NTNU
Norwegian University of
Science and Technology

Shortreads (SR) analysis pipeline and correlation between SR and microRNA (miRNA) expression.

Fedor Fadeev

Master of Science in Computer Science
Submission date: June 2016
Supervisor: Pål Sætrom, IDI

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

NTNU – Trondheim
Norwegian University of
Science and Technology

Correlation between ShortRead expression in miRNAs with different length of biogenesis pathway and ShortRead analysis pipeline.

Fedor Fadeev

June 2016

PROJECT / MASTER THESIS

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

Supervisor: Professor Pål Sætrom

Preface

This project was carried out during spring semester of 2016 as a finishing part of master's programme at NTNU.

This report is aimed at everyone interested in topics around MicroRNA and requires little special background. However some background in molecular biology and understanding of basic programming concepts will ease the reading.

Trondheim, 2016-06-10

Fedor Fadeev

Acknowledgment

I would like to thank my supervisor Pål Sætrom for the guidance and help he provided.

Abstract

MicroRNA are small non-coding RNA molecules that execute post-transcriptional regulation of gene expression for over half of human and mammalian genes.

This report investigates shortreads (11-15 nucleotides) that align to start or end of microRNAs. Their expression and correlation to miRNA grouped by length of half-lives. An attempt was made to implement a processing pipeline for this type of analysis - "Shores". Shores is described and discussed in the report and is used for all the analysis on shortreads.

Contents

Preface . . . i

Acknowledgment . . . ii

1 Introduction 2

1.1 Aim of work . . . 2

1.2 DNA makes RNA makes protein [13] . . . 3

1.3 Micro-RNA [11] . . . 4

1.4 Shortreads . . . 4

2 Methods and Materials 6

2.1 Data . . . 6

2.1.1 4SU dataset . . . 6

2.1.2 Fantom5 dataset . . . 6

2.2 Shores . . . 6

2.2.1 Prerequisites . . . 8

2.2.2 Organisation . . . 8

2.2.3 Example use-case . . . 8

2.3 Details on implementation of particular shores functions . . . 9

2.3.1 Correlation in time between SR and miRNA expressions in 4SU . . . 9

2.3.2 Scatter plot . . . 10

2.3.3 Grouped SR . . . 10

2.4 Tools and resources used in shores . . . 11

2.4.1 Bowtie . . . 11

2.4.2 MiRBase . . . 11

3 Results 12

3.1 4SU Shortreads 12

3.1.1 4SU dataset peculiarities 12

3.1.2 Differences in shortread levels 12

3.1.3 Correlation between shortread and miRNA expression levels 14

3.2 Shores 16

3.2.1 Scatter plot 18

3.2.2 Grouped SR plot 18

A Acronyms 28

B GSM entries of 4SU dataset 29

C Low RC in Fantom5 30

C.1 SR/miRNA ratio = 0.75 31

C.2 SR/miRNA ratio = 0.5 32

Chapter 1

Introduction

This short chapter begins with explaining the aim of work. It then quickly guides through the concepts required for understanding the work presented in other chapters.

1.1 Aim of work

In my preliminary study I explored shortreads in humane Fantom5[14] dataset. During that time a number of scripts were written. Aim of this work is to solidify all the individual code units that were implemented into a generic processing pipeline and make it accessible possibly for external interested parties. It is further to be applied to a dataset published with a paper on MiRNA stability[10].

Aim of this work is to explore whether shortreads expression significantly correlates with the expression of corresponding MiRNAs in a time series sample. The dataset from mentioned article is a time series dataset with 7 samples per experiment and is used for that purpose. The article also provided grouping of MiRNAs by their decay rate. Which made it possible to analyze whether shortreads expression trends differ between MiRNA groups with different degradation time.

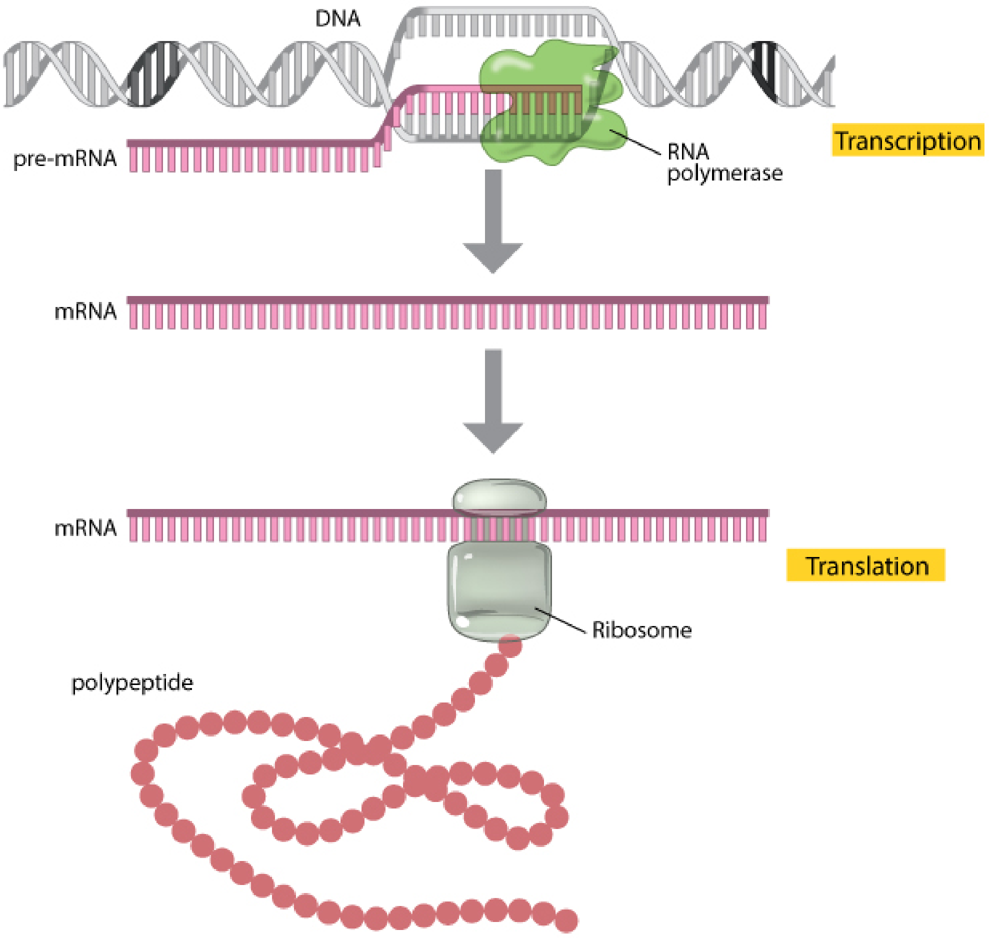


Figure 1.1: Gene expression through transcription and translation [13]

1.2 DNA makes RNA makes protein [13]

DNA is a molecule that carries genetic information in all living organisms. DNA is a long sequence of 4 basic monomers called nucleotides - adenine, cytosine, guanine and thymine (`A, C, G, T`). RNA is also a sequence of same basic nucleotides as is DNA, except instead of thymine it contains uracil. Proteins are generated in the following manner: first a sequence from DNA is *transcribed* into a piece of RNA called *messenger-RNA(mRNA)*. mRNA is a template for building a protein. The second step is *translation* into protein. In this step the mRNA sequence is processed in such a way that each 3 nucleotides generate a building block in the translated protein - one of the 21 amino acids. In the process of translation the protein will fold based on the properties of its amino acids. This process is illustrated on fig. 1.1.

1.3 Micro-RNA [11]

Not all translated DNA sequences yield protein-producing mRNA. Such RNAs are called non-coding RNAs (ncRNA). Specific group of ncRNA of interest is the micro-RNAs (miRNA). These short (ca 22 nucleotides long) sequences have been discovered to have a role in post-transcriptional regulation of gene-expression. This role is fulfilled through participation in RNA-Induced Silencing Complex (pp. 286-289 [11]). A single miRNA can target multiple genes. More than 60% of human protein-coding genes are targets of miRNA.

Following is a very simplified description of how translated RNA strand interacts with RISC.

When transcribed from DNA, future-miRNA forms imperfectly-aligned hairpin-structure - see fig. 1.2. A After exporting to cytoplasm and being processed by DICER, two base-paired RNA-strands are left. One of those strands will be loaded into Argonaut (AGO) protein and thus become a *mature miRNA*. The base-paired strands are referred to as guide-(the one that will become a mature miRNA) and passenger-(the one that will be discarded) strands. After being accepted into AGO, mature miRNA fulfills its biological role by helping target mRNA and interfere with the protein transcription.

1.4 Shortreads

Mossin in 2014 has discovered short ∼10 nt reads(shortreads or SR) that align to mature miRNA [2]. Before that, such shortreads were mostly considered unimportant degradation products and discarded. They do however appear to be products of undescribed mechanism of miRNA biogenesis. Analysis of these shortreads is the focus of the paper.

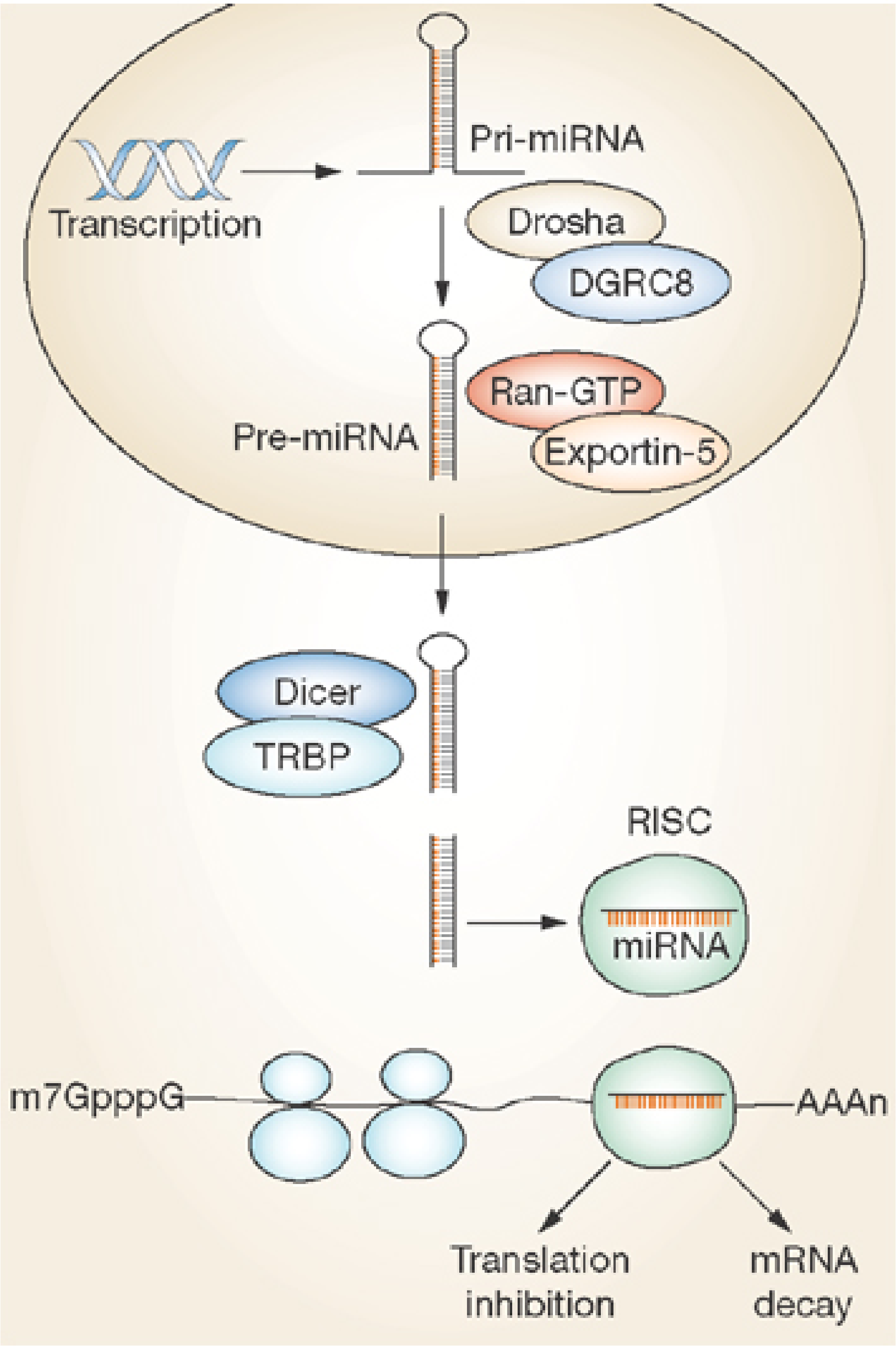


Figure 1.2: Lifespan of miRNA [16]

Chapter 2

Methods and Materials

2.1 Data

2.1.1 4SU dataset

What's called *4SU-dataset* in the rest of the paper is 3 experiments as time series with 2h difference between subsequent samples. Each experiment spans 14h and consists of 7 samples. They were released as additional materials for the paper on miRNA stability [10].

List of GSM-, SRR-entries and descriptions are available in appendix B.

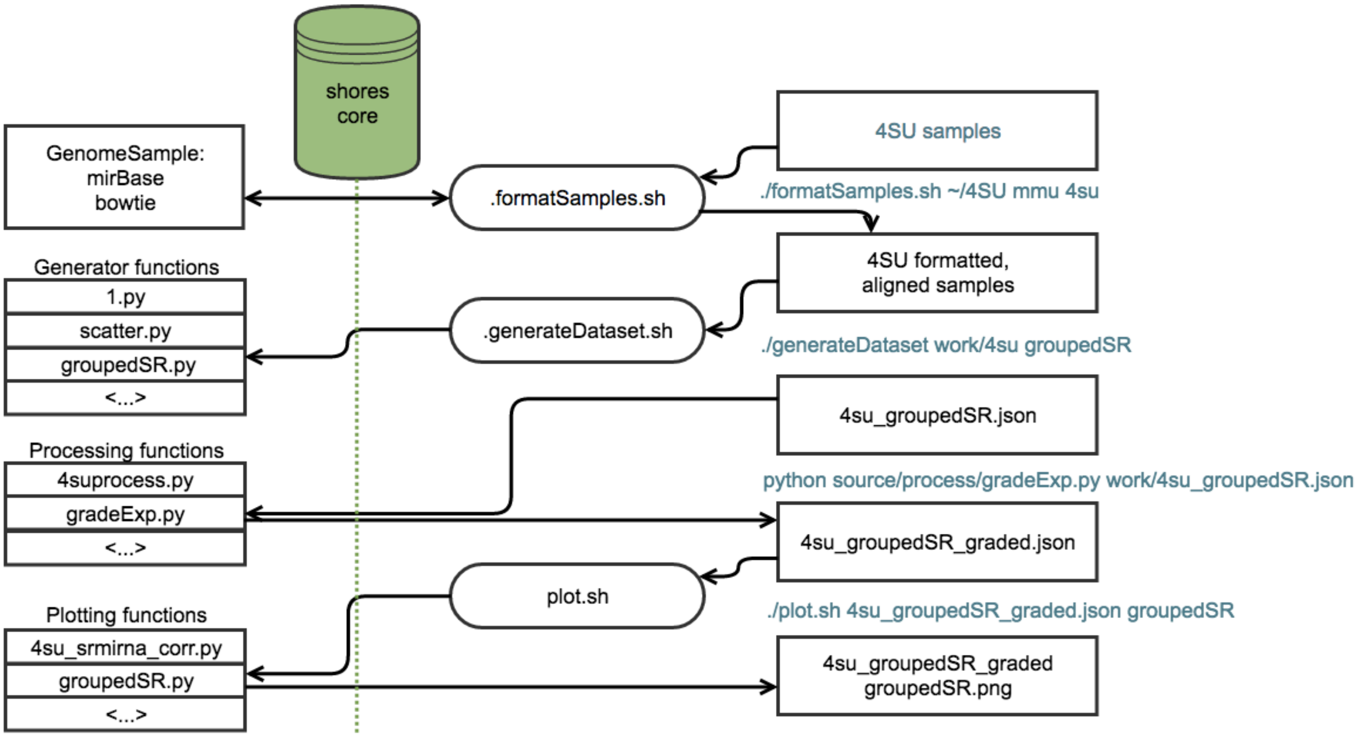
Provided samples are mouse-samples.

2.1.2 Fantom5 dataset

For reference and for demonstration of *shores* (section 2.2) on some plots Fantom5[14], [15] humane dataset was used. I used the same dataset in my preliminary work[9].

2.2 Shores

Shores [8] is a processing pipeline built to simplify and empower the process of exploring the data. It is built around *findShortReatsMiRNA.py* written by Kristin Wahl for her master's thesis [3].



```
./formatSamples.sh ~/4SU mmu 4su
./generateDataset work/4su groupedSR
python source/process/gradeExp.py work/4su_groupedSR.json
./plot.sh 4su_groupedSR_graded.json groupedSR
```

Figure 2.1: User mostly interacts with shores by running *.sh* scripts. Text marked with blue are the command issued by user through the sequence. Most *.sh* scripts act through running shores core scripts which in turn call the necessary functions implemented in atomic scripts in column to the left. All the *.json* datasets generated throughout the sequence are easily explorable and modifiable by the user.

It consists of a number of shell-scripts that run python-code and effectively allow placing a bunch of files (collapsed, with removed adapter-sequence) and producing *.png* plots by running 3 simple shell-scripts. It was also attempted to make it highly transparent and extensible so that each intermediate step can be modified and easily delved into. Each intermediate `.json` dataset can be loaded from command line with `python load.py <dataset>`, which loads a python REPL(appendix A) with dataset loaded into `data` pandas variable. Also introducing own processing functions is made straightforward.

Shores was initially a way to improve code organisation in what code was reused from project on shortreads done in 2015 [9]. It grew however into an honest attempt at an organised data-processing pipeline. See fig. 2.1 for a breakdown of parts for a single case of building a *groupedSR* type graph.

2.2.1 Prerequisites

Shores requires a number of prerequisites that regrettably doesn't exactly let user just clone it from Github and run it after that. Prerequisites include:

- Python 2.7
- numpy, scipy, biopython, seaborn python packages [5],[6]. All these are available through *pip* python package manager.
- bowtie (see section 2.4.1)

2.2.2 Organisation

To make use of the pipeline some conventions should be followed. The most important convention is the folder-structure:

1. `/sampleSets` - folder where input datasets should be put. Example on format of input dataset is provided: */work/SampleUserData*.
2. `/work` - here all the intermediate datasets are stored along with output plots.
3. `/source` - folder containing python source code (further separated into `/generate` for dataset-generator functions, `/process` for singular scripts used to transform existing *.json* datasets, `/plot` for plotting-scripts).

The full description of *shores* is available on the github repository [8].

2.2.3 Example use-case

This section describes the steps needed to produce a simple scatter-plot of shortreads for provided example-dataset:

- `./formatSamples.sh hsa work/sampleUserData/ testset`
  format test dataset into proper form
- `./generateDataset.sh sampleSets/testset/ scatter`
  generate the dataset from */sampleSets/testset*

- `./plot.sh work/testset_scatter_*.json scatter`
  generate a scatter plot from the *.json* dataset

2.3 Details on implementation of particular shores functions

This section describes in detail core analysis modules in Shores that were implemented.

2.3.1 Correlation in time between SR and miRNA expressions in 4SU

This plot function is in many ways hardwired to work with 4SU dataset and needs major adjustments for it to work with arbitrary timeseries datasets.

```
./generateDataset.sh sampleSets/<sampleSet> 1
```

The most generic generator-function is `1`. It almost directly records all data from `findShortreadsMiRNAs.py`. Output is written in form of pandas[7] `.json` dataset.

```
py source/process/4suprocess.py <dataset.json>
```

Processor-function written specifically for 4SU dataset. It assigns experiment numbers by hardwired sample names. It also aggregates all shortreads for a particular miRNA, prime and position into one record.

```
py source/process/4su_fastslow_correlation.py <dataset.json>
```

This processor reads `work/miRNAlists/4sumirnas.txt` - list of miRNAs with their classes that were extracted from supplementary materials from the paper[10] where 4SU dataset originates from. For each miRNA with sufficient records of shortreads a single entry is produced: description of miRNA, SR position and prime, sum of expression values and correlation scores. Single miRNA can produce up to three records - one for each experiment.

```
./plot.sh <dataset.json> 4su_srmirna_correlation
```

Four plots are built from the dataset:

1. Boxplot on correlation values for different positions of shortreads and classes of miRNA. fig. 3.5.
2. Boxplot on relation between SR and miRNA expression. fig. 3.2.
3. Mean rpm (sum per 7 records in a single timeseries) barchart.
4. Boxplot with absolute expression values for SR and miRNAs (also sum per 7 records in each case). fig. 3.3.

2.3.2 Scatter plot

```
./generateDataset.sh sampleSets/<sampleSet> scatter
```

Scatterplot generator script records all miRNAs that have shortreads. For each miRNA in dataset the total rpm of start-sr and end-sr are recorded. For each entry also the raw readcounts are recorded. The resulting dataset is saved as `.json` file.

```
./plot.sh /Users/r/bio/shores/work/<dataset.json> scatter
```

Finally the plot-function is called. It filters entries by number of raw-readcounts with $RawRC(miRNA) + RawRC(SR)$ (See section 3.2.1 for the reasoning behind raw-readcount threshold). Threshold is set to 10. Then either all samples are plotted onto one image or each sample of the dataset is plotted onto own image. See fig. 3.9.

2.3.3 Grouped SR

Firstly the *groupedSr* json dataset is generated from a set of compiled samples:

```
./generateDataset.sh sampleSets/<sampleSet> groupedSR
```

It goes through each miRNA in sample and gathers expression values for miRNA and aligned shortreads. Based on those values group (*Equal/Different*) and subgroup (*high/low/both/none*) are decided and record is added to pandas[7] dataset. The records for each miRNA in each sample constitute the final pandas dataset that is saved as `.json` file.

```
py source/process/gradeExpression.py <dataset.json> 10
```

The `dataset.json` from previous step is processed with `gradeExpression.py`. Last parameter is the number of buckets of expression levels to separate the values into. Each record (miRNA) gets a grade based on relation to maximum expression value in sample.

```
./plot.sh <gradedDataset.json> groupedSR
```

Finally the graded dataset from previous step is provided to groupedSR plot-function. It compiles eight values per sample - one per each *group.subgroup*. The value denotes share of miRNAs of given subgroup of total number of miRNAs in the group.

Examples of such a plot as well as detailed explanation of grouping is presented in section 3.2.2: see fig. 3.10 and fig. 3.11.

2.4 Tools and resources used in shores

The processing pipeline uses a number of important tools provided by community.

2.4.1 Bowtie

Bowtie is a command-line application for sequence alignment [1]. Version 1.1.2 was used.

2.4.2 MiRBase

MicroRNA data was downloaded from miRBase. Release 21 was used in this report.

Chapter 3

Results

The results of this work can be separated in two parts: discussion on the attempts to build a shortread-analysis pipeline in section 3.2 and application of it onto a specific dataset in section 3.1.

In a recently published study[10] half-lives of miRNA were studied. The half-life times were discovered to differ drastically. MiRNAs were grouped by the results into *fast*(4-14h), *slow*(>24h) and *other* groups.

Sequencing results were available as time-series spanning 12h (with 2h difference between subsequent samples) for three experiments. Analyzing these timeseries for shortreads with regards to described grouping of miRNA follows in section 3.1.

3.1 4SU Shortreads

3.1.1 4SU dataset peculiarities

4SU dataset has significantly more shortreads aligned to offsets other than $-1, 0, 1$ (see fig. 3.1) in comparison to my earlier study (figure 3.5 in [9]. 4SU samples belong to mouse and not human as the samples from that project.

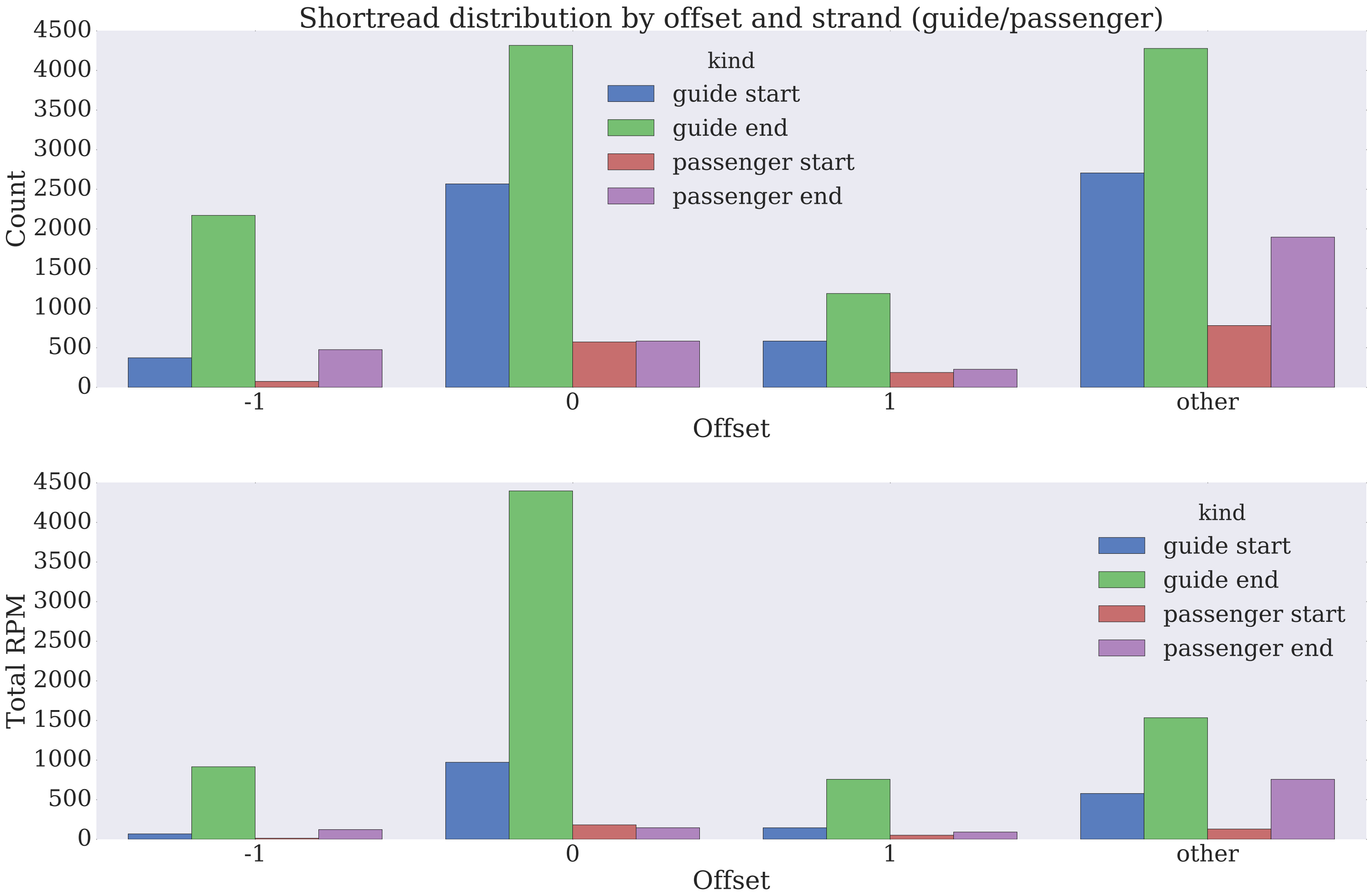


Figure 3.1: Figure shows distribution of shortreads between offsets and strands (guide/passenger). The top graph displays count of all sr-occurrences on y-axis, while the bottom graph on y-axis displays sum of reads per million of shortreads.

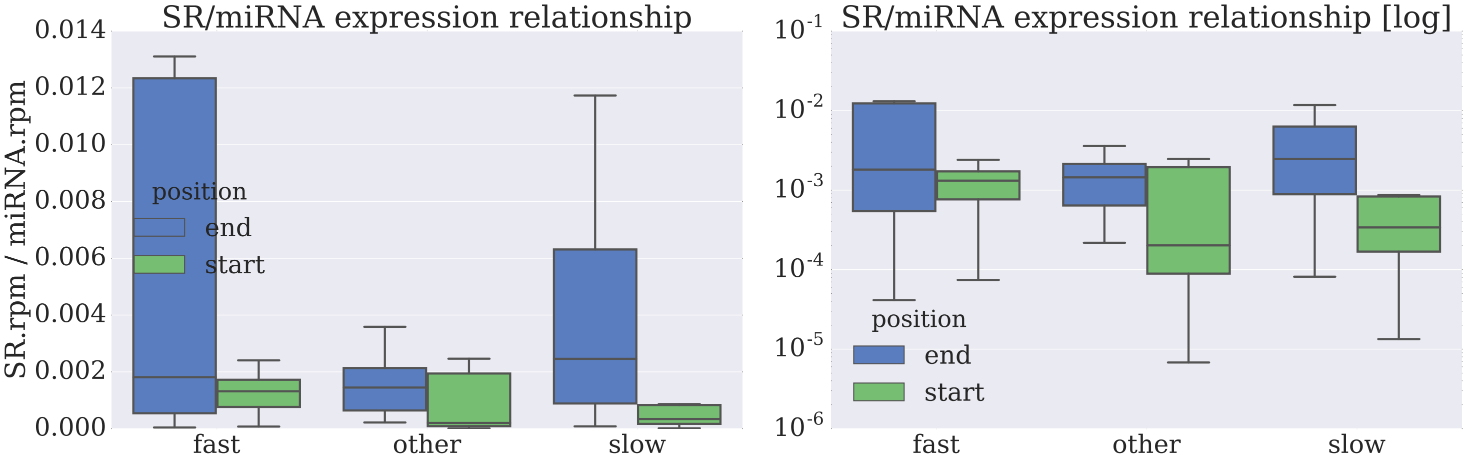


Figure 3.2: Figure depicts difference between relations $\frac{RPM(shortread)}{RPM(MiRNA)}$ in three groups that miRNAs were divided into in [10]

3.1.2 Differences in shortread levels

See fig. 3.2 and fig. 3.3. There hardly is much difference in levels of miRNA expression between *fast* and *slow* groups. There also is no notable difference in levels of end-shortreads. There is however apparent difference when it comes to start-shortreads. For *slow* miRNA group start-shortreads tend to have lower expression, which can possibly be a result of lower degradation activity of *slow* group of miRNAs.

3.1.3 Correlation between shortread and miRNA expression levels

The target of this inquiry is whether dynamics of shortread expression correlate with the expression of their miRNAs. Firstly miRNAs with shortreads persisting through all samples were picked and plotted as timeseries with respective shortreads: fig. 3.4. On this particular figure only miRNAs classified as *fast* are present. The figure does display some degree of correlation.

Further each miRNA with sufficient persistence of shortreads was given a correlation score per experiment - that is each of subgraphs on fig. 3.4 received correlation score. Pearson and Spearman methods were used. fig. 3.5 displays distribution of the correlation scores per class of miRNAs and the position of shortread. There is definite correlation between shortreads

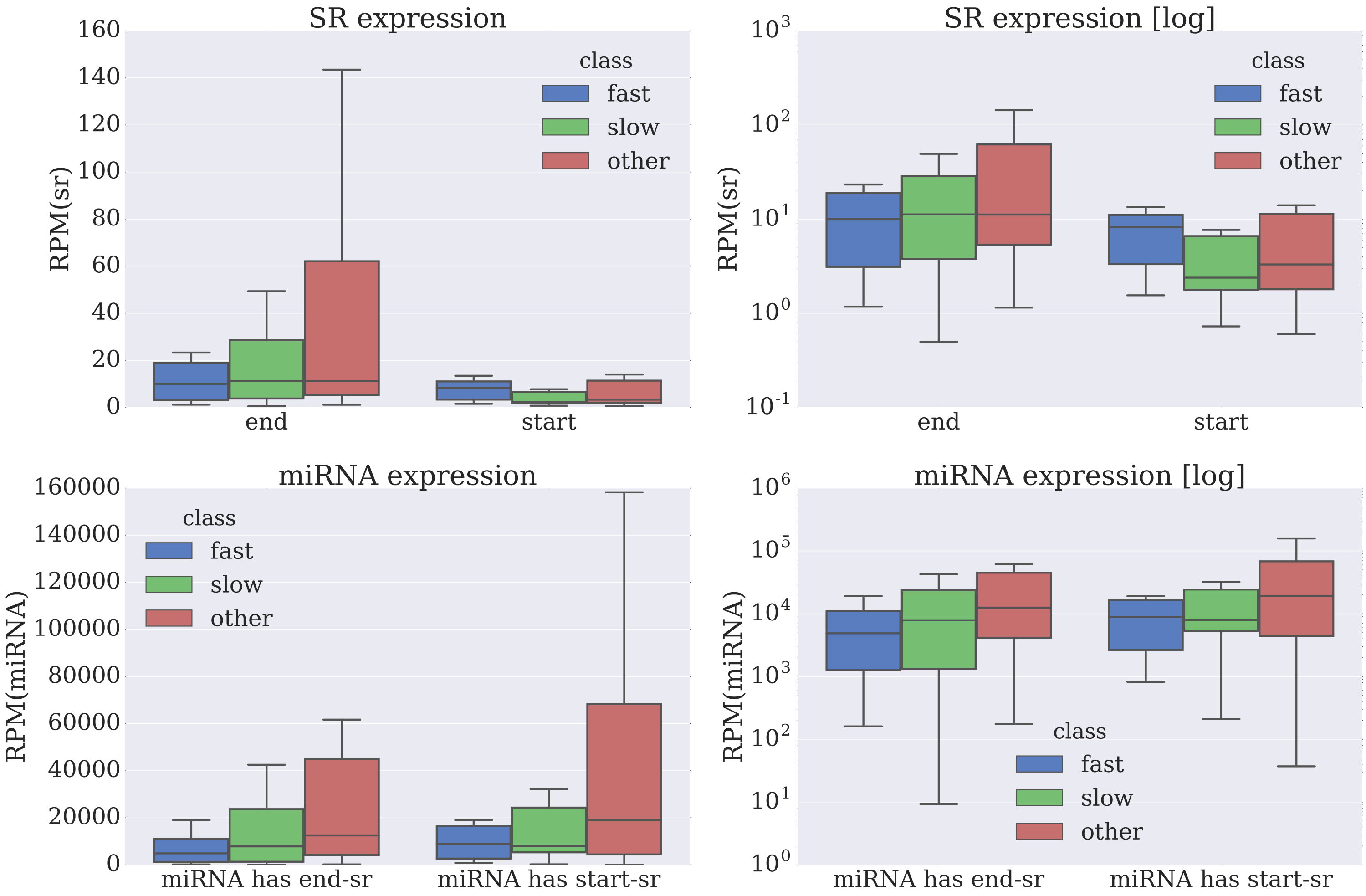


Figure 3.3: Box-plot with RPM values of expression of shortreads and miRNAs.

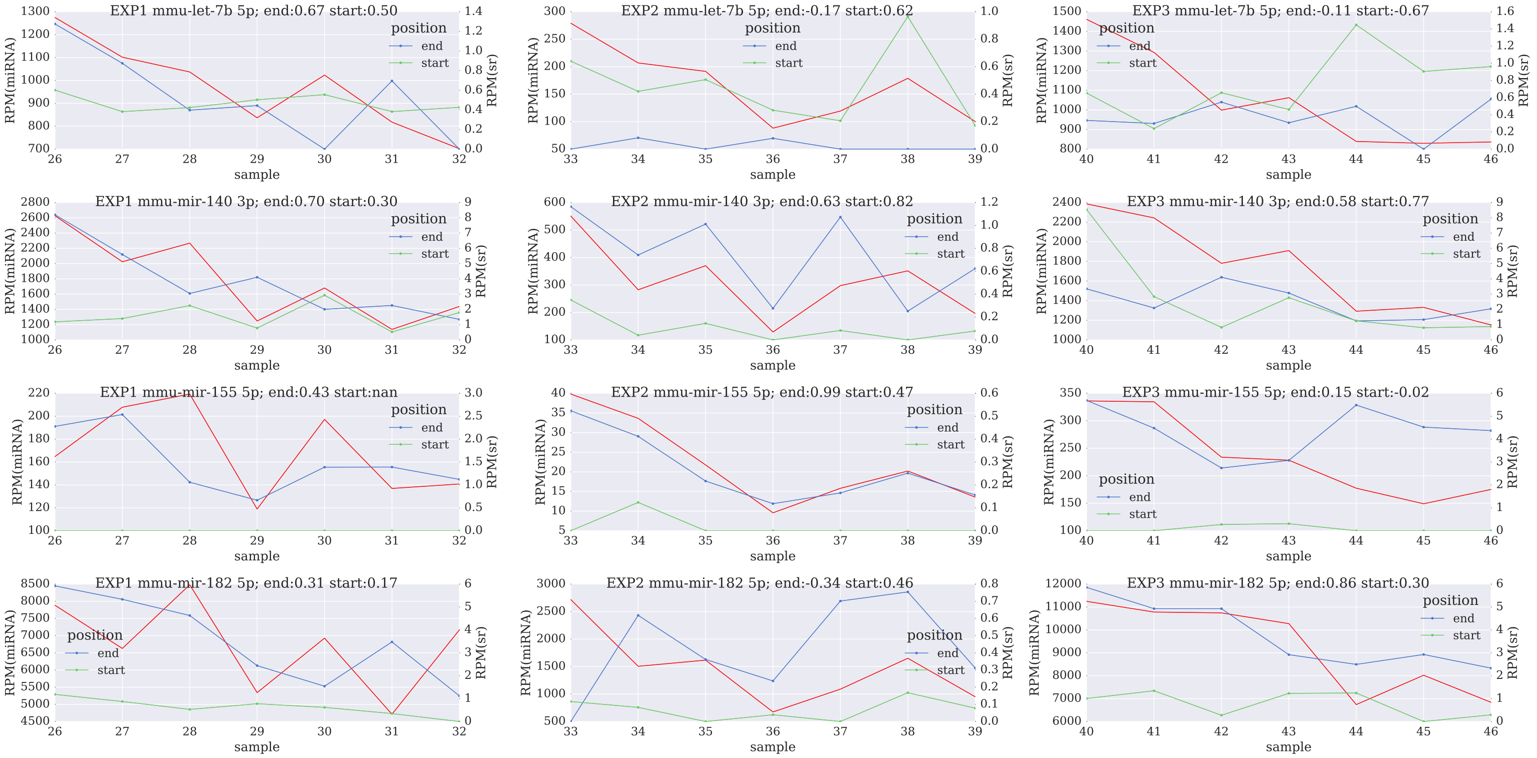


Figure 3.4: Expression dynamics of shortreads (blue and green lines, right y-axis) and corresponding miRNA (red line, left axis). Difference between each subsequent samples is 2 hours.

levels and miRNA levels in general. For details on implementation see section 2.3.1. Much to the contrast is the lack of correlation for start-SR in miRNAs classified as *slow* - miRNAs that have low degradation and long biogenesis path. It supports the claim that binds start-reads with degradation of miRNA. On the other hand end-reads are consistent in their high correlation with miRNA levels in all three classes.

In the earlier section 3.1.2 it was mentioned that start-shortreads have lower expression in *slow* group. This could mean that noise dominates the expression pattern in *slow*, unlike in other groups where noise is overshadowed by the expression pattern that is expressed with significantly higher read values.

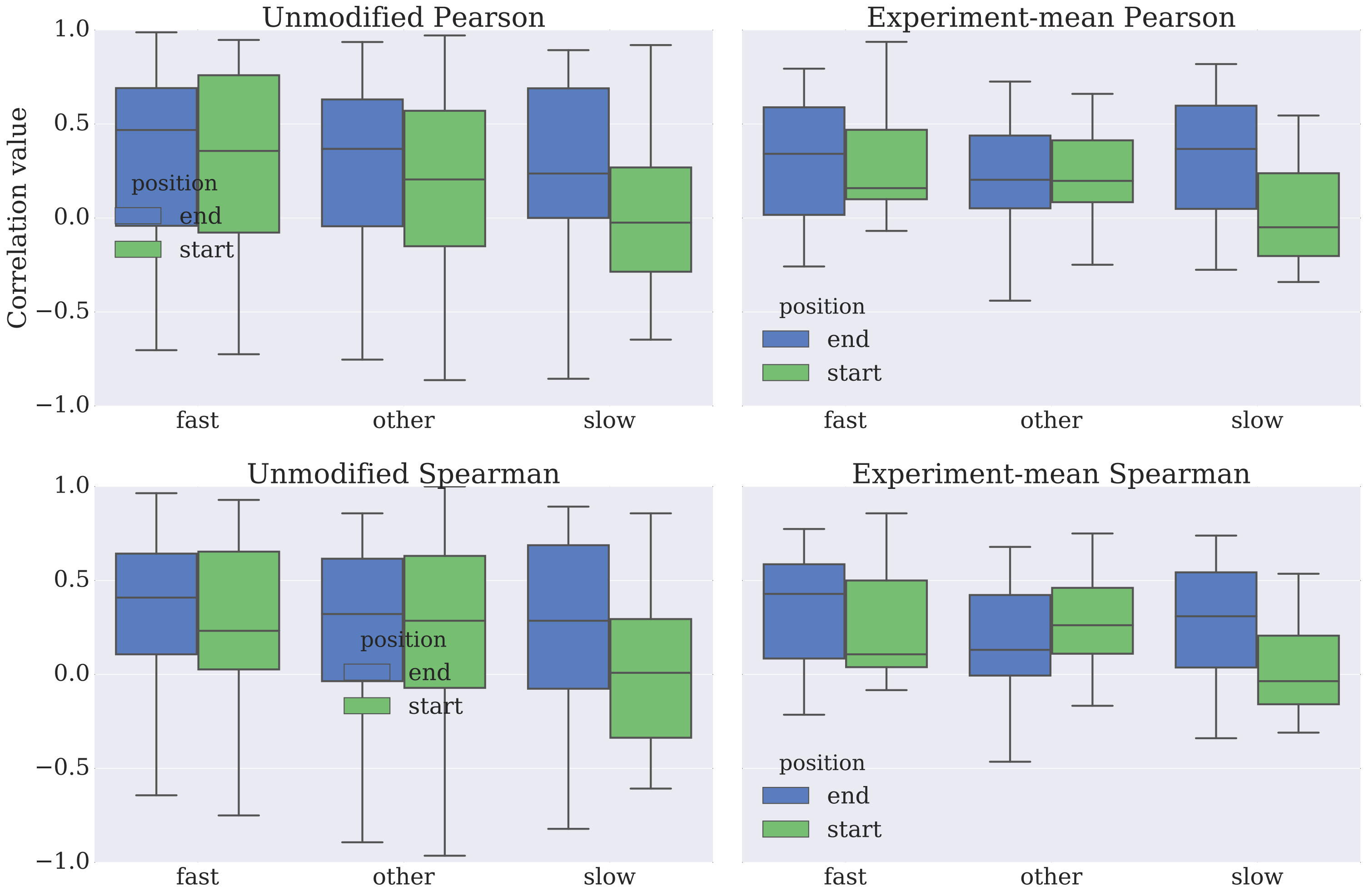


Figure 3.5: Correlation values for series of expression values of shortreads contra corresponding miRNA

3.2 Shores

This section goes through functions implemented in shores section 2.2 as well as discussion on the results displayed by aplpying these functions to a number of datasets.

3.2.1 Scatter plot

Shores scatterplot is meant to display relation between shortreads expression and related miRNA expression by plotting each encountered shortread. It was later decided to plot sum of shortreads per hairpin per position rather than each shortread individually, whereas the tables displayed later in this section assume plotting each individual shortread. This technicality can be disregarded as the following discussion still applies directly.

On the fig. 3.6 a number of "lines" of values are visible. It is easy to suspect that those lines correspond to discrete values of relation $\frac{Exp(SR)}{Exp(miRNA)} = \frac{1}{3}; \frac{1}{2}; 1; 2; etc.$ Which is really suspicious as there must be a reason for the shortreads to align so well.

Under closer inspection those turned out to be artifacts caused by small number of rawreads that after normalisation still produced such discrete values. See appendix C. It displays some of the entries for lines $\frac{Exp(SR)}{Exp(miRNA)} = \frac{3}{4}$ and $\frac{Exp(SR)}{Exp(miRNA)} = \frac{1}{2}$.

See fig. 3.7 for overview of some of those discrete values on scaterplot.

Finally the fig. 3.8 displays how limiting raw reads $(sr + miRNA)$ to > 10 can get rid of all the discussed lines of discrete values. It was decided to screen the entries this way in the *scatter* plots in *shores.*

One can either plot all samples within a given dataset onto one scatterplot or plot each sample into a single file. Example of the latter is a scatterplot from $4SU$ dataset: fig. 3.9. There is a line at $Exp(SR)$ 1.6. It is also explained by low number of raw reads that makes values end up on the same place on y-axis. In this particular plot there are 51 values with $Exp(SR) = 0.167347$. Most of those values under inspection were revealed to have $RawReads = 1$. The rest had $\frac{RawReads}{Alignments} = 1$.

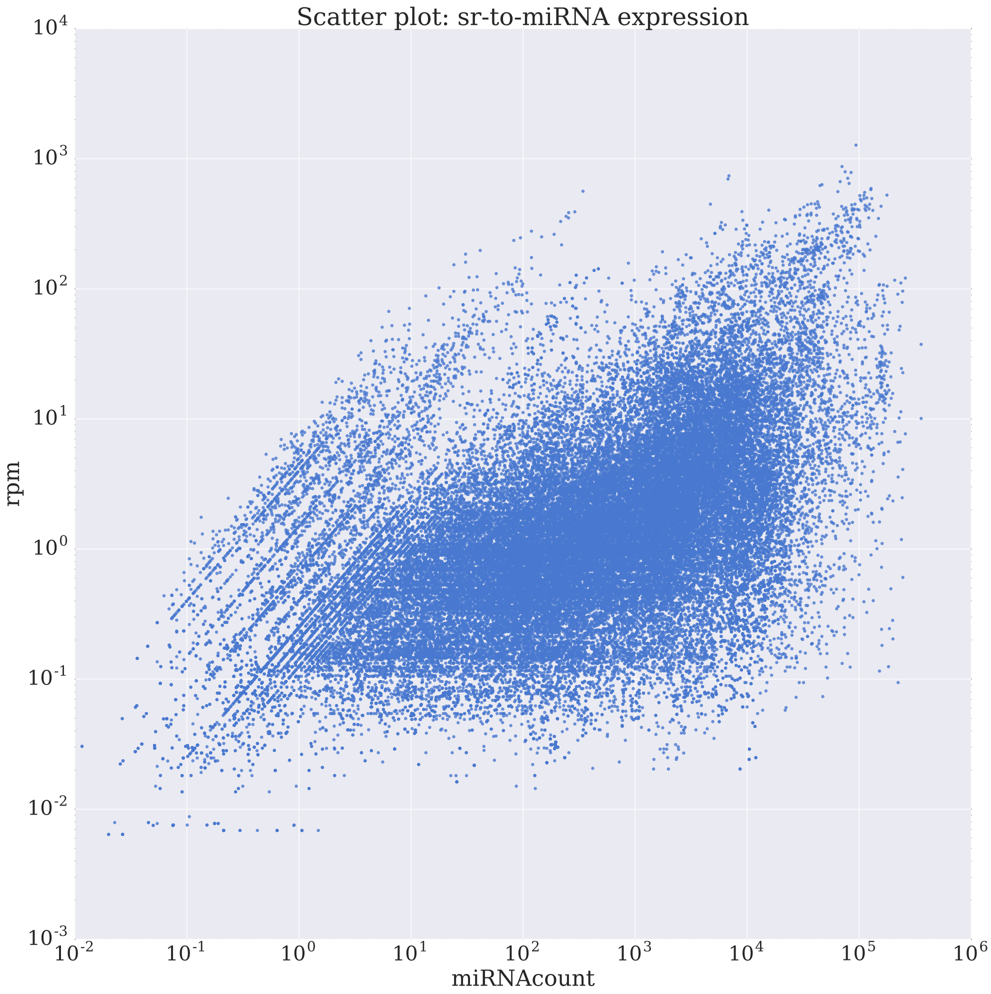


Figure 3.6: Scatterplot for all samples in fantom5 [14] dataset

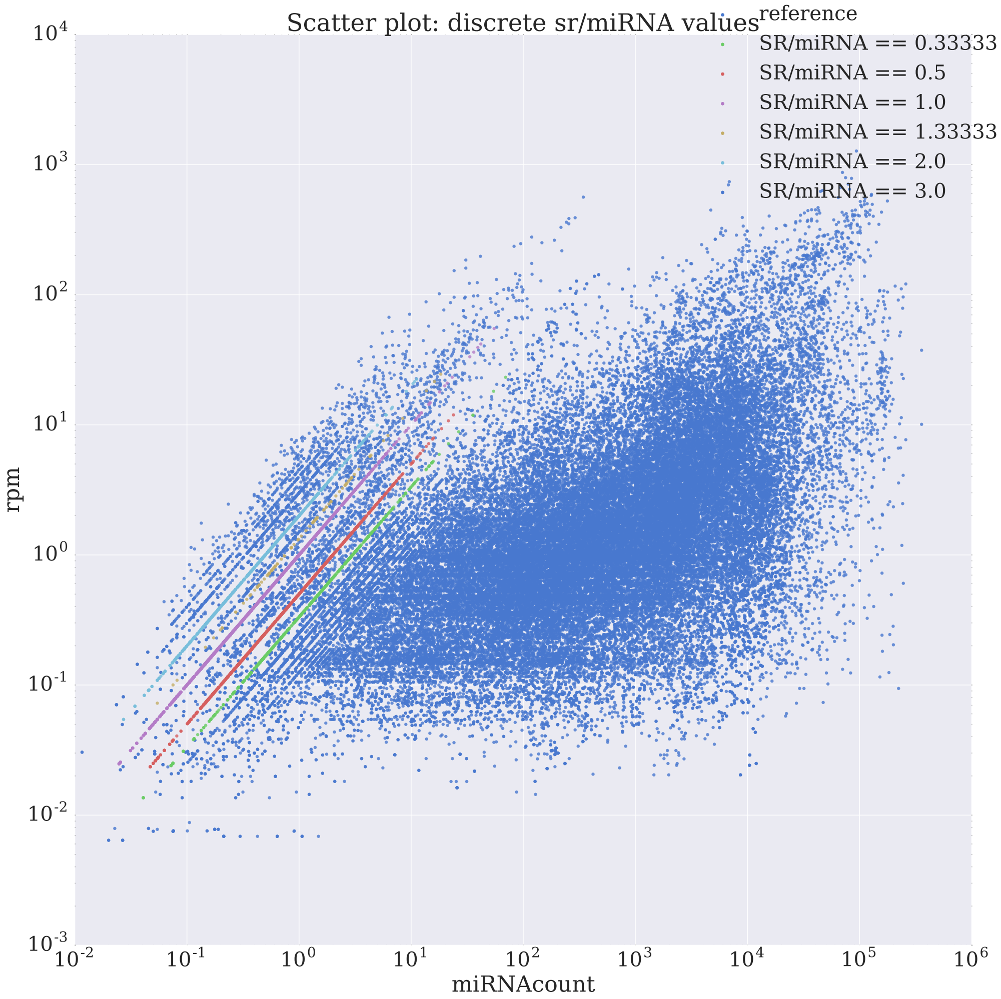


Figure 3.7: Scatterplot of fantom5 with discrete values highlighted

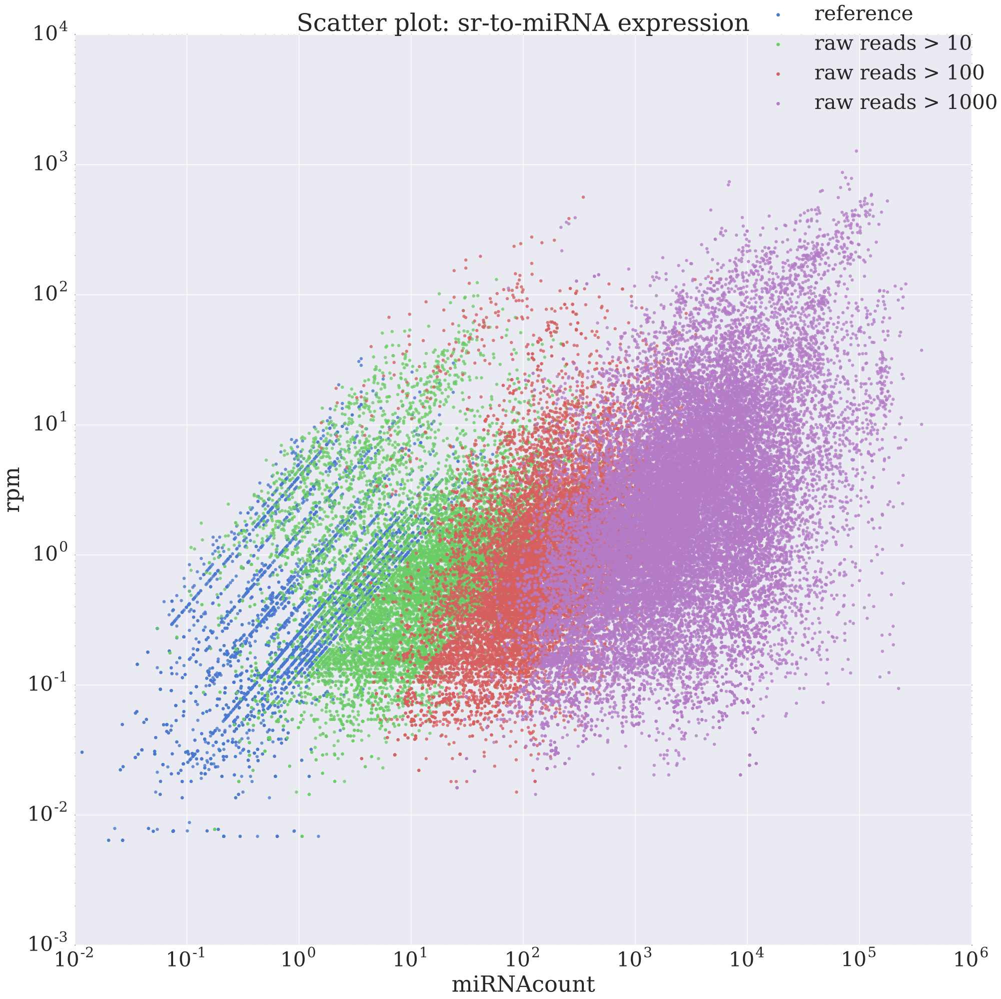


Figure 3.8: Scatterplot of fantom5 with grouping by sum of raw read values

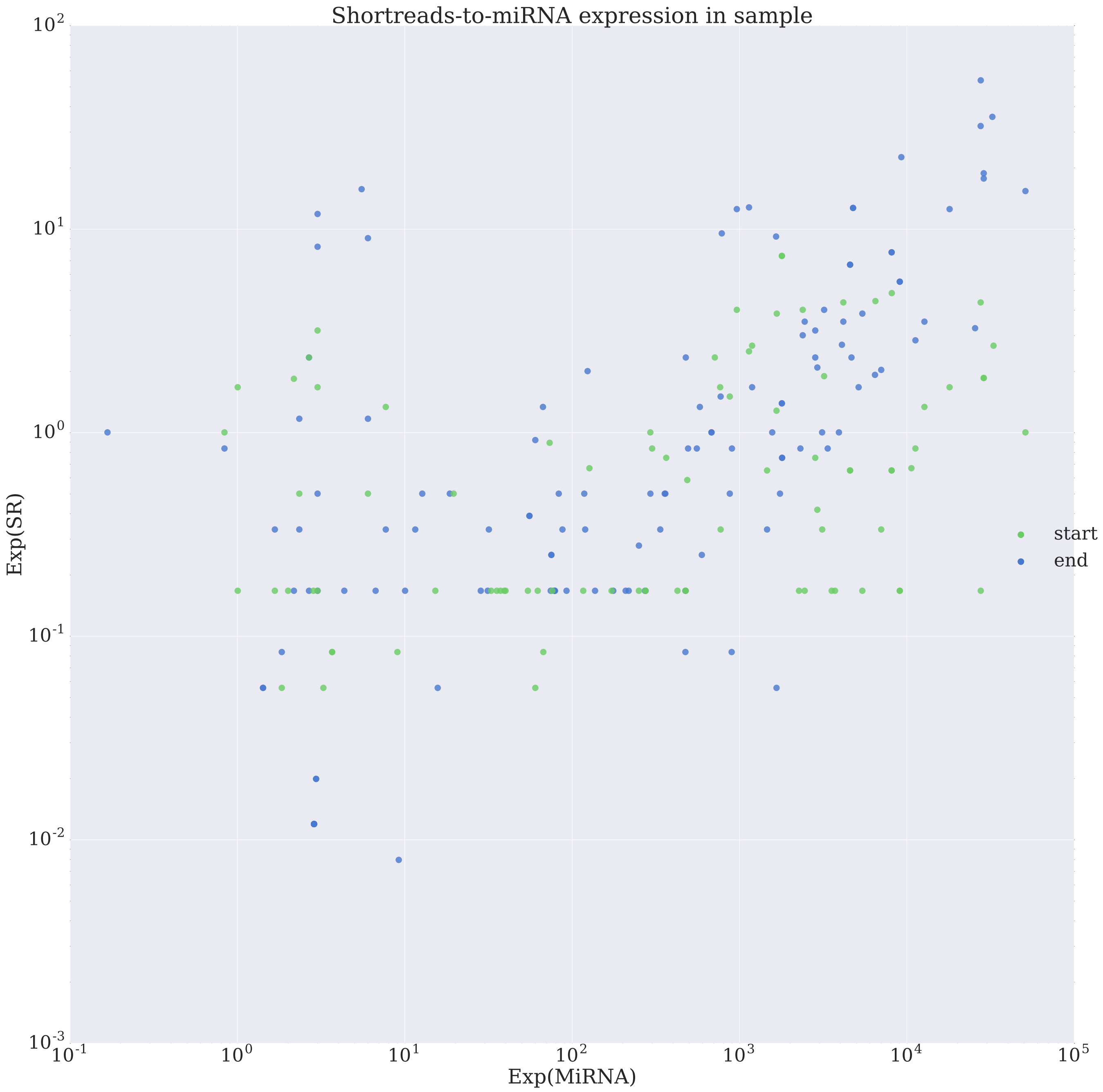


Figure 3.9: Scatterplot of a sample from 4SU

3.2.2 Grouped SR plot

In the report written for preliminary study [9] I explored patterns of presence of shortreads on a specific grouping for miRNAs. That type of plot was implemented in *shores* with some improvements. See fig. 3.10 for the improved plot of that from the preliminary study.

The aim of this analysis is to explore the difference in presence of shortreads in specifically grouped miRNAs. The grouping is done in the following manner: for each miRNA in a sample the expression of guide- and passenger- strands are compared. Based on the relation $\frac{Exp(guide)}{Exp(passenger)}$ the miRNA with it's shortreads are placed either in *Equal* (the difference is within 5 times) or *Different* (expression difference is higher than 15 times) groups.

Further each miRNA was placed into one of four subgroups based on the presence of shortreads: *High* (only guide strand has SR), *Low* (only passenger strand has SR), *Both* (both strands have SR), *None* (neither passenger nor guide strand have shortreads).

Then for each sample in the dataset the number of miRNAs for each group-subgroup pair were counted and divided by the total number of miRNAs in each group. These values are plotted on a box-and-whisker plot.

The improvement done in comparison with the plots in my earlier work is that within each sample miRNA expressions were graded based on the expression values, which enabled exploration of distributions for different percentiles of expression values of miRNAs.

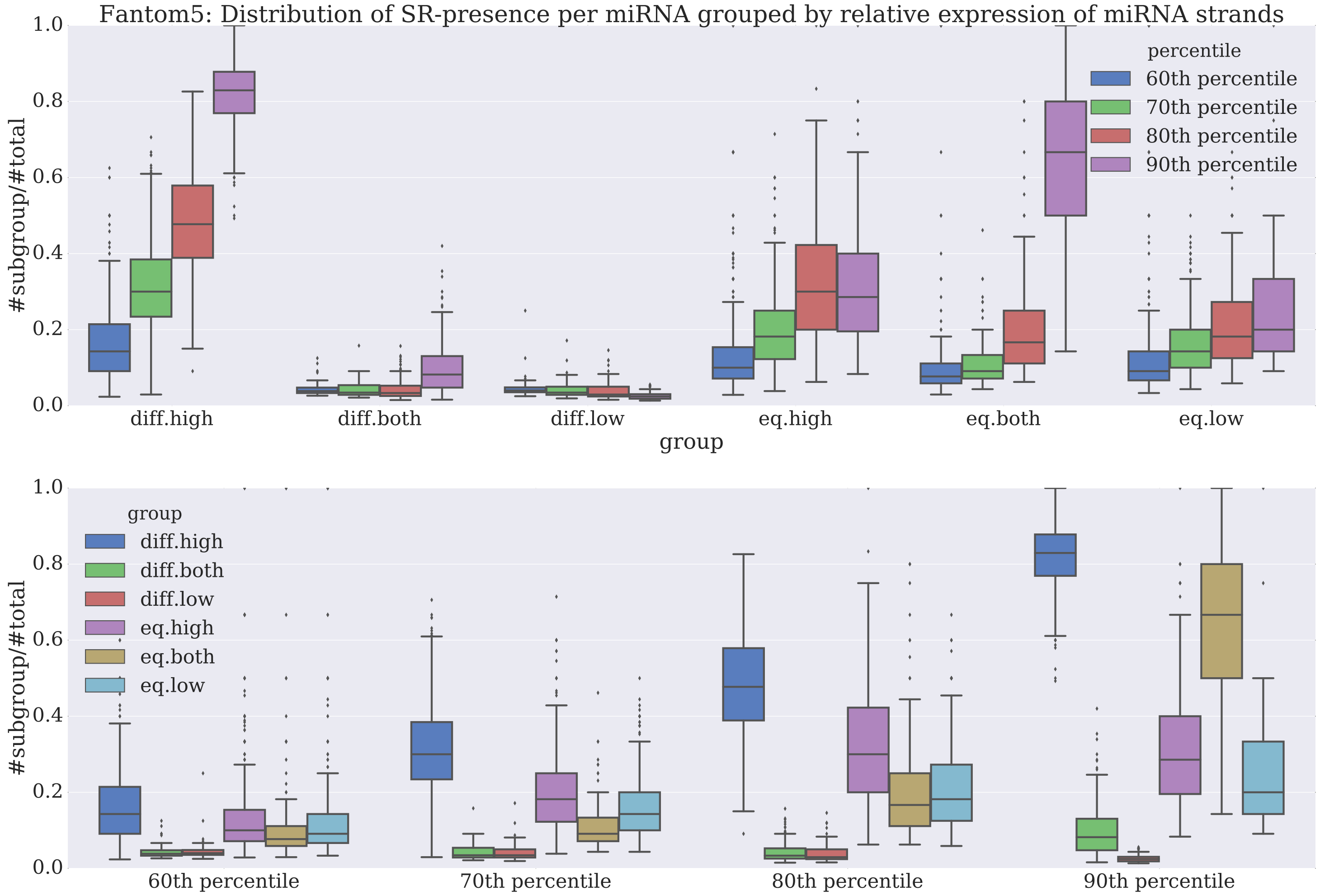


Figure 3.10: MiRNAs divided in groups as discussed in section 3.2.2. Each sample in the dataset produces a single value for each group - how prominent this group was in comparison with other groups (for *Equal* and *Different* all 4 subgroups sum up to 100%). The distribution of values are presented with a box-and-whiskers plot. Each percentile group contains only entries from its base to the base of next percentile group. That is ”70th percentile” group contains only values with levels from 70% to 80% of the maximum.

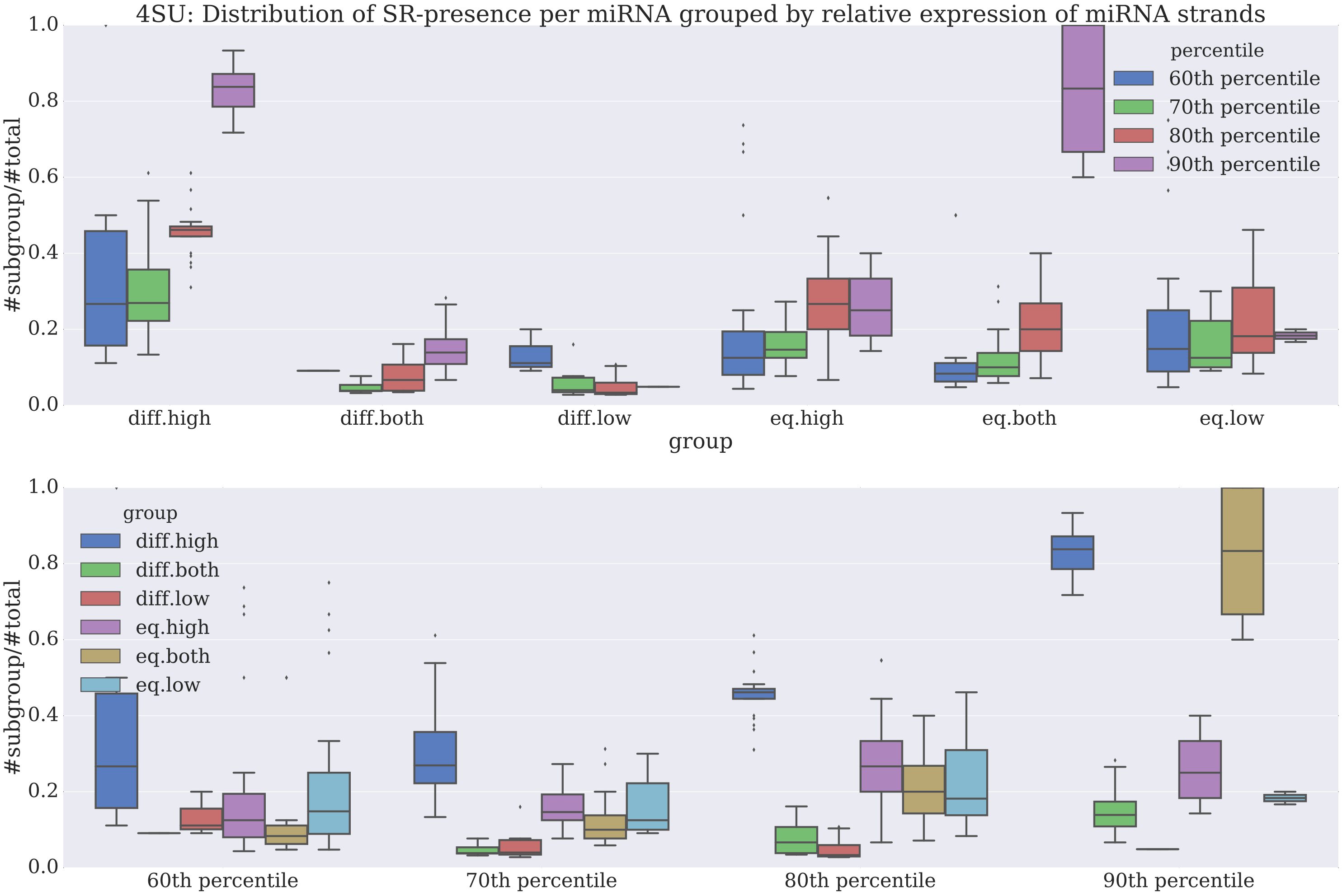


Figure 3.11: Simillar to fig. 3.10 *groupedSR* plot for 4SU dataset.

Bibliography

[1] Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Langmead B, Trapnell C, Pop M, Salzberg SL.* Genome Biol 10:R25.

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

[3] Identifying miRNA short reads as potential markers for biologically active miRNAs. *Wahl, Kristin.* NTNU, 2015

[4] miRBase: annotating high confidence microRNAs using deep sequencing data. *Kozomara A, Griffiths-Jones S.* Nucleic Acids Res. 2014 42:D68-D73

[5] Data Structures for Statistical Computing in Python *Wes McKinney* Proceedings of the 9th Python in Science Conference. 2010.

[6] Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Cock, Peter JA, et al.* Bioinformatics 25.11 2009: 1422-1423.

[7] Pandas, Python Data Analysis Library. 2015. *McKinney, W.* Reference Source (2014).

[8] Shores processing pipeline *Fedor F* https://github.com/fedyaFade/shores

[9] (Unpublished preliminary study) Shortread expression patterns in miRNA *Fedor F* NTNU, 2015

[10] Degradation dynamics of microRNAs revealed by a novel pulse-chase approach *Matteo J. Marzi, Francesco Ghini, Benedetta Cerruti, Stefano de Pretis, Paola Bonetti, Chiara*

Giacomelli, Marcin M. Gorski, Theresia Kress, Mattia Pelizzola, Heiko Muller, Bruno Amati, and Francesco Nicassio Cold Spring Harbor Laboratory Press

[11] MicroRNAs: genomics, biogenesis, mechanism, and function. *Bartel, David P.* cell 116.2 (2004): 281-297.

[12] Most mammalian mRNAs are conserved targets of microRNAs. *Friedman, Robin C., et al.* Genome research 19.1 (2009): 92-105.

[13] Translation: DNA to mRNA to protein *Clancy, Suzanne and Brown, William* Nature Education, 2008

[14] A promoter-level mammalian expression atlas. *Consortium, The FANTOM.* Nature 507.7493: 462-470, 2014

[15] Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Arner, Erik, et al.* Science 347.6225: 1010-1014, 2015

[16] MicroRNAs, the immune system and rheumatic disease. *Tili, Esmerina, et al.* Nature clinical practice Rheumatology 4.10 534-541, 2008

Appendix A

Acronyms

SR Shortread

miRNA MicroRNA

mRNA messenger RNA

RC Read Count

RNA Ribonucleic acid

DNA Deoxyribonucleic acid

AGO Argonaute protein

REPL Read–eval–print loop

Appendix B

GSM entries of 4SU dataset

Experiment	Timepoint	GSM-	Filename(SRR-)
EXP1	0h	GSM1867376	SRR2230026.collapsed.gz
	2h	GSM1867377	SRR2230027.collapsed.gz
	4h	GSM1867378	SRR2230028.collapsed.gz
	6h	GSM1867379	SRR2230029.collapsed.gz
	8h	GSM1867380	SRR2230030.collapsed.gz
	10h	GSM1867381	SRR2230031.collapsed.gz
	12h	GSM1867382	SRR2230032.collapsed.gz
EXP2	0h	GSM1867383	SRR2230033.collapsed.gz
	2h	GSM1867384	SRR2230034.collapsed.gz
	4h	GSM1867385	SRR2230035.collapsed.gz
	6h	GSM1867386	SRR2230036.collapsed.gz
	8h	GSM1867387	SRR2230037.collapsed.gz
	10h	GSM1867388	SRR2230038.collapsed.gz
	12h	GSM1867389	SRR2230039.collapsed.gz
EXP3	0h	GSM1867390	SRR2230040.collapsed.gz
	2h	GSM1867391	SRR2230041.collapsed.gz
	4h	GSM1867392	SRR2230042.collapsed.gz
	6h	GSM1867393	SRR2230043.collapsed.gz
	8h	GSM1867394	SRR2230044.collapsed.gz
	10h	GSM1867395	SRR2230045.collapsed.gz
	12h	GSM1867396	SRR2230046.collapsed.gz

Appendix C

Low RC in Fantom5

This appendix presents examples of numerous entries with low raw readcounts in fantom5 that end up producing the same precise values of SR-to-miRNA expression relationship.

Normalized expression(reads per million) is approximated by the following formula: $Exp = \frac{Reads}{Alignments} * \frac{10^6}{TotalReadsInSample}$

C.1 SR/miRNA ratio = 0.75

Hairpin ID	Read kind	Alignments	Raw readcount	Normalized readcount
hsa-mir-1275	mirna	1	8	14.0487174399
	shortread	1	6	10.5365380799
hsa-mir-4286	mirna	1	4	4.167591351
	shortread	1	3	3.1256935132
hsa-mir-224	mirna	1	4	4.15526134
	shortread	1	3	3.116446005
hsa-mir-665	mirna	1	4	3.9176739986
	shortread	1	3	2.9382554989
hsa-mir-1275	mirna	1	4	2.9186960725
	shortread	1	3	2.1890220544
hsa-mir-7641-1	mirna	2	4	2.8585476291
	shortread	2	3	2.1439107218
hsa-mir-7641-2	mirna	2	4	2.8585476291
	shortread	2	3	2.1439107218
hsa-mir-1275	mirna	1	4	2.8005026902
	shortread	1	3	2.1003770177
hsa-mir-6087	mirna	1	4	2.4616111737
	shortread	1	3	1.8462083803
hsa-mir-7641-1	mirna	2	4	2.4059368899
	shortread	2	3	1.8044526674
	shortread	2	3	1.8044526674
	shortread	2	3	1.8044526674
	mirna	2	4	2.4059368899
	shortread	2	3	1.8044526674
	shortread	2	3	1.8044526674
	shortread	2	3	1.8044526674
hsa-mir-3653	mirna	1	4	2.3302663145
	shortread	1	3	1.7476997359
hsa-mir-7641-2	mirna	1	2	1.744856599
	shortread	2	3	1.3086424492
hsa-mir-4425	mirna	1	4	1.477710583
	shortread	1	3	1.1082829372
hsa-mir-3607	mirna	1	8	1.1902716125
	shortread	1	6	0.8927037094
hsa-mir-4485	mirna	1	4	1.0461298817
	shortread	1	3	0.7845974113
hsa-mir-4508	mirna	1	2	0.8782854463
	shortread	2	3	0.6587140847
hsa-mir-1277	mirna	1	4	0.6403922274
	shortread	1	3	0.4802941706
hsa-mir-3665	mirna	1	1	0.5825665786
	shortread	4	3	0.436924934
	shortread	4	3	0.436924934

C.2 SR/miRNA ratio = 0.5

Hairpin ID	Read kind	Alignments	Raw readcount	Normalized readcount
hsa-mir-33a	mirna	1	8	21.4922037031
	shortread	1	4	10.7461018516
	mirna	1	18	18.6986760299
	shortread	1	9	9.3493380149
hsa-mir-7641-2	mirna	2	2	3.1882468468
	shortread	2	1	1.5941234234
	shortread	2	1	1.5941234234
	shortread	2	1	1.5941234234
	shortread	2	1	1.5941234234
hsa-mir-33a	mirna	1	12	12.0881711202
	shortread	1	6	6.0440855601
hsa-mir-7641-2	mirna	1	6	11.9622232988
	shortread	1	3	5.9811116494
	shortread	1	3	5.9811116494
	shortread	2	6	5.9811116494
hsa-mir-4286	mirna	1	2	11.2391121101
	shortread	1	1	5.6195560551
hsa-mir-132	mirna	1	2	10.4816858744
	shortread	1	1	5.2408429372
hsa-mir-652	mirna	1	2	10.1782724418
	shortread	1	1	5.0891362209
hsa-mir-671	mirna	1	2	10.1782724418
	shortread	1	1	5.0891362209
hsa-mir-33a	mirna	1	2	10.1782724418
	shortread	1	1	5.0891362209
	mirna	1	4	8.4350992917
	shortread	1	2	4.2175496458
hsa-mir-503	mirna	1	2	2.0771192848
	shortread	1	1	1.0385596424
hsa-mir-328	mirna	1	6	7.8646545385
	shortread	1	3	3.9323272692
hsa-mir-23b	mirna	1	2	7.7608720116
	shortread	1	1	3.8804360058
hsa-mir-3609	mirna	1	2	7.7608720116
	shortread	1	1	3.8804360058
hsa-mir-3607	mirna	1	2	7.3237270447
	shortread	1	1	3.6618635223
hsa-mir-185	mirna	1	2	7.1663782199
	shortread	1	1	3.58318911
	shortread	1	1	3.58318911
hsa-mir-92a-1	mirna	1	2	7.0703606945
	shortread	1	1	3.5351803472
hsa-mir-3607	mirna	1	2	7.0703606945
	shortread	1	1	3.5351803472