Effects of Hybridization on Heterochromatic Small Interfering RNA and Gene Expression in Zea mays

by

Travis Korry Coleman

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Plant Agriculture

Guelph, Ontario, Canada

© Travis Coleman, January, 2017
ABSTRACT

EFFECTS OF HYBRIDIZATION ON HETEROCHROMATIC SMALL INTERFERING RNA AND GENE EXPRESSION IN ZEA MAYS

Travis Korry Coleman
University of Guelph, 2016

Advisor:
Associate Professor Lewis N. Lukens

Despite decades of research, the molecular nature of heterosis is not completely understood. Heterosis for quantitative traits is controlled by the cumulative effects of multiple genes and regulatory elements, each of which may exhibit differing modes of action. While dominance and over-dominance theories can account for some of the genetic control of heterosis, neither provides a complete account. There is a growing body of evidence that implicates small RNAs as non-coding elements which may play a role in the manifestation of heterosis upon hybridization. In particular, 24-nt heterochromatic small interfering RNAs (hetsiRNAs) may play a role in mediating trans-genomic interactions via DNA methylation when two genomes come into contact in an F1 nucleus. Recent research in Arabidopsis thaliana has demonstrated that hetsiRNAs show non-additive expression upon hybridization, with the majority of hetsiRNAs being downregulated. This research seeks to examine such trends in non-additive hetsiRNA expression in commercial maize germplasm. Small RNA were isolated and deep sequenced from leaf tissue samples of two inbred lines and their F1 hybrid. In order to examine the effects of reducing hetsiRNAs in hybridization, each genotype was also sampled in each of two mediator of paramutation 1 (mop1) allelic states. Results demonstrate that mean hetsiRNA expression is significantly reduced in the hybrid relative to its parents and that hetsiRNA expression in mop1-1
is almost entirely absent. The F1 hybrid also demonstrated a marked reduction in hetsiRNA diversity, with 41% of hetsiRNA clusters being expressed in one or both parents but not the hybrid. Similarly, classifying hetsiRNA expression in the hybrid relative to the expected midparent value demonstrated that the majority of hetsiRNA clusters are expressed below the midparent level, with many below the low parent. Since hetsiRNAs are known to be involved in the regulation of asymmetric DNA methylation, this general downregulation may be related to mechanisms that evolved to promote genome integrity when two independent genomes come into contact. These results have further implications for the development of crop species and the notion of an epigenome that may affect phenotype yet may not exhibit patterns of Mendelian inheritance along with the genome.

Recent findings have also noted the enriched localization of hetsiRNAs to regions of the genome flanking genes. This raises the possibility that hetsiRNA may play a role in regulating gene transcription by influencing DNA methylation and heterochromatin formation around those genes. This research also seeks to examine the association between gene transcript expression and expression of hetsiRNAs in hybrid maize. Two commercial maize inbred lines and their F1 hybrid were subjected to deep sequencing for small RNA and mRNA. In order to examine the effect of knocking out hetsiRNA production on gene expression, both Mop1 and mop1-1 allelic states were sequenced for each genotype. Association of hetsiRNA reads and mRNA transcription was done by clustering analysis. Localization of hetsiRNA clusters within the boundaries of a genic frame demonstrated a strong enrichment of hetsiRNA clusters in the +/- 1 kb flanks of genic regions. Expression level analyses indicate a mean downregulation of both hetsiRNA and mRNA expression relative to expected midparent values for a small but significant number of genes. Furthermore, knockout of hetsiRNA production in the mutant
mop1-1 demonstrated a trend toward upregulation of genes relative to the wildtype. Together these results suggest a model in which hetsiRNAs function to regulate heterochromatin formation in gene boundaries which may subsequently regulate gene expression for a subset of genes.
Acknowledgements

I would like to thank my advisor, Dr. Lewis Lukens for his support and guidance throughout the course of my research. I would also like to thank my advisory committee, Dr. Liz Lee, Dr. Steven Rothstein, and Dr. Daniel Ashlock for their time and advice.

I would also like to thank DuPont Pioneer for providing financial support and work time flexibility, without which this degree would not have been possible. I would especially like to thank Dr. Steve King and Dr. Gustavo Garcia who encouraged me to pursue a Ph.D. degree and provided me with much needed support throughout this process.

Most importantly I would like to thank my wife Alison and my two daughters, Madelyn and Nora for their love and patience throughout this journey. Juggling the commitments of a young family, a demanding career, and a Ph.D. degree has been an often difficult process and I could not have done it without their amazing support.
# Table of Contents

Abstract

ii

Acknowledgements

v

Table of Contents

vi

List of Tables

viii

List of Figures

ix

List of Abbreviations

xi

Chapter 1: Thesis Overview

1

Chapter 2: Literature Review

4

2.1: Heterosis in Crop Species

4

2.2: Small RNA Classification, Biogenesis, and Function

9

2.3: RNA-directed DNA Methylation

16

2.4: RdDM and Gene Expression

21

2.5: Heritability of Epigenetic States

26

2.6: Epigenetics, Hybridization, and Heterosis

31

Chapter 3: hetsiRNA are Downregulated and Expressed non-Additively in Hybrid Maize

35

3.1: Abstract

35

3.2: Introduction

36

3.3: Materials and Methods

39

3.3.1: Genetic Material

39

3.3.2: Genotyping

43

3.3.3: Tissue Sampling and RNA Extraction

43

3.3.4: siRNA Sequencing and Analysis Pipeline

44
List of Tables

Table 1: Phenotypic effects of mopl on flowering time__________________________41
Table 2: Total reads sequenced per sample and resultant 24-nt reads _______________52
Table 3: Raw mRNA paired-read sequencing and alignment statistics_______________84
Table 4: Gene expression statistics for mRNA-sequencing alignments______________85
Table 5: Gene Ontology term enrichment analysis for genes significantly upregulated
   in mopl-1 backgrounds____________________________________________________90
Table 6: Annotated genes showing a significant deviation from expected midparent value
   and detectible siRNA expression in inbred A or B yet a complete lack of siRNA
   production in hybrids____________________________________________________101
List of Figures

Figure 1: Schematic representation of samples used for small RNA and mRNA sequencing experiments .......................................................... 42

Figure 2: Workflow of small RNA sequencing read processing pipeline ................. 47

Figure 3: Diagrammatic representation of stages 1 and 2 of siRNA meta-cluster analysis ... 48

Figure 4: Raw small RNA read distribution as a mean value across all samples after quality control and adapter trimming ...................................... 53

Figure 5: Expression of 24-nt siRNA prior to cluster analysis and filtering of conserved mop1 independent reads detected in both wildtype Mop1 and mutant mop1-1 samples ........................................................................ 54

Figure 6: Frequency distribution of 24-nt siRNA sequences by read count ............... 55

Figure 7: Size distribution of siRNA meta-clusters as a mean of all genotypes after Stage 2 of siRNA meta-cluster analysis .................................................. 59

Figure 8: Unique raw 24-nt reads prior to clustering analysis and filtration of mop1 independent siRNA species ......................................................... 60

Figure 9: Global net 24-nt siRNA meta-cluster expression as a result of a) replicated analysis and b) pooled replicates .......................................................... 61

Figure 10: Venn diagrams demonstrating co-expression of unique 24-nt hetsiRNA meta-clusters in wildtype Mop1 and mutant mop1-1 prior to filtering of mop1 independent 24-nt siRNA species .................................................. 62

Figure 11: Comparison between unique 24-nt hetsiRNA meta-clusters among genetic inbred and hybrid backgrounds and within mop1 genotypes prior to filtering mop1-1 independent 24-nt siRNA species ................................................ 63
Figure 12: siRNA meta-cluster expression deviation from expected midparent value in Hybrid AB.

Figure 13: Replicated siRNA meta-cluster analysis of hetsiRNA expression in Hybrid AB relative to expected midparent value derived from inbreds A and B.

Figure 14: Pooled siRNA meta-cluster analysis of Hybrid AB expression relative to high, low, and expected mid-parent values of parental inbreds A and B.

Figure 15: Workflow of mRNA-seq processing, alignment, and differential expression analysis pipeline.

Figure 16: Diagrammatic representation of three stage siRNA/gene meta-cluster analysis.

Figure 17: Differential expression of genes in Mop1 relative to mop1-1 genotypes.

Figure 18: Distribution of differential gene expression in Inbred A, Inbred B, and Hybrid AB at individual genes found to be significantly different between Mop1 and mop1-1.

Figure 19: Significant Gene Ontology term enrichment for biological processes associated with differentially expressed genes between Mop1 and mop1-1.

Figure 20: Significant Gene Ontology term enrichment for molecular functions associated with differentially expressed genes between Mop1 and mop1-1.

Figure 21: Localization of siRNA meta-clusters within siRNA/gene meta-clusters.

Figure 22: siRNA/gene meta-cluster expression deviation from expected midparent value for 75 loci showing statistically significant deviation from expected midparent value.

Figure 23: Standardization of 24-nt read datasets for hetsiRNAs detected in inbred A or B but absent from hybrid AB.
**List of Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>Base Pair</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>Chromatin Immunoprecipitation Sequencing</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy Number Variant</td>
</tr>
<tr>
<td>DMR</td>
<td>Differentially Methylated Region</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>easiRNA</td>
<td>Epigenetically Activated Small Interfering RNA</td>
</tr>
<tr>
<td>epiRIL</td>
<td>Epigenetic Recombinant Inbred Line</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>FGS</td>
<td>Filtered Gene Set</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragments Per Kilobase Million</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>hetsiRNA</td>
<td>Heterochromatic Small Interfering RNA</td>
</tr>
<tr>
<td>HP</td>
<td>High Parent</td>
</tr>
<tr>
<td>IBM</td>
<td>Intermated B73 x Mo17</td>
</tr>
<tr>
<td>kb</td>
<td>Kilobase</td>
</tr>
<tr>
<td>LP</td>
<td>Low Parent</td>
</tr>
<tr>
<td>miRNA</td>
<td>Micro RNA</td>
</tr>
<tr>
<td>MITE</td>
<td>Miniature Inverted-repeat Transposable Element</td>
</tr>
<tr>
<td>mop1</td>
<td>mediator of paramutation 1</td>
</tr>
<tr>
<td>MPV</td>
<td>Mid-Parent Value</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
</tbody>
</table>
mRNA-seq  Messenger RNA Sequencing  
Mu-PCR  Mutator Polymerase Chain Reaction  
natsiRNA  Natural-antisense Small Interfering RNA  
NIL  Near-Isogenic Line  
nt  Nucleotide  
OPV  Open Pollinated Variety  
PAV  Presence-Absence Variant  
PCR  Polymerase Chain Reaction  
pre-miRNA  Pre-Micro RNA  
pri-miRNA  Primary Micro RNA  
QC  Quality Control  
QTL  Quantitative Trait Locus  
RdDM  RNA-directed DNA Methylation  
RISC  RNA Induced Silencing Complex  
RNA  Ribonucleic Acid  
RPM  Reads Per Million  
rRNA  Ribosomal RNA  
siRNA  Small Interfering RNA  
SNP  Single Nucleotide Polymorphism  
tasiRNA  Trans-Acting Small Interfering RNA  
TIR  Terminal Inverted Repeat  
tRNA  Transfer RNA
Chapter 1: Thesis Overview

Heterosis generally refers to the phenotypically observable phenomenon wherein the hybrid progeny of two inbred lines exhibit superior attributes to that of the parental lines. This superiority may be in the form of enhanced adaptation, greater fitness, increased biomass, or any number of other quantitative traits of interest. In particular to applied researchers and plant breeders, the power of heterosis has been harnessed in certain crop species to both develop a consistent and predictable crop seed production system and to enhance various agronomic traits relating to yield and adaptation. As such, an accurate and complete understanding of the underlying mechanisms of heterosis is a desirable goal and one with great potential benefit to both society as a whole and also to industry. An understanding of the molecular basis of heterosis in particular would facilitate use of modern and novel molecular techniques to further harness the power of heterosis for crop development beyond what has been already accomplished in the eras of traditional and current molecular plant breeding.

Despite decades of research, heterosis is still not entirely understood, likely due to the complexity of and interactions among genetic, epigenetic, and environmental factors that underlie the phenomenon. A significant component of heterosis can be attributed to the function of non-additive gene action in the F1 progeny. This is perhaps one of the more elusive aspects of genetic analysis and characterization from the perspective of those seeking to utilize heterosis for crop improvement. Maize in particular offers a number of opportunities as both a model system and a crop to elucidate some of these mechanisms.

This thesis generally seeks to explore a number of aspects relating to the genetic and epigenetic factors influencing heterosis. Specifically the research is concerned with analyzing
broad patterns and global effects of hybridization on gene transcript expression, small interfering RNA expression, and the interplay between the two.

Chapter 2 will provide an overall literature review on the general topic of heterosis in plants and will focus particularly on those aspects of heterosis related to this research. Some of these topics include modes of gene expression related to heterosis, epigenetic interactions and imprinting, and the known interactions between small interfering RNA and DNA methylation and gene expression. Known types of small RNA, in particular small interfering RNA will be discussed, including modes of biogenesis, genetic structure, and their effects on gene action and heterochromatin formation.

Chapter 3 is comprised of the portions of this research pertaining to the effects of hybridization on small interfering RNA populations and expression. A panel of three inbreds and their hybrids were subjected to deep sequencing of small RNA species collected in leaf tissue samples. The small RNA sequences were analyzed in a processing pipeline that included a series of filtering processes as well as alignment and clustering analyses. Furthermore, the study employed the use of inbred and hybrid near-isogenic lines and the mutant mediator of paramutation 1 (mop1-1) to examine the effect of reduced small interfering RNA. The objective of this portion of the research was to quantify siRNA production in terms of additive and non-additive expression profiles as compared to expected Midparent Value (MPV). Other goals included examining presence/absence expression profiling in the inbreds and the hybrids.

Chapter 4 deals with the interplay between the mRNA analyses and the siRNA analysis presented in the preceding chapter. Small interfering RNA were further subjected to a meta-clustering analysis that associated siRNA cluster alignments with annotated genes. By doing this, siRNA expression profiles could be compared to mRNA transcript profiles at the level of
individual genes and associations between siRNA expression and non-additive gene expression examined. Furthermore this analysis facilitates a detailed review of the most highly significant and/or interesting loci.

Chapter 5 provides an overall discussion and synthesis of results presented in the previous chapters. Overall findings will be discussed as well as limitations to the research and potential avenues for future study.
Chapter 2: Literature Review

2.1: Heterosis in Crop Species

Heterosis, also known as hybrid vigour, refers to an increase in overall size or growth rate of an organism as compared to that of its parents. Often heterosis is considered in the context of economically or agronomically beneficial traits such as grain yield, leaf initiation, or plant biomass, among many others. It also enables consistent and reliable seed production practices as parental seed may be conveniently maintained as pure inbred lines and seed to be used for grain or biomass production may be created by hybridization of those inbred lines. Maize (Zea mays L.) is one such species in which breeders and crop growers have been able to harness the power of heterosis to drive enormous gains in crop and food production. Hybrid maize has indeed been quite a success story in terms of yield gains over time. Since the development and introduction of hybrid maize in the 1920’s, grain yield potential of hybrids has increased at an estimated rate of 0.95 bushels/acre/year (Duvick et al., 2004). Similarly, real world on-farm averages for United States corn production have seen comparable increases in yield at 0.95 bushels per acre for the period between 1930-2001 which represents the era of modern corn hybrid breeding (Duvick, 1984; Duvick et al., 2004). It is important to note however, that the data of United States corn production are non-linear with respect to yield gains over years. In the period between 1930 and 1950 the regression slope is decidedly lower and it is not until hybrid corn was more universally adopted in North America (~1950-1960’s on) that the regression slope both increases and becomes linear. This can be attributed in part to the fact that the Open Pollinated Varieties (OPV) in widespread use prior to hybrid maize were not able to take full advantage of heterosis. While the major factors influencing yield gains over the same period are improvement in genetic potential (via breeding) and agronomic advances (estimated at 50-60% genetic improvement 40-
50% agronomic improvement, (Russell, 1991)), it is heterosis and hybrid crops that have facilitated the rate of change on the genetic improvement side by allowing breeding programs to make concurrent gains in both heterotic groups. As such, it is of particular interest to further understand the underlying molecular basis of heterosis in order to take advantage of modern molecular techniques and to integrate them into crop breeding and production systems with the aim of increasing desirable traits in a more targeted and specific fashion.

While heterosis has been a topic of genetic interest since the early 20\textsuperscript{th} century (Bruce, 1910; Shull, 1909, 1910), over 100 years of research has not provided a clear picture of the underlying genetic or molecular basis. Two major theories persist: dominance and over-dominance.

The theory of dominance as a mechanism for heterosis was first postulated by Jones, (1917). By understanding the Mendelian laws of genetics and Morgan’s theory of linkage, Jones described a theory whereby various genes throughout the genome have a complementary and cumulative effect on a trait demonstrating heterosis. Over time and within a population, these genes accumulate favourable dominant alleles and unfavourable recessive alleles via \textit{de novo} mutation. Thus it stands to reason that heterosis would be manifested as the cumulative effects of dominance at multiple loci (Davenport, 1908; East, 1936; Jones, 1917). For example, consider the cross AxB where genes $C$, $D$, and $E$ have a purely dominant, favourable effect and contribute to a heterotic phenotype. We allow the dominant alleles to have a favourable quantitative effect of +1 each and null recessives to have a quantitative effect of 0. We also allow inbred line A to have genotype CCddEE and inbred line B to have genotype ccDDee. Thus inbred A has a phenotypic value of 2 and inbred B a value of 1. The resulting hybrid F1 progeny are
homogeneous with a genotype of CcDdEe and have a phenotype attributed to the cumulative effect of the dominant C, D, and E alleles (a phenotypic value of 3).

The original dominance theory was developed at a time when the molecular and genetic knowledge base was quite limited. In modern times, the theory can also be extended to include copy number variation (CNV) and presence-absence variation (PAV). Maize has been demonstrated to have many instances of low genetic collinearity or synteny among intraspecific inbred lines (Fu and Dooner, 2002). Further studies have estimated that CNVs between the maize inbred lines B73 and Mo17 are on the order of 100’s of genes while PAVs are estimated to be on the order of 1000’s of genes (Springer et al., 2009). This theoretically could have a significant contribution to any dominance effects as the accumulation of multiple genes with dominance effects (in the case of CNVs) or the complementation of numerous null or nonexistent alleles with a functional copy (in the case of PAVs) could have a large effect on a heterotic phenotype for those loci with dominant effects contributing to heterosis.

One major limitation of the dominance theory is that it is contingent upon inbred lines containing a certain load of deleterious or unfavourable recessive alleles in order to maintain the heterotic response (Birchler et al., 2010; East, 1936). This however is counter to the principles of natural selection and even more so artificial selection and breeding. Considering the previous example, if allele C has a dominant and favourable effect it would be selected for in breeding crosses involving inbred B (genotype ccDDDee) if the other parent in the cross carried the dominant C allele. Progeny resulting from the cross would be fixed as genotype CCDDDee and when crossed to inbred A (CCddEE) the heterotic effect would be lost at locus C. Expanded to all genes globally this would have a quantitative effect on complex traits such as grain yield and one would expect a reduction in the magnitude of heterosis with breeding and selection, over
time. Furthermore a model consisting of pure dominance as the underlying mechanism for heterosis would imply that through selection one would be able to recover an inbred line with vigour equal to that observed in the hybrid. This was not the case however when Duvick et al. analyzed the rate of hybrid yield gain over time and compared it compared with the rate of parental inbred yield gain, concluding that the rates of increase were equivalent (Duvick, 1999). This suggests that the magnitude of heterosis as defined by hybrid yield relative to their parents has not diminished over time. Thus it is reasonable to conclude that a scenario of simple dominance, while it may represent a significant element of heterosis is certainly not a complete explanation.

Another prevalent theory on the underlying basis of heterosis is that of “over-dominance” (Shull, 1908). The main supposition to the over-dominance theory is that the actions of individual heterozygous loci cause hybrid vigour. An example of this type of gene action in maize comes from the pl locus which encodes a transcription factor controlling anthocyanin production (Hollick and Chandler, 1998). In this case, the Pl’-mah allele exhibits dominant low-expression when paired with the high-expression but unstable Pl-Rh allele. When paired with other pl alleles in a heterozygous state however, the Pl’-mah allele exhibits an expression level above the high homozygous parent. Furthermore, the Pl-Rh allele is known to participate in epigenetic interactions and the phenomenon of paramutation which is an interaction between two alleles of a particular locus, where one allele causes a heritable change in the other allele (Hollick et al., 2000). Results such as these demonstrate that heterotic effects are not exclusively the result of dominant allele complementation and that a certain degree of heterosis is controlled by context dependent and allele specific interactions on the molecular level. Furthermore, examples such as this illustrate the possibility that epigenetics can play a role in establishing
hybrid vigour, a concept that will be discussed later. Unlike the dominance hypothesis however, overdominance requires that an inbred line with vigour equal to that of the hybrid can never be recovered from the hybrid owing to the fact that overdominant loci will be fixed in a homozygous state in any inbred lines. Overdominance has also been implicated in a number of other studies including tomato *SINGLE FLOWER TRUSS (SFT)* and its orthologue Arabidopsis *FLOWERING LOCUS T* (Krieger et al., 2010). Examples of overdominance tend to focus on specific examples such as these although there are cases, specifically that of paramutation, where there is growing evidence of more widespread effects across the genome (Hollick et al., 1997; Regulski et al., 2013).

Aside from dominance and over-dominance, linkage disequilibrium can also play a role in heterosis. Dominant alleles in repulsion phase can result in a situation termed pseudo-overdominance where dominant gene action at each tightly linked locus results in a situation where heterozygous loci appear to be the cause of increased vigour where in fact it is the cumulative action of dominance, confounded by linkage disequilibrium. In at least one example, a QTL for heterosis was observed to have overdominant action but upon completion of fine mapping and further analysis it was discovered to be the action of two closely linked dominant genes (Graham et al., 1997; Stuber et al., 1992). Linkage disequilibrium also plays a role in the perseverance of those minor effect deleterious alleles (described as “genetic load” by some) that are a condition of dominant heterosis (large effect deleterious alleles were presumably purged from germplasm in crop species that have undergone extensive selection) (Charlesworth and Charlesworth, 1999; Schnable and Springer, 2013). These deleterious alleles may persist largely because of negatively correlated trait effects and linkage disequilibrium, as selection for one locus causes linkage drag at the locus linked in repulsion phase. Deleterious alleles may also
persist because they are located in pericentromeric regions with very low recombination rates. Although it seems unlikely at first under the dominance theory for deleterious alleles to persist in the face of intense selection pressure, it must be noted that these instances of linkage disequilibrium along with negative correlations in selecting among multiple traits certainly provide a case for inbreds retaining a number of undesirable deleterious alleles.

On the whole, heterosis is likely to result from an accumulation of both dominant and over-dominant effects in summation across many genes. Moving beyond the scope of classical quantitative theories of dominance and over-dominance, there is a growing body of evidence that suggests heterotic effects may be furthermore subject to epistatic and epigenetic effects. Non-coding elements of the genome are also likely to play a role in trans-mediated epigenetic effects related to heterosis, as will be discussed further.

2.2: Small RNA Classification, Biogenesis, and Function

Small RNAs are non-coding transcripts produced from endogenous coding or non-coding genomic loci. Evidence suggests that they originally evolved as a response to viral attacks and that subsequent duplication of genes encoding proteins involved in small RNA biogenesis allowed subfunctionalization and evolution of various subclasses of small RNA, each with different modes of biogenesis and different functions of the mature molecules (Borges and Martienssen, 2015; Mallory and Vaucheret, 2006). Generally, small RNAs characterized to date range from approximately 20-24 nt in length and are derived from helical RNA duplex precursor molecules. Most subclasses of small RNAs share similarities in processing which begins with the formation of double stranded RNA duplexes from complementary RNA sequences. DICER-LIKE (DCL) proteins are then involved in cleavage of the RNA duplexes and ARGONAUTE
(AGO) proteins are involved in the formation of an RNA Induced Silencing Complex (RISC) which binds the mature, single stranded small RNA molecule. Depending on the specific small RNA sub-group and pathway, the RISC and bound RNA subsequently target either RNA molecules for degradation or translational inhibition, or genomic DNA sequences for initiation and maintenance of RNA-directed DNA Methylation (RdDM) (Borges and Martienssen, 2015).

Classification of small RNA groups can be made on a number of different criteria although typically they are grouped according to a combination of a) biogenesis mechanism of the immature RNA intermediates and b) function of the mature small RNA molecule (Axtell, 2013). Many variations exist on these groups and further sub-classification exists (Axtell, 2013) although for the purposes herein, small RNA will be classified into four main categories according to biogenesis and function: microRNA (miRNA), natural-antisense RNA (natsiRNA), secondary siRNAs, and heterochromatic siRNA (hetsiRNA).

The first major class of small RNA (miRNA) is defined by a conserved structure of self-complementarity in the precursor RNA molecule, the absence of RNA-dependent RNA polymerase in the biogenesis pathway and by their involvement with post transcriptional gene silencing in trans. miRNAs were discovered in Caenorhabditis elegans as a molecular mechanism underlying developmental regulation (Lee et al., 1993). They are now are relatively well understood, partly owing to their highly conserved biogenesis and specific mode of gene targeting action. miRNA are initially transcribed as methylguanylate capped and polyadenylated primary-miRNA (pri-miRNA). The pri-miRNA contain specific regions of sequence self-complementarity which cause formation of an imperfect loop or hairpin structure with a segment of double stranded RNA. The 5’ cap and polyA tail are removed from the pri-miRNA to form pre-miRNA which are then exported from the nucleus. Pre-miRNA then mainly rely on DCL1
proteins to process the hairpin precursor into mature miRNAs by cleaving the double-stranded region into smaller, generally 21-nt, RNA duplexes. After DCL catalyzes the release of miRNA, HUA ENHANCER1 (HEN1) methyltransferase methylates 2’-OH of duplexes to protect them from poly-uridylation and subsequent degradation (Yang et al., 2006). One strand of the mature miRNA is then loaded into an AGO-containing RISC (Voinnet, 2009). In Arabidopsis, 10 AGO proteins have been characterized and of them, AGOs 1, 2, 7, and 10 have been shown to function in miRNA-related post transcriptional gene silencing (Borges and Martienssen, 2015; Brodersen et al., 2008). In plants, the mature miRNA guides the RISC to a target mRNA and post transcriptional gene silencing occurs by either direct cleavage and subsequent degradation of the target mRNA molecule or by translational inhibition (Baumberger and Baulcombe, 2005; Chen, 2004; Llave et al., 2002). A certain sub-class of miRNA has also been discovered which are highly similar to miRNA in biogenesis and function although they lack some of the conserved features of the pre-miRNA hairpin-loop structure and they produce a slightly longer mature small RNA molecule, up to 24 nt in length. They are formed from RNA molecules which form a longer hairpin-loop structure, also due to partial self-complementarity. Evidence suggests that in this form of miRNA, multiple DCL enzymes (in Arabidopsis, all four DCL) are involved in cleavage of the primary RNA whereas canonical miRNA rely on DCL1 (Vazquez et al., 2008).

MicroRNAs have a strong role in regulating developmental functions, with many targeting transcription factors in particular. For example in maize, the gene Corngrass1 has a dominant Cg1-R allele which causes plants to initiate more leaves than wildtype and those leaves retain a slender juvenile morphology even upon transition to reproductive phase (Poethig, 1988; Whaley and Leech, 1950). The additional leaf initiation results from tiller formation at each axial meristem. Other developmental functions are also affected as root growth is initiated from axial
meristems and reproductive tissues are altered (Chuck et al., 2007). \textit{Cg1} was cloned and found to encode two \textit{miR156} genes that are overexpressed in meristem tissues containing the \textit{Cg1-R} allele (Chuck et al., 2007). Furthermore, it was demonstrated that overexpression of \textit{miR156} resulted in reduced levels of \textit{miR172} which is known to target transcripts of genes involved in developmental and phase change processes. Taken together, these results demonstrate both the dominant action of miRNAs and epistatic interactions that they participate in, both between other miRNAs and gene transcripts.

The second category, natsiRNA shares similarities with both miRNA and other forms of small RNA. In this form of siRNA biogenesis, complementary mRNA are produced by Pol II and/or Pol IV from opposing transcription of both DNA strands at a genic locus. The resulting self-compatible mRNA transcripts form a double stranded RNA molecule which is then cleaved into 21 or 22-nt double stranded mature siRNA by DCL1 and loaded into a RISC in a fashion similar to that described above (Mallory and Vaucheret, 2006; Sunkar et al., 2007). In this way natsiRNA share some similarity with miRNA in that they are both processed by DCL1 and they both function in post-transcriptional gene silencing by mRNA degradation or translational inhibition. In plants, only \textit{cis} natsiRNA have been discovered (i.e. double stranded RNA precursors arising from transcription of opposing DNA strands, resulting in self-complementarity). No examples of natsiRNA have been shown to arise in \textit{trans} from RNA transcribed at separate genomic loci and indeed, few examples of natsiRNA have been discovered overall to date.

Of the limited examples of natsiRNA, they generally have been demonstrated to be involved with responses to external stimuli such as abiotic stress or viral infection, rather than being involved with developmental processes such as miRNA (Borsani et al., 2005; Katiyar-
One of the earlier and best characterized natsiRNA examples is involved in response to salinity in Arabidopsis (Borsani et al., 2005). In this example, two complementary transcripts are produced from convergent transcription at a single locus with the first transcript coming from the pyrroline-5-carboxylate dehydrogenase (P5CDH) gene and the other from a gene known as SRO6 on the opposite strand. The transcripts overlap by 760 nt and produce natsiRNA which cleave and downregulate P5CDH expression. SRO6 is salt-stress induced and therefore so is the natsiRNA. Since P5CDH functions in proline metabolism, downregulation allows accumulation of proline which in turn confers salinity tolerance.

The third category of small RNA is that of secondary siRNA (which include sub-categories of phased-siRNA or phasiRNA, trans-acting siRNA or tasiRNA, and epigenetically activated siRNA or easiRNA) which result as secondary products of miRNA cleavage of a target primary RNA molecule (Borges and Martienssen, 2015; Fei et al., 2013). Similar to the previous small RNA examples, the primary RNA molecule is transcribed by Pol II and may either be a coding mRNA or a non-coding transcript. A defining characteristic of secondary siRNA is the mode of mature small RNA creation as secondary siRNAs require miRNA cleavage of the primary RNA. Production of the secondary small RNA precursors requires targeting and cleavage of the primary transcript by either two 21-nt or one 22-nt miRNA (Axtell et al., 2006; Yoshikawa et al., 2005). Unlike previously described small RNA classes, double stranded RNA formation requires the action of an RNA-dependent RNA polymerase operating on products of the primary transcript’s cleavage by miRNAs (Peragine et al., 2004). A. thaliana RDR6 accomplishes this function and has multiple putative orthologues in maize, presumably arising from genome duplication events and possibly allowing diversification in the pathway via subfunctionalization (Li et al., 2010). Secondary cleavage of the double-stranded RNA products
usually occurs in a phased manner with targeted cleavage by DCL2 and DCL4 producing multiple 21-nt or 22-nt mature secondary siRNA (Allen et al., 2005). These double stranded RNA are often loaded into an AGO-containing RISC as in the case of miRNAs and can act in *trans* (i.e. trans-acting siRNA or “tasiRNA”) to downregulate target mRNA by target degradation or by translational inhibition (Yoshikawa et al., 2005). Trans-acting siRNAs, although distinct from miRNAs, share a number of key elements. Generally they diverge in the mode of biogenesis of the double stranded mature small RNA molecule but have many commonalities in their mode of action and subsequent processing and targeting of mRNA molecules in *trans*. In anthers and pollen of rice and other grasses, evidence also exists of a subfunctionalized alternate pathway for phased tasiRNA production which involves a unique DCL3 variant that is distinct from the previously characterized tasiRNA pathway (Song et al., 2012). This pathway utilizes cleavage of primary RNA transcripts by miR2118 for production of 21-nt siRNAs and by miR2275 for the production of 24-nt siRNA. The primary RNA products are then made double stranded and finally cleaved in a phased manner by DCL4 or DCL5, respectively, to produce mature tasiRNA.

Finally, evidence suggests that easiRNAs are processed in a manner similar to other 21 or 22-nt secondary siRNAs and also involves RDR6, DCL4, and AGO1 (McCue et al., 2012; Nuthikattu et al., 2013). Recent studies have shown that preference for production of easiRNA versus other secondary siRNA may in part be the result of monouridylation of 22-nt miRNAs such as miR170 and miR171a by UTP:RNA URIDYLYLTRANSFERASE (URT1) which triggers the production of easiRNAs (Tu et al., 2015; Zhai et al., 2013). The defining characteristic that sets easiRNAs apart from other secondary siRNAs is that they arise from
transcriptionally active retrotransposons and function in the RdDM pathway in a manner similar to but distinct from hetsiRNA (Creasey et al., 2014).

The final main group of small RNA is that of heterochromatic siRNA (hetsiRNA) which are primarily defined by a biogenesis pathway that is similar to secondary siRNA yet involves distinct machinery and does not require mRNA cleavage. hetsiRNA precursors are transcribed from repetitive, non-coding, and intergenic loci of the genome and are 24-nt long. They are the only known siRNA precursors to be transcribed by RNA Polymerase IV (Pol IV), as opposed to secondary siRNA biogenesis which requires Pol II (Mosher et al., 2008). Primary RNA precursors are made double stranded by RDR2 instead of RDR6 as in the case of other secondary siRNAs, and double stranded RNA cleavage is accomplished by DCL3 as opposed to DCL2 and DCL4. In maize, hetsiRNA accumulation is dependent on the RDR2 ortholog mop1 (Nobuta et al., 2008). This results in 24-nt mature hetsiRNA instead of 21 or 22-nt secondary siRNA such as tasiRNA and easiRNA. hetsiRNA are often the most abundant small RNA species in sequencing data. In contrast to miRNA, which number on the order of 100’s of distinct sequences and constitute only a small proportion of the small RNA transcriptome, hetsiRNA have been detected as distinct sequences numbering approximately 100,000 in Arabidopsis (Lu et al., 2005; Mosher et al., 2008).

The other main distinguishing feature of hetsiRNA that sets them apart from all other small RNA is their exclusive involvement in DNA methylation via the RNA-directed DNA methylation (RdRM) pathway, as they have not been demonstrated to play a direct role in post-transcriptional gene silencing. hetsiRNAs are associated with the deposition of epigenetic marks in the form of 5-methyl cytosine in both symmetric (CG, CHG) and asymmetric (CHH) sequence contexts (H = A, C, or T) as well as H3K9 histone methylation (Law and Jacobsen, 2010;
Matzke and Mosher, 2014). It was previously postulated that hetsiRNAs were created and remained in the nucleus since that is where both biogenesis and functional action occur. However, like miRNAs, hetsiRNAs may act in \textit{trans} and are not restricted to acting solely on the repetitive elements from which they arise nor are they restricted to the nucleus from which they arise. Indeed, various research has demonstrated that export of the mature hetsiRNA molecule into the cytoplasm occurs and that formation of the of AGO4/hetsiRNA complex is a requirement for import back into the nucleus (Ye et al., 2012). Furthermore, it is hypothesized that this export/import requirement provides a mechanism for specific cell selectivity and regulation via a nuclear localization signal. In fact, it has been demonstrated that hetsiRNA are mobile across various cell types and tissues via the phloem, providing further evidence of widespread hetsiRNA activity in many specific contexts across many cell types and tissues (Melnyk et al., 2011).

As will be discussed in the following section, hetsiRNA form part of a complex regulatory network involving interactions between transposable elements, miRNA, easiRNA, hetsiRNA, and methylation machinery in the RdDM pathway in order to maintain and establish \textit{de novo} patterns of genome-wide methylation. This has important implications for hybridization of genomes as easiRNAs, hetsiRNAs, and RdDM affect gene expression, transposable element activity, and silencing or activation of foreign and divergent DNA coming into contact as the result of hybridization.

\textbf{2.3: RNA-directed DNA Methylation}

RNA-directed DNA methylation arises from a complex network of epigenetic machinery involving various small RNAs, DNA sequences, specific RNA polymerases, and DNA
methylation proteins. RdDM was first discovered in tobacco plants that were transformed with non-coding potato spindle tuber viroid RNA (Wassenegger et al., 1994). Southern blot analysis implicated a mechanism of targeted de novo methylation wherein only DNA loci corresponding to the integrated viroid sequence were methylated and methylation status of flanking native genomic DNA remained unaffected.

Subsequent work implicated the action of hetsiRNAs wherein targeting and recruitment of methylation machinery requires another RNA polymerase, Pol V, which produces long non-coding transcripts that serve as scaffold RNA at the hetsiRNA target site (Wierzbicki et al., 2008). The canonical model of RdDM involves first the production of 24-nt hetsiRNAs via the Pol IV / RDR2 / DCL3 / AGO4 pathway and secondly, the recruitment of Pol V at target sites which transcribes long non-coding scaffold RNA. The hetsiRNA/AGO4 complex then binds to the Pol V produced scaffold RNA transcript and recruits DNA methylation and chromatin modification machinery which includes, among others, DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) (Matzke et al., 2015; Wierzbicki et al., 2012). In this way, the hetsiRNA population is a function of Pol IV activity yet the population of target sites is a function of both sequence specificity between hetsiRNA and scaffold RNA as well as any factors influencing Pol V transcription. It remains unclear exactly what the base pairing requirements are for hetsiRNA binding and how Pol IV and Pol V preferentially target sites although chromatin immunoprecipitation sequencing (ChIP-seq) experiments have shown evidence that Pol V preferentially occupies promoter regions in close proximity to transposons in general and especially to evolutionarily young transposons (Zhong et al., 2012). Furthermore it has been demonstrated that Pol IV can occupy alternate gene promoters that overlap with the
main Pol II promoter that is used for gene transcription, implicating RdRM as a mechanism for transcriptional gene silencing (Sasaki et al., 2014).

RdDM has been demonstrated to act in _de novo_ methylation establishment. In fact, DNA methylation at target loci of RdDM was shown to operate on all cytosine residues (CG, CHG, and CHH contexts) as methylation status of all cytosine residues were altered in the aforementioned tobacco plants that were transformed with potato spindle tuber viroid RNA (Pélissier et al., 1999). RdDM and siRNA pathways have been demonstrated to have the capacity to generate _de novo_ initiation of DNA methylation in both symmetrical and asymmetrical contexts and are associated with H3K9 modification (Chan et al., 2004; Zilberman et al., 2004).

In the case of symmetric CG methylation, cytosines on both strands of DNA are methylated allowing the one hemimethylated strand to serve as a template for methylation of the newly synthesized and unmethylated strand during DNA replication. Once established, maintenance of these symmetric epigenetic marks does not require the RdDM pathway as evidenced by studies using transgenic 35S:GFP containing tobacco that was transfected with RNA viruses with sequences targeting either the 35S promoter or the GFP coding region (Jones et al., 2001). Here the authors demonstrated that transfection with targeted RNA induced methylation at both the promoter and coding regions and that methylation status at the promoter was heritably maintained in subsequent generations while asymmetrical methylation of the coding region was lost. Maintenance of symmetric epigenetic marks required _Met1_ methyltransferase activity but the RNA trigger was only required for _de novo_ methylation and not for maintenance of methylation in the symmetric context. It was thus observed that symmetric CG methylation is actively maintained by non-RdRM processes which do not involve small RNA but rather involve
chromatin remodeling, histone deacetylation and methylation by MET1 methyltransferase (Chan et al., 2005; Jones et al., 2001).

In contrast to CG methylation, methylation in the asymmetric CHH context results in only one strand of DNA being methylated. Thus when replicated, one template strand retains a hemimethylated cytosine and may act as a trigger for maintenance methylation while the other strand does not possess a methylated cytosine to use as a template. Yet clearly, there exists an active mechanism by which asymmetric methylation is retained during DNA replication. This mechanism was discovered in independent studies to require the RDR2 / DCL3 / AGO4 pathway and that the hetsiRNAs was able to both generate de novo methylation as well as maintain methylation status in asymmetric contexts (Chan et al., 2004; Zilberman et al., 2003). All de novo methylation, whether symmetric or asymmetric, requires the activity of DRM2 for establishment where maintenance of asymmetric methylation also notably requires DRM2. Thus maintenance of asymmetric methylation may be thought of as something of a continual re-establishment rather than a separate set of processes for maintenance. Notably, an alternate CHH methylation pathway exists via DEFICIENT IN DNA METHYLATION 1 (DDMI) and CHROMOMETHYLASE 2 (CMT2) which may provide some functional duplication to the RdDM pathway although it remains unclear as to the extent of functional overlap between these two pathways (Zemach et al., 2013).

Further to the canonical RdDM model, evidence is emerging for an increasingly complex model of methylation establishment and maintenance which involves interactions between hetsiRNA, easiRNA and by association to the latter, miRNA pathways. A study by Creasey et al., (2014), used small RNA sequencing in DECREASED DNA METHYLATION 1 (ddm1) mutants to determine that easiRNAs arose from specific retroelements and that specificity was
conferred by miRNA targeting of active retroelement transcripts in easiRNA biogenesis. They also found that there was an antagonistic relationship between easiRNA and hetsiRNA production, as increased easiRNA production in the *ddm1* mutant also resulted in decreased hetsiRNA production by approximately 50 percent. Conversely, when easiRNA biogenesis was disrupted in *ddm1 rdr6* and *ddm1 dcl1* double knockouts, many transposable elements showed increased hetsiRNA production, even at loci where there was no hetsiRNA production detected in the single *ddm1* mutant. Furthermore, it was demonstrated by methyl-sequencing that decreases in easiRNA production in the *ddm1 rdr6* mutant result in increased CHH context methylation (as effected by increased hetsiRNA production). This suggests an interplay between miRNA, easiRNA, and hetsiRNA that has interesting implications for developmental and tissue specific expression of miRNA as a mechanism for fine control of transposable element regulation and genome-wide methylation. Of particular note is the loss of RdDM and methylation in pollen grains, resulting in activation of those transposable elements that are normally regulated by hetsiRNAs and the RdDM pathway but not those targeted by easiRNAs. A mechanism such as this would provide control over which transposable elements are activated in male gamete formation and therefore which are re-methylated and inactivated upon fertilization by the maternal hetsiRNA transcriptome and RdDM machinery. This provides a mechanism by which hybridization would be controlled by the sequence divergence and targeting specificity between activated transposable elements in the male gamete and the maternal hetsiRNA transcriptome’s ability to re-silence them (Calarco et al., 2012; Mosher et al., 2008).

Indeed, it remains unclear as to the extent that Pol II and Pol IV/V driven pathways interact to regulate DNA methylation. Arising from genome duplication and subsequent sub-functionalization and specialization, Pol IV and Pol V are plant-specific and do not exist in
animals (Luo and Hall, 2007; Ream et al., 2009). In yeast for example, Pol II has activity encompassing both Pol IV and Pol V as it can transcribe both siRNA primary precursor RNA and long non-coding scaffold RNA that is required for AGO4/hetsiRNA targeting (Keller and Bühler, 2013). Likewise in plants, Pol II has been demonstrated to have activity resulting in transcription of long non-coding scaffold RNA that is able to recruit AGO4/siRNA complexes and that Pol II can also recruit Pol IV and Pol V at heterochromatic loci to increase hetsiRNA biogenesis (You et al., 2013; Zheng et al., 2009). This fits with an emerging model of a highly complex network involving multiple RNA polymerases and interactions among main miRNA, easiRNA, and hetsiRNA pathways. This blurs the lines between functional small RNA classifications and more importantly may provide the basis of RdDM-mediated epigenetic phase shifts which act in a highly context specific manner and have important implications for metabolic and developmental processes, hybridization, and evolution in plants.

2.4: RdDM and Gene Expression

In contrast to the relatively small and low complexity genome of Arabidopsis, the maize genome is large, highly complex and contains large intergenic regions with high levels of transposable elements and repetitive sequences that are often in the vicinity of genic regions (Schnable et al., 2009). Indeed, the maize B73 genome contains approximately 32,000 predicted genes and over 1 million transposable elements (Schnable et al., 2009). Of the ~ 2.3 gigabases in the maize genome, evidence shows that ~1/3 of cytosine residues are methylated (Montero et al., 1992). CG and CHG methylation contexts are the most prevalent, with CHH methylation patterns associated with RdDM and 24-nt hetsiRNAs accounting for only approximately 5% of genome methylation (Regulski et al., 2013). The Arabidopsis DNA Methyltransferase CMT2 is
able to mediate maintenance of methylation status at transposons in an RdDM independent pathway although there is no known maize CMT2 ortholog and it appears that maize relies primarily on RdDM to maintain methylation status in the asymmetrical CHH context (Zemach et al., 2013). Whereas symmetric CG and CHG methylation appear to be a more broadly distributed feature of repetitive DNA and transposable elements across the genome, heterochromatin formation at CHH context sequences and related to RdDM in maize is particularly enriched in long terminal repeats of retrotransposons, DNA transposons, and in particular it appears to be associated with heterochromatic regions in close proximity to genes, as will be discussed.

A recent study by Gent et al. (2013) used bisulphite sequencing to characterize methylation patterns associated with CG, CHG, and CHH methylation contexts. They found that CG and CHG methylation were associated with transposable elements across a broad genomic distribution, dense heterochromatin formation, and lack of transcription. On the contrary, CHH methylation and RdDM were strongly enriched in 1 kb flanking regions of genes, which the authors dubbed “CHH islands”. The authors had also previously reported 24-nt hetsiRNA enrichment in these flanking regions and that there was a strong positive correlation between gene proximity and siRNA expression (Gent et al., 2011). When gene expression levels were examined, the authors also found a strong negative correlation between CG and CHG methylation and gene expression and a positive correlation between CHH methylation and gene expression in these flanking regions (Gent et al., 2013). This further implies the involvement of RdDM as a regulator of gene expression and chromatin state. In fact, certain transposable elements, namely Class II DNA transposons of the Tc1/Mariner and PIF/Harbinger families (which include Miniature Inverted Repeat Transposable Elements, or MITEs) were among the most overrepresented in the 1 kb flanking regions but overall results indicated that transposon
classification was not a requirement for enriched CHH methylation, but rather gene proximity was the driving force (Gent et al., 2013).

Further study by Li et al. (2015) corroborates these findings and also provides further insight into the role of CHH islands and RdDM in regulating gene expression and transposable elements. Here the authors proposed two hypotheses for the function of CHH islands as borders between genic regions and nearby transposons. The first hypothesis was that methylation of CHH islands blocks heterochromatin formation from spreading into protein coding genic regions since CHH methylation is more closely associated with euchromatic than heterochromatic states (Zheng et al., 2013). This hypothesis would suggest then that mop1 mutants that are impaired in 24-nt hetsiRNA production, and thus impaired in RdDM, would demonstrate downregulated gene expression at loci associated with CHH-islands, yet this was not found to be true. The alternative hypothesis was that CHH islands function to prevent the spread of euchromatin outward from the genic region and into transposable elements which would then become activated, possibly disrupting nearby genes. In this context, CG and CHG methylation were notably reduced in the region 5' to the CHH island, supporting this alternative hypothesis.

Further analysis of RNA-seq data also demonstrated that a subset of transposons had upregulated expression in mop1-1 backgrounds and that these genes are significantly closer to CHH islands than transposons that remained silent.

Additional support for this hypothesis comes from Gent et al. (2014) who used bisulphite sequencing and small RNA sequencing to establish that loci undergoing RdDM as evidenced by hetsiRNA accumulation and increased CHH methylation represented approximately 2% of the genome and that these loci were much closer to genes than non-RdDM loci. Furthermore, CHH methylation at loci associated with RdDM decreased by 4-fold in the mop1 mutant which is
impaired in hetsiRNA production, while other non-RdDM, uncharacterized, and genic loci only exhibited a decrease of 1.4-fold. They further utilized micrococcal nuclease sequencing (which tests for DNA accessibility as a function of nuclease digestion) and chromatin immunoprecipitation sequencing to determine that loci associated with RdDM were substantially more accessible to transcriptional machinery than non-RdDM loci and that this accessible chromatin was associated with H3K9 methylation. These results suggest that not all heterochromatin is “created equal” and that intergenic regions associated with dense chromatin formation and transcriptional inactivation are different than intermediate and more gene-like heterochromatic regions flanking genes.

Further evidence for epigenetic regulation of heterochromatin formation and gene expression by hetsiRNAs, RdDM, and CHH methylation comes from work by Wei et al. (2014) in rice (Oryza sativa). Rice is similar to maize in that it has a relatively large and complex genome that is rich in transposable elements and other repetitive sequences that constitute approximately 35% of the genome (International Rice Genome Sequencing Project, 2005). Here the authors examined Miniature Inverted Repeat Transposable Elements (MITEs), their relationship to hetsiRNA producing loci, and their effects on gene expression. MITEs are highly-conserved, short (< 600 bp), non-autonomous DNA transposons and represent the most abundant transposable elements in many plant genomes (Han et al., 2013). They belong to the Class 2 group of transposable elements and the majority belong to the Tcl/Mariner and PIF/Harbinger superfamilies (Han et al., 2013). Similarly, Gent et al. (2013), Li et al. (2015), and Wei et al. (2014) found that hetsiRNA production was associated with MITE-rich gene flanking regions. In the OsDCL3a mutant, which is the rice homolog of Arabidopsis RDR2 and maize mop1, reduction in hetsiRNA production by RNAi caused a depression in heterochromatin formation in
MITE-associated gene flanking regions and increased expression of nearby genes. The authors detected 859 differentially expressed genes in hetsiRNA deficient lines with 79% of those genes being upregulated. Of particular note were genes related to gibberellin and brassinosteroid homeostasis which function in controlling a wide array of physiological processes and have significant implications for agricultural traits. Indeed, RNAi knockdown of OsDCL3a resulted in shortened plant stature at heading, increased flag leaf angle, and smaller panicle size with reduced panicle branching. Exogenous application of GA$_3$ rescued the dwarf phenotype of hetsiRNA deficient lines, confirming the role of hetsiRNAs and RdDM in regulation of gibberellin-controlled phenotypic traits. Upon examination of a specific gene known to have a MITE-rich 5’ region and function controlling flag leaf angle, the study found, by using bisulphite sequencing, a reduction in hetsiRNAs at flanking regions, a reduction in H3K9 methylation, and an increase in gene expression. Taken together, these studies suggest an emerging model whereby hetsiRNA production and RdDM regulate formation of intermediate heterochromatin in gene flanking regions that are rich in Class 2 transposable elements (DNA transposons). The exact functional nature of this phenomenon remains unclear however, as some have speculated that the main function of RdDM is to delineate regions of dense intergenic heterochromatin and prevent the spread of euchromatin, while others have speculated that the function is to fine tune regulation of nearby genes. Others still have observed stress-related induction of RdDM and hypothesized that rare but large evolutionary bursts of transposable element activity play a role in creation of de novo variation from which plants may evolve highly specific context-related responses (Naito et al., 2009; Springer et al., 2016). While these hypotheses are not mutually exclusive, it remains to be seen as to the extent that RdDM plays a role in each of these processes.
2.5: Heritability of Epigenetic States

Transgenerational inheritance of epigenetic changes was first observed by Barbara McClintock in her pioneering work with maize transposable elements. Here she observed what she described as “changes of phase” in Ac and Ds transposons as the elements switched between inactive (later to be discovered to be hypermethylated) and active (hypomethylated) states. She also observed that these changes in epigenetic state were heritable across generations (McClintock, 1950, 1961). As described above, recent evidence suggests that this transgenerational inheritance is regulated by a complex network involving hetsiRNA and RdDM. Establishment of methylation in all CG, CHG, and CHH contexts is mediated in part by RdDM. Once established, maintenance of symmetric CG and CHG methylation appears to be RdDM-independent and operates via a feedforward loop involving DNA methyltransferases and chromomethylases. In contrast, CHH methylation does require the ongoing activity of hetsiRNAs and RdDM for active maintenance (Bond and Baulcombe, 2014; Heard and Martienssen, 2014).

The activity of RdDM establishment of symmetric methylation was evidenced by a study examining the FLOWERING WAGENINGEN (FWA) gene in Arabidopsis (Soppe et al., 2000). FWA is a homeodomain transcription factor that contains a pair of direct repeats in the promoter region. When these repeats are hypomethylated, FWA expression is increased and initiation of flowering is delayed. Transformation with siRNA targeting the repeat sequences in the FWA promoter was demonstrated to induce RdDM and cause hypermethylation of the promoter region. Thus, FWA expression was downregulated and plants flowered earlier than wildtype counterparts. Furthermore, epigenetic silencing was stably inherited even in the absence of the
transgene. Further work indicated that while establishment of the silenced state required siRNA mediated RdDM, disruption of the non-RdDM methylation maintenance methyltransferase MET1 returned the FWA promoter to the hypomethylated state. This suggests a role in RdDM in establishment but not maintenance of symmetric methylation at this locus (Kinoshita et al., 2007).

Strong parallels exist between these findings and the maize Vegetative to generative transition 1 (Vgt1) locus which is a major effect QTL controlling flowering time (Buckler et al., 2009; Chardon et al., 2004). Vgt1 is a conserved non-coding sequence that is located ~70 kb upstream from an Ap2-like flowering repressor called ZmRap2.7 and acts as a cis-regulatory element. The Gaspé Flint allele of Vgt1 contains a MITE insertion and confers extremely early flowering (Salvi et al., 2007). Methyl-sequencing data recently observed hypermethylation at the Vgt1 MITE insertion in the line C22 (Gaspé Vgt1 allele introgression into N28 background) as well as the surrounding region of the non-coding sequence that exists downstream of the insertion site, when compared to the N28 near isogenic line (Castelletti et al., 2014). While the authors did not directly evaluate siRNA production in these materials, mapping siRNA reads from Nobuta et al. (2008) to the lines used in this study demonstrated that siRNA were only associated with the MITE itself and not with the Vgt1 non-coding element outside of the MITE insertion site. Methylation patterns were also analyzed in the F1 progeny of C22xN28 and methylation status for each allele in the F1 was observed to retain fidelity to those observed in their parental lines. This demonstrates a lack of trans-chromosomal epigenetic silencing at this locus but this is to be expected in this particular context as the results suggest the MITE is the primary target of hypermethylation and methylation of the Vgt1 non-coding sequence is a secondary result of methylation spreading from the MITE. Since the N28 allele does not contain
a MITE insertion, it is likely that there is no primary target for RdDM to operate on. In the C22 allele however, elevated CHH methylation was observed versus N28 which does suggest a role for RdDM in establishment and maintenance of epigenetic silencing for this allele across generations.

While establishment and maintenance of cis-mediated methylation status seems to be prevalent on a genome-wide scale, it remains unclear as to what extent trans-chromosomal interactions play a role in shaping genome-wide methylation status. Conceivably a hetsiRNA and RdDM mediated system could exist wherein hetsiRNA produced from one chromosome could target homologous loci and mediate epigenetic silencing in trans. As RdDM is able to establish methylation in all contexts, symmetric methylation could become established that would then rely on the methyltransferase feedforward loop and no longer require the presence of the initial hetsiRNA trigger and RdDM for active maintenance. This type of system has been proposed to, at least in part, explain the classical epigenetic system of paramutation.

Paramutation was first detected at the maize r1 locus where the R-r allele produces kernels with red pigment (Brink, 1958). When in a heterozygous state with the R-stippled allele, pigmentation is reduced by a trans-chromosomal interaction between the R-stippled and R-r alleles, producing an R-r’ allele. This trans-chromosomal interaction is heritable and confers reduced pigment even in the absence of R-stippled (Brink et al., 1960). Thus the R-r allele is termed paramutable and the R-stippled allele is termed paramutagenic. Additionally, the paramutant R-r’ allele then becomes paramutagenic itself, retaining the ability to induce paramutation in R-r’/R-r genotypes but R-r’ does revert to R-r when in a hemizygous state (Brown and Brink, 1960).
Various other examples of well-characterized paramutagenic loci exist in maize. Many of these loci defy a unifying theory as to an exact mechanism as there are often context specific irregularities associated with them, however most examples studied to date seem to implicate trans-chromosomal effects that are governed by RdDM and hetsiRNAs in some way. Various studies have examined the role of hetsiRNA biogenesis genes in the establishment and maintenance of paramutation among classical paramutation gene examples. While mutations at some hetsiRNA biogenesis genes (including rdr2 and the maize ortholog mop1) consistently affect paramutation, loss-of-function mutation at other genes often produces partial phenotypes or has no discernable effect on paramutation (Hollick, 2012). Analogous to many of the examples previously described here, paramutation has consistently been demonstrated to affect transcription and is associated with RdDM-mediated methylation status changes at enhancer sequences flanking coding genes (Hollick et al., 2000; Sidorenko and Chandler, 2008).

Further evidence for heritability of epigenetic states comes from (Li et al., 2014) where the authors found differential methylation between B73 and Mo17 parental lines at 962 loci which they termed Differentially Methylated Regions (DMRs). They then classified DMRs into categories based on the pattern of methylation inheritance in a NIL population that had been developed by backcrossing B73/Mo17 hybrids to either recurrent parent. Statistical comparison by t-test comparing each DMR between B73-backcrossed and Mo17-backcrossed NILs facilitated classification of methylation inheritance type. If methylation of a DMR was significantly different between B73-backcrossed and Mo17-backcrossed NILs then it was further classified by predicting the genotype of NILs with a genomic introgression within the DMR based on observed methylation. If the probability of the actual DMR genotype equaling the predicted genotype was low it was termed cis or trans methylation inheritance. DMRs matching...
the introgressed genotype were classified as *cis* and those matching the background genotype were classified as *trans*. Of these categories, a pattern of *cis* inheritance consistent with maintenance of symmetric methylation conferred by the local genotype was found to represent the vast majority of DMRs (similar to the FWA and *Vgt1* examples noted above). Few loci represented paramutation-like (13 of 800 classified DMRs) or *trans*-like patterns (25 of 800 classified DMRs) although it should be noted that ~38% of DMRs fell into a “complex” category of inheritance that could not be classified as either *cis* or *trans*-like. Furthermore, the authors here only examined loci that were introgressions in a population of NILs. As such the results are more applicable in the context of intergenerational heritability wherein chromosomal segments from one species, variety, or inbred line would be physically integrated into the genome of another via homologous recombination. Another context that was not directly examined would be that of heterosis in an F1 hybrid crop, where homologous chromosomes do not recombine but instead may exhibit *trans*-chromosomal epigenetic effects independent of recombination events.

In contrast to the previous study evidence for a more fragile and dynamic system of epigenetic heritability comes from the study of epigenetic mosaic recombinant inbred lines (epiRILs) in *Arabidopsis* (Reinders et al., 2009). Here the authors used the elegant approach of crossing a *met1* null mutant to its isogenic wildtype ecotype and subsequently selfing down a population of RILs. Since a *met1* knockout is sufficient to virtually eliminate all CG methylation and results in partial loss of CHH methylation, the resulting F2 progeny were genetically isogenic yet epigenetically mosaic and were segregating for epialleles much in the way a traditional RIL population would segregate genetically. In order to immediately restore *MET1* activity, only homozygous *MET1* individuals were selected in the F2 generation before inbreeding to produce epiRILs. Thus, the authors were able to observe evidence of unstable
inheritance of epialleles by various methods of methylation status quantification. Indeed, an approximately 30% deviation from midparent value toward the wildtype methylation status was observed. This suggests an active mechanism of trans-chromosomal methylation at genetically homologous loci. Interestingly the authors also observed segregation distortion and the preservation of heterozygous epigenetic states through the inbreeding process which led them to speculate that there is some mechanism whereby meta-stable epialleles continue to interact and persist in a heterozygous state. Taken together these results demonstrate the high degree of context dependency that exists in epigenetic characterization. That there is no obvious unifying mechanism of intergenerational inheritance or epigenome stabilization is not surprising given what is known about the highly complex modes of small RNA biogenesis, function, and their interactions with genome features via RdDM.

2.6: Epigenetics, Hybridization, and Heterosis

Traditional analysis of heterosis has largely attributed effects to dominance and over-dominance, however neither of these factors have been able to fully account for heterosis in general (Birchler et al., 2010). While dominance and over-dominance clearly play a major role, epistatic interactions are also thought to play a role and recently increasing evidence suggests that epigenetic effects mediated by small RNA and RdDM play an important role in establishing heritable epigenetic changes in hybrids and intraspecific crosses. Indeed, the magnitude of heterosis is largely considered to increase with increasing genetic distance between parents in an intraspecific cross yet this is not always the case. Arabidopsis hybrids from a cross between C24 and Ler ecotypes exhibit a 250% seed yield increase despite having highly similar genomes (96%) (Meyer et al., 2004; Schneeberger et al., 2011). This suggests the action of dominance or
over-dominance at few large effect heterotic QTL at the divergent 4% of the genome and/or epigenetic effects. Indeed, evidence from diallel analysis of Arabidopsis ecotypes suggests that the epistatic interaction between alleles of FLOWERING LOCUS C (FLC) and FRIGIDA (FRI) can account for most of the heterotic effect in the aforementioned C24 x Ler example (Moore and Lukens, 2011). Some additional evidence for large effect heterotic QTL does exist, such as maize CNR1 and CNR2 where downregulation results in heterosis-like effects in hybrids as a result of increased cell number but not cell size (Guo et al., 2010), however it is mostly inconsistent with observations that heterosis for yield is a complex trait, governed by many loci (Birchler et al., 2010; Groszmann et al., 2013; Schnable and Springer, 2013). Evidence supporting epigenetic factors as heterotic mechanisms is emerging from the analysis of trans-chromosomal methylation when two parental genomes and epigenomes come into contact in the nucleus of an F1 hybrid.

Greaves et al. (2012) examined trans-chromosomal methylation patterns and their relationship to siRNA production and gene expression in the aforementioned Arabidopsis C24 x Ler F1 hybrids. Here it was observed that non-additive epigenetic changes tended to occur more frequently at loci where methylation status differed between parental epialleles. At those loci, symmetric CG methylation tended to increase while asymmetric CHH methylation tended to decrease. Thus they proposed a model in which transchromosomal methylation occurs in the CG context and is maintained by RdDM independent pathways whereas loss of CHH methylation is associated with transchromosomal de-methylation as a result of a reduction in RdDM. These findings are supported by evidence from a genome-wide methylome map of maize produced by (Regulski et al., 2013). Here the authors examined bisulphite sequence data from the Intermated B73 x Mo17 (IBM) population and detected 9635 epipolymorphic regions that differed
significantly in their methylation status among RILs. Of these, 1772 were found to be non-additive as tracking SNPs associated with the region revealed epigenetic switches where the epigenotype and methylation status were more similar to the other parent than the parental genetic haplotype contributing the SNP. Overall, approximately 5-7% of highly differentially methylated regions were associated with a gain in methylation while approximately 2% lost methylation. Interestingly the authors also found differential methylation to have a significant effect on alternative mRNA splicing, adding another layer of complexity to the interactions between methylation and gene expression.

Evidence of trans-chromosomal methylation status changes correlates with similar studies in maize and Arabidopsis examining hetsiRNA production in hybrids where 24-nt siRNA were observed to be globally downregulated in hybrids as compared to expected mid-parent values (Barber et al., 2012a; Groszmann et al., 2011a). Like methylation patterns, hetsiRNA downregulation in the hybrid increased with increasing differences in parental hetsiRNA expression. In all of these studies, non-additive methylation and siRNA production were associated with approximately 1 kb intervals flanking genic regions (Barber et al., 2012a; Greaves et al., 2012; Groszmann et al., 2011a; Regulski et al., 2013).

Taken together, these results suggest a model wherein the majority of heterosis is controlled by dominant and over-dominant effects on a global scale yet epigenetic effects may play a smaller yet significant role in regulating gene expression and controlling heterotic effects at a subset of loci. At these loci, hetsiRNA would then mediate transchromosomal and intergenomic establishment of epigenetic methylation patterns via the RdDM pathway. In F1 hybrids, hetsiRNA are mainly expressed in a non-additive manner which may contribute to non-additive establishment of the epigenome. This has interesting implications for hybrid crop
development both from an inbred development and hybrid development standpoint. In the case of inbred development, two distinct genomes come into contact in the F1 and are recombined many times through the inbreeding process resulting in progeny that are a genetic mosaic of the two parental genomes. If epigenetic switches occur in the F1 as in the case of Regulski et al. (2013) these may persist in the epigenome of the progeny despite the loss of the parental genome and even the small RNA transcriptome because once established, the epigenome could be maintained by RdDM independent pathways. If these epigenetic switches have an effect on gene expression and are stably inherited, yet not be associated with the parental genotype from which they originated, it would confound breeding efforts, especially in the context of techniques such as MAS or GWAS where associations between genotype and phenotype drive estimation and prediction of phenotypic effects at particular loci. Such was the case at some loci in tomato where establishment of non-additive epigenetic marks in crosses between domesticated (Solanum lycopersicum) and wild (Solanum pennellii) lines resulted in heritable establishment of transgressive gene expression, due to epigenetic rather than genetic changes (Shivaprasad et al., 2012). The authors here also proposed a model in which hetsiRNA mediate epigenetic methylation status changes that persist in the progeny for several generations.
Chapter 3: hetsiRNA are Downregulated and Expressed non-Additively in Hybrid Maize

3.1: Abstract

Heterosis is controlled by the cumulative effects of many genes and regulatory elements, each which may exhibit differing modes of action. Recent research in Arabidopsis thaliana has demonstrated that 24-nt heterochromatic small interfering RNAs (hetsiRNAs) show non-additive expression upon hybridization, with the majority of hetsiRNAs being downregulated. This Chapter examines such trends in non-additive hetsiRNA expression in commercial maize germplasm. Small RNA were isolated and deep sequenced from leaf tissue samples of two inbred lines and their F1 hybrid. In order to examine the effects of reducing hetsiRNAs in hybridization, each genotype was also sampled in each of two mop1 allelic states. Results demonstrate that mean hetsiRNA expression is significantly reduced in the hybrid relative to its parents and that hetsiRNA expression in mop1-1 is almost entirely absent. The F1 hybrid also demonstrated a marked reduction in hetsiRNA diversity, with 41% of hetsiRNA clusters being expressed in one or both parents but not the hybrid. Similarly, classifying hetsiRNA expression in the hybrid relative to the expected midparent value demonstrated that the vast majority of hetsiRNA clusters are expressed below the midparent level, with many below the low parent. Since hetsiRNAs are known to be involved in the regulation of asymmetric DNA methylation, this general downregulation may be related to mechanisms that evolved to promote genome integrity when two independent genomes come into contact. These results have further implications for the development of hybrid crop species and the notion of an epigenome that may affect phenotype by participating in trans-genomic interactions upon hybridization.
3.2: Introduction

Heterosis is the phenomenon by which out-crossed progeny exhibit superior and transgressive phenotypes as compared to their more inbred parental lines. It is an important and widely exploited mechanism of enhancement that is utilized in agriculturally significant crop species including fresh vegetable species, sunflower, rice, canola, and especially maize. Although heterosis was first defined over a century ago (Bruce, 1910; Shull, 1908), the underlying molecular basis of heterosis still remains elusive today. Generally, theories describing heterosis refer to heterotic effects as the cumulative action of dominant and/or over-dominant genes and alleles (Birchler et al., 2010; Schnable and Springer, 2013). While it is likely that dominance and over-dominance play a major and significant role in the quantitative manifestation of heterotic effects, some challenges to these theories still persist. For example, under a completely dominant model, heterotic phenotypes in an F1 should be recoverable over time in inbred progeny as favourable dominant alleles would accumulate through selection. While linkage disequilibrium can hinder this process if desirable alleles for different traits are maintained in repulsion phase, in a dominant model, one should nonetheless expect to observe a reduction in the magnitude of heterosis over time. This would be especially true in species that have undergone intensive selection pressure such as maize, however this has not been observed as the magnitude of heterosis has remained relatively constant through decades of North American corn breeding (Duvick, 1984). Overdominance attempts to reconcile this limitation of the dominance theory by supposing that hybrid vigour is resultant from an accumulation of heterozygous alleles. These two theories are not mutually exclusive and likely groups of genes are involved with one or the other on a case-by-case basis. While overdominance theoretically resolves some limitations of the dominance theory, in practice relatively few published studies
have demonstrated the action of overdominant loci. Of particular note are certain alleles involved in paramutation, such as purple plant (pl), that exhibit an overdominant phenotype as the result of \textit{trans} interactions in a heterozygous state (Hollick, 2012; Hollick and Chandler, 1998). While it remains likely simple dominant or overdominant gene action across many loci contributes the majority of heterotic effects to the whole, growing evidence suggests that small RNAs and \textit{trans} chromosomal methylation may play a significant role in controlling the portion of heterosis that cannot be predicted by more classic quantitative genetic theories.

Small RNAs exist in many forms as classified by their mode of biogenesis and action of the mature molecule (Axtell, 2013). While these modes of biogenesis and function are somewhat diverse, commonalities include formation of 20-24 nt double stranded RNA (either by self-complementarity or by an RNA-Dependent RNA Polymerase), cleavage by a Dicer-like enzyme, and incorporation of the mature small RNA into an AGO-containing RISC. The RISC then targets specific genomic loci or mRNA transcripts for cleavage, translational inhibition, or DNA methylation via the RdDM pathway (Borges and Martienssen, 2015). This thesis primarily focuses on 24-nt hetsiRNA that are part of a distinct small RNA pathway where they are transcribed by RNA Polymerase IV, made double stranded by RDR2 (\textit{mop1} in maize), and cleaved by DCL3 (Mosher et al., 2008). The key distinguishing feature of mature hetsiRNAs is that they do not function to directly cleave or inhibit translation of mRNA transcripts. They do however, target machinery of the RdDM pathway to elicit methylation status changes (Matzke and Mosher, 2014; Matzke et al., 2015). RdDM is able to establish DNA methylation in all sequence contexts (CG, CHG, and CHH) but once established is only required for maintenance of CHH methylation (Pélissier et al., 1999). As such, these changes in methylation status exhibit
a heritability that is often independent of the loci responsible for inducing DNA methylation in the first place (Reinders et al., 2009).

Recent evidence has implicated hetsiRNAs in trans chromosomal methylation status changes that may exist and function on a larger scale than previously thought (Regulski et al., 2013). In this case, hetsiRNAs and CHH methylation were associated with paramutation-like trans chromosomal methylation switches that may function at loci contributing to broad scale heterotic effects. This is consistent with evidence from the maize mutant mediator of paramutation 1 (mop1) which demonstrates that hetsiRNAs are required for paramutation at various loci (Hollick and Chandler, 1998; Sidorenko and Chandler, 2008). As such, it is hypothesized that when divergent genomes of two inbred lines come into contact in the F1 nucleus, they each contribute a certain pool of 24-nt hetsiRNA producing loci. These hetsiRNA pools originating from either parent would then have broad effects on genome-wide methylation status and heterochromatin formation. Furthermore, hetsiRNAs originating from each parental genome may have regulatory effects upon hetsiRNA producing loci in the other parental genome, resulting in a hetsiRNA population in the F1 that is unique and non-additive. Initial evidence for this has been demonstrated in Arabidopsis and maize where hetsiRNAs were found to be downregulated in F1 hybrids as compared to expected midparent values (Barber et al., 2012a; Groszmann et al., 2011a).

The purpose of this study was to characterize the extent and nature of hetsiRNA regulation in maize F1 hybrids as compared to their parental inbred lines. Furthermore, the purpose was to examine the effect of the mop1-1 null mutation on these patterns of hetsiRNA accumulation. To accomplish this, small RNA deep sequencing was utilized on a panel of three inbred lines and their hybrids in the wildtype Mop1 state, as well in the mop1-1 mutant. As seen
in previous work in *Arabidopsis*, it was hypothesized that hetsiRNAs would be downregulated in the hybrid as compared to expected midparent values and that the degree of downregulation would correlate with expression level differences in the parental lines. What remains unclear, and what this study seeks to address, is whether or not the trend toward hetsiRNA downregulation arises from reduced expression at all hetsiRNA producing loci or whether it results from selective expression of hetsiRNAs resulting in an altered population of hetsiRNAs in the hybrid, not just in frequency of each molecule, but also in the diversity of the pool as a whole. As previous work examined global trends, it is currently unclear as to the extent to which individual hetsiRNA loci are downregulated in hybrids, relative to their expression levels in each parent.

3.3: Materials and Methods:

3.3.1: Genetic Material:

Plant material used in this study consisted of inbreds developed by Pioneer Hi-Bred. Introgressions of the mutant *mop1-1* allele were developed by backcrossing public Maize Genetics Stock accession 204D (Chandler et al., 2007) into 3 Pioneer inbred lines (termed A, B, C). Inbred A is mainly B73-derived (Stiff Stalk Synthetic heterotic group). Inbred B is derived from Mo17 and Iodent material (Non-Stiff Stalk heterotic group). Inbred C is primarily non-Iodent, non-Mo17 material (Non-Stiff Stalk heterotic group). To create wildtype *Mop1/Mop1* and mutant *mop1-1/mop1-1* introgression lines, accession 204D was backcrossed to each recurrent parent to the BC6F1 generation with selection by Mu-PCR to maintain the *mop1-1* allele. Plants were then selfed to BC6F2. Since *mop1* is involved in transposon silencing
(Alleman et al., 2006; Nobuta et al., 2008; Slotkin and Martienssen, 2007), lines were never maintained in \textit{mop1-1/mop1-1} homozygous mutant state for more than the initial generation after selfing (i.e. the sampling generation) in order to avoid confounding effects of uncontrolled methylation status alterations and widespread transposon activation in a prolonged, multi-generation \textit{mop1-1} state. Rather, BC6F2 populations were grown out and plants were genotyped to determine homozygous \textit{Mop1/Mop1} or \textit{mop1-1/mop1-1} individuals that were then use for data and tissue sample collection. To create wildtype \textit{Mop1/Mop1} and mutant \textit{mop1-1/mop1-1} hybrids, inbreds A, B, and C were used in a half-diallel scheme by crossing BC6F1 x BC6F1 plants. In this situation, \textit{mop1} segregated 1:2:1 and homozygous individuals were similarly determined by Mu-PCR for use in data and tissue sample collection. The resulting material thus consisted of Inbreds A, B, and C, each in homozygous wildtype \textit{Mop1} and mutant \textit{mop1-1} states; and Hybrids AB, AC, and BC, each in homozygous wildtype \textit{Mop1} and mutant \textit{mop1-1} states. The number of plants identified in each class is shown in Table 1. Days from seeding to silking and anthesis were recorded on the BC6F2 NILs during the growing season and comparisons between \textit{Mop1} wildtype plants and \textit{mop1-1} mutant plants were evaluated for significance by t-test at $\alpha = 0.05$. 


Table 1: Phenotypic effects of *mop1* on flowering time. Days to silking and anthesis are presented from time of seeding. * indicates comparisons between *Mop1* and *mop1-1* within background that were significant at $\alpha = 0.05$.

<table>
<thead>
<tr>
<th>Background</th>
<th>Genotype</th>
<th>n</th>
<th>Mean</th>
<th>$\Delta$</th>
<th>p-value</th>
<th>n</th>
<th>Mean</th>
<th>$\Delta$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td><em>mop1-1</em></td>
<td>10</td>
<td>87.8</td>
<td>+7.1</td>
<td>0.000*</td>
<td>6</td>
<td>84.8</td>
<td>+5.3</td>
<td>0.000*</td>
</tr>
<tr>
<td></td>
<td><em>Mop1</em></td>
<td>12</td>
<td>80.7</td>
<td></td>
<td></td>
<td>12</td>
<td>79.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A/B</td>
<td><em>mop1-1</em></td>
<td>17</td>
<td>81.2</td>
<td>+5.3</td>
<td>0.000*</td>
<td>18</td>
<td>79.1</td>
<td>+3.7</td>
<td>0.000*</td>
</tr>
<tr>
<td></td>
<td><em>Mop1</em></td>
<td>10</td>
<td>75.9</td>
<td></td>
<td></td>
<td>10</td>
<td>75.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A/C</td>
<td><em>mop1-1</em></td>
<td>20</td>
<td>79.4</td>
<td>+1.6</td>
<td>0.016*</td>
<td>20</td>
<td>80.1</td>
<td>n.s.</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td><em>Mop1</em></td>
<td>9</td>
<td>77.8</td>
<td></td>
<td></td>
<td>9</td>
<td>79.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td><em>mop1-1</em></td>
<td>17</td>
<td>87.2</td>
<td>+4.8</td>
<td>0.000*</td>
<td>17</td>
<td>85.3</td>
<td>+3.7</td>
<td>0.000*</td>
</tr>
<tr>
<td></td>
<td><em>Mop1</em></td>
<td>17</td>
<td>82.4</td>
<td></td>
<td></td>
<td>17</td>
<td>81.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B/A</td>
<td><em>mop1-1</em></td>
<td>21</td>
<td>84.5</td>
<td>+6.1</td>
<td>0.000*</td>
<td>21</td>
<td>83.0</td>
<td>+5.0</td>
<td>0.000*</td>
</tr>
<tr>
<td></td>
<td><em>Mop1</em></td>
<td>12</td>
<td>78.4</td>
<td></td>
<td></td>
<td>12</td>
<td>78.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B/C</td>
<td><em>mop1-1</em></td>
<td>12</td>
<td>84.0</td>
<td>+4.0</td>
<td>0.000*</td>
<td>12</td>
<td>84.0</td>
<td>+2.5</td>
<td>0.000*</td>
</tr>
<tr>
<td></td>
<td><em>Mop1</em></td>
<td>13</td>
<td>80.0</td>
<td></td>
<td></td>
<td>13</td>
<td>81.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td><em>mop1-1</em></td>
<td>13</td>
<td>84.8</td>
<td>n.s.</td>
<td>0.731</td>
<td>15</td>
<td>83.5</td>
<td>n.s.</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td><em>Mop1</em></td>
<td>13</td>
<td>84.6</td>
<td></td>
<td></td>
<td>10</td>
<td>84.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C/A</td>
<td><em>mop1-1</em></td>
<td>9</td>
<td>83.1</td>
<td>+3.1</td>
<td>0.034*</td>
<td>9</td>
<td>82.9</td>
<td>+2.9</td>
<td>0.011*</td>
</tr>
<tr>
<td></td>
<td><em>Mop1</em></td>
<td>10</td>
<td>80.0</td>
<td></td>
<td></td>
<td>10</td>
<td>80.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C/B</td>
<td><em>mop1-1</em></td>
<td>19</td>
<td>83.8</td>
<td>+3.8</td>
<td>0.000*</td>
<td>19</td>
<td>82.9</td>
<td>+1.6</td>
<td>0.003*</td>
</tr>
<tr>
<td></td>
<td><em>Mop1</em></td>
<td>13</td>
<td>80.0</td>
<td></td>
<td></td>
<td>13</td>
<td>81.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grand Mean</td>
<td></td>
<td>82.0</td>
<td>+4.5</td>
<td></td>
<td></td>
<td>81.4</td>
<td>+3.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1: Schematic representation of samples used for small RNA and mRNA sequencing experiments. Inbreds A, B, and C were intermated in a half-diallel to create hybrids AB, AC, and BC. Each inbred was crossed as a BC6F2 segregating for mop1 alleles, resulting in 36 total samples when replicated x3.

<table>
<thead>
<tr>
<th>Inbreds</th>
<th>mop1 Genotypes</th>
<th>Replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hybrids

AB
AC
BC

Mop1 Wildtype
mop1-1 Mutant

Pooled tissue from 3 plants
x3 Samples

x
3.3.2: Genotyping:

DNA was extracted from leaf tissue at the V6 stage of development in order to determine \textit{mop1} state. Mu-PCR was conducted as a 3-primer reaction using the following primers:

Forward: CCACCTTGATCCACGCAGACCGTGAT
Reverse: AGCGCTCAGCGGTACTGCATCTTATTAC
Mu-TIR: AGAGAAGCCAACGCCAWCGCCTCYATTTCGTC

PCR reagents were as follows: 2.5 mM MgCl$_2$, 200 µM each dNTP, 0.4 µM forward primer, 0.4 µM reverse primer, 0.4 µM Mu-TIR primer, 1.5 units \textit{Taq} polymerase, 20-50 ng DNA. Cycling conditions were: 60 s initial denaturation at 94 ºC, followed by 35 cycles of 60 s at 94 ºC, 30 s at 58 ºC, 60 s at 72 ºC, and a final extension for 3 m at 72 ºC. The wildtype allele amplifies with only the forward and reverse primers, producing a single band at 821 bp. The mutant allele amplifies with the two combinations of Forward + Mu-TIR and Reverse + Mu-TIR producing 2 bands at 446 bp and 517 bp. The \textit{mop1-1} allele also appears to be in high linkage disequilibrium with a maternally inherited floury endosperm phenotype. This is presumably caused by tight coupling phase linkage between \textit{mop1-1} and the \textit{fl1} allele of the \textit{fl1} gene (floury endosperm 1), which may aid in genotype confirmation through observation of co-segregation.

3.3.3: Tissue Sampling and RNA Extraction:

Plants for sampling and analysis were grown under field conditions at DuPont Pioneer Corn Research, Tavistock, ON, Canada in the summer of 2012. After genotyping to determine \textit{mop1} state, leaf tissue was sampled for subsequent RNA extraction and sequencing at the V8 growth stage (Iowa State University Extension, 2009). An approximately 3 cm x 3 cm square of leaf tissue was cut from the leaf margin, mid-way along the length of the youngest fully
developed leaf (fully extended with a leaf collar). Tissue was immediately preserved in 2 mL of Ambion RNAlater® solution (Thermo Fisher Scientific) and stored at 4 °C before being transferred to -80 °C for longer term storage. Tissue was sampled for 3 biological replicates per genotype. Each biological replicate consisted of a tissue pool from 3 plants. Total RNA was extracted using an Ambion miRvana™ kit (Thermo Fisher Scientific) according to the manufacturer’s protocol.

3.3.4: siRNA Sequencing and Analysis Pipeline

For a graphical summary of the small RNA analysis pipeline workflow, refer to Figure 2. Small RNA library preparation was done by The Clinical Genomics Centre in Toronto, ON, Canada using the same 36 total RNA samples as previously described. Small RNA libraries were generated size selecting RNA <=50 bases by gel electrophoresis. Library constructs were generated using Illumina TruSeq adapters and sequenced by Illumina HiSeq. Adapter trimming was done with CutAdapt v1.2.1 (Martin, 2011). A minimum target string of 10 nt was used to identify adapter sequence to be trimmed. Reads containing <10 nt were also discarded. This resulted in an overall small RNA dataset containing reads between 10-40 bases. Since 24-nt hetsiRNA were the species of interest, a custom BASH script was used to filter and retain 24-nt reads. Quality statistics for 24-nt reads were assessed using FASTX toolkit v0.0.13. In order to filter out fragments of known miRNA, rRNA, and tRNA a custom pipeline was used. Briefly, the 24-nt small RNA dataset was aligned to known miRNA, rRNA, and tRNA using PatMaN v1.2.2 (Prüfer et al., 2008). Small RNA hits were then filtered and retained for alignment and further analysis using a custom awk script and PYFASTA v0.4.5.
Alignment of 24-nt small RNA reads to the *Zea mays* B73 RefGen_v2 assembly (Release 5b) (Wei et al., 2010) was done with Bowtie v2.0.6 (Langmead and Salzberg, 2012). Since hetsiRNA are generated from multiple repetitive loci, parameters were set to allow unlimited alignments per read as opposed to arbitrarily limiting read alignments to a specific number of hits. Read alignments were constrained to only perfectly matching hits. Complete parameter options were as follows: -N 0 -L 24 -f --score-min L,0,0 --end-to-end --no-unal. Alignment output .sam files were then processed by SAMtools v0.1.18 (Li et al., 2009). Output .sam files were converted to .bam format using samtools view, sorted by alignment position using samtools sort, and converted to .bed format using the bamToBed function within the BEDTools suite v2.17.0 (Quinlan and Hall, 2010).

In this set of analyses, siRNA clustering was done in a two stage process using the mergeBed function of BEDTools (Figure 3). Aligned siRNA reads were grouped into a primary cluster if reads were located within 100 nt of each other and at least 100 nt away from another primary cluster (-d 100). The exact terminal positions of primary siRNA clusters however are specific to each sample and genomic alignment. Therefore, in order to establish common intervals and compare siRNA clusters among genotypes, a second-stage clustering process was completed by aligning primary clusters that were immediately adjacent or overlapping (-d 0). This siRNA “meta-cluster” establishes a common positional interval at each siRNA producing locus across alignments and allows direct comparison of siRNA producing loci across multiple genotypes. Total reads mapping to each siRNA meta-cluster were retained through the pipeline and summed using a custom awk script. siRNA meta-cluster expression at each locus was calculated as primary 24-nt siRNA aligned reads per million raw 24-nt reads (herein simply referred to as Reads per Million or RPM). To reduce the effects of background noise, only meta-
clusters $\geq 100$ nt in length and with expression $\geq 0.1$ RPM were retained for subsequent analysis.
Figure 2: Workflow of small RNA sequencing read processing pipeline.
Figure 3: Diagrammatic representation of stages 1 and 2 of siRNA meta-cluster analysis. In Stage 1, siRNA reads mapping to within 100 nt of each other form a primary cluster. In Stage 2, primary clusters are merged to a common frame of reference if they overlap between different genotypes.

**Clustering Stage 1: siRNA Cluster**

![Diagram of siRNA Cluster](image1)

**Clustering Stage 2: siRNA Meta-Cluster**

![Diagram of siRNA Meta-Cluster](image2)
3.4: Results

3.4.1: Phenotypic Effects of mop1-1

Contrary to a previous report (Barber et al., 2012b) that heterotic phenotypes were not associated with mop1 loss-of-function, in this study it was observed that mop1-1 mutants exhibited a statistically significant delayed flowering phenotype in all 9 inbred and hybrid backgrounds with the exception of inbred C (see Table 1). The overall effects of mop1-1 mutation across genetic backgrounds were + 4.5 days to silking and + 3.5 days to anthesis, relative to wildtype Mop1. In some backgrounds, the mutation added as much as ~7 days to flowering time (i.e. days to silk in the inbred A background). mop1-1 lines were notably shorter in total stature than Mop1 wildtypes, although a severe wind event occurred post-flowering that caused extensive root lodging and prevented collection of accurate plant height data.
3.4.2: siRNA Read Characterization

Examination of raw small RNA reads after trimming revealed a number of length classes that were enriched in the dataset (Figure 4). As expected, 21 and 22-nt classes of small RNA exhibited an expression peak, presumably corresponding to non-hetsiRNA species (miRNA, tasiRNA, and easiRNA). Interestingly, in all genotypes and mop1 states, 24-nt species were of lower expression levels than 21 and 22-nt species. Also of particular interest was a very large peak at 13-nt as well as peaks in the 16-17 nt range and 33-36 nt range that do not correspond to current known small RNA classes in maize. Inbred C as well as its hybrid combinations AC and BC exhibited no apparent difference between Mop1 and mop1-1 for mean 24-nt siRNA expression (Figure 5). This combined with the lack of significantly delayed flowering phenotype and atypical mRNA expression results (as further described in Chapter 4) suggest either an incompletely penetrant effect of mop1-1 in background C, or some form of contamination, whether from pollen or seed contamination in line development, or from the RNA sequencing process. As such, Inbred C as well as Hybrids AC and BC were removed from further analyses and examination.

A potential limitation of the preceding analysis is that abundance is calculated as reads per million (RPM) within an individual sample. As all reads were filtered to include only reads that were 24-nt in length, there is no need to normalize for varying transcript length, as when normalizing mRNA-seq datasets using Fragments Per Kilobase Million (FPKM). The limitation may occur if relatively few sequences of small RNA are extremely highly expressed in a given sample as it would cause other, less abundant sequences to be underrepresented. This may pose an issue when attempting to draw comparisons among samples that have been independently normalized. Thus most small RNAs from a sample with few highly expressed sequences could
appear to be expressed at lower levels than those from samples that may lack those few highly expressed sequences, even though their absolute abundance could be in fact equal. In order to examine the potential for this to occur, the distribution of small RNA read frequencies were examined for 24-nt small RNA sequences for Inbred A, Inbred B, and Hybrid AB. Reads per unique 24-nt sequence were tallied using the fastx_collapser function within FASTX toolkit v0.0.13. These counts were then used to calculate the overall proportion of reads for a given bin of counts (Figure 6). Overall, a slight enrichment in reads was observed in the 10-100 read count and 100-1,000 read count bins for mop1-1 relative to wildtype Mop1. This is expected as the 24-nt RNA pool diversity is decreased in mop1-1. No genotypes demonstrated a high frequency of sequences beyond the 1,000 reads bin which is evidence to support the direct comparison of normalized siRNA expression levels among different genotypes.
Table 2: Total reads sequenced per sample and resultant 24-nt reads that were filtered from total 10-40 nt small RNA sequencing libraries.

<table>
<thead>
<tr>
<th>Background</th>
<th>Genotype</th>
<th>Raw Reads</th>
<th>Raw Reads - 24nt</th>
<th>% Reads - 24nt</th>
<th>Unique 24-nt Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Mop1</td>
<td>88.3 million</td>
<td>3.4 million</td>
<td>3.9</td>
<td>839,100</td>
</tr>
<tr>
<td>AA</td>
<td>mop1-1</td>
<td>92.1 million</td>
<td>2.7 million</td>
<td>2.9</td>
<td>264,387</td>
</tr>
<tr>
<td>BB</td>
<td>Mop1</td>
<td>187.8 million</td>
<td>7.2 million</td>
<td>3.8</td>
<td>876,446</td>
</tr>
<tr>
<td>BB</td>
<td>mop1-1</td>
<td>89.2 million</td>
<td>2.1 million</td>
<td>2.3</td>
<td>155,300</td>
</tr>
<tr>
<td>AB</td>
<td>Mop1</td>
<td>57.0 million</td>
<td>2.0 million</td>
<td>3.4</td>
<td>441,873</td>
</tr>
<tr>
<td>AB</td>
<td>mop1-1</td>
<td>82.9 million</td>
<td>2.2 million</td>
<td>2.6</td>
<td>185,298</td>
</tr>
</tbody>
</table>
Figure 4: Raw small RNA read distribution as a mean value across all samples after quality control and adapter trimming. Read expression is standardized as reads per million sequenced.
Figure 5: Expression of 24-nt siRNA prior to cluster analysis and filtering of conserved mop1 independent reads detected in both wildtype Mop1 and mutant mop1-1 samples. Error bars represent the standard error of 3 biological replicates.
Figure 6: Frequency distribution of 24-nt siRNA sequences by read count. Reads per unique sequence were tallied and proportions of overall read counts per sequence were calculated for bins with 1^10 read count progression.
3.4.3: siRNA Reduction in Hybrids and \textit{mop1-1}

Raw small RNA reads sequenced ranged from ~57-189 million for each genotype. Of those, approximately 2-7 million (or 2.3-3.9 \% of total) represented 24-nt small RNA. Within these 24-nt subsets, approximately 155,000-876,000 unique raw sequences were observed (Table 2). siRNA meta-clustering analyses produced clusters primarily between 100-500 nt in length. Cluster lengths have an approximate Poisson distribution as frequency diminishes with increasing cluster length (Figure 7). Very few siRNA clusters were detected greater than 500 nt in length with none greater than 800 nt.

Comparisons between \textit{Mop1} and \textit{mop1-1} genotypes revealed fold reductions of unique 24-nt small RNA reads in the mutant of 3.2x, 5.6x, and 2.4x for Inbred A, Inbred B, and Hybrid AB, respectively. When considering siRNA expression in the hybrid vs inbreds, in the \textit{Mop1} state, Hybrid AB exhibited a 1.9 fold reduction in unique 24-nt siRNA as compared to the midparent value of A and B. However, in the \textit{mop1-1} state, there was only a 1.1 fold reduction (Figure 8).

Despite \textit{mop1} loss of function in the mutants, a number of 24-nt producing siRNA expressing loci were detected and retained through the clustering analysis. When siRNA meta-clusters showed expression in all genotypes, including \textit{mop1-1} mutants, these loci were termed “conserved” siRNA meta-clusters. Expression at these loci must operate via a \textit{mop1}-independent pathway (i.e. non-RDR2 driven) as previously described. In order to accurately characterize the effect of \textit{mop1-1} derived hetsiRNAs, these conserved siRNA loci were excluded from further analyses.
Despite the specific request to select for RNA inserts in the range of 18-26 nt, datasets produced by The Clinical Genomics Centre consisted of reads across the entire 0-50 nt spectrum generated by the protocol in the absence of size selection. This was unintended and represented a potential issue with reduced sequencing depth at 24-nt. In order to evaluate the effect of reduced 24-nt sequencing depth, comparisons were analyzed utilizing two different strategies: a) the original experimental design with 3 replicates per genotype, aligned independently, and b) a strategy to increase sequence depth by pooling the 3 replicates prior to alignment. For subsequent analyses where statistical comparisons between genotypes are not considered, pooled results are reported to maximize alignment accuracy. For comparisons between genotypes where a test of significance is required, both replicated and pooled results are reported and indicated as such.

siRNA meta-cluster analysis revealed an almost complete lack of hetsiRNA meta-cluster expression in mop1-1 hybrids as was expected. Supporting the results from analysis of unique reads, significant decreases in 24-nt small RNA cluster abundance in the hybrid AB as compared to the inbreds A and B were detected in both the replicated and pooled alignment analyses (Figure 9). Results from the replicated and pooled analyses were highly similar although the replicated analysis revealed a large variance associated with expression in Inbred A. This could be due to reduced sequence depth though it is unclear as to why variances in other lines were substantially smaller.

To determine the extent to which individual siRNA meta-clustered loci were co-expressed in inbreds and hybrids, Venn diagrams were generated (Figures 10 & 11). When comparing Mop1 wildtype with the mop1-1 mutant in all 3 genetic backgrounds, approximately 13-16% of loci were commonly expressed, with 45-64% of meta-clusters being expressed in
solely the wildtype and 2-4% of meta-clusters being expressed solely in the mutant (Figure 10). When comparing the two parental inbreds and one hybrid, (all Mop1), 42% of all meta-clusters were expressed among all three. 10-12% of meta-clusters were uniquely expressed in one inbred or another and 19% were expressed in both A and B. Combined, this accounts for 41% of meta-clusters that were not detected in the wildtype hybrid at all. In contrast, only 2% of meta-clusters were detected exclusively in the hybrid. When the same backgrounds were compared in the mutant mop1-1 context, very few clusters are shared, with only 1-9% of meta-clusters representing any particular co-expression group and 70% of meta-clusters not represented in any background (Figure 11).
Figure 7: Size distribution of siRNA meta-clusters as a mean of all genotypes after Stage 2 of siRNA meta-cluster analysis. Meta-cluster analysis resulted in a mean cluster length of 238 nt with an approximate Poisson distribution and few siRNA clusters greater than 500 nt in length.
Figure 8: Unique raw 24-nt reads prior to clustering analysis and filtration of mop1 independent siRNA species.
Figure 9: Global net 24-nt siRNA meta-cluster expression as a result of a) replicated analysis and b) pooled replicates.
Figure 10: Venn diagrams demonstrating co-expression of unique 24-nt hetsiRNA meta-clusters in wildtype $Mop1$ and mutant $mop1-1$ prior to filtering of $mop1$ independent 24-nt siRNA species. Meta-clusters were considered co-expressed if detected in both genotypes.
Figure 11: Comparison between unique 24-nt hetsiRNA meta-clusters among genetic inbred and hybrid backgrounds and within mop1 genotypes prior to filtering mop1-1 independent 24-nt siRNA species.
3.4.4: Non-additive siRNA Expression

Previous reports detected a trend in siRNA expression where, at loci with large differences in expression between parents, an associated reduction in siRNA expression in the F1 relative to the expected midparent was observed (Barber et al., 2012b; Groszmann et al., 2011b). To test for a similar trend in this dataset, siRNA expression fold change between inbreds A and B was calculated as HP/LP where HP is the expression level of the high parent and LP is that of the low parent. Fold deviation from expected midparent value (MPV) was calculated as \( \log_2(\text{AB}/\text{MPV}) \) where AB is the expression of the F1 hybrid. Results show an overall reduction in siRNA expression in the hybrid as compared to MPV (Figure 12). There is no trend indicating that large differences in hetsiRNA expression between the two inbred parents are associated with reductions in hetsiRNA expression in the F1 relative to the expected midparent value (Figure 12). (Linear regression statistics were \( m = -0.017, R^2 = 0.018, p\text{-value} = 0.999 \)).

To examine siRNA meta-cluster expression profiles in the context of parental levels, each meta-cluster expression was categorized in relation to its expression in the high parent (HP), low parent (LP), and mid-parent value (MPV). Replicated alignments were utilized to enable statistical comparison and category determination was done by t-test (\( \alpha = 0.1 \)) comparing hybrid meta-cluster expression levels to parental expression levels. According to statistical significance in these comparisons, each hybrid meta-cluster was assigned to one of the following categories: below LP, equal to LP, equal to both LP and MPV (ambiguous), equal to MPV, equal to both MPV and HP (ambiguous), equal to HP, and above HP (Figure 13a). The majority of meta-clusters were expressed at below midparent levels, with 65% being expressed below MPV and 22% expressed at or lower than LP. Since statistical significance testing resulted in ambiguous classification (i.e. meta-cluster expression equivalent to multiple groups), these ambiguous
classifications were also examined. In this case, 58% of non-ambiguous meta-clusters were expressed below MP with only 4% above MPV (Figure 13b). Similar results were observed when utilizing pooled samples in meta-cluster analysis (Figure 14).
Figure 12: siRNA meta-cluster expression deviation from expected midparent value in Hybrid AB, expressed as $\log_2(AB/\text{MPV})$ and fold change between parents A and B, expressed as $\max(A,B)/\min(A,B)$. 
Figure 13: Replicated siRNA meta-cluster analysis of hetsiRNA expression in Hybrid AB relative to expected midparent value derived from inbreds A and B. Classes were defined by t-test at $\alpha = 0.1$. Non-ambiguous classes are reported as those siRNA meta-clusters that were statistically significant in only one unique class.
Figure 14: Pooled siRNA meta-cluster analysis of Hybrid AB expression relative to high, low, and expected mid-parent values of parental inbreds A and B.
3.5: Discussion

While heterosis for agriculturally important traits in crop species has been directly studied for over a century, the underlying molecular basis remains relatively unknown. Typically, heterosis for complex traits has been associated the summation of additive, dominant, and over-dominant effects contributing to a quantitative phenotype consistent with Fisher’s infinitesimal model (Fisher, RA, 1918). Emerging evidence from studies on the molecular basis of heterosis is beginning to point to a highly complex and integrated network of molecular regulation wherein the mode of action at any particular gene may be highly context-dependent (Birchler et al., 2010; Schnable and Springer, 2013). As such, it is unlikely that there exists some grand unifying explanation for the molecular basis of heterosis. The development of deep sequencing technology and rapid technical advances in the field of genomics have enabled the study of epigenetic changes at a resolution far greater than ever before. Indeed, studies of global genomic epigenetic trends and effects have implicated small RNA-mediated epigenetic changes such as DNA methylation and chromatin modifications as being involved with the manifestation of heterosis in ways that move beyond classic studies of additive or dominant effects (Greaves et al., 2015; Groszmann et al., 2013; Ng et al., 2012).

This study sought to characterize 24-nt hetsiRNA expression patterns in maize resulting from hybridization and to examine the effect of the mop1-1 mutation on those patterns. Groszmann et al. (2011b) had observed a global net downregulation of 24-nt siRNAs in Arabidopsis hybrids as compared to expected midparent values as calculated from observed inbred expression. Furthermore, it was observed that the more divergent the parental expression levels were, the more siRNA expression was downregulated. In the present study it was similarly observed that 24-nt siRNAs demonstrated a net global downregulation of expression upon
hybridization with an accompanying loss of diversity in unique hetsiRNA species. Furthermore, from these results it can be reasonably concluded that these globally downregulated 24-nt siRNA trends are indeed related to the Pol IV/RDR2 (mop1) dependent pathway as these trends persisted upon the removal of conserved, mop1 independent, 24-nt RNAs that were detected in the mutant mop1-1 (Figure 8, Figure 9). The nature of these conserved siRNAs is unclear although evidence suggests that 24-nt siRNA production is not exclusively limited to the Pol IV/RDR2 pathway and that some may be produced by a Pol II dependent pathway that does not involve maize mop1 (You et al., 2013; Zheng et al., 2009). Interestingly these conserved 24-nt siRNA reads exhibited a much higher mean expression level than mop1-dependent 24-nt hetsiRNAs.

Certain factors could suggest a global siRNA reduction in hybrids when in fact such a reduction does not exist. For example, the result of global net downregulation of hetsiRNAs in hybrids in this study is likely influenced by the relative degree of coancestry between B73 (i.e. the reference genome) and the germplasm utilized in this study. Inbred A is an Iowa Stiff Stalk Synthetic derivative whereas Inbred B has no direct pedigree relationship to B73. As such one could possibly expect confounding results such as a correlation between expression level and genetic distance to B73. In this case however, it would be expected that the hybrid AB would show results that are intermediate to inbreds A and B, yet this was consistently not the case. Further evidence of a true reduction in hybrid hetsiRNA comes from the analysis of unique raw 24-nt sequences where an approximately 50% reduction was observed in an analysis that was not contingent upon alignment to a reference genome (Table 2, Figure 8). Venn analysis of hetsiRNA clusters in inbreds and hybrids demonstrated interesting trends as well with greater than 40% of hetsiRNA clusters being detected in parental inbreds but not in the hybrid. This
result supports the hypothesis that despite Mendelian inheritance of siRNA producing loci from parental inbreds, there exists some silencing mechanism in the hybrid whereby hetsiRNAs are not expressed.

Studies by Barber et al. (2012b) and Groszmann et al. (2011b) describe results that were similar to this experiment regarding the reduction of 24-nt siRNAs in maize hybrids. In both studies, the authors described a system in which hetsiRNAs are globally downregulated with a trend toward a greater reduction in expression relative to the midparent with increasingly divergent expression in the parents. In the present study, this was not observed to be the case. The results here clearly show the same global net downregulation however there was no correlation between this and parental expression divergence. These differing results may have to do with dynamic hetsiRNA expression profiles that are dependent upon specific tissue and developmental stage.

In the study by Groszman et al. (2016b), whole seedlings were sampled whereas in the study by Barber et al. (2012b) comparative analyses were done using RNA pools from both the developing ear and the shoot apical meristem. In the case of the developing ear dataset, the association between divergent parental expression and hybrid downregulation was observed. In the shoot apical meristem, it was not observed, despite it being a rich source of siRNAs (Nobuta et al., 2008). In the present study, vegetative leaf tissue was sampled under the working hypothesis that establishment of epigenetic marks is not an instantaneous process and that siRNA expression profiles were more likely to be stabilized in mature tissue types. Taken together, these results demonstrate context specificity where non-additive hetsiRNA expression profiles are likely highly dependent upon tissue type and developmental stage. In particular, the siRNA expression profiles observed in the developing ear may be a result of increased activation of
transposable elements in germline cells (Slotkin and Martienssen, 2007). Interestingly there is
some evidence that a specific subset of 24-nt hetsiRNAs (p4-siRNAs) are initiated in the
maternal gametophyte, that they are maternally expressed in the developing ear, and that they
play a role in establishing the initial distinction between self and non-self when two genomes
come into contact (Mosher et al., 2009). These p4-siRNAs could conceivably play some role in
the overall reduction in hetsiRNAs in F1 hybrids. As such it would be interesting to compare
pools of hetsiRNAs between reciprocal hybrids to see if maternal inheritance has an effect on
hetsiRNA pools.

Further to evidence presented by Barber et al. (2012b) and Groszmann et al. (2011a)
where only net global expression levels were reported, analysis of hybrid hetsiRNA cluster
expression relative to parental expression at individual clusters revealed a strong distribution
toward hybrid cluster expression at or below levels detected in the low parent. Both replicated
and pooled analyses pointed to significant downregulation as approximately 80% of specific
hetsiRNA clusters showing expression at or below midparent. Furthermore, 10% of loci were
expressed below the low parent, suggesting a transgressive mechanism of silencing.

While the previously reported correlation between differential parental expression and
hybrid downregulation was not detected in this study, alterations in methylation status have
previously been observed at loci showing siRNA expression reduction in hybrids (Gent et al.,
2013; Groszmann et al., 2011b; He et al., 2010). It is possible that high expression levels in one
parent are sufficient to establish a switch from non-symmetric CHH maintenance of methylation
status by RdDM to symmetric maintenance of methylation by the non-RdDM feedforward loop.
Thus hetsiRNAs and RdDM may initiate trans-chromosomal methylation in all contexts and an
accumulation of methylated nucleotides would cause downregulation of further siRNA
production at that locus, switching it to feed-forward maintenance by symmetric methylation. It remains unclear however, as to how highly expressive siRNA persist without silencing in the inbred. These results fit with recent observations suggesting that trans-chromosomal interactions between siRNA producing loci may produce paramutation-like events that have non-additive effects on DNA methylation and possibly gene expression (Greaves et al., 2012; Regulski et al., 2013; Shivaprasad et al., 2012).

One limitation of the dominance theory is that it fundamentally requires that heterotic phenotypes should be recoverable in inbreds over time due to segregation and accumulation of dominant alleles by outcrossing and subsequent inbreeding. As noted above, this however has not been observed, as the magnitude of heterosis in commercial maize has not diminished over time despite intensive breeding efforts (Duvick, 1984; Duvick et al., 2004). Should hetsiRNA mediated epialleles be involved in heterosis, it would represent an explanation for the lack of reduction in heterosis by net accumulation of dominant alleles. While these dominant alleles would accumulate on the genetic level, their epigenetic states would not necessarily co-segregate as they may become altered in a non-Mendellian fashion via hetsiRNA mediated pathways. Such may be the case where Reinders et al. (2009) demonstrated a system where epialleles in Arabidopsis do not necessarily co-segregate with their genetic alleles. Furthermore, these epialleles, once established in the F1 may also demonstrate unstable and dynamic changes in subsequent generations beyond the F1 hybrid (Ha et al., 2009).
Chapter 4: hetsiRNA Effects on Gene Expression

4.1: Abstract

Heterochromatic small interfering RNAs (hetsiRNAs) have been demonstrated to play an active role in the establishment and maintenance of DNA methylation via the RNA-dependent DNA methylation pathway. Recent findings have noted the enriched localization of hetsiRNAs to regions of the genome flanking genes. This raises the possibility that hetsiRNA may play a role in regulating gene transcription by influencing DNA methylation and heterochromatin formation around those genes. This chapter examines the association between gene transcript expression and expression of hetsiRNAs in hybrid maize. Two commercial maize inbred lines and their F1 hybrid were subjected to deep sequencing for small RNA and mRNA. In order to examine the effect of knocking out hetsiRNA production on gene expression, both Mop1 and mop1-1 allelic states were sequenced for each genotype. Association of hetsiRNA reads and mRNA transcription was done by clustering analysis. Localization of hetsiRNA clusters within the boundaries of a genic frame demonstrated a strong enrichment of hetsiRNA clusters in the +/- 1 kb flanks of genic regions. Expression level analyses indicate a mean downregulation of both hetsiRNA and mRNA expression relative to expected midparent values for a small but significant number of genes. Furthermore, knockout of hetsiRNA production in the mutant mop1-1 demonstrated a trend toward upregulation of genes relative to the wildtype. Together these results suggest a model in which hetsiRNAs function to regulate heterochromatin formation at gene boundaries which may subsequently regulate expression for a subset of genes.
4.2: Introduction

RNA-dependent DNA methylation represents a highly complex network of interacting factors that serve to regulate heterochromatin formation at specific genomic loci. The canonical model of RdDM involves 24-nt hetsiRNA production via the Pol IV / RDR2 pathway which target specific genomic sequences (Matzke and Mosher, 2014). An additional plant-specific RNA polymerase (Pol V) functions to transcribe long non-coding scaffold RNA at the target site that are essential for recruitment of the hetsiRNA/AGO4 complex (Wierzbicki et al., 2008, 2012). Proteins associated with chromatin modification and methylation are then recruited to the target site and establish methylation in all three sequence contexts (CG, CHG, and CHH) (Péllissier et al., 1999). While hetsiRNAs and RdDM are not required for maintenance of symmetric CG and CHG methylation, they are required to maintain asymmetric CHH methylation (Chan et al., 2004; Jones et al., 2001; Zilberman et al., 2003).

Recent studies have demonstrated that hetsiRNAs and CHH methylation are preferentially clustered around the flanking regions of genes, raising the possibility that they function in regulating heterochromatin/euchromatin boundaries and possibly gene expression (Gent et al., 2011, 2013, 2014; Li et al., 2015). Analysis of methylation status and transposon activation has suggested that hetsiRNAs and CHH methylation serve to establish a type of intermediate heterochromatin that serves as a boundary against the excessive spread of euchromatin outside of genic regions (Li et al., 2015; Zheng et al., 2013). Indeed, it was further established that DNA associated with hetsiRNA production and CHH methylation was more accessible to transcriptional machinery than those loci not undergoing active RdDM and that these loci represented approximately 2% of the genome (Gent et al., 2014). Thus, in these studies there is a positive correlation between hetsiRNA production, CHH methylation, and gene
expression. Contrary to these results, others have found a negative correlation between hetsiRNA production and gene expression. (Wei et al., 2014) examined a rice OsDCL3a (a maize mop1 ortholog) mutant and determined that reduction in hetsiRNA mediated methylation of MITE transposons near genic regions was associated with gene upregulation at those loci.

Trans-generational inheritance of epigenetic states has been well documented in plants. It was first discovered as McClintock’s identification of transposable elements and observations of “changes of phase” in maize (McClintock, 1950). More recent studies have established the link between feed-forward mechanisms of symmetric methylation maintenance and RdDM mediated maintenance of asymmetric methylation (Bond and Baulcombe, 2014; Heard and Martienssen, 2014). Various studies have utilized transgene technology to establish methylation by RdDM but then go on to observe the maintenance of methylation status even in the absence of the transgene (Soppe et al., 2000). Similar examples exist in the study of paramutation where a paramutagenic allele alters the methylation status and expression of a paramutagenic allele in a heritable way. In many paramutation examples, RdDM and hetsiRNAs have also been implicated as playing a role in stable inheritance of epigenetic states (Hollick, 2012).

Building on evidence that hetsiRNA expression and accumulation of methylated residues in genic regions can have a significant effect on gene expression, it is therefore of particular interest to determine the extent to which hetsiRNA-mediated trans-chromosomal methylation affects gene expression in hybrids. It has been documented here and in previous reports (Barber et al., 2012b; Groszmann et al., 2011b) that hetsiRNAs are expressed non-additively in hybrids. As such it follows that they could play a significant role in regulation of non-additive gene expression in hybrids.
With respect to hybrid crop development, epigenetic modifications could play a significant role in two similar yet distinct contexts in the breeding process. The first context is that of inbred development where two parental genomes from within a heterotic group come into contact in the nucleus of an F1 plant. Here, hetsiRNA populations from one parental genome could mediate trans-chromosomal methylation of alleles from the other parental genome. In this situation, heritability is a factor as lines are subsequently inbred, resulting in distinct inheritance of an allele whose epigenetic state may be determined by unlinked loci, as was the case in Arabidopsis epi-RILs (Greaves et al., 2012; Reinders et al., 2009). Thus, the correlation between genotype and phenotype may change if the epigenetic state becomes altered by inter-genomic interactions. Recent evidence from studies in maize also point to the widespread potential for trans-chromosomal events to shape the epigenome in a way that cannot be inferred from genetic information such as SNP analysis and other techniques currently in use (Regulski et al., 2013).

The second context is similar in that trans-chromosomal methylation events shape the F1 epigenome in a way that is dependent on the interactions between genetic and epigenetic inheritance yet inheritance of these altered epialleles beyond the F1 is not a primary concern. An example of this would be crosses between inbreds of different heterotic groups, such as in maize F1 hybrids that are commercially grown. Here it would only be of importance to accurately predict epigenetic effects on phenotype in the F1.

While Chapter 3 focused on characterization of hetsiRNA expression in hybrids, the purpose of this study was to examine the relationship between hetsiRNA production and gene expression in both inbreds and hybrids. The study first utilized mRNA-seq and mop1 null mutant NILs to examine the effects of eliminating hetsiRNA production. hetsiRNA were subjected to a clustering analysis that allowed the association of specific hetsiRNA loci with specific genes.
Here it was hypothesized that hetsiRNA clusters map to the flanking regions of genic intervals and that they are associated with control of gene expression. Since these hetsiRNA are generally associated with changes in chromatin state at gene boundaries, it was hypothesized that hetsiRNA reduction in the mutant would lead to net reduction of gene expression at target loci relative to the wildtype. Building on the results from Chapter 3, where hetsiRNA exhibited non-additive expression levels in the hybrid below midparent values, it was also hypothesized that downregulation of hetsiRNA at genic loci in hybrids causes a decrease in gene expression relative to MPV.

4.3: Materials and Methods

4.3.1: mRNA Sequencing and Alignment Pipeline:

Genetic material, tissue sampling, and RNA extraction were from the same materials as previously described in sections 3.3.1 to 3.3.3. Library preparation and mRNA sequencing was done by The Clinical Genomics Centre in Toronto, ON, Canada. A total of 36 independent mRNA libraries were prepared from total RNA using Illumina HiSeq 2000 protocols (Figure 1). Libraries were sequenced on the Illumina HiSeq 2000 platform and run through the CASAVA post-processing pipeline to perform de-multiplexing and basic file conversion. Reads were 102 bases long.

Initial read quality was assessed using FASTX toolkit v0.0.13. Alignments were made to the Zea mays B73 RefGen_v2 assembly (Release 5b) (Wei et al., 2010) using Tophat v2.0.7 (Kim et al., 2013; Langmead et al., 2009) and Bowtie v2.0.6 (Langmead and Salzberg, 2012). See Figure 15 for a schematic representation of the mRNA processing pipeline workflow. Mate inner distance was set to 179 nt (-r 179) to reflect the mean insert size in the library. Since it was
necessary to compare gene expression among different genotypes, use of a common set of genes was necessary. The –G option was invoked to preferentially map splice junctions to the B73 RefGen_v2 Filtered Gene Set (FGS) and splice junctions were restricted to only those transcripts using the --no-novel-juncs option. This analysis was done using paired reads and conservative alignment parameters were used to limit mapping to a) alignments where both mate pairs mapped to the reference (--) and b) mate pairs were concordant (--). All other parameters were default settings. Transcript expression was normalized as Fragments per Kilobase Million (FPKM). Normalization was done with Cufflinks v2.0.2 (Trapnell et al., 2012) using default parameters. Comparative differential expression analysis was done with Cuffdiff which is contained within the previously described Cufflinks package. Multiple comparison testing was controlled by using a false discovery rate of 0.04 (FDR 0.04).

GO term enrichment analysis was done using agriGO (Du et al., 2010) using the Singular Enrichment Analysis feature. The query dataset consisted of differentially expressed genes between Mop1 and mop1-1, in hybrid AB (504 genes, of which 359 had GO annotations). The reference dataset consisted of all genes with detectable expression (19,368 genes, of which 12,192 had GO annotations). GO annotations were from the Zea mays V5a dataset available from maizesquence.org.

### 4.3.2: siRNA/gene Meta-Clustering Analysis:

In order to compare siRNA meta-clusters with annotated genes, a three stage siRNA clustering process was used. siRNA pre-processing and the initial two clustering stages are previously described in Chapter 3. Third stage clustering combined siRNA meta-clusters and annotated genes on the basis of siRNA-metacluster distance using a custom R script. Since
multiple siRNA meta-clusters may reside within flanking regions of a gene, the script compares all possible pair-wise combinations of siRNA meta-cluster and annotated gene locations. Previous results indicate that 24-nt siRNAs and CHH methylation islands are found to be enriched within 1 kb of genic regions (Groszmann et al., 2011b; Li et al., 2015). Therefore, if an siRNA-metacluster was located within +/- 1 kb of a gene, the two were combined into an siRNA/gene meta-cluster (Figure 16). Thus, both localization of the siRNA relative to the genic region and co-localization of siRNA and mRNA expression could be examined.

To examine internal localization of siRNA clusters within an siRNA/gene metacluster, a normalized genic frame was established that consisted of 1000 nt of upstream flanking sequence, a genic region standardized to the mean nucleotide length of all genes in the dataset (5,374 nt), and a downstream 1000 nt flanking sequence. Each siRNA cluster position was calculated as a proportional distance along this frame, from the start position (-1000) to the midpoint of the siRNA cluster. siRNA clusters with midpoints in the -1,000-0 range were termed upstream, those between 0-5,374 were termed genic and those in the 5,374+ range were termed downstream. Those siRNA clusters not mapping to +/- 1 kb from genic regions were termed intergenic.

4.3.3: siRNA/gene Meta-Cluster Expression:

Correlation between non-additive gene and siRNA expression within siRNA/gene metaclusters was examined by first testing for statistically significant deviation of mRNA expression in the hybrid AB from expected midparent value as calculated from expression levels of inbreds A and B. This was done by conducting a two-tailed, one-sample t-test using MPV FPKM expression as the hypothesized mean and hybrid AB FPKM expression as the observed mean where MPV was the mean of AA FPKM and AB FPKM. The t-test was conducted with n -
1 = 2 degrees of freedom (equation 1, where $t$ is the test statistic, $\bar{x}$ is the observed mean, $\mu_0$ is the hypothesized mean, $s$ is the sample standard deviation, and $n$ is the sample size).

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Filters were applied such that genes with only 1 biological replicate showing detectable expression were excluded from the analysis. No minimum FPKM threshold was applied. Correction for multiple testing was done by utilizing the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) with a False Discovery Rate of 0.05.

![Workflow diagram](image_url)

Figure 15: Workflow of mRNA-seq processing, alignment, and differential expression analysis pipeline.
Clustering Stage 1: siRNA Cluster

Clustering Stage 2: siRNA Meta-Cluster

Clustering Stage 3: siRNA/gene Meta-Cluster

Figure 16: Diagrammatic representation of three stage siRNA/gene meta-cluster analysis. Stage 1 involves cluster siRNA into primary clusters, Stage 2 establishes a meta-cluster as a common frame of reference between genotypes. Stage 3 associates siRNA meta-clusters with nearby or overlapping annotated gene features within a +/- 1 kb frame.
4.4: Results

4.4.1: RNA-seq Alignment

Deep sequencing of mRNA across all 36 samples (3 inbreds + 3 hybrids x 2 mop1 states x 3 biological replicates) resulted in raw read numbers between 18 and 40 million reads per sample (Table 3). In order to generate alignments and estimate gene expression, raw reads were aligned to the B73 RefGen v2 using Tophat and FPKM was calculated using Cufflinks. Within samples, successful alignment of paired reads as a ratio of total QC-passed reads ranged from 0.7x to 2.5x, with an average of 0.9x. Unmapped reads ranged from 26 to 40% (Table 3). The B73 RefGen_v2 Filtered Gene Set (FGS) contains 39,656 annotated genes. When RNA-seq alignment was constrained to the FGS, expression of approximately 11,000-13,000 genes (or roughly 1/3 of annotated genes) were detected. Each genotype had approximately 100-200 uniquely expressed genes that were not detected in any other genotype.

Considering reads within a genetic background (inbred or hybrid genotypes) and comparing Mop1 versus mop1-1, samples exhibited expression of approximately 1,000-2,000 unique genes expressed in either mop1 state. There was no clear pattern of unique gene expression preference in Mop1 or mop1-1 (i.e. neither Mop1 nor mop1-1 showed an appreciable number of uniquely expressed genes as compared to the other). As previously described, some irregularities were observed with Inbred C both phenotypically and in small RNA datasets (Table 1, Table 4). In the RNA-seq datasets, this trend is also observable in excessively high and abnormal maximum FPKM data as well as high numbers of uniquely expressed genes (Table 4). This likely points to errors in library preparation and the RNA sequencing pipeline. As in the small RNA dataset analyses, for subsequent analyses in this chapter, inbreds A, B and hybrid AB will be considered further.
Table 3: Raw mRNA paired-read sequencing and alignment statistics.

<table>
<thead>
<tr>
<th>Background</th>
<th>Genotype</th>
<th>Raw Reads</th>
<th>QC-Passed Reads (%)</th>
<th>Reads Mapped</th>
<th>Properly Paired (%)</th>
<th>Unmapped Reads (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>mop1-1</td>
<td>29,015,116</td>
<td>99.66</td>
<td>21,842,124</td>
<td>97.34</td>
<td>32.19</td>
</tr>
<tr>
<td>A</td>
<td>mop1-1</td>
<td>19,732,484</td>
<td>99.67</td>
<td>13,971,746</td>
<td>98.00</td>
<td>33.84</td>
</tr>
<tr>
<td>A</td>
<td>mop1-1</td>
<td>24,501,330</td>
<td>99.67</td>
<td>19,716,198</td>
<td>96.91</td>
<td>34.45</td>
</tr>
<tr>
<td>B</td>
<td>mop1-1</td>
<td>21,697,966</td>
<td>99.67</td>
<td>19,456,470</td>
<td>95.58</td>
<td>39.30</td>
</tr>
<tr>
<td>B</td>
<td>mop1-1</td>
<td>22,303,268</td>
<td>99.67</td>
<td>14,615,954</td>
<td>97.86</td>
<td>40.24</td>
</tr>
<tr>
<td>B</td>
<td>mop1-1</td>
<td>20,971,628</td>
<td>99.67</td>
<td>51,498,766</td>
<td>91.02</td>
<td>36.70</td>
</tr>
<tr>
<td>C</td>
<td>mop1-1</td>
<td>18,321,498</td>
<td>99.67</td>
<td>15,821,328</td>
<td>95.89</td>
<td>34.90</td>
</tr>
<tr>
<td>C</td>
<td>mop1-1</td>
<td>24,272,752</td>
<td>99.66</td>
<td>18,508,650</td>
<td>91.02</td>
<td>36.80</td>
</tr>
<tr>
<td>C</td>
<td>mop1-1</td>
<td>24,713,248</td>
<td>99.66</td>
<td>11,424,622</td>
<td>94.99</td>
<td>35.42</td>
</tr>
<tr>
<td>A/B</td>
<td>mop1-1</td>
<td>28,024,890</td>
<td>99.66</td>
<td>23,661,332</td>
<td>96.78</td>
<td>32.80</td>
</tr>
<tr>
<td>A/B</td>
<td>mop1-1</td>
<td>19,529,234</td>
<td>99.66</td>
<td>15,892,140</td>
<td>96.81</td>
<td>31.29</td>
</tr>
<tr>
<td>A/B</td>
<td>mop1-1</td>
<td>29,695,640</td>
<td>99.67</td>
<td>30,700,288</td>
<td>95.42</td>
<td>29.57</td>
</tr>
<tr>
<td>A/C</td>
<td>mop1-1</td>
<td>34,319,764</td>
<td>99.66</td>
<td>36,325,178</td>
<td>95.09</td>
<td>33.14</td>
</tr>
<tr>
<td>A/C</td>
<td>mop1-1</td>
<td>34,144,408</td>
<td>99.66</td>
<td>28,011,128</td>
<td>96.78</td>
<td>32.87</td>
</tr>
<tr>
<td>A/C</td>
<td>mop1-1</td>
<td>38,729,990</td>
<td>99.67</td>
<td>29,025,394</td>
<td>97.11</td>
<td>38.23</td>
</tr>
<tr>
<td>B/C</td>
<td>mop1-1</td>
<td>35,770,220</td>
<td>99.66</td>
<td>25,070,740</td>
<td>98.06</td>
<td>35.42</td>
</tr>
<tr>
<td>B/C</td>
<td>mop1-1</td>
<td>30,711,178</td>
<td>99.67</td>
<td>21,521,150</td>
<td>97.32</td>
<td>39.05</td>
</tr>
<tr>
<td>B/C</td>
<td>mop1-1</td>
<td>35,138,246</td>
<td>99.66</td>
<td>17,895,052</td>
<td>96.43</td>
<td>60.05</td>
</tr>
<tr>
<td>A</td>
<td>Mop1</td>
<td>35,600,666</td>
<td>99.67</td>
<td>28,965,006</td>
<td>97.64</td>
<td>26.29</td>
</tr>
<tr>
<td>A</td>
<td>Mop1</td>
<td>36,550,504</td>
<td>99.67</td>
<td>28,675,020</td>
<td>97.96</td>
<td>27.83</td>
</tr>
<tr>
<td>A</td>
<td>Mop1</td>
<td>33,209,094</td>
<td>99.66</td>
<td>26,760,374</td>
<td>97.74</td>
<td>27.75</td>
</tr>
<tr>
<td>B</td>
<td>Mop1</td>
<td>33,201,298</td>
<td>99.67</td>
<td>29,163,756</td>
<td>96.96</td>
<td>32.85</td>
</tr>
<tr>
<td>B</td>
<td>Mop1</td>
<td>32,610,956</td>
<td>99.67</td>
<td>24,694,634</td>
<td>97.90</td>
<td>33.02</td>
</tr>
<tr>
<td>B</td>
<td>Mop1</td>
<td>31,638,398</td>
<td>99.66</td>
<td>24,684,652</td>
<td>97.17</td>
<td>35.49</td>
</tr>
<tr>
<td>C</td>
<td>Mop1</td>
<td>34,505,758</td>
<td>99.76</td>
<td>35,078,792</td>
<td>94.77</td>
<td>34.51</td>
</tr>
<tr>
<td>C</td>
<td>Mop1</td>
<td>38,602,304</td>
<td>99.74</td>
<td>57,467,310</td>
<td>91.90</td>
<td>35.52</td>
</tr>
<tr>
<td>C</td>
<td>Mop1</td>
<td>38,453,512</td>
<td>99.75</td>
<td>33,309,864</td>
<td>95.99</td>
<td>35.01</td>
</tr>
<tr>
<td>A/B</td>
<td>Mop1</td>
<td>33,375,570</td>
<td>99.75</td>
<td>21,968,374</td>
<td>95.40</td>
<td>52.23</td>
</tr>
<tr>
<td>A/B</td>
<td>Mop1</td>
<td>38,208,840</td>
<td>99.76</td>
<td>30,456,554</td>
<td>97.72</td>
<td>30.17</td>
</tr>
<tr>
<td>A/B</td>
<td>Mop1</td>
<td>39,954,292</td>
<td>99.74</td>
<td>30,676,364</td>
<td>98.06</td>
<td>31.22</td>
</tr>
<tr>
<td>A/C</td>
<td>Mop1</td>
<td>37,215,822</td>
<td>99.76</td>
<td>32,472,578</td>
<td>97.07</td>
<td>29.12</td>
</tr>
<tr>
<td>A/C</td>
<td>Mop1</td>
<td>38,198,168</td>
<td>99.76</td>
<td>30,127,164</td>
<td>97.53</td>
<td>28.97</td>
</tr>
<tr>
<td>A/C</td>
<td>Mop1</td>
<td>34,558,374</td>
<td>99.76</td>
<td>39,394,388</td>
<td>95.29</td>
<td>30.22</td>
</tr>
<tr>
<td>B/C</td>
<td>Mop1</td>
<td>36,876,360</td>
<td>99.75</td>
<td>36,550,786</td>
<td>95.42</td>
<td>33.65</td>
</tr>
<tr>
<td>B/C</td>
<td>Mop1</td>
<td>33,448,104</td>
<td>99.76</td>
<td>25,446,132</td>
<td>97.45</td>
<td>32.79</td>
</tr>
<tr>
<td>B/C</td>
<td>Mop1</td>
<td>37,397,696</td>
<td>99.76</td>
<td>65,454,434</td>
<td>91.83</td>
<td>32.51</td>
</tr>
</tbody>
</table>
Table 4: Gene expression statistics for mRNA-sequencing alignments.

<table>
<thead>
<tr>
<th>Background</th>
<th>Genotype</th>
<th>Max FPKM</th>
<th>Expressed Genes</th>
<th>Unique Genes (Entire Dataset)</th>
<th>Unique Genes Expressed - mop1-1</th>
<th>Unique Genes Expressed – Mop1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>wildtype</td>
<td>4,438</td>
<td>11,711</td>
<td>109</td>
<td>361</td>
<td>1,774</td>
</tr>
<tr>
<td>AA</td>
<td>mop1-1</td>
<td>7,524</td>
<td>12,342</td>
<td>212</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BB</td>
<td>wildtype</td>
<td>6,239</td>
<td>12,282</td>
<td>150</td>
<td>285</td>
<td>1,334</td>
</tr>
<tr>
<td>BB</td>
<td>mop1-1</td>
<td>6,005</td>
<td>11,778</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>wildtype</td>
<td>22,208</td>
<td>26,033</td>
<td>10,194</td>
<td>10,786</td>
<td>1,054</td>
</tr>
<tr>
<td>CC</td>
<td>mop1-1</td>
<td>4,860</td>
<td>12,771</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>wildtype</td>
<td>6,456</td>
<td>12,652</td>
<td>145</td>
<td>320</td>
<td>1,888</td>
</tr>
<tr>
<td>AB</td>
<td>mop1-1</td>
<td>5,060</td>
<td>13,015</td>
<td>151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td>wildtype</td>
<td>5,138</td>
<td>12,846</td>
<td>175</td>
<td>414</td>
<td>1,710</td>
</tr>
<tr>
<td>AC</td>
<td>mop1-1</td>
<td>4,455</td>
<td>12,404</td>
<td>212</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>wildtype</td>
<td>5,417</td>
<td>11,758</td>
<td>137</td>
<td>291</td>
<td>1,923</td>
</tr>
<tr>
<td>BC</td>
<td>mop1-1</td>
<td>10,877</td>
<td>12,648</td>
<td>129</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.4.2: Differential Gene Expression in *Mop1* vs. *mop1-1*

To examine the effect of *mop1-1* mutation on gene transcript levels, *Mop1* versus *mop1-1* samples within each genetic background were analyzed using the Cuffdiff function within Cufflinks. Of approximately 20,000 genes that passed internal quality thresholds, 1,881 genes were differentially expressed (p-value < 0.05, FDR = 0.04) between *Mop1* and *mop1-1* in Inbred A, 744 in Inbred B, and 504 in Hybrid AB. Further examination of significantly differentially expressed genes in *Mop1* vs *mop1-1* revealed that genes in the *mop1-1* genotype are upregulated far more often than downregulated with Hybrid AB demonstrating the largest proportional upregulation (Figure 17). In order to examine the distribution of these significant genes, differential expression was calculated as log$_2$(FPKM *Mop1*/ FPKM *mop1-1*). Differential expression distributions for each inbred genetic background are depicted in Figure 18. These results further support the trend towards a distribution that favours upregulation of genes in the *mop1-1* state (i.e. a distribution skewed toward the left of the plot). Although there was one case where the maximal count class resides on the positive side of the log$_2$(Mop1/mop1-1) distribution, indicating upregulation in *Mop1* (e.g. Inbred A), the bulk of the distribution and magnitude of effects in this case both trend toward *mop1-1* gene upregulation. When the hybrid AB was considered, a similar yet stronger trend was observed (Figure 18). Indeed, the mean differential expression in each background was negative, indicating an overall trend toward gene upregulation in *mop1-1* with the largest number of genes being observed in the hybrid (Figure 18).
Figure 17: Differential expression of genes in *Mop1* relative to *mop1-1* genotypes. Count indicates the number of statistically significant genes. Blue bars indicate the proportion of genes upregulated in *Mop1* relative to *mop1-1*. Red bars indicate the proportion of genes upregulated in *mop1-1* relative to *Mop1*.
Figure 18: Distribution of differential gene expression in Inbred A, Inbred B, and Hybrid AB at individual genes found to be significantly different between *Mop1* and *mop1-1*.
4.4.3: Gene Ontology Enrichment Analysis

In order to examine biological and molecular processes preferentially associated with differentially expressed genes in Mop1 vs mop1-1, Gene Ontology (GO) enrichment analysis was performed by agriGO utilizing significant differentially expressed genes between Mop1 and mop1-1 in hybrid AB (504 genes, of which 359 had GO annotations) and comparing to a reference set consisting of all genes with detectable expression (19,368 genes, of which 12,192 had GO annotations). A total of 6 biological process GO terms and 13 molecular function GO terms were significant at an FDR of 0.05 (Table 5). No significant GO enrichment was found for genes significantly upregulated in Mop1 backgrounds. Biological processes associated with “negative regulation of molecular function” and “catalytic activity” represented one enriched group under regulation of molecular function while genes associated with “lipid transport and localization” as well as “oxidation/reduction” represented another group of enrichment (Table 5, Figure 19). Various molecular functions were also enriched in differentially regulated genes. These represented general categories of nucleic acid or protein binding, and catalytic activity (Figure 20). Most notably, genes related to transcription factor activity and methyltransferase activity were significantly enriched. This points to a role for mop1 in regulation of DNA methylation status establishment and maintenance as well as a possible role in gene regulation via regulation of transcription factors, although one cannot discount the possibility of epistatic interactions in cascading gene networks as a root cause of this enrichment.
Table 5: Gene Ontology term enrichment analysis for genes significantly upregulated in *mop1-1* backgrounds. No significant GO enrichment was found for genes significantly upregulated in *Mop1* backgrounds.

<table>
<thead>
<tr>
<th>GO term</th>
<th>Ontology</th>
<th>Description</th>
<th>Number in input list</th>
<th>Number in BG/Ref</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0044092</td>
<td>Biological Process</td>
<td>negative regulation of molecular function</td>
<td>7</td>
<td>24</td>
<td>0.0000046</td>
<td>0.0017</td>
</tr>
<tr>
<td>GO:0043086</td>
<td>Biological Process</td>
<td>negative regulation of catalytic activity</td>
<td>7</td>
<td>24</td>
<td>0.0000046</td>
<td>0.0017</td>
</tr>
<tr>
<td>GO:0006869</td>
<td>Biological Process</td>
<td>lipid transport</td>
<td>7</td>
<td>36</td>
<td>0.00008</td>
<td>0.015</td>
</tr>
<tr>
<td>GO:0010876</td>
<td>Biological Process</td>
<td>lipid localization</td>
<td>7</td>
<td>36</td>
<td>0.00008</td>
<td>0.015</td>
</tr>
<tr>
<td>GO:0055114</td>
<td>Biological Process</td>
<td>oxidation reduction</td>
<td>47</td>
<td>875</td>
<td>0.00016</td>
<td>0.024</td>
</tr>
<tr>
<td>GO:0006633</td>
<td>Biological Process</td>
<td>fatty acid biosynthetic process</td>
<td>9</td>
<td>70</td>
<td>0.00023</td>
<td>0.029</td>
</tr>
<tr>
<td>GO:0020037</td>
<td>Molecular Function</td>
<td>heme binding</td>
<td>24</td>
<td>228</td>
<td>0.00000015</td>
<td>0.000026</td>
</tr>
<tr>
<td>GO:0046906</td>
<td>Molecular Function</td>
<td>tetrapyrrole binding</td>
<td>24</td>
<td>230</td>
<td>0.00000018</td>
<td>0.000026</td>
</tr>
<tr>
<td>GO:0005506</td>
<td>Molecular Function</td>
<td>iron ion binding</td>
<td>32</td>
<td>355</td>
<td>0.00000066</td>
<td>0.000026</td>
</tr>
<tr>
<td>GO:0042802</td>
<td>Molecular Function</td>
<td>identical protein binding</td>
<td>8</td>
<td>34</td>
<td>0.0000056</td>
<td>0.00062</td>
</tr>
<tr>
<td>GO:0004497</td>
<td>Molecular Function</td>
<td>monooxygenase activity</td>
<td>17</td>
<td>167</td>
<td>0.000013</td>
<td>0.0012</td>
</tr>
<tr>
<td>GO:0009055</td>
<td>Molecular Function</td>
<td>electron carrier activity</td>
<td>23</td>
<td>318</td>
<td>0.00012</td>
<td>0.0087</td>
</tr>
<tr>
<td>GO:0016747</td>
<td>Molecular Function</td>
<td>transferase activity, transferring acyl groups other than amino-acyl groups</td>
<td>13</td>
<td>133</td>
<td>0.00019</td>
<td>0.011</td>
</tr>
<tr>
<td>GO:0008171</td>
<td>Molecular Function</td>
<td>O-methyltransferase activity</td>
<td>5</td>
<td>19</td>
<td>0.00019</td>
<td>0.011</td>
</tr>
<tr>
<td>GO:0004252</td>
<td>Molecular Function</td>
<td>serine-type endopeptidase activity</td>
<td>11</td>
<td>107</td>
<td>0.00037</td>
<td>0.017</td>
</tr>
<tr>
<td>GO:0008289</td>
<td>Molecular Function</td>
<td>lipid binding</td>
<td>9</td>
<td>75</td>
<td>0.00039</td>
<td>0.017</td>
</tr>
<tr>
<td>GO:0016491</td>
<td>Molecular Function</td>
<td>oxidoreductase activity</td>
<td>52</td>
<td>1062</td>
<td>0.00065</td>
<td>0.025</td>
</tr>
<tr>
<td>GO:0003700</td>
<td>Molecular Function</td>
<td>transcription factor activity</td>
<td>25</td>
<td>406</td>
<td>0.00069</td>
<td>0.025</td>
</tr>
<tr>
<td>GO:0016746</td>
<td>Molecular Function</td>
<td>transferase activity, transferring acyl groups</td>
<td>13</td>
<td>165</td>
<td>0.0014</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Figure 19: Significant Gene Ontology term enrichment for biological processes associated with differentially expressed genes between Mop1 and mop1-1. Numbers listed below significant GO terms indicate the number of GO terms per total genes in the query (359 genes) and reference (12,192 genes) datasets.
Figure 20: Significant Gene Ontology term enrichment for molecular functions associated with differentially expressed genes between \textit{Mop1} and \textit{mop1-1}. Numbers listed below significant GO terms indicate the number of GO terms per total genes in the query (359 genes) and reference (12,192 genes) datasets.
4.4.4: Localization of siRNA Clusters

Trends in hetsiRNA cluster analysis have thus far pointed to non-additive expression in hybrids with a global downregulation of siRNA cluster expression as compared to inbreds and midparent values. In order to determine the potential effect of hetsiRNA cluster expression on gene expression, co-localization of hetsiRNA meta-clusters and annotated gene regions was accomplished by a three stage siRNA clustering analysis. The first and second stages consist of primary siRNA cluster formation and secondary siRNA meta-cluster formation as previously described in Chapter 3. The third stage of siRNA clustering maps co-localization of siRNA meta-clusters to a +/- 1 kb frame flanking annotated genes. These clusters are termed siRNA/gene meta-clusters. As hetsiRNA and RdDM have been associated with flanking regions of genes, it is hypothesized that hetsiRNA induced methylation of gene flanking regions could result in transcriptional inhibition either by enhanced heterochromatin formation or the establishment and spread of DNA methylation by RdDM in regulatory sequences (Gent et al., 2013). As such, the distribution of primary siRNA clusters within each siRNA/gene meta-cluster was examined by normalizing each annotated siRNA/gene meta-cluster to the mean sequence length and calculating the midpoint of siRNA clusters associated with it, relative to the normalized genic frame which was 5,374 nt (Figure 21). These results show a striking distribution with siRNA cluster distribution localized directly adjacent to genes both on the upstream and downstream flanks, with distinct boundaries between genic and flanking regions and siRNA cluster distribution diminishing gradually with increasing distance from the gene. siRNA clusters co-localizing to flanking regions represented 12% of all siRNA meta-clusters with an equal split between upstream and downstream flanks. While 14% of siRNA clusters co-localized within genic regions, the majority of these were of lower density (low number of
siRNA hits per nt of genic frame) and distributed at the borders of genic regions, similar to flanking clusters. 74% of siRNA clusters were associated with intergenic regions, as was expected due to hetSiRNA association with transposable elements. Excluding clusters associated with intergenic, non-flanking regions and comparing frequencies of siRNA cluster distribution within the normalized frame, one would expect by chance alone that only 27% of clusters map to flanking regions while 73% map to genic regions. These data demonstrate a strong deviation from the frequencies expected by chance as clusters in flanking regions represented 46% of those mapped and those mapping to genic regions only 54%. Even so, those siRNA clusters mapping to genic regions are distributed at the boundaries, suggesting that siRNA producing loci overlap with coding regions to some degree at the borders (i.e. there is a gradual shift in distribution at gene borders rather than an absolute boundary). It should be noted however that the reduction in siRNA frequencies around the +/- 1 kb marks are more likely due to the nature of the clustering process wherein siRNA primary clusters are only associated with a gene if within 1000 nt. As such, it is possible that many primary siRNA clusters that reside at positions less than -1000 nt or greater than +1000 nt were not included in the analysis.
Figure 21: Localization of siRNA meta-clusters within siRNA/gene meta-clusters. Genic frame was standardized as an upstream flanking region of 1,000 nt, followed by a genic region corresponding to the mean of 5,374 nt, and a downstream flanking region of + 1,000 nt.
4.4.5: Correlation of non-Additive Gene Expression and siRNA

Localization of hetsiRNA clusters to the flanking regions of genes has been implicated in the establishment of methylation status as well as setting boundaries between euchromatin and heterochromatin (Li et al., 2015). These flanking regions often contain cis-regulatory elements, and as such any change in methylation or chromatin status may have a direct effect on gene transcription. Furthermore as hetsiRNA are generally downregulated in hybrids (see Chapter 3), trans-chromosomal methylation may reflect a mechanism of regulating non-additive gene expression. To examine this potential effect, RNA-seq expression data was analyzed to determine deviation from expected midparent value and then correlated with hetsiRNA expression that co-localized within siRNA/gene meta-clusters. MPV was calculated as the mean of inbred A and B FPKM for each gene in the B73 RefGen v2. A one-sample t-test with correction for multiple comparison (FDR = 0.05) was used to determine if hybrid AB gene expression differed significantly from expected MPV. Of 7,705 siRNA meta-clusters tested, 1,501 were associated with 1 or more genic regions (siRNA/gene meta-clusters). Only 75 of those genes survived multiple comparison testing and were determined to deviate significantly from MPV (see Appendix). GO term enrichment analysis did not reveal any significant enrichment of biological processes or metabolic functions for these 86 genes.

To determine the putative effect of hetsiRNA co-localization at these non-additive genes, the percent change in Hybrid AB from expected MPV was plotted for both mRNA expression (FPKM) and hetsiRNA expression (RPM) at each siRNA/gene meta-cluster (Figure 22). There was a strong trend toward downregulation in siRNA expression, as expected from the analysis of global siRNA clusters. Surprisingly though, there was also a strong trend toward downregulation
of gene expression at these loci with 89% of loci exhibiting downregulation of both siRNA and mRNA in the hybrid compared to MPV. These data do not support the hypothesis that reduced hetsiRNA production in hybrids leads to reduced methylation in genic flanking regions that in turn leads to increased transcription. Instead these data suggest that reduced hetsiRNA production in localized gene flanking regions may in fact lead to reduced gene expression.

Quantifying expression relative to MPV requires both hetsiRNA and mRNA expression to be detected in both parental inbreds and the hybrid. Previously it was found (Chapter 3) that hetsiRNA sequence diversity (as measured by unique reads) is reduced by 49% in hybrids and thus many hetsiRNA metaclusters are not represented in the previous analysis. This presents the possibility that hetsiRNA and mRNA expression can be influenced by the complete absence of certain hetsiRNAs in the hybrid. In order to examine the localization and potential effects of these hybrid-absent hetsiRNAs, any reads expressed in mop1-l were removed from wildtype datasets to isolate mop1-dependent hetsiRNAs. Trimmed and QC-filtered 24-nt reads from hybrid AB were then used to filter and retain the subset of 24-nt reads from inbreds A and B that had no detectable expression in the hybrid. A limitation of this approach is a severe reduction in depth of sequencing that remains after these filters, even with pooling replicates to increase read coverage. As such, a potential source of error could result from stochasticity and sampling error in a shallow dataset. Indeed, the vast majority of reads in each dataset were only detected once (AA = 89%, BB = 85%, and AB = 93%). To help control for this, sequencing depth was standardized by randomly sampling reads from the datasets with highest overall read counts (AA and BB) such that datasets for all three genotypes had read depth equal to the lowest dataset (AB = 428,388 24-nt reads). Furthermore, only reads detected at least 5 times within a reduced dataset
were used in the analysis. This resulted in unique 24-nt read counts of 4,113, 4,244, and 2,249 for inbreds AA, BB, and the hybrid AB respectively (Figure 23).

To determine if hetsiRNAs absent from the hybrid co-localized with non-additively expressed genes, each reduced 24-nt dataset was aligned to the B73 RefGen v2. And siRNA/gene meta-clusters were determined by 3-stage clustering analysis as previously described. Of the previously identified 75 non-additively expressed genes in RNA-seq data, 13 were associated with hetsiRNA clusters absent in the hybrid (Table 6). Many of these genes were associated with large-scale metabolic processes and networks. Thus, these data suggest that although siRNA are globally downregulated, their direct effect on gene expression by transcriptional repression is likely limited to something on the order of 10% of non-additively expressed genes.
Figure 22: siRNA/gene meta-cluster expression deviation from expected midparent value for 75 loci showing statistically significant deviation from expected midparent value. Percent deviation from MPV was calculated as the difference between hybrid and MPV divided by MPV, thus a negative value indicates downregulation as compared to expected MPV.
Figure 23: Standardization of 24-nt read datasets for hetsiRNAs detected in inbred A or B but absent from hybrid AB. a) Unique raw reads after adapter trimming; b) Unique raw reads after filtering out *mop1* independent reads and random sampling to standardize each dataset to the minimum number of reads; c) Unique raw reads after removal of sequences with < 5 reads; and d) total read counts after removal of sequences with < 5 reads.
Table 6: Annotated genes showing a significant deviation from expected midparent value and detectible siRNA expression in inbred A or B yet a complete lack of siRNA production in hybrids. Direction indicates mRNA expression in hybrid AB relative to MPV.

<table>
<thead>
<tr>
<th>Zea mays gene_id</th>
<th>Dir</th>
<th>GO Molecular Function</th>
<th>GO Biological Process</th>
<th>Ortholog</th>
<th>Species</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G399320</td>
<td>D</td>
<td>GTPase activity, GTP binding</td>
<td>mRNA export from nucleus, protein import into nucleus, ribosome biogenesis</td>
<td>AT1G06720</td>
<td>Arabidopsis thaliana</td>
<td>P-loop containing nucleoside triphosphate hydrolases superfamily protein</td>
</tr>
<tr>
<td>AC217977.3_FG001</td>
<td>D</td>
<td>extracellular matrix structural component</td>
<td>extracellular matrix organization, response to auxin, lateral root morphogenesis</td>
<td>AT3G07390</td>
<td>Arabidopsis thaliana</td>
<td>AIR12 auxin-responsive family protein</td>
</tr>
<tr>
<td>GRMZM2G165597</td>
<td>D</td>
<td>Protein kinase activity, ATP binding</td>
<td>Protein phosphorylation</td>
<td>GSMUA_Achr4 G00140_001</td>
<td>Musa acuminata</td>
<td>receptor-like serine/threonine-protein kinase</td>
</tr>
<tr>
<td>GRMZM5G858454</td>
<td>D</td>
<td>Aconitate hydratase activity, 4 iron, 4 sulfur cluster binding</td>
<td>Metabolic process</td>
<td>PGSC0003DMG 400008740</td>
<td>Solanum tuberosum</td>
<td>Aconitase</td>
</tr>
<tr>
<td>GRMZM2G135337</td>
<td>D</td>
<td>ATP binding, Small protein activating enzyme activity</td>
<td>Cellular protein modification process</td>
<td>OS03G0294900</td>
<td>Oryza sativa</td>
<td>Ubiquitin-activating enzyme E1-3</td>
</tr>
<tr>
<td>GRMZM2G145870</td>
<td>D</td>
<td>Catalytic activity, Indole-3-glycerol-phosphate synthase activity, Copper ion binding</td>
<td>Tryptophan metabolic process, Metabolic process</td>
<td>AT2G04400</td>
<td>Arabidopsis thaliana</td>
<td>Indole-3-glycerol phosphate synthase</td>
</tr>
<tr>
<td>GRMZM2G126545</td>
<td>D</td>
<td>transcription, DNA-templated, regulation of transcription, DNA-templated</td>
<td>methyl-CpG binding</td>
<td>AT5G52230</td>
<td>Arabidopsis thaliana</td>
<td>MBD13 Methyl-CpG-binding domain-containing protein</td>
</tr>
<tr>
<td>GRMZM2G159587</td>
<td>D</td>
<td>Nucleotide binding, Oxidoreductase activity, Cofactor binding, NAD binding</td>
<td>Metabolic process, Oxidation-reduction process</td>
<td>AT1G79870</td>
<td>Arabidopsis thaliana</td>
<td>Glyoxylate/hydroxypyruvate reductase A HPR2</td>
</tr>
<tr>
<td>GRMZM5G834303</td>
<td>D</td>
<td>Transferase activity</td>
<td>Metabolic process</td>
<td>PGSC0003DMG 402027132</td>
<td>Solanum tuberosum</td>
<td>UDP-glucosyltransferase family 1 protein</td>
</tr>
<tr>
<td>GRMZM2G029396</td>
<td>D</td>
<td>farnesyl-diphosphate farnesyltransferase activity, transferase activity, transferring alkyl or aryl (other than methyl) groups</td>
<td>lipid biosynthetic process, biosynthetic process</td>
<td>AT4G34640</td>
<td>Arabidopsis thaliana</td>
<td>SQS1 Squalene synthase</td>
</tr>
<tr>
<td>GRMZM2G056103</td>
<td>D</td>
<td></td>
<td></td>
<td>AT5G13390</td>
<td>Arabidopsis thaliana</td>
<td>NEF1 no exine formation 1</td>
</tr>
<tr>
<td>GRMZM2G099528</td>
<td>U</td>
<td></td>
<td></td>
<td>AT3G17780</td>
<td>Arabidopsis thaliana</td>
<td>B-cell receptor-associated 31-like</td>
</tr>
<tr>
<td>AC198353.5_FG004</td>
<td>D</td>
<td>RNA binding, methyltransferase activity</td>
<td>regulation of transcription, DNA-templated, rRNA processing, tRNA metabolic process, chloroplast organization, protein targeting to chloroplast</td>
<td>AT3G13180</td>
<td>Arabidopsis thaliana</td>
<td>NOL1/NOP2/sun family protein antitermination NusB domain-containing protein</td>
</tr>
</tbody>
</table>
4.5: Discussion

The maize genome is large, highly complex, and rich in transposable elements and repetitive sequences with 32,000 annotated genes and over 1 million transposable elements (Schnable et al., 2009). Of the approximately 2.3 gigabases in the maize genome, it is estimated that about one third of all cytosine residues are methylated (Montero et al., 1992). While DNA methylation is traditionally associated with genome stability and silencing of transposable elements by transcriptional repression and heterochromatin formation, small RNA mediated DNA methylation is enriched around the flanks of genic regions (Gent et al., 2013; Li et al., 2015). It is postulated that these hetsiRNA producing loci serve to regulate boundaries between euchromatin and heterochromatin, with hetsiRNA production preventing euchromatin expansion from spreading beyond genic regions and possibly activating nearby transposable elements. As such, it is conceivable that alterations in hetsiRNA production at genic flanking regions could have an impact on gene expression of neighbouring loci by alterations in methylation status and/or changes in chromatin state.

The data presented here confirm the enriched distribution of hetsiRNAs at genic flanking regions although hetsiRNA clusters were not shown to have absolute boundaries at the ends of genic regions. Rather, a gradual diminishment of hetsiRNA production was observed as overlap with genic sequence increased. This supports the hypothesis of (Gent et al., 2014) who determined that loci associated with RdDM and hetsiRNA production were substantially more accessible to transcriptional machinery than non-RdDM loci. Their results further indicated a positive correlation between hetsiRNA production, CHH methylation, and transcript abundance at flanking genes. The “soft” boundary delineations detected in these data therefore also raise the interesting possibility that alternate isoforms could be controlled in part by RdDM at genes with
multiple promoters by fine tuning heterochromatin formation immediately inside the boundaries of genes where these promoters reside or by regulating alternate exon transcription on the downstream flank. Furthermore, there may exist extraneous regulatory elements in the form of enhancers and repressors that reside in the flanking regions of genes that may also be affected by this hetsiRNA distribution and cause either up- or down-regulation of the adjacent target genes.

Here it was demonstrated across both hybrid and inbred genotypes that inhibition of hetsiRNA biogenesis by null mutation at mop1 results in a net global upregulation of gene expression. Epistatic effects and biochemical interactions resulting from the transcriptional alteration of certain key genes may have cascading effects and certainly mop1 inhibition and elimination of 24-nt hetsiRNA production is sufficient to alter transcriptional networks, with approximately 5-10% of genes being significantly affected. Gene Ontology enrichment analysis revealed the subset of genes upregulated in mop1-1 mutants to be significantly enriched for genes involved in diverse molecular functions. Of particular note is the enrichment of transcription factors in the subset of genes upregulated in mop1-1, suggesting a role for hetsiRNAs in repression of transcription factors and furthering the hypothesis that hetsiRNAs play a role in regulation of a smaller subset of genes that may have cascading pleiotropic effects.

Clustering of hetsiRNAs into siRNA/gene meta-clusters revealed approximately 1,500 annotated genes in the wildtype hybrid that were co-located with hetsiRNA production. Of these, only 75 showed a significant deviation from expected midparent value in their mRNA expression profile. Partly this is due to the stringency required of multiple comparison correction and it is likely that a dataset with greater statistical power would reveal more siRNA/gene meta-clusters that deviate from expected MPV for mRNA expression. Interestingly enough, the majority of non-additively expressed siRNA/gene meta-clusters showed a distinct trend where almost 90%
of clusters were downregulated for both hetsiRNA and mRNA expression as compared to MPV of the wildtype Hybrid AB. This corroborates evidence from both (Gent et al., 2014; Veiseth et al., 2011) and suggests that an increased hetsiRNA in genic flanking regions is associated with an increase in transcript levels. Possible mechanisms by which this could occur include CHH methylation, H3K9 dimethylation, and intermediate heterochromatin formation, resulting in increased accessibility of neighboring genes and transcriptional activation. In a mutant such as mop1-1 where hetsiRNA production is reduced or eliminated, under this hypothesis, one would then expect the opposite: decreased CHH methylation and intermediate heterochromatin formation leading to a decrease in gene transcription. Analysis of mRNA-seq data here however showed that most transcripts are upregulated in a mop1-1 background where hetsiRNA are greatly reduced or absent. It remains a possibility that a small number of loci (as detected by the MPV analysis) are directly affected by hetsiRNA and that large-scale changes in transcript expression are more the result of pleiotropic and down-stream effects.

While the results presented here complement a model in which hetsiRNA upregulate transcriptional activation via altered chromatin state, examples have been noted that suggest the opposite in that hetsiRNA production is associated with transcriptional inactivation (Wei et al., 2014). A key difference may be the presence of MITE transposons in the flanking genic region. As evidenced by increased neighbouring gene expression in the rice OsDCL3a mutant, in this example, reduction of hetsiRNAs caused an increase in gene expression. A similar situation is observed in the case of Vgt-1 where presence of a MITE in a non-coding element causes an increase in methylation and inactivation of the regulatory conserved non-coding region (Castelletti et al., 2014; Salvi et al., 2007). Due to the role of DNA methylation in maintaining epigenetically silenced states at transposons it is possible that these situations are the result of a
tipping balance between hetsiRNAs promoting accessible euchromatin formation in gene flanking regions and the necessity of ensuring genomic stability by repressing transposable elements. hetsiRNAs are able to establish methylation in CH, CHG, and CHH contexts and feedforward loops exist in the maintenance of methylation in the former two symmetric contexts. It is therefore possible that the presence of a MITE in the flanking region causes sufficient establishment of symmetric DNA methylation and feed-forward maintenance that a certain threshold is passed and dense heterochromatin is formed rather than intermediate heterochromatin. Since the former is associated with transcriptional repression and the latter, activation, this would provide a scenario in which hetsiRNA may regulate gene expression in either direction and in a context-specific manner.

A defining characteristic of the hetsiRNA clustering results presented in Chapter 3 is the absence of many unique clusters in the hybrid. As such, one potential limitation to the preceding approach of studying siRNA/gene meta-clusters in the hybrid is that many clusters are not represented in that dataset. Since complete absence of hetsiRNA production would conceivably have a large effect on gene expression, an examination of those clusters expressed in inbreds, but not the hybrid was conducted. The results here indicated that 13 of the previously identified 75 non-additively expressed genes (as defined by statistically significant deviation of hybrid gene expression compared to MPV) were associated with a complete lack of hetsiRNA expression in the hybrid as compared to its inbred parents. Like siRNA/gene metaclusters that showed downregulation of both siRNA and mRNA, these genes were also predominantly downregulated (12 of 13) in the hybrid as compared to MPV (Table 6).

Genes downregulated by the complete loss of siRNA expression in the hybrid were also associated with a wide array of molecular processes and suggest that siRNA regulation of a
limited subset of genes may have broad effects upon genome wide transcriptional networks. Of particular note here is GRMZM2G126545 which encodes a putative methyl-CpG-binding domain protein with homology to *Arabidopsis* MBD13. Recent characterization of MBD proteins has demonstrated their roles in the interpretation of DNA methylation and regulation of chromatin state at methylated loci (Springer and Kaeppler, 2005; Zemach et al., 2013). It has been speculated that gene duplication throughout plant evolution has allowed subfunctionalization and specialization among multiple MBD protein-encoding genes which in turn confers context specific interpretation of DNA methylation and enables a balance between transposable element suppression and gene transcription (Hendrich and Tweedie, 2003). The results presented in Chapter 3 indicate a global downregulation of siRNA production. While the mechanism controlling this downregulation is yet to be elucidated, two plausible scenarios may exist. The first is that changes in siRNA expression in hybrids are mediated by *trans*-chromosomal interactions between complementary or opposing pools of siRNA being generated from each parental genome. The second scenario is that there may be a specific global switch or switches that regulate siRNA and/or gene expression. If one or more methyl-CpG-binding domain proteins are indeed affected by siRNA suppression during hybridization then it could provide an intriguing mechanism whereby a global switch dictating interpretation of genome-wide methylation and heterochromatin formation may further drive changes in siRNA production and gene expression.
Chapter 5: Overall Conclusions and Future Direction

While heterosis likely results from the accumulation of many factors including additive, dominant, and overdominant effects, there also remains an unexplained series of factors influencing heterosis. Increasing evidence from recent advances in sequencing technology and methodology point to a complex regulatory network of involving interplay between various forms of small RNAs, heterochromatin, and gene transcription. The results presented here describe a model in which 24-nt hetsiRNAs are preferentially located in the flanking regions of genes and have a significant effect on global gene expression in inbreds and F1 hybrids. It is unclear whether hetsiRNAs directly affect a wide array of genes or if widespread changes in gene expression are the result of direct effects upon a limited number of genes with pleiotropic and cascading effects.

Furthermore, hetsiRNAs exhibit a global net downregulation and a reduction in diversity resulting from hybridization. Co-localization of hetsiRNA clusters with genes indicated that approximately 1,500 siRNA clusters were associated with genic regions. Most of these genic regions were expressed in an additive fashion (not significantly different from expected midparent value), yet a subset of siRNA/gene meta-clusters were identified where mRNA expression significantly deviated from the expected midparent value. For the vast majority of these siRNA/gene meta-clusters, expression of siRNAs were either completely absent from the hybrid or also downregulated as compared to midparent value. It can be speculated that reduction in hetsiRNA expression at gene flanking regions resulting from hybridization in turn regulate formation of intermediate heterochromatin and serve to regulate gene expression in a context specific manner while preserving genome integrity by silencing transposable elements.
The results presented in this thesis differ somewhat from previously published results in that they did not detect a correlation between differential parental siRNA expression and deviation from expected midparent value in the hybrid. Previous studies (Barber et al., 2012b; Groszmann et al., 2011b) demonstrated that as the magnitude of the difference in hetsiRNA expression between parents increases, hetsiRNA expression in the F1 progeny decreases. Evidence suggests that hetsiRNA expression is highly context dependent and varies with tissue and growth stage.

Future research should focus on a wide array of tissues such as the shoot apical meristem, the developing ear, or perhaps whole seedlings. This would not only provide a more comprehensive analysis of hetsiRNA and their effects on gene expression across tissues, but it would make the results more directly comparable to the published literature. The effects of material siRNA expression in female gametophytes translating into heritable epigenetic changes in mature tissues and subsequent generations also remains an interesting avenue for research. Similar analyses could be conducted on reciprocal hybrids to determine if there are differing results as a consequence of the inheritance of a maternal pool of siRNA. Of course, future research could focus on validation of specific siRNA/gene meta-clusters identified in this study in order to present causative rather than associative evidence of hetsiRNA control of gene expression at these loci. In order to validate and expand on bioinformatic approaches, quantitative RT-PCR could be done on a subset of interesting gene candidates. Of particular interest would be the subset of 75 genes that were non-additively expressed as compared to expected MPV in hetsiRNA/gene meta-clusters. Also, the 13 genes that showed differential expression and were co-located with hetsiRNAs not expressed in the hybrid would be interesting candidates for further inspection. Finally, bisulphite treated methyl-sequencing datasets were
also generated from DNA samples taken from lines used in this study. Future analysis of these
datasets should provide insight into methylation status at these individual loci could provide
further evidence of the model presented herein.
Chapter 6: References


Shull, G.H. (1908). The composition of a field of maize. J. Hered. 296–301.


