UNDERSTANDING SMALL RNA BIOGENESIS THROUGH ANALYSIS OF SEQUENCING DATA

Teo Ren Yi
A0140768Y

Undergraduate Research Opportunities in Science
PROJECT REPORT
submitted to the
Department of Biological Sciences
National University of Singapore

LSMX3289
26 March 2018
4793 words

Abstract

sRNA are non-coding RNAs of 20-30 nt that associates with AGO proteins with the potential of regulating gene expression. To understand sRNA biogenesis, we analysed sRNA-seq data involving key proteins in the maturation of miRNA and siRNA, namely the Dicer family proteins and their binding partners loquacious and R2D2 in *Drosophila melanogaster*. We rewrote a sRNA analysis pipeline dedicated for the discovery of miRNA and endo-siRNA from a collection of bash scripts into python and R scripts using the snakemake workflow engine. After updating the annotation details to the current genome assembly curated in FlyBase, we managed to replicate a prior finding where loqs-pd rescue is responsible for a substantial increase of reads mapped to hpRNA, a type of endo-siRNA. To complement our findings, we queried online sRNA-seq datasets of similar studies targeting sRNA biogenesis. We observed different sRNA profile under the various experimental condition and drew biological insights into sRNA biogenesis. The findings suggest the possibility that R2D2 may be involved in endo-siRNA biogenesis, despite being canonically dedicated towards exo-siRNA biogenesis. This study also paves the way for a future project which seeks to provide an even in-depth investigation of sRNA biogenesis including differential expression analysis of sRNA types.

Introduction

Small RNA Biogenesis and Loquacious

Small RNA (sRNA) is a term often used to describe non-coding RNA of approximately 20 – 30 nt in length (Carthew and Sontheimer, 2009), particularly of interest are those that engages the RNA interference (RNAi) machinery (Czech and Hannon, 2010) though interactions with Argonaute proteins (AGO), this primarily includes micro-RNA (miRNA), and small interfering RNA (siRNA) and PIWI-interacting (piRNA) (Lai and Okamura, 2008). While the biogenesis of miRNA is better understood and well characterised, siRNA and piRNA biogenesis appears to be more convoluted as there appear to be a diverse types of siRNA precursors converging on RNA-induced silencing complexes (RISCs) in the cytoplasm (Carthew and Sontheimer, 2009) with the upstream events poorly understood (Lai and Okamura, 2008).

RISC-loading depends on Dicer-processing of precursor of sRNA molecules, in *Drosophila melanogaster* Dicer exists in two forms Dicer-1 (Dcr-1) and Dicer-2 (Dcr-2) (Czech et al., 2009). Canonically, miRNAs are processed by Dcr-1 and loaded onto AGO1 which regulates translational regulation while siRNAs are processed by Dcr-2 associates with AGO2 and engages in RNAi. It is to note that some miRNA particularly the star-strand may be loaded onto AGO2 (Czech et al., 2009) and recent findings in (Daugaard and Hansen, 2017) suggests the increasing complexity of miRNA biogenesis distinct from the canonical model. Unlike AGO1, other AGO proteins loaded with endogenous siRNA (endo-siRNA) or piRNA may possess slicer activity which can causes degradation of target RNA (Lai and Okamura, 2008) and may be responsible for transposon defence (Siomi et al., 2011).

Unlike piRNA which are derived from single-stranded precursors, sRNA species originating from a dsRNA intermediary requires processing by the RNase III enzyme (Dicer) during their biogenesis with a helper protein that containing double stranded RNA binding domain (dsRBD) (Forstemann et al., 2005) & (Saito et al., 2005). In mammals Dicer's partner is the protein TRBP (Carthew and Sontheimer, 2009) while in

D. melanogaster which possess two Dicer proteins, R2D2 was first identified as Dcr-2's partner while Dcr-1's partner was discovered by (Forstemann et al., 2005) & (Saito et al., 2005) to be protein Loquacious (loqs), which was also verified as the equivalent of mammalian TRBP in *Drosophila*.

Recent updates on FlyBase (Tweedie et al., 2009) has identified six potential isoforms for loqs, ranging from isoforms A to F. However little is known about the predicted loqs isoforms C, E and F. loqs-PA and loqs-PB protein consists three dsRBD and associates with Dcr-1 to facilitate miRNA processing while loqs-PD contains of only two dsRBD and appears to be functional paralog of R2D2 interacting with Dcr-2 (Hartig and Förstemann, 2011) & (Haac et al., 2015).

Endo-siRNA and the Role of Loquacious

In the resultant effector RISCs, while the primary role of the short RNA sequence appears to be target recognition, the information encrypted within is much more profound. The sequence is indirectly involved in many key step of RNAi, formation of precursor RNA molecules, thermos-stability of the double-stranded RNA (dsRNA) which is known to affect the strand selection and sorting into AGO (Czech and Hannon, 2010) & (Tants et al., 2017). Although the mechanisms are unclear it is strongly suggested that upstream biogenesis of precursor molecules may prescribe downstream regulatory events of RISC. Furthermore, due to their convergence in downstream silencing pathways and association with the few key proteins, the functional boundaries between sRNA types appears to overlap (Carthew and Sontheimer, 2009) & (Zhou et al., 2008) and cross-talk within different sRNA biogenesis pathways may be occurring(Marques et al., 2010). Thus, the need for characterisation of sRNA types before biological insights can be derived.

siRNA can be further subdivided into two categories endogenous siRNA (endo-siRNA) and exogenous siRNA (exo-siRNA), endo-siRNA arise from precursors originating from a variety of sources within the genome, notably the transposable elements and repetitive elements while exo-siRNA matures from with precursors that are experimentally introduced

or of viral origin (Haac et al., 2015). Studies by (Hartig and Förstemann, 2011) suggest that loqs-PD is the preferred binding partner of Dcr-2 in endo-siRNA biogenesis while Dcr-2 partners with R2D2 during exo-siRNA biogenesis, though their biogenesis occurs independently, they might compete for Dcr-2 binding. However, (Marques et al., 2010) argues that loqs-PD is necessary for processing of dsRNA into mature siRNA while R2D2 is required for the siRNA-loading into AGO2 with both protein working in a sequential manner during RNAi. The interrelatedness of loqs-PD and R2D2 is further confirmed with deep sequencing studies of sRNA by (Tants et al., 2017) suggesting that loqs-PD can replace R2D2 as a molecular sensor during the strand selection prior to AGO-loading for both endo-siRNA and exo-siRNA.

Interrogating the Role of Loquacious on siRNA Biogenesis with Small-RNA-Seq Data

Many of the earlier findings (Czech et al., 2008); (Ghildiyal et al., 2008) and (Marques et al., 2010) on the interaction between loquacious and siRNA biogenesis relies on the size distribution of sRNA and the probing of specific genes in northern blots to assay the alteration of different sRNA biogenesis pathways depending on the experimental conditions. In our study, we instead characterise sRNA classes from gene annotations details from curated databases and relied on sequence alignment of small RNA-sequencing (sRNA-seq) data to observe changes of sRNA profile under different experimental conditions when key proteins involved in sRNA biogenesis are altered.

From drawing reference to a pre-existing small RNA-sequencing (sRNA-seq) pipeline used in (Lim et al., 2016), we rewrote an analysis pipeline using snakemake, a bioinformatics workflow engine while updating the annotation details. Additionally, we adapted parts of the script and performed retrospective analysis on published sRNA-seq datasets from GEO. From our primary dataset we affirmed the initial finding that loqs-PD rescue was associated with a most prominent increase of long hairpin RNA (hpRNA) reads compared to other loqs-isoforms. Furthermore, by comparing the results of three other datasets analysed with

a similar apporach, we were able to glean further insights into the interaction between loqs-PD and R2D2 in endo-siRNA biogenesis.

Materials and Methods

Small RNA Sequencing

Primary sRNA Libraries

The design of the analysis pipeline centred around a pre-existing pipeline by Dr Chak Li Ling described in (Lim et al., 2016). Similarly, the primary dataset is based on the same set of experiment where deep sequencing was performed on a stable loqs-null mutant *D. melanogaster* cell line with varying loqs-isoforms knocked-in.

GEO datasets

The following datasets GSE26230, GSE17171 and GSE37443 described in this paper were obtained from the Gene Express Omnibus (GEO) (Barrett et al., 2013). Datasets were first pre-processed in the into a suitable format for our designed pipeline with basic text processing and bioinformatics tools. As format of the sRNA-seq data in the public repository varied across datasets, each dataset was processed with a custom script adopting a similar workflow from our main pipeline.

A Snakemake Based Analysis Pipeline

As the project was in part intended to be learning experience to create bioinformatics workflow for beginners, we were tasked to rewrite as well as update the relevant annotation details of the pre-existing bash scripts using Snakemake (Köster and Rahmann, 2012), a workflow engine dedicated to aid the design of bioinformatics pipeline. With the flexibility of Snakemake, we wrote our analysis pipeline with a combination of python (Python

Software Foundation, n.d.) and bash (Free Software Foundation, 2007) which offered us access to a wide support of libraries and packages from both platforms. We replaced parts of the original scripts with simple loops and reduced generation of temporary files and improved the overall readability of the script using Snakemake rules. Additionally, Snakemake allowed for the incorporation of R (R Core Team, 2013) scripts into its workflow, which expanded the access to the packages provided by the R community, this was especially helpful for downstream data analysis.

After several revisions to streamline the workflow, we have segregated the scripts into three major process namely, Index generation, Sequential mapping and downstream analysis. A list of the packages used along with their version detail will be added to the appendix while the scripts can be found on https://github.com/RY-T/Empty_loqsproject.

Index Generation Script

The index generation script is designed to extract gene annotation details from a variety of curated sources and compile them into bowtie indexes used for read-mapping of the sRNA-seq data.

To prepare the mapping indexes, annotation details were extracted primarily from FlyBase (Tweedie et al., 2009) in the form of Gene Feature Format (.gff) from the latest release of *D. melanogaster* reference genome (Dm6) (Santos et al., 2015) while annotation details for repetitive elements were extracted from the output of RepeatMasker (Smit, AFA, Hubley, R & Green, 2015) by repeat-masking using Dm6 FlyBase reference genome using the giri Repbase library (Bao et al., 2015). The feature specific files were then converted into .bed extension with the help of tools like gffutils (Dale, 2013), gff2bed (Neph et al., 2012) bedtools (Quinlan and Hall, 2010). They were subsequently converted to .fasta format using bedtools-getfasta with the Dm6 genome. Through the process, tools such as the python pandas library (McKinney, 2010) and basic UNIX commands were used for text processing and file format conversion. EDirect from NCBI (Kans, 2008) was used to query and download the fasta sequence for *Drosophila* rRNA (M21017_1.fasta) from

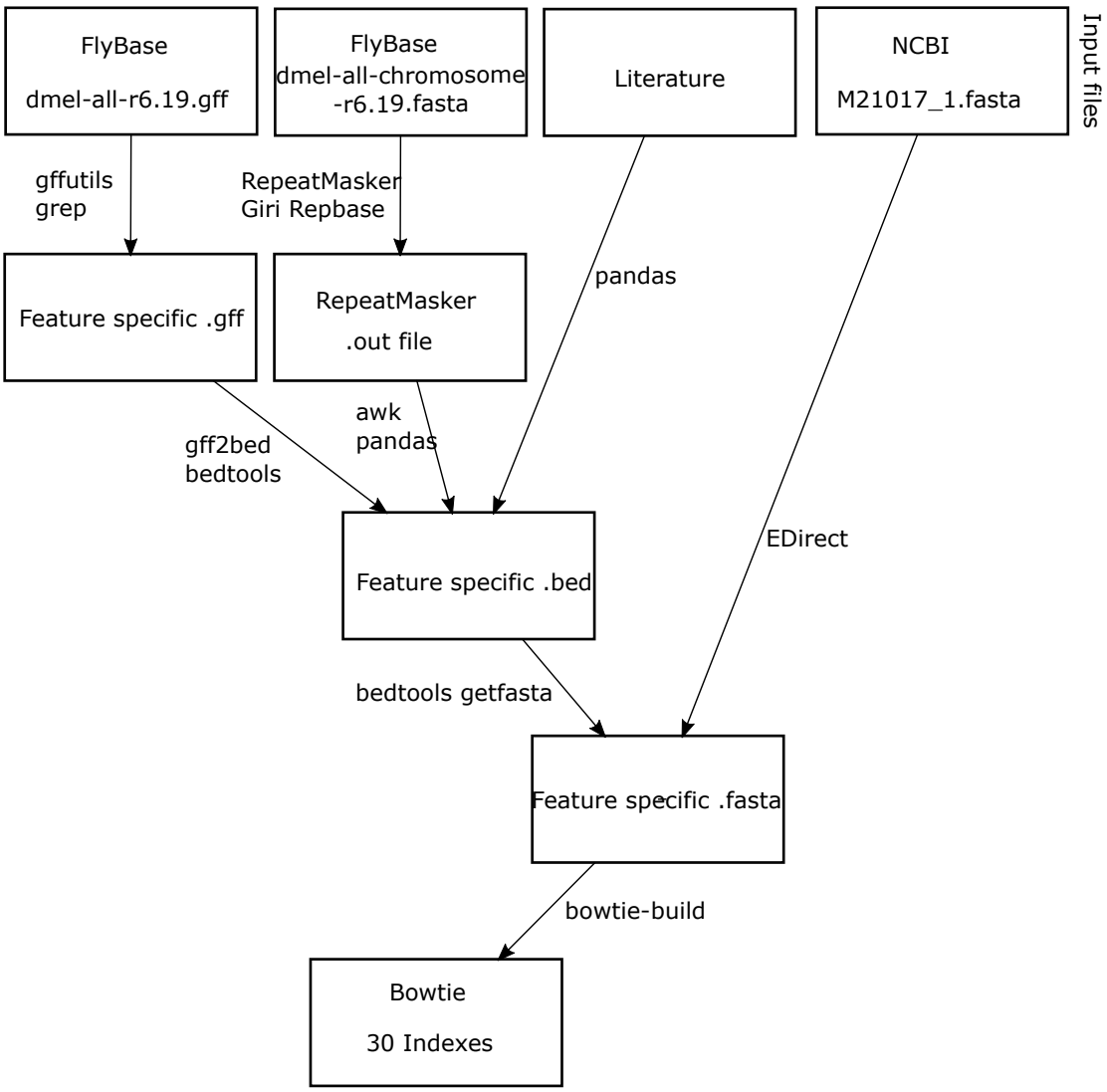


Figure 1: Schematic diagram of the Index generation script

NCBI database. Once the feature specific bed files were generated, they were converted to the fasta format and bowtie-build (Langmead et al., 2009) was then used to generate the indexes.

Table 1 summarises the preparation of bowtie indexes, indexes that had a 'overlap' indicated were prepared using bedtools-merge of the corresponding feature types. For the last four indexes, a second round of bed-tools merge with the anti-sense matching option enabled was applied to pre-existing bed files of the corresponding feature types that were derived from the previous steps, Index 26 to 30 forecasts potential cis-NATs regions in the genome by searching of genomic overlaps of opposing strands against across regions where transcription occurs. In indexes 21 to 25, information from the literature regarding chromosomal location of the gene loci were extracted and converted into bed format and processed along with the pipeline with the other .bed extension files.

Sequential Mapping Script

For quality control of the sRNA libraries, FastQC (Babraham Bioinformatics, 2018) was performed on fastq format files. Subsequently, pre-processing including of adapter trimming, quality filtering and collapsing of identical sequences were performed on the sRNA libraries using the collection of command line tool from Hannon's Lab (Hannon's Lab, 2018). In addition to text-processing and formatting as before, the pandas library was also used for tabulating the output of the read mapping to generate the gene feature count files from the collapsed reads for downstream data analysis.

To initiate the sequential mapping, the pre-processed reads were first mapped to Index0 bowtie index (Dm6 fasta sequence) using bowtie with the unmapped reads counted and discarded. Reads that mapped to Dm6 was then used to map against Index1, this time the reads mapped would escape the cycle while reads that are unmapped will used as starting point for the read mapping of the next bowtie index, this same process then proceeds sequentially from Index1 to Index30. All bowtie indexing was performed with the option -v 0 –best prescribing for zero mismatches and if multiple mappings was present one

Table 1: Description of Bowtie indexes

Bowtie Indexes	sRNA Category	Source	Description
Index0	Dm6 genome	FlyBase dmel-all-chromosome-r6.19.fasta	Dm6 genome
Index1	background RNAs	RepeatMasker	rRNA
Index2	background RNAs	Flybase dmel-all-r6.19.gff	exon-rRNA overlap
Index3	background RNAs	NCBI M21017.1	Ribosomal DNA
Index4	background RNAs	RepeatMasker	RNA
Index5	background RNAs	Flybase dmel-all-r6.19.gff	exon-tRNA overlap
Index6	background RNAs	Flybase dmel-all-r6.19.gff	exon-snRNA overlap
Index7	background RNAs	Flybase dmel-all-r6.19.gff	exon-snoRNA overlap
Index8	mirprecursor	Flybase dmel-all-r6.19.gff	pre-miRNA
Index9	hpRNA	Flybase dmel-all-r6.19.gff	hpRNA
Index10	retrotransposon	RepeatMasker	LTR
Index11	retrotransposon	RepeatMasker	LINE
Index12	DNA transposon	RepeatMasker	DNA
Index13	other repeats	RepeatMasker	Satellite
Index14	other repeats	RepeatMasker	Low_complexity
Index15	other repeats	RepeatMasker	Rolling Circle
Index16	other repeats	RepeatMasker	Simple_repeat
Index17	other repeats	RepeatMasker	Other
Index18	other repeats	RepeatMasker	Unknown
Index19	other repeats	Flybase dmel-all-r6.19.gff	Transposable Elements
Index20	other repeats	RepeatMasker	Artefact
Index21	other known siRNA loci	(Okamura et al., 2008) Nature Structural & Molecular Biology,	Reported siRNA loci
Index22	other known siRNA loci	(Czech et al.,2008) Nature	Reported siRNA loci
Index23	other known siRNA loci	(Kawamura et al., 2008) Nature	Reported siRNA loci
Index24	other known siRNA loci	(Ghildiyal et al., 2008) Science	Reported siRNA loci
Index25	other known siRNA loci	FlyBase	CR14033
Index26	new cisNAT loci	Flybase dmel-all-r6.19.gff	mRNA exon-mRNA exon overlap
Index27	new cisNAT loci	Flybase dmel-all-r6.19.gff	mRNA-exon ncNA exon overlap
Index28	new cisNAT loci	Flybase dmel-all-r6.19.gff	mRNA-exon tRNA exon overlap
Index29	new cisNAT loci	Flybase dmel-all-r6.19.gff	mRNA-exon snoRNA exon overlap
Index30	new cisNAT loci	Flybase dmel-all-r6.19.gff	mRNA-pseudogene exon overlap

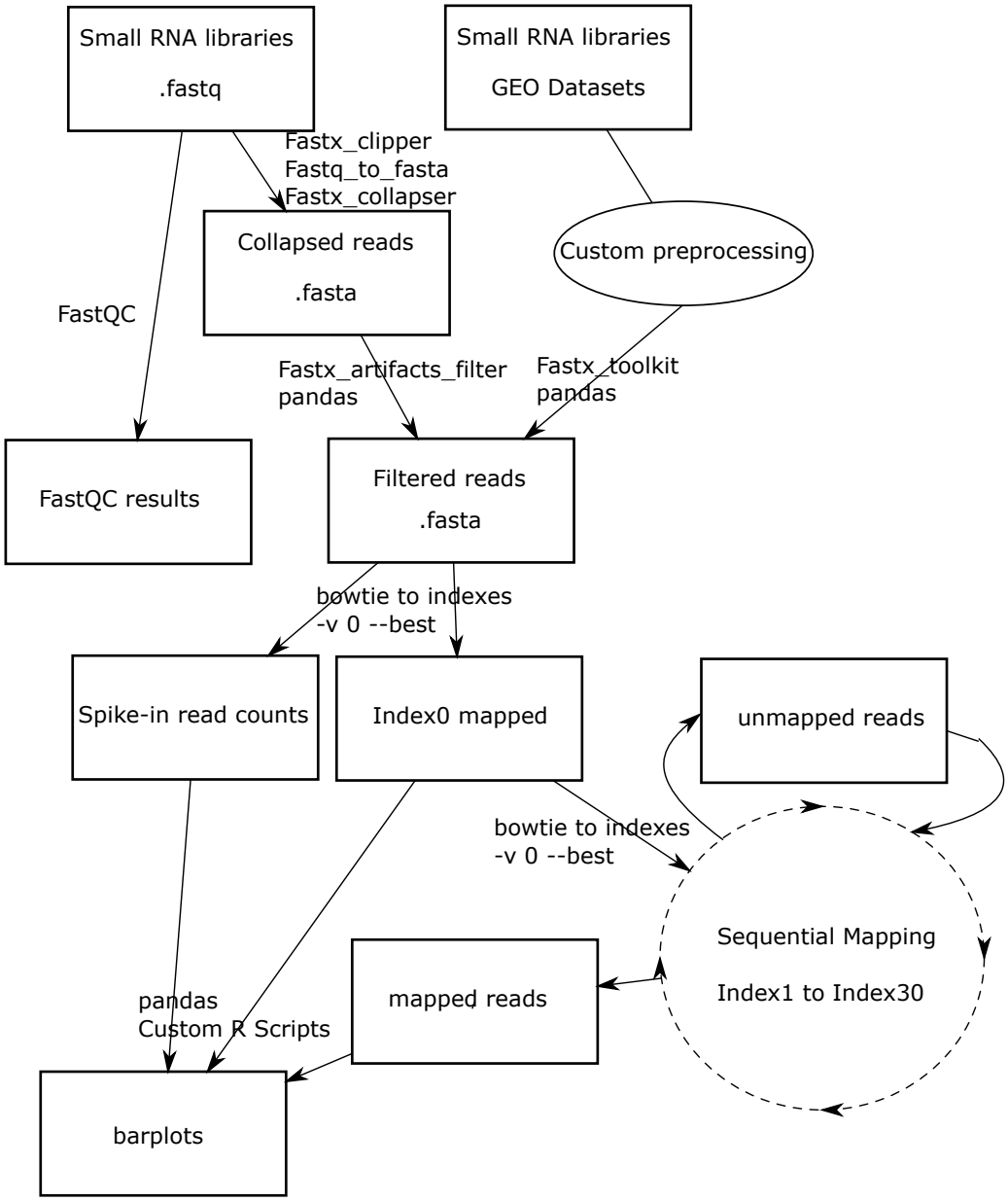


Figure 2: Schematic diagram of the Sequential mapping script

location with the best alignment score would be chosen at random.

Downstream Analysis using R Scripts

Graphical outputs were generated using the Tidyverse packages (Wickham, 2017) and cowplot (R Core Team, 2013) on top of basic R functions within snakemake. For our preliminary analysis of differential gene expression using edgeR(Robinson et al., 2010) on the various sRNA categories (Law et al., 2016)'s script on RNA-seq analysis was adapted and repurposed for sRNA-seq.

Results

Replicating Initial Findings of loqs-PD and hpRNA

The primary motivation of this project was prompted by a curiosity uncovered in preliminary analysis of deep sequencing results in (Lim et al., 2016) where a subclass of siRNA, long hairpin RNA (hpRNA) demonstrated significantly higher fold change in the loqs-PD rescue sRNA libraries when compared to other loqs-isoforms rescue in the same dataset. Thus, we sought to replicate this finding with updated annotation details and with a more concise workflow.

Though our analysis pipeline we are able to replicate the initial finding, whereby when specific sRNA types broken down into different categories, the mean fold change of sRNA reads mapped to hpRNA was markedly higher in loqs-PD rescue compared to the rest. In addition, loqs-PD libraries also showed slight increase of reads mapped to the genes related to 'DNA Transposon' compared across other loqs-isoforms. This finding along with other technical verifications that was conducted throughout the process while reconstructing the analysis pipeline seem to indicate that we were able to replicate the parts of the analysis pipeline by Dr Chak.

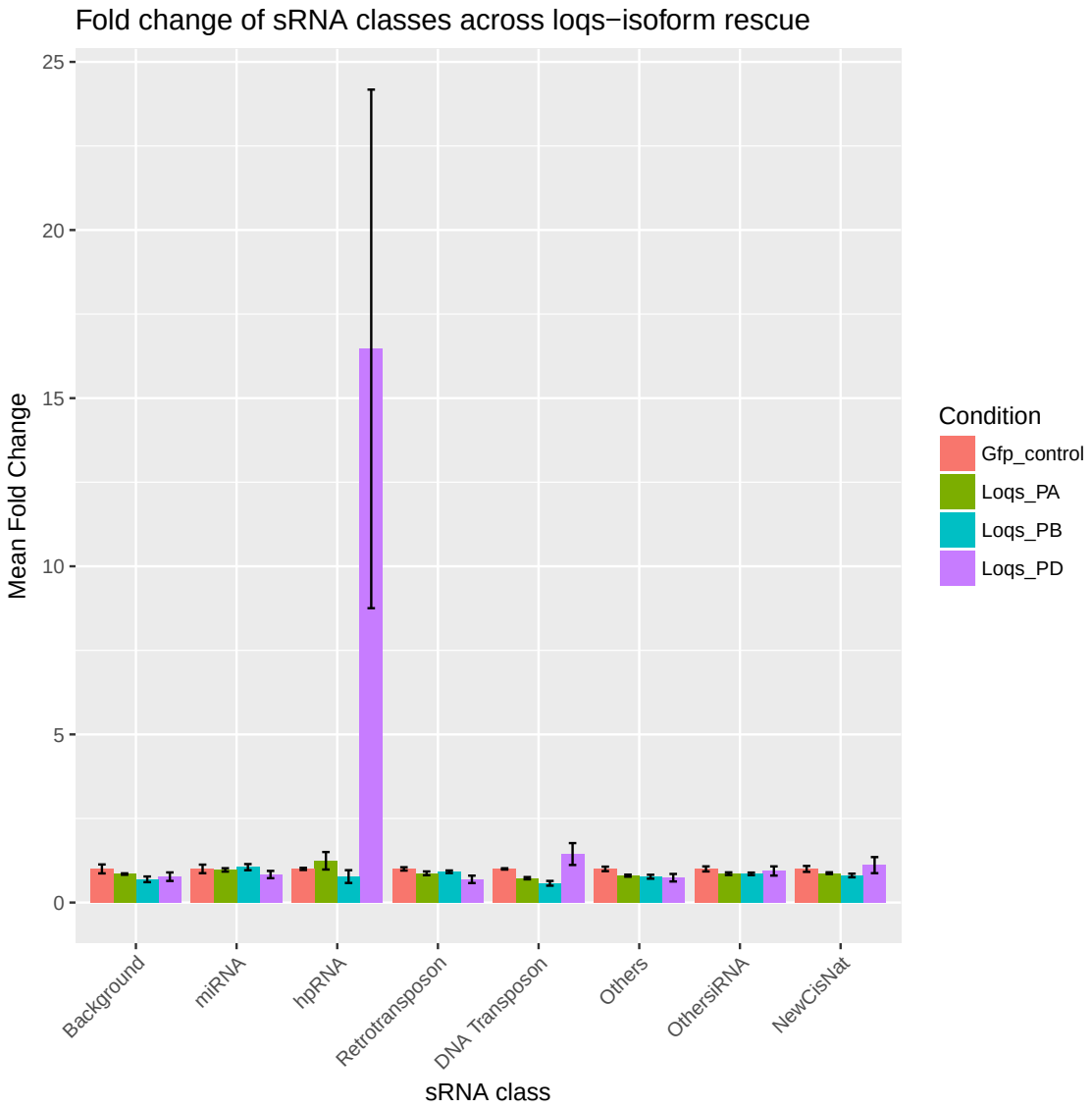


Figure 3: Mean fold change analysis of loqs isoform-specific rescue on different classes of small RNAs

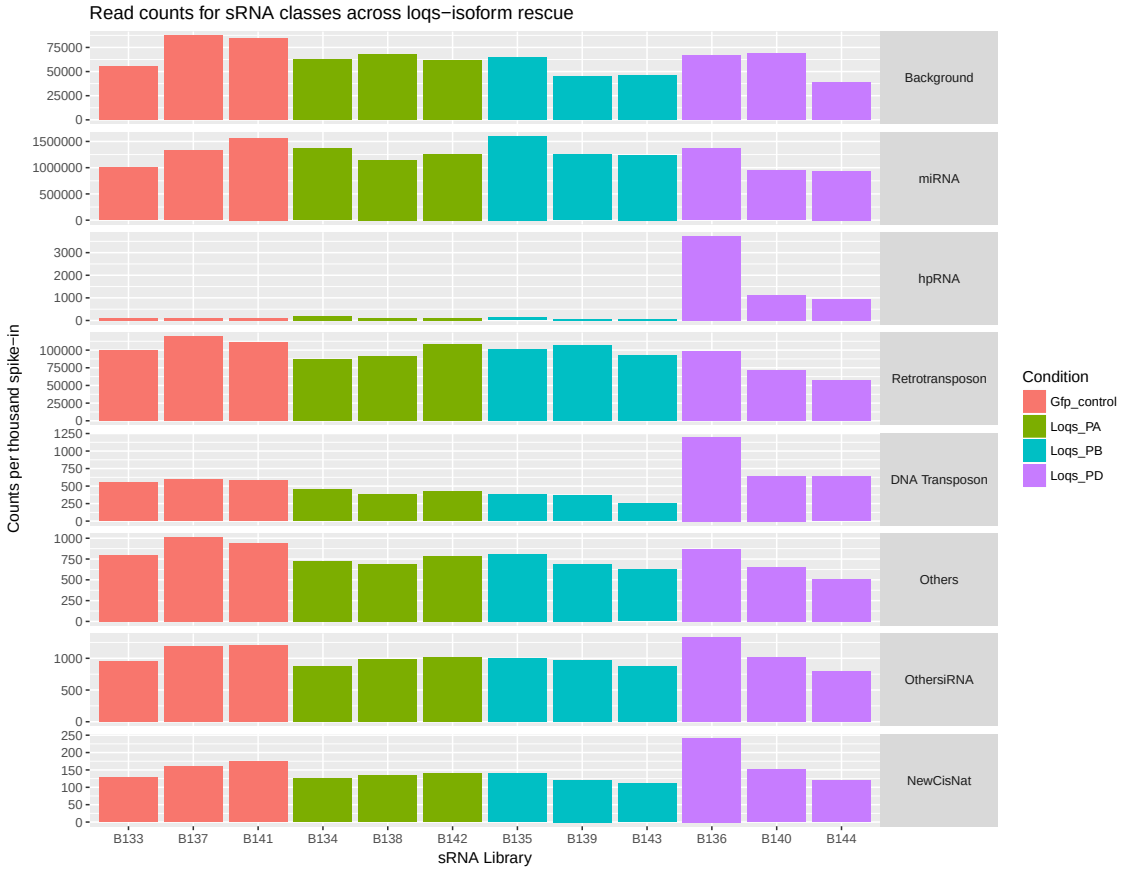


Figure 4: Per thousand spike-in read counts of loqs isoform-specific rescue on our primary sRNA libraries

Furthering the Pipeline

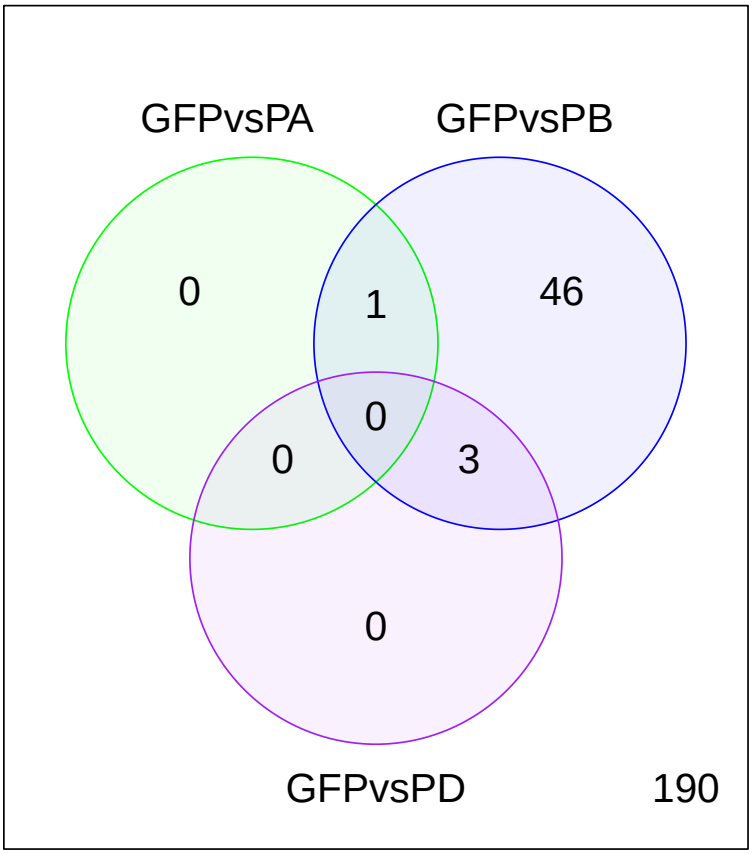


Figure 5: Preliminary DGE analysis on isoform-specific rescue on features counts mapped to miRNA index presented as a venn diagram

As part of an ongoing work, we have also embarked on ways to dissect the dataset even further. As a proof of concept, figure 5 and 6 exhibits the notion that we are able to utilise some of the output files of the analysis pipeline to perform differential gene expression (DGE) experiments for a specific class of sRNA. This opens the exciting possibility of using a wide array of bioinformatic tools available in the Bioconductor(Huber et al., 2015) and other packages in R for downstream analysis. While there exist sRNA-seq analysis pipeline that capable of performing DGE in the literature, most are biased towards the discovery of miRNA but we believe that studying DGE of other sRNA types may hold

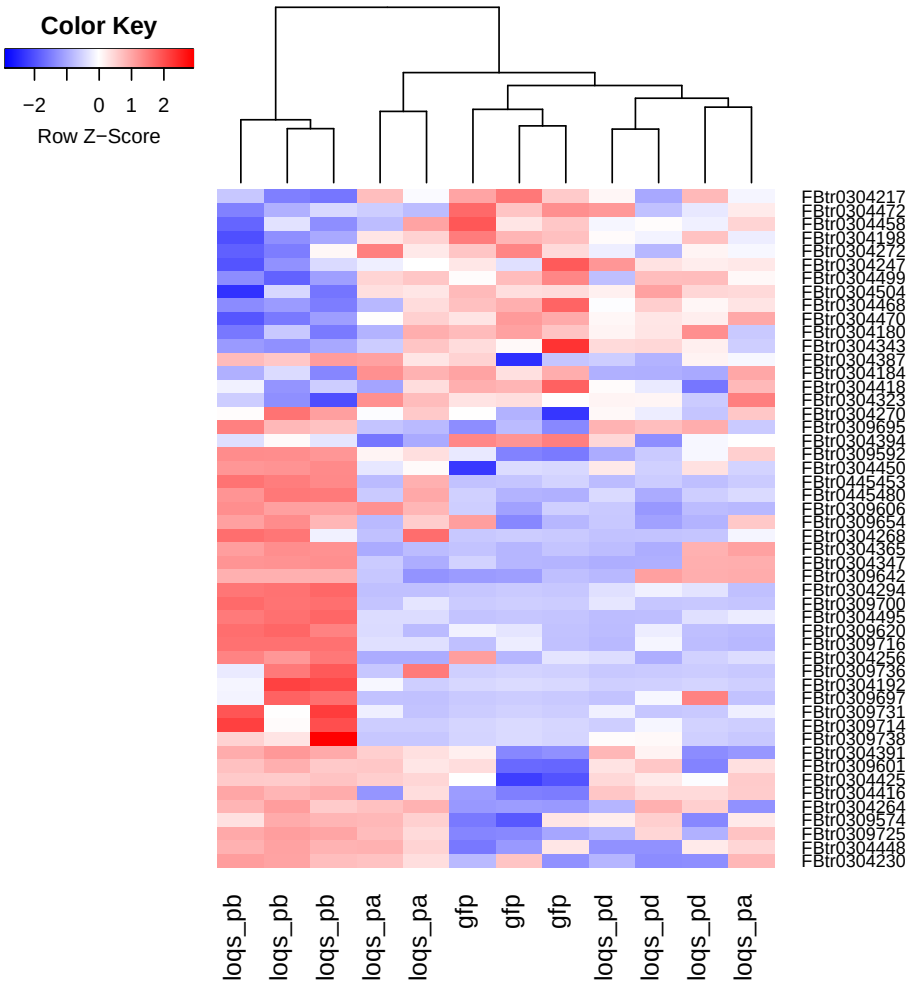


Figure 6: Preliminary DGE analysis on isoform-specific rescue on features counts mapped to miRNA index presented in a heatmap

important biological insights.

Sequential Mapping and definition of sRNA Classes

Most of the figures in this report reference upon the categorisation of sRNA types into various categories according to Table 1. Which was used in the analysis of (Lim et al., 2016). Unlike conventional sequencing where reads are mapped onto a single pre-built index, read mapping was arranged in the manner that starting from index 1, only reads that do not map to the current index it will continue to be cycled through the remaining indexes, imposing a filtering effect on the reads as they proceed down the different categories.

The first category is the background RNAs, consisting of rRNA, tRNA, snRNA and snoRNA as these small stretches of RNA molecules are naturally found in abundance in the cell. Although they and are mostly non specific to sRNA biogenesis, they may show up as reads during the size-selection process. However, it should be noted that ribosomal DNA and other repetitive elements may harbour sites for miRNA (Chak et al., 2015). The next category is miRNA, where the majority of the sequencing reads would be mapped to, followed by hpRNA which will be described in detail in later section. Subsequently, retrotransposons and DNA transposon and 'other repeats' are annotated by giri Repbase using RepeatMasker. While repetitive elements are usually left out of sequencing analysis, Transposable elements (TE) are both associated with piRNA (Siomi et al., 2011) and siRNA (Lai and Okamura, 2008) biogenesis and thus were included. The final two categories were arbitrarily conceived, the 'other siRNA' category defines the region where known siRNA have been previously described in the literature while 'new cisNAT' describes all potential cis-NAT junctions. Cis-natural antisense transcripts siRNA (cis-NATs siRNA) are loqs-dependent siRNA originating from gene loci where transcription occurs both on the top and bottom strand (Okamura, Balla, et al., 2008). Due to the way the sequential mapping was structured, 'other siRNA' would capture endo-siRNA that has been previously described but has yet to be integrated into the official FlyBase database while reads mapped to the 'new cisNAT' index are potentially novel

cisNAT-siRNAs.

Online GEOdatasets

A secondary objective of the project was to utilise this analysis pipeline on online repositories like GEO to pool publicly available sRNA-seq data to form a coherent narrative with our original finding. However, as it is not a common practice to include spike in oligonucleotides for normalisation, the graphical output of the online datasets is normalised to counts per million (cpm) instead counts per thousand spike-in. cpm reflects the proportion of certain sRNA class when it is compared to the total number of reads mapped to the fly genome.

GSE37443

An in-depth analysis of loqs-isoform rescue was performed by (Fukunaga et al., 2012) towards quantification of miRNA and siRNA products on flies carrying the GMR-wIR transgene, the sRNA libraries includes results from fly ovaries and fly heads in loqs-ko flies, consisting of different combinations of loqs-isoform, knocked-in. In fly ovaries the different experimental conditions were dedicated towards measuring the effect of loqs-isoform rescue on maintenance germline stem cell (GSC) while in the fly heads the effects of loqs-isoform rescue were observed through assaying a white eye phenotype. The GMR-wIR transgene encodes an inverted repeat hairpin RNA of the *white* gene, which when processed into mature siRNA prevents the formation of red pigments in the eye, this appears to act as a proxy for an indicator of hpRNA biogenesis.

From their analysis, they determined that loqs-PB rescue was necessary for the maintenance of GSC while loqs-PD rescue was necessary to observe the white eye phenotype in loqs-ko flies. Our analysis of their sequencing data demonstrated that the various knock in conditions that resulted in a white eye phenotype indeed corresponds with elevated levels of hpRNA and endo-siRNA. However, there seem to be some discrepancy against the sRNA

profile of the fly head and the fly ovaries, suggesting heterogenous expression amongst the different tissue types.

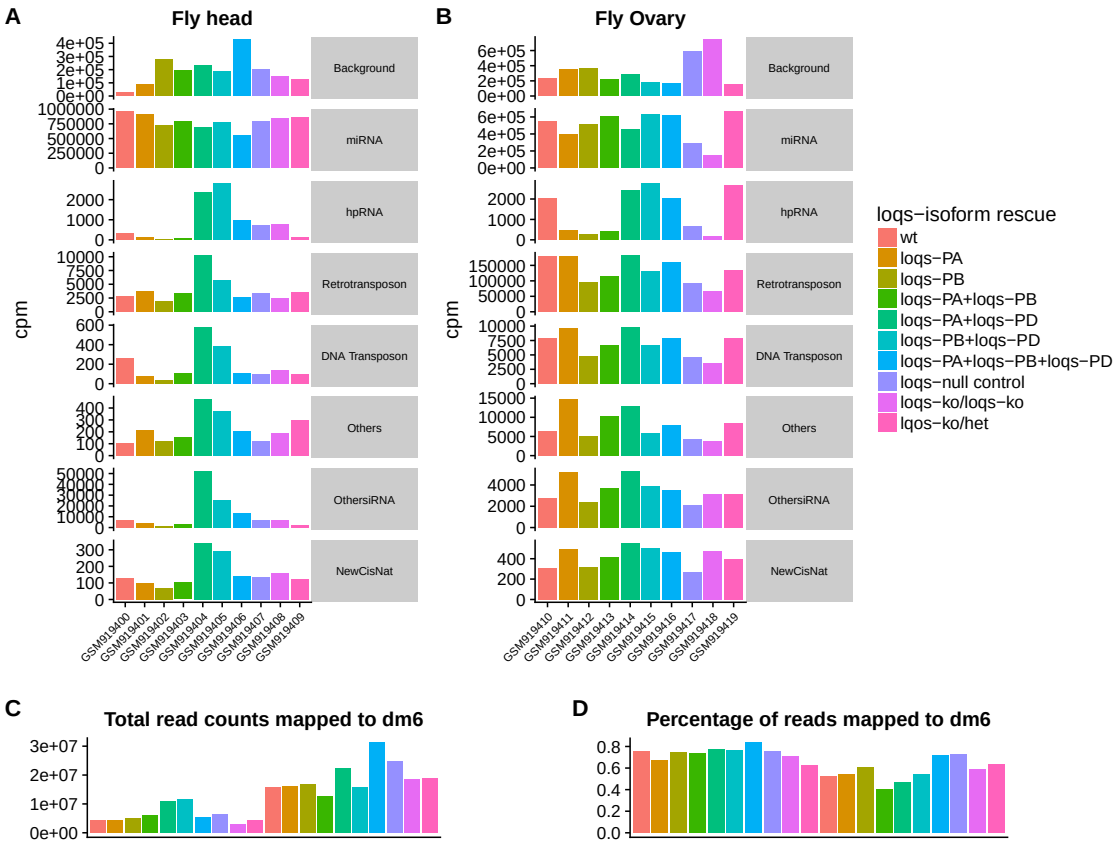
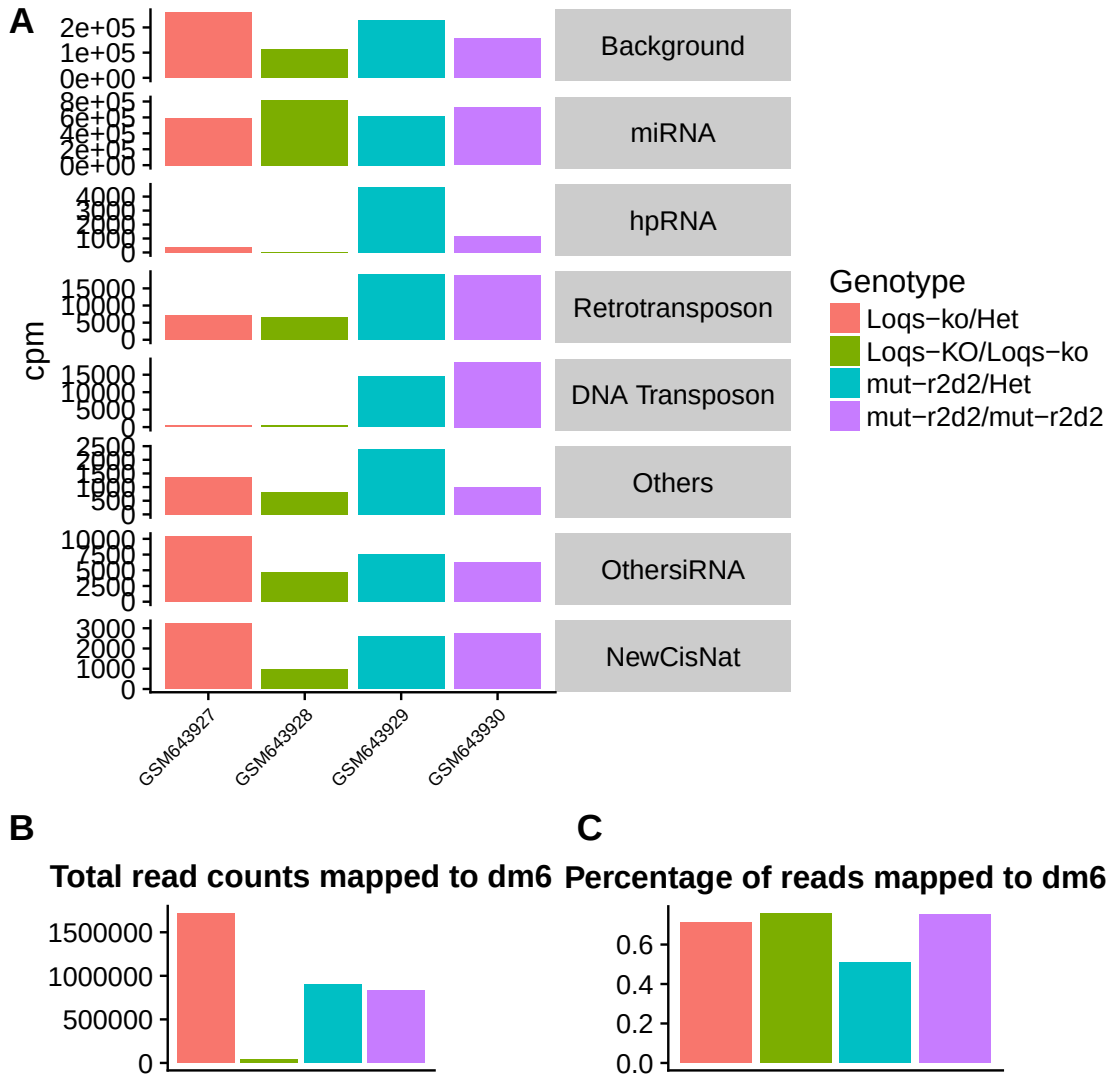


Figure 7: A) Graphical output of counts per million for GSE37443 fly head, B) Graphical output of counts per million for GSE37443 fly ovary, C) Total number of reads mapped to dm6, D) Percentage of reads mapped to dm6

GSE26230

In (Hartig and Förstemann, 2011) a series of sRNA-seq was performed on heads and thorax samples of *D. melanogaster* which were either heterozygous or double mutants of either loqs or R2D2. To restore miRNA biogenesis of the loqs-knockout (loqs-ko) flies and ensure their viability, transgene expression of loqs-PB was permitted for the loqs-ko samples. From their analysis, as reads mapping to either transposons or CG4068 (associated with hpRNA) were present in double knockouts of loqs-PD or double mutant R2D2, they rejected the notion that R2D2 and loqs-PD operates in a sequential manner and that absence of either does not definitively halt endo-siRNA biogenesis.



In our analysis pipeline, compared to (Hartig and Förstemann, 2011) instead of relying on a specific gene feature we have expanded the collection of endo-siRNA. From which, it can be observed that there is a more pronounced change amongst the endo-siRNA classes across the dataset, which will be described in the later section. However, it is to be noted that due to the low total read counts in the double loqs-ko library in this dataset, the resultant sRNA profile may be distorted.

GSE17171

GSE17171 from (Zhou et al., 2009) is a collection of sRNA libraries generated from dsRNA knockdown of specific proteins related to sRNA biogenesis performed on a commonly

used *D. melanogaster* cell line. From their analysis, along with verification using northern blots they concluded that miRNA biogenesis is affected by Dcr-1 knockdown, while Dcr-2 and loqs knockdown resulted in substantial reduction of reads mapped to repeats and transposon regions as well as a moderate reduction of *klarsicht* loci, a known cis-NAT siRNA region. The main changes in sRNA profile observed by (Zhou et al., 2009) was recapitulated in Figure 9 through our workflow.

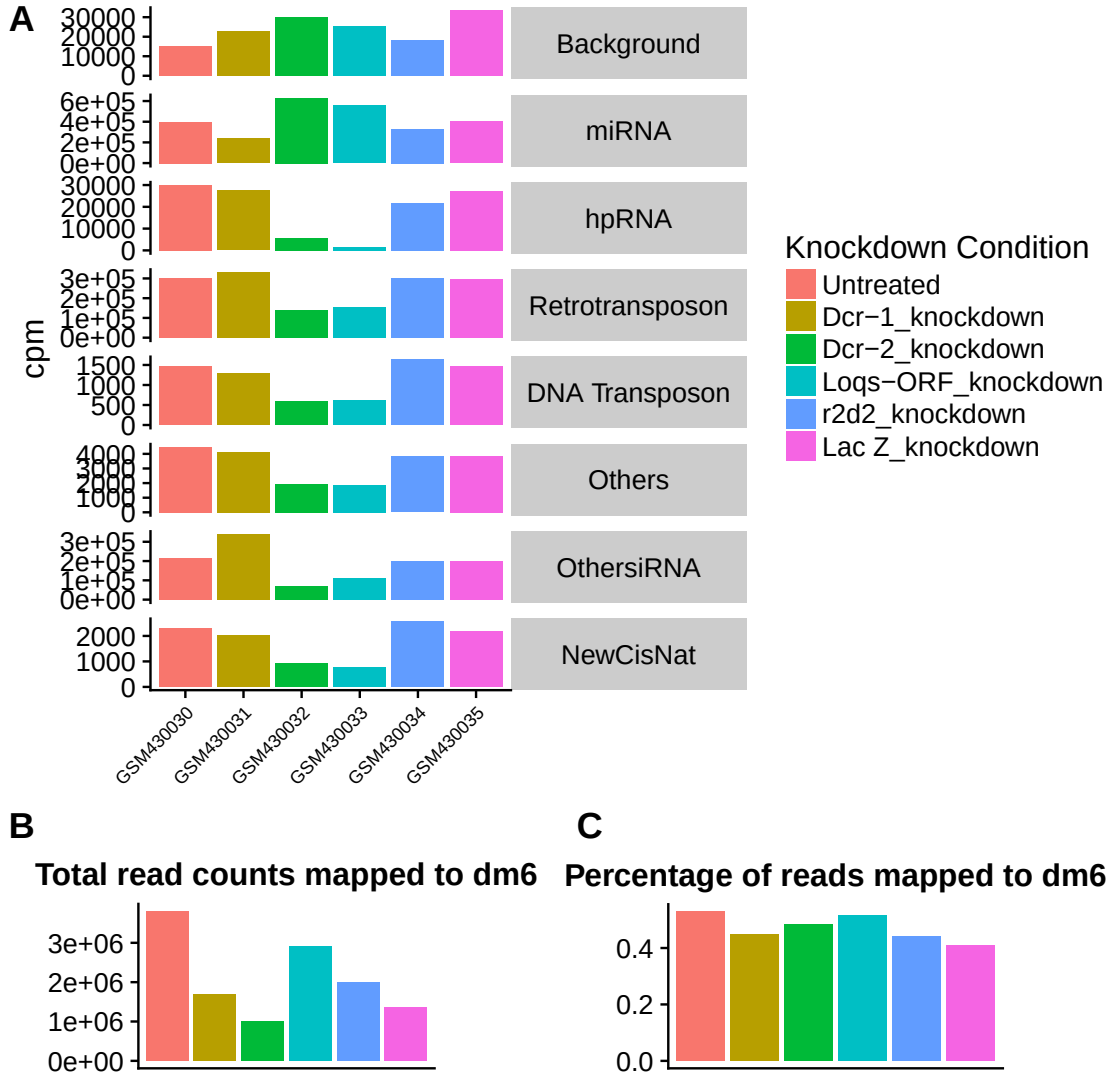


Figure 8: A) Graphical output of counts per million for GSE17171, B) Total number of reads mapped to dm6, C) Percentage of reads mapped to dm6

Discussion

Key findings

hpRNA Biogenesis is Profoundly Affected by loqs-PD

hpRNA is a class of sRNA described in (Okamura, Chung, et al., 2008), it hpRNA are distinct from miRNA originating from longer precursor molecules, it is the first description of an endogenous RNAi in *Drosophila* and engages the AGO2 machinery instead of AGO1. Further described in (Okamura, Chung, et al., 2008), the known binding partner of Dcr-1, loquacious, interferes with Dcr-2 processing of hpRNA when previously only the known partner of Dcr-2 is R2D2.

Figure 3 and 4 suggest that hpRNA biogenesis which is seemingly absent in the loqs-null mutant stable cell line and is either restored or enhanced when loqs-PD isoform is specifically rescued. When we consider figure 7 and the experimental setup of GSE37443, it seems imply that whenever the combination of knock-ins includes loqs-pd, a spike in the proportion of reads mapped to the hpRNA class would be observed. Similarly, in GSE26230 and GSE17171 whenever loqs is absent in the genotype or when loqs is knocked down, the proportion of hpRNA reads decreased.

Interestingly, the output of GSE26230 and GSE17171 seems to imply that hpRNA biogenesis may be subdivided into hpRNA that are processed by either R2D2 or loqs-PD. In GSE26230, loss of heterozygosity of R2D2 resulted in a marked decrease of hpRNA, this is unusual since cannonically R2D2 is assumed to be dedicated towards producing exo-siRNA and not endo-siRNA. This abnormally is unlikely to be attributed to the effects of normalisation as the total read counts of both library appears to be similar. In GSE17171, while hpRNA levels are most affected during knockdown of Dcr-2 or loqs we do see a slight decrease in the hpRNA proportion in the R2D2 knockdown compared to the untreated and lac-Z knockdown. This appears to go against the prevailing notion that endo-siRNA biogenesis is solely due loqs-PD interaction with Dcr-2. Perhaps within hpRNA species

some may require different Dcr-2 partners while some requires both Dcr-2 partners to be present with the R2D2-loqs- Dcr-2 trimer complex described in (Miyoshi et al., 2010).

To fulfil our curiosity, while we do not have a robust way to perform DGE on specific sRNA class we experimented with the same workflow described in figure 5 and 6 for hpRNA and generated a Venn-diagram and heatmap (see appendix for figure 10 and 11). 5 significant hits in the loqs-PD resuce for hpRNA out of a total collection of 22 hpRNA was revealed but unfortunately, as the regulatory target of hpRNA is not well understood, the biological significance of this finding could not be ascertained from FlyBase. However, it does suggest for the possibility that availability of different Dcr-2 partner may be alter the expression for different hpRNA subspecies since basal level of hpRNA expression was observed in the main dataset despite absence of loqs-PD. Perhaps when a more robust DGE analysis pipeline for the various subclasses of sRNA is developed we would be able to verify this findings across the various GEO datasets we have analysed.

Endo-siRNA Biogenesis and Loqs-PD rescue

While the primary dataset and GSE37443 appears to have similar experimental setup on specific loqs-isoform rescue, there appears to be some discrepancy in the sRNA profile. While both concur that loqs-PD rescue is associated with higher proportion of hpRNA, the other endo-siRNA appear to follow a different trend. In the main dataset, apart from hpRNA and 'DNA Transposon' in loqs-pd rescue, there seem to be no discernible change in other sRNA classes compared to the control, while in GSE37443, the last four panels appear appears to follow the trend of hpRNA, indicating that the loqs-PD rescue affects overall endo-siRNA biogenesis. A possible explanation for this could be that mutant loqs-null stable cell line (Lim et al., 2016) are actively dividing while cells in GSE34773 may not be, thus the accumulation of endo-siRNA was not as pronounced.

As suggested in (Hartig and Förstemann, 2011) the competitive binding of Dcr-2 to either loqs-PD or R2D2 binding might toggle Dcr-2 activity determining the preference for endo or exo-siRNA biogenesis while (Fukunaga et al., 2012) suggests that loqs-PD binding only

increases enzyme efficiency of Dcr-2 increasing steady state levels for both exo-siRNA and endo-siRNA in a sequence independent manner. Taken together, perhaps independent of loqs-PD and a hypothetically low R2D2 level, Dcr-2 is capable of balancing endo or exo-siRNA biogenesis, producing basal amount of endo-siRNA sufficient for the cell's survival. Hence, the observed elevated levels of hpRNA might be a form of redundant expression, the loqs-PD rescue serviced only to bias the steady state increase of endo-siRNA.Taking a closer look at figure 7, sidewise comparison of sRNA profile for wildtype and heterozygous-loqs cells in fly ovary appear to have an elevated level of endo-siRNA compared to the fly head. One possible explanation could be that isoform specific expression of loqs-PD may only be present during the fly's development, by performing loqs-PD rescue we may redundantly toggle Dcr-2 to favour the accumulation of endo-siRNA.

miRNA Biogenesis in the context of Loqs

Loqs is vital in maintaining cell viability in normal *D. melanogaster* due to its role in miRNA biogenesis as the partner of Dcr-1 (Fukunaga et al., 2012), specifically the loqs-PB isoform is able to rescue fertility through GSC maintenance whereas the loqs-PA isoform is only able to restore viability. Therefore, in figure 7 there does not exist a sRNA library consists of only loqs-PD rescue, as loqs-null is conventionally lethal. For our main dataset, the effects of loqs-PB rescue has been detailed in (Lim et al., 2016). Although we have yet to develop a robust way to ascertain DGE our preliminary analysis in figure 5 and 6 indicate that indeed some miRNA genes are differently expressed when we contrast loqs-PA with loqs-PB rescue.

As miRNA composed of most of the sequencing reads of size 18-30 nt, sRNA-seq analysis involving loqs is especially problematic in as loqs directly affects miRNA biogenesis. Therefore, the total read counts as well as an indicator for the percentage of reads that perfectly mapped to the genome was provided, so that the results can be properly inferred.

Queries About the sRNA-Seq Analysis Pipeline

Non-Specific reads

It is to be noted that during the library preparation of the sRNA-seq, the RNA molecules were mainly selected base on their size as such products of mRNA decay may be captured as a valid read. However, in our experiment we operate under the caveat that each perfect match is a valid read and an indicator to actual sRNA level. Especially for the reads mapped to the 'newCisNAT' class which are in derived from active gene regions where transcription occurs, specificity may be an issue. While it is not possible that every read mapped to an sRNA category truly associates with RISC and is capable of gene regulation, the relative abundance of mapped reads is used as a proxy to indicate a corresponding increase in expression as well as maturation of the sRNA species.

Furthermore, in each category the read counts are pooled, if each class truly contains subclasses that are differentially expressed, any apparent change in read counts may be masked. Alternatively, we may mistakenly infer the change in sRNA biogenesis changes that occurred to only a specific subset of genes was generalised. Perhaps a more robust downstream DEG needs to be in place before we can truly ascertain our findings.

Sequential Mapping

Conventional read mapping is usually performed by aligning sequencing data with only one index, all features required would be compiled into a multi-line fasta file then compiled into a single index, feature information would be then extracted downstream. Instead, our indexes were separately prepared from feature details upstream of the read mapping.

To test the for influence of sequential mapping, we compared the graphical output from a sequential-mapping against mapping the reads to all the indexes separately (Figure 12 and Figure 13). From it, it appears that sequential mapping affects only the last three categories the most with this the higher number of reads in the latter classes is mostly in line with our expectations. Indirectly, this demonstrates the increased mean fold change of hpRNA

observed in figure 3 was not a result of the way sequential mapping was constructed. Perhaps most surprising is the finding that some substantial portion of reads are mapped to 'other siRNA' class, while it has been about nine years, some siRNA sequences that were previously described have yet to be catalogued in the recent database, suggesting that publicly available online datasets may still harbour a wealth of information that has yet been uncovered.

Our Modifications to the Analysis Pipeline

By using the snakemake workflow engine instead of solely relying on bash scripts, we were able to gain access to a wide array of tools developed for python, bash and R, this relieved the need of 'hardcoding' the text processing and file formatting steps. However, a significant portion of the time had to be spent deciphering and understanding the initial pipeline to 'reinvent the wheel'.

Other modification made to the pre-existing pipeline includes complexing of multiple intermediate text processing steps to reduce the overall number of intermediate files, inclusion of assembly scaffolds in *Drosophila* genome to make full use of the annotation details in the database. Lastly, we reduced the amount of annotation resources as FlyBase is regularly updated was found to contain a more comprehensive set of features than other resources. Another point of departure from the original pipeline was leaving out the option to check for multi-mappers during read alignment with bowtie. While this may affect the accuracy of downstream DGE analysis, due to the way the sequential mapping the structured the final graphical output would remain the same, since potential multiple mappers would still be pooled and collated in the same index. Pertaining DGE, we would perhaps benefit from adopting a different read quantification strategy by performing read mapping with other bioinformatics tools like miRDeep*(An et al., 2013) and ShortStack(Axtell, 2013) which are designed with specific algorithms to perform read quantification for multiple-mappers in the near future.

Unorthodox Use of Snakemake

Perhaps a critique on the pipeline we generated could be that we failed to use snakemake in the way it was prescribed by the developers. To run our analysis, we have a collection of just three scripts to process the entire workflow, whereas snakemake was intended to be used as a collection of smaller scripts which could then be used by a wrapper script. Files extensions should be used for the conversion from one file type to another as the sequencing data is processed. Instead, we relied on generating flags files after the end of a rule and used the creation of flags to control the logic of the pipeline. Perhaps, some future works could be dedicated towards decomposing the three scripts into smaller rules targeted at handling file extensions, which snakemake is optimised for.

References

An, J., Lai, J., Lehman, M.L., Nelson, C.C., 2013. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Research 41, 727–737. https://doi.org/10.1093/nar/gks1187

Axtell, M.J., 2013. ShortStack: Comprehensive annotation and quantification of small RNA genes. RNA 19, 740–751. https://doi.org/10.1261/rna.035279.112

Babraham Bioinformatics, 2018. FastQC.

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6, 11. https://doi.org/10.1186/s13100-015-0041-9

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

Carthew, R.W., Sontheimer, E.J., 2009. Origins and Mechanisms of miRNAs and siRNAs. Cell 136, 642–655. https://doi.org/10.1016/j.cell.2009.01.035

Chak, L.-L., Mohammed, J., Lai, E.C., Tucker-Kellogg, G., Okamura, K., 2015. A deeply conserved, noncanonical miRNA hosted by ribosomal DNA. RNA 21, 375–384. https://doi.org/10.1261/rna.049098.114

Czech, B., Hannon, G.J., 2010. Small RNA sorting: matchmaking for Argonautes. Nature Reviews Genetics 12, 19.

Czech, B., Malone, C.D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J. a, Sachidanandam, R., Hannon, G.J., Brennecke, J., 2008. An endogenous small interfering RNA pathway in Drosophila. Nature 453, 798–802. https://doi.org/10.1038/nature07007

Czech, B., Zhou, R., Erlich, Y., Brennecke, J., Binari, R., Villalta, C., Gordon, A., Perrimon, N., Hannon, G.J., 2009. Hierarchical rules for Argonaute loading in Drosophila. Molecular Cell 36, 445–56. https://doi.org/10.1016/j.molcel.2009.09.028

Dale, R., 2013. Gffutils.

Daugaard, I., Hansen, T.B., 2017. Biogenesis and Function of Ago-Associated RNAs. Trends in Genetics 33, 208–219. https://doi.org/10.1016/j.tig.2017.01.003

Forstemann, K., Tomari, Y., Du, T., Vagin, V.V., Denli, A.M., Bratu, D.P., Klattenhoff, C., Theurkauf, W.E., Zamore, P.D., 2005. Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. PLoS Biology 3, e236.

Free Software Foundation, 2007. Bash (3.2.48) [Unix shell program].

Fukunaga, R., Han, B.W., Hung, J.-H., Xu, J., Weng, Z., Zamore, P.D., 2012. Dicer partner proteins tune the length of mature miRNAs in flies and mammals. Cell 151, 533–46. https://doi.org/10.1016/j.cell.2012.09.027

Ghildiyal, M., Seitz, H., Horwich, M.D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E.L.W., Zapp, M.L., Weng, Z., Zamore, P.D., 2008. Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells. Science (New York, N.Y.) 320, 1077–1081. https://doi.org/10.1126/science.1157396

Haac, M.E., Anderson, M.A.E., Eggleston, H., Myles, K.M., Adelman, Z.N., 2015. The hub protein loquacious connects the microRNA and short interfering RNA pathways in mosquitoes. Nucleic acids research 43, 3688–700. https://doi.org/10.1093/nar/gkv152

Hannon's Lab, 2018. FASTX-Toolkit.

Hartig, J.V., Förstemann, K., 2011. Loqs-PD and R2D2 define independent pathways for RISC generation in Drosophila. Nucleic acids research 39, 3836–51. https://doi.org/10.1093/nar/gkq1324

Huber, W., Carey, J., V., Gentleman, R., Anders, S., Carlson, M., Carvalho, S., B., Bravo, C., H., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, D., K., Irizarry, A., R., Lawrence, M., Love, I., M., MacDonald, J., Obenchain, V., Ole's, K., A., Pag'es, H., Reyes, A., Shannon, P., Smyth, K., G., Tenenbaum, D., Waldron, L., Morgan, M., 2015. {O}rchestrating high-throughput genomic analysis with {B}ioconductor. Nature Methods 12, 115–121.

Kans, J., 2008. Entrez Direct: E-utilities on the UNIX Command Line.

Köster, J., Rahmann, S., 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Lai, E.C., Okamura, K., 2008. Endogenous small interfering RNAs in animals. Nature Reviews Molecular Cell Biology 9, 673. https://doi.org/10.1038/nrm2479

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25. https://doi.org/10.1186/gb-2009-10-3-r25

Law, C.W., Alhamdoosh, M., Su, S., Smyth, G.K., Ritchie, M.E., 2016. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Research 5, 1408. https://doi.org/10.12688/f1000research.9005.2

Lim, M.Y.T., Ng, A.W.T., Chou, Y., Lim, T.P., Simcox, A., Tucker-Kellogg, G., Okamura, K., 2016. The Drosophila Dicer-1 Partner Loquacious Enhances miRNA Processing from Hairpins with Unstable Structures at the Dicing Site. Cell Reports 15, 1795–808. https://doi.org/10.1016/j.celrep.2016.04.059

Marques, J.T., Kim, K., Wu, P.-H., Alleyne, T.M., Jafari, N., Carthew, R.W., 2010. Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila. Nature Structural & Molecular Biology 17, 24–30. https://doi.org/10.1038/nsmb.1735

McKinney, W., 2010. Data Structures for Statistical Computing in Python, in: Walt, S. van der, Millman, J. (Eds.), Proceedings of the 9th Python in Science Conference. pp. 51–56.

Miyoshi, K., Miyoshi, T., Hartig, J.V., Siomi, H., Siomi, M.C., 2010. Molecular mechanisms that funnel RNA precursors into endogenous small-interfering RNA and microRNA biogenesis pathways in Drosophila. RNA (New York, N.Y.) 16, 506–15. https://doi.org/10.1261/rna.1952110

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., Stamatoyannopoulos, J.A., 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics 28, 1919–1920.

Okamura, K., Balla, S., Martin, R., Liu, N., Lai, E.C., 2008. Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. Nature Structural and Molecular Biology 15, 581–590. https://doi.org/10.1038/nsmb.1438

Okamura, K., Chung, W.-j., Ruby, J.G., Guo, H., Bartel, D.P., Lai, E.C., 2008. The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. Nature 453, 803–6. https://doi.org/10.1038/nature07015.The

Python Software Foundation, n.d. Python Language Reference, version 2.7.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616

Saito, K., Ishizuka, A., Siomi, H., Siomi, M.C., 2005. Processing of pre-microRNAs by the Dicer-1-Loquacious complex in Drosophila cells. PLoS biology 3, e235. https://doi.org/10.1371/journal.pbio.0030235

Santos, G. dos, Schroeder, A.J., Goodman, J.L., Strelets, V.B., Crosby, M.A., Thurmond, J., Emmert, D.B., Gelbart, W.M., 2015. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Research 43, D690–D697.

Siomi, M.C., Sato, K., Pezic, D., Aravin, A.A., 2011. PIWI-interacting small RNAs: the vanguard of genome defence. Nature Reviews Molecular Cell Biology 12, 246.

Smit, AFA, Hubley, R & Green, P., 2015. RepeatMasker Open-4.0.

Tants, J.-N., Fesser, S., Kern, T., Stehle, R., Geerlof, A., Wunderlich, C., Juen, M., Hartlmüller, C., Böttcher, R., Kunzelmann, S., Lange, O., Kreutz, C., Förstemann, K., Sattler, M., 2017. Molecular basis for asymmetry sensing of siRNAs by the Drosophila Loqs-PD/Dcr-2 complex in RNA interference. Nucleic Acids Research 45, 12536–12550. https://doi.org/10.1093/nar/gkx886

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., Zhang, H., 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Research 37, D555–D559. https://doi.org/10.1093/nar/gkn788

Wickham, H., 2017. tidyverse: Easily Install and Load the 'Tidyverse'.

Zhou, R., Czech, B., Brennecke, J., 2009. Processing of Drosophila endo-siRNAs depends on a specific Loquacious isoform. RNA 15, 1886–1895. https://doi.org/10.1261/rna.

1611309.processed

Zhou, R., Hotta, I., Denli, A.M., Hong, P., Perrimon, N., Hannon, G.J., 2008. Comparative Analysis of Argonaute-dependent Small RNA Pathways in Drosophila. Molecular cell 32, 592–599. https://doi.org/10.1016/j.molcel.2008.10.018

Acknowledgements

Firstly, I would like to thank my UROPS supervisor Prof Greg Tucker-Kellogg for his patience and kind guidance throughout this project and providing us with the chance to work on the project despite us having no programming background initially. Despite his busy schedule he found time to walk through some of our codes and his open-door policy in the Lab allowed for timely consultations. He gave us the freedom to work independently and learn from our mistakes but nudges in the right direction and kept us on track. Even at the last leg of this project, he went beyond his way and assisted me in typesetting this document in markdown format.

Lee Jin Wee, a fellow student who worked with me on this project during the early phase of understanding the script and preparing the indexes as part of his ISM project. Jin Wee is a good sounding board and a fellow comrade who also went through the same growing pains of trying to learn programming only as an undergrad.

Next, we would also like to thank Okamura's lab for the collaboration with Prof Greg, providing us with the sRNA-Seq data and the initial analysis pipeline and also Dr Chak Li Ling who took the time to explain to the scripts and entertained our questions.

Kenneth who assisted us in technical know-hows about scripting in snakemake as well as the rest of GTK-Lab who provided us with insightful comments, the sharing knowledge during weekly lab meetings as well as contributing to the liveliness of place.

Lastly, I would also like to thank the countless individuals who actively participate in online forums such as Stack Overflow and Bio-Stars for their questions and answers which doubles as a 24/7 support for the project. Also, the developers and maintainers of bioinformatics tools and to the curators of online databases such as FlyBase around the globe.

Appendices

Table 2: Details of Program versions

Program Name	Environment	Version
Python	Python	2.7.14
GNU bash	Bash	4.3.48
R	R	3.4.4
Snakemake	Python	4.0.0
RepeatMasker	Bash	4.0.7
gffutils	Bash	0.8.8
gff2bed	Bash	2.4.26
bedtools	Bash	2.25.0
Pandas	Python	0.22.0
eutils	Bash	
FastQC	Bash	0.11.4
FASTX_Toolkit	Bash	0.0.14
dplyr	R	0.7.4
ggplot2	R	2.2.1
plyr	R	1.8.4
cowplot	R	0.9.2
EdgeR	R	3.20.9

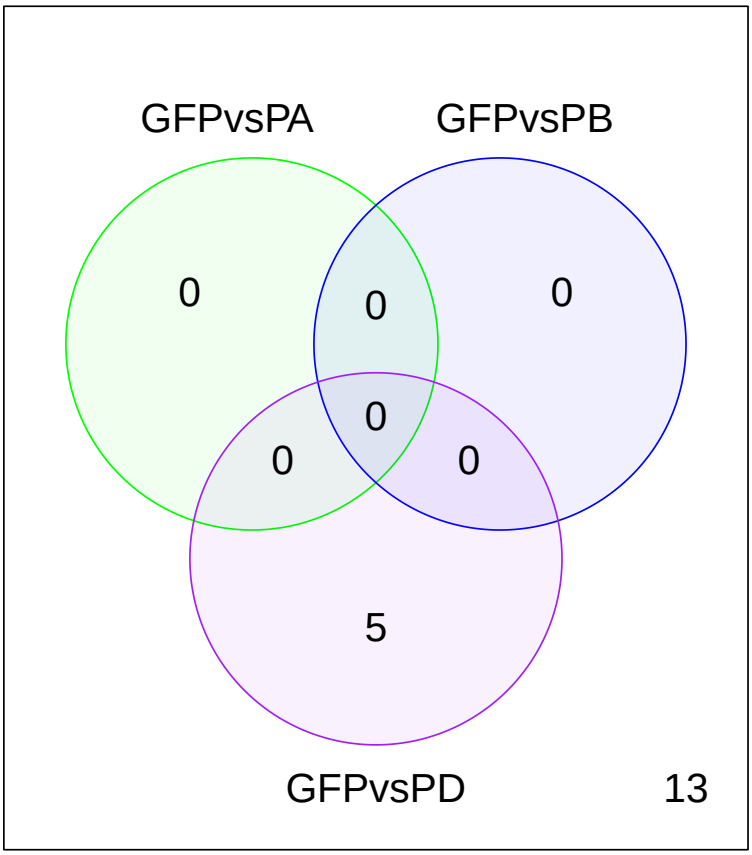


Figure 9: Preliminary DGE analysis on isoform-specific rescue on features counts mapped to miRNA index on a venn diagram

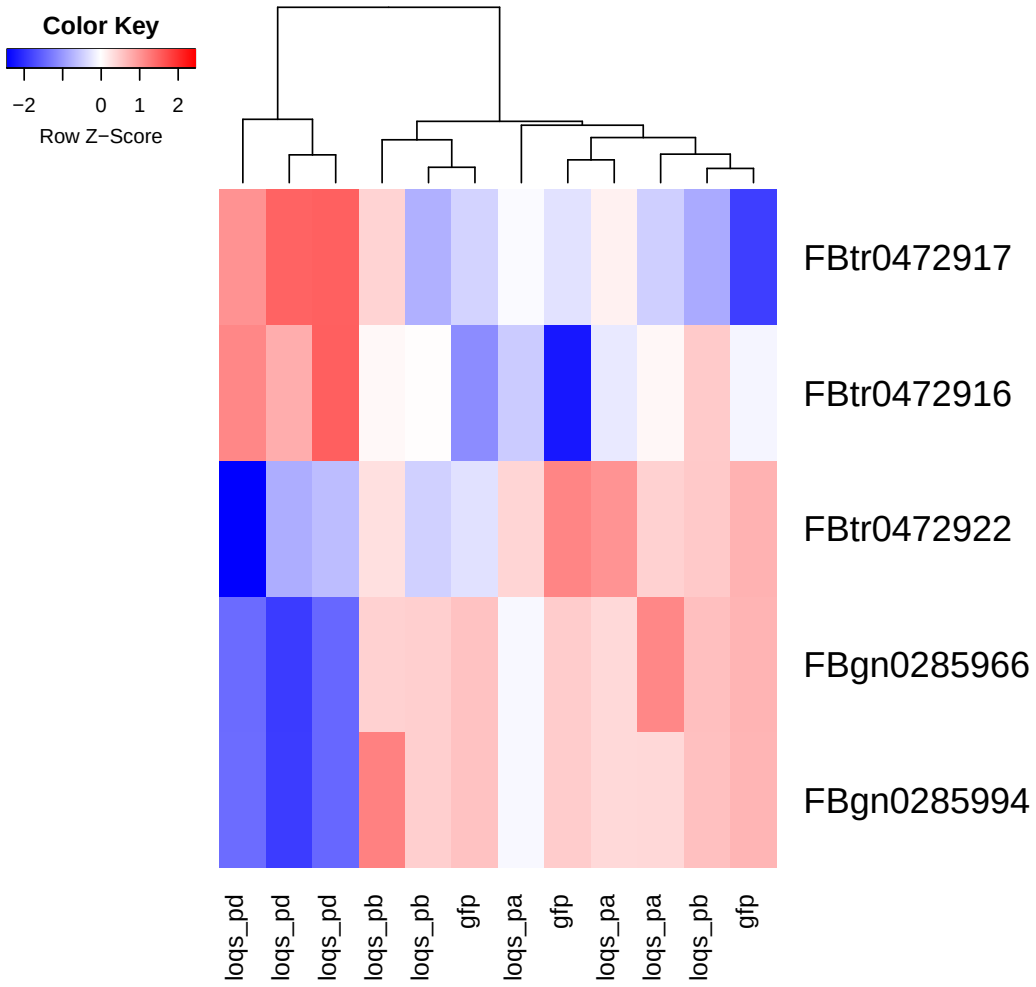


Figure 10: Preliminary DGE analysis on isoform-specific rescue on features counts mapped to miRNA index on a heatmap

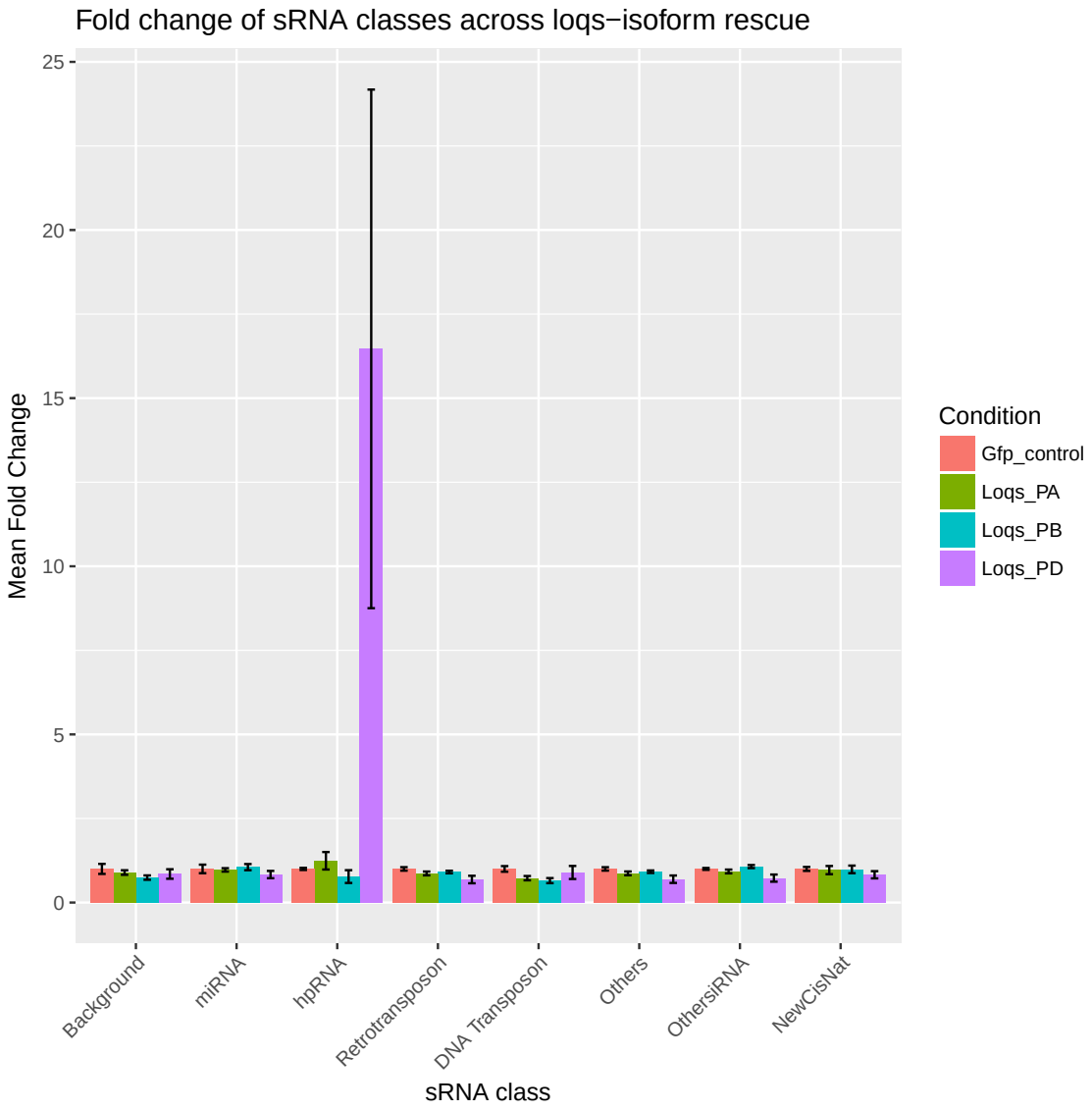


Figure 11: Replication of Figure 3 without sequential mapping

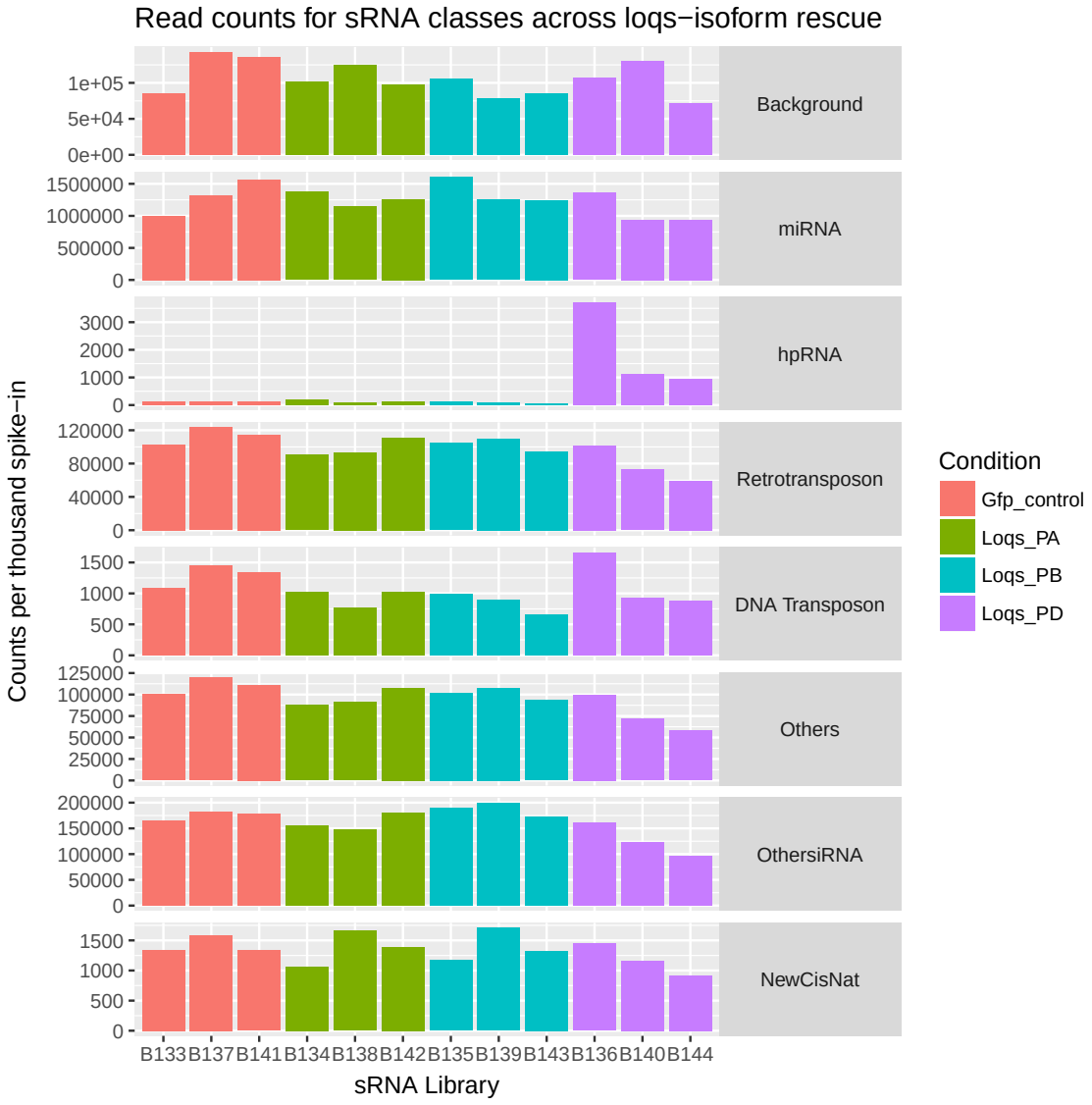


Figure 12: Replication of Figure 4 without sequential mapping

TEO_REN_YI_A0140768Y_LSM 3289

by TEO REN YI e0004678

Submission date: 24-Mar-2018 05:44PM (UTC+0800)
Submission ID: 935454770
File name: TEO_REN_YI_A0140768Y_LSM3289.pdf
Word count: 8395
Character count: 43711

Figure 13: Turnitin