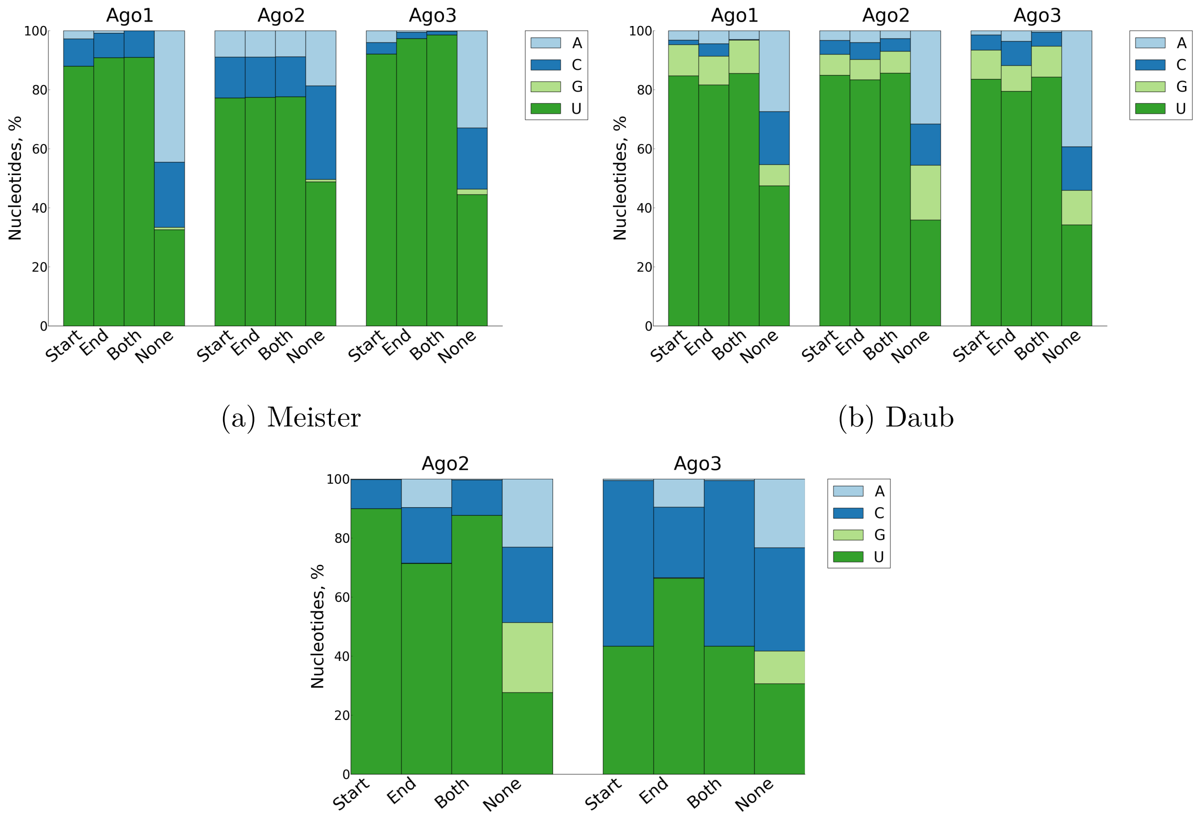
4.4 Terminal nucleotide preferences

As discussed in Section 2.2.3, terminal nucleotides of miRNAs have been studied to investigate possible preferences of the different Argonautes. In mice, a preference for 3' adenine and uracil has been reported for Ago1 and Ago2, respectively, while a preference for 5' adenine and uracil has been predicted for humans due to structural requirements of the Argonaute proteins. To investigate whether different terminal nucleotides might be associated with short read association of miRNAs, the 5' and 3' terminal nucleotide of miRNAs was analysed.

For the human data sets, Figure 4.7 presents the 5' terminal nucleotides of miRNAs of four distinct groups: miRNAs with only start reads, only end reads, both start and end reads and miRNAs without short reads aligned to them. For the analysis an expression level threshold of $rpm = 0.5$ was used for both miRNAs and short reads. As expected, the vast majority of miRNAs with short reads, either start, end, or both, have a 5' terminal uracil. An exception is Ago3 in *Rajewsky*, where uracil accounts for approximately 40%. An interesting observation is that for all data sets, the miRNAs not associated with short reads have a different nucleotide distribution than those associated with short reads. If uracil is important for miRNAs to bind with Argonautes, it may indicate that miRNAs not capable of binding with Argonautes are also not associated with short reads, and as such, short reads are associated with miRNAs that are active in the cell. A closer investigation reveals that the actual expression level of miRNAs without short reads compared to those associated with short reads are low, corresponding to the findings in Section 4.3 of highly expressed miRNAs being associated with short reads, introducing some uncertainty to the reliability of comparisons between the two sets. Still, the pattern is an interesting observation.

For the *Lundbæk* data set, all samples follow the same pattern as the human data sets. The same is observed for *Corry*, but an important note is that this pattern can only be verified for miRNAs with start reads as there are hardly any miRNAs with end reads of an expression above the threshold. The results for *Corry* and *Lundbæk* are found in Appendix C.

Figure 4.8 presents the 3' terminal nucleotides of miRNAs from the human data sets, where the same threshold of $rpm = 0.5$ is used and the groups defined in the same manner. For the 3' end, consistent nucleotide preference across the data sets cannot be found. The nucleotide distribution for Ago2 are quite similar for *Meister* and *Daub*, where the majority of miRNAs associated with short reads have either a 3' terminal cytosine or uracil, while miRNAs not associated with short reads seem to have a higher share of 3' adenine and a lower share of cytosine. Ago1 and Ago3 show no similarities across the samples, and the *Rajewsky* set do not share any patterns with any of the other human data sets. The only pattern found for all samples is that miRNAs without short reads have a different nucleotide distribution than those with short reads aligned to them. The degree of difference is not the same in all samples, but still clear. However, as with the 5' terminal nucleotides, the group of miRNAs not associated with short reads have a drastically lower expression level, and comparing the groups introduces some uncertainty to the reliability of the results. Still, the pattern is interesting and might indicate a functional difference between miRNAs with and without short reads.

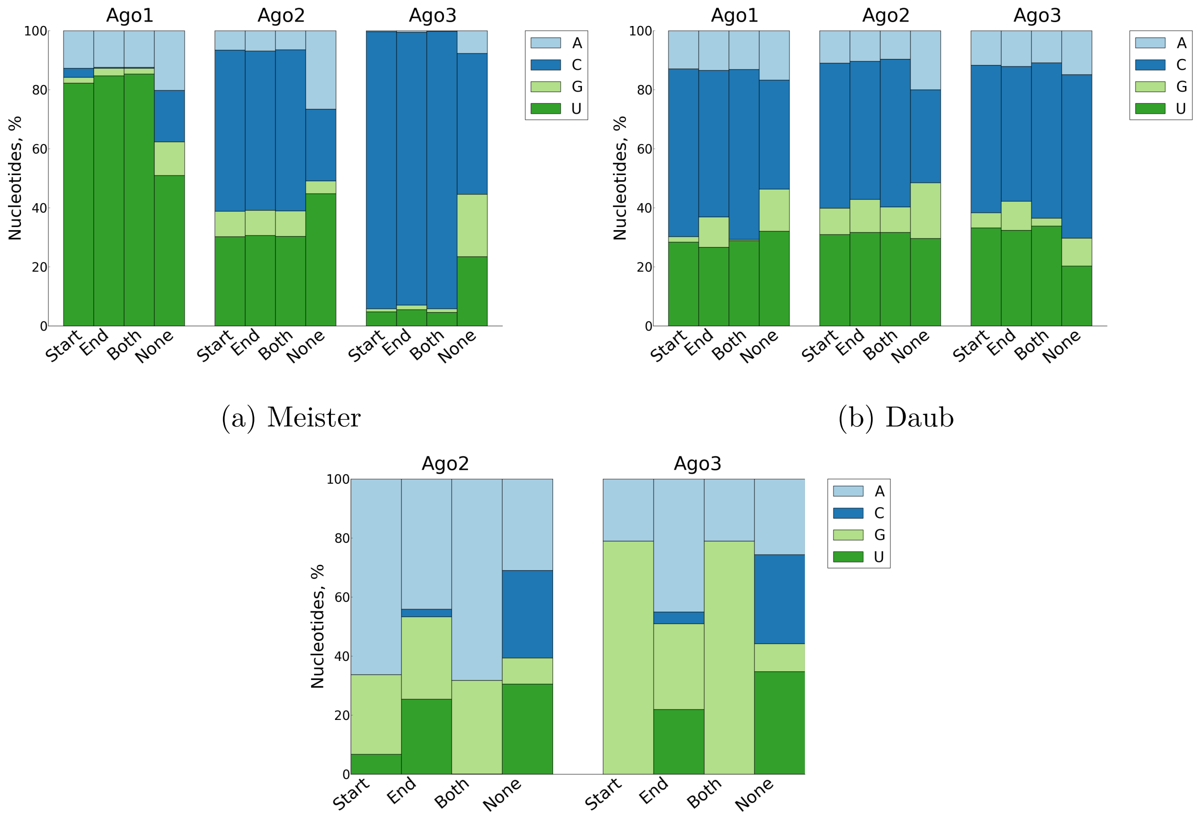


(a) Meister

(b) Daub

(c) Rajewsky

Figure 4.7: 5’ terminal nucleotide preferences of human miRNAs associated with start reads, end reads, both start and end reads or neither.



(a) Meister

(b) Daub

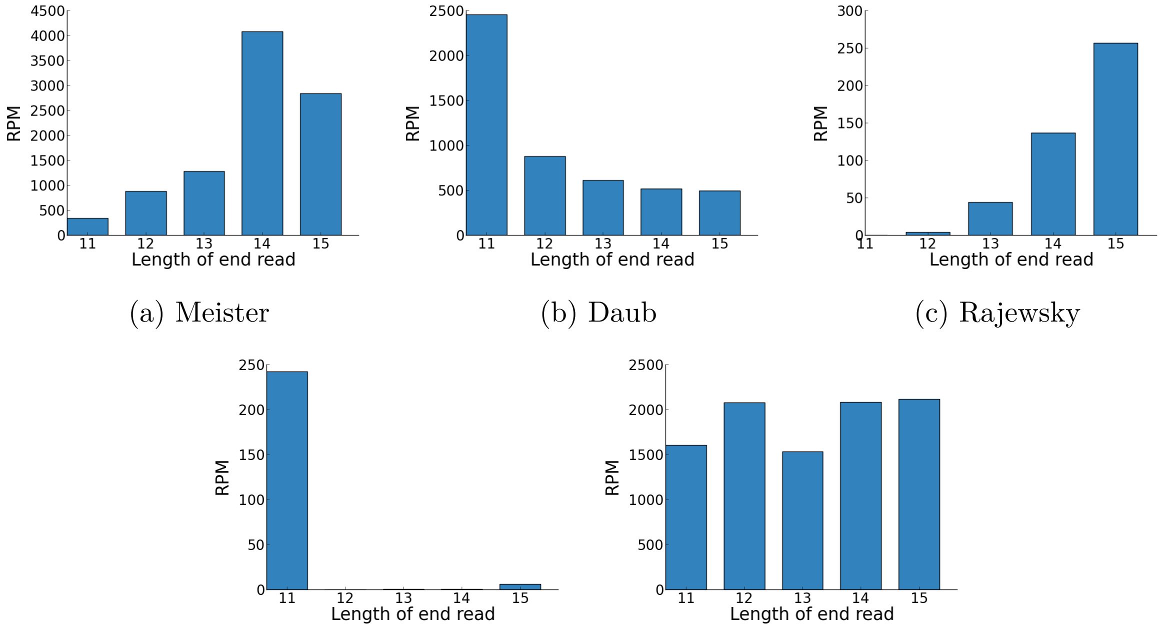
(c) Rajewsky

Figure 4.8: 3' terminal nucleotide preferences of human miRNAs associated with start reads, end reads, both start and end reads or neither.

For the 3' terminal nucleotides of the mouse data sets, *Lundbæk KO* show the same trends as for 5' terminal nucleotides, while *Lundbæk WT* show incosistent results among its internal samples. *Rui* and *Corry* contain too few miRNAs associated with start or end reads above the threshold to be included in this analysis. The results for *Corry* and *Lundbæk* are found in Appendix C.

4.5 Analysing short read lengths

When analysing short reads, it would be interesting to investigate whether short reads of a particular length are more frequent than others. To enable such an analysis, the total expression level of all short reads of each length within the range 11-15 nucleotides were summarized and plotted for each length. As in earlier sections, *Rui* is omitted due to low expression levels of short reads. *Corry* reveals a preference for start reads of length 12 and end reads of length 11, *Daub* for start and end reads of length 11, *Rajewsky* for start and end of length 15, while *Meister* and *Lundbæk* show less clear preferences. The distributions are not similar across any multiple data sets, and the results do not provide any insight regarding the length of short reads. The results for end reads are presented in Figure 4.9, while the results for start reads are found in Appendix D, Figure D.1.

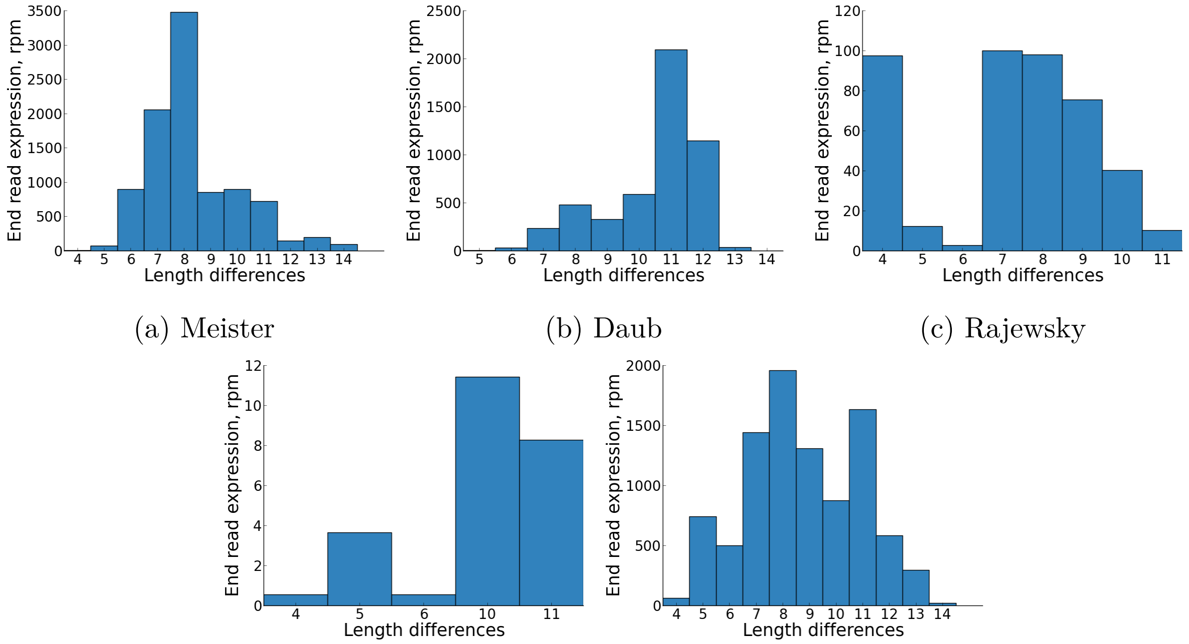


(a) Meister (b) Daub (c) Rajewsky

(d) Corry (e) Lundbæk

Figure 4.9: The total expression level of end reads of each length within the range 11-15 nucleotides.

Even though the length alone do not seem to be an important feature of short reads, the length of end reads could have a more sophisticated importance. Short reads are of length 11-15 nts, and their corresponding miRNAs of approximately 22 nts. The seed region of miRNAs is the first ~8 nucleotides of the sequence, which implicates a possibility that end reads of e.g. length 14 is the remaining nucleotides after the seed region of a miRNA of length 22. To investigate whether this might be true, the difference between the lengths of end reads and their corresponding mature miRNAs was calculated, and the total expression level for all end reads of each length difference summarized. The results are presented in Figure 4.10. *Meister* seems to have a clear preference for a difference of 8, *Daub* for 11 and *Lundbæk* for 8 and 11. *Rajewsky* and *Corry* do not show any clear preference. A closer investigation of the individual samples of each set reveals that all subsamples of *Meister* and *Daub* show a clear preference for 8 and 11 nts, respectively, while the individual samples of the other data sets show much variation. Taken together,



(a) Meister

(b) Daub

(c) Rajewsky

(d) Corry

(e) Lundbæk

Figure 4.10: Differences between the length of miRNAs and their associated end reads. The total expression level of end reads for each length difference is plotted in rpm.

these results are not significant and do not provide any direct insight, however the clear preference of *Daub* and *Meister* are interesting.

Investigating the lengths of short reads and end reads in particular could not provide any consistent insight into the function or traits of short reads, other than dismiss the possibility that their lengths are of central importance.

Table 4.4: ANOVA p-value results

Feature	**Meister**	**Daub**	**Rajewsky**	**Corry**	**Lundbæk**
Ago	**0.000795**	0.080591	0.814369	0.158942	**0.005045**
Strand	0.552156	**0.000177**	**0.000281**	**3.81e-10**	**0.000643**
Position	**0.010027**	0.284557	0.303438	0.416926	**3.83e-10**
Offset	0.166807	0.309072	0.179639	**4.12e-10**	0.410241
Ago + strand	0.167533	0.341845	0.955350	0.818357	**0.000742**
Ago + position	**0.020789**	0.736115	0.086723	0.314100	0.848687
Strand + position	0.185666	0.453279	0.554527	0.350389	0.921602
Ago + offset	0.226724	0.899623	0.623590	0.131213	0.995590
Strand + offset	**0.034654**	0.730168	0.787048	**0.000779**	0.242640
Position + offset	0.895815	0.385174	0.713991	**0.001790**	0.205713
Ago + strand + position	0.178897	0.743960	0.236442	0.613452	0.641173
Ago + strand + offset	0.120423	0.993507	0.961543	0.828930	0.998795
Ago + position + offset	0.527726	0.930109	0.327872	0.655669	0.999994
Strand + position + offset	0.064324	0.347805	0.156358	0.097808	0.175154
Ago + strand + position + offset	0.295503	0.849407	0.864046	0.949222	0.999757

4.6 Investigating significantly differentiating features

In the prior sections, tendencies and patterns of short read expression and alignment to miRNAs have been studied. An approach taken next is an attempt at identifying statistically significantly differentiating features of short read expression, by calculating the *Analysis of Variance* (ANOVA) test as described in Section 3.6.2. The conditional value to be evaluated is the expression level of each short read, and the features evaluated for influencing the expression level of a short read are its alignment offset, start/end position, residential strand, and which Argonaute protein it is associated with. The p-values from the resulting ANOVA tables are presented in Table 4.4 for all data sets except *Rui*, due to its low expression levels of short reads. The significance level is set to $p = 0.05$, and values below this threshold are presented in bold.

As seen in the results, only the Argonaute protein, residential strand, position and combination of strand and offset are found statistically significant across multiple data sets. The Ago protein is found significant in the *Meister* and *Lundbæk* data sets, however the *Lundbæk* data set represent different cell lines and a knock out of Ago2, rather than different Argonautes, and this feature is thus not comparable with the *Meister* data set. The position is found significant in the *Meister* and *Lundbæk* data set, while the combination of residential strand and offset is found significant in the *Meister* and *Corry* data set. However, both features generates p-values way above the significance level in the other three sets, which renders the features not significant on a more global level.

The residential strand of short reads is found significantly different in all data sets except Meister. This finding is supported by prior results, which revealed higher expression levels of short reads aligning to the 5' strand than to the 3' strand. The total expression level of

Table 4.5: Total expression level miRNAs from each hairpin strand.

Strand	Corry	Daub	Lundbæk	Meister	Rajewsky	Rui
5’	367,529	708,216	2,689,325	1,508,977	601,526	1,270,883
3’	34,853	206,801	1,333,518	290,165	585,016	218,516

each strand is presented in table 4.5, which reveals that all five data sets contain higher expression levels of miRNAs from the 5’ strand than the 3’ strand, some more than a five-fold difference. The Wilcoxon signed-rank test provides a p-value of 0.02770, implying a significant difference between the expression of the 5’ and 3’ strand. This indicates that the strand itself is not a significant feature, but rather that the expression of short reads correlates with the expression of miRNAs, regardless of the strand.

The ANOVA results could not provide any consistent significantly differentiating features for all data sets, however significantly higher expression levels of miRNAs of one strand was found to yield significantly higher expression levels of short reads for that strand. The complete ANOVA tables for each data set are found in Appendix E, Figures E.1 through E.6.

4.7 Hairpin classification

Prior sections have investigated features of short reads and the correlation of short reads and miRNAs. In this section, an approach at identifying features of short read associated miRNAs is presented, attempted by investigating the short read association of miRNA hairpins in correlation with the hairpin strand preference. For the following subsections, first a classification scheme for miRNA hairpins based on their strand preference is established, second the short read association of the highest and lowest expressed strands of each class are investigated. Third, the correlation of hairpin and short read expressions are analysed. Fourth, the different short read alignment scenarios for hairpins are presented, along with an investigation of the correlation between scenarios and hairpin expression. Finally, the hairpin fold change of the different scenarios are compared.

4.7.1 Classification scheme

For all hairpin classification analyses to be presented, expressed miRNAs are divided into two classes based on their fold change (FC). A miRNA hairpin is annotated with two strands, and if both are expressed, the degree of preference for one or the other must be established. This ratio between the strands of a hairpin is denoted fold change (FC) throughout this report. The classes used are 'Different' and 'Equal', representing the scenarios where there is a clear differential preference for one strand over the other, and where there is no such preference. The classification scheme is given by

$$miRNA \in \begin{cases} E, & FC < 10 \\ D, & \text{otherwise} \end{cases} \tag{4.1}$$

where D represents the 'Different' class and E the 'Equal' class. The classification threshold is set to $FC = 10$, where FC is given by

$$FC = \frac{max(rpm_{5p}, rpm_{3p})}{min(rpm_{5p}, rpm_{3p})} \tag{4.2}$$

where rpm_{5p} and rpm_{3p} represent the total expression level for the 5' and 3' strands of the miRNA, respectively. Following, a miRNA is of class 'Different' if the total expression level of one strand is at minimum ten-fold the total expression level of the other; otherwise it is of the 'Equal' class. All miRNA hairpins with only one expressed strand are discarded.

The short read association of miRNAs of the two classes focuses on which strand the short reads align to. Short reads either align to the highest or lowest expressed strand, and if the two strands are equally expressed, the 5' strand is denoted as the highest expressed strand. Following, hairpins can be associated with short reads on the guide strand, passenger strand, both strands, or neither strand. An expression level threshold of $rpm = 0.5$ is set for all miRNAs and short reads, discarding reads not qualified.

4.7.2 Short read-association of strands

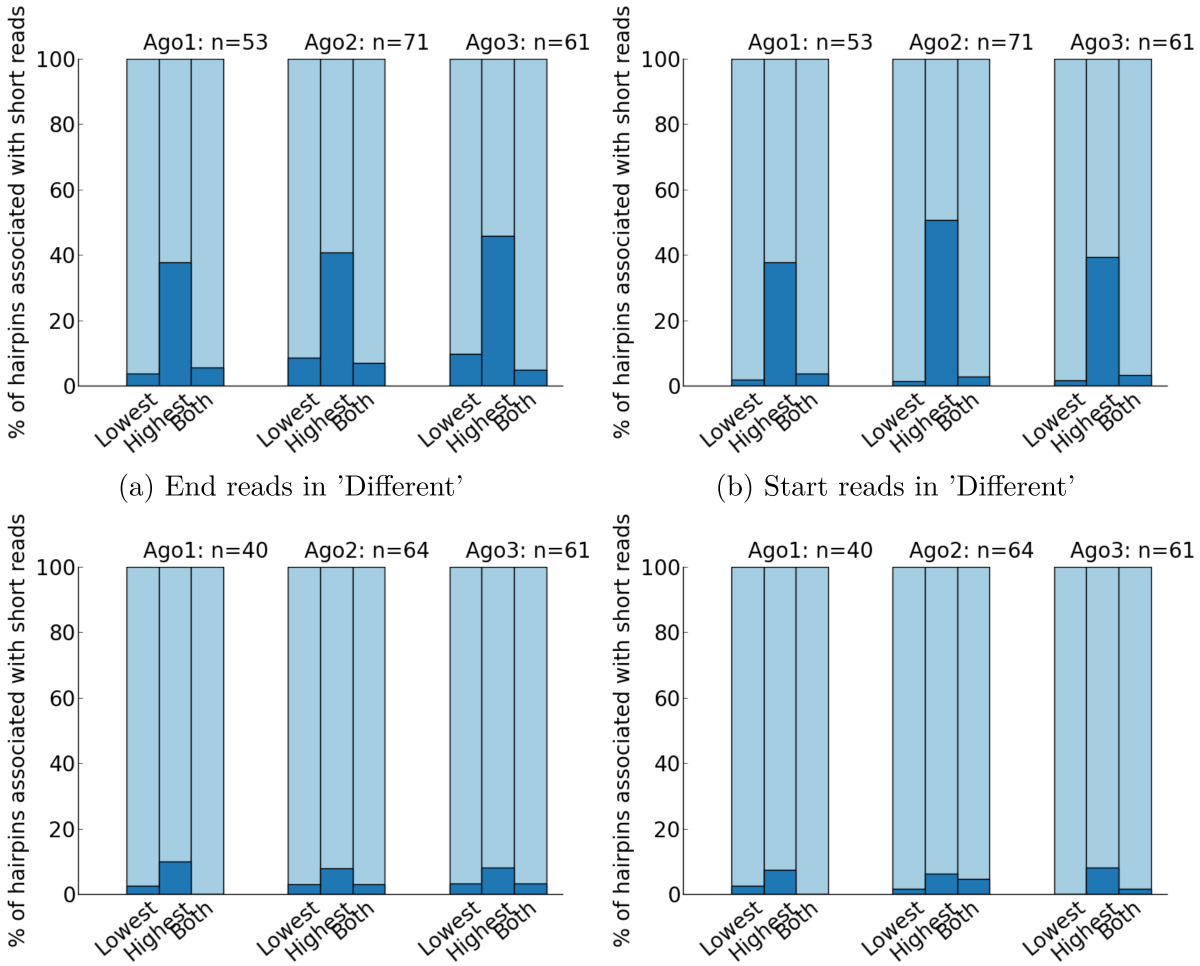
After classifying all expressed miRNAs as described in the last section, the short read association of miRNAs is investigated. A miRNA might be associated with short reads in both strands, the highest or the lowest expressed strand. When plotting the percentage of miRNAs associated with short reads on either or both strands, the same tendency is found in all data sets. The results from the *Daub* data set are presented in Figure 4.11 as an example, where the percentage of miRNAs associated with short reads is plotted for all subsamples, and the set size n is printed above the bars for reference. In all data sets, the vast majority of short read associated miRNAs are of the 'Different' class, of which the majority is associated with short reads on the highest expressed strand. MiRNAs with a clear strand preference are more likely to perform active functions in a cell, as they are more susceptible to be incorporated into Argonautes, thus the results indicate that short reads are related to active miRNA hairpins. As in Section 4.3, the majority of short read associated miRNAs reside in the highest expressed strand, the guide strand that are incorporated into Argonaute proteins, further supporting a possible perception of short reads being related to active miRNAs in the cell.

Meister, *Rajewsky* and *Lundbæk* show tendencies of higher association of end reads than start reads, while Daub show no difference and Corry a preference for start reads. Thus, no consistent preference for start or end reads are found. The results for the other five data sets are found in Appendix F.1, Figures F.1 through F.5.

4.7.3 Correlation of short read and hairpin expression

The last section identified a trend where miRNAs with a clear strand preference are more prone to be associated with short reads than miRNAs without a clear strand preference. To investigate whether there is a difference in also the correlation between the expression of short reads and hairpins between the 'Different' and 'Equal' class, this correlation of short reads aligning to the lowest, highest and both strands of a hairpin were plotted for start and end reads, 'Different' and 'Equal', separately. The results for all data sets are found in Appendix F.2, Figures F.6 through F.11.

The vast majority of short read associated hairpins belong to the 'Different' class, and as such, the majority of expressed short reads are expected to also be in the 'Different' class. This is true, however a non-negligible share of the expressed reads reside in the 'Equal' class in all data sets. Four out of five data sets still show higher expression levels of hairpins in 'Different' than 'Equal' class, the exception being Rajewsky, and the same four except end reads in Corry show also higher expression levels of short reads in 'Different' than 'Equal'. The few numbers of short reads aligning to the passenger strand renders these not suitable for comparison across the sets. The majority of short reads have been found to align to the guide strand, and as such, these serve the best basis for comparison. The results for start and end reads were not found significantly different, and as presented in the last section, the share of short read associated miRNAs is also similar for start and end reads. Following, this analysis do not separate short reads into start and end reads. The combined results are presented in Table 4.6, where the correlation coefficients, regression line slopes and p-values for short reads aligning to the guide strands of the 'Different' and 'Equal' class are summarized and presented separately. The p-value



(a) End reads in 'Different'

(b) Start reads in 'Different'

(c) End reads in 'Equal'

(d) Start reads in 'Equal'

Figure 4.11: Short read association of miRNAs of classes 'Different' and 'Equal' from the *Daub* data set. A miRNA can be associated with short reads on its highest and/or lowest expressed strand, resulting in three groups: highest, lowest and both. The percentage of miRNAs associated with either is plotted in this figure, with the set size of each subsample denoted above.

Table 4.6: Correlation coefficients (r), regression line slopes (s), *p-values* and n for the correlation between the expression of short reads aligned to the guide strand and hairpins of the 'Different' and 'Equal' class for all six sample data sets. The p-value represents the null hypothesis that the slope is zero.

	Different				Equal			
Data set	r	s	p	n	r	s	p	n
Corry	0.230	0.181	0.229	29	0.879	0.805	0.121	4
Daub	0.374	0.363	0.000	97	0.201	0.155	0.410	19
Lundbæk	0.511	0.384	0.000	460	0.199	0.082	0.0297	119
Meister	0.558	0.491	0.000	81	0.180	0.153	0.474	18
Rajewsky	0.317	0.220	0.015	58	0.623	0.314	0.004	19

represents the probability value of the null hypothesis test that the slope is zero. The same table for start and end reads separately is found in Appendix F.2, Table F.1.

For the 'Different' class, the correlation calculations are performed with p-values below the significance level of $p = 0.05$ in all data sets except *Corry*, indicating that the slope representing the correlations in the other data sets are not zero, and a correlation exists. When comparing the correlation coefficients across the two classes, *Corry* and *Rajewsky* show higher values in 'Equal', while the other data sets show a clear preference for 'Different'. When comparing regression line slopes between 'Different' and 'Equal', the same is observed: *Corry* and *Rajewsky* show steeper slopes in 'Equal', while the other data sets show clearly steeper slopes in 'Different'. As expected from the last section, the group size, n, is much higher in 'Different' than 'Equal' for all data sets. An important observation is the high p-values in the 'Equal' class compared to 'Different' for all data sets except *Corry*, indicating that a significant correlation of short read and hairpin expression for the guide strand is found in 'Different', but not in 'Equal'.

The correlation coefficient is expected to decrease as n increases, as fitting a regression line to a set of data points is usually harder as the number of points increase. An interesting observation is thus that except for *Corry* and *Rajewsky*, the correlation coefficients generally seem to slightly increase when n increases, and there is not a consistent proportionality between n and r. r seems to be independent of n, possibly implying that the correlation of short read and hairpin expression is a true biological connection and not statistical artefacts.

This analysis finds a greater expression level of short reads and hairpins in the 'Different' class than 'Equal', and visualizes a clear correlation between the two. This correlation is found significant in the 'Different' class except for *Corry*, and for some of the data sets also in the 'Equal' class. When comparing the two classes, the regression line slope is found steeper in 'Different' than 'Equal' for three of the data sets. The correlation coefficient is found independent of the group size n, indicating that the observations are more likely to be true observations and not artefacts.

Table 4.7: Short read alignment scenarios for a hairpin. For following analyses, each scenario is represented by the colour given in this table.

Scenario	Definition	Colour
A	Short reads only on the passenger strand	Green
B	Short reads only on the guide strand	Blue
C	Short reads on both strands	Orange
D	Short reads on neither strand	Red

4.7.4 Hairpin expression of alignment scenarios

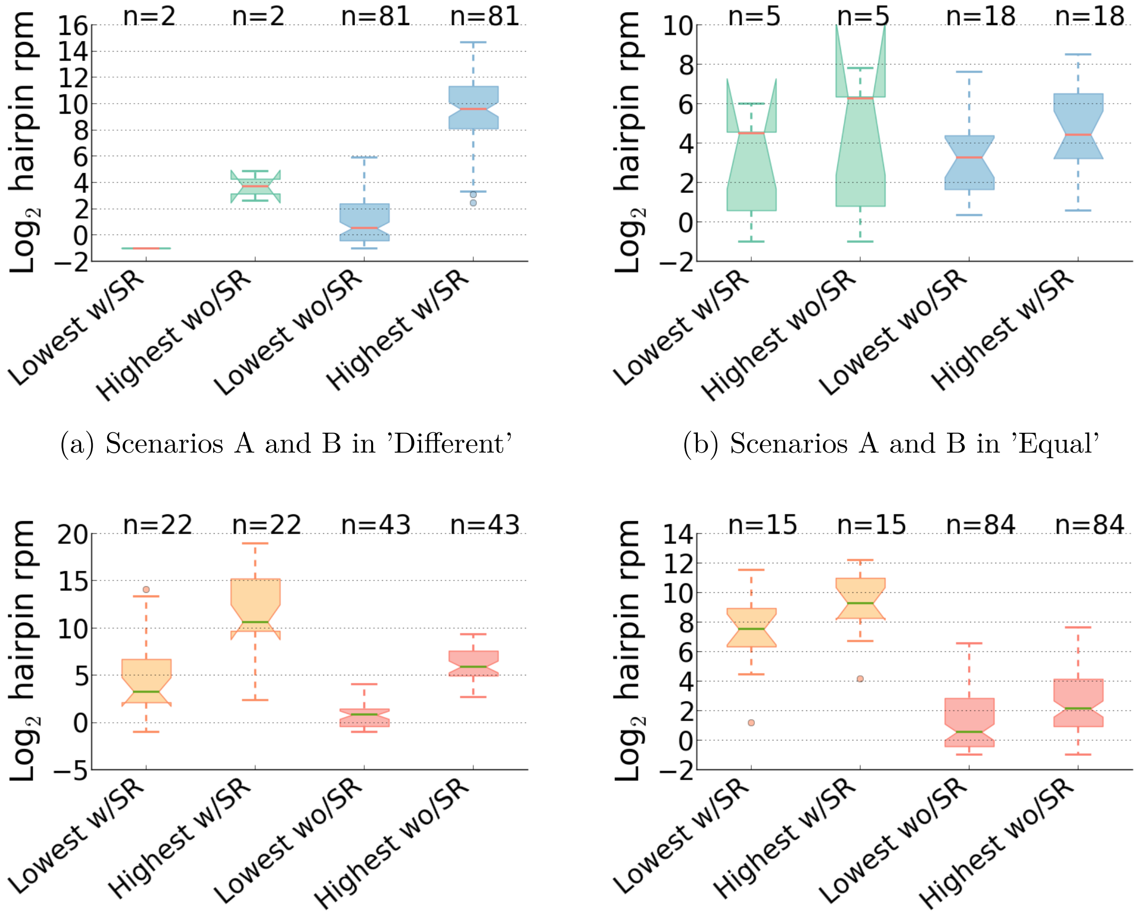
The search for features of short read associated miRNAs has so far revealed that the majority of short read associated miRNAs belong to the 'Different' class, and especially reside in the guide strand, and that the expression of short reads and hairpins correlates in a partly linear manner, especially for the 'Different' class and with a steeper correlation in 'Different' than 'Equal'. The last section only covers the expression of hairpins with short reads associated with it, when in reality there is a set of hairpin expressions not yet investigated. Next, the different short read alignment scenarios for a hairpin are investigated. A miRNA hairpin has two strands, and when both are expressed, there are in total four possible short read alignment scenarios, given in Table 4.7.

For each scenario, hairpin expressions are plotted in a notched boxplot (section 3.4.2) separately for 'Different' and 'Equal'. The results for all data sets are given in Appendix F.3, Figures F.12 through F.17, while the results for *Meister* are presented in Figure 4.12 as an example.

For scenario B and D, a first observation is the obvious difference in group size between the two classes, where short reads yield 81 and 43 hairpins for B and D in 'Different', compared to 18 and 84 hairpins in 'Equal', respectively. Simultaneously, the share of hairpins in A compared to B is much lower in 'Different' than 'Equal', while the share of hairpins in C compared to D is much lower in 'Equal' than 'Different', even though the expression level of scenario C is generally higher than D. For 'Different', scenario B represents the majority of hairpins, while for 'Equal', scenario D represents the majority. A very interesting observation is the drastically smaller number of hairpins of scenario A than B, implicating that hairpins with short reads only on the passenger strand are rare if there is a clear strand preference. Closer investigations of the short reads aligned to these hairpins reveal low short read expression levels, and might illustrate the level of noise in the samples. These observations are consistent across all data sets.

As explained in section 3.4.2, the median of two boxes differs with 95% confidence if their notches do not overlap. There are in total 16 boxes representing hairpin expressions, and two different comparisons among these are interesting to investigate: the same scenario between the classes, and the same strand between the scenarios of the same class. These comparisons have been performed for all data sets, however both *Rui*, *Corry* and *Rajewsky* contain scenario boxes with $n = 0$, and comparing the other sets with these are not always reliable.

Comparisons of the same scenario between the classes did not provide significant findings



(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal'

Figure 4.12: The expression levels of each hairpin strand of scenario A (green), B (blue), C (orange) and D (red) for the *Meister* data set in the 'Different' and 'Equal' class.

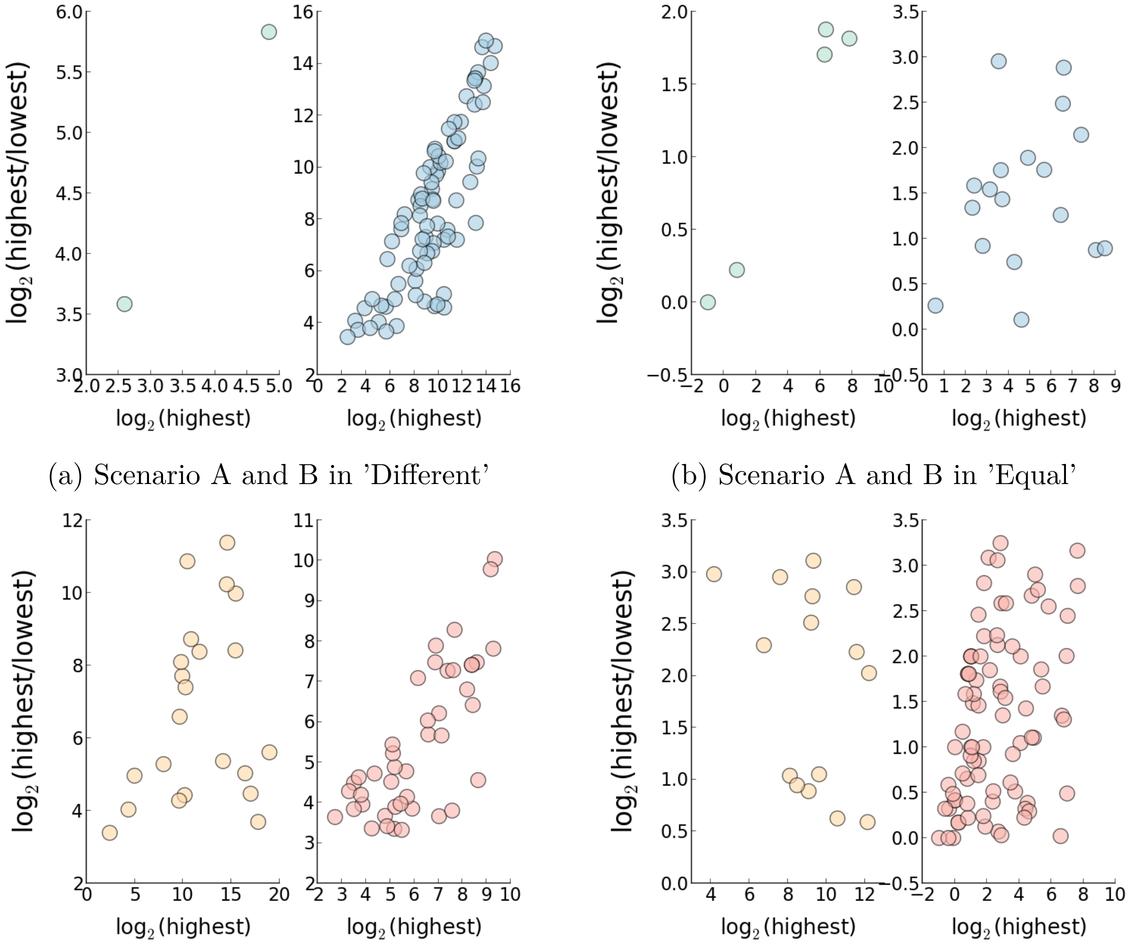
across all data sets, however when omitting *Rajewsky* and *Corry*, neither the passenger nor guide strand of scenario B overlapped across the classes. The notches for the guide strand of scenario C only overlapped between the classes for the *Meister* data set, while the guide strand of scenario D did not overlap in any data sets. Comparing the same strand between the scenarios of the same class revealed that no strands of scenario C and D overlapped in either 'Different' or 'Equal'. The guide strand of scenario A and B did not overlap in the 'Difference' class of any data set except *Corry*.

In total, this implies that there is a significant difference in expression level of hairpins from both strands of scenario B, the guide strand of C, and the guide strand of D, between the 'Different' and 'Equal' class, indicating that the strand preference of a hairpin significantly influences its expression level. The short read association of a guide strand was found to yield significantly different expression levels across scenarios A and B in 'Different', and the same was found for both strands of scenarios C and D in both classes.

4.7.5 Fold change of hairpin groups

The previous section summarized the expression level of each strand of hairpins in the different scenarios. Next, the strand preference of hairpins of the different scenarios is analysed by calculating their fold changes (FC), the ratio between the expression level of its guide and passenger strand. The fold change is known to be a factor for predicting active miRNAs in a cell, where higher expression levels predict more active miRNAs. Comparing the fold changes for the different scenarios, and of the different classes, might yield some insight into whether short read association also might be a factor for predicting active miRNAs.

For the four scenarios presented in the last section, the fold change of every hairpin is calculated, log_2 transformed and plotted against the log_2 transformed expression level of its guide strand. Note that for this analysis, start and end reads have been combined. The results for all data sets are found in Appendix F.4, Figures F.18 through F.23, while the results for *Daub* are given in Figure 4.13 as an example.



(a) Scenario A and B in ’Different’

(b) Scenario A and B in ’Equal’

(c) Scenario C and D in ’Different’

(d) Scenario C and D in ’Equal’

Figure 4.13: Fold change of scenarios A (green), B (blue), C (orange) and D (red) of the ’Different’ and ’Equal’ class of Meister

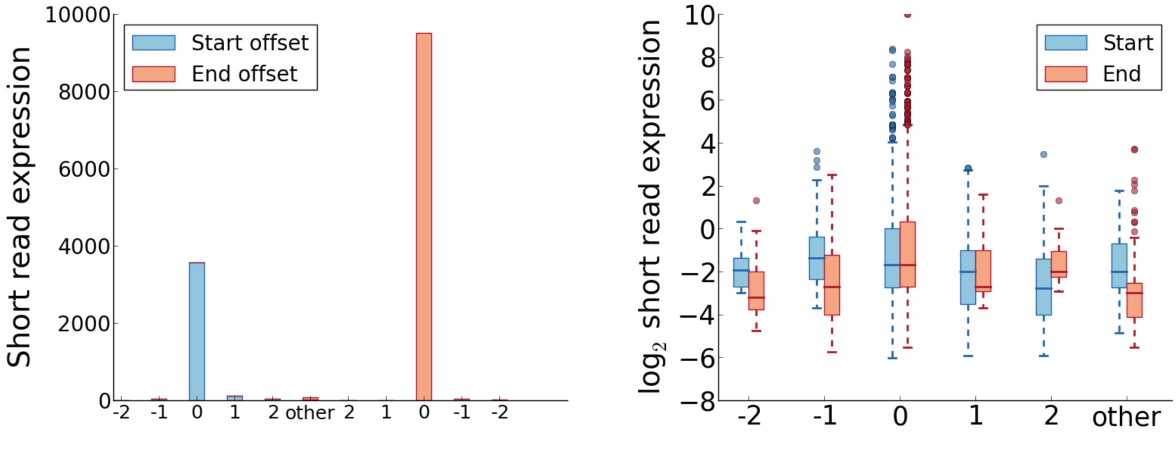
The classification threshold of $FC = 10$ presented in section 4.7.1 is observable in the figure, as all FC values for ’Different’ class is greater than and all FC values for ’Equal’ are less than 3.32, the log_2 transform of 10. Scenario A, with small n-values in all data sets, shows a random distribution in both 'Different' and 'Equal', with no visible difference. Scenario B shows a more linear distribution in the 'Different' class, while a random distribution in the 'Equal' class. Scenario C shows random distributions in all sets. Scenario D shows a partially linear distribution in the 'Different' class, while a random distribution in 'Equal'. By far, scenario B in 'Different' reveals the highest fold change values, followed by scenario C, and scenario D. An interesting observation is the generally lower fold changes of hairpins of scenario A than B.

In total, no distributions of the 'Equal' class seem causal, while scenario B and D appear at minimum partially causal in the 'Different' class, of which scenario B are both more linear and contain much higher fold change values. The fold change of a hairpin cannot alone fully predict the short read association of a miRNA, however the hairpins with the largest fold changes are associated with short reads in primary the guide strand, secondary in both. Most hairpins of scenario D in the 'Different' class have FC values half the maximum values of scenario B. To some extent, this thus indicates that the combination of high fold change and high expression levels of a hairpin can predict short read association of miRNAs.

4.8 Evaluating all expressed isomiRs

All prior analyses have been performed by regarding only the highest expressed miRNA or isomiR of each annotated hairpin strand, discarding all isomiRs lower expressed. The last years, the function of isomiRs have gained more research focus, and multiple reports state that isomiRs may be involved in biological functions just as the mature miRNAs, and that different isomiRs of the same miRNA might mediate different functions. If multiple isomiRs and not only the mature sequences perform functions in a cell, and so far my results indicate that short reads are associated with active miRNAs, it would be interesting to analyse the effect of including all isomiRs in the analyses. As explained for step four in Section 3.4.1, the processing of each data set can be done regarding all expressed unannotated isomiRs above a threshold of $rpm = 0.5$, essentially aligning short reads to the isomiR with primarily the best alignment, secondarily the highest expression level. All prior analyses have been reproduced using these settings, and the results are presented and discussed in this section: processing statistics, alignments, coexpression of short reads and isomiRs, short read association of isomiRs, terminal nucleotide preferences, short read lengths, ANOVA results, short read association of and correlation with isomiRs based on classification of fold change, and short read alignment scenarios of fold change classification.

First, the processing statistics for isomiRs are presented in Section 4.1, where the number of unannotated isomiRs range from 1,500 - 4,550 for the human samples, and 617-1,849 for the mouse samples, of which all samples contain clearly higher number of isomiRs than mature sequences. Generally, the share of miRNAs and isomiRs associated with short reads are reduced, which is expected as the number of isomiRs increase while the number of short reads are the same. An important note is that earlier, a short read aligning to a specific strand of a hairpin was only aligned to the mature sequence, resulting in many short reads aligning to the same sequences. Now, short reads align to the best sequence, and due to this the *Daub* data set actually shows an increased share of short



(a)

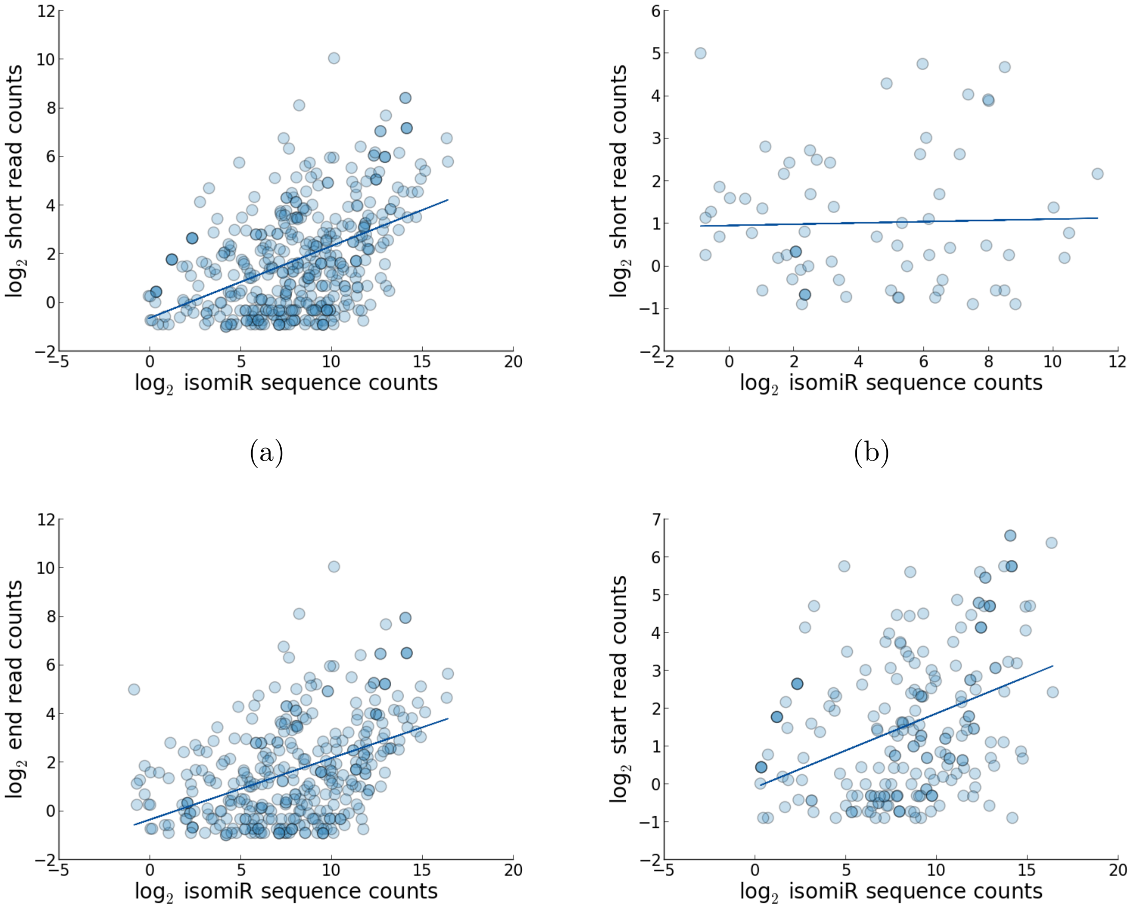
(b)

Figure 4.14: Short read alignment to isomiRs for the Meister dataset, with total expression levels (a) and quartile distributions (b).

read associated miRNA sequences, indicating a large group of miRNAs aligning better to isomiRs than mature sequences.

Second, the alignments presented in Section 4.2.1 are reproduced. The expected outcome is naturally better alignments, as short reads are now aligned to the best fitted isomiR, and the results for all data sets support this. The share of short reads aligning with offset '0' has drastically increased, and all other offsets, are drastically decreased. Generally, the quartiles of all offsets except '0' has decreased, along with all whiskers and outliers, while offset '0' has increased quartiles, whiskers and outliers, indicating that many of the short reads earlier not aligning to offset '0' now aligns perfectly with another isomiR. As many of the upper outliers have shifted accordingly, it seems not only lowly expressed short reads have been affected. If short reads truly represents functional miRNAs, it seems isomiRs might also be functional. This shift is found in all data sets, and the results are presented in Appendix G.1, Figures G.1 through G.6. The results for the *Meister* data set are presented in Figure 4.14 as an example.

Third, analysing the shift in the correlation of short read and isomiR expression is important. Short reads aligning perfectly with isomiRs might just illustrate that there are many lowly expressed isomiRs of the same strand, and at least one of them happens to align perfectly with a short read. If this is the case, the expected results would be a non-linear expression correlation, and an increased number of lowly expressed isomiRs aligning to short reads of random expression levels. All data sets do indeed reveal a reduced regression line slope for isomiRs, however the distribution is still linear. Additionally, the isomiRs are not only found at the bottom end of the isomiR expression scale, but rather evenly distributed. In total, this indicates that indeed some isomiRs might be lowly expressed and perhaps randomly align well with short reads, however the even distribution and continued linear correlation rather indicates that many isomiRs truly correlates with short reads and their function of origin. The majority of short reads still align to the guide strand, and a slight preference in expression levels for end reads are observable. The original results for *Daub* were presented in Figure 4.4, and the updated results are presented in Figure 4.15 as an illustration of patterns found in all data sets. The updated results for all data sets are found in G.7.



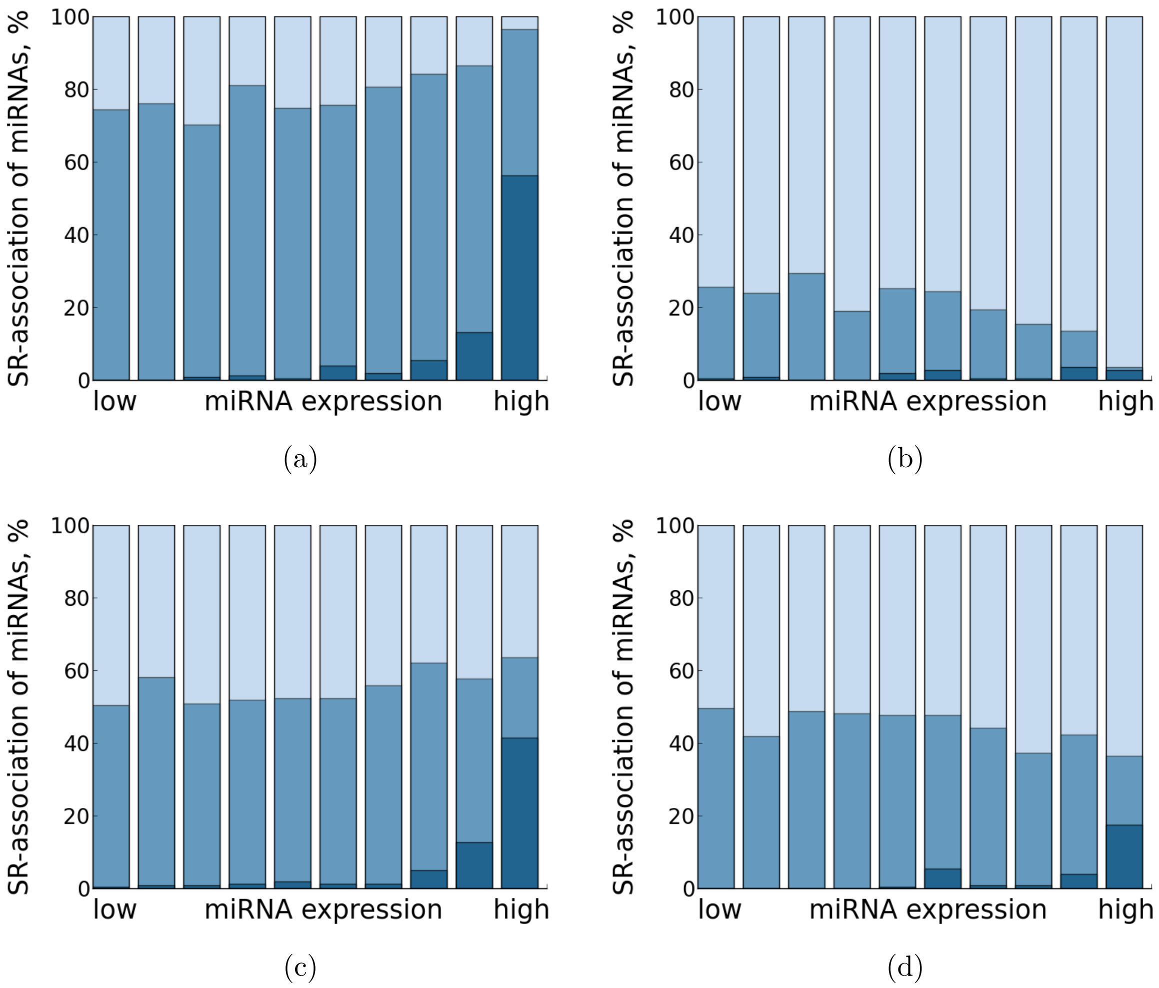
(a)

(b)

(c)

(d)

Figure 4.15: Correlation of short reads and isomiRs for the Daub data set, for the (a) guide strand, (b) passenger strand, (c) end reads and (d) start reads.



(a)

(b)

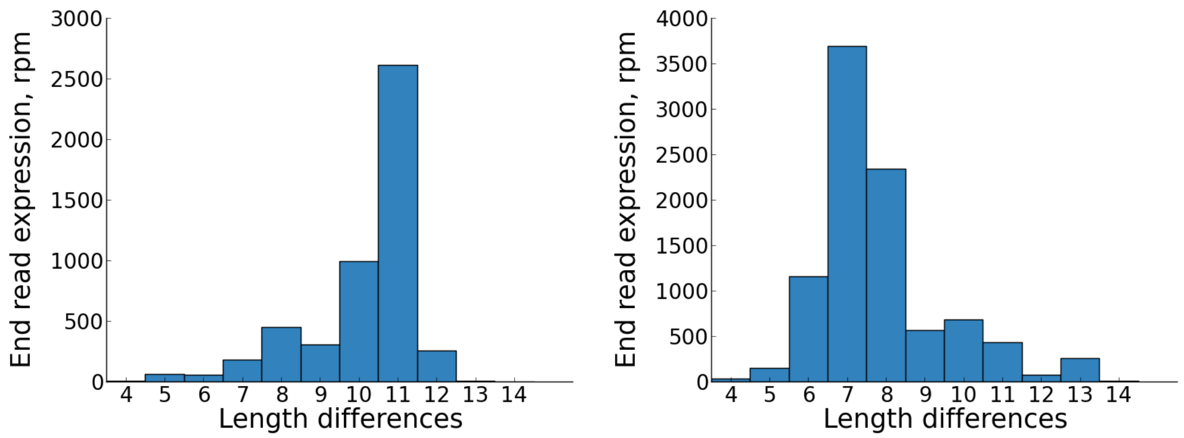
(c)

(d)

Figure 4.16: Short read association of isomiRs compared to expression levels for the (a) guide strand, (b) passenger strand, (c) 5' strand, and (d) 3' strand for the Daub data set

Fourth, the short read association of isomiRs are expected to be generally lower than for mature sequences, as the number of miRNA sequences has drastically increased while the number of short reads is still the same. As reported for the processing statistics, the share of miRNA sequences associated with short reads has generally decreased. This is also seen in the updated graphs for short read association, where the share of short read associated miRNAs is generally lower. An interesting observation is that the decrease is smaller for the 10% highest expressed isomiRs, which appears more preferred. This is however not that surprising, as the mature miRNAs still are the highest expressed miRNAs, and the majority of the additional isomiRs are found in the lower expression bins. The fact that the short read association of miRNAs and isomiRs still increases with the expression level supports the prior indications of short read associated isomiRs being actually functional and not just random sequences that happen to align with short reads. The prior findings of miRNAs residing in the guide strand being more associated with short reads are observed also for isomiRs. The results for the *Daub* data set are presented in Figure 4.16 as an illustration of the trend found in all data sets.

Fifth, the terminal nucleotides of miRNAs and isomiRs are investigated as in Section 4.4. Other than an observation of isomiRs obtaining similar nucleotide distributions as



(a) (b)

Figure 4.17: Length difference between end reads and their corresponding isomiRs for the (a) Daub and (b) Meister data set.

mature miRNAs for both the 5' and 3' end, no consistent changes are observed across the samples. For the *Daub* data set, the terminal nucleotide preferences of isomiRs related to their short read association is presented in Figure G.9, and the comparison of terminal nucleotide distributions of mature sequences and isomiRs are presented in Figure G.8, as an illustration of the tendencies observed for all data sets.

Sixth, the length of short reads are investigated anew in the same manner as in Section 4.5. As the set of short reads are the same, the length distributions are identical to the original results. The difference between isomiR length and end read length is thus potentially very interesting. If the additional isomiRs are the true aligned miRNA sequences for short reads, the length difference distribution should be more true and representative of the actual relationship between short reads and isomiRs. However, as with the original results, no consistent findings are observed across the data sets. Some data sets obtain a shift in the preferred length and some obtain a clearer preference for the original length, however the preferred lengths are still not similar across the sets. The results for *Daub* and *Meister* are presented in Figure 4.17 as an illustration of the differences between the data sets.

Seventh, the attempt at identifying statistically significantly differentiating features of short reads performed in Section 4.6 did fail to identify a differentiating feature, but rather illustrated the expression difference of mature miRNAs between the 5' and 3' strand of hairpins in the samples. The ANOVA test was repeated to include all isomiRs, and the updated results are presented in Table 4.8. As the set of short reads are still the same, and rather only the position and offset might have changed, the results should be similar for most features. As in the original results, the strand does still yield p-values below the significance level of $p = 0.5$ in all data sets except *Meister*, and when comparing the total expression levels of both strand, the fold change is still similar. However, in addition to *Meister* and *Lundbæk*, now also *Corry* contain significant p-values for both the Argonaute and the position of short reads, thus three out of five data sets find these features significant. This is interesting, and as reported in prior sections, a tendency for higher expression levels for end reads and steeper correlations between end reads and miRNAs might support the position being a significant feature, however for *Corry*, a

Table 4.8: ANOVA p-value results for isomiRs

Feature	**Meister**	**Daub**	**Rajewsky**	**Corry**	**Lundbæk**
Ago	**0.000872**	0.090556	0.971286	**0.013734**	**0.007652**
Strand	0.555809	**0.000246**	**0.000428**	**5.26e-11**	**0.000498**
Position	**0.009786**	0.214910	0.871202	**0.000643**	**2.75e-12**
Offset	0.478778	0.555482	0.225339	0.595797	0.097293
Ago + strand	0.160115	0.390001	0.993178	0.295148	**0.000292**
Ago + position	**0.020399**	0.683498	0.057967	0.259445	0.860303
Strand + position	0.145760	0.468465	0.119010	**0.038085**	0.915800
Ago + offset	0.843927	0.956229	0.692706	0.395862	0.999768
Strand + offset	0.094984	0.874761	0.681754	**0.020003**	0.622750
Position + offset	0.957971	0.785608	0.547024	0.222115	0.368370
Ago + strand + position	0.125507	0.737716	0.163986	0.582698	0.519110
Ago + strand + offset	0.214729	0.991144	0.957237	**0.033914**	0.999905
Ago + position + offset	0.847583	0.994175	0.368769	0.892783	1.000000
Strand + position + offset	0.255565	0.425632	0.229679	**0.043782**	0.377377
Ago + strand + position + offset	0.686131	0.956324	0.864973	0.892852	1.000000

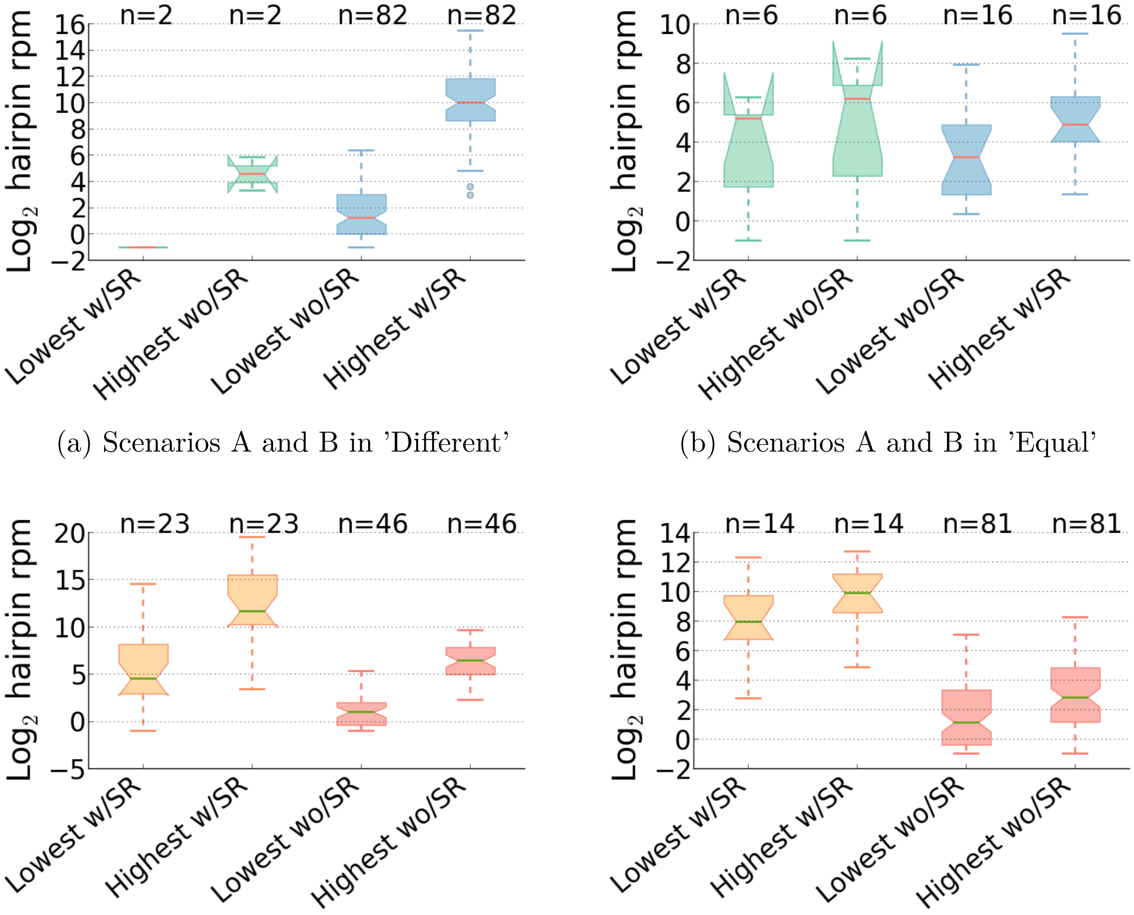
preference for start reads are found. No consistent differences between Argonautes have been found however. Three out of five data sets is not sufficient to conclude the features significant, and except illustrating the continued preference for 5' strands, these results do not provide any new insight.

Eighth, the classification scheme presented in Section 4.7.1 is applied to all isomiR alignments, and first the short read association of strand in both the 'Different' and 'Equal' class is investigated. The same pattern is found as in the original results of Section 4.7.2, however the class sizes are updated, and generally, the number of hairpins in the 'Different' class has increased while 'Equal' has decreased. The correlation of expression of short reads aligning to the guide strand and their corresponding isomiRs yielded the results presented in Table 4.9, where compared to the original results in Table 4.6, the slopes of all data sets except *Corry* are steeper in D than E. Correlations that evolve more significant as n increases should statistically be more correct, and this might indicate that there is indeed a more causal relationship between the expression of short reads and hairpins if the hairpins has a high fold change.

Nineth, the short read alignment scenarios defined in Section 4.7.4 are applied to the classification scheme, and the expression of hairpins of each scenario is plotted again for each class. When compared to the original results, the quartiles of scenarios A and B are slightly increased for the 'Different' class of all data sets, while scenarios C and D are slightly increased for the 'Equal' class of all data sets. The drastically low number of hairpins of scenario A is still observable, and the preference of 'Different' class for scenario B and 'Equal' class for scenario D still holds. Figure 4.18 presents the updated results for the *Meister* data set as an illustration of the observations in all data sets. As for the fold changes of the different scenarios in the two classes, there are no changes from the original results of Section 4.7.5.

Table 4.9: isomiRs

	Different				Equal			
Data set	r	s	p	n	r	s	p	n
Corry	0.220	0.201	0.252	29	0.880	0.692	0.120	4
Daub	0.408	0.394	0.000	98	0.194	0.116	0.456	17
Lundbæk	0.532	0.407	0.000	463	0.255	0.103	0.006	115
Meister	0.565	0.527	0.000	82	0.381	0.330	0.146	16
Rajewsky	0.410	0.305	0.001	65	0.492	0.104	0.104	12



(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal'

Figure 4.18: The expression levels of each isomiR hairpin strand of scenario A (green), B (blue), C (orange) and D (red) for the *Meister* data set in the 'Different' and 'Equal' class.

In total, isomiRs align well with short reads, their expression correlates with short reads, higher expression levels of isomiRs predict higher association of short reads, the nucleotide preferences of isomiRs are similar to mature sequences, the strand preference of hairpins including isomiRs predicts a steeper correlation of short read and hairpin expression, and the expression of alignment scenarios are slightly increased. These findings indicate that isomiRs yield more significant results than mature sequences alone, and support prior reports of isomiRs serving biological functions in the cell.

Chapter 5

Discussion and evaluation

The rationale for this study, as presented in Section 2.6, was to reproduce and extend my experiments from Wahl (2014). No prior studies of short reads have been reported, due to a common assumption that short reads are products of either Ago2 cleavage or known degradation processes. This chapter will provide a discussion regarding the credibility of these assumptions and the implications of my findings presented in Chapter 4, and the discussion is threefold. First, the credibility of the prior assumptions and my findings in Wahl (2014) are discussed. Second, a discussion of short read and miRNA features is provided. Third, possible Argonaute dependencies are discussed. Fourth, a novel model for miRNA activity in relation with Argonaute proteins is presented. Fifth, the limitations and reliability of the results are discussed, before finally, a conclusion of the overall findings ends this chapter.

5.1 Credibility of prior assumptions

The results presented in Section 4.2.1 and 4.2.2 verify the findings of Wahl (2014). Short reads are found to align well with either the start or end of mature miRNA sequences, and this tendency is found to not only hold for a small subset of miRNAs, but is a general tendency across all expressed miRNAs. Additionally, the coexpression of short reads and their associated mature miRNAs are found slightly linear, where the majority of short reads align to the guide strand, and short reads on the passenger strand yield random coexpression with miRNAs. These results are found for all data sets, and provide enhanced significance to the original findings of Mossin (2014) and Wahl (2014).

As discussed in Wahl (2014), the prior assumptions of short reads being either products of Ago2 cleavage or known degradation processes cannot hold for the observed results. As the report stated, for these assumptions to hold, the alignment of short reads should be randomly distributed along the hairpin strands, the coexpression of short reads and miRNAs should be strictly linear, and short reads should align to passenger strands. If Ago2 cleavage yielded short reads, there should only be a correlation of short reads and miRNAs for Ago2 samples, and there should be a significant difference between the Ago2 KO and WT samples of *Lundbæk*. As the results of all data sets of this study contradict these requirements, the prior assumptions regarded short reads are strongly discouraged.

In addition to the original analyses in Wahl (2014), results of this study further discourages the prior short read assumptions. The short read association of miRNAs are found to increase with the expression level of miRNAs (Section 4.3), and again the majority of short read associated miRNAs reside in the guide strand. In Section 4.6, the expression of miRNAs are found to be significantly differentiating for the expression level of short reads. When applying the fold change classification scheme in Section 4.7, miRNAs with large fold change values were found to be more prone to short read association, and miRNAs with short reads only on the passenger strand were merely found for miRNAs with high strand preferences. Together, this implies that short reads are associated with highly expressed guide strands of miRNAs, which is incompatible with the current assumptions on the origin of short reads.

5.2 Features of short reads associated miRNAs

All observations in this study discourage the prior assumptions of short reads being products of either Ago2 cleavage or known degradation processes. The origin of short reads and what they actually represent is unknown, and different approaches have been taken to attempt revealing the nature of short reads and short read associated miRNAs.

Short reads of this study are all within 11-15 nucleotides, align well with miRNAs, and their expression levels correlate with the expression level of miRNAs. Four data sets contained higher expression levels of end reads, while two data sets contained higher expression levels of start reads, thus the position is not found significant on a global level. However, all three human data sets contained higher end read levels, while the two mouse IP data sets contained higher start read levels, indicating a possible species difference. No significant, global miRNA independent features of short reads have been identified across the data sets, illustrated by the ANOVA results in Section 4.6. Features of individual short reads do not appear relevant, but rather the features of short read associated miRNAs.

For all expressed hairpins in all six data sets, the 5' strand is clearly higher expressed than the 3' strand, and the association and expression of short reads are accordingly higher for the 5' strand. The expression level of miRNAs were in Section 4.2.2 found to correlate with the expression level of short reads in a partly linear manner, and in Section 4.6, this correlation was found to be significant. One of the most important, consistent findings of this study is the significant difference of short read association of the guide strand compared to the passenger strand. This pattern is found in coexpression analyses, short read association analyses and miRNA classification by fold change analyses, where the number of short read associated miRNAs and their expression levels are significantly higher for the guide strand. In Section 4.7.4, short reads aligning to the passenger strand of hairpins with clear strand preferences were very rare; passenger strands were mostly associated with short reads if the hairpin lacked a clear strand preference or if also the guide strand was associated with short reads. This strongly discourages prior assumptions of the origin of short reads, and indicates a more significant relationship of short reads and miRNAs.

Different features and patterns of miRNAs with and without short reads have been analysed, and in Wahl (2014), an attempt at classifying individual miRNAs based on their short read association failed. In this study, another approach has been pursued, providing more significant results. When analysing short read association of miRNAs in Section 4.3, the vast majority was found within the 20% most expressed miRNAs, and the majority of these resided in the guide strand. In Section 4.4, miRNAs without short reads were found to possess a different terminal nucleotide distribution than miRNAs with short reads, notably with some constraints as to the reliability of the comparison due to drastically differences in expression levels between the groups of miRNAs. When classifying miRNA hairpins based on their strand preference in Section 4.7.2, the vast majority of short read associated miRNAs was found to originate from hairpins with clear strand preferences, and of these, the correlation of the expression level of short reads and guide strands were found more significant and steep in Section 4.7.3. The short read association of hairpin strands was found to be a differentiating feature for hairpin expression in Section 4.7.4, and to some degree, higher fold change combined with higher expression levels of hairpins were found to predict miRNA short reads in Section 4.7.5. All these observations have one united implication: short reads are mainly associated with biologically active miRNAs.

Traditionally, high expression levels and clear strand preferences are signals of biologically active miRNAs, and the combination of high fold change and expression levels of guide strands conventionally predict active miRNAs. If the majority of short reads are associated with active miRNAs, then in addition to high expression and fold change values, short reads can actually be markers for active miRNAs.

5.3 Argonaute dependencies

Possible Argonaute dependencies were initially interesting as short reads were assumed to partly originate from Ago2 cleavage of passenger strands. From the beginning of my work with miRNA short reads in 2014, all analyses have included a differentiation on the associated Ago protein to enable comparisons. In Wahl (2014), I could not determine an Argonaute independent short read existence, but rather found that different Argonautes provided different levels of influence on short read expression. In this study I find no results indicating Ago dependency: there are no consistent terminal nucleotide differences for the different Argonautes among the data sets, short read alignment distributions are similar, short read and miRNA coexpressions are similar, length differences of miRNAs and short reads are similar, short read association of miRNAs are similar, short read association of different strands of the 'Different' and 'Equal' class are similar, and expression level of short reads aligning to the different Ago proteins are different across the data sets.

The ANOVA results of Section 4.6 identified the Argonaute protein as significant for the *Lundbæk* and *Meister* data set, which were the two data sets analysed in Wahl (2014), however this feature was not found significant for the other four data sets. The individual data sets show some variation among its different Argonaute IP samples, which separately are promising for an existence of Ago dependency, however these differences are not consistent across the data sets, and can not be concluded as significant. As such, I find no consistent, global indications of an Argonaute dependency of short read expression for the data available, however as subsamples of different Agos vary within data sets, I cannot prove an Argonaute independency of short read expression either.

Even though no consistent differences among the Argonaute proteins have been observed across the data samples, an important observation is the increased expression levels of short reads for Ago2 WT compared to Ago2 KO in the *Lundbæk* data set, presented in Figure 4.2. This might indicate that the presence of Ago2 influences the expression of short reads. As short read association is primarily found for the highest expressed miRNAs, the 75th and 90th percentile of expressed miRNAs of the KO and WT samples are compared, and the result are found in Appendix H, Table H.1. For the DOC cell line, both the 75th and 90th percentile is increased for the WT sample compared to the KO sample, and accordingly, the total short read expression level is increased. For the GH cell line, the percentiles are slightly increased in the KO sample, but not the short read expression. The DOC samples are genetically identical, whereas the GH cell samples are from two individuals of different gender, and as such, comparing the KO and WT samples of DOC are more reliable than GH. The short read alignment distribution is similar for Ago2 KO and WT from both cell lines. In total this indicates that Ago2 is not critical for short read expression and alignment, but rather the expression level of miRNAs per se. This might illustrate that short reads are associated with all three Argonaute proteins investigated in this study, and that a lack of any one of the three will reduce the level of short reads.

One drawback of the data sets analysed in this study is the non-negligible difference in the preparations and methods used to generate the data, a limitation discussed in Section 5.5.1. Different protocols have been used, and different Argonaute proteins are immunoprecipitated. This does not provide for a solid comparison of such delicate relations as the Argonaute-short read dependency appears to be. To gain a more reliable, informative comparison and investigation of Argonaute dependency of short read expression, a more substantial amount of comparable data is needed.

5.4 Modified model for miRNA activity

The current biogenesis model of miRNAs, as presented in Section 2.2, states that after the Dicer-TRBP complex cleaves off the pre-miRNA hairpin loop, the miRNA:miRNA* duplex is loaded into an Ago protein. The Argonaute must then determine which strand to retain, usually the strand with the thermodynamically less stable 5' end, a specific terminal nucleotide preference or base pairing in specific locations, and when the miRNA strand is selected, the miRNA* strand is evicted. The mature miRNA strand is then strongly protected by the Argonaute protein, while the miRNA* strand is rapidly degraded, resulting in the commonly observed high accumulation of miRNAs compared to miRNA*s. Short reads are assumed to be the remnants of different stages of this degradation of the miRNA* strand. The results of this study have firmly established that the majority of the observed short reads align to miRNA sequences rather than miRNA* sequences, and following, the current model is an insufficient representation of the reality of miRNA biogenesis. In this section, I first present a modified model based on the results of this study. Following, whether the results of this study supports the model is discussed. Lastly, a brief discussion on existing literature supporting the model and its implications are presented.

5.4.1 Model definition

Based on the results of this study, I suggest a modified model of miRNA activity and degradation, presented in Figure 5.1. The figure illustrates four scenarios that yield different origins of short reads. As in the original model, The Dicer-TRBP complex cleaves off the pre-miRNA hairpin loop, and the resulting miRNA:miRNA* duplex is loaded into an Argonaute protein, resulting in pre-RISC. Usually, the strand preference is already destined upon loading, and the 5' end of the guide sequence is immediately securely

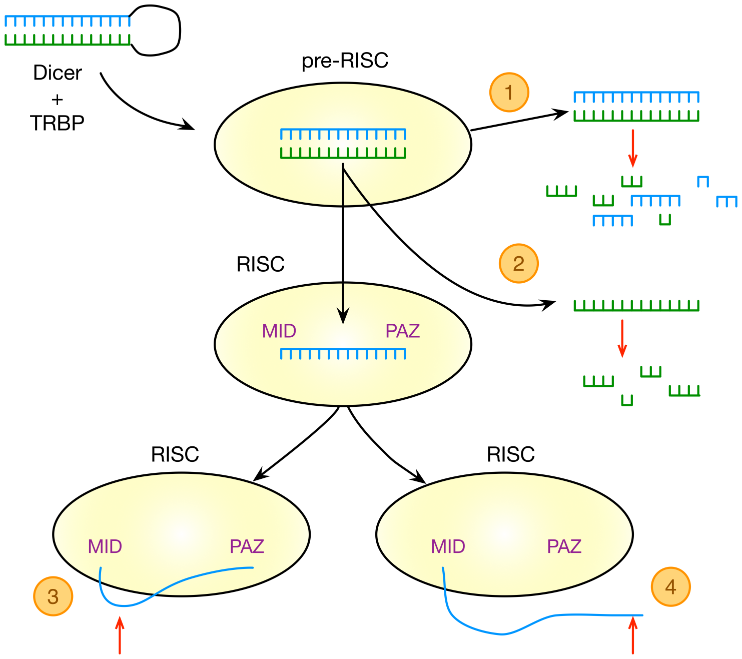


Figure 5.1: Modified miRNA model based on results of this study, where miRNA and miRNA* strands are represented in blue and green, respectively. The different origins of short reads are marked with number, where (1) represents miRNA:miRNA* duplexes released from Ago, (2) miRNA* strands evicted by maturation of RISC, (3) miRNA strands evicted from RISC by alteration of seed region, and (4) miRNA strands evicted from RISC by 3' alterations.

bound to the MID domain of the Argonaute protein. The duplex is then dissociated, and the unwinding is done by either cleavage by Ago2, or due to mismatches in specific locations. However, if the duplex is tightly bound the Argonaute might be unable to separate the strands, and the whole duplex is released from Ago, illustrated by scenario 1 in the figure.

If the unwinding succeeds, the 5' end of the guide strand is securely bound in the MID domain and the 3' end securely bound in the PAZ domain of the Argonaute, while the passenger strand is released, illustrated by scenario 2 in the figure. When the guide strand is incorporated into the mature RISC, the miRNA mediated gene regulation is carried out by base pairing the seed region to target mRNAs. In mature RISC, miRNA 5' end is always tightly bound. However, the seed is on open display to bind to mRNAs, and though not validated by any experiments, this might enable the seed region to be cleaved and modified by nucleases in the cell. If so, the remaining miRNA sequence would be released from Ago, illustrated by scenario 3 in the figure.

Following the two-state model presented in Section 2.2.3, the miRNA 3' end is released from Ago during base pairing with mRNA, enabling commonly observed 3' modifications. Additions of 3' adenine or uracil are common, as well as removal of terminal nucleotides. When the 3' end of the miRNA is altered, the Argonaute may reject rebinding the 3' end, and the whole sequence is released, illustrated by scenario 4 in the figure.

Following this model, short reads found in a miRNA associated sample may originate from four different scenarios, as illustrated in the figure. The first two scenarios leave either the passenger strand or both strands in the cell, unprotected and subject to degradation by multiple nucleases, and might be both cleaved, trimmed and modified. If the passenger strand of scenario 2 is evicted from Ago2, it is already cleaved before eviction. These miRNA and miRNA* sequences might equally yield start and end reads of different, random alignments. The last two scenarios involve degradation of incorporated mature miRNAs, where the first yields perfectly aligned end reads by altering the seed region, and the second yields perfectly aligned start reads by trimming the 3' end.

Caused by the four scenarios of short read origins, both start and end reads are expected to be observed, where the two first scenarios might yield random alignments, and the two last perfect alignments. As guide strands accumulate in significantly higher numbers than passenger strands, the expression of approximately perfectly aligned guide strand short reads are expected to be significantly higher than randomly aligned short reads and passenger strand short reads.

5.4.2 Support for model in results

The modified miRNA model is supported by multiple results of this study. Regarding short reads, the existence of start and end reads from both strands are observed in all data sets, and perfectly aligned short reads from the guide strand are much higher expressed than imperfectly aligned short reads, and short reads from the passenger strand. These findings support the model, and illustrate the existence of short reads of different origins. As the guide strand accumulates into higher expression levels, the short read association of miRNAs are expected to increase with the miRNA expression, which is found in the results. Hairpins with equal expression of both strands in a sample have been reported to often actually exist as miRNA:miRNA* duplexes in the cell, not unwinded and not degraded. Duplexes not easily unwinded and often rejected from pre-RISC might thus exist in the cell in both duplex and degraded forms, yielding full-length miRNAs of both strands, and short reads aligning to either end of them both. This is supported by the findings of this study, as the 'Equal' class of the hairpin classification based on fold change values of Section 4.7 actually represents, to some extent, the group of equally expressed miRNA strands of miRNA:miRNA* duplexes. Additionally, some hairpins might yield isomiRs of different strands that both can be incorporated into RISC, and as such, equally expressed pairs of the 'Equal' class hairpins might both yield short reads through targeting mRNA. Short reads are observed for the 'Equal' class, however in lower numbers, and highest expressed hairpins are those associated with short reads on both strands, further supporting the model.

In the 'Different' class, representing hairpins incorporated into mature RISC, the highest expression levels are observed for hairpins associated with short reads on the guide strand or both, where the first are greater in numbers than the latter. Few hairpins are associated with only passenger strand short reads, and might illustrate the level of noise in the data or an isomiR yielding a different strand preference of an Argonaute, essentially resulting in a perfectly aligned short read for the passenger strand. Not all hairpins with a clear strand preference are associated with short reads, which may be caused by the same fact reducing short reads in the data: the data analysed in this study contain only exact matches, and modified miRNAs are thus not included. As tailing are common during degradation, especially in 3' modifications, a substantial amount of short reads might not be present in these analyses.

The tendency for terminal nucleotide differences of miRNAs to be dependent on their short read association might indicate that inactive miRNAs do not yield short reads, which is supported by the model. The analysis of differences between miRNA and end read lengths in Section 4.5 did not reveal a consistent pattern, however after the end read is released from Ago, it would undergo further trimming and degradation, resulting in different lengths of end reads. In total, there are no strong discrepancies between the observations in this study and the suggested model.

5.4.3 Literature support and implications

As discussed in the rationale for this study, there are scarce amounts of literature on the subject of miRNA degradation. The eviction of miRNA:miRNA* duplexes if the Argonaute is unable to unwind the duplex is commonly agreed upon, however the process of degradation and eviction is not understood. The eviction of a passenger strand upon selection of a guide strand is conventional and frequently reported, where the passenger strand is cleaved if Ago2 is involved, and rapidly degraded when evicted. The degradation processes involved is not fully understood.

The model introduces two conditions for yielding short reads of incorporated guide strands. To yield end reads, the seed region of a miRNA is assumed attacked, of which no significant evidence in the literature is found. As of my knowledge however, no evidence exists of the contrary either, and my findings are supported by the literature review by Ruegger and Grosshans (2012) which actually presents the possibility of target mRNA binding of miRNAs to be a factor for miRNA degradation. To yield start reads, the miRNA 3' end must be exposed for attack. The *two-state* model accounts for 3' end exposure, and releasing of the miRNA 3' end is reported by multiple experiments. Regarding degradation processes, 3' tailing and 3'-to-'5 trimming of miRNAs are reported to be commonly observed for miRNAs with an extensive complementarity to target mRNAs (Ameres et al., 2010). If both the seed region and 3' end of guide strands are exposed, the strand might yield both start and end reads. The determining factor for what sequence end that is attacked is unknown, however the level of complementarity between the miRNA and target mRNA might be of importance.

However scarce, this model is supported by the available literature, and its implications are intriguing. First, as remarked by Ruegger and Grosshans (2012), if target mRNAs influence the stability of miRNAs, the current concept of miRNAs regulating mRNAs must be altered, into a more complex mutual regulation of miRNAs and mRNAs. Second, the short read association of highly expressed miRNAs implicates that short reads might be used as markers for biologically active miRNAs, and when combined with fold change values, a more precise prediction is achieved.

5.5 Evaluation

This section first presents a discussion on the basis of comparison between the data sets, and an overall presentation of weaknesses and limitations of the analyses presented in this report. Second, the strengths and findings supporting reliable patterns are discussed, concluding which results are not reliable, and which represent true, biological relations.

5.5.1 Limitations

For this study, all six NGS data sets presented in Section 3.2 have been processed as described in Section 3.3 and parsed as described in Section 3.4.1. As such, all data sets have been equally processed in my analyses, and should be comparable across the data sets. However, the procedures and technologies used in the different laboratories to isolate miRNAs, generate and process the RNA sequencing data, were not identical, and the basis of comparison might not be fully present.

For the human data sets, immunoprecipitation was performed by using known antibodies and agarose beads in *Meister*, known antibodies and polymer beads in *Daub*, and tagged Ago proteins and magnetic beads in *Rajewsky*. For the mouse samples, immunoprecipitation was performed by known antibodies however with unknown bead types in *Corry*, and known antibodies and agarose beads in *Rui*. *Lundbæk* performed knock out of Ago2, and provides KO and WT samples of two different cell lines. The starting point for comparing these data sets is thus uncertain.

A major difference is in the IP procedure performed by *Rajewsky* compared to the others, where Ago proteins are tagged and bound to magnetic beads. Magnetic beads are perceived to reduce background noise compared to agarose beads, however tagging proteins might obscure their natural function and even introduce new functionality to the proteins, and the true relevance of immunoprecipitating tagged proteins is uncertain. As the *Rajewsky* data set differs from the other two human data sets in many analyses, both in number of unique alignments, read depth, and short read expression and behaviour, ruling out eventual significant differences due to preparations is not possible, and comparison of the *Rajewsky* data set with the others are not fully reliable.

The *Meister*, *Daub*, *Corry* and *Rui* are all produced using antibodies, and non-magnetic beads. However, the data sets show significant differences, including variations in the number of unique alignments and read depth, where *Rui* represents the extreme. Other unknown parameters might be of influence, such as washing procedures, incubation times and centrifugation, as well as the technology and parameters used for sequencing. The extreme difference in unique alignments for *Rui* indicates specific settings not known, and the low levels renders complex comparisons with this data set unreasonable. The *Corry* data set contain triple samples of supposedly the same procedure and technique, however the results for the individual samples show great variations, even in the number of unique alignments and expression level of hairpins, where e.g. a standard deviation of 23,000 is present for the unique alignments of the three duplicate samples of Ago2. *Corry* yields different results than the other data sets in multiple analyses of this study, which might illustrate the unreliable content of this data set or merely the level of noise possibly present in all data sets.

Caution is advised for comparisons of both *Corry*, *Rajewsky* and *Rui*, and as *Lundbæk* does not contain IP data, the base of comparison across all six data sets is not convincingly strong. The data sets contain IP or KO/WT of different Argonautes, and even though no significant Ago difference was found, they might not be truly comparable. As comparing the data sets in the first place introduces uncertainty, there might actually exist Ago dependencies not visible, and combining the different Ago samples and comparing the total samples might enhance the behaviours common for the different Agos, while obscure behaviours that are not common. This might result in some of the anomalies observed in the results.

The methodology used in my data processing might also be flawed. Requiring exact match in sequence alignment was intended to increase accuracy and reduce noise, however this might also limit the possibility of obtaining holistic results. Requiring exact matches inhibits non-templated isomiRs and modified reads, such as tailing of short reads, and thus many interesting sequences might be lacking from the analyses performed in this study. If start reads are results of guide strands modified and trimmed at the 3' end while in RISC, then a significant amount of start reads might not be included in the data, possibly explaining a tendency for higher expression levels of end reads than start reads in the human samples.

Another possible flaw of my data processing is the use of expression level thresholds, where $t = 0.5rpm$ has been consistently used. Whether this threshold is a sufficient level is unknown to me, however applying a threshold generally introduces a more reliable basis for comparison across the data sets. Another threshold of question is the fold change threshold for hairpin classification, where $FC = 10$ was used. Whether other threshold values are more realistic in true biological settings are unknown, and there might exist more sophisticated classification schemes, however the significant difference in fold change patterns of the different alignment scenarios of hairpins above and below the threshold indicates a rational classification.

5.5.2 Strengths and reliability

Based on the limitations presented in the last section, finding differing results across the data sets in many of the analyses presented in this paper is not surprising, such as nucleotide preferences, short read lengths, significantly differing features, and group size and expression of alignment scenarios of classified hairpins. However, not all results varied among the data sets. The alignment of short reads to miRNAs was indisputably well in all data sets, even in *Rui* which contain very few short reads. The correlation of short reads and miRNAs were also slightly linear in all data sets. The share of short read associated miRNAs were similar for all sets except *Rui*, and the short read association of miRNAs increased in accordance with miRNA expression levels for the same data sets. In all five data sets the expression of short reads and the share of short read associated miRNAs were greater for the guide strand, and all data sets revealed a tendency for nucleotide differences among miRNAs without associated short reads. The short read association of the different strands of the 'Different' and 'Equal' class were found to be significantly different in all data sets, and besides *Corry*, the correlation of hairpin and short read expression for the guide strand was found better and steeper in the 'Different' class. The short read association of hairpin strands was found differentiating for the hairpin expression level in all data sets.

In addition to the results from analysing mature miRNAs, when including all templated isomiRs, most results were similar or more significant. This indicates that even though the data may be limited, the paramount patterns observed are likely to be true, as they converge when additional data is introduced to the analyses. Some exceptions of detailed analyses exist, illustrating the noise and inaccuracy of compound and complex relationships in the data. However, even though a certain level of noise is likely present, the paramount patterns are clear and reproducible in all data sets: short reads align well with miRNAs, almost perfectly when regarding all templated isomiRs, and the majority of short reads align to highly expressed miRNAs, and especially the guide strand. The strand preference and fold change of hairpins are found to be strong predictors for short read association in all data sets, and the coexpression of short reads and miRNAs are significant in all data sets.

Importantly, no results support the prior assumptions of short reads being products of either Ago2 cleavage or known degradation processes of passenger strands in the cell. In total, the concluded findings of this report are reliable, and especially, the presented modified model of miRNA is supported by the results, independently of the processing limitations.

5.6 Conclusion

By conducting the experiments of this study, I have not found any evidence supporting the prior assumptions of short reads being merely products of Ago2 cleavage or known degradation processes of passenger strands. A more complex relationship of short reads and miRNAs appear to be present, where the majority of short reads align to guide strands, and the majority of short read associated miRNAs are the highest expressed miRNAs of the samples, while the existence of passenger strand short reads and imperfect aligned short reads also exist in lower numbers. In total, this indicates that there might be different origins for short reads, and a modified model of miRNA activity and degradation are presented, based upon the reliable findings of this study. The implication of this model is that the majority of short reads, which are highly expressed and perfectly aligned, are markers for biologically active miRNAs in the cell.

Possible Argonaute dependencies of short read expression and alignments are not found across the data sets. However, the individual samples of the data sets contain variations, indicating data set specific Argonaute differences, which might not be reproducible among the data sets due to noise, a constrained base of comparison and the fact that not all data sets contain the same Argonaute samples. Hence, this study proves neither Argonaute dependency nor independency of short read expression. However, the paramount patterns observed and correlation of short reads and miRNAs are existent in all Argonaute samples, indicating that this is a general relation. Specific features and more complex behaviours related to specific Argonautes might exist, even though not observable on a global level in this study.

Chapter 6

Conclusion and future work

The outset for this study was to assess the credibility of my findings in Wahl (2014) as well as the common assumptions of short reads being merely products of Ago2 cleavage or degradation of passenger strands. My results and the discussion in the last Chapter strongly discourage these assumptions, and the findings of Wahl (2014) are verified and their significance enhanced. Additionally, short reads are found to mainly associate with the highest expressed miRNAs, especially guide strands of hairpins with clear strand preferences. The conclusion implicates that rather than being unimportant remnants of passenger strands, short reads are actually markers for biologically active miRNAs. A modified model of miRNA activity and degradation based on the results of this study is presented, where different origins for short reads are elaborated.

The scope of this project has been restricted, and the results highlight potential directions for further research. First, an important requirement for future work on the subject would be to reproduce my analysis on data sets with a more reliable base of comparison, especially regarding Argonaute dependencies. Including more data sets of the same genome and conduct genome-specific experiments might be an alternative. Second, requiring exact matches omits all non-templated isomiRs and short reads with 3' tailing, thus the data analysed in this study might not be complete. Including additional non-templated reads might yield more significant results, and might enable an interesting analysis of the differences between start and end reads, where non-templated reads are expected to yield especially higher numbers of start reads.

Third, the proposed model of miRNA activity and degradation introduces new research areas. Studies to investigate whether miRNA/mRNA interactions are mutual and not a one-way process could yield most interesting results, altering the current understanding of the role of miRNAs. Whether the seed region of incorporated miRNAs actually is prone to modifications is important to verify the modified model.

Fourth, testing the sanity of my results by analysing an arbitrary set of small RNA sequences of an independent data set would be interesting. Especially, extracting the miRNAs associated with short reads should yield highly expressed, functional miRNAs. Simultaneously, comparing the set of short read associated miRNAs with the set of biologically active miRNAs should result in a high degree of overlapping miRNAs.

Whatever the approach, the goal should be to verify my findings, pursue a substantiated understanding of miRNA activity and degradation, and perform extensive analyses to identify more complex relations, features and possible Argonaute dependent behaviours.

Bibliography

Ambros, V. (2004). The functions of animal microRNAs. *Nature*, *431*(7006), 350-355.

Ameres, S. L., Horwich, M. D., Hung, J.-H., Xu, J., Ghildiyal, M., Weng, Z., & Zamore, P. D. (2010). Target RNA-directed trimming and tailing of small silencing RNAs. *Science*, *328*, 1534-1539.

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

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, *136*(2), 215-233.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., & Canaider, S. (2013). An estimation of the number of cells in the human body. *Annals of Human Biology*, *40*(6), 463–471.

Burroughs, A. M., Ando, A., de Hoon, M. J., Tomaru, Y., Nishibu, T., Ukekawa, R., Funakoshi, T., Kurokawa, T., Suzuki, H., Hayashizaki, Y., & Daub, C. O. (2010, October). A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness. *Genome Research*, *20*(10), 1398-1410.

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

Burrows, M., & Wheeler, D. J. (1994). *A Block-Sorting Lossless Data Compression Algorithm* (Tech. Rep.). Digital SRC Research Report.

Carthew, R. W., & Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, *136*(4), 642-655.

Cenik, E. S., & Zamore, P. D. (2011). Argonaute proteins. *Current Biology*, *21*(12), 446-449.

Choo, K. B., Soon, Y. L., Nguyen, P. N. N., Hiew, M. S. Y., & Huang, C.-J. (2014). MicroRNA-5p and -3p co-expression and cross-targeting in colon cancer cells. *Journal of Biomedical Science*, *21*(1), 95.

Cloonan, N., Wani, S., Xu, Q., Gu, J., Lea, K., Heater, S., Barbacioru, C., Steptoe, A. L., Martin, H. C., Nourbakhsh, E., Krishnan, K., Gardiner, B., Wang, X., Nones, K., Steen, J. A., Matigian, N. A., Wood, D. L., Kassahn, K. S., Waddell, N., Shepherd, J., Lee, C., Ichikawa, J., McKernan, K., Bramlett, K., Kuersten, S., &

Grimmond, S. M. (2011). MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biology*, *12*(12), R126.

Dueck, A., Ziegler, C., Eichner, A., Berezikov, E., & Meister, G. (2012). MicroRNAs associated with the different human Argonaute proteins. *Nucleic Acids Research*, *40*(19), 9850-9862.

Elkayam, E., Kuhn, C.-D., Tocilj, A., Haase, A. D., Greene, E. M., Hannon, G. J., & Joshua-Tor, L. (2012). The Structure of Human Argonaute-2 in Complex with miR-20a. *Cell*, *150*(1), 100-110.

Farazi, T. A., Brown, M., Morozov, P., ten Hoeve, J. J., Ben-Dov, I. Z., Hovestadt, V., Hafner, M., Renwick, N., Mihailović, A., Wessels, L. F. A., & Tuschl, T. (2012). Bioinformatic analysis of barcoded cDNA libraries for small RNA profiling by next-generation sequencing. *Methods*, *58*(2), 171-187.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd.

Gusfield, D. (1997). *Algorithms on strings, trees and sequences: computer science and computational biology.* Cambridge University Press.

Hoopes, L. (2008). Introduction to the gene expression and regulation topic room. *Nature Education*, *1*(1), 160.

IHGSC. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860-921.

Kaboord, B., & Perr, M. (2008). Isolation of Proteins and Protein Complexes by Immunoprecipitation. *Methods in Molecular Biology*, *424*(1), 349-364.

Kandeel, M., Al-Taher, A., Nakashima, R., Sakaguchi, T., Kandeel, A., Nagaya, Y., Kitamura, Y., & Kitade, Y. (2014, May). Bioenergetics and Gene Silencing Approaches for Unraveling Nucleotide Recognition by the Human EIF2C2/Ago2 PAZ Domain. *PLoS One*, *9*(5), e94538.

Kozomara, A., & Griffiths-Jones, S. (2014). MiRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, *42*(Database Issue), 68-73.

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

Langmead, B., Trapnell, C., & Pop, S., Mihai andSalzberg. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), 25.

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

Leinco Technologies, I. (2015). *Immunoprecipitation Protocol.* `http://www.leinco.com/`

immunoprecipitation. (Accessed: 20.04.2015)

Lewin, B. (2006). *Essential Genes.* Pearson Education.

Mah, S. M., Buske, C., Humphries, R. K., & Kuchenbauer, F. (2010). miRNA*: a passenger stranded in RNA-induced silencing complex? *Critical Reviews in Eukaryotic Gene Expression*, *20*(2), 141-148.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1).

Mattick, J. S., & Makunin, I. V. (2006). Non-coding RNA. *Human Molecular Genetics*, *15 (suppl 1)*, 17-29.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., & Tuschl, T. (2004, July). Human Argonaute2 Mediates RNA Cleavage Targeted by miRNAs and siRNAs. *Molecular Cell*, *15*(2), 185-197.

Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C., Chen, S., Hannon, G. J., & Qi, Y. (2008, April). Sorting of Small RNAs into Arabidopsis Argonaute Complexes Is Directed by the 5' Terminal Nucleotide. *Cell*, *133*(1), 116-127.

Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C. J., & Marra, M. A. (2008, April). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research*, *18*(4), 610-621.

Mossin, J.-P. S. (2014). *Studying differential isomiRs in a high-throughput sequencing data identifies miRNA end-reads as a novel, putative miRNA degradation product* (Unpublished master's thesis). Norwegian University of Science and Technology.

National Library of Medicine US, NLM. (2015). *Genetics home reference [internet].* Retrieved from http://ghr.nlm.nih.gov/handbook/illustrations/dnastructure (Accessed: 03.03.2015)

Polikepahad, S., & Corry, D. B. (2013, January). Profiling of T helper cell-derived small RNAs reveals unique antisense transcripts and differential association of miRNAs with argonaute proteins 1 and 2. *Nucleic Acids Research*, *41*(2), 1164-1177.

Ruegger, S., & Grosshans, H. (2012). MicroRNA turnover: when, how, and why. *Trends in Biochemical Sciences*, *37*(10), 436 - 446.

Rutz, S., & Scheffold, A. (2004). Towards in vivo application of RNA interference - new toys, old problems. *Arthritis Research & Therapy*, *6*(2), 78-85.

Rybak-Wolf, A., Jens, M., Murakawa, Y., Herzog, M., Landthaler, M., & Rajewsky, N. (2014, November). A variety of dicer substrates in human and C. elegans. *Cell*, *159*(5), 1153-1167.

Saito, T., & Saetrom, P. (2010). MicroRNAs - targeting and target prediction. *New Biotechnology*, *27*(3), 243-9.

Sasaki, H. M., & Tomari, Y. (2012). The true core of RNA silencing revealed. *Nature Structural and & Molecular Biology*, *19*, 657 - 660.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., & Gibrat, J. F. (2012). Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of Computational Biology*, *19*(6), 796-813.

Seong, Y., Lim, D.-H., Kim, A., Seo, J. H., Lee, Y. S., Song, H., & Kwon, Y.-S. (2014, October). Global identification of target recognition and cleavage by the Microprocessor in human ES cells. *Nucleic Acids Research*, *42*(20), 12806-12821.

Soifer, H. S., Rossi, J. J., & Saetrom, P. (2007). MicroRNAs in Disease and Potential Therapeutic Applications. *Molecular Therapy*, *15*(12), 2070-2079.

Sung, W.-K. (2010). *Algorithms in bioinformatics: a practical introduction.* Chapman & Hall/CRC.

Tan, G. C., Chan, E., Molnar, A., Sarkar, R., Alexieva, D., Isa, I. M., Robinson, S., Zhang, S., Ellis, P., Langford, C. F., Guillot, P. V., Chandrashekran, A., Fisk, N. M., Castellano, L., Meister, G., Winston, R. M., Cui, W., Baulcombe, D., & Dibb, N. J. (2014). 5ísomiR variation is of functional and evolutionary importance. *Nucleic Acids Research*, *42*(14), 9424-9435.

Tili, E., Michaille, J.-J., Costinean, S., & Croce, C. M. (2008). MicroRNAs, the immune system and rheumatic disease. *Nature Clinical Practice Rheumatology*, *4*(10), 534-541.

Tomari, Y., Du, T., & Zamore, P. D. (2007). Sorting of Drosophila Small Silencing RNAs. *Cell*, *130*(2), 299-308.

Wahl, K. (2014, December). *Studying miRNA short reads as possible products of either a novel degradation process or an unknown biological function.* (Report for specialization project in TDT4501 at NTNU)

Wang, D., Zhang, Z., O'Loughlin, E., Lee, T., Houel, S., O'Carroll, D., Tarakhovsky, A., Ahn, N. G., & Rui, Y. (2012, April). Quantitative functions of Argonaute proteins in mammalian development. *Genes & Development*, *26*(7), 693-704.

Wang, Y., Juranek, S., Li, H., Sheng, G., Wardle, G. S., Tuschl, T., & Patel, D. J. (2009). Nucleation, propagation and cleavage of target RNAs in argonaute silencing complexes. *Nature*, *461*(7265), 754 - 761.

Wang, Z. (2009). Epitope tagging of endogenous proteins for genome wide Chromatin immunoprecipitation analysis. *Methods in Molecular Biology*, *567*, 87-98.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57-63.

Wiki Kids Ltd. (2014). `http://wonderwhizkids.com/conceptmaps/Central_Dogma_of_Molecular_Biology.html`. (Accessed: 06.12.2014)

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80-83.

Wyman, S. K., Knouf, E. C., Parkin, R. K., Fritz, B. R., Lin, D. W., Dennis, L. M., Krouse, M. A., Webster, P. J., & Tewari, M. (2011). Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Research*, *21*(9), 1450-1461.

Zhang, Z., Qin, Y. W., Brewer, G., & Jing, Q. (2012). MicroRNA degradation and turnover: regulating the regulators. *Wiley Interdisciplinary Reviews: RNA*, *3*(4), 593-600.

Appendix A

Sample processing statistics

Table A.1: Processing of Meister

Sample		Ago1 IP	Ago2 IP	Ago3 IP	Mean	SD
Alignments	Unique	10,226	13,497	9,159	10,961	2,260
Reads	Top 10	141,379	77,654	80,468	99,834	29,399
Expression	5’	549,286	343,729	616,013	503,009	115,876
	3’	67,927	166,746	55,526	96,733	49,765
Hairpins	Expressed	590 (31.6%)	577 (30.9%)	582 (31.2%)	583 (31.2%)	7
	SR-associated	178 (30.2%)	224 (38.8%)	177 (30.4%)	193 (33.1%)	27
MiRNAs	Annotated	393 (15.3%)	410 (16.0%)	410 (16.0%)	404 (15.8%)	10
Matures	Unannotated	484	517	457	486	30
	SR-associated	197 (22.5%)	266 (28.7%)	197 (22.7%)	220 (24.7%)	40
IsomiRs	Unannotated	1,781	2,612	1,700	2,031	505
	SR-associated	447 (20.57%)	669 (22.14%)	380 (18.01%)	499 (20.2%)	151
Short reads	Candidates	1,346	1,844	970	1,387	438
	Actual	1,183	1,685	870	1,246	411

Table A.2: Processing of Daub

Sample		**Ago1 IP**	**Ago2 IP**	**Ago3 IP**	**Mean**	**SD**
Alignments	Unique	111,448	60,228	87,293	86,323	25,624
Reads	Top 10	32,652	98,879	36,471	56,001	30,360
Expression	5’	129,617	414,433	164,226	236,092	126,895
	3’	42,356	108,970	55,535	68,953	28,803
Hairpins	Expressed	487 (26.1%)	716 (38.4%)	556 (29.8%)	586 (31.4%)	117
	SR-associated	174 (35.7%)	235 (32.8%)	196 (35.3%)	202 (34.3%)	31
MiRNAs	Annotated	409 (16.0%)	562 (21.9%)	452 (17.6%)	474 (18.5%)	79
Matures	Unannotated	394	608	444	482	112
	SR-associated	199 (24.8%)	278 (23.8%)	237 (26.5%)	238 (24.9%)	40
IsomiRs	Unannotated	1,222	1,880	1,407	1,503	339
	SR-associated	366 (22.44%)	677 (27.72%)	482 (25.93%)	508 (25.7%)	157
Short reads	Candidates	1,232	2,119	1,449	1,600	462
	Actual	966	1,980	1,289	1,412	518

Table A.3: Processing of Rajewsky

Sample		**Ago2 IP**	**Ago3 IP**	**Mean**	**SD**
Alignments	Unique	31,075	37,041	34,058	4,219
Reads	Top 10	544,613	478,609	511,611	33,002
Expression	5’	290,194	311,432	300,813	10,619
	3’	287,999	297,107	292,553	4,555
Hairpins	Expressed	1,306 (70.0%)	1,310 (70.2%)	1,308 (70.1%)	3
	SR-associated	471 (36.1%)	406 (31.0%)	439 (33.6%)	46
MiRNAs	Annotated	1,378 (53.8%)	1,407 (54.9%)	1,393 (54.4%)	21
Matures	Unannotated	1,350	1,400	1,375	35
	SR-associated	649 (23.8%)	572 (20.4%)	611 (22.1%)	54
IsomiRs	Unannotated	4,598	4,506	4,552	65
	SR-associated	1,077 (18.02%)	1,123 (18.99%)	1,100 (18.5%)	33
Short reads	Candidates	3,090	2,932	3,011	112
	Actual	2,715	2,746	2,731	22

Table A.4: Processing of Corry Ago1

Sample		**Ago1a IP**	**Ago1b IP**	**Ago1c IP**	**Mean**	**SD**
Alignments	Unique	120,849	114,106	154,278	129,744	21,513
Reads	Top 10	37,795	19,118	54,278	37,064	14,363
Expression	5’	22,222	36,802	5,094	21,372	12,959
	3’	3,386	4,859	509	2,918	1,806
Hairpins	Expressed	397 (33.5%)	422 (35.6%)	385 (32.5%)	401 (33.8%)	19
	SR-associated	155 (39.0%)	197 (46.7%)	197 (51.2%)	183 (45.6%)	24
MiRNAs	Annotated	438 (20.7%)	516 (24.4%)	425 (20.1%)	460 (21.8%)	49
Matures	Unannotated	328	349	313	330	18
	SR-associated	180 (23.5%)	245 (28.3 %)	219 (29.7%)	215 (27.2%)	33
IsomiRs	Unannotated	560	834	236	543	299
	SR-associated	179 (17.99%)	279 (20.71%)	205 (31.11%)	221 (22.0%)	52
Short reads	Candidates	600	957	1,187	915	296
	Actual	403	660	804	622	203

Table A.5: Processing of Corry Ago2

Sample		**Ago2a IP**	**Ago2b IP**	**Ago2c IP**	**Mean**	**SD**
Alignments	Unique	170,735	210,866	171,799	184,467	22,869
Reads	Top 10	122,068	110,303	154,930	129,100	18,885
Expression	5’	49,607	90,423	163,496	101,175	47,113
	3’	3,951	7,219	15,037	8,736	4,651
Hairpins	Expressed	440 (37.1%)	469 (39.6%)	523 (44.1%)	477 (40.3%)	42
	SR-associated	184 (41.8%)	214 (45.6%)	228 (43.6%)	209 (43.8%)	22
MiRNAs	Annotated	450 (21.3%)	542 (25.7%)	625 (29.6%)	539 (25.5%)	88
Matures	Unannotated	404	434	488	442	43
	SR-associated	220 (25.8%)	266 (27.3%)	295 (26.5%)	260 (26.5%)	38
IsomiRs	Unannotated	461	636	972	690	260
	SR-associated	240 (26.37%)	356 (30.25%)	421 (26.38%)	339 (27.6%)	92
Short reads	Candidates	739	1,142	1,280	1,054	281
	Actual	571	884	1,049	835	243

Table A.6: Processing of Rui

Sample		**Ago1 IP**	**Ago2 IP**	**Ago3 IP**	**Mean**	**SD**
Alignments	Unique	4,624	4,052	4,890	4,522	428
Reads	Top 10	56,758	46,702	67,864	57,108	8,643
Expression	5’	434,002	422,371	414,568	423,647	7,985
	3’	75,758	67,948	74,880	72,862	3,493
Hairpins	Expressed	538 (45.4%)	499 (42.1%)	575 (48.5%)	537 (45.3%)	38
	SR-associated	25 (4.7%)	14 (2.8%)	14 (2.4%)	18 (3.4%)	6
MiRNAs	Annotated	531 (25.1%)	495 (23.4%)	563 (26.7%)	530 (25.1%)	34
Matures	Unannotated	432	416	493	447	41
	SR-associated	25 (2.6%)	14 (1.5%)	14 (1.3%)	18 (1.8%)	6
IsomiRs	Unannotated	1,693	2,005	1,850	1,849	156
	SR-associated	26 (1.17%)	15 (0.6%)	15 (0.62%)	19 (0.8%)	6
Short reads	Candidates	30	17	17	21	8
	Actual	29	16	15	20	8

Table A.7: Processing of Lundbæk DOC KO

Sample		**DOC KO1**	**DOC KO2**	**DOC KO3**	**Mean**	**SD**
Alignments	Unique	26,847	24,116	38,722	29,895	7,765
Reads	Top 10	10,525	10,574	16,967	12,689	3,025
Expression	5’	251,266	237,130	318,298	268,898	35,404
	3’	66,373	80,611	76,035	74,340	5,935
Hairpins	Expressed	335 (28.3%)	309 (26.1%)	355 (30.0%)	333 (28.1%)	23
	SR-associated	93 (27.8%)	82 (26.5%)	112 (31.6%)	96 (28.8%)	15
MiRNAs	Annotated	356 (16.9%)	346 (16.4%)	397 (18.8%)	366 (17.3%)	27
Matures	Unannotated	244	207	262	238	28
	SR-associated	105 (17.5%)	89 (16.1%)	132 (20.0%)	109 (18.0%)	22
IsomiRs	Unannotated	1,388	1,439	2,262	1,696	491
	SR-associated	159 (9.12%)	148 (8.19%)	200 (7.52%)	169 (8.2%)	27
Short reads	Candidates	436	456	698	530	146
	Actual	253	246	390	296	81

Table A.8: Processing of Lundbæk DOC WT

Sample		**DOC WT1**	**DOC WT2**	**DOC WT3**	**Mean**	**SD**
Alignments	Unique	29,253	35,432	32,246	32,310	3,090
Reads	Top 10	16,784	10,478	10,884	12,715	2,882
Expression	5’	242,629	126,400	160,564	176,531	48,775
	3’	132,619	96,636	121,748	117,001	15,068
Hairpins	Expressed	403 (34.0%)	394 (33.3%)	382 (32.2%)	393 (33.2%)	11
	SR-associated	108 (26.8%)	120 (30.5%)	110 (28.8%)	113 (28.8%)	6
MiRNAs	Annotated	488 (23.1%)	465 (22.0%)	462 (21.0%)	472 (22.3%)	14
Matures	Unannotated	296	274	287	286	11
	SR-associated	123 (15.7%)	137 (18.5%)	125 (16.7%)	128 (16.9%)	8
IsomiRs	Unannotated	1,959	1,739	1,740	1,813	127
	SR-associated	227 (9.27%)	246 (11.2%)	232 (10.54%)	235 (10.3%)	10
Short reads	Candidates	647	776	755	726	69
	Actual	418	457	469	448	27

Table A.9: Processing of Lundbæk GH KO

Sample		**GH KO1**	**GH KO2**	**GH KO3**	**Mean**	**SD**
Alignments	Unique	33,751	24,581	26,014	28,115	4,933
Reads	Top 10	16,869	14,835	14,022	15,242	1,197
Expression	5’	256,876	301,528	296,602	285,002	19,989
	3’	102,106	136,008	139,796	125,970	16,945
Hairpins	Expressed	368 (31.1%)	360 (30.4%)	378 (31.9%)	369 (31.1%)	9
	SR-associated	109 (29.6%)	94 (26.1%)	95 (25.1%)	99 (26.8%)	8
MiRNAs	Annotated	438 (20.7%)	399 (18.9%)	424 (20.1%)	420 (19.9%)	20
Matures	Unannotated	265	270	297	277	17
	SR-associated	122 (17.4%)	109 (16.3%)	106 (14.7%)	112 (16.1%)	9
IsomiRs	Unannotated	1,694	1,777	1,783	1,751	50
	SR-associated	233 (10.88%)	204 (9.38%)	207 (9.43%)	215 (9.9%)	16
Short reads	Candidates	665	480	531	559	96
	Actual	427	351	384	387	38

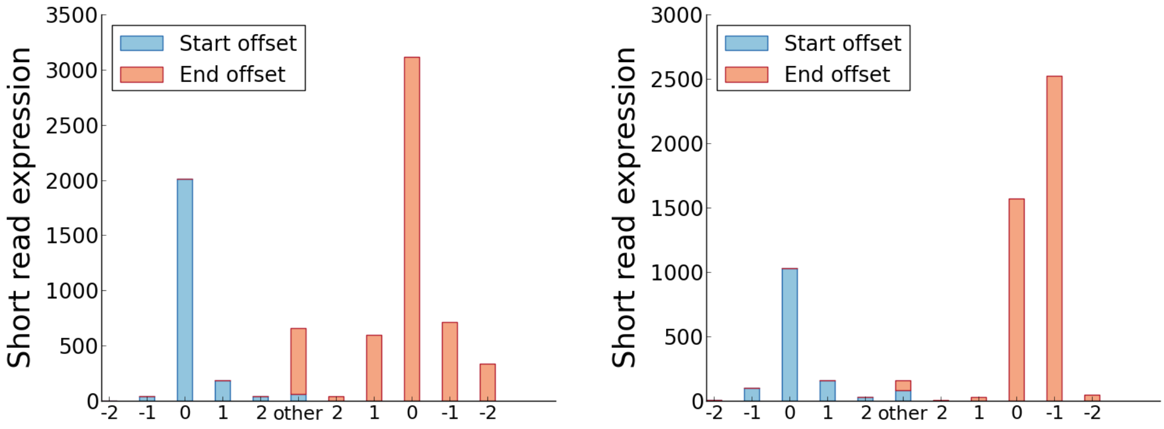
Table A.10: Processing of Lundbæk GH WT

Sample		**GH WT1**	**GH WT2**	**GH WT3**	**Mean**	**SD**
Alignments	Unique	29,658	33,601	34,618	32,626	2,620
Reads	Top 10	12,088	16,677	16,755	15,173	2,182
Expression	5’	130,569	179,545	187,988	166,034	25,313
	3’	107,606	135,196	138,830	127,211	13,941
Hairpins	Expressed	355 (30.0%)	392 (33.1%)	381 (32.2%)	376 (31.7%)	19
	SR-associated	94 (26.5%)	136 (34.7%)	127 (33.3%)	119 (31.6%)	22
MiRNAs	Annotated	388 (18.4%)	429 (20.3%)	438 (20.7%)	418 (19.8%)	27
Matures	Unannotated	301	309	248	286	33
	SR-associated	108 (15.7%)	152 (20.6%)	140 (20.4%)	133 (18.9%)	23
IsomiRs	Unannotated	1,775	1,889	1,855	1,840	59
	SR-associated	231 (10.73%)	305 (13.16%)	281 (12.26%)	272 (12.0%)	38
Short reads	Candidates	739	845	818	801	55
	Actual	486	605	569	553	61

Appendix B

Short read alignments

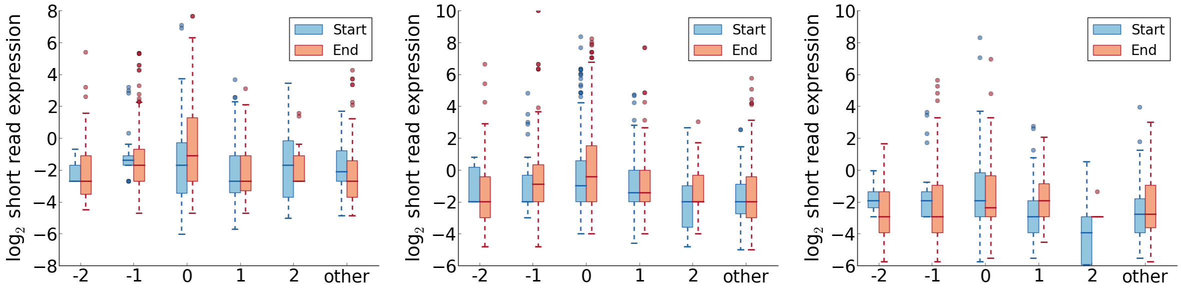
B.1 Meister alignments



(a) 5’ strand alignment

(b) 3’ strand alignment

Figure B.1: Total short read alignment for the Meister data set, shown for (a) the 5’ strand and (b) the 3’ strand.



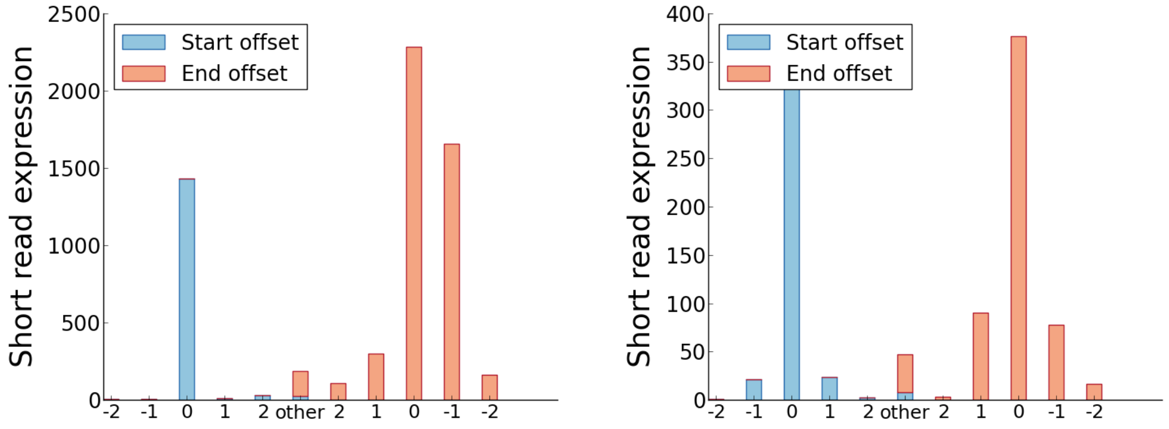
(a) Ago1 IP

(b) Ago2 IP

(c) Ago3 IP

Figure B.2: Alignment boxplots for the Meister data set, shown for (a) Ago1 IP, (b) Ago2 IP and (c) Ago3 IP.

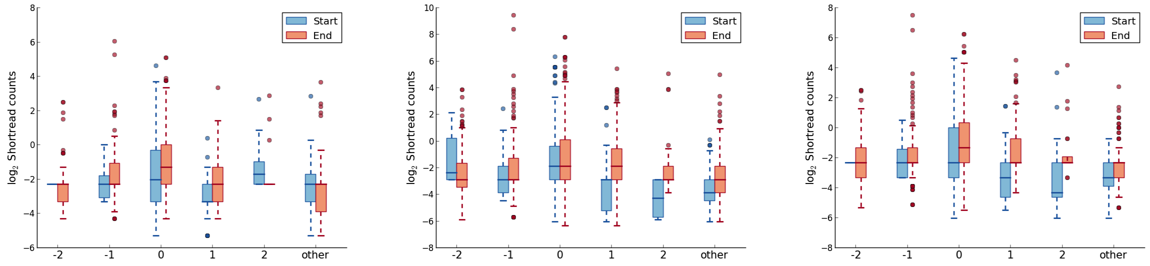
B.2 Daub alignments



(a) 5' strand alignment

(b) 3' strand alignment

Figure B.3: Total short read alignment for the Daub data set, shown for (a) the 5' strand and (b) the 3' strand.



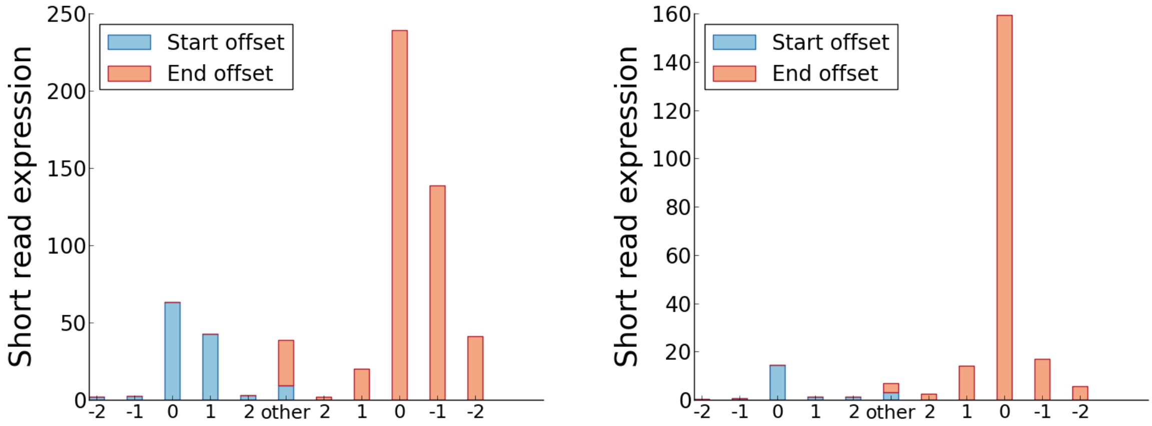
(a) Ago1 IP

(b) Ago2 IP

(c) Ago3 IP

Figure B.4: Alignment boxplots for the Daub data set, shown for (a) Ago1 IP, (b) Ago2 IP and (c) Ago3 IP.

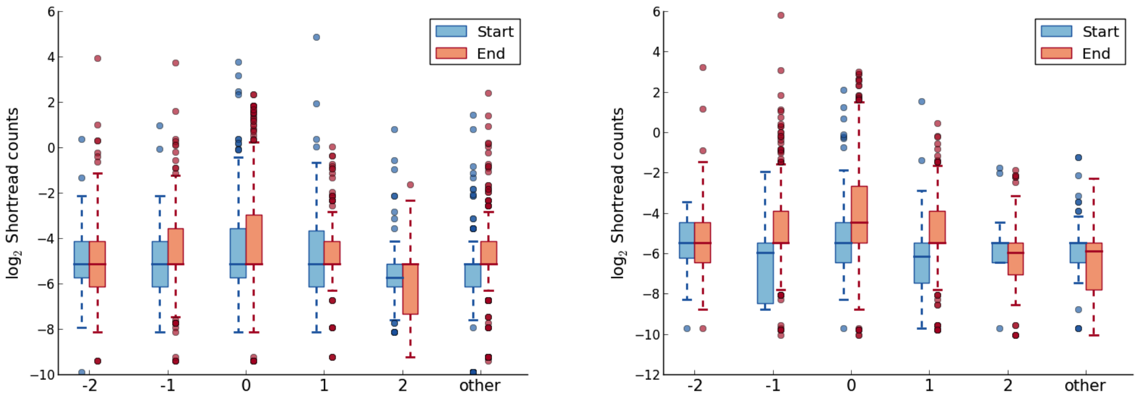
B.3 Rajewsky alignments



(a) 5' strand alignment

(b) 3' strand alignment

Figure B.5: Total short read alignment for the Rajewsky data set, shown for (a) the 5' strand and (b) the 3' strand.

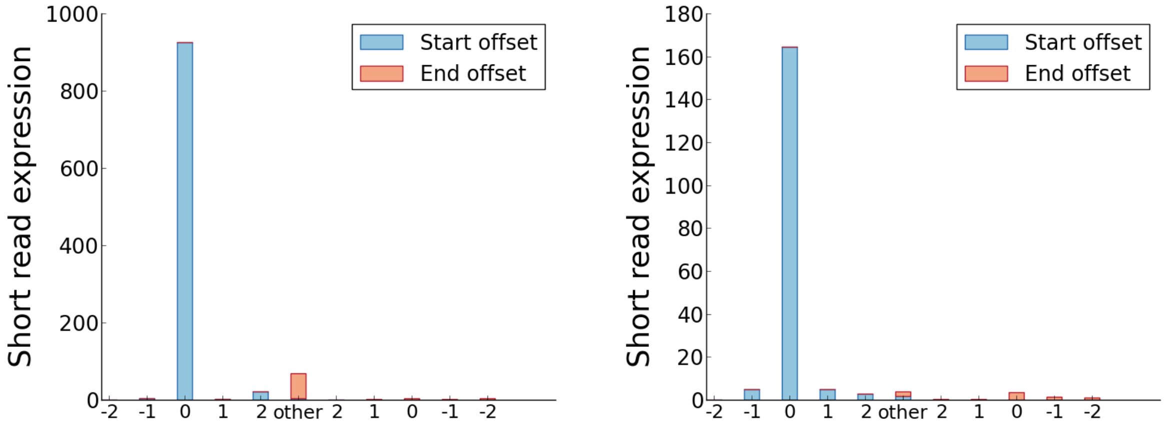


(a) Ago2 IP

(b) Ago3 IP

Figure B.6: Alignment boxplots for the Rajewsky data set, shown for (a) Ago2 IP and (b) Ago3 IP.

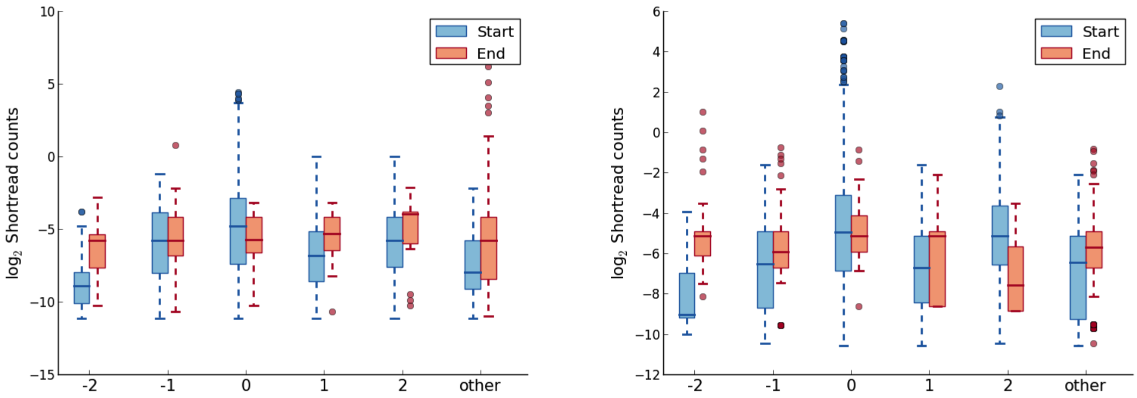
B.4 Corry alignments



(a) 5’ strand alignment

(b) 3’ strand alignment

Figure B.7: Total short read alignment for the Corry data set, shown for (a) the 5’ strand and (b) the 3’ strand.

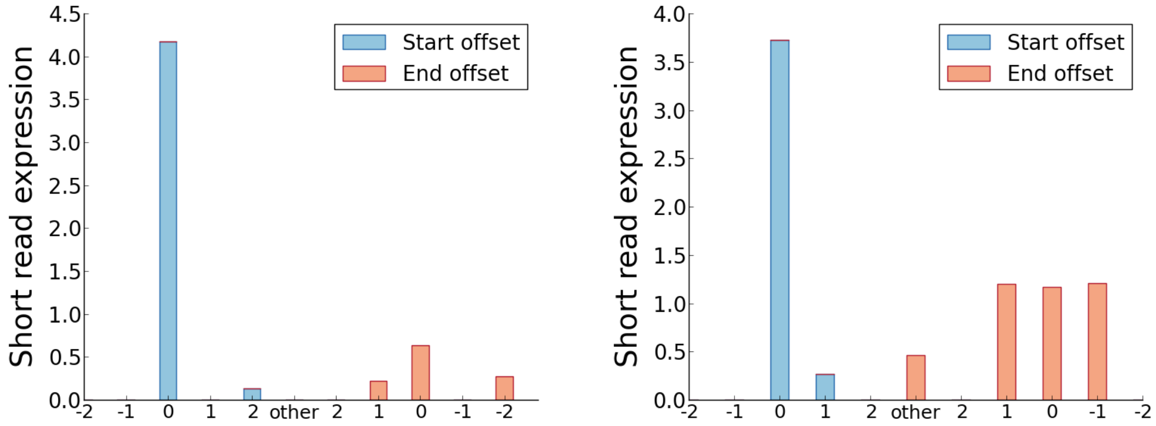


(a) Ago1 IP

(b) Ago2 IP

Figure B.8: Alignment boxplots for the Corry data set, shown for (a) the sum of Ago1 IP samples, (b) the sum of Ago2 IP samples.

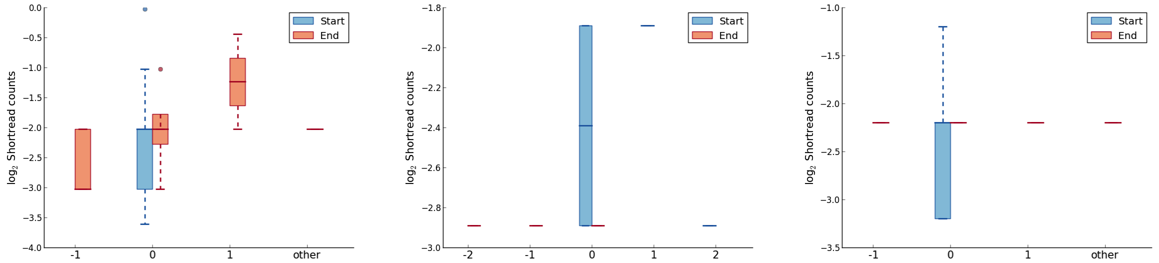
B.5 Rui alignments



(a) 5’ strand alignment

(b) 3’ strand alignment

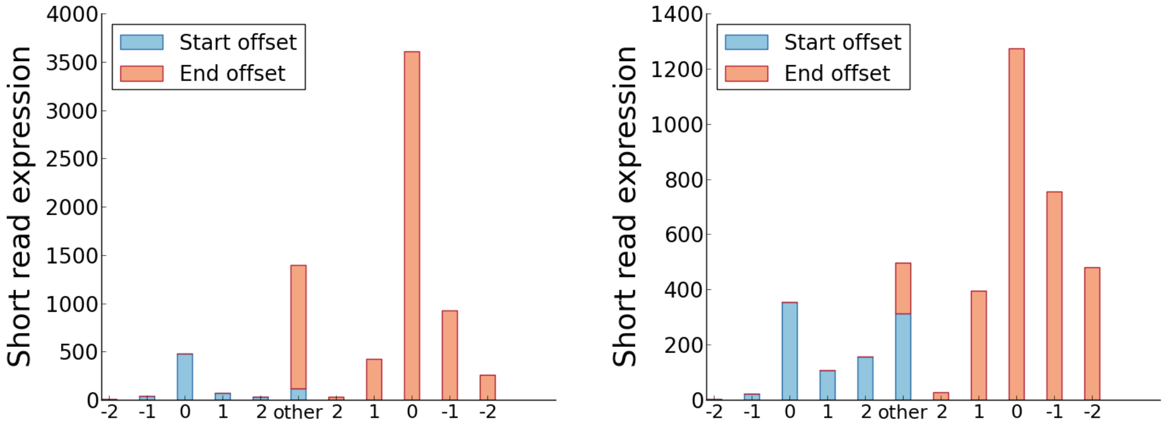
Figure B.9: Total short read alignment for the Rui data set, shown for (a) the 5’ strand and (b) the 3’ strand.



(a) Ago1 IP (b) Ago2 IP (c) Ago3 IP

Figure B.10: Alignment boxplots for the Rui data set, shown for (a) Ago1 IP, (b) Ago2 IP and (c) Ago3 IP.

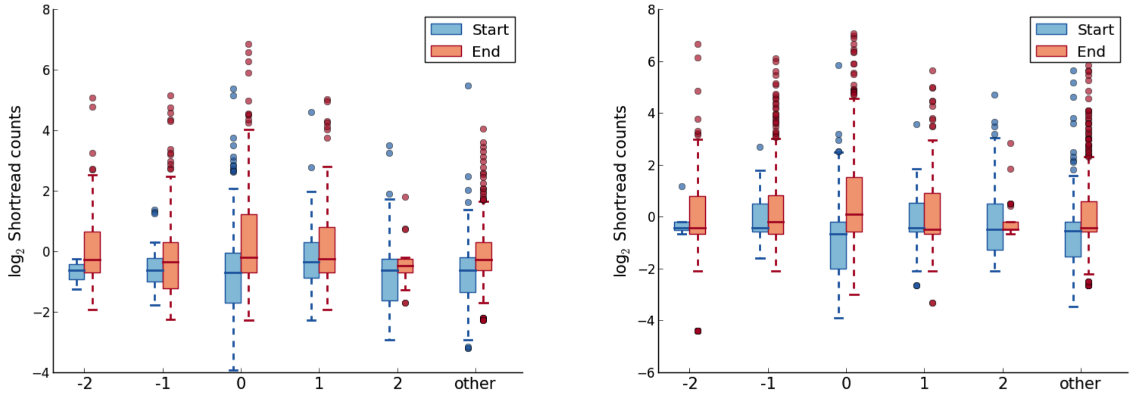
B.6 Lundbæk alignments



(a) 5’ strand alignment

(b) 3’ strand alignment

Figure B.11: Total short read alignment for the Lundbæk data set, shown for (a) the 5’ strand and (b) the 3’ strand.



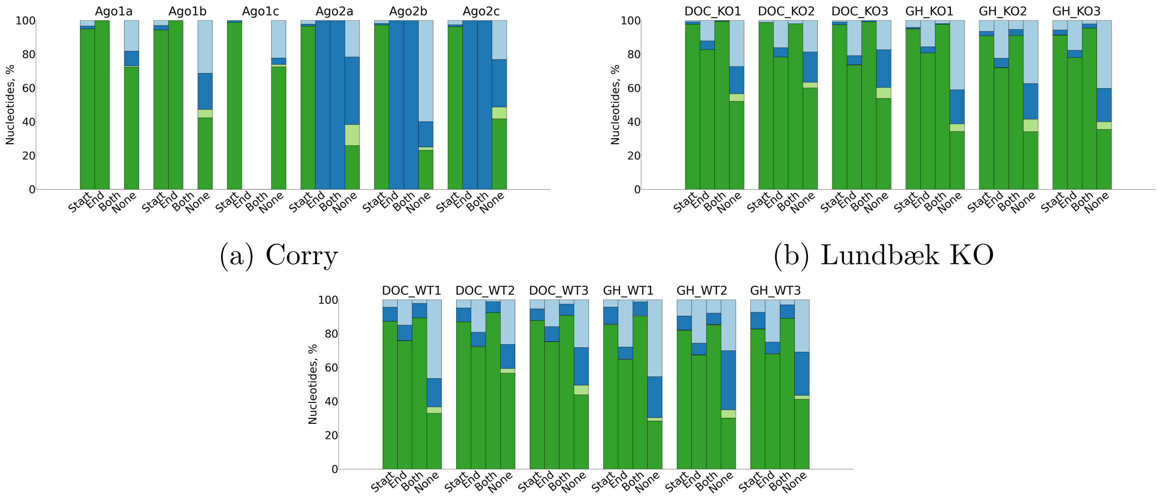
(a) Ago2 KO

(b) WT

Figure B.12: Alignment boxplots for the Lundbæk data subsets, shown for (a) Ago2 KO and (b) WT.

Appendix C

Nucleotide preferences

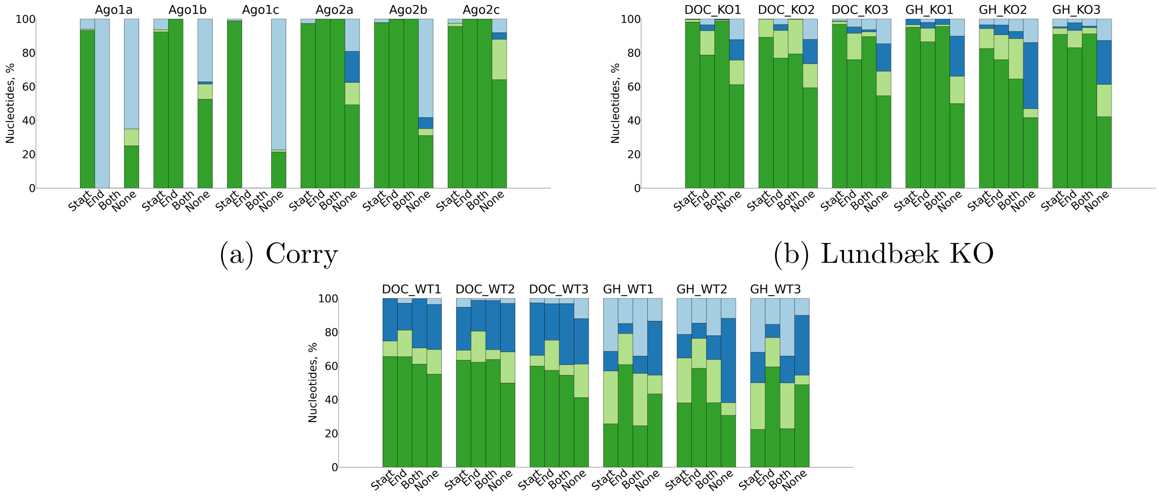


(a) Corry

(b) Lundbæk KO

(c) Lundbæk WT

Figure C.1: 5' terminal nucleotide preferences of mouse miRNAs associated with start reads, end reads, both start and end reads or neither.



(a) Corry

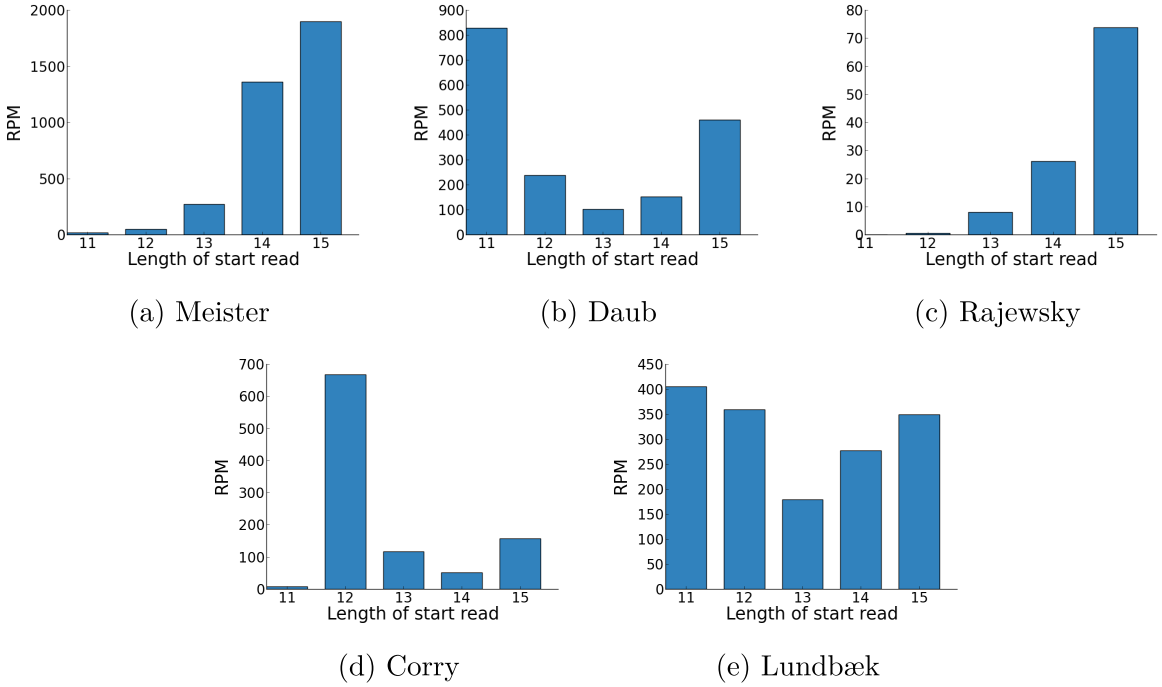
(b) Lundbæk KO

(c) Lundbæk WT

Figure C.2: 3' terminal nucleotide preferences of mouse miRNAs associated with start reads, end reads, both start and end reads or neither.

Appendix D

Short read lengths



(a) Meister

(b) Daub

(c) Rajewsky

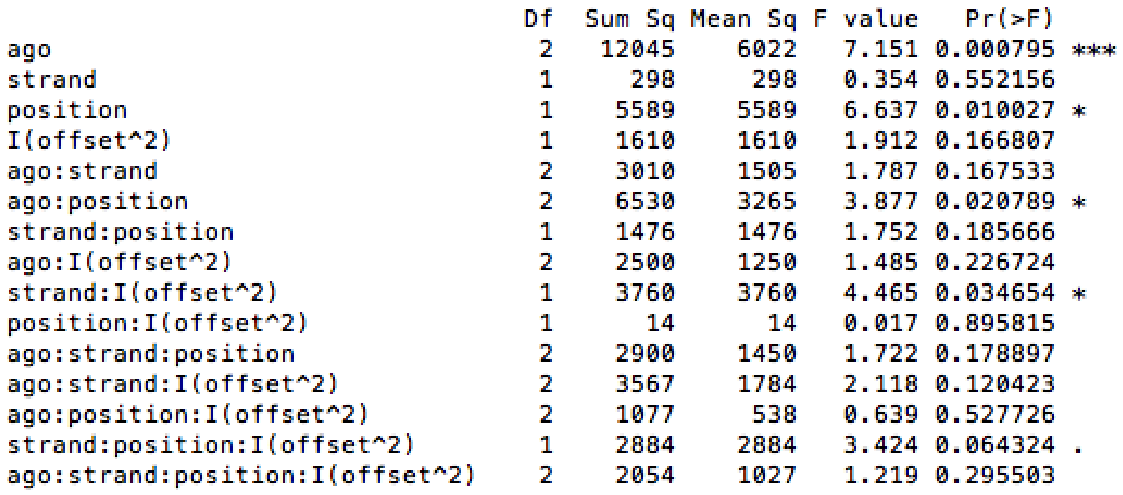
(d) Corry

(e) Lundbæk

Figure D.1: The total expression level of start reads of each length within the range 11-15 nucleotides.

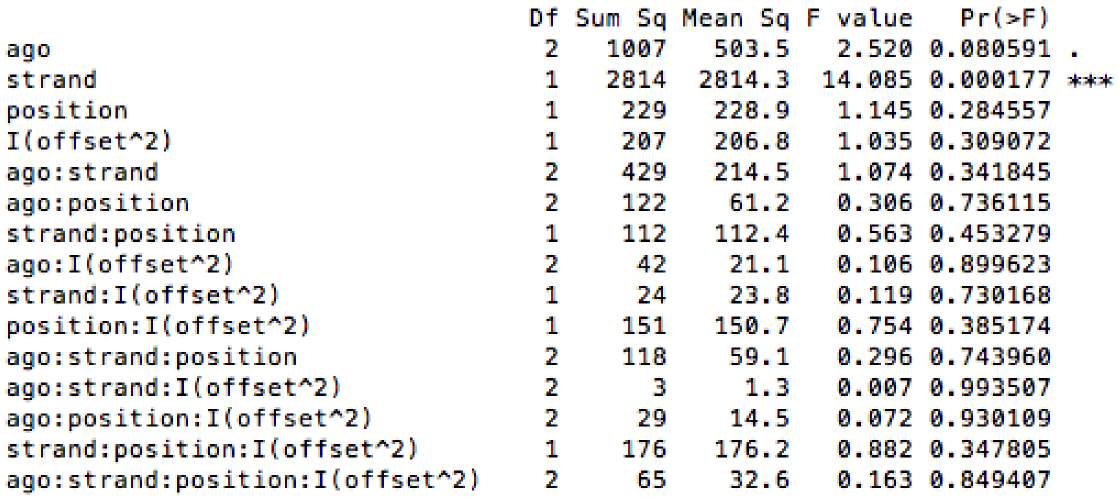
Appendix E

Anova results



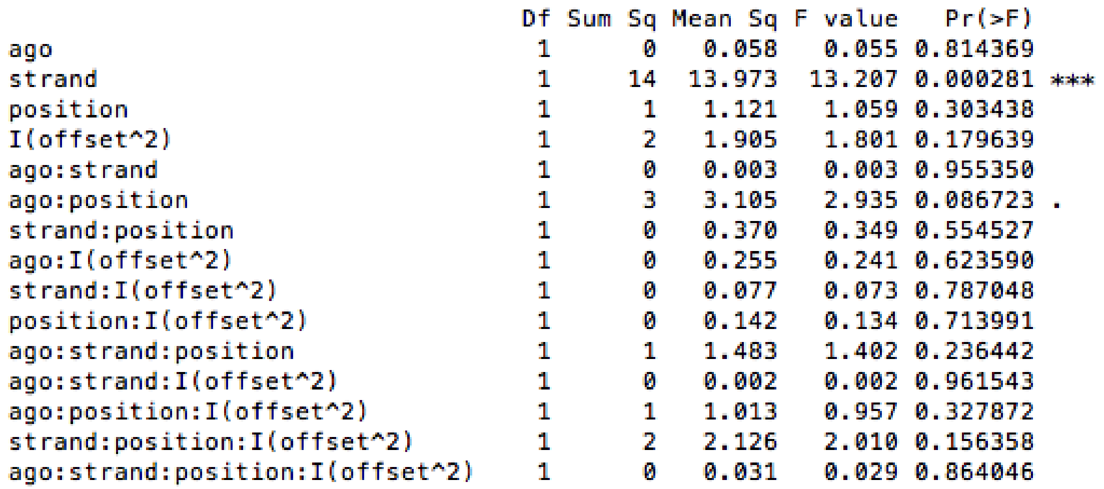
```
                                   Df  Sum Sq Mean Sq F value   Pr(>F)    
ago                                 2   12045    6022   7.151 0.000795 ***
strand                              1     298     298   0.354 0.552156    
position                            1    5589    5589   6.637 0.010027 *  
I(offset^2)                         1    1610    1610   1.912 0.166807    
ago:strand                          2    3010    1505   1.787 0.167533    
ago:position                        2    6530    3265   3.877 0.020789 *  
strand:position                     1    1476    1476   1.752 0.185666    
ago:I(offset^2)                     2    2500    1250   1.485 0.226724    
strand:I(offset^2)                  1    3760    3760   4.465 0.034654 *  
position:I(offset^2)                1      14      14   0.017 0.895815    
ago:strand:position                 2    2900    1450   1.722 0.178897    
ago:strand:I(offset^2)              2    3567    1784   2.118 0.120423    
ago:position:I(offset^2)            2    1077     538   0.639 0.527726    
strand:position:I(offset^2)         1    2884    2884   3.424 0.064324 .  
ago:strand:position:I(offset^2)     2    2054    1027   1.219 0.295503    
```

Figure E.1: Anova results for Meister



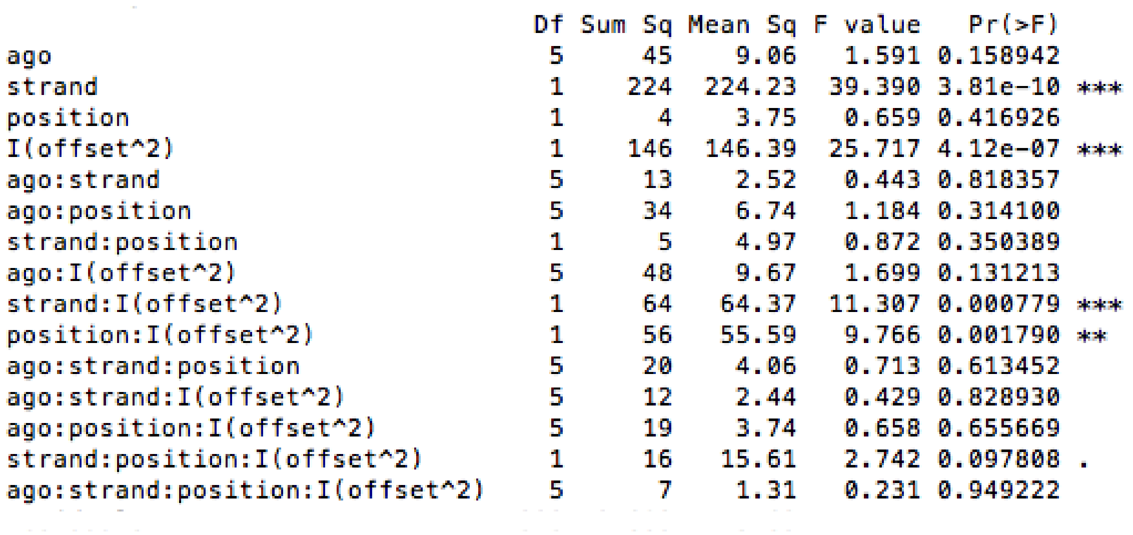
```
                                   Df Sum Sq Mean Sq F value   Pr(>F)    
ago                                 2   1007   503.5   2.520 0.080591 .  
strand                              1   2814  2814.3  14.085 0.000177 ***
position                            1    229   228.9   1.145 0.284557    
I(offset^2)                         1    207   206.8   1.035 0.309072    
ago:strand                          2    429   214.5   1.074 0.341845    
ago:position                        2    122    61.2   0.306 0.736115    
strand:position                     1    112   112.4   0.563 0.453279    
ago:I(offset^2)                     2     42    21.1   0.106 0.899623    
strand:I(offset^2)                  1     24    23.8   0.119 0.730168    
position:I(offset^2)                1    151   150.7   0.754 0.385174    
ago:strand:position                 2    118    59.1   0.296 0.743960    
ago:strand:I(offset^2)              2      3     1.3   0.007 0.993507    
ago:position:I(offset^2)            2     29    14.5   0.072 0.930109    
strand:position:I(offset^2)         1    176   176.2   0.882 0.347805    
ago:strand:position:I(offset^2)     2     65    32.6   0.163 0.849407    
```

Figure E.2: Anova results for Daub



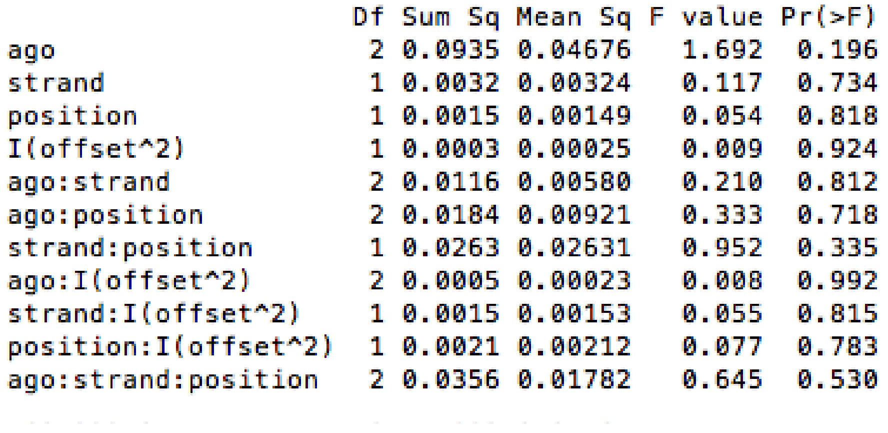
```
                            Df Sum Sq Mean Sq F value   Pr(>F)
ago                          1      0   0.058   0.055 0.814369
strand                       1     14  13.973  13.207 0.000281 ***
position                     1      1   1.121   1.059 0.303438
I(offset^2)                  1      2   1.905   1.801 0.179639
ago:strand                   1      0   0.003   0.003 0.955350
ago:position                 1      3   3.105   2.935 0.086723 .
strand:position              1      0   0.370   0.349 0.554527
ago:I(offset^2)              1      0   0.255   0.241 0.623590
strand:I(offset^2)           1      0   0.077   0.073 0.787048
position:I(offset^2)         1      0   0.142   0.134 0.713991
ago:strand:position          1      1   1.483   1.402 0.236442
ago:strand:I(offset^2)       1      0   0.002   0.002 0.961543
ago:position:I(offset^2)     1      1   1.013   0.957 0.327872
strand:position:I(offset^2)  1      2   2.126   2.010 0.156358
ago:strand:position:I(offset^2)  1  0   0.031   0.029 0.864046
```

Figure E.3: Anova results for Rajewsky



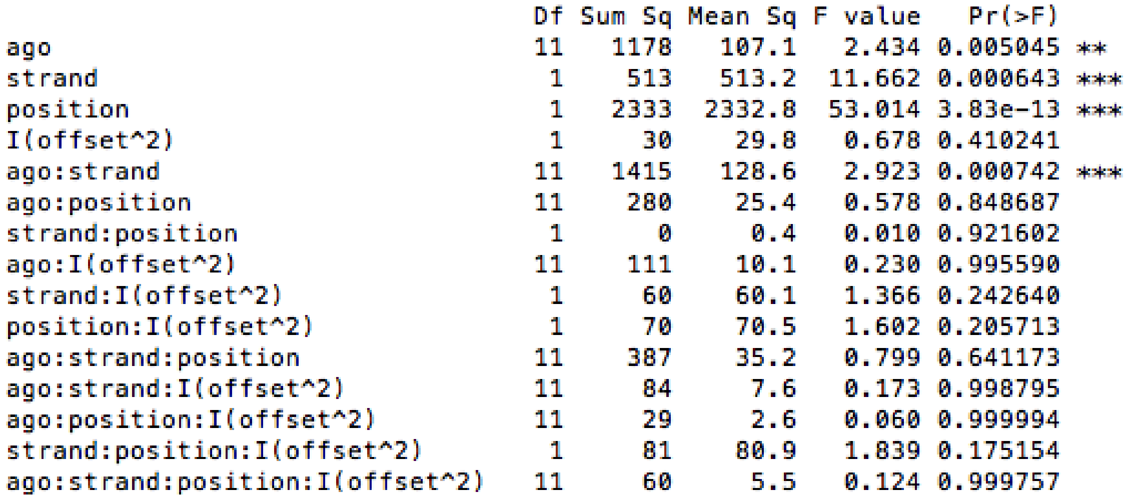
```
                                Df Sum Sq Mean Sq F value   Pr(>F)
ago                              5     45    9.06   1.591 0.158942
strand                           1    224  224.23  39.390 3.81e-10 ***
position                         1      4    3.75   0.659 0.416926
I(offset^2)                      1    146  146.39  25.717 4.12e-07 ***
ago:strand                       5     13    2.52   0.443 0.818357
ago:position                     5     34    6.74   1.184 0.314100
strand:position                  1      5    4.97   0.872 0.350389
ago:I(offset^2)                  5     48    9.67   1.699 0.131213
strand:I(offset^2)               1     64   64.37  11.307 0.000779 ***
position:I(offset^2)             1     56   55.59   9.766 0.001790 **
ago:strand:position              5     20    4.06   0.713 0.613452
ago:strand:I(offset^2)           5     12    2.44   0.429 0.828930
ago:position:I(offset^2)         5     19    3.74   0.658 0.655669
strand:position:I(offset^2)      1     16   15.61   2.742 0.097808 .
ago:strand:position:I(offset^2)  5      7    1.31   0.231 0.949222
```

Figure E.4: Anova results for Corry



```
                     Df Sum Sq Mean Sq F value Pr(>F)
ago                   2 0.0935 0.04676   1.692  0.196
strand                1 0.0032 0.00324   0.117  0.734
position              1 0.0015 0.00149   0.054  0.818
I(offset^2)           1 0.0003 0.00025   0.009  0.924
ago:strand            2 0.0116 0.00580   0.210  0.812
ago:position          2 0.0184 0.00921   0.333  0.718
strand:position       1 0.0263 0.02631   0.952  0.335
ago:I(offset^2)       2 0.0005 0.00023   0.008  0.992
strand:I(offset^2)    1 0.0015 0.00153   0.055  0.815
position:I(offset^2)  1 0.0021 0.00212   0.077  0.783
ago:strand:position   2 0.0356 0.01782   0.645  0.530
```

Figure E.5: Anova results for Rui



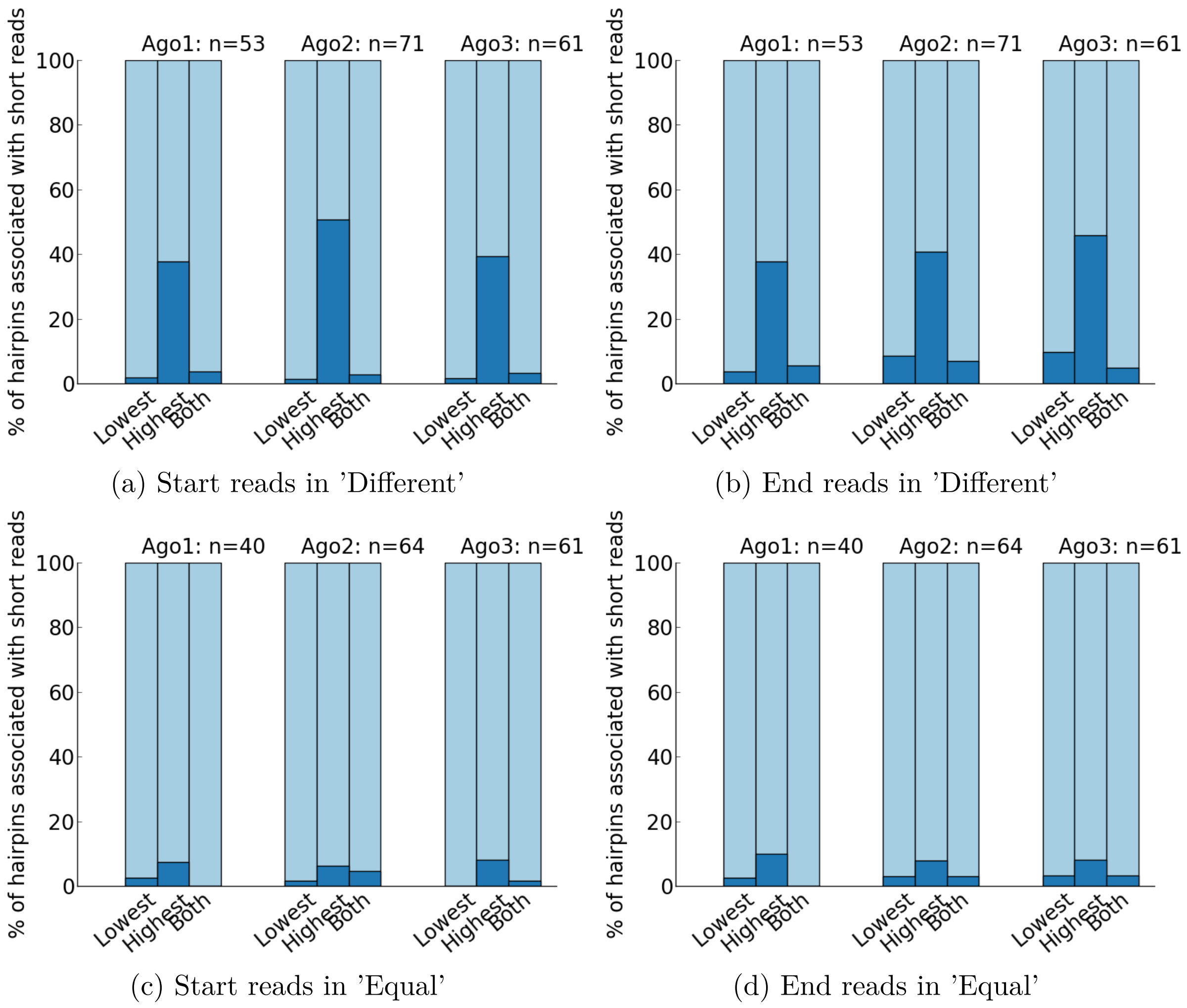
```
                                 Df Sum Sq Mean Sq F value   Pr(>F)    
ago                              11   1178   107.1   2.434 0.005045 ** 
strand                            1    513   513.2  11.662 0.000643 ***
position                          1   2333  2332.8  53.014 3.83e-13 ***
I(offset^2)                       1     30    29.8   0.678 0.410241    
ago:strand                       11   1415   128.6   2.923 0.000742 ***
ago:position                     11    280    25.4   0.578 0.848687    
strand:position                   1      0     0.4   0.010 0.921602    
ago:I(offset^2)                  11    111    10.1   0.230 0.995590    
strand:I(offset^2)                1     60    60.1   1.366 0.242640    
position:I(offset^2)              1     70    70.5   1.602 0.205713    
ago:strand:position              11    387    35.2   0.799 0.641173    
ago:strand:I(offset^2)           11     84     7.6   0.173 0.998795    
ago:position:I(offset^2)         11     29     2.6   0.060 0.999994    
strand:position:I(offset^2)       1     81    80.9   1.839 0.175154    
ago:strand:position:I(offset^2)  11     60     5.5   0.124 0.999757    
```

Figure E.6: Anova results for Lundbæk

Appendix F

Hairpin classification

F.1 SR association of hairpins



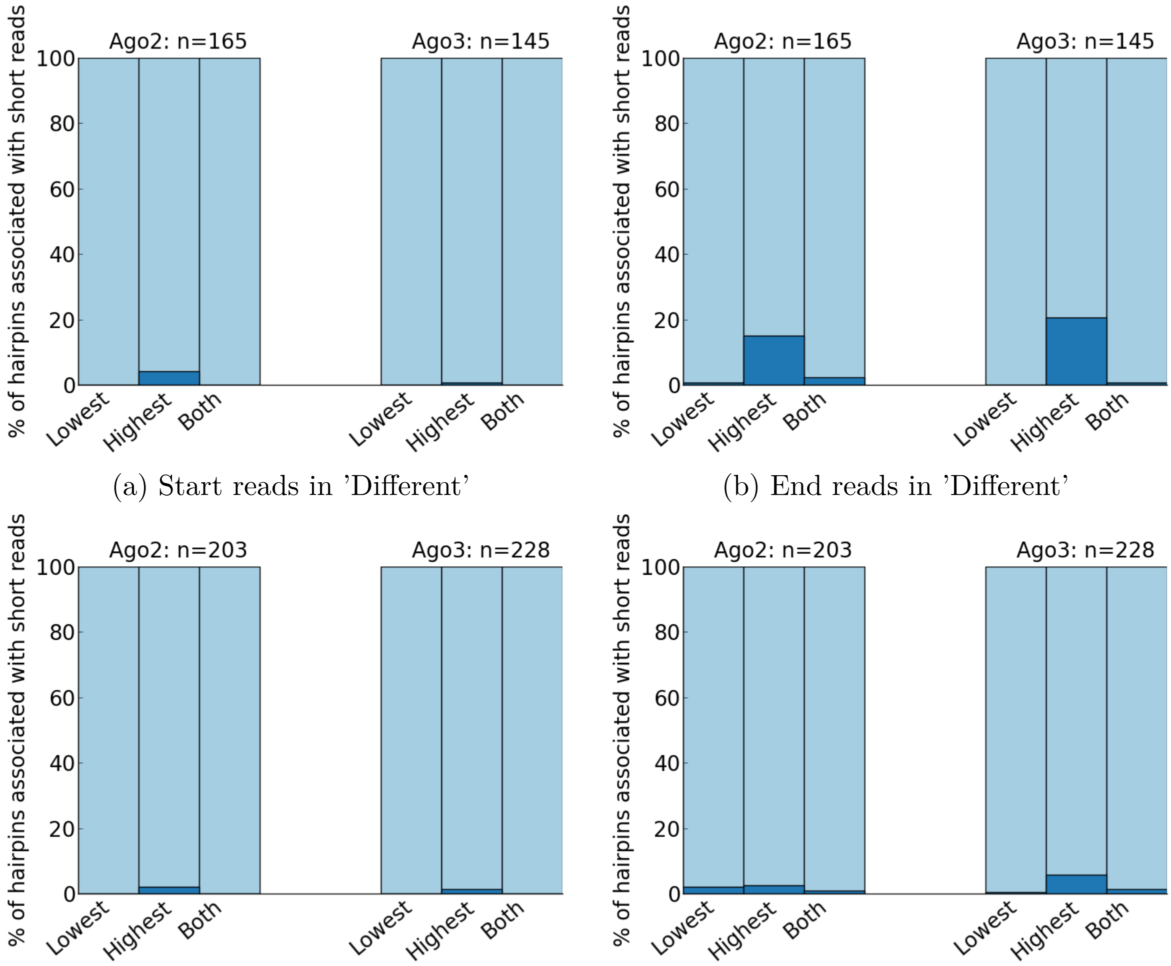
(a) Start reads in 'Different'

(b) End reads in 'Different'

(c) Start reads in 'Equal'

(d) End reads in 'Equal'

Figure F.1: Short read association of miRNAs of classes 'Different' and 'Equal' from the *Daub* data set.



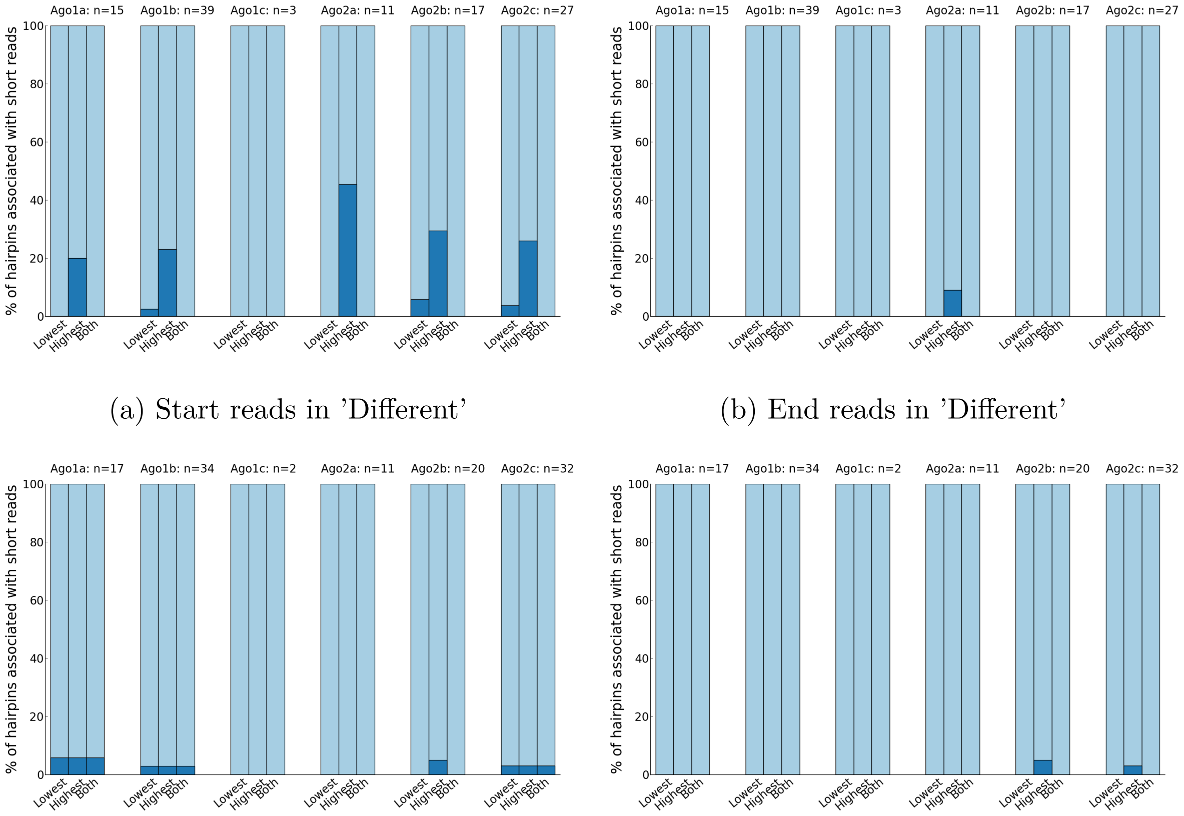
(a) Start reads in 'Different'

(b) End reads in 'Different'

(c) Start reads in 'Equal'

(d) End reads in 'Equal'

Figure F.2: Short read association of miRNAs of classes 'Different' and 'Equal' from the *Rajewsky* data set.



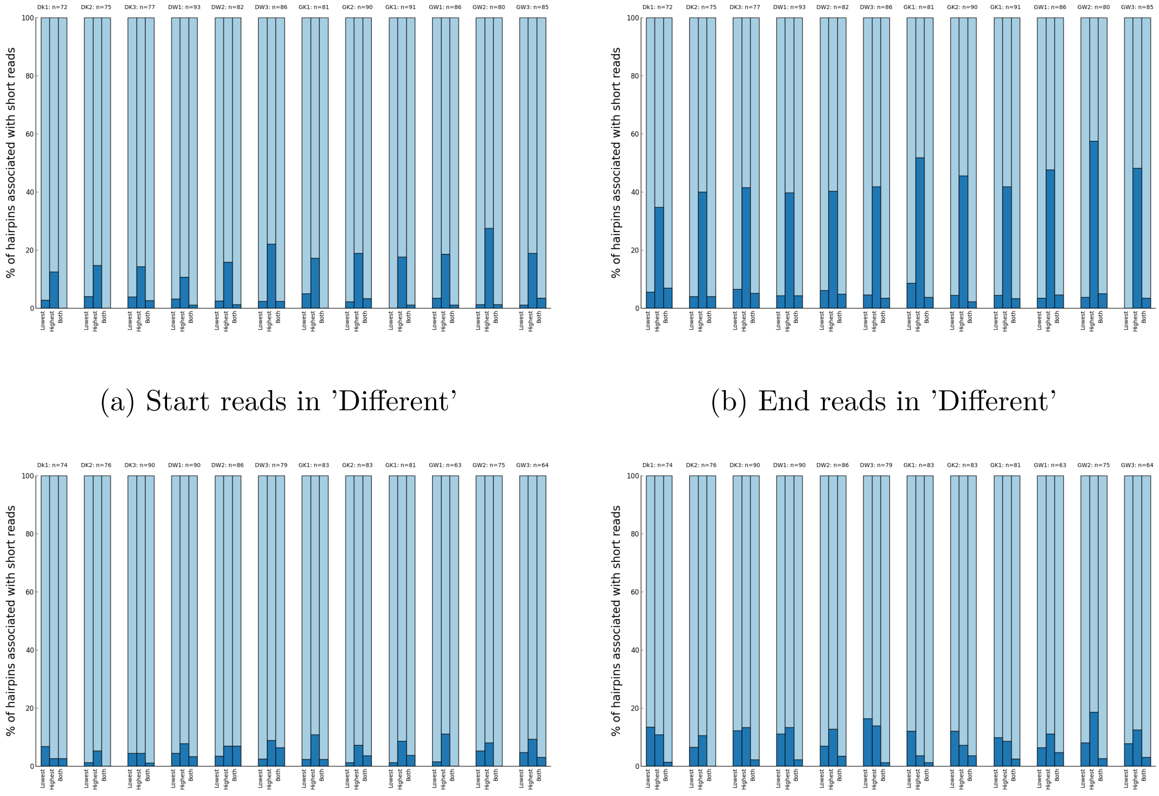
(a) Start reads in 'Different'

(b) End reads in 'Different'

(c) Start reads in 'Equal'

(d) End reads in 'Equal'

Figure F.3: Short read association of miRNAs of classes 'Different' and 'Equal' from the *Corry* data set.



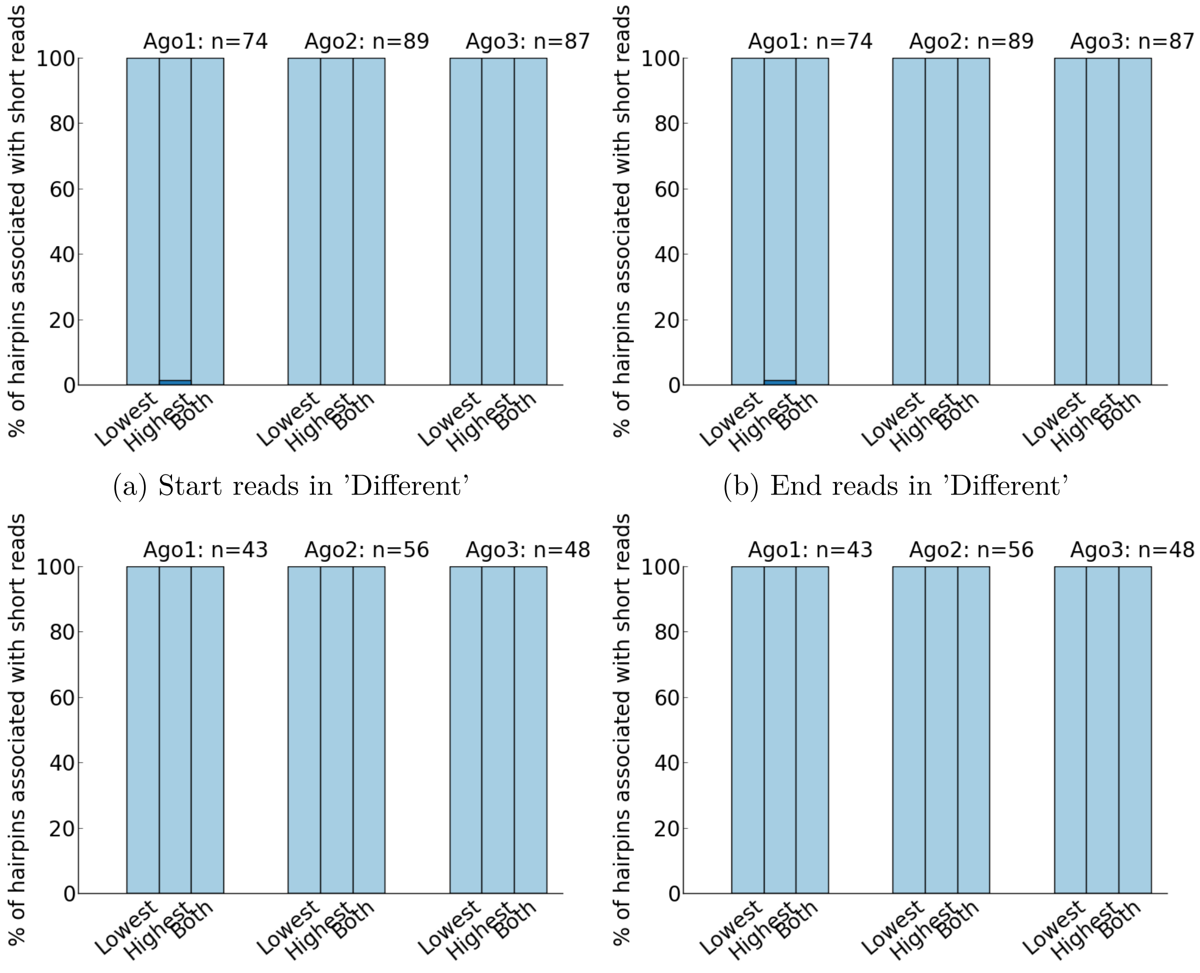
(a) Start reads in 'Different'

(b) End reads in 'Different'

(c) Start reads in 'Equal'

(d) End reads in 'Equal'

Figure F.4: Short read association of miRNAs of classes 'Different' and 'Equal' from the *Lundbæk* data set.



(a) Start reads in 'Different'

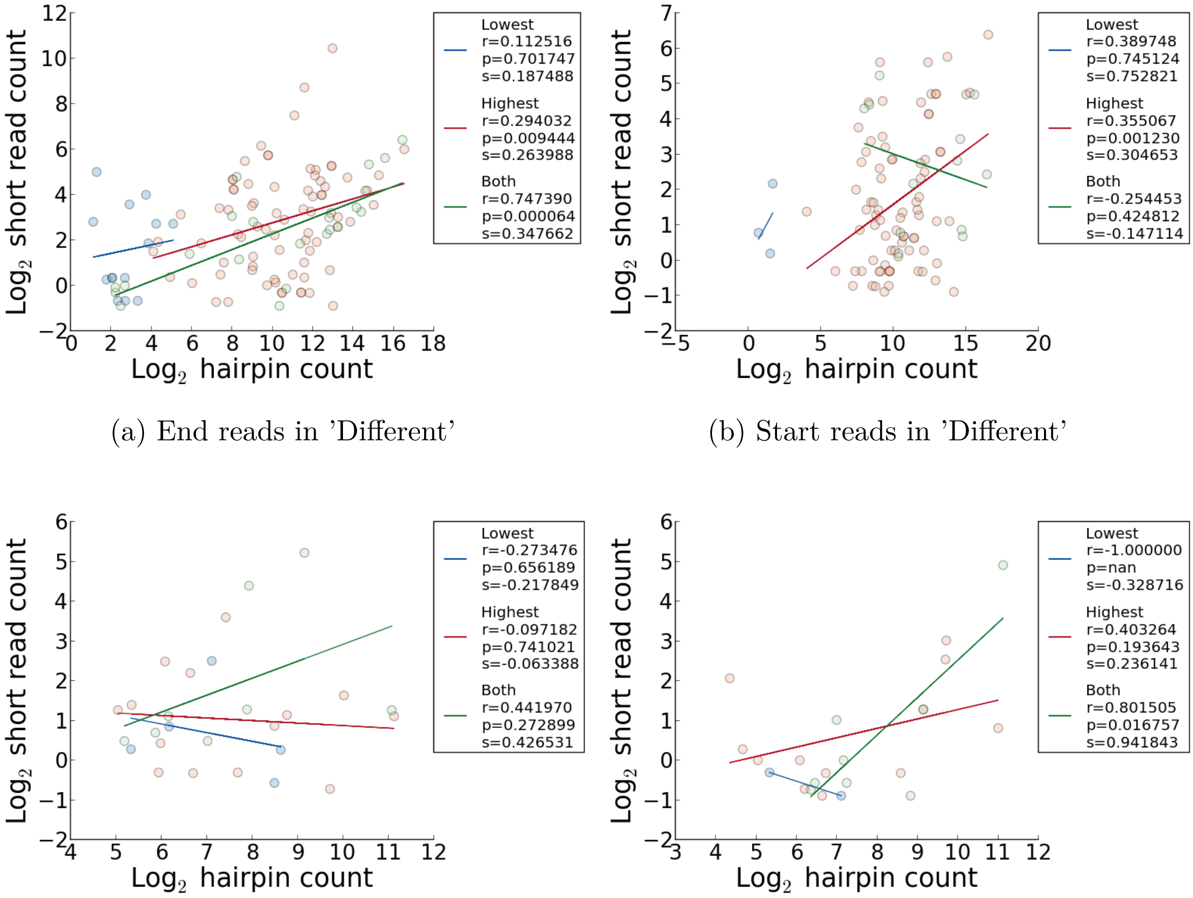
(b) End reads in 'Different'

(c) Start reads in 'Equal'

(d) End reads in 'Equal'

Figure F.5: Short read association of miRNAs of classes 'Different' and 'Equal' from the *Rui* data set.

F.2 Hairpin and SR correlation



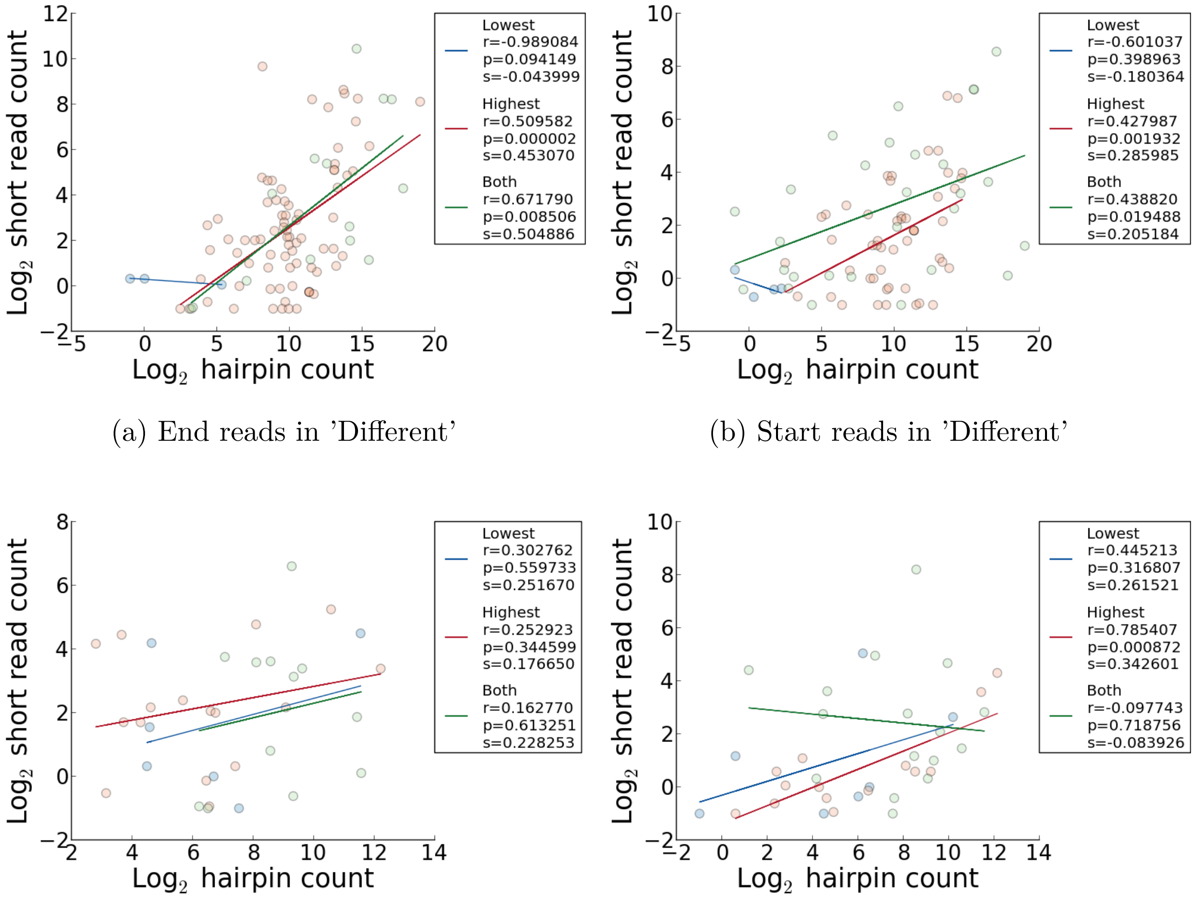
(a) End reads in 'Different'

(b) Start reads in 'Different'

(c) End reads in 'Equal'

(d) Start reads in 'Equal'

Figure F.6: Correlation of hairpin and short read expression for the *Daub* data set.



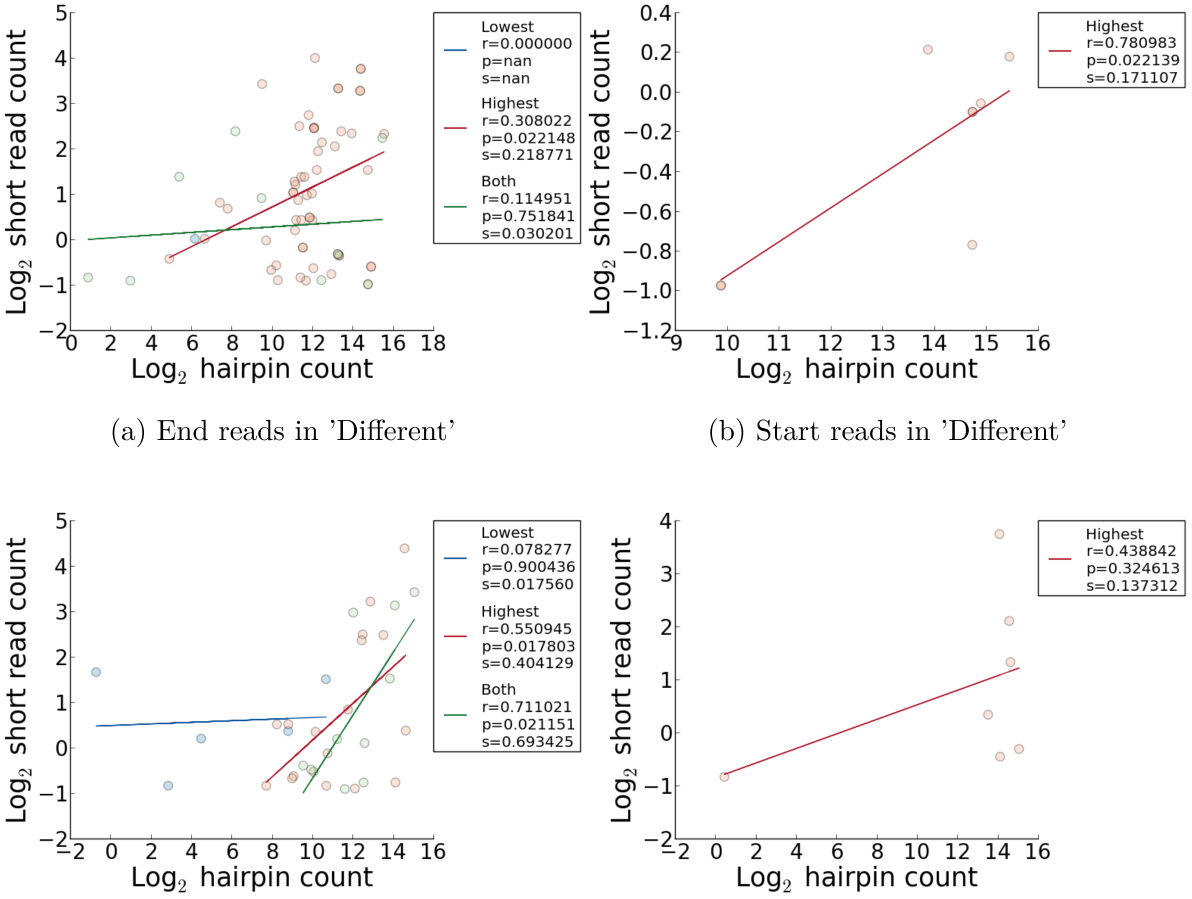
(a) End reads in 'Different'

(b) Start reads in 'Different'

(c) End reads in 'Equal'

(d) Start reads in 'Equal'

Figure F.7: Correlation of hairpin and short read expression for the *Meister* data set.



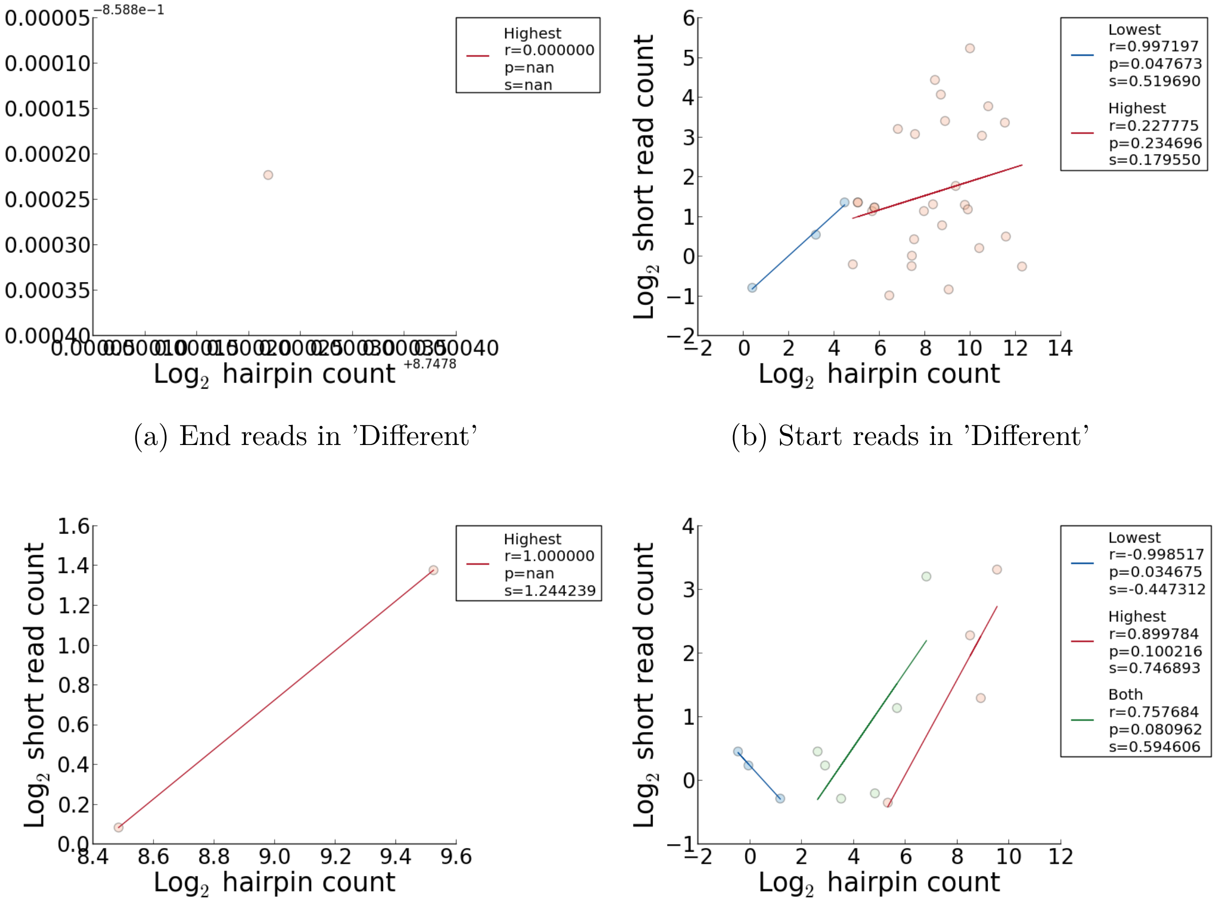
(a) End reads in 'Different'

(b) Start reads in 'Different'

(c) End reads in 'Equal'

(d) Start reads in 'Equal'

Figure F.8: Correlation of hairpin and short read expression for the *Rajewsky* data set.



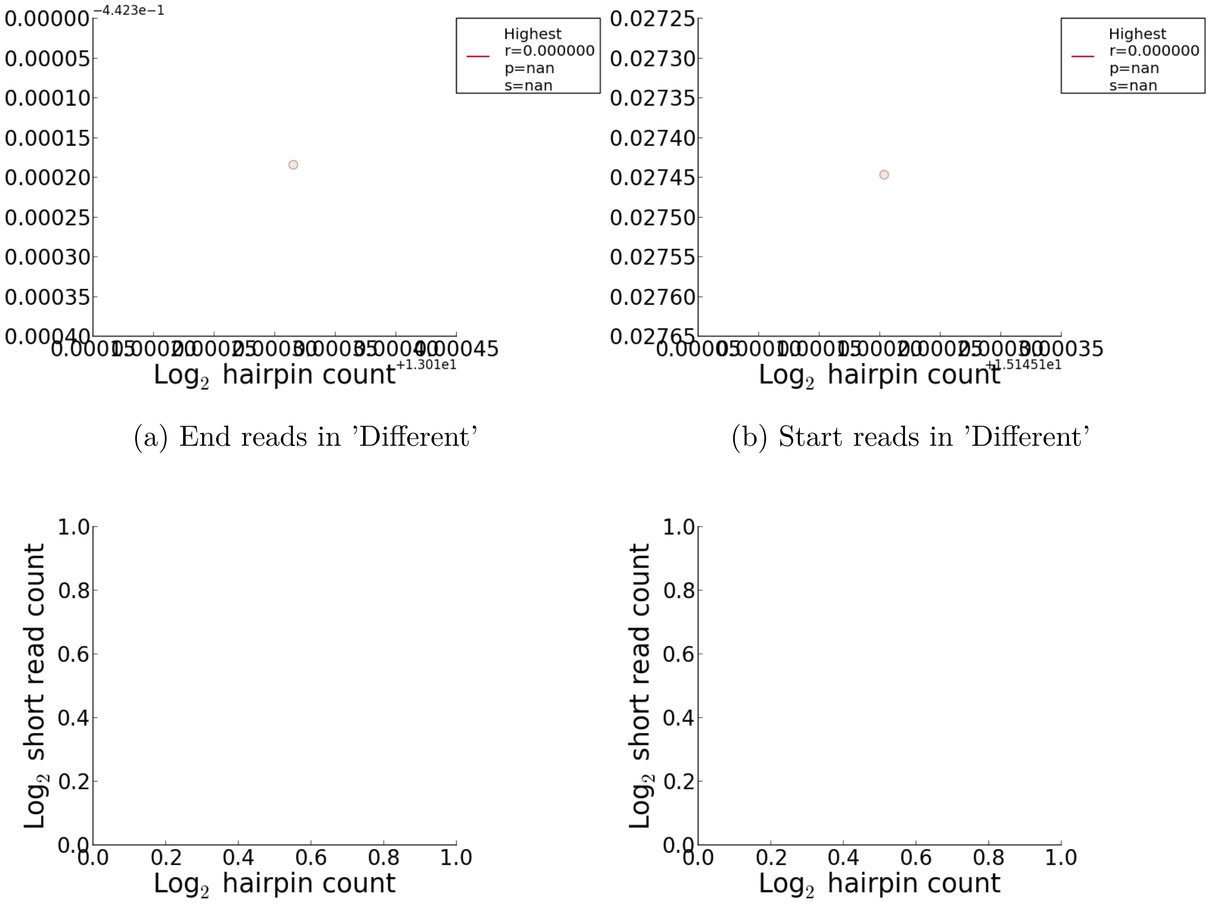
(a) End reads in 'Different'

(b) Start reads in 'Different'

(c) End reads in 'Equal'

(d) Start reads in 'Equal'

Figure F.9: Correlation of hairpin and short read expression for the *Corry* data set.



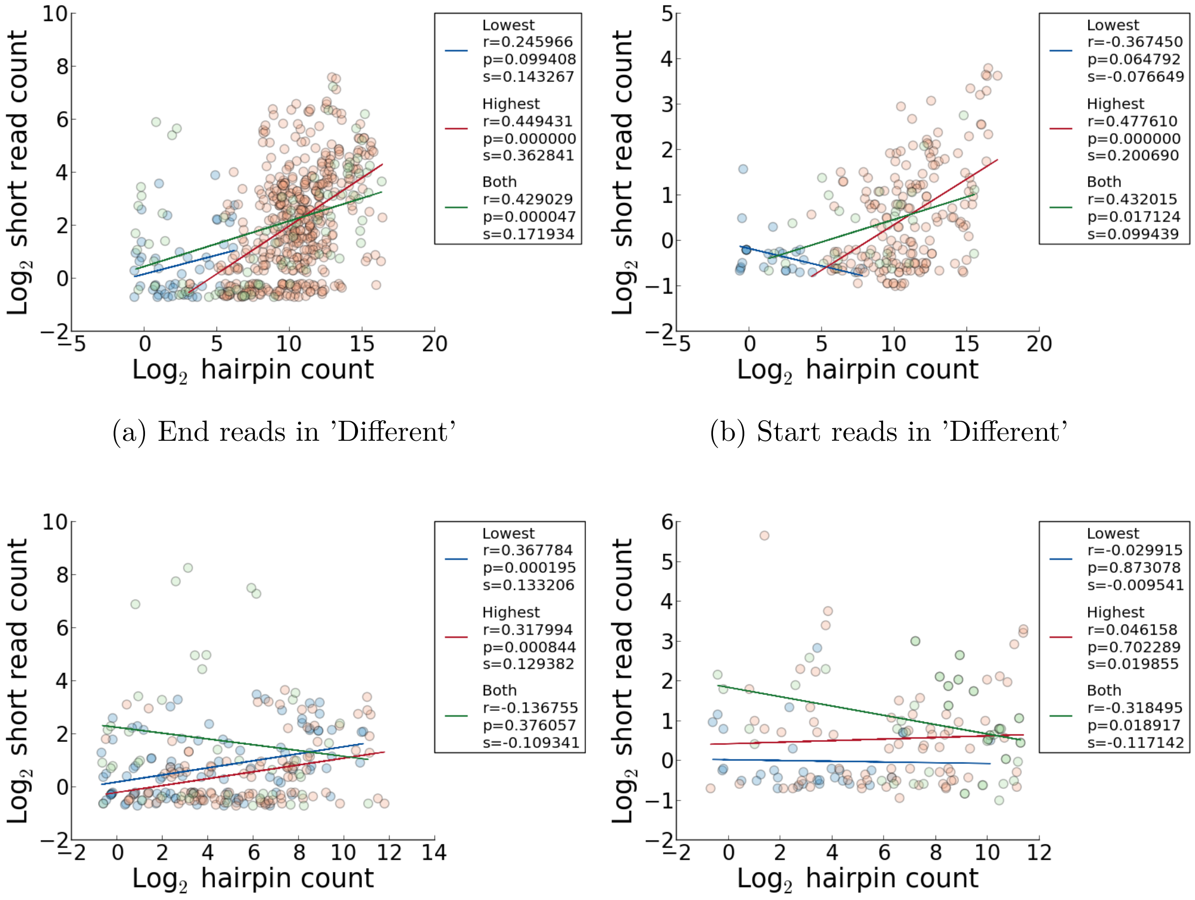
(a) End reads in 'Different'

(b) Start reads in 'Different'

(c) End reads in 'Equal'

(d) Start reads in 'Equal'

Figure F.10: Correlation of hairpin and short read expression for the *Rui* data set.



(a) End reads in 'Different'

(b) Start reads in 'Different'

(c) End reads in 'Equal'

(d) Start reads in 'Equal'

Figure F.11: Correlation of hairpin and short read expression for the *Lundbæk* data set.

Table F.1: Correlation coefficients (r), regression line slopes (s), *p-values* and n for the correlation between the expression of short reads aligned to the guide strand and hairpins of (a) the 'Different' class and (b) the 'Equal' class for all six sample data sets. The p-value represents the null hypothesis that the slope is zero.

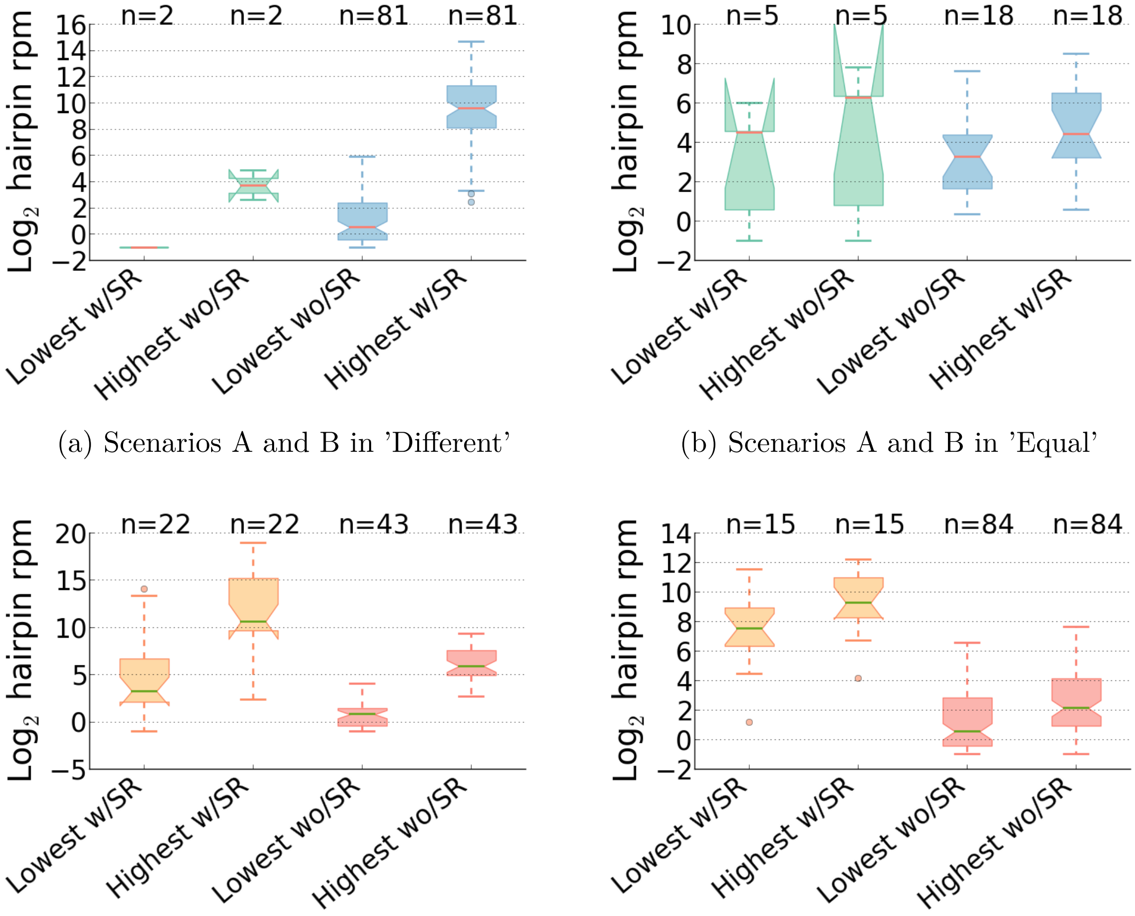
	End reads				Start reads			
Data set	r	s	p	n	r	s	p	n
Daub	0.294	0.264	0.009	77	0.355	0.305	0.001	80
Meister	0.510	0.453	0.000	77	0.428	0.286	0.002	50
Rajewsky	0.308	0.219	0.022	55	0.781	0.171	0.022	8
Corry	0.000	nan	nan	1	0.228	0.180	0.235	29
Lundbæk	0.449	0.363	0.000	442	0.478	0.201	0.000	174
Rui	0.000	nan	nan	1	0.000	nan	nan	1

(a) 'Different' class

	End reads				Start reads			
Data set	r	s	p	n	r	s	p	n
Daub	-0.097	-0.063	0.741	14	0.403	0.236	0.194	12
Meister	0.253	0.177	0.345	16	0.785	0.343	0.001	14
Rajewsky	0.551	0.404	0.018	18	0.439	0.137	0.325	7
Corry	1.000	1.244	nan	2	0.900	0.747	0.100	4
Lundbæk	0.318	0.129	0.001	107	0.046	0.020	0.702	71
Rui	nan	nan	nan	0	nan	nan	nan	0

(b) 'Equal' class

F.3 Alignment scenarios



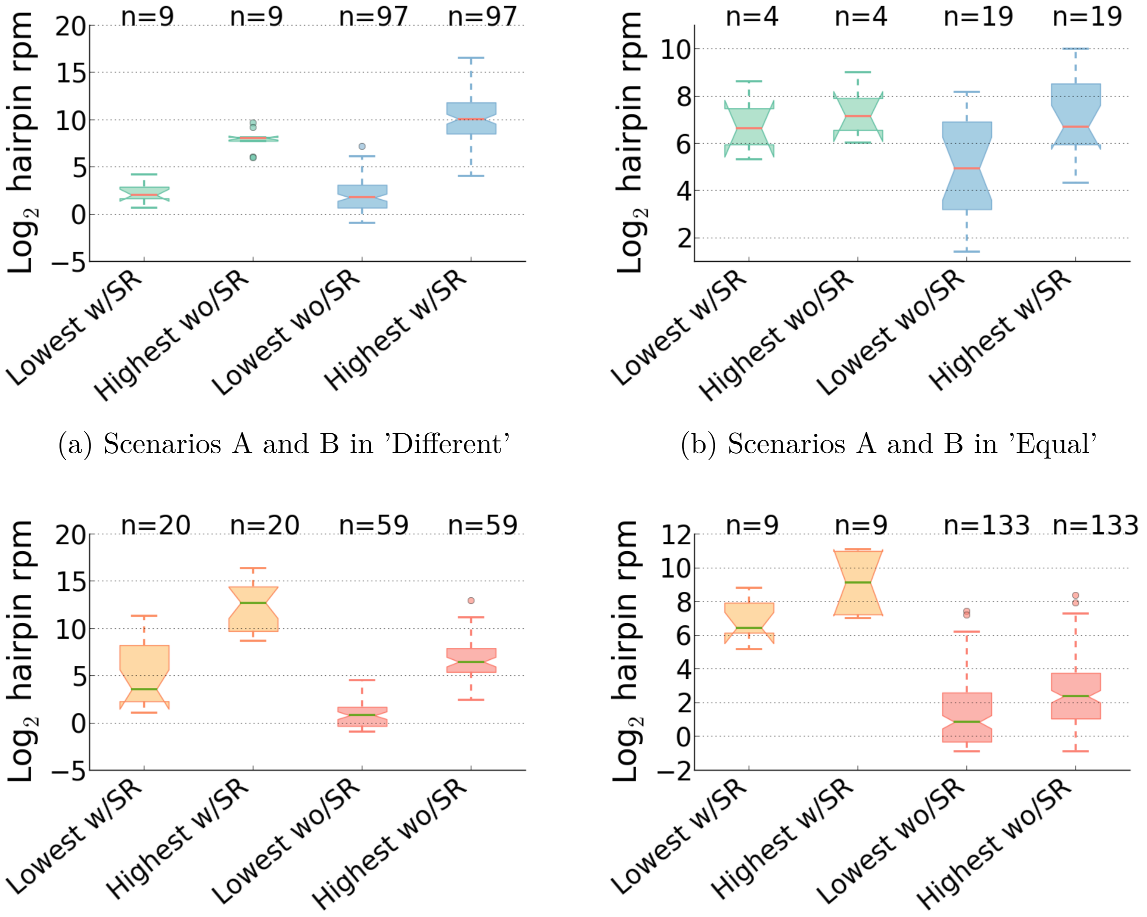
(a) Scenarios A and B in ’Different’

(b) Scenarios A and B in ’Equal’

(c) Scenarios C and D in ’Different’

(d) Scenarios C and D in ’Equal’

Figure F.12: The expression levels of each hairpin strand of scenario A (green), B (blue), C (orange) and D (red) for the *Meister* data set in the ’Different’ and ’Equal’ class.



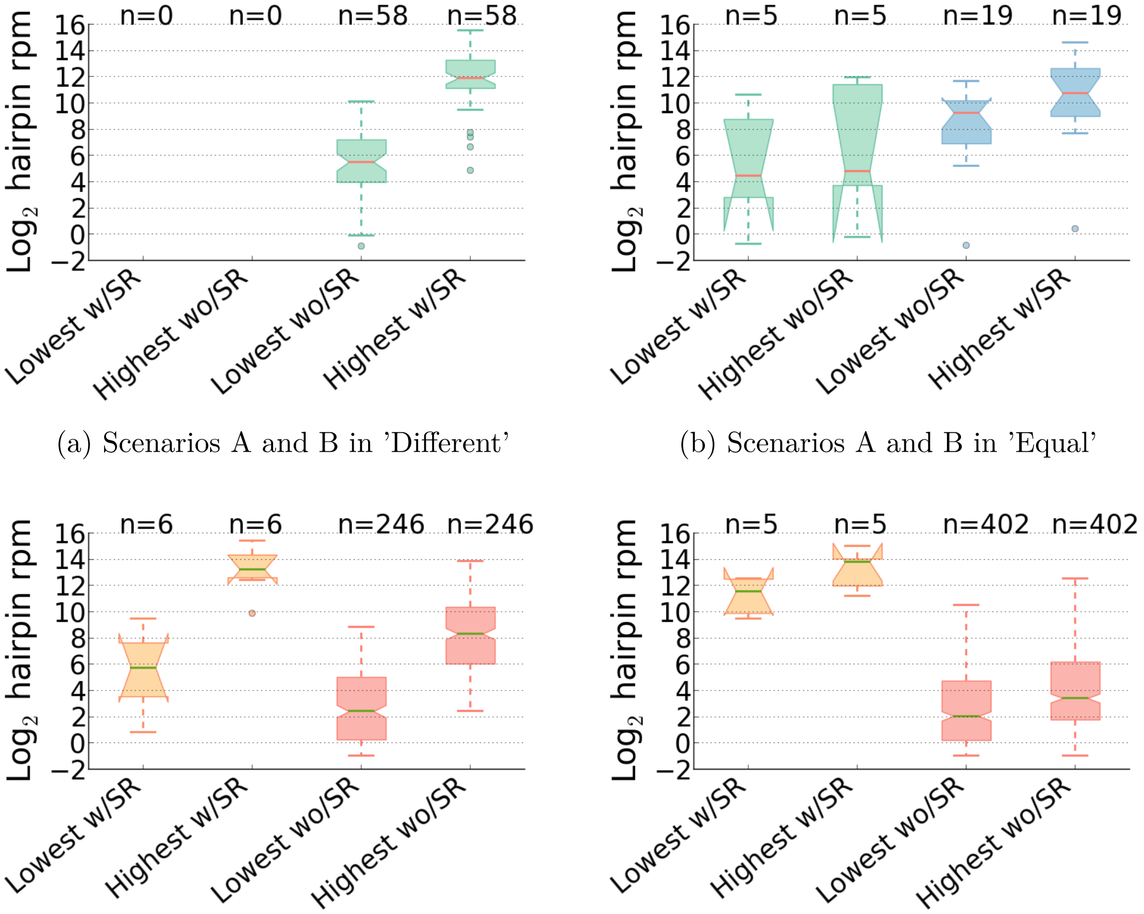
(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal'

Figure F.13: The expression levels of each hairpin strand of scenario A (green), B (blue), C (orange) and D (red) for the *Daub* data set in the 'Different' and 'Equal' class.



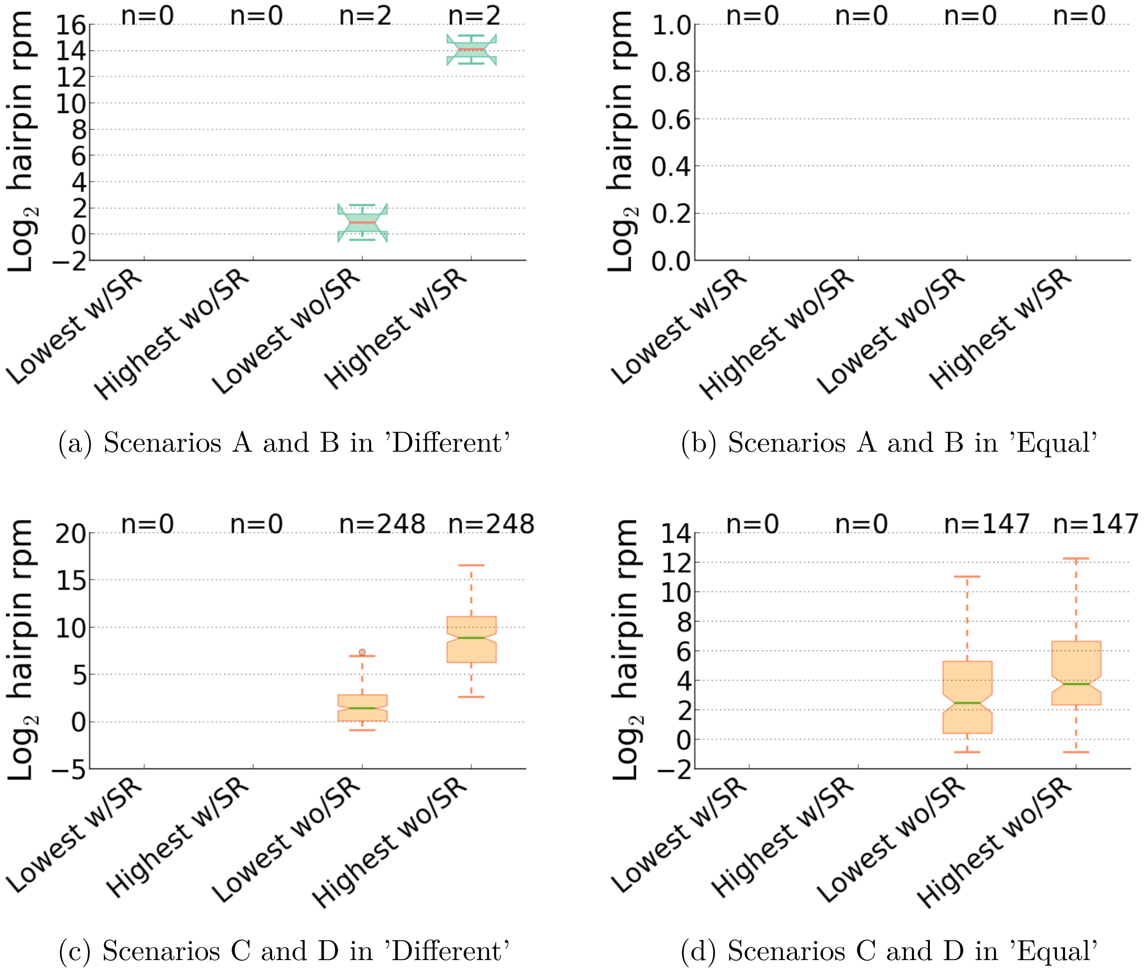
(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal'

Figure F.14: The expression levels of each hairpin strand of scenario A (green), B (blue), C (orange) and D (red) for the *Rajewsky* data set in the 'Different' and 'Equal' class.



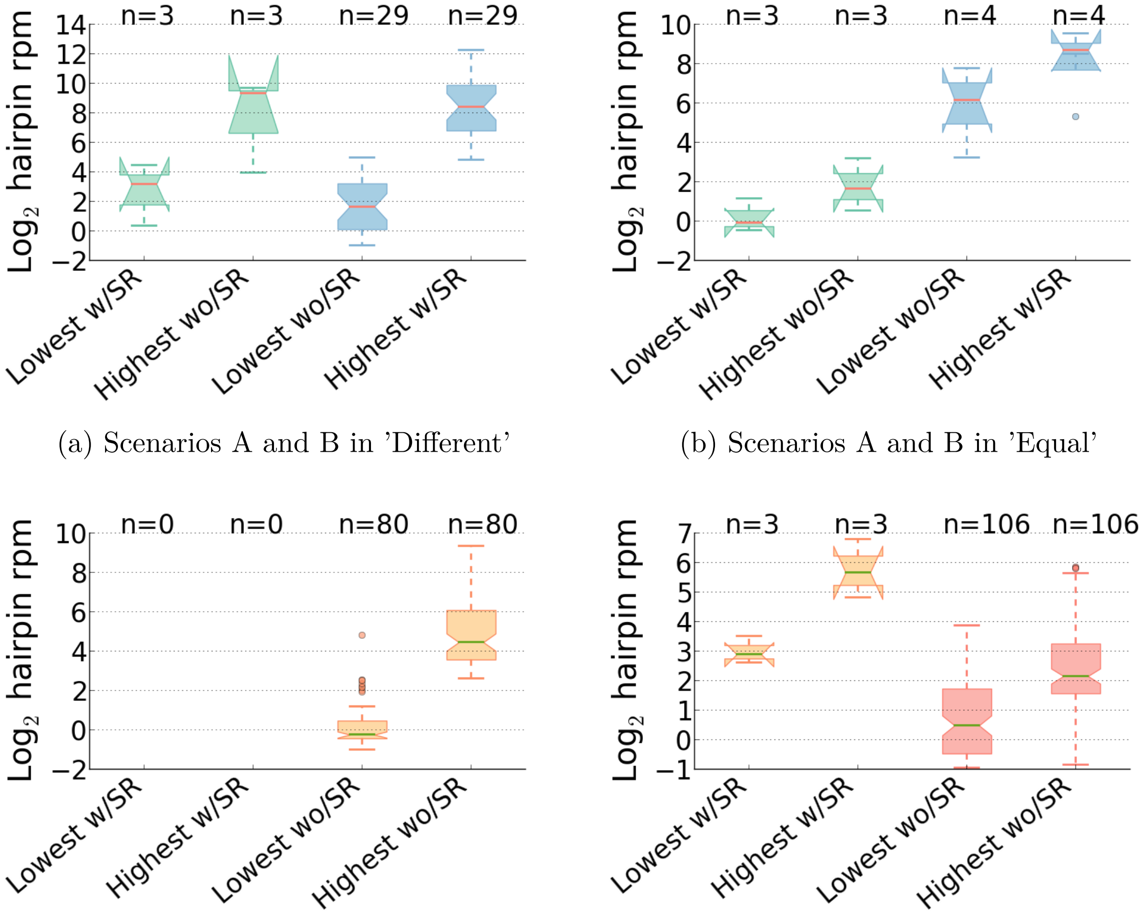
(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal'

Figure F.15: The expression levels of each hairpin strand of scenario A (green), B (blue), C (orange) and D (red) for the *Rui* data set in the 'Different' and 'Equal' class.



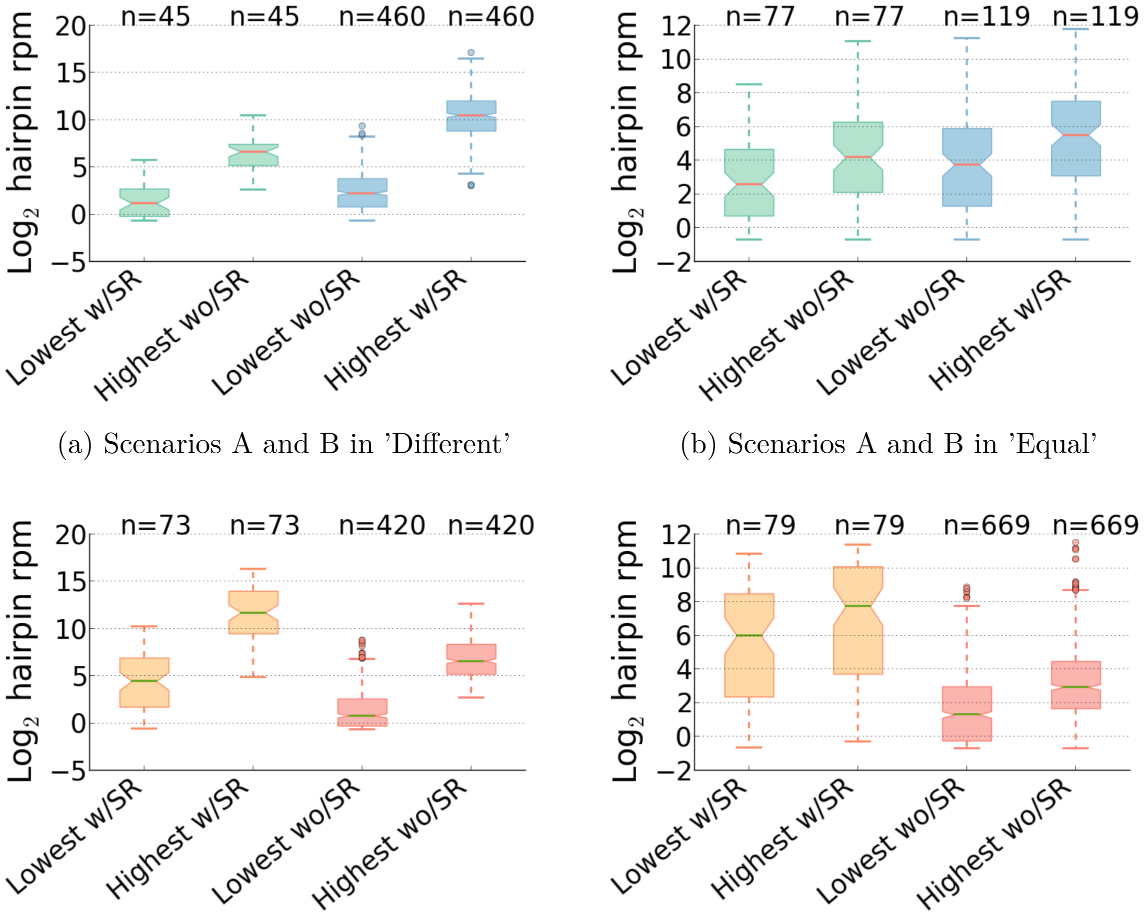
(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal'

Figure F.16: The expression levels of each hairpin strand of scenario A (green), B (blue), C (orange) and D (red) for the *Corry* data set in the 'Different' and 'Equal' class.



(a) Scenarios A and B in 'Different'

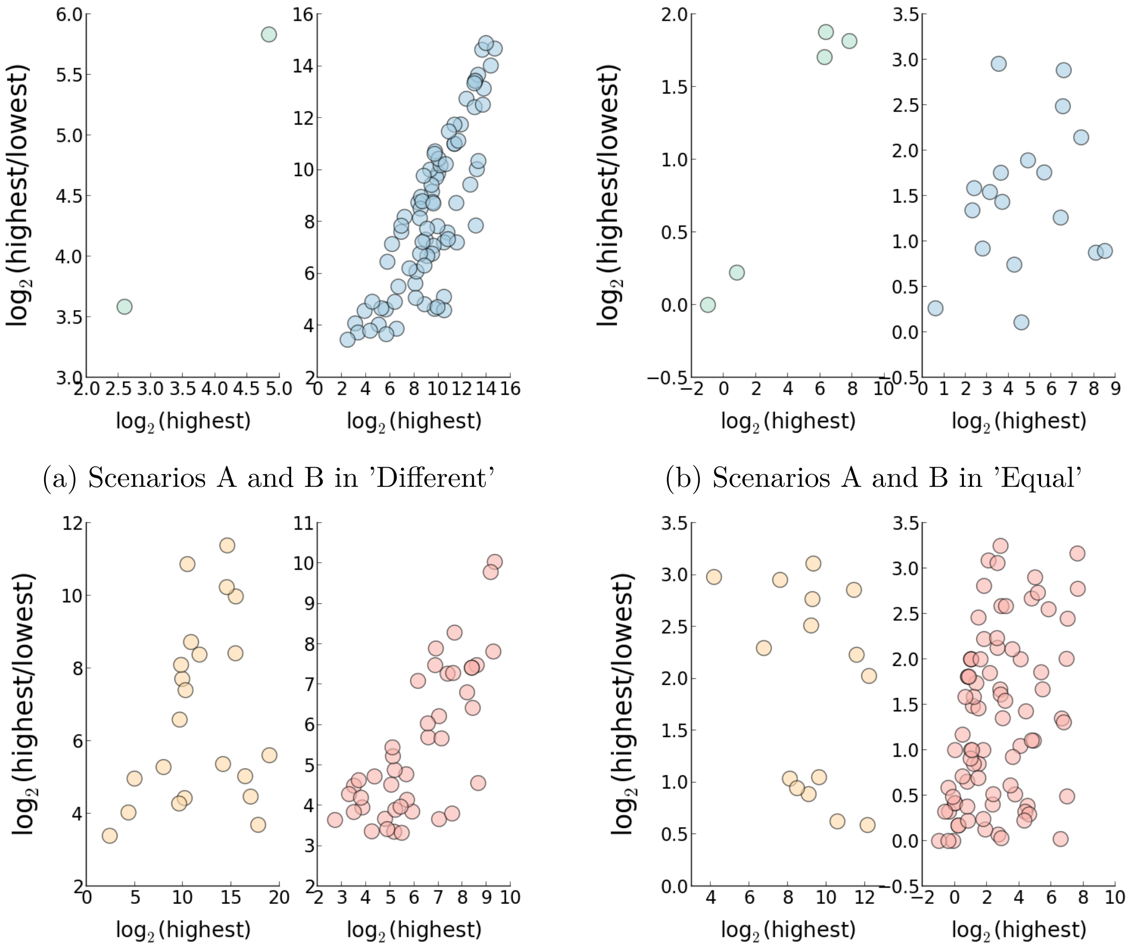
(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal'

Figure F.17: The expression levels of each hairpin strand of scenario A (green), B (blue), C (orange) and D (red) for the *Lundbæk* data set in the 'Different' and 'Equal' class.

F.4 FC of alignment scenarios



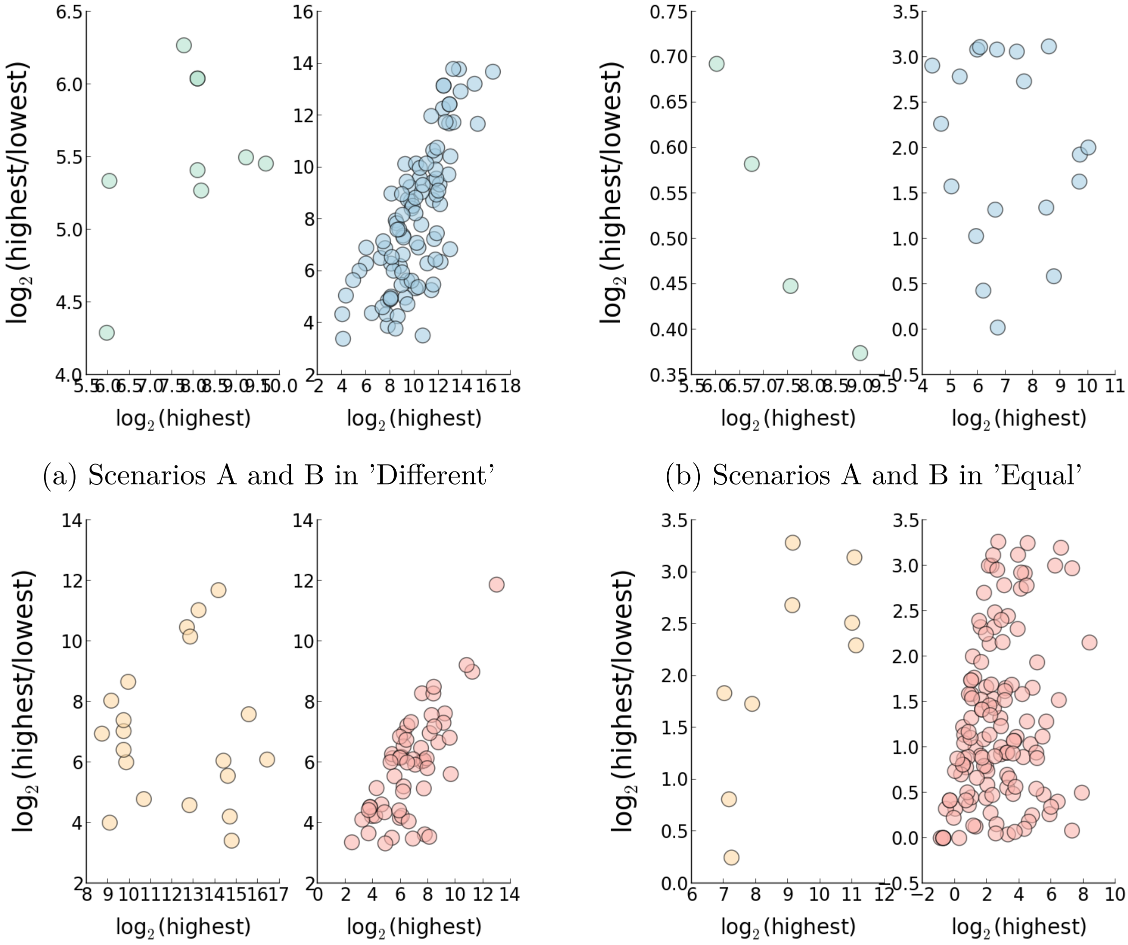
(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal '

Figure F.18: Fold change of scenarios A (green), B (blue), C (orange) and D (red) for the *Meister* data set.



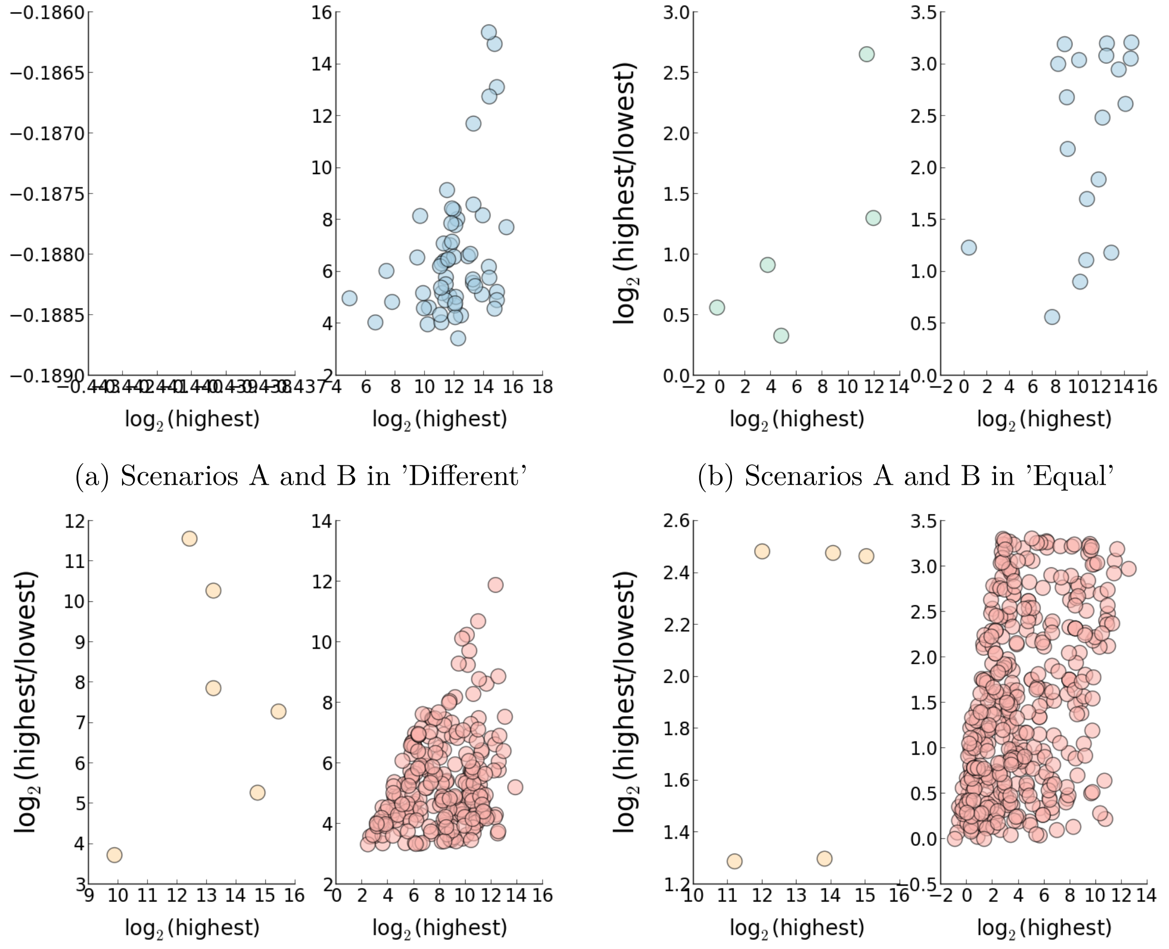
(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal '

Figure F.19: Fold change of scenarios A (green), B (blue), C (orange) and D (red) for the *Daub* data set.



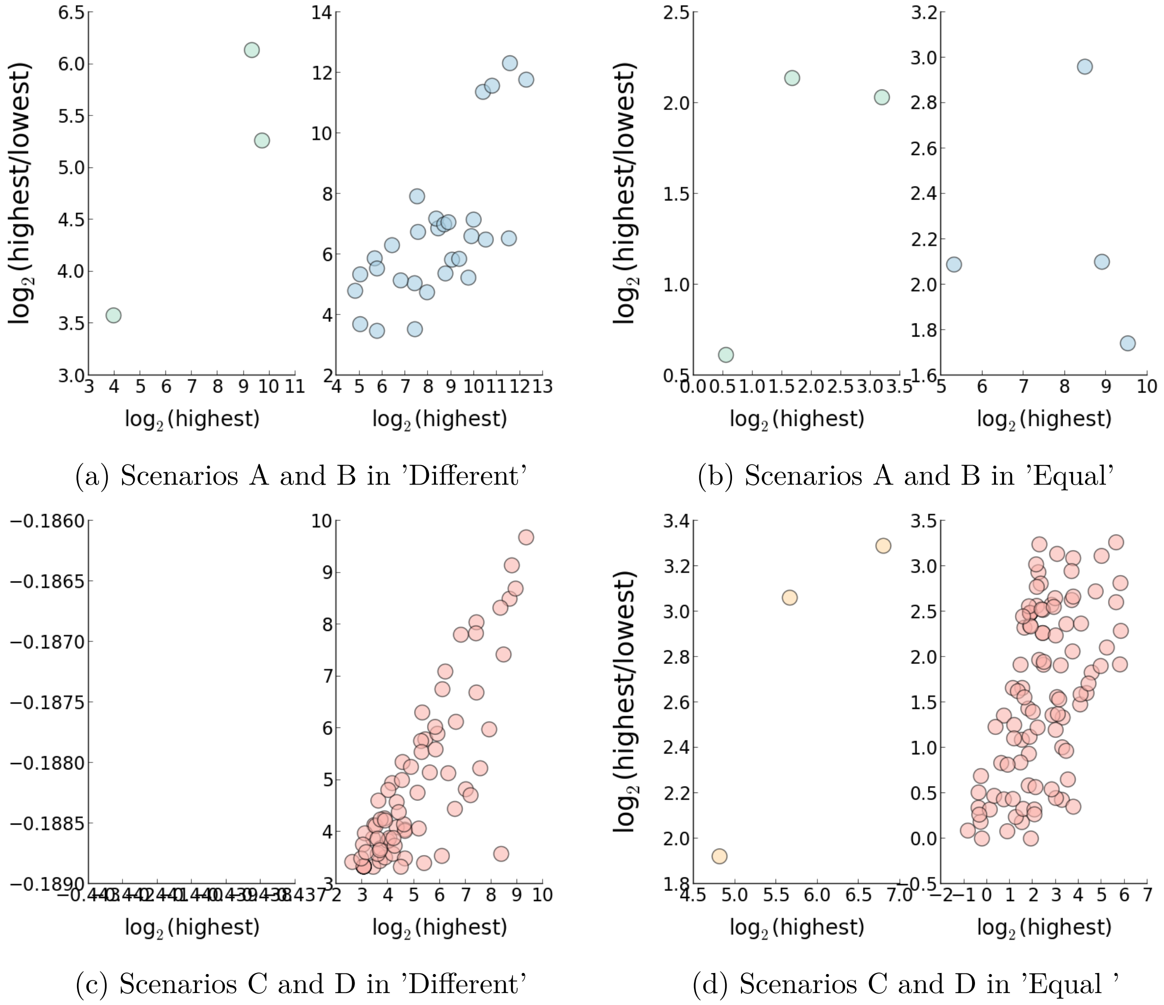
(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal '

Figure F.20: Fold change of scenarios A (green), B (blue), C (orange) and D (red) for the *Rajewsky* data set.



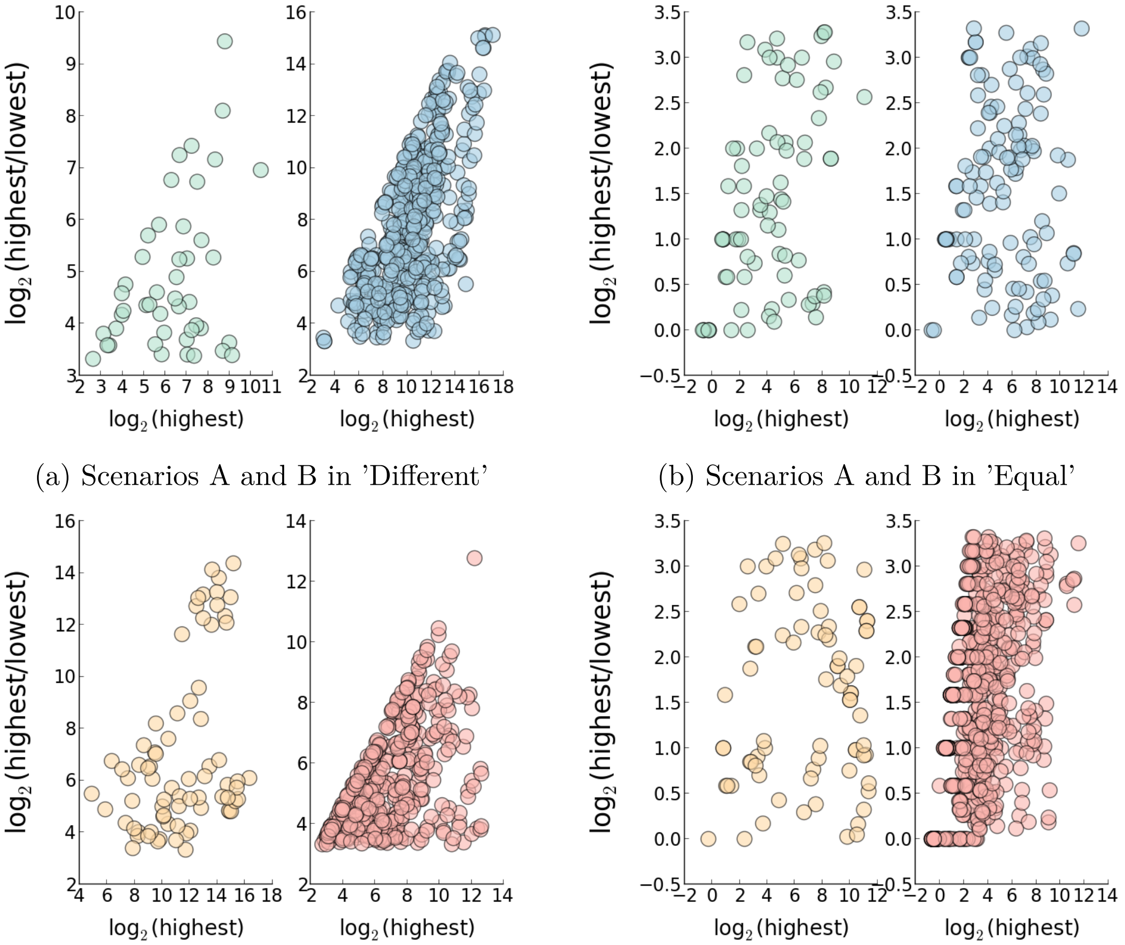
(a) Scenarios A and B in ’Different’

(b) Scenarios A and B in ’Equal’

(c) Scenarios C and D in ’Different’

(d) Scenarios C and D in ’Equal ’

Figure F.21: Fold change of scenarios A (green), B (blue), C (orange) and D (red) for the *Corry* data set.



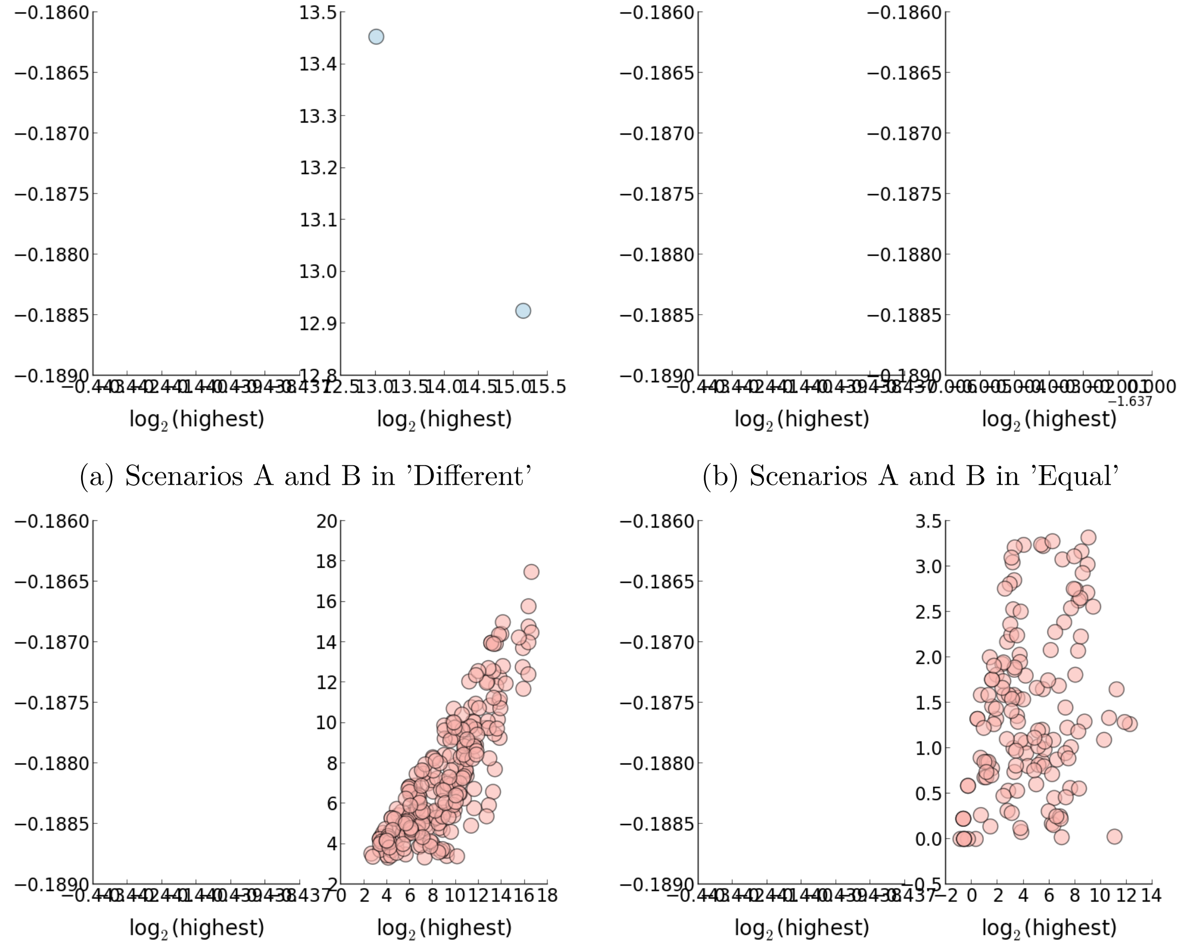
(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal '

Figure F.22: Fold change of scenarios A (green), B (blue), C (orange) and D (red) for the *Lundbæk* data set.



(a) Scenarios A and B in 'Different'

(b) Scenarios A and B in 'Equal'

(c) Scenarios C and D in 'Different'

(d) Scenarios C and D in 'Equal '

Figure F.23: Fold change of scenarios A (green), B (blue), C (orange) and D (red) for the *Rui* data set.

Appendix G

IsomiR results

G.1 Alignments

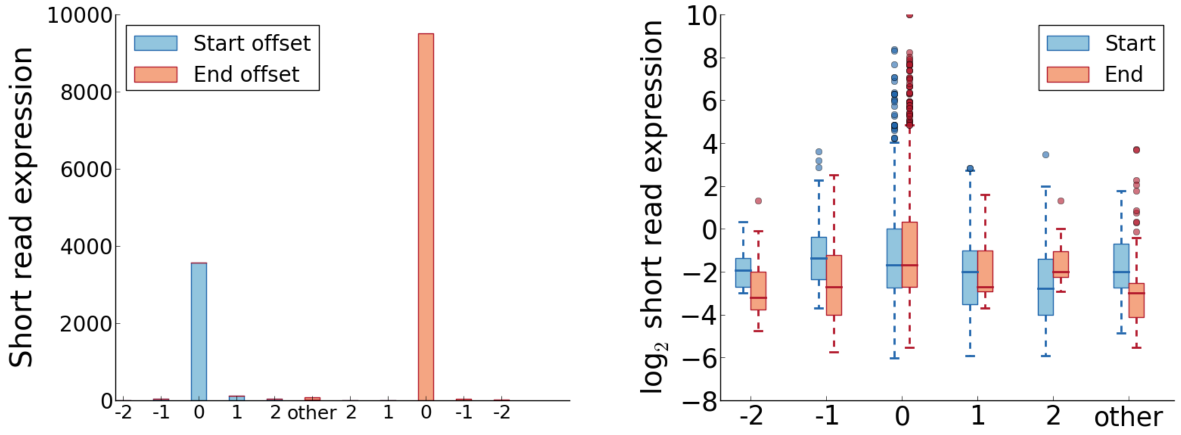


Figure G.1: Short read alignments to all isomiRs in total rpm (left) and quartiles (right) for the Meister data set

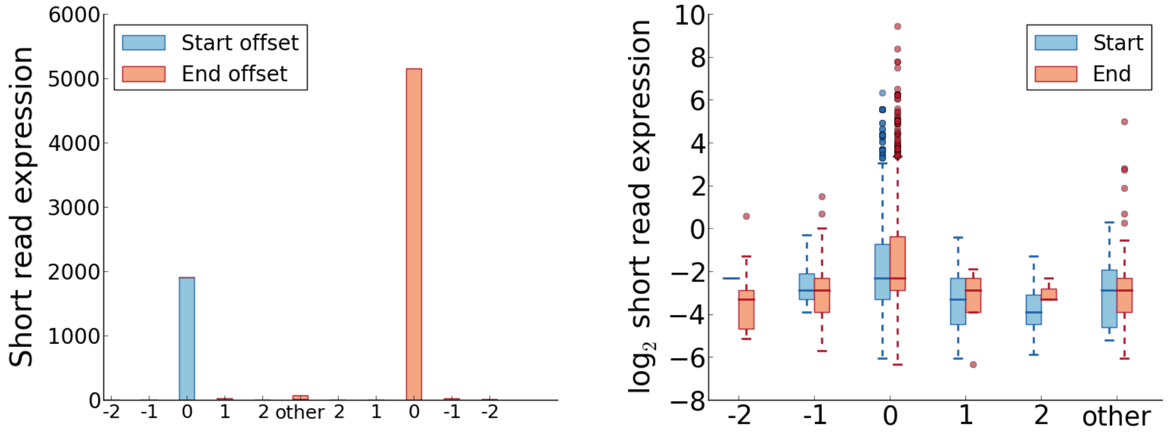


Figure G.2: Short read alignments to all isomiRs in total rpm (left) and quartiles (right) for the Daub data set

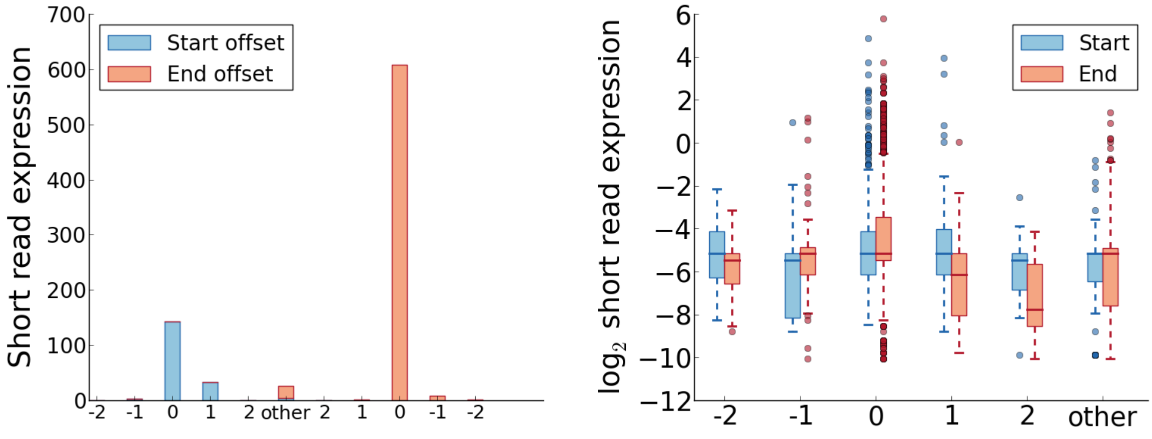


Figure G.3: Short read alignments to all isomiRs in total rpm (left) and quartiles (right) for the Rajewsky data set

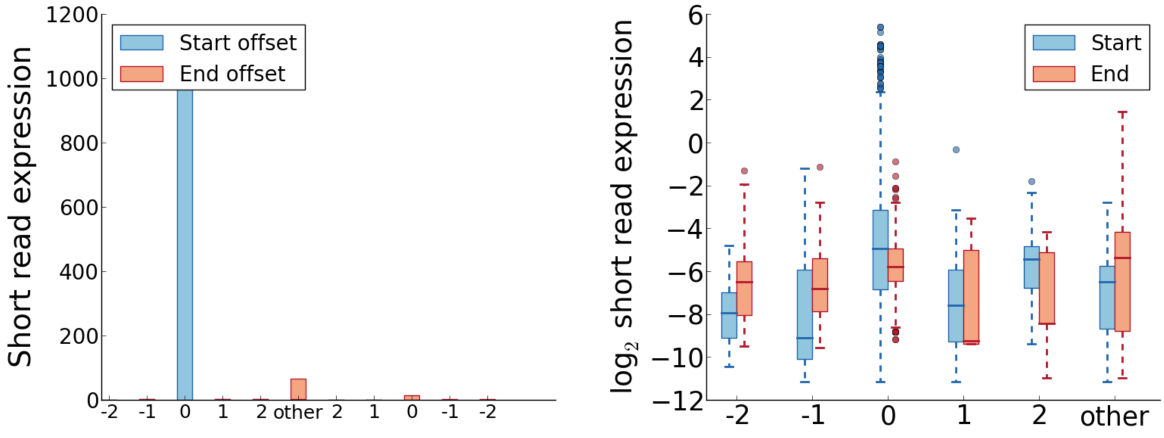


Figure G.4: Short read alignments to all isomiRs in total rpm (left) and quartiles (right) for the Corry data set

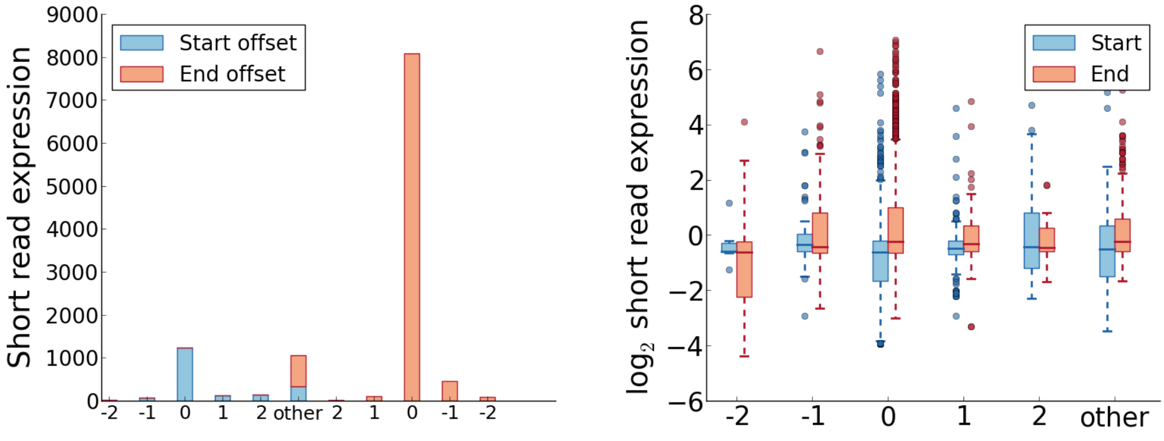


Figure G.5: Short read alignments to all isomiRs in total rpm (left) and quartiles (right) for the Lundbæk data set

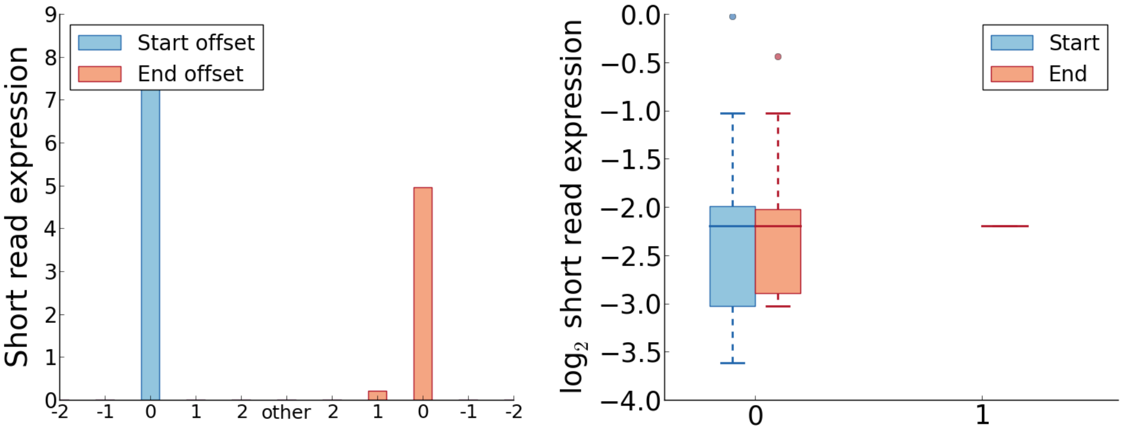
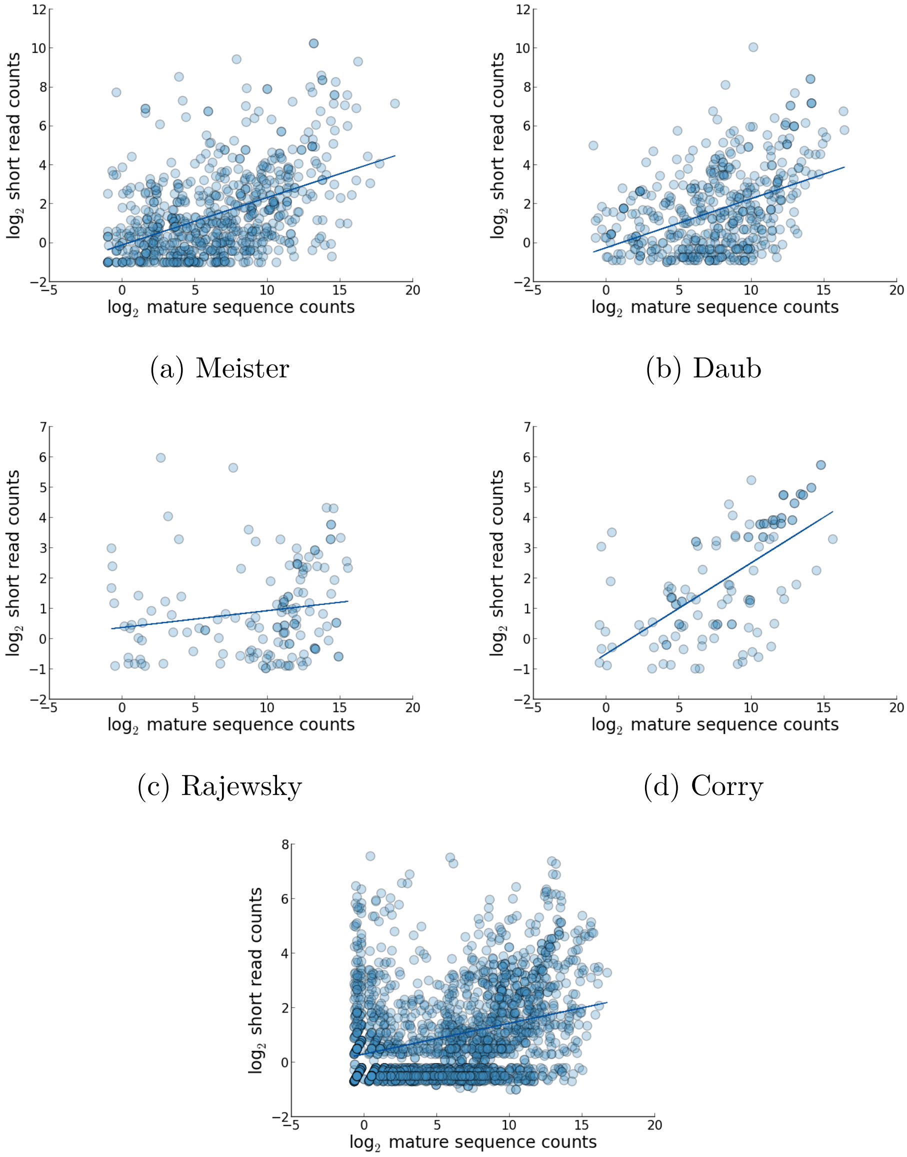


Figure G.6: Short read alignments to all isomiRs in total rpm (left) and quartiles (right) for the Rui data set

G.2 SR and miRNA correlation



(a) Meister

(b) Daub

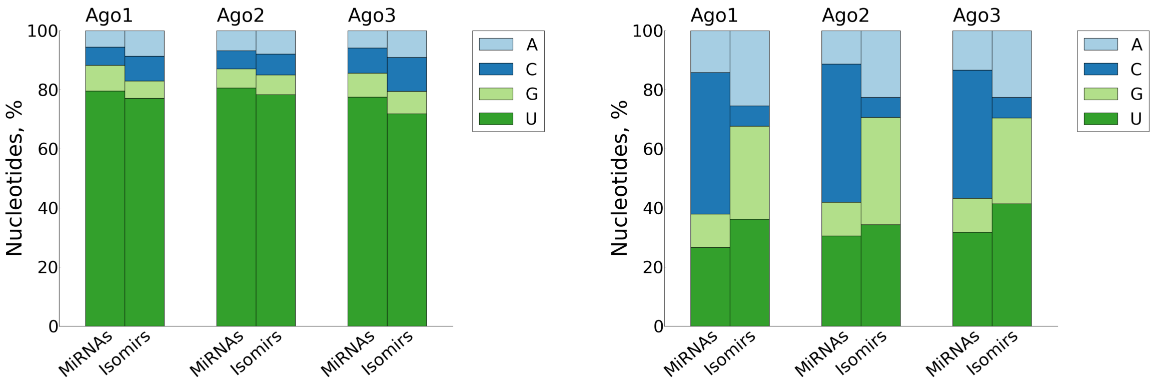
(c) Rajewsky

(d) Corry

(e) Lundbæk

Figure G.7: Correlation between expression levels of short reads and corresponding isomiRs for (a) Meister, (b) Daub, (c) Rajewsky, (d) Corry and (e) Lundbæk.

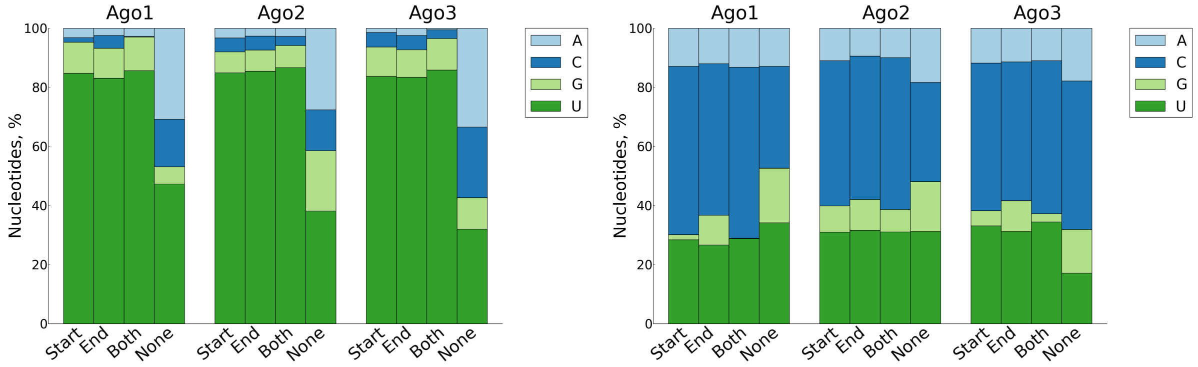
G.3 Nucleotide preferences



(a)

(b)

Figure G.8: The nucleotide distribution of mature sequences and isomiRs for the 5' (a) and 3' (b) strand of hairpins in the *Daub* data set.



(a)

(b)

Figure G.9: Terminal nucleotide preferences of isomiRs associated with start reads, end reads, both start and end reads or neither for the 5' (a) and 3' (b) strand of the *Daub* data set.

Appendix H

Argonaute dependencies

Table H.1: Comparison between the knock-out and wild-type samples of *Lundbæk*. The short read column represents the total, normalised expression level of short reads. The matures column represent the results of regarding only mature sequences, while the isomiRs column represents results from regarding all isomiRs. For the isomiRs column, the expression level of all isomiRs of the same miRNA are summarized, for the matures column the mature sequences of the same miRNA are summarized. For both columns, the values are log2 transformed, and the 75th and 90th percentile calculated and presented in the corresponding columns in the table.

	Short reads	**Matures**		**IsomiRs**	
		75th	90th	75th	90th
DOC KO	1,518	7.25	10.63	7.88	11.12
DOC WT	3,735	7.35	11.04	7.95	11.42
GH KO	2,435	7.39	11.06	8.07	11.67
GH WT	3,643	6.96	10.96	7.66	11.80