Small RNA-seq Analysis of

Drosophila melanogaster loqs Mutants

LEE JIN WEE

A0140556J

Independent Study Module

(ISM)

Project Report

submitted to the

Department of Biological Sciences

National University of Singapore

UIS 3923 (4MC)

Dec 2017

1. Introduction

While originally thought to be a process exclusive to exogenous double-stranded RNA (dsRNA), RNA interference (RNAi) has emerged to be a crucial endogenously initiated process which regulates numerous aspects of gene expression. The small RNA molecules involved in endogenous RNAi tend to be around ∼20–30 nucleotides and can be differentiated into 3 distinct classes: **1)** microRNAs (miRNAs), **2)** small-interfering RNAs (siRNAs) and **3)** piwi-interacting RNAs (piRNAs) (Carthew & Sontheimer, 2009).

In the study of small RNAs, given the high number of paralogs it shares with humans, *Drosophila melanogaster* has been used as a central model organism. In *D. melanogaster* most research efforts have been directed towards both understanding miRNAs and siRNAs, with piRNAs being known for their elusive role in germline and transposon silencing (Iwasaki & Siomi, 2015). When it comes to miRNAs and siRNAs, there are several distinguishing attributes in their biogenesis and processing. Mature miRNAs range from ~21–25 nucleotides after being processed from stem-looped precursors (Hannon & He, 2004). On the other hand, endogenous mature siRNAs tend to carry a 21 nucleotide signature and are canonically derived from 3 types of pre-cursors : **1)** Transposable Elements (TEs), **2)** Cis/Trans Natural Antisense RNA (Cis-nat RNAs) and **3)** Hairpin RNAs (hpRNAs) (Lai & Okamura, 2008).

When it comes to their maturation/processing pathways, although both miRNAs and siRNAs require the involvement of the RNase III *Dicer* (*Dcr*) protein family, they diverge in another important aspect (Tijsterman & Plasterk, 2004). Associating with *Dicer* proteins in the mature miRNA and siRNA processing pathway are members of the *Loquacious* (*Loqs)* protein family, which consists of 3 isoforms: **1)** Loqs-PA, **2)** Loqs-PB

and **3)** Loqs-PD. Both Loqs-PA and Loqs-PB have been shown to associate with Dcr-1 to form the miRNA-dicing complex that processes the hairpin pre-miRNA structure (Jiang et al., 2005; Saito et al., 2005). However it has been shown that both Loqs-PA and -PB are non-essential in mature miRNA production, with in fact only around ~40% of miRNAs showing total Loqs dependence (Liu et al., 2007; Lim et al., 2016). The remaining Loqs-PD isoform playing an exclusive role in mature endo-siRNA production, with the production of a subset of mature endo-siRNAs being fully dependant on Loqs-PD (Hartig et al., 2009; Zhou et al., 2009).

Although previous studies have outlined the distinct roles each Loqs isoform, there has yet to be a specific transcriptomic analysis of Loqs-PD function and its overarching involvement in the various siRNA biogenesis pathways remains unknown. As such, *D. melanogaster* cell lines lacking the *loqs* gene locus were generated by Professor Katsutomo Okamura's lab, which resulted in 3 additional isoform-specific loqs rescue-lines (PA, PB, PD). Small RNA libraries were then prepared from KO and rescue cell lines, with 3 replicates each. The initial bioinformatics analysis performed by Dr Chak Li Ling in 2013 found that only for the Loqs-PD rescue libraries, a significantly higher amount of reads were mapped to hpRNA features, with no equivalent trend being observed for other siRNA features. This preliminary analysis thus suggested that not only is Loqs-PD is exclusively involved in the processing of hpRNA, it could also be essential to this process.

Despite this promising preliminary finding, there were several latent issues with Dr Li Ling's bioinformatics analysis pipeline. Firstly, the current pipeline inevitably uses an outdated version of the *D. melanogaster* genome (DM3), as the latest release of the

DM6 *D. melanogaster* genome was 2014. Secondly, Dr Li Ling's pipeline comprises of multiple bash scripts and a few thousand lines of highly repetitive code – a somewhat unavoidable consequence of given the circumstances at the time. This makes modifying the pipeline for alternative analysis extremely cumbersome and could in fact cause a significant hinderance down the road.

The main objective of this project was to then update the small RNA mappings using the DM6 genome while also rewriting Dr Li Ling's pipeline into a python-based Snakemake workflow. While the bulk of the project was to be highly technical without much emphasis on scientific discovery, we also aimed to determine if the same spike in hpRNA counts could be observed in the Loqs-PD rescue libraries post-DM6 mapping. Lastly, with our highly customable Snakemake workflow, we also planned to perform specific reconfigurations to Dr Li Ling's sequential mapping algorithm to test the validity of her initial findings.

2.Materials and Methods

2.1. Generation of *loqs* cell lines

The *loqs* mutant cell line was established from existing *loqs* mutant embryos and each rescue was performed by singly re-expressing *Loqs-PA*, *Loqs-PB* or *Loqs-PD* in these cell lines.

2.2. Generation of small RNA libraries and initial analysis pipeline.

The initial analysis pipeline written by Dr Liling pipeline was modelled after the method that was previously described by Chak et al. (2015). Initial processing was carried out with Fastxtoolkit v0.0.13.2 (http://hannonlab.cshl.edu/fastx_toolkit), Bedtools v2.19.0 (Quinlan & Hall, 2010) and custom shell scripts. The reads are first filtered for a length requirement of 18–30 nt and sequentially mapped to reference sequences with bowtie v1.0.0 (Langmead et al., 2009) with no mismatches allowed.

Several sets of reference sequences were used in the sequential mapping. RepeatMasker was used to identify Transposable Elements and other repeated sequences (Tarailo-Graovac & Chen, 2002). Known siRNA sequences were subsequently obtained from several published papers(Czech et al., 2008 ; Kawamura et al., 2008; Ghildiyal et al., 2008 ; Okamura et al., 2008 ; Okamura et al., 2008) miRNA features were obtained from pre-defined sequences on defined based on mirbase20 (Kozomara & Griffiths-Jones, 2014). The remaining features were derived from the Flybase General Feature Format (GFF) file (Tweedie et al., 2009). All reference sequences used in the initial analysis were based on Release 5 of the *Drosophila melanogaster* genome.

Following mapping, all raw read counts were first adjusted by the number of mapped alignment hits for a read sequence within the bowtie index. These were then normalized to per million mapped reads (RPM).

2.3. Conversion of pipeline to Snakemake workflow

The bash scripts from the initial pipeline were then re-written into a python-based Snakemake workflow (Köster & Rahmann, 2012). Additionally, the Python Data Analysis Library (*pandas*) was used for data processing (http://pandas.pydata.org/).

3. Results and Discussion

Given the nature of this project, before rewriting and reupdating the analysis, a significant amount of time had to first be spent understanding the rubrics of Dr Liling's initial pipeline. We then deconstructed her method to 3 main components: 1) Pre-processing and filtering, 2) Bowtie index generation and sequential mapping, 3) Normalization. Next, given the 1 semester time frame of this project, we chose to focus our efforts on components 1) and 2), knowing that they would give us a somewhat accurate picture of the updated analysis.

The next objective of the project was to then rewrite both components 1) and 2) to a modularized Snakemake workflow, while also updating the coordinates of all genome mappings to the newer Release 6 of the *Drosophila melanogaster* genome. The results of the initial mapping are shown in Fig. 1. Despite lacking normalization and correction for multiple mappers, the results produced by our updated pipeline bears still bears strong similarities to that generated by Dr Liling. As seen in Fig. 1, the most observable trend would be the clear spike in reads mapped to the hpRNA sequences and DNA transposon sequences for all 3 Loqs-PD rescue libraries. This is a finding that was previously shown in Dr Liling's analysis and its presence in our re-written Snakemake workflow is a positive indicator that the foundation of the analysis remains true to original.

Although this initial finding requires further normalization and validation it is certainly a novel finding. While previous findings with *Loqs-PD* have shown that it does indeed regulate siRNA biogenesis, a complete transcriptomic profile of its siRNA targets have yet to be performed (Hartig et al, 2009). Interestingly, while there was an

observable *Loqs-PD* dependant trend in both hpRNAs and DNA transposon, no such trend was observed for the other published "other siRNA" sequences. We suspected that this could be due to the arrangement of Bowtie indexes in Dr Liling's initial sequential mapping algorithm, with the "Other siRNA" indexes being mapped sequentially after the other transposon indexes. As a result, reads which could have mapped to "Other siRNA" would have been mapped to the transposon indexes. This suspicion is not unfounded, considering how siRNAs also function to silence transposons as a way to protect the genome (Lai & Okamura, 2008). We therefore decided to alter the mapping sequence such the "Other siRNA" sequences were mapped before transposon sequences. However, as shown in Fig. 2, there did not appear to be any observable changes in trend for both "Other siRNA" and transposon mappings even after changing the mapping sequence. This finding is a significant one as it not only suggests that *Loqs-PD* exclusively regulates the biogenesis of hpRNAs and small DNA Transposon-derived sequences, it also suggests that these small DNA transposon-derived sequences are currently uncharacterized as siRNAs and exploration of these sequences could potentially lead to the discovery of novel siRNAs.

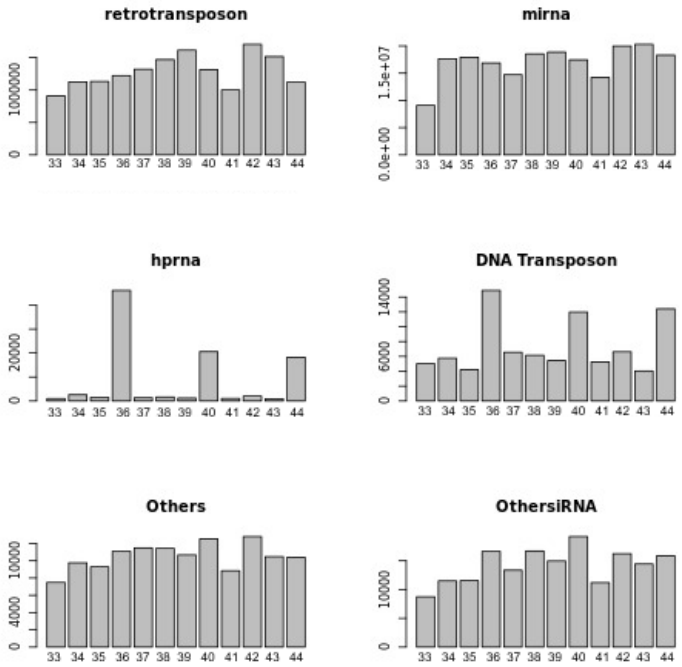

Figure 1. Results of mapping to DM6 sequences as performed using Dr Liling's original sequential order. The biological replicates are as represented: 1) *loqs-null:* 33,37,41 , 2) *loqs-PA:* 34,38,42 , 3) *loqs-PB:* 35,39,43 and 4) *loqs-PD:* 36,40,44.

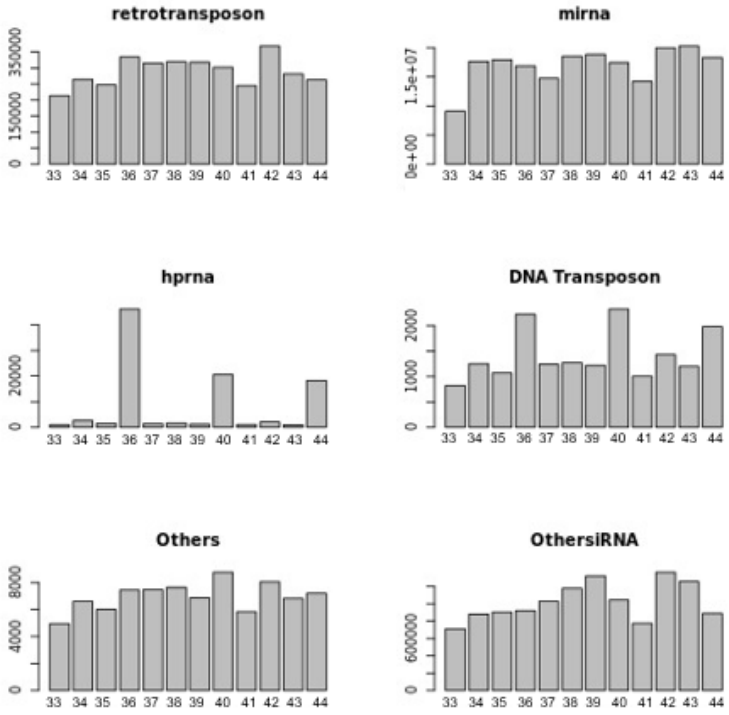

Figure 2. Results of mapping to DM6 sequences as performed using the modified sequential order. The biological replicates are as represented: 1) *loqs-null:* 33,37,41 , 2) *loqs-PA:* 34,38,42 , 3) *loqs-PB:* 35,39,43 and 4) *loqs-PD:* 36,40,44.

4. Conclusion

While the extent of scientific discovery in this project was limited by the extensive amount of technical work required to update and rewrite the original data analysis pipeline, a considerable amount of meaningful work was accomplished. We managed to somewhat replicate Dr Liling's initial findings, while also further validating the role that *Loqs-PD* plays in the biogenesis of hpRNAs and potentially DNA Transposon-derived siRNAs. However before any conclusive statements can be made, additional work must be done in terms of normalization and correction for multiple mappers – a key consideration for any small RNA seq analysis.

Works Cited

Carthew, R. W., & Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, *136*(4), 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*(3), 375–384. https://doi.org/10.1261/rna.049098.114

Czech, B., Malone, C. D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., … Brennecke, J. (2008). An endogenous small interfering RNA pathway in *Drosophila*. *Nature*, *453*(7196), 798. https://doi.org/10.1038/nature07007

Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., … Zamore, P. D. (2008). Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells. *Science*, *320*(5879), 1077–1081. https://doi.org/10.1126/science.1157396

Hannon, G. J., & He, L. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, *5*(7), 522. https://doi.org/10.1038/nrg1379

Hartig, J. V., Esslinger, S., Böttcher, R., Saito, K., & Förstemann, K. (2009). Endo-siRNAs depend on a new isoform of loquacious and target artificially introduced, high-copy sequences. *The EMBO Journal*, *28*(19), 2932–2944. https://doi.org/10.1038/emboj.2009.220

Iwasaki, Y. W., Siomi, M. C., & Siomi, H. (2015). PIWI-Interacting RNA: Its Biogenesis and Functions. *Annual Review of Biochemistry*, *84*(1), 405–433. https://doi.org/10.1146/annurev-biochem-060614-034258

Jiang, F., Ye, X., Liu, X., Fincher, L., McKearin, D., & Liu, Q. (2005). Dicer-1 and R3D1-L catalyze microRNA maturation in Drosophila. *Genes & Development*, *19*(14), 1674–1679. https://doi.org/10.1101/gad.1334005

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., … Siomi, H. (2008). Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature*, *453*(7196), 793. https://doi.org/10.1038/nature06938

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, *42*(D1), D68–D73. https://doi.org/10.1093/nar/gkt1181

Lai, E. C., & Okamura, K. (2008). Endogenous small interfering RNAs in animals. *Nature Reviews Molecular Cell Biology*, *9*(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

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*(8), 1795–1808. https://doi.org/10.1016/j.celrep.2016.04.059

Liu, X., Park, J. K., Jiang, F., Liu, Y., McKearin, D., & Liu, Q. (2007). Dicer-1, but not Loquacious, is critical for assembly of miRNA-induced silencing complexes. *RNA*, *13*(12), 2324–2329. https://doi.org/10.1261/rna.723707

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*(6), 581. 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*(7196), 803. https://doi.org/10.1038/nature07015

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

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*(7), e235. https://doi.org/10.1371/journal.pbio.0030235

Tarailo-Graovac, M., & Chen, N. (2002). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi0410s25

Tijsterman, M., & Plasterk, R. H. A. (2004). Dicers at RISC: The Mechanism of RNAi. *Cell*, *117*(1), 1–3. https://doi.org/10.1016/S0092-8674(04)00293-4

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., … Zhang, H. (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Research*, *37*(suppl_1), D555–D559. https://doi.org/10.1093/nar/gkn788

Zhou, R., Czech, B., Brennecke, J., Sachidanandam, R., Wohlschlegel, J. A., Perrimon, N., & Hannon, G. J. (2009). Processing of Drosophila endo-siRNAs depends on a specific Loquacious isoform. *RNA*, *15*(10), 1886–1895. https://doi.org/10.1261/rna.1611309