COMPUTATIONAL ANALYSIS OF SMALL RNAs IN MAIZE MUTANTS WITH DEFECTS IN DEVELOPMENT AND PARAMUTATION

by

Reza Hammond

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics and Computational Biology

Spring 2014

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

COMPUTATIONAL ANALYSIS OF SMALL RNAs IN MAIZE MUTANTS WITH DEFECTS IN DEVELOPMENT AND PARAMUTATION

by

Reza Hammond

Approved: _____
Blake Meyers, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Cathy Wu, Ph.D.
Chair of the Department of Bioinformatics and Computational Biology

Approved: _____
Babatunde Ogunnaike, Ph.D.
Dean of the College of College of Engineering

Approved: _____
James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

I would like to first thank my parents and my siblings for their endless support, encouragement and love. You have each played an instrumental role in the development of my life and career and I am so thankful for that. I would also like to thank my girlfriend, Elaina Welch, for your motivation and support. Whenever I've felt uncertain, you have always been there to reassure me or even just listen to me while I worked through problems. I can't thank you enough for everything.

My friends, there are far too many of you to list, but too important to not mention. Thank you for your endless support in my times of need. I don't know what I would do without you all.

To all of my fellow lab members, old and new, I greatly appreciate all of your guidance and support. Your kindness and passion creates the work environment of dreams. As we all grow and move on, I hope that we will all stay in touch and remain friends.

My undergraduate advisers: Li Liao, Michela Taufer, John Cavazos, Terrence Harvey and Jacqueline Aldridge. The guidance that I received from each of you will never be forgotten.

Finally my adviser, Blake Meyers: you took me into your lab with no practical biological experience and gave me a chance to work with so many amazing people. You have had so much patience with me and have helped me more than anyone can ever imagine. It has been a pleasure working with you and I hope that as I begin to work on my PhD with you, the experience will be just as great.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Small RNAs (sRNAs) are short segments of RNA which can induce gene silencing through interactions with homologous sequences by controlling either the stability or translation of mRNA. There are various classes of small RNA which differ in both their biogenesis and modes of suppression including microRNAs (miRNA) and heterochromatic short interfering RNAs (siRNA). The maize mutant f*uzzy tassel* exhibits a wide range of developmental defects has been found to contain a mutation in *DICER-LIKE1*, a critical enzyme involved in the biogenesis pathway of miRNAs. In addition, numerous maize mutants identified as being deficient in paramutation have been shown to be severely deficient in the accumulation of heterochromatic siRNAs as compared to normal maize plants. In order to determine the overall effects on small RNAs of these mutants, we have developed two separate computational pipelines to investigate the severity and specificity for which each of these mutants affect the biogenesis of miRNAs and heterochromatic siRNAs.

Chapter 1

INTRODUCTION

Small RNAs (sRNAs) are short segments of RNA which can induce gene silencing through interactions with homologous sequences. This silencing is effected by controlling either the stability or translation of mRNA. Not all sRNAs are created equal, however; there are various classes of sRNAs which differ in their biogenesis and modes of silencing. In plants, regulatory sRNAs range in size from 20-24 nucleotides (nt) in length and are produced both in response to exogenous viral infections as well as endogenous entities of the plant transcriptomes (Axtell, 2013). Two classes of sRNAs that are known to be important in the silencing of genes are short interfering RNAs (Elbashir, *et al.*, 2001) and microRNAs (Chen and Rajewsky, 2007). Short interfering RNAs (siRNAs) come in various classes which are all created via processing by a Dicer-like (DCL) enzyme which catalyzes the cleavage of double-stranded RNA (dsRNA). Similar to siRNAs, microRNAs (miRNAs) are processed by a DCL enzyme, however their precursors differ greatly as precursor miRNAs are produced from single-stranded RNAs that form a template for Dicer as a hairpin structure (Axtell, 2013).

Several maize mutants have been identified via screens for defects in either normal developmental processes, or in paramutation, an epigenetic phenomenon

involving small RNAs (see below). Some of these mutants have defects in the biogenesis pathways of miRNAs (i.e. those with defects in development) or heterochromatic siRNAs (those with defects in paramutation). One of the developmental mutants is called *fuzzy tassel* (for its eponymous phenotypic effect) which contains a mutation in the gene encoding DICER-LIKE1, affecting miRNA biogenesis. The other mutants have defects in NUCLEAR DNA-DEPENDENT RNA POLYMERASE IV subunit 1 and NUCLEAR DNA-DEPENDENT RNA POLYMERASE IV/V subunit 2, affecting heterochromatic siRNA biogenesis. Through the use of next generation sequencing, it is possible to map the abundance of small RNAs in target locations of the genome. Then, via bioinformatics-based analysis, it can be possible to determine how phenotypic changes are propagated by examining sRNA abundances in various genomic regions.

The goal of my work described in this thesis is to analyze these maize mutants to observe the effects for which these mutant maize plants have on specific classes of small RNAs. Pipelines to analyze their biogenesis and targeting pathways have been developed to allow for not only rapid analyses of these mutants, but for the rapid analysis of other mutants of maize and other organisms. What is presented in this thesis are the tools that have been created thus far, how they have been used, and what has been uncovered using these tools.

Chapter 2

MAIZE MUTANTS WITH DEFECTS IN DEVELOPMENT

Small RNA Biogenesis in Plants

Arabidopsis thaliana is the model plant organism for which numerous studies have been used to test hypotheses about the biology of plants – lessons applicable to a wide range of other plant species including crops. The small RNA biogenesis pathway of *Arabidopsis* involves numerous proteins and enzymes; mutants in the genes which encode these have been used as the basis of understanding for small RNA biogenesis. Some of the key players in these pathways in *Arabidopsis* include: ten ARGONAUTE family proteins, six RNA-dependent RNA polymerases (RDRP), four DICER endonucleases and two plant-specific DNA-dependent RNA polymerases (Pol IV and Pol V) (Pikaard, *et al.*, 2012). Pol IV and (RDR2) are required for the production of 24-nt siRNAs and thus in the absence of either of these polymerases, 24-nt siRNAs cease to be produced in Arabidopsis (Pikaard, *et al.*, 2012). The absence of these proteins would result in the lack of ability to prevent the transcription of repetitive genomic regions (Stonaker, *et al.*, 2009).

Maize *fuzzy tassel* Mutants

The *fuzzy tassel (fzt)* maize mutant was isolated by Beth Thompson during a screen of an M2 population of A619 EMS mutagenized plants, while she was a post-doc in Sarah Hake's lab at UC Berkeley. When analyzing these plants, a striking amount of phenotypic defects were observed including: shorter plant, reduced number of leaves and fewer and shorter internodes. In addition to these developmental defects, a large amount of reproductive defects in both male and female reproductive tissues resulting in complete sterility were observed (Thompson, *et al.*, in revision).

In order to pinpoint the causes of these defects, *fzt* was positionally cloned by the Thompson lab which mapped it to the maize ortholog of *Arabidopsis DICER-LIKE1*. In plants, primary miRNA transcripts are processed by DICER-LIKE1 (DCL1) to eventually create mature miRNAs (Zhu, *et al.*, 2013). Mutations in the miRNA biogenesis pathway in *Arabidopsis*, including *DICER-LIKE*, have previously been identified to produce pleiotropic developmental defects (Schauer, *et al.*, 2002). From this, we have hypothesized that a mutation in maize *DCL1* resulting in the misregulation of miRNAs could explain the pleiotropic phenotypic variation seen in *fzt*. In order to validate this, several analyses were performed to determine the level of impact that this *fzt* mutation has on the biogenesis of miRNAs.

Analysis of Maize *fzt* Mutant

In *Arabidopsis*, mutations in *DCL1* typically result in the improper processing of miRNAs resulting in repressed expression of various miRNAs (Bouché, *et al*.,

2006). Before investigating miRNA expression data, we first examined the overall expression of small RNAs to see if there actually was a significant impact on any size class of small RNA in *fzt*. Small RNA libraries were created from both seedling and tassel primordia tissue of both the *fzt* mutant and its non-mutant siblings. Three biological replicates were generated for all comparisons of small RNA abundances. Within both the seedling and tassel tissues, there is a statistically significant decrease in 21/22-nt sRNA abundance, these being the typical size class of miRNAs (Figures 9 & 10).

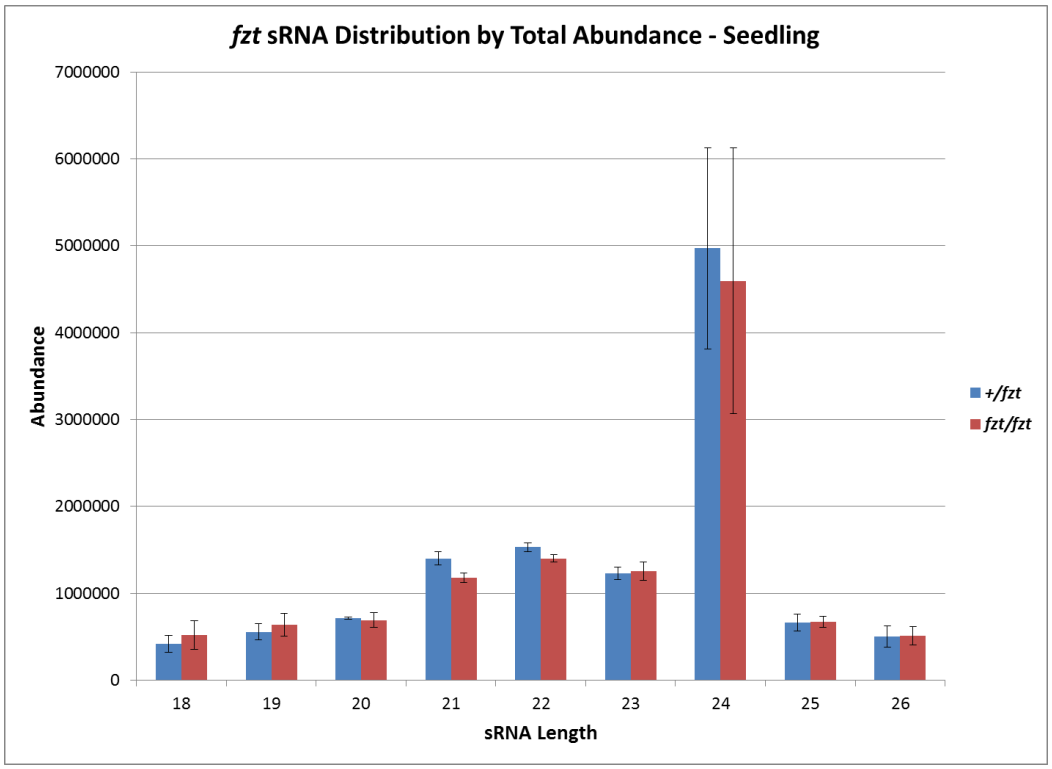

Figure 1: Distribution of total small RNA abundance by size class in seedling
sRNA distribution plot comparing total expressed small RNAs from three normal and three dcl1-*fzt* replicates.

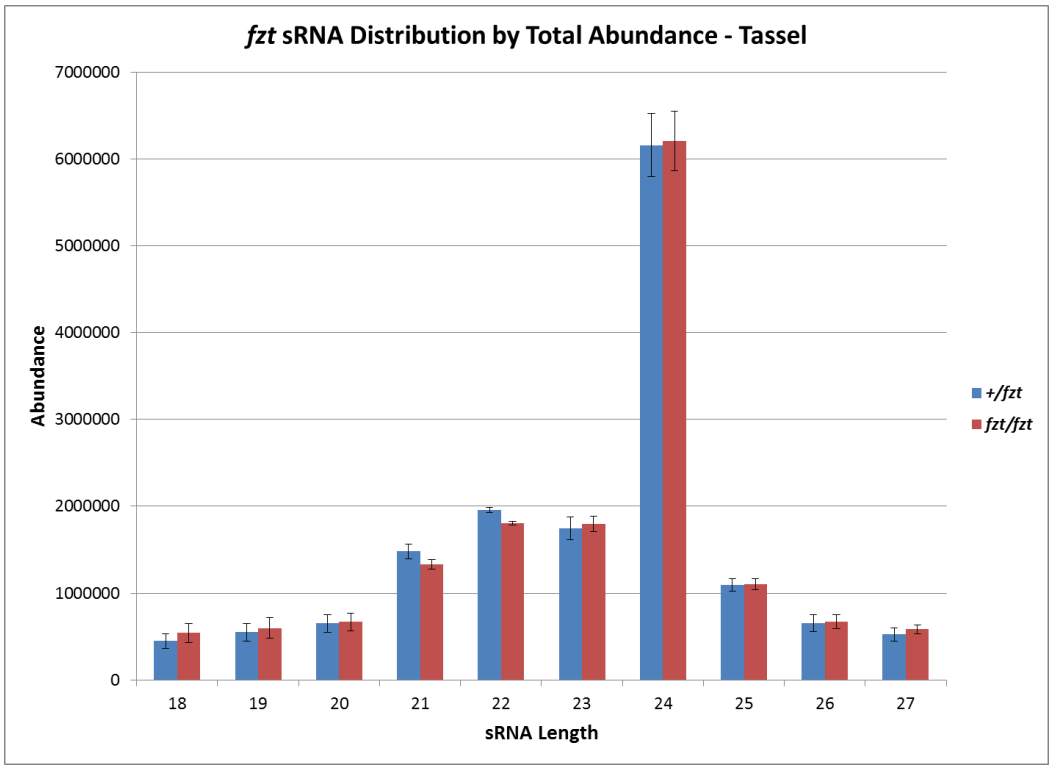

Figure 2: Distribution of total small RNA abundance by size class in tassel
sRNA distribution plot comparing total expressed small RNAs from three normal and three dcl1-*fzt* replicates.

The reduction in abundance of 21 and 22 nt sRNAs in *fzt* is statistically significant, though the change is quite small. This mutation of *DCL1* (identified as *DCL1-FZT*) is not a loss-of-function and thus we might expect that miRNAs are still produced in some effect; this is indicative of the slight reduction we see in the abundance of 21 and 22 nt sRNAs. Thus, we may hypothesize that all miRNAs are reduced a small amount or at least a subset of miRNAs are affected by this mutation of *DCL1*.

To investigate this further, individual miRNA expression from both the mutant and non-mutant were compared. Abundances of annotated miRNAs from miRBase, a repository for previously identified miRNAs, were extracted for both the *fzt* mutant

and its normal sibling in seedling and tassel tissues. The fold differences of miRNA expression were compared to determine how significant an impact miRNAs really exhibit in the *fzt* mutant. The 50 most abundant miRNAs in seedling and 47 most abundant in tassel were compared in expression between mutant and non-mutant. In seedling, all 50 investigated miRNAs have a reduced fold change, though about 45 of those are statistically significant (Figure 11). In tassel primordia, 37 of the 47 investigated miRNAs are reduced to a significant margin (Figure 12).

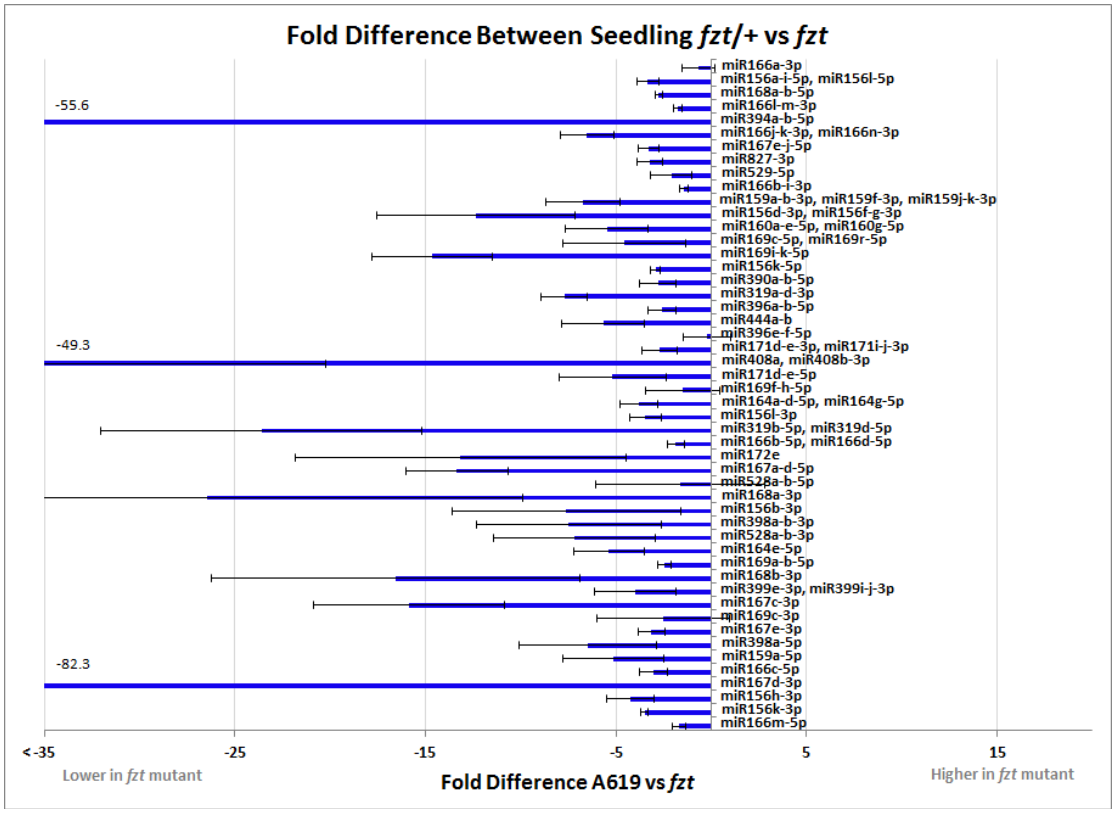

Figure 3: Comparison of miRNA abundance levels in *fzt*/+ and *fzt*/*fzt* in seedling
Fold difference of 50 most abundance annotated miRNAs from miRBase in *dcl1-fzt* as compared to normal controls. The miRNAs have been sorted by greatest abundance, so the higher the miRNA in the chart, the greater its overall abundance is in all libraries.

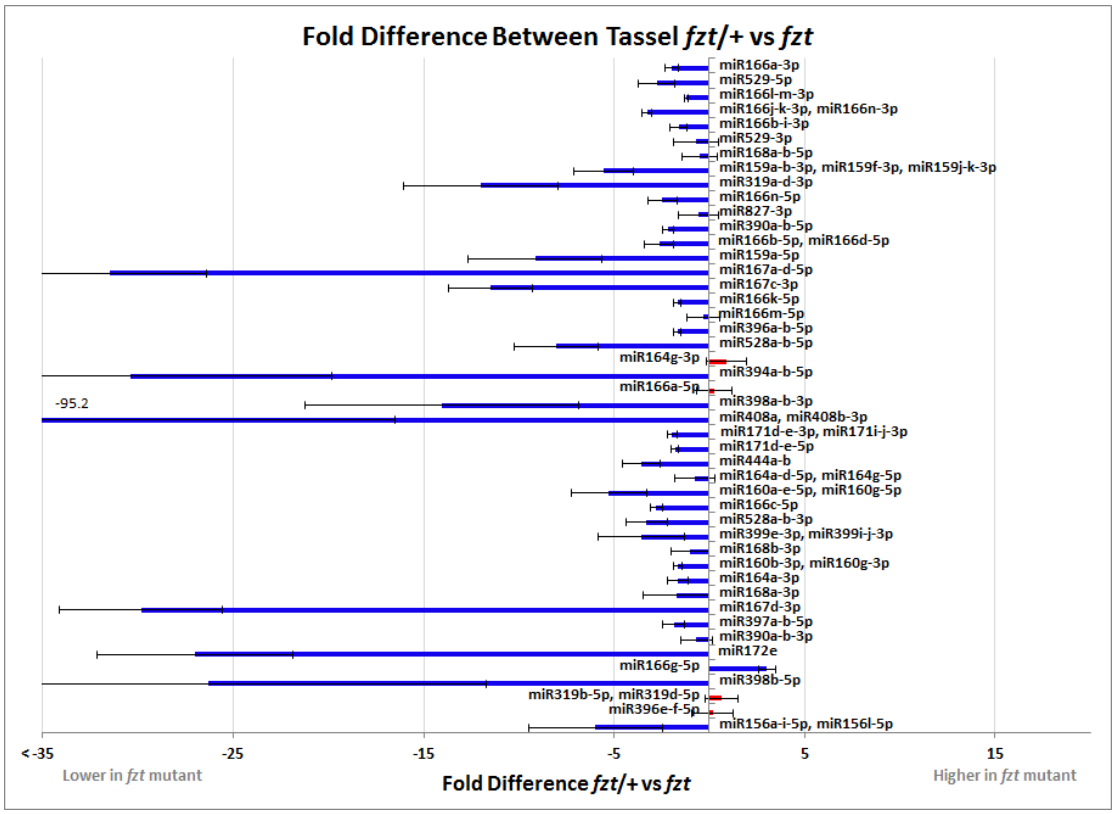

Figure 4: Comparison of miRNA abundance levels in *fzt*/+ and *fzt*/*fzt* in tassel
Fold difference of 47 most abundance annotated miRNAs from miRBase in *dcl1-fzt* as compared to normal controls. The miRNAs have been sorted by greatest abundance, so the higher the miRNA in the chart, the greater its overall abundance is in all libraries.

It can be seen that the abundance of the majority of miRNAs is downregulated in the *fzt* mutant at statistically significant values; the level of effect, however, is quite sporadic and non-uniform. From this analysis, we can determine two things. The first is that DCL1-fzt does in fact affect the biogenesis of miRNAs in maize. The second is that DCL-fzt does not affect all miRNAs the same.

Plant miRNAs suppress gene expression by binding to mRNA with near perfect complementarity, leading to target mRNA cleavage and decay (Finnegan and Matzke, 2003). While we have examined *fzt* suppressing the biogenesis of miRNAs,

we wanted to investigate the impact on mRNA expression. By creating RNA-seq libraries, it becomes possible to analyze the transcriptome, which includes mRNA expression, of *fzt*.

In order to identify the targets of suppressed miRNAs, TargetFinder, a tool which can predict sRNA targets, was modified and used to identify likely targets of the *fzt* miRNAs expressed within these libraries. The most suppressed miRNAs in *fzt* were selected from both seedling (23 total) and tassel primordial (14 total) to investigate the impact that these miRNAs have on the actual expression of mRNA. The modified TargetFinder was used to predict numerous targets of these miRNAs which are assigned a score between 0 and 7. When examining this data, one should expect that the targets of maize miRNAs be expressed in greater amounts in the *fzt* mutant as the lower abundance of these miRNAs implies less suppression of their target mRNAs. When analyzing this data, we see that the majority of the predicted targets in both seedling and tassel tissues appear to be relatively unchanged (Figures 13 & 14). There are, however, a few targets which are impacted quite dramatically compared to the other targets in both seedling and tassel. It may be possible that the targets that are relatively unaffected are not real targets of these miRNAs, but further analysis must be performed to remove such potential false positives.

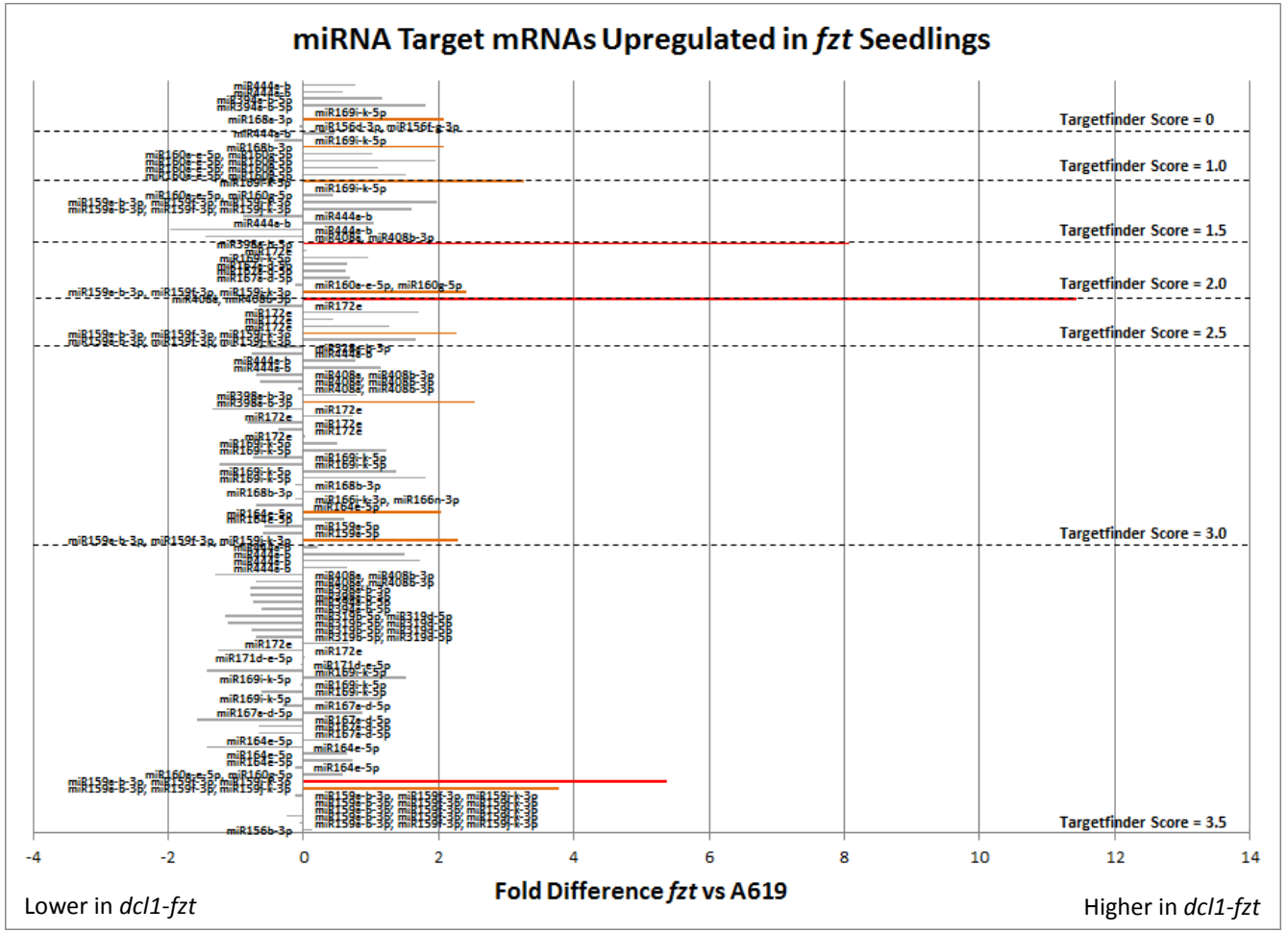

Figure 5: Comparison of miRNA target abundance levels in *fzt*/+ and *fzt*/*fzt* in seedling
Fold difference of predicted miRNA targets in *dcl1-fzt* as compared to normal controls. Targets separated by Targetfinder score where the lower scores indicate better homology.

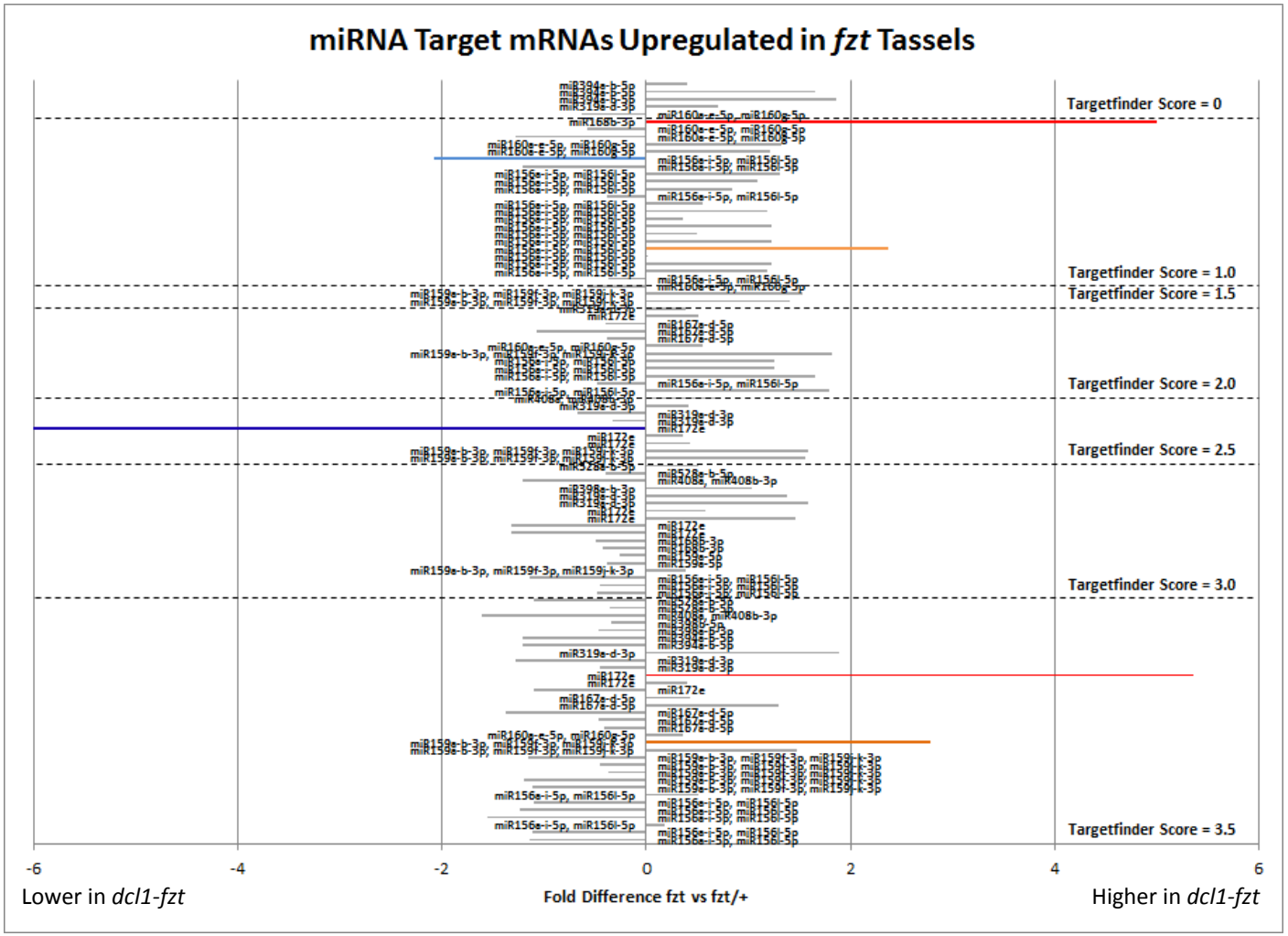

Figure 6: Comparison of miRNA target abundance levels in *fzt*/+ and *fzt*/*fzt* in tassel
Fold difference of predicted miRNA targets in *dcl1-fzt* as compared to normal controls. Targets separated by Targetfinder score where the lower scores indicate better homology.

One curious phenomenon in this data is the presence of two heavily downregulated targets in the tassel primordial RNAseq data. Such a phenomenon should be unexpected and implies incoherent regulation of these miRNAs and their targets (Jeong and Green, 2013). This is not unprecedented, however, and has been seen previously, though these targets must be validated and investigated further before speculating too much on this target.

Conclusions of *fzt* Analysis

In our analysis of small RNA expression levels of the *fzt* mutant, we were able to determine that there was a small but statistically significant decrease of 21-nt sRNAs. From this, it can be hypothesized that either all miRNAs are reduced a small amount or only a subset of miRNAs are affected by this mutation in *DCL1*. In order to determine the level of effect this mutant had on miRNAs, numerous miRNAs from miRBase were identified and their expression levels in *fzt* were compared to the expression levels of their non-mutant siblings. We were able to determine that nearly all identified miRNAs were reduced, though some miRNAs were impacted more than others. This supports the hypothesis that only a subset of miRNAs is affected.

Plant miRNAs suppress gene expression by binding to mRNA with near perfect complementarity, leading to target mRNA cleavage and decay (Finnegan and Matzke, 2003). miRNAs suppress gene expression, so by predicting target locations of miRNAs, we should be able to identify genes for which *DCL1* indirectly regulates. With the use of RNA-seq libraries, we analyzed the level of mRNA expression

mapped to regions for which miRNAs are predicted to target. Using a modified version of TargetFinder, http://carringtonlab.org/resources/targetfinder, we were able to predict targets for miRNAs within the entire maize genome. Then, using this list of targets, we identified the expression level of the targets within the mutant and non-mutant libraries to build a comparison for seedling and tassel libraries. While false positives may exist, we observed that the majority of largely affected targets were expressed in greater abundance in the *fzt* mutant. This further implies that the underlying cause of the phenotypic defects in the *fzt* mutant is due to the misregulation of miRNAs by DCL1-fzt.

Future Work

Primary transcripts (pri-miRNAs) are produced by RNA polymerase II (Pol II) which are then processed by DCL1 to eventually create miRNAs. The exact mechanism for which DCL1 identifies and cleaves is not known, but it is believed that there are unique structures of the pri-miRNA that DCL1 recognizes to either cleave to create the miRNA (productive processing) or a non-miRNA fragment (abortive processing) (Zhu, *et al.*, 2013).

This analysis shows that *fzt* impacts the function of DCL1 by inhibiting the production of some miRNAs. While useful, the cause for why some miRNAs are impacted and others are not remains unknown. I suspect that the underlying cause of the variation in the miRNA abundances is the structure of certain pri-miRNA transcripts. Thus, it may be plausible that the miRNAs with mostly unchanged

abundance levels have primary transcripts that are cleaved properly by DCL1-FZT. The heavily impacted miRNAs could have structures that are no longer recognized by DCL1-FZT forcing the aborted processing of the miRNAs rather than their productive processing (Zhu, *et al.*, 2013). It may be possible that this mutation affects the ability of DCL1 to cleave which results in depleted abundances of some miRNAs while others remain largely unchanged. If the pri-miRNAs of the miRNAs can be performed, we can identify whether there exist some similarities to the reduced miRNAs causing the non-uniform decrease of miRNAs in the *fzt* mutant.

While work is being done within the Meyers' lab to improve the target prediction process, the targets predicted by TargetFinder are only computationally predicted and don't rely on much more than DNA pairing patterns. In order to remove the possibilities of false positives in the target analysis, previously validated targets could be used as a high-confidence dataset, though this would not necessarily be representative of all targets for our data. Parallel Analysis of RNA Ends (PARE) libraries could thus be constructed. These PARE validated targets would be a better representative of the predicted targets of TargetFinder or the previously validated targets from other analyses.

Chapter 3

MAIZE MUTANTS WITH DEFECTS IN PARAMUTATION

Paramutation in Maize

Paramutation, as first described by Alexander Brink in 1956 for the maize *red1* gene, is the non-Mendelian segregation of alleles due to one allele's epigenetic change of the same gene in the other allele (Arteaga-Vazquez, *et al.*, 2010). This results in the change of phenotype even though the genetics was not coded for such an occurrence (Adams and Meehan, 2013). This phenomenon is well-studied, though not fully understood, and the mechanism for how it occurs can only be postulated. Much work in the field of maize research is starting to help unveil how paramutation is carried out and its molecular basis.

In maize, paramutation has been observed at the *b1, r1,* and *pl1* loci (Lisch, et. al, 2002). Numerous genes required for paramutation have been identified in maize mutants in multiple labs; several mutants identifiable as *mediator of paramutation* (*mop*) have been isolated in the *b1* system while several more identifiable as *required to maintain repression* (*rmr*) have been isolated in the *pl1* system. Thus far, all genes characterized as being required for paramutation were identified via forward genetic screens and all encoded proteins associated with the biogenesis of siRNAs in other species (Arteaga-Vazquez, *et al.*, 2010). Therefore, it would appear that siRNAs could

play a role in the molecular mechanism of paramutation and thus these mutants can be used to unveil the role of small RNAs, if any, in paramutation.

siRNA Association with Paramutation in Maize

Multiple maize mutants have been used to identify genes required for maize paramutation, with all currently-published mutants associated with the siRNA biogenesis pathways. The first gene cloned was *mop1*, which was predicted to be an ortholog of the *Arabidopsis thaliana RDR2* (Alleman, *et al.*, 2006). Following this, more mutants in maize showing repressed paramutation were identified and shown to encode proteins with high similarity to subunits of the plant-specific RNA polymerases (putative orthologs of subunits of the previously mentioned Arabidopsis Pol IV and Pol V complexes). These mutants include: *rmr6/mop3* (the dual name resulting from the identification of the same gene in both screens, rmr from the Hollick lab and mop from the Chandler lab), believed to encode for the largest subunit of Pol IV, *NUCLEAR DNA-DEPENDENT RNA POLYMERASE IV* (*NRPD1*), and *rmr7/mop2*, believed to encode the shared second largest subunit of Pol IV and PolV, *NUCLEAR DNA-DEPENDENT RNA POLYMERASE IV/V* (*NRPD2/E2*). All of these mutants have been implicated in some way with the *Arabidopsis* 24-nt heterochromatic siRNA biogenesis pathway, typically associated with the silencing of repetitive region (Axtel, 2013).

Previous Analyses of *rdr2$_{mop1}$* Mutant

The *mop1* mutant (referred to as *rdr2$_{mop1}$* for its mutation in the maize gene orthologous to *Arabidopsis RDR2* [RNA-Dependent RNA Polymerase 2]) is a homozygous recessive mutation that prevents paramutation at multiple loci, including the *b1, r1,* and *pl1* loci (Dorweiler, *et al.*, 2000), which has been shown to be required in the biogenesis of 24-nt heterochromatic siRNAs (Lu, *et al.*, 2006). Two maize alleles, including *B′*, responsible for the sporadic purple color of maize plants, and *B-I*, responsible for the dark-purple plant color, have been identified as markers that give a clear indication of the occurrence of paramutation. In the absence of MOP1, paramutation is unable to occur which allows for the expression of both the *B′* and the *B-I* phenotype. In the presence of MOP1, however, the paramutagenic allele *B′* is able to change the paramutable allele *B-I* to *B′* 100% of the time (Lisch, *et al.*, 2002). In other words, in the absence of MOP1 the gametophytes of a heterozygous individual *B*/*B-I* will produce either *B* or *B-I*. In the presence of MOP1, however, a heterozygous individual *B/B-I* will produce a gametophyte with either *B* or *B**, where the *B** indicates the *B-I* allele has paramutated to *B*.

In 2008, a paper from the Meyers lab and collaborators (Nobuta et al., 2008) showed that there is a significant reduction of 24-nt heterochromatic siRNAs as well as an enrichment of both 21 and 22-nt sRNAs in *rdr2$_{mop1}$* mutants. The enrichment of 21-nt sRNAs was consistent with Arabidopsis *rdr2* mutants, but the increase in 22-nt sRNAs was unique to maize. Upon analyzing various miRNAs from both the sequencing data and northern blots, it was determined that the enrichment of miRNAs

seen in the sequencing data was not apparent in the blot data. The interpretation was that the miRNA abundances only appeared to be increased due to a similar sampling depth but the absence of the major class of 24-mers. Based on this conclusion, the data was normalized to the level of the average level of wildtype expression (Nobuta, et al., 2008). This analysis had been performed prior to the completion of the maize genome sequencing, so two contigs comprising 14 Mb were used as reference for the maize homoeologous regions (Nobuta, et al., 2008; Bruggmann, et al., 2006). To confirm that this analysis holds true for the current version of the maize genome, this analysis was recreated using new $rdr2_{mop1}$ libraries mapped to the whole genome. The results remain consistent with the earlier conclusions (Figure 1).

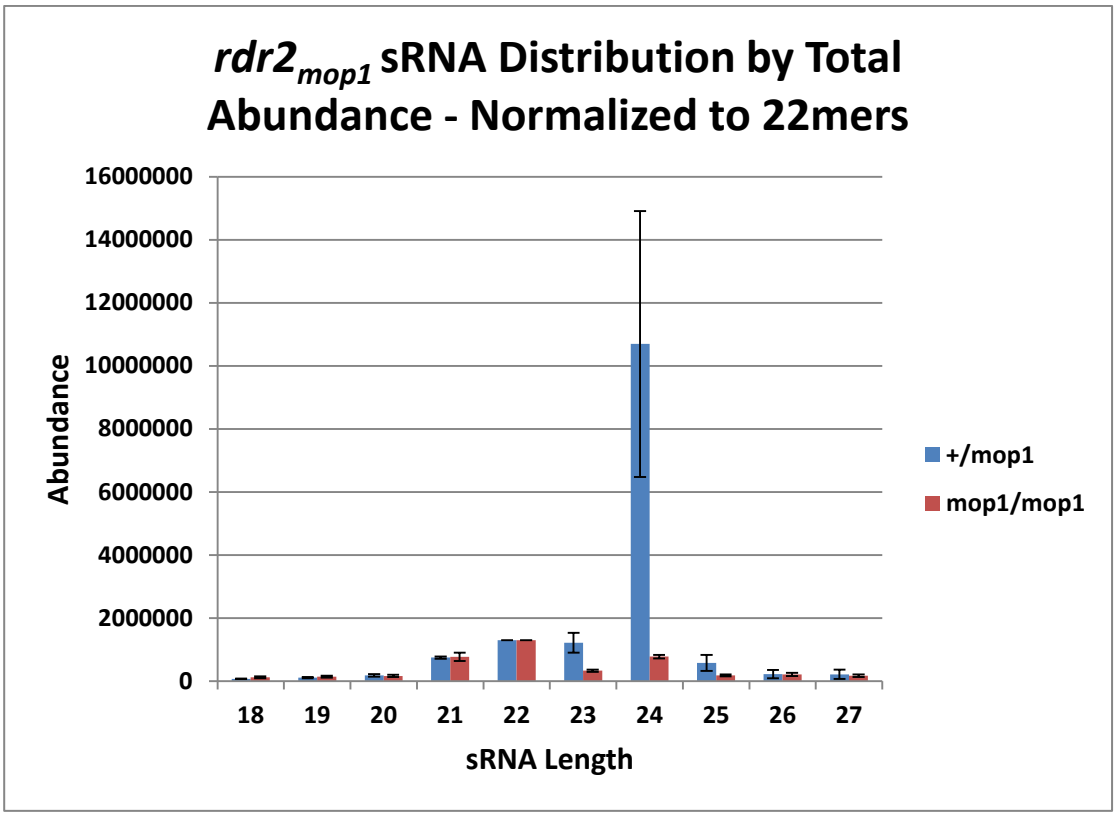

Figure 7: Distribution of total small RNA abundance by size class in *rdr2*$_{mop1}$
sRNA distribution plot comparing total expressed small RNAs from three *+/mop1* (+) and four *mop1/mop1* (-) siblings

nrpd1$_{rmr6/mop3}$ and *nrpd2/e2*$_{rmr7/mop2}$

rmr6 (referred to as n*rpd1*$_{rmr6}$ for its mutation in the maize gene orthologous to *Arabidopsis NRPD1*) and *rmr7* (referred to as n*rpd2/e2*$_{rmr7}$ for its mutation in the maize gene orthologous to *Arabidopsis NRPD2/E2*) function within the RNA-directed DNA methylation pathway which produces 24-nt siRNAs. The *nrpd1*$_{rmr6}$ maize mutant disrupts the gene orthologous to *Arabidopsis NRPD1*, the largest subunit of Pol IV, for which only one copy exists. *nrpd1*$_{rmr6}$ mutations have been seen to act specifically on the *Pl1-Rhoades* alleles by maintaining the repressed state of *Pl'* (Hollick, *et al.*, 2005). Similar to the *Arabidopsis rdr1* and maize *rdr2*$_{mop1}$ mutant phenotypes, a

dramatic loss of 24-nt heterochromatic siRNAs was observed (roughly 18% of non-mutant levels) (Lu, *et al.*, 2006). This is expected as Pol IV is required for the biogenesis of all *Arabidospsis* 24-nt siRNAs and in *Arabidopsis*, this subunit is required for proper Pol IV function.

The *NRPD2/NRPE2* gene encodes for the second largest subunit which is shared between Pol IV and Pol V, at least in *Arabidopsis* – hence the "D2/E2" designation (Pol IV is NRPD and Pol V is NRPE). While *Arabidopsis* has one functional *NRPD2/E2* gene, maize has three copies of the *NRPD2/E2*-like gene. Stonaker, *et al.* (2009) and Sidorenko, *et al.* (2009) independently identified recessive mutants of the same *NRPD2/E2-like* gene in maize, which both showed mutant phenotypes. This suggests that the three *NRPD2/E2-like* genes of maize are not fully redundant as a loss of function of two or three of the genes would have been required to see the mutant phenotype. While full redundancy does not appear likely, there remains the possibility that there is some partial redundancy in the maize *NRPD2/NRPE2-like* gene. Much like both the *nrpd1* and *rdr2* maize mutants, Stonaker, *et al.* (2009) and Sidorenko, *et al.* (2009) described significant losses of 24-nt siRNAs. While loss-of-function mutations of *NRPD1* and *NRPD2* genes cripple Pol IV, as both are catalytic subunits, these mutants do not produce the same phenotype. Phenotypic defects examined by both groups showed that the *NRPD2/NRPE2-like* loss-of-function mutation is not as severe as in the *nrpd1* mutants suggesting that there is more than one NRPD2/NRPE2-like protein that may interact with NRPD1 to form the Pol IV with reduced functionality (Pikaard and Tucker, 2009).

Analysis of Maize *rmr* Mutants

In the *rdr2*$_{mop1}$ analysis, a large reduction in 24-nt sRNAs was observed. *nrpd1*$_{rmr6}$ and *nrpd2/e2*$_{rmr7}$ are believed to have interruptions of genes involved in the same heterochromatic siRNA production pathway, therefore their loss of function should also result in the same reduction in 24-nt sRNA abundance. We began this analysis by creating a small RNA distribution profile to examine how overall small RNA expression was affected in each of the *nrpd1*$_{rmr6}$ and *nrpd2/e2*$_{rmr7}$. The n*rpd1*$_{rmr6}$ small RNA distribution profile replicates the trend seen in the *rdr2*$_{mop1}$ profile; there is a clear reduction of 24-nt sRNAs with slight enrichment of 21 and 22-nt sRNAs. Due to the enrichment of 21 and 22-nt sRNAs, all small RNAs were normalized to the averaged ratio of mutant/non-mutant abundance (Nobuta, *et al.*, 2008). Figure 2 shows the normalized distribution which mimics northern blots where substantial reductions in 24-nt sRNAs were seen in the *rdr2*$_{mop1}$ mutants (Sidorenko, *et al.*, 2009; Stonaker, *et al.*, 2009). The sRNA distribution profile of *nrpd2/e2*$_{rmr7}$ also shows a slight enrichment of 21 and 22-nt sRNAs calling for the normalization of the mutant to the levels of the average non-mutant 22-nt sRNA levels. While there is a reduction in 24-nt sRNA expression level, it is nowhere near as dramatic of a loss as n*rpd1*$_{rmr6}$ and *rdr2*$_{mop1}$ (Figure 3). This varies quite significantly from Stonaker, *et al.* (2009) and Sidorenko, *et al.* (2009) publications which both described decreases in 24-nt siRNA levels akin to n*rpd1*$_{rmr6}$ and *rdr2*$_{mop1}$ as measured by northern blot analysis.

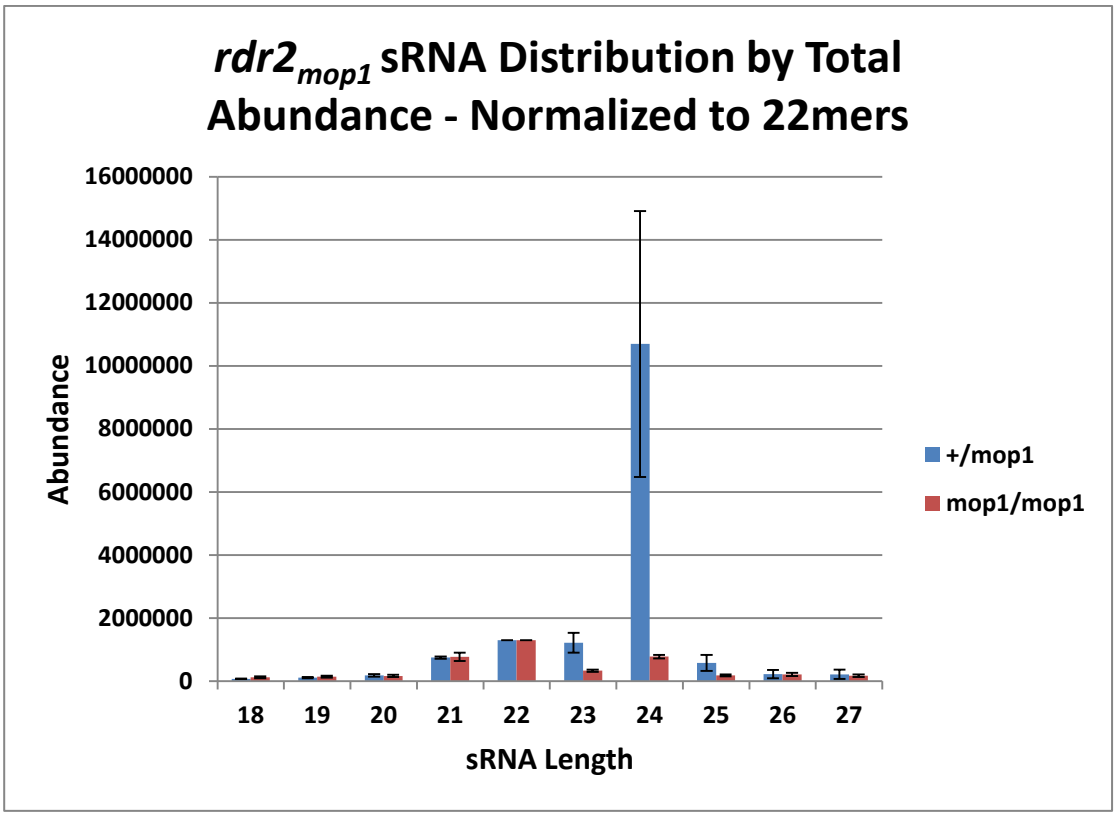

Figure 8: Distribution of total small RNA abundance by size class in n*rpd1*$_{rmr6}$
sRNA distribution plot comparing total expressed small RNAs from five *+/rmr6-1* (+) and four *rmr6-1/rmr6-1* (-) siblings

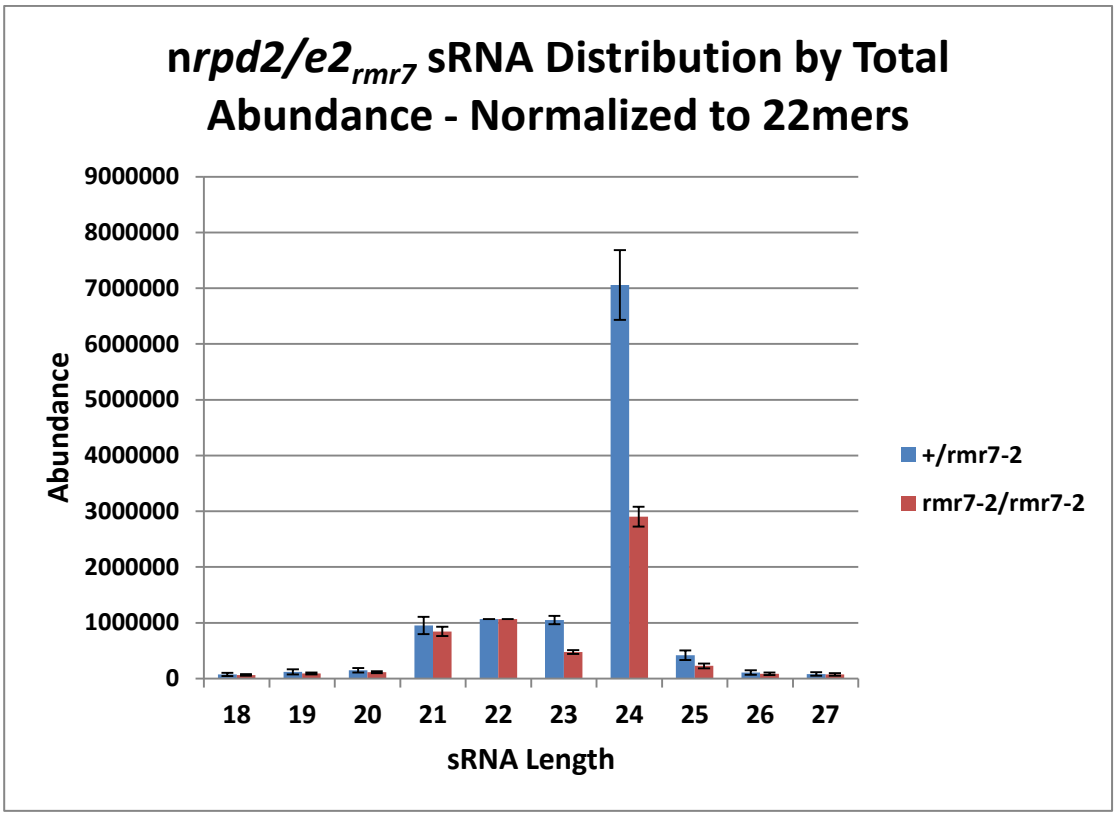

Figure 9: Distribution of total small RNA abundance by size class in n*rpd2/e2*$_{rmr7}$
sRNA distribution plot comparing total expressed small RNAs from four *+/rmr7-2* (+) and six *rmr7-2/rmr7-2* (-) siblings

The sRNA distribution plot shows an overview of all sRNAs expressed, though not all of these will be mapped to the genome. To see the effect of genomic matched sRNAs, we mapped all 24-nt sRNAs to the genome and clustered them into 500 basepair windows for which we can perform library-by-library comparisons. This analysis allows us to determine how these mutations affect small RNAs in various regions of the genome in each mutant comparison. From these pairwise comparisons, we can also perform a qualitative analysis of how *nrpd2/e2*$_{rmr7}$ differs from both *rdr2*$_{mop1}$ and *nrpd1*$_{rmr6}$.

In the cluster analysis of all small RNAs, distinct differences in the expression of small RNAs between mutant and non-mutant of both *rdr2*$_{mop1}$ and *nrpd1*$_{rmr6}$ and *nrpd2/e2*$_{rmr7}$ can be seen (Figure 4). While the majority of clusters appear to be unchanged between all mutant and non-mutant siblings, roughly 10% of clusters in the *rdr2*$_{mop1}$ and *nrpd1*$_{rmr6}$ mutants are downregulated compared to their non-mutant siblings. It also appears that the *nrpd1*$_{rmr6}$ mutation is impacted on a whole genome level more than *rdr2*$_{mop1}$.

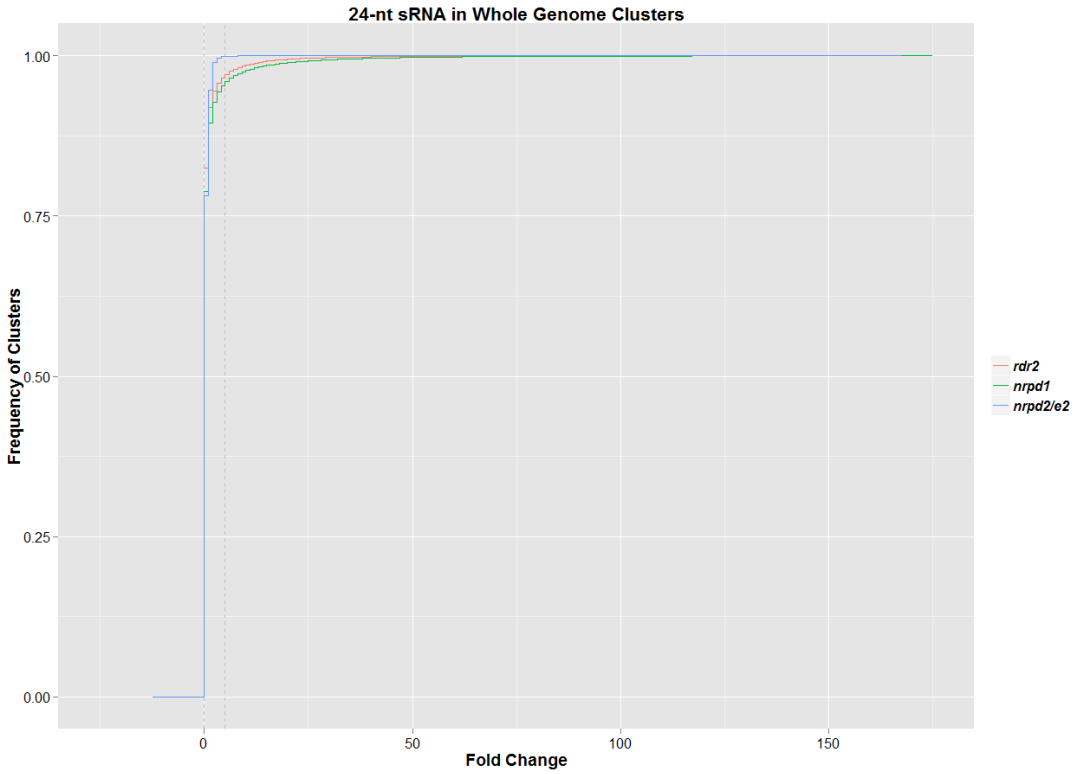

Figure 10: Empirical cumulative distribution function of static clusters
Empirical cumulative distribution function showing the frequency of fold difference across all clusters in each mutant compared to its non-mutant siblings. Positive numbers indicate abundance is greater in non-mutant. While the majority of clusters for all comparisons are relatively unchanged, it can be observed that roughly 10% of *rdr2*$_{mop1}$ and *nrpd1*$_{rmr6}$ mutant clusters are downregulated while the *nrpd2/e2*$_{rmr7}$ mutant doesn't appear to deviate from its non-mutant sibling.

We know that heterochromatic siRNAs are derived from repeat regions of the genome (Axtell, 2013) and that the *rdr2*$_{mop1}$ mutant displayed significant reductions of 24-nt sRNAs in repeat regions (Nobuta, *et al.*, 2008), so we decided to create clusters based on the repetitive regions of the genome. To do this, we use RepeatMasker to find the coordinates of the repetitive regions of the genome and used these to create new clusters. In the repeat regions we again see a similar trend in all mutants (Figure 5). It appears that the *nrpd2/e2*$_{rmr7}$ mutant is quite unchanged while roughly 15% of the *rdr2*$_{mop1}$ and *nrpd1*$_{rmr6}$ mutant clusters are downregulated compared to their non-mutant siblings. In order to determine the level of impact in the repetitive regions, we examined repeat clusters annotated as retrotransposons and transposons individually (Figures 6 & 7).

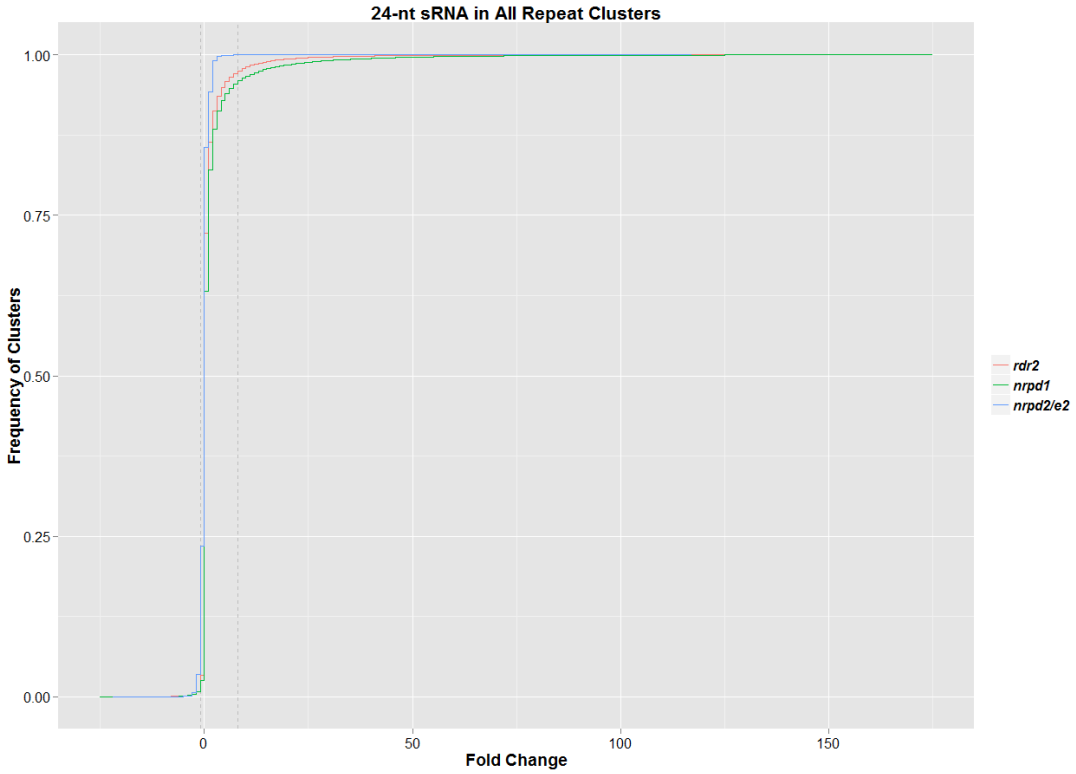

Figure 11: Empirical cumulative distribution function of repeat clusters
Empirical cumulative distribution function showing the frequency of fold difference across all repeat clusters in each mutant compared to its non-mutant siblings. Positive numbers indicate abundance is greater in non-mutant. While the majority of clusters for all comparisons are relatively unchanged, it can be observed that roughly 15% of $rdr2_{mop1}$ and $nrpd1_{rmr6}$ mutant clusters are downregulated while the $nrpd2/e2_{rmr7}$ mutant doesn't appear to deviate from its non-mutant sibling.

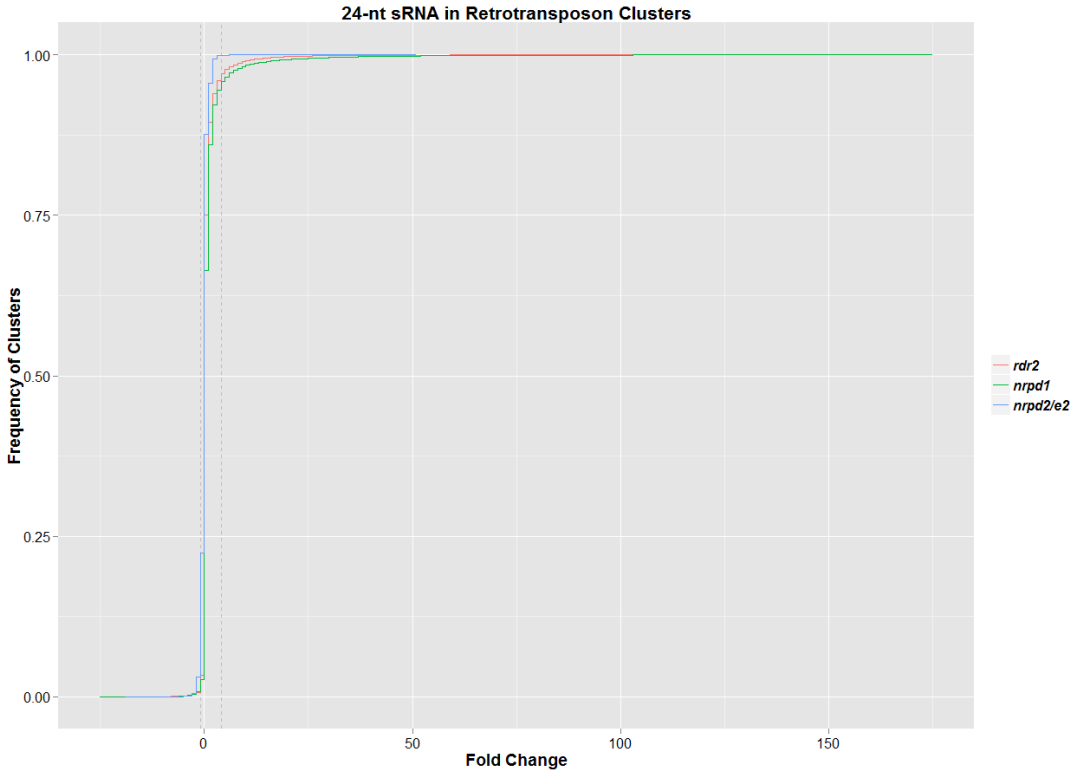

Figure 12: Empirical cumulative distribution function of retrotransposon clusters
Empirical cumulative distribution function showing the frequency of fold difference across all retrotransposon clusters in each mutant compared to its non-mutant siblings. Positive numbers indicate abundance is greater in non-mutant. While the majority of clusters for all comparisons are relatively unchanged, it can be observed that roughly 10% of $rdr2_{mop1}$ and $nrpd1_{rmr6}$ mutant clusters are downregulated while the $nrpd2/e2_{rmr7}$ mutant doesn't appear to deviate from its non-mutant sibling.

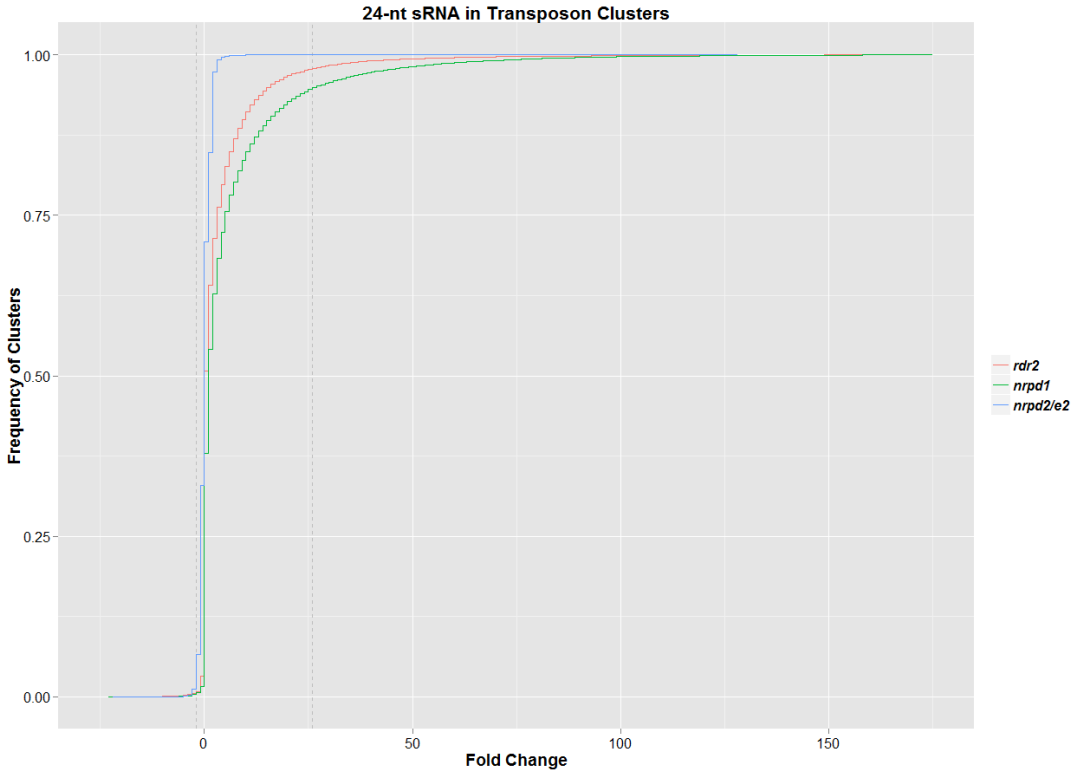

Figure 13: Empirical cumulative distribution function of transposon clusters
Empirical cumulative distribution function showing the frequency of fold difference across all repeat clusters in each mutant compared to its non-mutant siblings. Positive numbers indicate abundance is greater in non-mutant. While the majority of clusters for all comparisons are relatively unchanged, it can be observed that roughly 40% of $rdr2_{mop1}$ and $nrpd1_{rmr6}$ mutant clusters are downregulated while 20% of $nrpd2/e2_{rmr7}$ mutant clusters are downregulated, though the impact does not increase the same level that the other two comparisons do.

We observe that in the retrotransposons, $nrpd2/e2_{rmr7}$ is barely impacted while roughly 10% $rdr2_{mop1}$ and $nrpd1_{rmr6}$ clusters are expressed greater in non-mutants. This is an overall decrease from what was seen in all repeat clusters, so we look to the transposons separately to see the effect there. In the transposons, we observe that about 20% of $nrpd2/e2_{rmr7}$ mutant clusters are downregulated compared to non-mutant clusters. $rdr2_{mop1}$ and $nrpd1_{rmr6}$ are both heavily impacted in the DNA transposons, with about 40% of all mutant clusters being downregulated compared to non-mutant clusters. This finding is consistent with the findings of Nobuta, *et al.* (2008). What is

striking about these results is that while some *nrpd2/e2*$_{rmr7}$ mutant clusters are downregulated, the extent to which they are downgregulated is minimal. Both *rdr2*$_{mop1}$ and *nrpd1*$_{rmr6}$ lines deviate heavily from 1 (where 24-nt sRNA abundance is the same) whereas the *nrpd2/e2*$_{rmr7}$ does deviate a bit from 1, though the level of impact never increases heavily. This suggests that even in the regions for which *rdr2*$_{mop1}$ and *ndprd1*$_{rmr7}$ are impacted most significantly, *nrpd2/e2*$_{rmr7}$ is not affected to the same degree. There are a few more analyses that can be done on an individual loci basis, but due to the differing genetic backgrounds of these mutants, this analysis must be delayed till later. More will be said about this below.

Conclusions of Analysis of Mutants Deficient in Paramutation

Twenty-six libraries were used in this analysis containing small RNA data of *rdr2*$_{mop1}$, *nrpd1*$_{rmr6}$ and *nrpd2/e2*$_{rmr7}$ mutants and their non-mutant siblings. Throughout the analysis of these mutants deficient of paramutation, the decrease of 24-nt siRNAs has been a recurring event. While the near complete removal of 24-nt sRNAs was seen in plants with mutations in maize *RDR2* (Figure 1) and *NRPD1* (Figure 2), the same cannot be said of maize *NRPD2/E2* (Figure 3). From this analysis, there are a few non-mutually exclusive ways in which this data can be interpreted. The first would be that given there exist three paralogs of the maize *NRPD2/E2-like* gene, there may be little to no redundancy between these genes. This can be supported by the 24-nt siRNAs analysis of *nrpd2/e2*$_{rmr7}$ in which an effect was seen on these small RNAs, though the effect is not nearly as complete as in either

rdr2$_{mop1}$ or *nrpd1*$_{rmr6}$. There is little evidence to refute this hypothesis and while the data presented in this thesis may initially implicate partial redundancy, we cannot be sure that this is the case due to the lack of knowledge of how the complete lack of *NRPD2/E2-like* would affect 24-nt siRNA biogenesis.

We also know that Sidorenko *et al.* (2009) saw a massive reduction of 24-nt sRNAs in *nrpd2/e2*$_{mop2}$ +/-, though greater reductions in -/-. This indicates that the heterozygous sibling could already represent a depletion of 24-nt siRNAs and thus its use as a control for comparison to the -/- mutant would not allow us to observe the full loss of 24-nt siRNAs in the *NRPD2/E2-like* mutant. While our analysis of *rdr2*$_{mop1}$ and *nrpd1*$_{rmr6}$ -/- mutants show massive drops relative their +/- siblings, the *rdr2*$_{mop1}$, *nrpd1*$_{rmr6}$ and *nrpd2/e2*$_{rmr7}$ are all from different genetic backgrounds and such we cannot necessarily claim that the decrease in 24-nt siRNAs should be similar for all mutants. Potential to solve for such problems are presented below.

Future Work

Other mutant alleles of *nrpd2/e2*$_{rmr7}$ can be attained which can be instrumental for the determination of significance that maize NRPD2/E2 plays on the production of small RNAs. These alleles have various levels of impact on *NRPD2/E2-like* which should allow for us to learn more about the impact *NRPD2/E2-like* has on the production of 24-nt siRNAs. In addition, numerous libraries of each mutant exist in a different background, i.e. the *mop* mutants which also exhibit a loss of paramutation at various alleles. Due to their different backgrounds, we cannot perform library to

library comparisons of the *rmr* and *mop* mutants; however, similarities should exist as each *mop* mutant affects a locus with which an *rmr* mutant is known to affect as well. This analysis would allow us to learn the shared effects of the mutations despite differing genetic background.

Double mutants also exist in the *mop* mutants affecting both the *RDR2-like* and *NRPD1-like* genes as well as both the *NRPD1-like* and *NRPD2/E2-like* genes. With these double mutants, the combined effects of the mutant genes can be determined so that the individual effects of each gene can be determined. This should be instrumental in determining the significance of the *NRPD2/E2-like* gene in maize and how it affects paramutation. In conjunction with this small RNA data, we have access to numerous RNA-seq libraries for all *mop* and *rmr* mutants that have been mentioned thus far that have yet to be analyzed. With this data, we can identify differentially expressed genes and attempt to learn what loci they regulate through the use of the small RNA data.

While the analysis of all of this data will be quite extensive, the analytical pipeline (Materials and Methods) that has been generated in both projects will significantly speed up the process.

Chapter 4

MATERIALS AND METHODS

Short Read Sequencing of Maize Small RNAs

The maize genome is quite large with version two containing 2,066,432,718 bp, though most of this genome consists of nongenic, repetitive regions. One type of repeat element, DNA transposons, account for more than 75% of the genome (Llaca, *et al.*, 2011). With how repetitive the maize genome is, it is not surprising that a large amount of small RNAs can be mapped to a very high amount of genomic locations. Thus, without filtering out these small RNAs that map to the highly repetitive regions, it would be impossible to map and analyze the data. In order to make this data more manageable and alleviate this problem, small RNAs that map to more than 50 locations in the genome are filtered out of the dataset. There is the potential that this biases the results identified in the data, but about 87% of distinct RNAs are still captured while reducing the total matched locations by about 96%. With the great reduction of total matched locations, the data can be stored and accessed with relative ease while retaining high resolution of small RNA information in the genomic regions that are not highly repetitive.

Small RNA Distribution of *fzt* Libraries

Twelve total libraries were used in the analysis of the *fzt* mutant. These were made up of three biological replicate libraries created from 14 day old seedling and tassel primordia for both the *fzt* mutant and their normal siblings. The abundance of small RNAs was normalized to the sequencing depth as reads per ten million to allow for library to library comparisons. When sequencing was completed, the raw small RNA reads had their adapters removed and then were mapped to the maize genome (AGP version 2). The data was finally uploaded into a MySQL database that was queried to find the normalized counts of all expressed sRNAs from sizes 18-34 nucleotides in length. Replicates were then averaged their standard errors were used to create error bars.

miRNA Profile of *fzt* Libraries

In order to determine the effect of the *fzt* mutant on miRNAs, miRNAs had to be analyzed in both the mutant and its normal siblings calling for the creation of a miRNA profile. All identified maize miRNAs were downloaded into FASTA format from miRBase version 20. miRBase will separate some miRNAs with identical sequences but different targets into multiple entries. For the purpose of this analysis, however, only miRNA sequences are of importance and so multiple miRNA sequences with different names would produce duplicate results. In order to resolve this issue, a python script was written to query the MySQL database and merge all redundant mature miRNAs and their abundances into one entry. Of the 321 total

miRBase sequences, 140 were returned that had some abundance in any of the 12 libraries. For the actual miRNA analysis, seedling and tassel primordia datasets were separated as it could be possible for some miRNAs to be expressed greatly in one tissue but not the other. To limit the analysis to miRNAs with significant abundance, miRNAs with summed abundance across the six libraries had to be at least 50 to be included. In seedling, the top 50 of 74 miRNAs that met this criterion were selected for analysis; for tassel primordia, all 46 miRNAs meeting this criterion were selected for analysis. For each sibling replicate, a fold change was computed and stored. The fold changes across all three replicates were then averaged to compute the averaged fold change across all libraries. The standard errors of the three fold changes could then be computed to generate the error bars for charting.

Target Analysis of miRNAs

Twenty-three miRNAs in the seedling libraries and 14 in the tassel libraries had their targets analyzed to observe the effect of the of reduced miRNA production levels. TargetFinder, a tool which can predict sRNA targets, was modified to identify likely miRNA targets in the maize transcriptome. TargetFinder scores miRNA-target pairs with a score from 0 to 7, with the lower score indicating the more likely match. Targets with scores ≥ 4 were eliminated from the analysis for their potential of being false positives. It is possible that many targets with scores lower than this, even at a score of 0, are not real targets, but no elimination procedure is attainable at this time to clean up this data.

Small RNA Distribution of *rmr* Libraries

Twenty-six total libraries were used in the analysis of the *rmr* mutants created from the immature ears (3-5 cm in length). The 26 libraries consisted of four biological replicate libraries of the $nrpd1_{rmr6}$ mutant and five biological replicates of its normal siblings, six biological replicate libraries of the $nrpd2/e2_{rmr7}$ mutant and four replicates of its normal siblings, and four biological replicate libraries of the $rdr2_{mop1}$ mutant and three replicates of its normal siblings. The abundance of small RNAs was normalized to the sequencing depth as reads per eight million allowing for library to library comparisons. Similar to the *fzt* mutant, when sequencing was completed, the adapters were removed and then small RNAs were mapped to the maize genome. Once in the MySQL database, small RNAs were queried to find the normalized counts of all expressed sRNAs from sizes 18-34 nt. Replicates were once again averaged and their standard errors were used for error bars.

Genome-Wide Clustering of Maize Small RNAs

The process of static clustering begins by segmenting the genome is broken up into fixed or "static" window sizes (for this project 500 bp is used producing 4,132,872 clusters, though it can be modified to any size). Then, each of these clusters shows a representation of all small RNAs mapping to it as a hits normalized abundance for each cluster. The hits-normalized abundance is calculated simply as the summation of the small RNA expression abundance divided by its hits to the genome. This can then allow for a one-to-one comparison of clusters across multiple libraries.

Hits normalized abundance values were assigned to clusters for all small RNA abundances as well as only 21-nt, 22-nt and 24-nt sRNAs.

In addition to this static clustering, annotation-based clustering can also be done to cluster small RNAs based on repeat regions of the genome. The repeat-based clustering algorithm groups small RNAs together if they map to the same RepeatMasker predicted repeat regions in the genome (resulting in 1,268,787 clusters). This clustering algorithm will result in some small RNAs that will not be represented in a cluster. This allows for data to be minimized and queried in order to answer specific questions about particular regions of the genome.

Development of an Automated Cluster Analysis Pipeline

While the clustering helps reduce the data to a manageable size, it is still very difficult to look at all clusters to determine points of difference between mutant and wildtype libraries. This is not the only project which uses the cluster analysis, so to help speed up the process in the future for others lab members; I developed an automated pipeline to summarize what the cluster data returns. This pipeline contains a few Python scripts which requires limited interaction from the user which is mostly limited to the library names and which libraries are being compared. The process is outlined here for clarity.

The first script in this pipeline takes a list of libraries (requires a specific format) of samples and replicates. Using the sample names, columns are added to the MySQL clustering table of the replicate average hits-normalized abundance for each

sample for each cluster. The next step in the pipeline uses another input file of the libraries that are to be compared. This script then creates fold-change columns in the clustering table which can be accessed quickly to determine the clusters of with high amounts of differences. The next script in the pipeline looks for all comparison columns in the clustering table and selects the differentially expressed clusters in either direction. From this, numerous comma-separated values are created containing the cluster identifier, the hits normalized abundance values for each library, the fold difference for each library and the binary logarithm for each cluster. These files can be referred to for specific cluster values for manual analysis as well as the generation of the empirical cumulative distribution function figures. Each of these was generated by exporting these outputs to R and plotting the empirical cumulative distribution function with ggplot2. In addition, files containing the summary of fold differences as well as their binary logarithms across all library comparisons are returned for which fold difference histograms may be assembled. The final script of the pipeline is useful for the comparison of multiple mutants to one another. It allows for the identification of differentially regulated clusters in multiple libraries and can allow for the simple analysis of what is commonly differentially expressed and what is unique.

REFERENCES

Adams, I. R., & Meehan, R. R. (2013). From Paramutation to Paradigm, *9*(5), 9–10. doi:10.1038/nrg3435.3.

Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J. E., White, J., … Chandler, V. L. (2006). An RNA-dependent RNA polymerase is required for paramutation in maize. *Nature*, *442*(7100), 295–8. doi:10.1038/nature04884

Arteaga-vazquez, M., Sidorenko, L., Rabanal, F. A., Shrivistava, R., & Nobuta, K. (2010). RNA-mediated trans -communication can establish paramutation at the b1 locus in maize. doi:10.1073/pnas.1007972107/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1007972107

Axtell, M. J. (2013). Classification and comparison of small RNAs from plants. *Annual Review of Plant Biology*, *64*, 137–59. doi:10.1146/annurev-arplant-050312-120043

Bouché, N. (2010). New insights into miR398 functions in Arabidopsis. *Plant Signaling & Behavior, 5*(6), 684–686. doi:10.1111/j.1365-313X.2010.04162.x.e

Bruggmann, R., Bharti, A. K., Gundlach, H., Lai, J., Young, S., Pontaroli, A. C., … Messing, J. (2006). Uneven chromosome contraction and expansion in the maize genome, 1241–1251. doi:10.1101/gr.5338906.9

Chen, K., & Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews. Genetics*, *8*(2), 93–103. doi:10.1038/nrg1990

Dorweiler, J. E., Carey, C. C., Kubo, K. M., Hollick, J. B., Kermicle, J. L., & Chandler, V. L. (2000). Mediator of Paramutation1 Is Required for Establishment and Maintenance of Paramutation At Multiple Maize Loci. *The Plant Cell*, *12*(11), 2101–18. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=150161&tool=pmcentrez&rendertype=abstract

Elbashir, S. M. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Development*, *15*(2), 188–200. doi:10.1101/gad.862301

Finnegan, E. J., & Matzke, M. a. (2003). The small RNA world. *Journal of Cell Science*, *116*(Pt 23), 4689–93. doi:10.1242/jcs.00838

Hollick, J. B., Kermicle, J. L., & Parkinson, S. E. (2005). Rmr6 maintains meiotic inheritance of paramutant states in Zea mays. *Genetics*, *171*(2), 725–40. doi:10.1534/genetics.105.045260

Jeong, D, Green, P (2013) The Role of Rice microRNAs in Abiotic Stress Responses. Journal of Plant Biology, 56:187-197.

Lisch, D., Carey, C. C., Dorweiler, J. E., & Chandler, V. L. (2002). A mutation that prevents paramutation in maize also reverses Mutator transposon methylation and silencing. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(9), 6130–5. doi:10.1073/pnas.052152199

Llaca, V., Campbell, M. a., & Deschamps, S. (2011). Genome Diversity in Maize. *Journal of Botany*, *2011*, 1–10. doi:10.1155/2011/104172

Lu, C., Kulkarni, K., Souret, F. F., Muthuvalliappan, R., Tej, S. S., Poethig, R. S., … Meyers, B. C. (2006). MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase - 2 mutant, (302), 1276–1288. doi:10.1101/gr.5530106.1276

Nobuta, K., Lu, C., Shrivastava, R., Pillay, M., De Paoli, E., Accerbi, M., … Meyers, B. C. (2008). Distinct size distribution of endogeneous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(39), 14958–63. doi:10.1073/pnas.0808066105

Pikaard, C. S., Haag, J. R., Pontes, O. M. F., Blevins, T., & Cocklin, R. (2012). A transcription fork model for Pol IV and Pol V-dependent RNA-directed DNA methylation. *Cold Spring Harbor Symposia on Quantitative Biology*, *77*, 205–12. doi:10.1101/sqb.2013.77.014803

Pikaard, C. S., & Tucker, S. (2009). RNA-silencing enzymes Pol IV and Pol V in maize: more than one flavor? *PLoS Genetics*, *5*(11), e1000736. doi:10.1371/journal.pgen.1000736

Schauer, S. E., Jacobsen, S. E., Meinke, D. W., & Ray, A. (2002). DICER-LIKE1: blind men and elephants in Arabidopsis development. *Trends in Plant Science*, *7*(11), 487–91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12417148

Sidorenko, L., Dorweiler, J. E., Cigan, a M., Arteaga-Vazquez, M., Vyas, M., Kermicle, J., … Chandler, V. L. (2009). A dominant mutation in mediator of paramutation2, one of three second-largest subunits of a plant-specific RNA polymerase, disrupts multiple siRNA silencing processes. *PLoS Genetics*, *5*(11), e1000725. doi:10.1371/journal.pgen.1000725

Stonaker, J. L., Lim, J. P., Erhard, K. F., & Hollick, J. B. (2009). Diversity of Pol IV function is defined by mutations at the maize rmr7 locus. *PLoS Genetics*, *5*(11), e1000706. doi:10.1371/journal.pgen.1000706

Zhu, H., Zhou, Y., Castillo-González, C., Lu, A., Ge, C., Zhao, Y.-T., … Zhang, X. (2013). Bidirectional processing of pri-miRNAs with branched terminal loops by Arabidopsis Dicer-like1. *Nat Struct Mol Biol*, *20*(9), 1106–1115. Retrieved from http://dx.doi.org/10.1038/nsmb.2646