2010

Analysis of differences in gene expression and the genetic and epigenetic regulation of transcript accumulation in maize

Yi Jia
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd

Part of the Cell and Developmental Biology Commons, and the Genetics and Genomics Commons

Recommended Citation
http://lib.dr.iastate.edu/etd/11302

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Analysis of differences in gene expression and the genetic and epigenetic regulation of transcript accumulation in maize

by

Yi Jia

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Plant Biology

Program of Study Committee:
Patrick S. Schnable, Major Professor
Hui-Hsien Chou
Dan Nettleton
Thomas A Peterson
Steven A Whitham

Iowa State University
Ames, Iowa

2010

Copyright © Yi Jia, 2010. All rights reserved.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ABSTRACT</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER 1. GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Dissertation organization</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>References</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 2. ALL POSSIBLE MODES OF GENE ACTION ARE OBSERVED IN A GLOBAL COMPARISON OF GENE EXPRESSION IN A MAIZE F₁ HYBRID AND ITS INBRED PARENTS</td>
<td>8</td>
</tr>
<tr>
<td>Abstract</td>
<td>8</td>
</tr>
<tr>
<td>Introduction</td>
<td>8</td>
</tr>
<tr>
<td>Results</td>
<td>9</td>
</tr>
<tr>
<td>Discussion</td>
<td>12</td>
</tr>
<tr>
<td>Materials and Method</td>
<td>13</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>16</td>
</tr>
<tr>
<td>References</td>
<td>16</td>
</tr>
<tr>
<td>CHAPTER 3. LOSS OF RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) FUNCTION CAUSES WIDESPREAD AND UNEXPECTED CHANGES IN THE EXPRESSION OF TRANSPONSES, GENES, AND 24-NT SMALL RNAS</td>
<td>29</td>
</tr>
<tr>
<td>Abstract</td>
<td>29</td>
</tr>
<tr>
<td>Author Summary</td>
<td>29</td>
</tr>
<tr>
<td>Introduction</td>
<td>30</td>
</tr>
<tr>
<td>Results</td>
<td>31</td>
</tr>
<tr>
<td>Discussion</td>
<td>35</td>
</tr>
<tr>
<td>Materials and Method</td>
<td>39</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>41</td>
</tr>
<tr>
<td>References</td>
<td>41</td>
</tr>
<tr>
<td>CHAPTER 4. NATURAL ANTISENSE TRANSCRIPTS CO-LOCALIZE WITH OPEN CHROMATIN MARKS AND SMALL RNAS AND ARE REGULATED BY A COMPLEX GENETIC ARCHITECTURE IN MAIZE</td>
<td>51</td>
</tr>
<tr>
<td>Abstract</td>
<td>51</td>
</tr>
<tr>
<td>Introduction</td>
<td>51</td>
</tr>
<tr>
<td>Results</td>
<td>53</td>
</tr>
<tr>
<td>Discussion</td>
<td>58</td>
</tr>
<tr>
<td>Materials and Method</td>
<td>61</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>65</td>
</tr>
<tr>
<td>References</td>
<td>65</td>
</tr>
<tr>
<td>CHAPTER 5. CELL-LAYER SPECIFIC GENE EXPRESSION IN MAIZE SHOOT APICAL MERISTEMS</td>
<td>76</td>
</tr>
<tr>
<td>Abstract</td>
<td>76</td>
</tr>
<tr>
<td>Introduction</td>
<td>76</td>
</tr>
<tr>
<td>Results</td>
<td>78</td>
</tr>
<tr>
<td>Discussion</td>
<td>82</td>
</tr>
<tr>
<td>Materials and Method</td>
<td>84</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>87</td>
</tr>
<tr>
<td>References</td>
<td>87</td>
</tr>
</tbody>
</table>
ABSTRACT

Much of phenotypic variation is the result of alteration in gene expression patterns. Maize exhibits extremely high levels of diversity in DNA sequences, gene expression and phenotypes. Transcript profiling studies were conducted to explore the relationship between genotype and gene expression patterns. Consistent with prior reports these studies demonstrated that gene expression can be controlled not only by genetic but also epigenetic determinants. Importantly, the identified widespread natural antisense transcripts (NATs) and hundreds of the NAT-eQTL regulating the expression of NATs further increase the complexity of gene expression. These observations and conclusions enhance our understanding of various fundamental biology questions. For example, heterosis is the phenomenon that the progeny of particular inbred lines exhibit superior performance as compared to both parents. We have hypothesized that such extreme phenotypic variations between inbred and hybrid are explained by multiple factors, including epigenetic factors such as those studied here.
CHAPTER 1. GENERAL INTRODUCTION

DISSERTATION ORGANIZATION

This dissertation includes the general introduction (Chapter 1), four journal papers (Chapters 2-5), and general conclusions (Chapter 6).

The paper in Chapter 2, which was published in the Proceedings of the National Academy of Sciences [4], describes a global gene expression study between maize inbreds and their hybrid. I contributed to the experimental design, performance of the experiments and analysis of the data. The co-first author Ruth A. Swanson-Wagner’s contributions included experimental design, performance of the experiment, and analysis of the data, as well as drafting the paper with Dr. Schnable.

The paper in Chapter 3, which was published in PLoS Genetics [38], describes a transcriptomic comparison between an epigenetic mutant and non-mutant controls using RNA-seq technology. I contributed to the experimental design, performed the experiment, analyzed the data and wrote the paper with Dr. Schnable.

The manuscript in Chapter 4 is to be submitted to Genome Research. In this manuscript, a genome-wide identification for natural antisense transcripts (NATs) was conducted using bioinformatics and RNA-seq approaches. In addition, a NAT-eQTL experiment was conducted and analyzed to illuminate the regulatory network responsible for the expression of NATs. I designed the experiment, conducted the experiments, analyzed the data and wrote the manuscript with Dr. Schnable.

The manuscript in Chapter 5 is to be submitted to PLoS Genetics. This manuscript describes the analysis of cell-layer specific gene expression within the maize shoot apical meristem. In this study, I conducted all the data analysis and wrote the manuscript with Dr. Schnable.

INTRODUCTION

Maize is a model crop species

Maize is an important economic crop and serves as a source of food, feed and energy. It is widely grown throughout the world and provides the highest yields among all grain crops. The US produces more than 40% of world’s maize grain. In the US alone, 2 billion bushels of maize grain with a value of $47 billion was produced in 2008 [1]. In addition to its role as a food and feed crop, in recent years maize has been used as a feedstock for ethanol production.

Maize exhibits extensive diversity in DNA sequences, gene expression patterns and phenotypes [1,2,3,4,5]. In addition to its economic importance, maize is an ideal biological system for various fundamental research topics, such as heterosis, epigenetic regulation, recombination, functional genomics, plant development and evolution.
Heterosis
Heterosis (or hybrid vigor) is the phenomenon that the offspring of two selected parents exhibit greater fitness than the parents [6,7]. This phenomenon was first described by Charles Darwin [8]. In the early 20th century, George H. Shull and M. East demonstrated heterosis in maize and laid the fundamental basis for applying heterosis to agricultural production [9,10]. Maize yields have been increased eight-fold over the last century, and 15%-60% of increase is attributed to heterosis [1].

The molecular mechanisms responsible for heterosis are not known although plant breeders and seed companies have exploited heterosis for almost one century. Two classic hypotheses were developed to explain heterosis: dominance and overdominance [6]. The dominance hypothesis states the deleterious alleles in the hybrid from one parent are complemented by superior alleles from the opposite parent of the hybrid. The overdominance hypothesis states that the interactions between the two alleles in heterozygotes results in superior performance relative to either of the homozygous parents.

Genetic architecture controlling gene expression
Regulation of transcription
In eukaryotic cells, DNA is assembled and packed into chromatin with histone proteins, which can form nucleosomes. Under specific process of development or different environments, different genes need to be turned on or off. During this process, the histone proteins and DNA can be modified or unmodified to change the chromatin to different status for either activating or inactivating gene expression [11]. Such regulatory events are controlled not just by a single regulator, but also by combinatorial mechanisms involving multiple steps that take place during transcription. Overall, the process of transcription can be divided into five different stages: 1) Pre-initiation [12]. The preinitiation complex is required for the initiation of transcription. At the beginning of this stage, the transcription factor II D (TFIID) binds to the TATA box via the TATA binding protein (TBP), five other general transcriptional factors (TFIIA, TFII B, TFIIE, TFIIF and TFI IH) and RNApolymerase bind around the TATA box. Once the preinitiation complex is formed, DNA helicase can separate the opposing strands of double stranded DNA to provide the single-stranded DNA for transcriptional template. 2) Initiation [13]. After the preinitiation complex is formed, a serial of transcriptional factors bind to the promoter. Then RNA polymerase can recognize and bind to the core promoter. This combination of various transcriptional factors and RNA polymerase is called the transcription initiation complex. 3) Promoter clearance [14]. The RNA polymerase clears the promoter after the first bond is synthesized. 4) Elongation [15]. The non-coding strand of DNA is used as template for RNA synthesis in the 5' → 3' orientation. A single template can be used for multiple rounds to produce multiple copies of RNAs. 5) Termination [16]. This is the final step of transcription when the newly synthesized RNA molecule detaches from the DNA template.
Approaches for measuring transcriptional levels

Transcription can be measured via a variety of different methods which are suitable for different experimental goals. The following is a non-exhaustive list of experimental approaches that can be used to measure transcriptional levels: 1) Nuclear run-on assays for measuring the transcription rate and mRNA stability via radioactivity [17]; 2) RNase protection assays for mapping transcribed regions [18]; 3) In situ hybridization experiments for localizing specific DNA or RNA sequence [19]; 4) Northern blots for studying the expression of genes or small RNAs; 5) RT-PCR for detecting transcript; 6) Microarray hybridization using labeled cDNAs or RNAs for measuring transcript accumulation in a high-throughput manner [20]; 7) High-throughput RNA (or cDNA) sequencing as a open platform for large-scale measurement of transcript accumulation [21].

Identification of regulatory elements for transcription

Genes can be turned on or expressed at different levels at different developmental stages or different tissues. In fact some genes are only turned on following specific stimuli or under specific stress conditions. Such specific regulation usually is controlled by sequences existing in the promoter called cis-acting elements. Proteins that bind to these cis-elements to thereby regulate the transcription of target genes are termed trans-acting regulatory factors [22].

Multiple approaches can be used to identify cis- elements or trans-factors that regulate transcription. The following is a non-exhaustive list: 1) Bioinformatics searches for motifs that are conserved across co-expressed genes or different organisms [23]; 2) Promoter deletion analyses for mapping specific functional segments [24]; 3) Chromatin immunoprecipitation (CHIP) chip or CHIP-seq to identify protein binding sites genome wide [25]; 4) Expression QTL (eQTL) analyses to simultaneously identify cis- and trans-genome intervals [26]; 5) Association mapping experiments in which transcript level are treated as traits [27].

Epigenetic landscape and its regulation of gene expression in plants

It has become clear that the genetic variation alone cannot explain all of phenotypic variation, such as cell type differentiation because the genetic information in most cells within one individual organism are essentially the same. Instead, such differential phenotypes are usually regulated by differential status of chromatin, such as open or closed by differential covalently modification of DNA and histone proteins [28]. As such the status of chromatin can be inherited through multiple generations of cell division [29]. This type of regulation that occurs in the absence of changes in DNA sequences is called epigenetic regulation. Many types of chromatin modifications have been identified for DNA and histones, such as methylation of DNA and histone acetylation, ubiquitylation, phosphorylation and sumoylation. [28]. These different types of modifications are classified as epigenetic marks, which are distributed on the chromosome in specific patterns related to genomic features, for example, non-genic sequences. Different
epigenetic marks are often observed to differentially affect different aspects of gene expression [28].

**Methods for epigenome profiling**

Research on global epigenome profiling has been conducted mainly from two aspects in recent years: DNA methylation and histone modification. In practice, each of epigenome profiling consists of two steps: one is the pretreatment of DNA segments associated with modified DNA or histone followed by quantification of the pretreated DNA samples [30]. For the pretreatment of DNA sequences associated with histone modifications, DNA segments associated with specific modified histone proteins are pull down via chromatin immunoprecipitation (ChIP) procedure using antibodies specific to the histone modifications of interest. More pretreatment options are available for DNA methylation analysis, including digestion with methylation sensitive enzymes, affinity enrichment via methylated DNA immunoprecipitation and sodium bisulphate conversion of unmethylated cytosines to uracil. The quantification step can be conducted by hybridizing pretreated and treated DNA onto genome tilling array or direct sequencing.

**Landscape of epigenetic modification in plants**

Several genome-wide epigenome profiling experiments have been conducted using different technologies with different resolution levels on the model plant Arabidopsis [21,31-34]. Most of our knowledge about the distribution of epigenetic modification on the genome or genic levels has been learned from this model. Rabinowicz reported that unlike the case in mammals, methylation levels between genes and transposons are extremely different using PCR-based techniques [35]. So it is not surprising to observe much higher methylation levels in the paracentromic regions of plant chromosomes that has been revealed by the use of next generation sequencing technologies. However, genome-wide methylation profiling experiments have demonstrated that not all genes are hypomethylated; indeed, about one-third of transcribed gene regions are hyper-methylated [33]. However, the function of such methylation in genic regions is not yet clear. Some hypotheses about the function of gene body methylation are potentially worth further testing. For example, Zilberman et al [34] hypothesized that gene body methylation may contribute to the inhibition of cryptic transcription initiation. Alternatively, Zhu [36] proposed that recombination or transposon insertion within genes can be largely excluded by methylation in the gene body.

Wang et al [37] recently reported on the integrated maize epigenome and thereby provided many valuable data sets, including DNA methylation, histone modifications, small RNA and mRNAs. This study confirmed most of the epigenome distribution patterns on the chromosome level, that were expected based on prior studies, such as the dense methylation in the repetitive regions and open chromatin markers (e.g. H3K4me3, H3K9ac and H3k36me3) associated with highly transcribed genes and closed chromatin markers (e.g. H3K27me3 and DNA methylation) located in inactive genes or regions. This study also offered an opportunity to compare the epigenome between different plant organisms. One interesting difference is the distribution of methylation within genes. In Arabidopsis, more than one third of genes
show high methylation levels in the transcribed regions. But in maize and similar in rice, only one high methylation peak was observed around the ATG regions. This difference in the methylation distribution within genes could be due to the significant different genome architecture between Arabidopsis and maize (or rice).

REFERENCES


CHAPTER 2. ALL POSSIBLE MODES OF GENE ACTION ARE OBSERVED IN A GLOBAL COMPARISON OF GENE EXPRESSION IN A MAIZE F₁ HYBRID AND ITS INBRED PARENTS

A paper published in The Proceedings of the National Academy of Sciences

Ruth A. Swanson-Wagner², Yi Jia², Rhonda DeCook, Lisa A. Borsuk, Dan Nettleton, and Patrick S. Schnable³

ABSTRACT

Heterosis is the phenomenon whereby the progeny of particular inbred lines have enhanced agronomic performance relative to both parents. Although several hypotheses have been proposed to explain this fundamental biological phenomenon, the responsible molecular mechanisms have not been determined. The maize inbred lines B73 and Mo17 produce a heterotic F₁ hybrid. Global patterns of gene expression were compared in seedlings of these three genotypes by using a microarray that contains 13,999 cDNAs. Using an estimated 15% false discovery rate as a cutoff, 1,367 ESTs (9.8%) were identified as being significantly differentially expressed among genotypes. All possible modes of gene action were observed, including additivity, high- and low-parent dominance, underdominance, and overdominance. The largest proportion of the ESTs (78%; 1,062 of 1,367) exhibited expression patterns that are not statistically distinguishable from additivity. Even so, 22% of the differentially regulated ESTs exhibited nonadditive modes of gene expression. Classified on the basis of significant pairwise comparisons of genotype means, 181 of these 305 nonadditive ESTs exhibited high-parent dominance and 23 exhibited low-parent dominance. In addition, 44 ESTs exhibited underdominance or overdominance. These findings are consistent with the hypothesis that multiple molecular mechanisms, including overdominance, contribute to heterosis.

INTRODUCTION

The hybrid progeny of selected pairs of inbred lines exhibit enhanced agronomic performance relative to both parents (1), a phenomenon that is termed heterosis or hybrid vigor. Heterosis is widely exploited in applied plant breeding. For example, ≈95% of U.S. maize acreage is planted to hybrids. Duvick (2) estimates that maize hybrids exhibit a 15% yield advantage relative to superior open-pollinated varieties and that worldwide heterosis accounts for an additional 55 million metric tons of grain yield annually. Despite the fact that heterosis has been widely exploited by plant breeders to the benefit of agriculture and society, the molecular mechanisms responsible for this basic biological phenomenon are not well understood.

Multiple models have been proposed to explain heterosis (3). The two most commonly invoked are the dominance (or complementation) hypothesis and the overdominance hypothesis. The first hypothesis (4, 5)

¹ Reprinted with permission of Proc. Natl. Acad. Sci. USA;
² Ruth A. Swanson-Wagner and Yi Jia contributed equally to this work;
³ To whom correspondence should be addressed (schnable@iastate.edu).
states that deleterious alleles at different loci in the two homozygous parental genomes are complemented in the heterozygous F1 hybrid. More recently, the special case that complementation of genes that differ in their presence and absence among maize lines may contribute to heterosis has been proposed (6). Complementation cannot by itself, however, explain heterosis because although the per se performance of inbred lines can be improved by purging them of detrimental alleles, doing so has little impact on heterosis (3). Additional evidence for this view comes from the findings that progressively more heterosis occurs in polyploids as the diversity of the component genomes increases and inbreeding depression in autotetraploids increases faster than homozygosity.

The overdominance hypothesis (1, 5, 7) states that the improved performance of an F1 hybrid relative to its inbred parents is a consequence of favorable allelic interactions at heterozygous loci that outperform either homozygous state. Although these classical hypotheses have provided guidance for experimentation (8–11), it is likely that heterosis depends on multiple mechanisms, including epigenetic phenomena. It is also possible that differential accumulation of allele-specific transcripts in hybrids may contribute to heterosis (12).

It has been hypothesized that differential gene expression in inbreds and hybrids may be responsible for heterosis (13, 14). For example, a hybrid could accumulate levels of transcript equal to the mid-parent (additivity), the high or low parent (high or low parent dominance), above the high parent (overdominance), or below the low parent (underdominance). Prior studies of gene expression in inbreds and their F1 hybrids have focused on relatively few genes.

Here, we apply global transcript profiling technology to examine the expression of thousands of genes in two inbred parents and their F1 hybrid to begin to understand the underlying mechanisms and complex regulatory network surrounding heterosis.

More than 1,300 ESTs exhibited significant differential expression patterns among the three genotypes at an estimated false discovery rate (FDR) of 15%. The most common mode of action was additivity, but several hundred genes exhibited high- or low-parent dominant, overdominant, or underdominant modes of gene action. The expression patterns of >90% of sampled genes were validated by using quantitative real-time PCR (qRT-PCR). The finding that all modes of gene action can be detected in inbreds and their F1 hybrid is consistent with the hypothesis that multiple molecular mechanisms, including overdominance, contribute to heterosis.

RESULTS

The maize F1 hybrid generated by crossing the inbred lines B73 and Mo17 is taller, matures more quickly, and produces higher grain yields than both parents (15). We elected to analyze global patterns of gene
expression in these three genotypes because this hybrid and its relatives are widely grown in the Corn Belt (16) and the genetic map of maize is based on recombinant inbreds developed from this hybrid.

Because heterosis affects most aspects of plant growth and development, one of the challenges in designing such an experiment is deciding which tissue to analyze. In making this decision, we sought a system in which we could tightly control environmental variability and that, therefore, would provide the statistical power to detect even subtle changes in gene expression that nevertheless may be biologically relevant. We elected to analyze seedlings because seedling dry weight exhibits a substantial degree of heterosis (Table 1), and seedlings can be grown under controlled conditions (see Methods). Although the B73 × Mo17 hybrid is used commercially, the Mo17 × B73 hybrid exhibits a greater degree of heterosis for seedling dry weight (Table 1) and, therefore, was selected for the profiling experiments. Above-ground seedling tissue was harvested from the three genotypes 14 days after planting. RNA extracted from these seedlings was reverse transcribed, labeled with fluorescent dyes, and hybridized to a cDNA array that contains 13,999 informative spots (see Methods). Nine biological replications were analyzed to provide a higher degree of statistical power.

**Statistical Analysis of Microarray Hybridization Data**

Data normalization and transformation were performed to reduce nonbiological variation, make signal intensities comparable across arrays, and achieve approximate normality and constant variance for statistical modeling (see Methods). A mixed model analysis of the data revealed genes having significant differences in gene expression levels in at least one of the three genotypes. The distribution of \( P \) values generated from the tests for genotype (Fig. 1) was used to detect genes with significant differential expression across genotypes. Multiple significance thresholds were investigated, providing significance lists of lengths 280 (1% FDR), 460 (3%), 990 (10%), and 1,367 (15%) genes. The estimated 15% FDR level was chosen to sample a large pool of significant genes for further analyses. At this threshold, 9.8% (1,367 of 13,999) of informative cDNAs on the microarray were differentially expressed among the genotypes. As a group, these genes are involved in a wide variety of cellular processes. The statistical power of this experiment made it possible to detect even small changes in gene expression. For example, the significant fold changes between the low- and high-expressing genotypes for the 1,367 differentially expressed ESTs ranged from 1.2 to 88 (Fig. 3).

**Analysis of Gene Action**

The 1,367 ESTs identified as differentially expressed at the 15% FDR level were further investigated to determine their modes of gene action (Table 4, which is published as supporting information on the PNAS web site). The estimated genotype means from the mixed model analyses were used to determine the expression pattern for each EST. To visualize the patterns of gene expression, a 2D polar coordinate plot of the 3D line mean patterns was implemented (Fig. 2). The radius at which a gene is plotted represents the
log₂ fold change between the highest and lowest expression levels among the three genotypes. The angle at which a gene is plotted represents the relationships among the means of the three genotypes. A gene plotted on the horizontal line exhibits pure additivity. A gene plotted on the vertical line in the top or bottom half of the figure exhibits overdominance or underdominance, respectively. A gene plotted on a solid line exhibits the same expression level in two genotypes, and a different expression level in a third genotype (see plot labels for direction of differential expression). A gene plotted between lines exhibits a mode of gene action that is intermediate to that indicated by the two nearest lines.

A large proportion of the ESTs (78%, 1,062 of 1,367) exhibited expression patterns that are not statistically distinguishable from additivity. The 305 ESTs that exhibited nonadditive mode of gene expression were further classified based on significant pairwise comparisons ($P < 0.05$). Of these 305 ESTs, 181 exhibited high-parent dominance and 23 exhibited low-parent dominance. Twelve ESTs exhibited modes of gene action intermediate between additivity and dominance. In addition, 34 ESTs exhibited clear overdominance, and 10 exhibited clear underdominance. Although it was possible to conclude that the remaining 45 ESTs exhibited nonadditive gene action, there was insufficient statistical power to assign them to one of the above categories. Genes that exhibited overdominance and underdominance exhibited near parent fold changes of 1.2–2.0 and 1.3–4.7, respectively, and are involved in multiple cellular processes (Table 2).

**Validation of Modes of Gene Action via qRT-PCR**

A sample of 45 genes identified in the mixed model analysis as having significant differences in gene expression across genotypes was selected for validation via qRT-PCR (Table 3). Genes that exhibited a variety of modes of gene action (i.e., from all 12 sectors of Fig. 2) were chosen for validation. Selected genes exhibited changes in gene expression from the low- to high-expressing genotype of 1.4- to 88-fold.

Primers were designed to specifically amplify each of the 45 genes (see Methods). These primers were used to conduct qRT-PCR on seven biological replications of RNA from the three genotypes. Using threshold cycle ($C_t$) values generated from qRT-PCR experiments (see Methods), the null hypothesis of equal expression across the B73, Mo17, and F₁ genotypes was tested. For 10 genes, there was not sufficient statistical power in the qRT-PCR experiments to determine a mode of gene action (Table 5, which is published as supporting information on the PNAS web site). There was sufficient statistical power to make conclusions regarding modes of gene action for the remaining 35 genes. For 15 of these genes, the modes of gene action detected in the qRT-PCR experiments were indistinguishable from the modes based on the microarray experiment ($P < 0.05$). For 17 of the genes, the mode of gene action obtained via the qRT-PCR experiments were at least consistent with the modes obtained from the microarray experiment. Hence, the qRT-PCR experiments either validated or were consistent with the modes of gene action exhibited by 91% (32 of 35) of the genes in the microarray experiment.
DISCUSSION

Despite its critical importance to agriculture, a mechanistic understanding of heterosis has not been achieved. As a step toward generating data needed to test existing hypotheses, prior studies have analyzed modes of gene action in small sets of maize genes (13, 17). Even so, a global understanding of the behavior of gene expression in inbreds and their F₁ hybrids is lacking. The current study used microarray technology to characterize the modes of gene action for 13,999 cDNAs.

Approximately 9.8% (1,367 of 13,999) of the ESTs assayed in this experiment exhibited differential expression among the three genotypes. The majority of these ESTs (n = 1,062) exhibited modes of gene action that could not be distinguished from additivity. The expression of these genes could be controlled by cis-acting regulatory elements and/or dosage-dependent trans-acting factors. The large number of genes that exhibited additive gene action is consistent with the complementation hypothesis of heterosis.

Approximately 22% (305 of 1,367) of the differentially regulated ESTs detected in this study exhibited nonadditive modes of gene expression. Most of these genes exhibited high-parent dominance (n = 181). However, low-parent dominance (n = 23), underdominance (n = 10), and overdominance (n = 34) also were observed. It was possible to validate via qRT-PCR the modes of gene action exhibited by 91% (32 of 35) of a sample of the differentially expressed genes.

Overall, 2.2% (305 of 13,999) of the ESTs analyzed in this survey of nearly 14,000 cDNAs exhibited nonadditive modes of gene action. These results differ substantially from a prior study of a smaller set of genes. Auger et al. (17) reported that 19 of the 30 genes (63%) exhibited nonadditive gene action. Although these two studies were conducted by using the same genotypes, Auger et al. (17) used gel blots to analyze RNA extracted from leaves of adult field-grown plants, whereas we used microarrays to analyze gene expression in seedlings grown under highly controlled environmental conditions. One explanation for the different rates of nonadditive gene expression observed in the two studies is differential sampling of the maize gene space. The 30 genes studied by Auger et al. (17) may be a less random sample than the 13,999 cDNAs present on our microarray. Alternatively (or in addition), the percentage of genes that exhibit nonadditive gene action may differ during development even though both stages of development analyzed in these studies exhibit heterosis. The nonadditive expression of these genes could be explained by dominant allelic and nonallelic epistatic control of transcript accumulation.

The existence of overdominant gene action has important implications for evolutionary theory, in particular the maintenance of genetic variability. The evidence for overdominant gene action however, has been limited so far. Some, but not all, experiments conducted by using Drosophila (18), and more recently by using C. elegans, (19) have uncovered evidence for overdominance. Although the results of the current study in maize are consistent with prevailing views that most loci exhibit additive, or less frequently, dominant gene action, the identification of 34 ESTs that exhibited overdominance suggests that
hypothesized genetic processes, including heterosis, that invoke overdominance cannot be excluded from consideration. Although it is not possible to exclude the possibility that the overdominant gene action observed in this study is the result of “pseudooverdominance,” caused by the combined action of linked loci, such blocks of genes would have similar effects on genetic processes as overdominant loci.

Analysis of the microarray experiment resulted in the identification of 44 ESTs that exhibit overdominance or underdominance. qRT-PCR experiments validated these modes of gene action for five of eight tested genes. The existence of genes that exhibit overdominant or underdominant modes of gene action in B73, Mo17, and their F₁ hybrid is at least consistent with the overdominance hypothesis of heterosis. Although heterosis is controlled by many genes, only a small fraction of all genes are involved (3). Hence, it is at least possible that some of these genes may contribute to heterosis. Consistent with this hypothesis, genes that exhibit overdominance include those that potentially affect a wide variety of regulatory steps, including splicing, translation, protein folding, modification, and degradation (Table 2); others are involved in stress response. All of these functions could contribute to posttranscriptional regulatory cascades contributing to heterosis.

Among other mechanisms, one attractive hypothesis for the existence of underdominant and overdominant gene action invokes the action of small interfering RNAs (siRNAs). siRNAs are typically derived from transposons and repeats, although some genes and other sequences can generate siRNAs (20). siRNAs can regulate gene expression by cleaving target mRNAs (21) and via transcriptional silencing (22). Maize inbreds differ radically in transposon and repeat content (6, 23, 24). In addition, in this study, at least two transposons exhibited >4-fold differences in expression between B73 and Mo17 (Table 4). Hence, inbreds are likely to differ in their complement of siRNAs. If siRNAs from one inbred do not match genes (e.g., repetitive sequences in 3’ UTRs) from the other inbred, the resulting hybrid could exhibit novel patterns of gene expression, including overdominance or underdominance. Consistent with this hypothesis, we observe profound differences in the accumulation of antisense RNAs in B73 and Mo17 (Y.J., R.A.S.-W., S. J. Emrich, R.D., L. Guo, Y. Fu, D. A. Ashlock, D.N., and P.S.S., unpublished data). Overall, our results are consistent with the hypothesis that multiple molecular mechanisms contribute to heterosis.

**METHODS**

**Genetic Stocks and Experimental Design**

The inbreds B73 (Schnable laboratory accession no. 660) and Mo17 (Schnable laboratory accession no. 2618) used in this study were derived by self-pollination from stocks originally obtained from Donald Robertson and Mike Lee (Iowa State University), respectively. Mo17 was crossed as a female by B73 to generate the F₁. Kernels from three different seed sources (ears) per genotype were used in the experimental design. Individual genotypes within a replication, however, were all derived from the same source. Before microarray analyses, genotypes were confirmed by using codominant IDP genetic markers that distinguish B73 from Mo17 (Y. Fu, T. J. Wen, Y. I. Ronin, D. I. Mester, Y. Yang, M. Lee, A. B.
Korol, D. A. Ashlock, and P.S.S., unpublished data). Ten biological replications of B73, Mo17, and their F_1 (Mo17 × B73) were grown under highly controlled conditions in a randomized complete block design. For each replication, the B73, Mo17, and F_1 samples were hybridized to three two-color cDNA microarrays by using a loop design such that each loop included all pairwise comparisons between genotypes. RNA pools for each genotype were alternately labeled providing dye balance within each loop. After hybridization, one biological replicate was removed because of poor quality. The final analysis incorporated 27 microarray slides (three slides for each of nine high-quality biological replicates).

**Plant Growth and RNA Isolation**

Kernels were planted in SB 300 Universal soil (Sun Gro Horticulture, Bellevue, WA) within a PGW-40 (Percival Scientific, Perry, IA) growth chamber that provided 15 h of light (25°C) and 9 h of dark (20°C). Light intensity was \(650–800 \mu \text{mol·m}^{-2}·\text{s}^{-1}\). Seedlings were watered as needed by using a 0.7 M calcium nitrate solution. Fourteen days after planting, six random healthy plants were harvested as a pool for each genotype-replication. All aboveground tissue was separated from root tissue and immediately submerged in liquid nitrogen. After separately grinding each genotype-replication pool in liquid nitrogen, RNA was extracted from \(10\) g of frozen tissue by using TRIzol reagent (Invitrogen) as per the manufacturer’s instructions, with slight modifications. RNA integrity was confirmed via gel electrophoresis. The OligoTex mRNA midi kit (Qiagen, Valencia, CA) was used to extract mRNA from 500 \(\mu\)g of RNA by using the manufacturer’s protocol, with slight modifications. mRNA yields were typically between 0.75 and 1.5% of the starting RNA.

**Microarray Printing**

The SAM1.1 cDNA array was printed on the UltraGAPs slide (Corning, Corning, NY) by using a PixSYS 5500 Arrayer (Cartesian Technologies, Irvine, CA). The GEO platform file for this chip is posted at www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc = GPL2613.

**Reverse Transcription, Labeling, and Microarray Hybridization**

Two \(\mu\)g of mRNA were labeled according to Nakazono et al. (25), with slight modifications. Specific cDNA samples were labeled with Cy3 and Cy5 fluorescent dyes in accordance with the experimental design.

**Microarray Data Acquisition, Normalization, and Analysis**

Replications 1–5 were scanned by using a ScanArray 5000 (Packard). Replications 6–9 were scanned by using a PerkinElmer Pro Scan Array HT. A minimum of six scans for each dye channel were completed at increasing photomultiplier tube gain and laser power settings. Only one of these scans was selected for analysis for each channel per slide. A set of scans was selected that had similar natural logged median intensity values for the Cy3 and Cy5 channels of each individual slide and across all slides.
The lowess normalization method was applied to the log of background-corrected raw signal intensities to remove signal-intensity-dependent dye effects on each slide (26, 27). Normalization was conducted separately for each slide to avoid introducing dependencies among biological replications. After lowess normalization, the normalized data for each slide/dye combination were median-centered so that expression measures would be comparable across slides. Median centering involves subtracting the median value for a particular slide/dye combination from each individual value associated with the particular slide/dye combination. Thus, negative (positive) values indicate that a particular transcript was expressed below (above) the median for a particular slide/dye combination.

The SAM1.1 maize cDNA array chip contains 19,200 spots. Before statistical analysis 897 “empty” and “bad-PCR” spots were removed from the data set. For each of the remaining 18,303 spots on the microarray, a mixed linear model analysis (28) of the normalized, log-scale signal intensities was conducted to identify transcripts whose expression differed significantly among genotypes. The mixed linear model included fixed effects of genotype and dye and random effects related to the experimental design. The \( P \) values generated from the tests for line effect (testing for equality of the three genotype means) were used to determine significance of differential expression. The estimated means from the mixed model for each genotype were used to identify the mode of gene action for all significant genes. Each gene was classified into a significant pattern category by using pairwise comparison tests (\( P < 0.05 \)). \( P \) values from the linear-in-genotype contrasts (testing for \( F_1 \) genotype mean equal to the average of the two parental line means) from the mixed models were used to classify significant genes into the categories of not distinguishable from additivity (\( P > 0.05 \)) and distinguishable from additivity (\( P < 0.05 \)). Genes in the latter group were further classified into more specific nonadditive categories by using the aforementioned significant patterns. Genes with an \( F_1 \) genotype mean not significantly different from one parent and significantly larger (smaller) than the other parent were said to exhibit high-parent (low-parent) dominance. Genes with an \( F_1 \) genotype mean that was significantly larger (smaller) than both B73 and Mo17 were said to exhibit “clear” overdominance (underdominance).

By following the statistical analysis, an additional 4,112 spots were removed from the data set because of concerns regarding the quality of the associated DNA sequences and 192 exogenous spots also were removed. As a result, this study reports the gene expression patterns of 13,999 “informative” spots.

**Validation of Gene Expression via qRT-PCR**

Primers were designed to amplify a sample of genes that exhibited statistically significant genotype effects in the analysis of the microarray data. Individual ESTs or EST contigs (if available) were compared to the MAGI 4.0 database (29) of assembled maize genomic sequences by using blast (30). The primers were designed by using primer3 (31). The design parameters were used as follows: \( T_m \), 58°C to 61°C, no difference >2°C between the primers in a pair; primer length, 19–24 bp; GC content, 45–55%; amplicon
length, 100–200 bp. Whenever possible, primers were designed to span introns. Only primer pairs having high scoring matches to a single MAGI were synthesized (Integrated DNA Technologies, Coralville, IA). Only primers yielding a single product in conventional PCR and qRT-PCR were used in the validation experiment.

RNA samples from seven biological replications of B73, Mo17, and the F₁ were treated with RNase-free DNase I (Stratagene), extracted with 1:1 phenol:chloroform, and purified with the RNeasy Mini Kit (Qiagen). qRT-PCR was conducted by using an Mx4000 multiplex quantitative PCR system (Stratagene). A human gene (GenBank accession no. AA418251) was spiked into each reaction as an external reference for data normalization.

qRT-PCR data were initially analyzed by using mx4000 analysis software. Genotype-specific $C_t$ values for each gene and control were calculated by using baseline-corrected, ROX-normalized parameters. Three technical replicates were included in each plate, and the average $C_t$ value for each genotype was normalized within a plate to the human control gene by computing $\Delta C_{t,\text{genotype}} = C_{t,\text{genotype}} - C_{t,\text{genotype(control)}}$ (32). The $\Delta C_{t,\text{genotype}}$ values from the seven biological replicates were analyzed with sas statistical software (SAS, Cary, NC) by using a mixed linear model that included the fixed effect of genotype and random effects relevant to the experimental design. The fixed effect of genotype (B73, Mo17, and F₁) was tested for significance ($P < 0.05$) and genes were classified into significant patterns by using pairwise comparison tests ($P < 0.05$).

**ACKNOWLEDGMENTS**

We thank Nathan Springer, Jim Birchler, and Rob Martienssen for fruitful discussions; Karthik Viswanathan, Pengcheng Lu, and Cheng-Ting Yeh for computational support; Sang-Duck Seo for implementing the design of Fig. 2; and an anonymous reviewer for useful suggestions regarding this figure. This research was funded in part by the Iowa State University Plant Sciences Institute, Hatch Act, and State of Iowa funds.

**REFERENCES**

TABLE 1

Heterosis for seedling dry weight

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Mean seedling dry weight (g)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>B73</td>
<td>0.351 ± 0.092</td>
</tr>
<tr>
<td>Mo17</td>
<td>0.310 ± 0.082</td>
</tr>
<tr>
<td>B73 × Mo17</td>
<td>0.392 ± 0.085</td>
</tr>
<tr>
<td>Mo17 × B73</td>
<td>0.517 ± 0.078</td>
</tr>
</tbody>
</table>

* Seedlings of each genotype were grown under the conditions described in Methods and harvested 14 days after planting. B73 × Mo17 and Mo17 × B73 designate reciprocal hybrids in which the female parent is B73 or Mo17, respectively. Dry weights were determined for 36 individual seedlings per genotype.
<table>
<thead>
<tr>
<th>GB no.</th>
<th>BLAST results (e-value)</th>
<th>Sig. pattern</th>
<th>NP P value</th>
<th>NP fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overdominant gene action</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG841554</td>
<td>Unknown protein ($6 \times 10^{-41}$)</td>
<td>B=M&lt;F</td>
<td>0.0313</td>
<td>2.01</td>
</tr>
<tr>
<td>CB334498</td>
<td>ns</td>
<td>B=M&lt;F</td>
<td>0.0060</td>
<td>0.63</td>
</tr>
<tr>
<td>DV550679</td>
<td>ns</td>
<td>B=M&lt;F</td>
<td>0.0140</td>
<td>1.42</td>
</tr>
<tr>
<td>DV489625</td>
<td>Putative serine-threonine protein kinase ($1 \times 10^{-42}$)</td>
<td>B=M&lt;F</td>
<td>0.0063</td>
<td>1.40</td>
</tr>
<tr>
<td>DV496092</td>
<td>Putative peroxidase ($2 \times 10^{-14}$)</td>
<td>B=M&lt;F</td>
<td>0.0163</td>
<td>1.40</td>
</tr>
<tr>
<td>CD661986</td>
<td>Putative jasmonate-induced protein ($2 \times 10^{-19}$)</td>
<td>B=M&lt;F</td>
<td>0.0185</td>
<td>1.35</td>
</tr>
<tr>
<td>CD484960</td>
<td>β-glucosidase aggregating factor precursor ($2 \times 10^{-37}$)</td>
<td>M&lt;B&lt;F</td>
<td>0.0007</td>
<td>1.33</td>
</tr>
<tr>
<td>DV494333</td>
<td>Pre-mRNA splicing factor ($3 \times 10^{-15}$)</td>
<td>B=M&lt;F</td>
<td>0.0036</td>
<td>1.29</td>
</tr>
<tr>
<td>DV549450</td>
<td>Putative SSR α subunit ($9 \times 10^{-75}$)</td>
<td>B=M&lt;F</td>
<td>0.0156</td>
<td>1.26</td>
</tr>
<tr>
<td>CD651121</td>
<td>β-glucosidase aggregating factor precursor ($3 \times 10^{-35}$)</td>
<td>B=M&lt;F</td>
<td>0.0496</td>
<td>1.26</td>
</tr>
<tr>
<td>DV549419</td>
<td>Single-strand DNA endonuclease ($4 \times 10^{-6}$)</td>
<td>B=M&lt;F</td>
<td>0.0271</td>
<td>1.25</td>
</tr>
<tr>
<td>DV550393</td>
<td>Unknown protein ($2 \times 10^{-11}$)</td>
<td>B=M&lt;F</td>
<td>0.0393</td>
<td>1.25</td>
</tr>
<tr>
<td>CB331016</td>
<td>Putative glucose-1-phosphate adenyltransferase ($6 \times 10^{-13}$)</td>
<td>B=M&lt;F</td>
<td>0.0311</td>
<td>1.24</td>
</tr>
<tr>
<td>BM379680</td>
<td>Superoxide dismutase ($1 \times 10^{-43}$)</td>
<td>B=M&lt;F</td>
<td>0.0104</td>
<td>1.24</td>
</tr>
<tr>
<td>CB280807</td>
<td>Proteasome component ($2 \times 10^{-18}$)</td>
<td>B=M&lt;F</td>
<td>0.0034</td>
<td>1.24</td>
</tr>
<tr>
<td>DV491691</td>
<td>ns</td>
<td>B=M&lt;F</td>
<td>0.0383</td>
<td>1.23</td>
</tr>
<tr>
<td>CB603924</td>
<td>Putative jasmonate-induced protein ($2 \times 10^{-19}$)</td>
<td>B=M&lt;F</td>
<td>0.0284</td>
<td>1.23</td>
</tr>
<tr>
<td>DV943290</td>
<td>Calcineurin B protein ($3 \times 10^{-51}$)</td>
<td>B=M&lt;F</td>
<td>0.0205</td>
<td>1.23</td>
</tr>
<tr>
<td>BM080645</td>
<td>Unknown protein ($1 \times 10^{-10}$)</td>
<td>M&lt;B&lt;F</td>
<td>0.0433</td>
<td>1.22</td>
</tr>
<tr>
<td>CB381444</td>
<td>Pollen 2-phosphoglycerate dehydrogenase 2 precursor ($6 \times 10^{-25}$)</td>
<td>B=M&lt;F</td>
<td>0.0103</td>
<td>1.22</td>
</tr>
<tr>
<td>DV622265</td>
<td>ADP-ribosylation factor ($5 \times 10^{-81}$)</td>
<td>B=M&lt;F</td>
<td>0.0349</td>
<td>1.21</td>
</tr>
<tr>
<td>DV489865</td>
<td>Expressed protein ($2 \times 10^{-39}$)</td>
<td>B=M&lt;F</td>
<td>0.0154</td>
<td>1.21</td>
</tr>
<tr>
<td>CB617229</td>
<td>Superoxide dismutase ($2 \times 10^{-84}$)</td>
<td>B=M&lt;F</td>
<td>0.0107</td>
<td>1.20</td>
</tr>
<tr>
<td>CD651750</td>
<td>β-glucosidase aggregating factor precursor ($4 \times 10^{-22}$)</td>
<td>B=M&lt;F</td>
<td>0.0430</td>
<td>1.20</td>
</tr>
<tr>
<td>BM073302</td>
<td>Superoxide dismutase ($7 \times 10^{-74}$)</td>
<td>B=M&lt;F</td>
<td>0.0182</td>
<td>1.19</td>
</tr>
<tr>
<td>CB381307</td>
<td>Putative 20-kDa chaperonin ($7 \times 10^{-47}$)</td>
<td>B=M&lt;F</td>
<td>0.0286</td>
<td>1.19</td>
</tr>
<tr>
<td>CB331033</td>
<td>Putative translation elongation factor ($2 \times 10^{-30}$)</td>
<td>B=M&lt;F</td>
<td>0.0495</td>
<td>1.18</td>
</tr>
<tr>
<td>CB329753</td>
<td>Superoxide dismutase ($4 \times 10^{-77}$)</td>
<td>B=M&lt;F</td>
<td>0.0330</td>
<td>1.18</td>
</tr>
<tr>
<td>CB380870</td>
<td>Putative jasmonate-induced protein ($6 \times 10^{-19}$)</td>
<td>M&lt;B&lt;F</td>
<td>0.0240</td>
<td>1.17</td>
</tr>
<tr>
<td>CB605313</td>
<td>β-glucosidase aggregating factor precursor ($6 \times 10^{-22}$)</td>
<td>B=M&lt;F</td>
<td>0.0427</td>
<td>1.17</td>
</tr>
<tr>
<td>CB886104</td>
<td>Histone H2B ($4 \times 10^{-35}$)</td>
<td>B=M&lt;F</td>
<td>0.0315</td>
<td>1.17</td>
</tr>
<tr>
<td>CD001350</td>
<td>Nucleosome chromatin assembly factor group D protein ($7 \times 10^{-15}$)</td>
<td>B=M&lt;F</td>
<td>0.0200</td>
<td>1.16</td>
</tr>
<tr>
<td>CD485186</td>
<td>DNA-binding protein ($1 \times 10^{-22}$)</td>
<td>B=M&lt;F</td>
<td>0.0344</td>
<td>1.16</td>
</tr>
<tr>
<td>DV492982</td>
<td>Putative translation elongation factor ($5 \times 10^{-40}$)</td>
<td>B=M&lt;F</td>
<td>0.0496</td>
<td>1.16</td>
</tr>
<tr>
<td>GB no.</td>
<td>BLAST results (e-value)†</td>
<td>Sig. pattern‡</td>
<td>NP $P$ value§</td>
<td>NP fold change¶</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------------------</td>
<td>--------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>DV493742</td>
<td>Zein ($2 \times 10^{-67}$)</td>
<td>F&lt;B=M</td>
<td>0.0137</td>
<td>4.27</td>
</tr>
<tr>
<td>CB281986</td>
<td>Putative nucleoside diphosphate kinase ($7 \times 10^{-31}$)</td>
<td>F&lt;B=M</td>
<td>0.0201</td>
<td>2.71</td>
</tr>
<tr>
<td>BM337640</td>
<td>15-kDa beta zein ($8 \times 10^{-44}$)</td>
<td>F&lt;B=M</td>
<td>0.0074</td>
<td>3.85</td>
</tr>
<tr>
<td>AI612441</td>
<td>ns</td>
<td>F&lt;B=M</td>
<td>0.0162</td>
<td>3.45</td>
</tr>
<tr>
<td>BM351629</td>
<td>ns</td>
<td>F&lt;B=M</td>
<td>0.0433</td>
<td>2.46</td>
</tr>
<tr>
<td>BM073434</td>
<td>Pathogenesis-related protein ($5 \times 10^{-66}$)</td>
<td>F&lt;B=M</td>
<td>0.0002</td>
<td>2.00</td>
</tr>
<tr>
<td>DV489785</td>
<td>Multidrug resistance protein ($6 \times 10^{-19}$)</td>
<td>F&lt;B=M</td>
<td>0.0047</td>
<td>1.45</td>
</tr>
<tr>
<td>DV942972</td>
<td>ns</td>
<td>F&lt;B=M</td>
<td>0.0068</td>
<td>1.61</td>
</tr>
<tr>
<td>BM073507</td>
<td>ns</td>
<td>F&lt;B=M</td>
<td>0.0252</td>
<td>1.58</td>
</tr>
<tr>
<td>DV622486</td>
<td>Branched silkless 1 ($2 \times 10^{-9}$)</td>
<td>F&lt;B=M</td>
<td>0.0341</td>
<td>1.24</td>
</tr>
</tbody>
</table>

The $P$ value for equality of means among the three genotypes was <0.0212 for all listed ESTs. The ESTs are listed in decreasing order of near-parent fold changes.

*Genbank accession number: Based on sequence similarity, the following groups of ESTs appear to correspond to single genes: CD661986, CB603924, CD651750, CB380870, and CB605313; CD484960 and CD651121; BM379680, CB617229, and BM073302; and CD001350 and CD485186.

†Individual ESTs or the corresponding EST contigs (if available) were screened against a copy of the NCBI nr database downloaded February 8, 2006 by using blastx. ns indicates no significant blast hits by using an $e$-value cutoff of $1 \times 10^{-5}$.

‡Indicates a failure to reject the null hypothesis that the values of the indicated genotypes are identical at $P < 0.05$. B, B73; M, Mo17; F, F1.

§Near-parent $P$ value from equality of means test between F1 and parent with the nearest level of gene expression.

¶Fold change between F1 and parent with the nearest level of gene expression shown as a ratio. For overdominant genes, the ratio of F1 to near parent expression level is listed; for underdominant genes, the ratio of the near parent to F1 expression level is listed.
### TABLE 3

The 35 genes identified as being differentially regulated among genotypes in the microarray analysis that had significant qRT-PCR validation results

<table>
<thead>
<tr>
<th>GenBank Accession No.</th>
<th>BLAST Results (e value)*</th>
<th>Microarray Results</th>
<th>qRT-PCR Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Fold-change †</td>
<td>Significant Pattern †</td>
</tr>
<tr>
<td>GenBank</td>
<td></td>
<td>Sector †</td>
<td>P-value</td>
</tr>
<tr>
<td>Equivalent gene action observed in microarray and qRT-PCR experiments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BM073941</td>
<td>Beta-amyrin synthase (2e-48)</td>
<td>87.98</td>
<td>10</td>
</tr>
<tr>
<td>BM340381</td>
<td>ns</td>
<td>10.73</td>
<td>10</td>
</tr>
<tr>
<td>BM338817</td>
<td>ns</td>
<td>9.79</td>
<td>10</td>
</tr>
<tr>
<td>BM074072</td>
<td>Heme A:farnesyltransferase (1e-14)</td>
<td>9.37</td>
<td>10</td>
</tr>
<tr>
<td>BM334691</td>
<td>ns</td>
<td>6.64</td>
<td>3</td>
</tr>
<tr>
<td>BM073390</td>
<td>ns</td>
<td>6.27</td>
<td>3</td>
</tr>
<tr>
<td>BM337350</td>
<td>ns</td>
<td>5.06</td>
<td>10</td>
</tr>
<tr>
<td>BM079864</td>
<td>Circulin B (8e-07)</td>
<td>4.31</td>
<td>3</td>
</tr>
<tr>
<td>DV942972</td>
<td>ns</td>
<td>3.86</td>
<td>9</td>
</tr>
<tr>
<td>DV550757</td>
<td>Unknown protein (2e-59)</td>
<td>3.47</td>
<td>3</td>
</tr>
<tr>
<td>DV549373</td>
<td>Putative chloroplast 50S ribosomal protein (4e-30)</td>
<td>2.40</td>
<td>11</td>
</tr>
<tr>
<td>BM080645</td>
<td>Unknown protein (1e-10)</td>
<td>2.08</td>
<td>11</td>
</tr>
<tr>
<td>BM073434</td>
<td>Pathogenesis related protein (5e-66)</td>
<td>2.05</td>
<td>6</td>
</tr>
<tr>
<td>BM072868</td>
<td>Unknown protein (2e-12)</td>
<td>2.01</td>
<td>11</td>
</tr>
<tr>
<td>DV490892</td>
<td>ns</td>
<td>1.76</td>
<td>2</td>
</tr>
<tr>
<td>BM348583</td>
<td>Putative cytochrome P450 (2e-6)</td>
<td>2.60</td>
<td>10</td>
</tr>
<tr>
<td>BM073611</td>
<td>ns</td>
<td>22.77</td>
<td>10</td>
</tr>
<tr>
<td>BG841239</td>
<td>Unknown protein (4e-76)</td>
<td>17.07</td>
<td>10</td>
</tr>
<tr>
<td>BM073284</td>
<td>Circulin B (7e-7)</td>
<td>16.57</td>
<td>3</td>
</tr>
<tr>
<td>BM073916</td>
<td>ns</td>
<td>5.01</td>
<td>9</td>
</tr>
<tr>
<td>BM336730</td>
<td>ns</td>
<td>4.18</td>
<td>9</td>
</tr>
<tr>
<td>BM337359</td>
<td>Putative wound-inducive gene (1e-14)</td>
<td>3.42</td>
<td>8</td>
</tr>
<tr>
<td>BM337880</td>
<td>ns</td>
<td>3.26</td>
<td>8</td>
</tr>
<tr>
<td>BM334498</td>
<td>Cytosolic aldehyde dehydrogenase RF2D (2e-49)</td>
<td>2.26</td>
<td>2</td>
</tr>
<tr>
<td>BM348968</td>
<td>Putative MYB29 protein (2e-48)</td>
<td>2.24</td>
<td>4</td>
</tr>
<tr>
<td>AI861151</td>
<td>Putative Xa1-like protein (2e-68)</td>
<td>2.22</td>
<td>4</td>
</tr>
<tr>
<td>DV942867</td>
<td>ns</td>
<td>2.20</td>
<td>8</td>
</tr>
<tr>
<td>BG842276</td>
<td>unknown protein (1e-10)</td>
<td>2.05</td>
<td>11</td>
</tr>
<tr>
<td>BG841156</td>
<td>Carboxypeptidase D (7e-74)</td>
<td>2.02</td>
<td>2</td>
</tr>
<tr>
<td>CB334498</td>
<td>Putative 32.7 kDa jasmonate-induced protein (2e-19)</td>
<td>1.65</td>
<td>1</td>
</tr>
<tr>
<td>CD661986</td>
<td>Putative serine-threonine protein kinase (1e-42)</td>
<td>1.59</td>
<td>12</td>
</tr>
<tr>
<td>DV489625</td>
<td>ns</td>
<td>1.41</td>
<td>12</td>
</tr>
</tbody>
</table>

Different modes of gene action detected in microarray and qRT-PCR experiments

<table>
<thead>
<tr>
<th>GenBank</th>
<th>BLAST Results (e value)*</th>
<th>Microarray Results</th>
<th>qRT-PCR Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Fold-change †</td>
<td>Significant Pattern †</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sector †</td>
<td>P-value</td>
</tr>
<tr>
<td>BG841472</td>
<td>Putative cystatin (3e-61)</td>
<td>4.66</td>
<td>3</td>
</tr>
<tr>
<td>BM073340</td>
<td>nonspecific lipid transfer protein (5e-23)</td>
<td>2.92</td>
<td>3</td>
</tr>
<tr>
<td>BM268642</td>
<td>transcription factor MADS57 (1e-23)</td>
<td>2.14</td>
<td>11</td>
</tr>
</tbody>
</table>
The 10 genes with nonsignificant patterns in the qRT-PCR experiment are listed in Table 5.
*Individual ESTs or the corresponding EST contigs (if available) were screened against a copy of the NCBI nr database downloaded February 8, 2006 by using blastx. ns indicates no significant blast hits by using an e-value cutoff of $1 \times 10^{-5}$; †Fold changes were calculated between highest- and lowest-expressing genotypes. ‡Sector location in Fig. 2; § indicates a failure to reject the null hypothesis that the values of the indicated genotypes are identical at $P < 0.05$. B, B73; M, Mo17; F, F$_1$. 
TABLE 4

Microarray analysis results, mode of gene action classified, and BLAST annotation for the 1,367 ESTs identified as differentially expressed among genotypes (15% FDR)

This table is available and published as supporting information on the PNAS web site at:
http://www.pnas.org/content/suppl/2006/04/14/0510430103.DC1/10430Table4.pdf
TABLE 5
Ten genes in the qRT-PCR experiments did not have sufficient statistical power to determine a significant pattern of gene action

<table>
<thead>
<tr>
<th>GenBank Accession No.</th>
<th>BLAST Results (e value)*</th>
<th>Fold-change †</th>
<th>Sector ‡</th>
<th>Significant Pattern §</th>
<th>qRT-PCR P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB815723</td>
<td>cysteine proteinase inhibitor (2e-55)</td>
<td>2.99</td>
<td>2</td>
<td>B &lt; M ≡ F</td>
<td>0.7942</td>
</tr>
<tr>
<td>AI782877</td>
<td>unknown protein (2e-87)</td>
<td>4.56</td>
<td>4</td>
<td>F ≡ B &lt; M</td>
<td>0.6929</td>
</tr>
<tr>
<td>AI665879</td>
<td>cytochrome c oxidase subunit 2 (1e-13)</td>
<td>1.97</td>
<td>5</td>
<td>F ≡ B &lt; M</td>
<td>0.1122</td>
</tr>
<tr>
<td>BM348600</td>
<td>ns</td>
<td>3.87</td>
<td>8</td>
<td>F &lt; B</td>
<td>0.1746</td>
</tr>
<tr>
<td>AW126440</td>
<td>unknown protein (3e-11)</td>
<td>6.36</td>
<td>10</td>
<td>M &lt; B ≡ F</td>
<td>0.098</td>
</tr>
<tr>
<td>CD484517</td>
<td>similar to H3 histone family (8e-12)</td>
<td>2.35</td>
<td>11</td>
<td>M &lt; B ≡ F</td>
<td>0.3528</td>
</tr>
<tr>
<td>AI714832</td>
<td>unknown protein (2e-36)</td>
<td>2.27</td>
<td>11</td>
<td>M &lt; F</td>
<td>0.4008</td>
</tr>
<tr>
<td>BM351629</td>
<td>ns</td>
<td>3.91</td>
<td>7</td>
<td>F &lt; B ≡ M</td>
<td>0.6863</td>
</tr>
<tr>
<td>DV489785</td>
<td>multidrug resistance associated protein (2e-18)</td>
<td>1.78</td>
<td>7</td>
<td>F &lt; B ≡ M</td>
<td>0.3865</td>
</tr>
<tr>
<td>DV493742</td>
<td>unknown protein (3e-75)</td>
<td>4.66</td>
<td>7</td>
<td>F &lt; B ≡ M</td>
<td>0.8502</td>
</tr>
</tbody>
</table>

*Individual ESTs or the corresponding EST contigs (if available) were screened against a copy of the NCBI nr database downloaded Feb 8, 2006 by using Blastx. ns indicates no significant BLAST hits using an E-value cutoff of 1e-5;
† Fold changes were calculated between highest and lowest expressing genotypes;
‡ Sector location in Fig. 2;
§ Significant pattern determined in microarray analysis. ≡ indicates a failure to reject the null hypothesis that the values of the indicated genotypes are identical at p<0.05. B, B73; M, Mo17; F, F1.
FIGURE LEGENDS

**Figure 1.** Distribution of $P$ values for the 13,999 gene-specific hypothesis tests for equality of means across genotypes.

**Figure 2.** Two-dimensional presentation of gene action and fold changes of ESTs that are differentially expressed in B73, Mo17, and their F$_1$ hybrid. ESTs falling directly on a dashed line between 1 and 7 o’clock, 3 and 9 o’clock, or 5 and 11 o’clock exhibited differences in expression from the low to the middle expressing genotype that is equivalent to the change from the middle to the high expressing genotype. ESTs falling on the horizontal and vertical lines exhibited pure additivity and over- (or under-) dominance, respectively (see plot labels). The radius at which an EST is plotted represents the log$_2$ of the fold change between the high- and low-expressing genotypes. ESTs associated with a FDR of 1%, 5%, and 15% are shown in red, blue, and black, respectively. To provide better resolution for those of the 1,367 differentially expressed ESTs with smaller fold changes, only the 1,361 ESTs that exhibited changes of <16-fold are plotted. The remaining six ESTs are listed in Table4.

**Figure 3.** Histogram of fold changes (differences between low- and high-expressing genotypes) for the 1,367 ESTs that exhibited statistically significant differences in expression among genotypes.
Figure 1 Swanson-Wagner, Jia et al., 2006
Figure 2 Swanson-Wagner, Jia et al., 2006
Figure 3 Swanson-Wagner, Jia et al., 2006
CHAPTER 3. LOSS OF RNA–DEPENDENT RNA POLYMERASE 2 (RDR2) FUNCTION CAUSES WIDESPREAD AND UNEXPECTED CHANGES IN THE EXPRESSION OF TRANSPOSONS, GENES, AND 24-NT SMALL RNAs

A paper published in *PLoS GENETICS*¹

Yi Jia², Damon R Lisch³, Kazuhiro Ohtsu², Michael J Scanlon⁴, Dan Nettleton² and Patrick S Schnable²,⁵

ABSTRACT

Transposable elements (TEs) comprise a substantial portion of many eukaryotic genomes and are typically transcriptionally silenced. RNA–dependent RNA polymerase 2 (RDR2) is a component of the RNA–directed DNA methylation (RdDM) silencing pathway. In maize, loss of *mediator of paramutation1* (*mop1*) encoded RDR2 function results in reactivation of transcriptionally silenced *Mu* transposons and a substantial reduction in the accumulation of 24 nt short-interfering RNAs (siRNAs) that recruit RNA silencing components. An RNA–seq experiment conducted on shoot apical meristems (SAMs) revealed that, as expected based on a model in which RDR2 generates 24 nt siRNAs that suppress expression, most differentially expressed DNA TEs (78%) were up-regulated in the *mop1* mutant. In contrast, most differentially expressed retrotransposons (68%) were down-regulated. This striking difference suggests that distinct silencing mechanisms are applied to different silencing templates. In addition, >6,000 genes (24% of analyzed genes), including nearly 80% (286/361) of genes in chromatin modification pathways, were differentially expressed. Overall, two-thirds of differentially regulated genes were down-regulated in the *mop1* mutant. This finding suggests that RDR2 plays a significant role in regulating the expression of not only transposons, but also of genes. A re-analysis of existing small RNA data identified both RDR2–sensitive and RDR2–resistant species of 24 nt siRNAs that we hypothesize may at least partially explain the complex changes in the expression of genes and transposons observed in the *mop1* mutant.

Author Summary

Shoot apical meristems (SAMs) are ultimately responsible for generating all above-ground plant tissues. Recent studies highlighted the effects of chromatin remodeling on the expression of various genes important to SAM development. The transposons that comprise a substantial portion of many eukaryotic genomes are typically transcriptionally silenced, presumably to promote genome stability. We demonstrate that a loss of a key component of the RNA–dependent DNA Methylation (RdDM) silencing pathway affects the expression of not only transposons but also thousands of genes, including nearly 80% of the chromatin-associated genes. Surprisingly, the expression of many transposons and genes is down-regulated via the loss of this component of the silencing pathway. In this study, we have shown that a maize mutation

---

¹ Reprinted with permission of *PLoS GENETICS*.
² Iowa State University, Ames, IA; ³ University of California, Berkeley, CA; ⁴ Cornell University, Ithaca, NY; ⁵ To whom correspondence should be addressed (schnable@iastate.edu).
of RDR2 causes significant changes in SAM morphology. In combination, these observations indicate the complexity of transcriptome regulation and the crucial roles of RDR2 on transcriptome regulation, chromatin modification, and SAM development.

INTRODUCTION

Repetitive sequences, including transposable elements (TEs) and tandem repeats, comprise a substantial fraction of many eukaryotic genomes. To protect genome integrity TEs are typically transcriptionally silenced via epigenetic mechanisms [1]–[3]. At the core of many of these mechanisms are a variety of small RNAs. Diverse small RNA pathways exist in most eukaryotes producing microRNAs and short-interfering RNAs (siRNAs) that function to negatively regulate gene expression and/or to suppress the activity of transposons. siRNAs derived from double-stranded RNAs via mechanisms that are either RNA-dependent RNA polymerase (RDR) dependent or independent. Various RDRs (e.g. RDR1, RDR2 and RDR6) are functional in different siRNA pathways and most siRNAs are biosynthesized from the heterochromatic loci, a process that generally requires RDR2 and DICER-LIKE3 (DCL3)[4]. In Arabidopsis, these heterochromatic 24 nt siRNAs, predominate the small RNA population. These siRNAs recruit chromatin-targeted RNA silencing components to form transcriptionally silent heterochromatin, which is derived mainly from TEs and tandem repeats [5], by cytosine methylation and various histone modifications, such as histone deacetylation and histone H3 lysine 9 dimethylation. RDR2 is required for the biogenesis of most of these 24 nt siRNAs [5],[6]. Thus, RDR2 is a key component of the chromatin-targeted RNA silencing process, which is also called the RNA dependent DNA Methylation (RdDM) pathway.

The maize homolog of AtRDR2, *mediator of paramutation1 (mop1)* [7], is required to establish and maintain paramutation at multiple genetic loci [8]. Paramutation is an epigenetic process in which the interaction of two alleles of a single locus causes a heritable change of one allele that is induced by the other allele. Paramutation was first observed on the *red1 (r1)* locus in maize by R.A. Brink in the 1950s in which specific weakly expressed alleles can heritably change other strongly expressed alleles to weakly expressed alleles [9]. In maize, paramutation is also observed in *b1* and *pl1* loci. The *mop1* gene, which encodes the maize version of RDR2, is required for paramutation in *b1*, *r1* and *pl1* loci, demonstrating that an RNA-dependent mechanism is important for paramutation in maize [7]. A tandem repeat 100 kb upstream of *b1* locus is required for paramutation in *b1* locus. This repeat does not share sequence similarity with the *b1* coding sequence, but has been demonstrated to physically interact with the *b1* transcription start site [10]. Mutations in *mop1* can also result in reactivation of transcriptionally silenced Mutator transposons and can substantially reduce the overall levels of 24 nt siRNAs, demonstrating that in maize RDR2 contributes to the silencing of repetitive elements and plays an important role in the biogenesis of 24 nt siRNAs [11-12].
Shoot apical meristems (SAMs) are responsible for the elaboration of all above-ground plant organs [13]. Maintenance of SAM identity and organogenesis are precisely regulated by a complex regulatory network involving various transcriptional factors and signal transduction proteins, as well as epigenetic factors [14]. In a previously conducted global gene expression analysis of maize SAMs and seedlings we identified 2,700+ genes as being preferentially expressed in SAMs [15]. This included ~60 retrotransposons that were substantially up-regulated in SAMs as compared to seedlings despite the fact that \textit{mop1} was expressed >100 fold higher in SAMs than in seedlings. It was not clear at that time why repetitive retrotransposons and \textit{mop1}, which contributes to the silencing of repetitive elements, would be both significantly up-regulated in SAMs. Given that retrotransposon-derived transcripts represent ~9% of all SAM transcripts (as opposed to ~0.3% of seedling transcripts)[15], the SAM is an ideal model for studying the effects of \textit{mop1} on the accumulation of retrotransposon-derived transcripts.

Here we report an analysis of an RNA-seq experiment that detected hundreds of transposons and thousands of the genes that are differentially expressed in \textit{mop1} mutant and non-mutant SAMs. This finding suggests that RDR2 plays a role in regulating the expression of not only transposons, but also of genes. Consistent with this observation, RDR2 mutants also exhibited a distinct SAM morphology relative to their wild type siblings, suggesting a role for RDR2 in normal SAM development. As expected based on its role in the RdDM pathway, loss of RDR2 function resulted in the up-regulation of many DNA TEs, retrotransposons and genes. However, some DNA TEs and many retrotransposons and genes were down-regulated in the \textit{mop1} mutant.

**RESULTS**

**Collection of SAMs via Laser Capture Microdissection (LCM) and RNA–seq**

A family segregating 1:1 for \textit{mop1} homozygotes and heterozygotes was planted in a growth chamber and harvested 14-days after planting. Individual seedlings were genotyped to identify \textit{mop1} homozygotes and heterozygotes (see Materials and Methods). To determine the effects of \textit{mop1} on SAM development the ratios of height versus width of mutants (homozygous) and non-mutants (heterozygous) SAMs were compared (Figure 1); mutant ratios were significantly smaller than non-mutant ratios (p-value = 0.006; Table S1). SAMs plus leaves at plastochron 0 (P0) and P1 stages were collected via LCM followed by RNA extraction and amplification according to our previously published procedures [15]. A pooled RNA sample from twelve mutant SAMs and a pooled RNA sample from ten non-mutant SAMs were subjected to Illumina/Solexa sequencing (see Materials and Methods). 2.8 million reads from \textit{mop1} mutants and 4.1 million reads from non-mutants could be uniquely mapped to the Maize Genome Sequencing Project's (MGSP's) B73 reference genome [16] (see Materials and Methods).

**Expression of various TEs is differentially affected by loss of RDR2 function**
More than 80% of the B73 reference genome is composed of TEs [16]. To ensure genome stability these elements are mostly suppressed via genome defense systems such as chromatin-based silencing, which is guided and reinforced by 24 nt siRNAs [17]. RDR2 plays an important role in the biogenesis of 24 nt siRNAs [18]. To test the hypothesis that the loss of RDR2 function in the mop1 mutant would result in the activation of TEs (i.e., an increased accumulation of TE-derived transcripts), the Illumina/Solexa reads obtained from the SAM were aligned to all annotated TEs in the B73 genome (see Materials and Methods). As expected based on the mechanism associated with the RdDM silencing pathway, many DNA TEs (class II) and retrotransposons (class I) were up-regulated in the mop1 mutant (Table 1). However, a significant fraction of DNA TEs and retrotransposons were down-regulated (Table 1). The most strongly up-regulated DNA TE super-families were Stowaway and Tourist and the most strongly down-regulated DNA TE family was CACTA (Figure S1). Both the Ty1/Copia-like and Ty3/Gypsy-like super-families of retrotransposons include some families that were up-regulated and others that were down-regulated (Figure S2).

To assess the effects of mop1 on specific groups of DNA TEs, each DNA TE super-family was divided into families based on a phylogenetic analysis conducted by the Maize Transposon Consortium that defined 797 unique families (i.e., monophyletic clades) [16]. Of these families, 22% (175/797) were differentially expressed in the mop1 mutant (Table 1; Table S2). Consistent with our hypothesis, most (78%; 136/175) of the differentially expressed DNA TEs were up-regulated in the mop1 mutant. In particular, of the 32/140 TE families annotated as Mutator-like elements (MULEs) that were differentially expressed, 29 (91%) were up-regulated. This is consistent with the report that silenced Mutator TEs can be reactivated in mop1 mutants [11]. Similarly, of the 31/228 hAT families that were differentially expressed, 74% (23/31) were up-regulated. Among the 58/151 differentially expressed CACTA families many (37/58) individual CACTA families were up-regulated. Even so, CACTA DNA TEs as a group were down-regulated by loss of RDR2 function (Figure S1). The up-regulation of one MULE element and the down-regulation of one hAT element (viz., an Ac-like element) were confirmed by quantitative real-time PCR (qRT-PCR) (Figure 2).

A similar analysis was performed on retrotransposons, which have been categorized by the Maize Transposon Consortium into super-families (e.g., RLG, retrotransposon LTR Gypsy), families (e.g., Huck) and sub-families (e.g., Ac186577_1525). 71% of the 608 unique retrotransposon sub-families were differentially expressed to various degrees in the mop1 mutant (Table S3). Consistent with current understanding of the RdDM silencing pathway, approximately one-third (32%) of the differentially expressed retrotransposon sub-families were up-regulated in mop1 mutants as compared to non-mutants. But inconsistent with expectations, the majority (68%) of the differentially regulated retrotransposons were down-regulated in the mutant (Table 1; Table S3). Some specific sub-families of retrotransposons were mostly up- or mostly down-regulated. For example, within the Ty1/Copia-like super-family, 12/16 Ji and 9/17 Opie sub-families were up-regulated. Within the Ty3/Gypsy-like super-family, 6/7 Flip and 19/20 Huck sub-families were up-regulated. In contrast, 38/41 Ty3/Gypsy-like Cinful-zeon sub-families and 9/10
annotated Ty3/Gypsy-like Prem1 sub-families were down-regulated in the mop1 mutant (Table S3). These distinct responses to loss of RDR2 function suggest that the expressions of different retrotransposon families are regulated via different mechanisms.

Many chromatin modification genes are down-regulated in the mop1 mutant

To study the impacts of mop1 on chromatin modification pathways, 386 maize genes annotated as being involved in chromatin modification were downloaded in Dec. 2008 from ChromDB [19]. In total, 361 of these genes can be uniquely mapped to the B73 reference genome (see Materials and Methods). RNA-seq reads that map to these chromatin-associated genes were used to conduct Fisher's exact tests. Nearly 80% (286/361) of these chromatin pathway genes were differentially expressed between the mop1 mutants and non-mutant siblings using a 5% false discover rate (FDR) cutoff, a frequency far higher than observed when considering all genes (24%). Approximately ¾ (76%) of differentially expressed chromatin genes were down-regulated in the mop1 mutant (Table S4). A wide variety of chromatin-associated genes exhibited differential expression between mop1 mutants and non-mutants, including those affecting various histone modifications, such as histone ubiquitination, methylation, acetylation and deacetylation (Table S4). As the maize homolog of AtRDR2, mop1 is expected to function in the RdDM pathway, which involves the biogenesis of 24 nt siRNAs, de novo methylation of DNA, maintenance of DNA methylation, and demethylation [20]. Almost all of the genes known to be implicated in the RdDM pathway were down-regulated in the mop1 mutants (Table 2). The maize genome contains two DCL3 paralogs, which are involved in the biogenesis of 24 nt small RNAs. One of these (Zmdcl3b) was down-regulated while the other (Zmdcl3a) was up-regulated (Table 2). These opposite responses suggest that the two DCL3 paralogs may be functionally distinct. This type of divergent gene expression pattern was also observed for DRM protein in which one maize homolog was down-regulated and another was up-regulated. In addition to DRM protein, AGO4, DRD1, and MET1proteins function in the de novo methylation pathway [20]. All maize homologs of these genes were down-regulated (Table 2). CMT3, MET1, DDM1, HDA6, SUVH4 function in the maintenance methylation pathway [20]. All maize homologs of these genes were down-regulated (Table 2). DNA demethylation is thought to regulate epigenome dynamics in opposition to the RdDM pathway. In Arabidopsis, ROS1 and DME remove DNA methylation [20]. There are two homologs of DME gene in maize and both were down-regulated as well (Table 2). The expression levels of several of genes important for epigenetic silencing (viz., met1, met3, three ago4 paralogs, and ddm1) were tested via qRT-PCR with results that were consistent with those obtained from RNA-seq (Figure 2). These observations demonstrate that there is widespread down-regulation of components in the RdDM pathway in the mop1 mutant, suggesting either that MOP1 positively regulates the entire pathway, or that genes involved in chromatin modification and DNA methylation are co-regulated in maize.

In addition to mop1, mutations in two other maize genes are known to affect the accumulation of both 24–26 nt siRNAs and DNA methylation. rmr1 (required to maintain repression1) encodes a SWI/SNF2 class
chromatin remodeling protein. Mutations in \textit{rmr1} have dramatic effects on accumulation of 24–26 nt siRNAs, maintenance of the repressed state of paramutant genes, and methylation of \textit{Mu} transposons [21],[22]. Unlike the related DDM1 orthologs (Table 2), expression of \textit{rmr1} was not significantly changed in the \textit{mop1} mutant. This observation is consistent with the suggestion that RMR1 may act genetically upstream and sometimes independently of RDR2 [22]. The second cloned gene that affects 24–26 nt siRNA accumulation in maize is \textit{rmr6}. This gene encodes the conserved Pol IV largest subunit (RPD1) and is required for paramutation [23]. Although expression of this gene was reduced in the \textit{mop1} mutant, this change was not significant.

\textbf{Widespread changes in gene expression following loss of RDR2 activity}

The finding that SAM morphology differs between \textit{mop1} mutant and non-mutants led us to hypothesize that \textit{mop1} affects not only the expression of TEs and components in RdDM pathway but also genes important to the development of the SAM. To test this hypothesis the Illumina/Solexa reads were mapped to the “filtered gene set” of maize generated by the MGSP (see Materials and Methods). Among reads that could be uniquely mapped to the genome, 2.2 million (78\%) from \textit{mop1} mutants and 3.2 million (79\%) from non-mutants aligned to gene models (Figure S3; Table S5). At least one Illumina/Solexa read from at least one of the two genotypes aligned to 24,743 of the 32,540 genes in the MGSP’s filtered gene set (Table S5). Of these genes, 6,016 (24\% of 24,743) could be declared to be differentially expressed between the mutant and non-mutant pools at an estimated 5\% FDR [24]. The ratio of number of genes that were up-regulated in the \textit{mop1} mutant to those that were down-regulated was ~1:2 (Table S6). Consistent with our finding that \textit{mop1} mutant SAMs differ morphologically from non-mutant siblings (Figure 1; Table S1), several key regulators of SAM development, including \textit{fasciated ear2} [25], \textit{terminal earl-like 2} [26], \textit{outer cell layer4} [27] and \textit{liguleless3} (encoding a \textit{Knotted} class 1 homeodomain protein) [28] were differentially expressed (Table 3). The differential expression of one of these genes, \textit{liguleless3}, was validated via quantitative real-time PCR (qRT-PCR) (Figure 2). This finding suggests that 24 nt siRNAs play a role in regulating (directly or indirectly) the expression of not only transposons, but also of genes.

\textbf{Re-analysis of existing 24 nt siRNA data}

The analyses described above demonstrate that loss of RDR2 function results in widespread changes in the accumulation of transcripts from DNA TEs, retrotransposons and genes. As expected based on a model in which RDR2 generates 24 nt siRNAs that suppress expression, many TEs and genes were up-regulated in the \textit{mop1} mutant. Interestingly, some DNA TEs and many retrotransposons and genes were actually down-regulated in the \textit{mop1} mutant, demonstrating new complexities in the regulation of the expression of transposons and genes.

To explore this complexity we undertook a re-analysis of 24 nt siRNAs isolated from immature ears of \textit{mop1} and non-mutant plants [18]. Nobuta et al. reported that \textit{mop1} mutants accumulated many fewer 24 nt siRNAs (as a proportion of all small RNAs) than do wild-type [18]. Their analysis treated all 24 nt siRNAs
as a group. We extended their analysis by considering the effects of the mop1 mutation on the accumulation of each individual species of 24 nt siRNAs in their data set. As such, our analysis enabled us to identify specific RNA species that make up a significantly greater proportion of the observed reads from one genotype than from the other (see Materials and Methods).

Considering the union of mop1 mutant and non-mutant reads, the Nobuta et al. data set contains >2.3M distinct 24 nt siRNA species. Many of these are present at very low abundance and as such it would not be possible to detect statistically significant differences between the two genotypes even if such differences exist. Of the 5% of RNA species for which 5 or more counts were recorded in the union of the two genotypes (125,344/2.3M), we found that 30% (38,564/125,344) of the 24 nt siRNAs were differentially expressed between the two genotypes. Consistent with the report of Nobuta et al., most of the 38,564 differentially regulated species of 24 nt siRNAs (33,614) were down-regulated in the mop1 mutant (Figure S4; Table S7). We term these “RDR2-sensitive 24 nt siRNAs”. Quite unexpectedly, 4,950 distinct species of 24 nt siRNAs were significantly “up-regulated” in the mop1 mutant (Figure S4; Table S7). Although some of these may be actually up-regulated, others may simply be less down-regulated in the mutant than are other species of 24 nt siRNAs (see Materials and Methods). We have therefore termed these “up-regulated” species “RDR2-resistant 24 nt siRNAs”.

**DISCUSSION**

RDR2 is an essential component of the heterochromatin silencing pathway in multiple species [5], [29] and functions in DNA and histone methylation, the biogenesis of 24 nt siRNAs and the silencing of repetitive DNAs [30]. The maize homolog of RDR2, mop1, was originally identified as a mutant that functions as an epigenetic regulator of a target gene via interactions with upstream tandem repeats [7]. mop1 is also required for the methylation of the terminal inverted repeats of Mu TEs and for the maintenance of silencing of MuDR transposons [31]. Based on its mutant phenotypes, it has been hypothesized that the mop1 gene regulates many loci [8]. To test this hypothesis and to examine the effect of RDR2 on the silencing of TEs in a large, complex genome, we conducted RNA-seq experiments on SAMs of mop1 mutant and non-mutant seedlings. SAMs were selected for analysis because they are responsible for the elaboration of all aerial organs [13], they have a complex transcriptome [32], and our prior analyses had revealed that multiple retrotransposons and mop1 are all substantially up-regulated in SAMs as compared to seedlings [15].

**Effect of RDR2 on gene expression**

We identified more than 6,000 genes whose expression differed between mop1 mutant and non-mutant SAMs. These widespread differences in gene expression are consistent with the multiple developmental defects associated with the loss of mop1 function in mutants [8].
Over several generations, maize lines that carry the *mop1* mutation can accumulate a variety of epimutant phenotypes (Lisch, unpublished data). In this study we controlled for the effects of any segregating epialleles by analyzing RNA from pools of *mop1* and non-mutant SAMs. However, our discovery that genes involved in a variety of silencing pathways including DNA methylation, histone modification and RNA-mediated silencing, are differentially regulated in the *mop1* mutant complicates any facile explanation for the origins of these phenotypes. Unlike *rdr2* mutants in Arabidopsis, *ddm1* and *met1* mutants can have severe effects on plant morphology [33], and the maize homologs of both of these genes are down-regulated in the *mop1* SAMs. It is not clear whether or not the down-regulation in the maize *mop1* mutants of so many genes involved in epigenetic regulation is consequential, but it does suggest that many of the phenotypes that arise in *mop1*-containing lines over multiple generations may not be the direct result of the loss of RDR2 activity. The same may well be true for at least some of the differences in gene expression that we observed.

**Effect of RDR2 on the accumulation of TE-derived transcripts**

RNA-seq data identified hundreds of DNA TEs and retrotransposons that are differentially regulated in SAMs. Based on a model in which RDR2 generates 24 nt siRNAs that silence DNA TEs and retrotransposons, our expectation was that loss of RDR2 function in *mop1* SAMs would result in the up-regulation of DNA TEs and retrotransposons. Although we did observe that the majority of differentially expressed DNA TEs (78%) were up-regulated in the *mop1* mutant, less than half of all differentially regulated retrotransposons (32%) were up-regulated. This suggests that at least some DNA TEs and retrotransposons are silenced via distinct mechanisms.

**RDR2-dependent silencing of pericentromeric TEs**

Pericentic heterochromatin is rich in TEs in many species, and these sequences are typically heavily methylated and associated with large numbers of 24 nt siRNAs [2]. Consistent with its role in the RdDM pathway loss of RDR2 function results in the up-regulation of certain TEs, including *Huck* elements which are members of the Ty3/Gypsy-like super-family of retrotransposons. Fluorescence *in situ* hybridization (FISH) experiments reveal that although *Huck* elements are dispersed along all chromosomes they are significantly enriched in the vicinity of, but not in, centromeres [34]. Indeed, in general Ty3/Gypsy-like sequences cluster in pericentromeric regions across all grass species [35]. Our observations provide evidence that at least one pericentromeric repeat (i.e., *Huck*) is transcriptionally silenced via the RdDM pathway.

**RDR2-sensitive and RDR2-resistant 24 nt siRNAs**

In Arabidopsis RNA gel blot experiments, the population of 24 nt siRNAs is almost entirely eliminated in the *rdr2* mutant [5], indicating RDR2 is required for the biogenesis of nearly all 24 nt siRNA. In the maize *mop1* mutant, the population of 24 nt siRNAs is dramatically reduced, but not eliminated [18]. Via a re-
analysis of an existing small-RNA data set we identified >33,000 unique “RDR2-sensitive” and ~5,000 unique “RDR2-resistant” 24 nt siRNAs that are “down-regulated and “up-regulated” in the mop1 mutant, respectively.

RDR2–independent silencing of DNA TEs and retrotransposons

In contrast to elements such as Huck, the silencing of some types of DNA TEs and retrotransposons (e.g., most Prem1 elements) does not appear to require RDR2, as evidenced by the fact that they are down-regulated in the mop1 mutant. The hypothesis that an RDR2-independent heterochromatin silencing pathway exists in maize is consistent with our previous observation that many retrotransposon are significantly up-regulated (some >1,000×) in SAMs as compared to seedlings even though mop1 transcripts accumulate in SAMs to a level 100× higher than in seedlings. On the other hand, because new retrotransposon insertions are quite rare in maize [36], we considered the possibility that a significant proportion of the retrotransposon-derived transcripts we detected in SAMs are generated via RDR2 activity itself [37], which can produce aberrant non-polyadenylated RNAs. If this were the case, these species would indeed be lost in the RDR2 mutant (along with associated siRNAs). However, because the procedures we used to construct our RNA-seq libraries preferentially target mRNA species this possibility seems unlikely.

We therefore considered other RDR2-independent mechanisms for silencing DNA transposons and retrotransposons in a complex genome such as that of maize. Because the expression of many genes is affected by the mop1 mutant, it is possible, for example, that some of these effects could be antagonistic to the direct effects of mop1 on gene silencing. In addition, Lippman et al. [38] reported that the epigenetic inheritance of different TEs differed from mutant to mutant in Arabidopsis and proposed the existence of distinct but interacting pathways responsible for transposon silencing via siRNAs and histone modifications. Observations from fission yeast offer a plausible possibility for an RDR2-independent pathway. In this yeast, inhibition of histone deacetyltransferases causes an inherited loss of heterochromatin [39]. Several genes encoding histone deacetyltransferases (HDACs) were up-regulated in the mop1 mutant. It is possible that enhanced expression of these HDACs could enhance silencing of some TEs. Similarly, the reduction in expression of the maize orthologs of ROS1 and DME1, both of which are required for demethylation of a variety of target genes in Arabidopsis [40], could result in the silencing of a variety of genes in mop1 mutants. Hence, our observations in the mop1 mutant of the down-regulation of some DNA TEs and many retrotransposons, enhanced expression of genes in the HDAC silencing pathway, and decreased expression of genes in the demethylation pathway are consistent with the existence of multiple silencing mechanisms, but suggest that these mechanisms can potentially interact antagonistically.

Nobuta et al. reported that 22 nt small RNAs are highly abundant in the mop1 mutant [18] and suggested that these small RNAs may be the result of an alternative mode of heterochromatic siRNA production that
is independent of, and may even be enhanced by, the loss of RDR2. Alternatively, or in addition, the RDR2-independent silencing we observed could be the result of the RDR2-resistant 24 nt siRNAs we identified. As discussed in the Materials and Methods section, these RDR2-resistant 24 nt siRNAs may actually be produced at higher levels in the mop1 mutant. If this were the case, then these RDR2-resistant siRNAs could be responsible for the enhanced silencing of some of the DNA TEs and retrotransposons we observed in the mop1 mutant. If, on the other hand, the RDR2-resistant siRNAs are simply less susceptible to loss of RDR2 function, they would need to be more effective at silencing in the mop1 background to explain the enhanced silencing of DNA TEs and retrotransposons we observed. RDR2-resistant 24 nt siRNAs might, for example, exhibit enhanced repressive activity in response to changes in chromatin structure resulting from loss of RDR2 activity.

Potential sources of RDR2-resistant siRNAs include novel combinations of sense/anti-sense transcripts and transcribed inverted repeats. In maize retrotransposons are often present in vast nested arrays [36]. Enhanced transcription of these nested retrotransposons (due perhaps to loss of RDR2-dependent silencing) could result in the production of novel combinations of sense and antisense RNAs that could be processed into biologically active siRNAs even in the absence of RDR2. Thus, the effects of the mop1 mutant on a given transposon family may be a reflection not just of its sequence, but of the physical distribution of that family within the genome. With that in mind, it is interesting to note that the one family of DNA TEs with a high proportion of up-regulated members (CACTA elements) exhibits a distinct chromosomal distribution relative to the other families of DNA transposons. CACTA elements are significantly more likely to be found in gene-poor, heterochromatic regions of the genome than are all other DNA TEs [16].

In addition, dosage effect has been reported in the maize Activator/Dissociation (Ac/Ds) and Mutator transposon families [41]–[44], demonstrating that the regulation of the TE transposition is complex; similar complexity likely contributes to our observation of the down-regulation of TEs in the mop1 mutant. For example, some of the changes in expression of TEs and genes observed in this study could be due to epigenetic interactions between TEs and genes. In several cases it has been demonstrated that expression of TEs can reduce expression of nearby genes [45]–[49]. For example, de-repression of an LTR retrotransposon flanking the BONSAI gene in Arabidopsis in a ddm1 mutant background results in epigenetic silencing of BONSAI [50]. This process involves the production of BONSAI-specific siRNAs. Given our observation that DNA TEs, which tend to be preferentially located in gene-rich regions of the genome are likely to be up-regulated in the mop1 mutants, it is possible that many of the negative effects on gene expression are due to similar interactions between genes and nearby TEs. To more comprehensively analyze the relationship between levels of gene expression, siRNA production, and DNA methylation, it will be necessary to analyze all of these variables in a single tissue. Further, given the number of variables involved, a clear understanding of cause and effect relationships between RDR2 activity and expression will require detailed analyses of individual transposons, retrotransposons and genes.
MATERIALS AND METHODS

Genetic stocks, plant growth conditions, genotyping, and RNA-seq

The mop1-1 allele used in this study has been described previously [12]. This mutant is within a mixed genetic background, including both the highly inbred a1-mum2 minimal Mutator line [51] and the Mutator line from which mop1-1 was first derived [8]. The mop1-1 mutation in this background was maintained through several generations via sib crosses, self fertilizations, or back-crosses with the a1-mum2 stock. Although the progenitors of this line contained active MuDR elements, these elements were no longer present in the line used for this study, which lacked detectable Mutator activity.

This genetic background is distinct from that analyzed by Nobuta et al. [18]. Importantly, the family used in these experiments is closely related to a mop1 mutant lineage that gave rise to a large number of unique morphological phenotypes not previously observed in mop1 mutant plants (Lisch, unpublished observation). Given this, and given the dramatic differences in TE composition between maize inbred lines, direct comparisons of transcript data between the current data set and that of Nobuta et al. should be treated with caution.

A plant having the genotype mop1-1/mop1-1 was crossed to Mop1/mop1-1 heterozygote and the resulting progeny kernels planted in growth chambers (PGW-40, Percival Scientific, http://www.percival-scientific.com). Temperature and light cycles were set as 25 degrees for 15 hours of light and 20 degrees for 9 hours of dark. During the light period the light intensity at the surface of the growth medium was maintained between 650 and 800 umol m$^{-2}$ sec$^{-1}$.

At 14-days after planting SAMs were collected using the PALM MicroBeam System (115V Z, P.A.L.M. Microlaser Technologies, http://www.palm-microlaser.com). Plants homozygous and heterozygous for the mop1-1 mutant allele were distinguished using two pairs of primers: a pair of mop1-specific primers consisting of RDRF3 (sequence: 5'-TCTCCACCGCCCACCTTGAT-3') and RDRR2 (sequence: 5'-ATGGCCAGCCGAGGTGCAGAT-3') and a primer pair consisting of the Mutator TIR primer Mu-TIR (5'-AGAGAAGCCAACGCCAWCGCTCYATTCGTC-3') and the mop1-specific primer RDRF3. Twelve mop1-1/mop1-1 and ten Mop1/mop1-1 SAMs were used to form mop1 mutant and non-mutant pools. Collected SAM tissues were used for RNA extraction, RNA amplification and synthesis of double stranded cDNAs according to our previous published procedures [15]. These procedures preferentially target polyadenylated transcripts. Illumina/Solexa libraries were constructed using these double stranded cDNAs following Illumina/Solexa's standard protocol for genomic library preparation. The resulting libraries were sequenced on the Solexa 1G Genome Analyzer at the Michael Smith Genome Sciences Centre (Vancouver, BC, Canada). Each library was sequenced using 2 lanes on a Solexa flow cell. The Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo) accession number for the data used in the paper is GSE16789.
Alignments of RNA–seq to the maize reference genome and TEs

The resulting Solexa reads were aligned to the maize B73 reference genome (Release 4a.53) (http://www.maizesequence.org) with the short read aligner NOVOALIGN (http://www.novocraft.com) using 32 base sequences. Low quality bases located at the end of reads were trimmed and only reads that mapped uniquely to the genome with a maximum of two mismatches including insertion/deletion (indel) across 32 bases were used for subsequent analyses. The “filtered gene set”, a collection of high-quality gene models developed by the MGSP, was projected onto the B73 reference genome.

In addition, the Illumina/Solexa reads were also aligned directly to the DNA TE families and retrotransposon subfamilies. Due to the repetitive property of the TEs, each read is allowed to be mapped to multiple DNA TE families or retrotransposon subfamilies but each read is only counted once within each family or sub-family with same alignment criteria as used for alignments to the reference genome. The 386 chromatin-associated genes were mapped to the maize B73 reference genome using criteria of 95% identity and 90% coverage. Reads that uniquely mapped to the reference genome were projected onto each of these chromatin-associated genes allowing us to detect differential expression.

Identifying differential expression via a likelihood ratio test and Fisher's exact test

Two statistical procedures to identify differentially expressed genes were compared and evaluated: a likelihood ratio test based on a Poisson model (below) [52] and Fisher’s exact test. Although the two procedures produced similar p-values (R = 0.9; Figure S5), the Fisher’s exact test was more conservative. It was therefore selected for use in this study.

The likelihood ratio test analysis generally followed the procedure described in Marioni et al. [52]. For each gene, the number of reads from the mop1-1 mutant sample and the non-mutant sample were modeled as independent Poisson random variables with mean $\lambda_m C_m$ for mutant and mean $\lambda_n C_n$ for non-mutant, where $C_m$ and $C_n$ denote counts of the total number of mapped reads for the mutant (2,156,241) and non-mutant (3,248,869) samples, respectively. It is straightforward to show that the likelihood ratio statistic for testing the null hypothesis of $H_0$: $\lambda_m = \lambda_n$ is $T = 2\{k_m \log (k_m/C_m)+k_n \log(k_n/C_n)−k \log(k/C)\}$, where $k_m$ is the number of mutant reads for the gene in question, $k_n$ is the number of non-mutant reads for the gene in question, $k = k_m+k_n$, and $C = C_m+C_n$. This statistic is distributed approximately as a Chi-square random variable with 1 degree of freedom when the null hypothesis of no differential expression is true. Thus, p-values were obtained by comparing the observed statistic for each gene to the Chi-square distribution with 1 degree of freedom. Ideally, sequencing would have been carried out separately for multiple independent biological replications of each genotype so that over dispersion relative to the Poisson distribution could have been assessed and accounted for using an analysis like that proposed by Robinson and Smyth [53]. Note that our qRT-PCR validation and analysis (discussed below) was based on separate measurements of independent biological replications.
Detection of RDR2–sensitive and –resistant 24 nt siRNAs
Each species of 24 nt siRNAs was tested whether the proportions in the library between mop1 mutant and non-mutant were significantly different via Fisher's exact test. Because we are able to measure only the abundance of each species in a genotype relative to the total number of reads for that genotype, it is difficult to formally distinguish 24 nt siRNAs that are up-regulated in the mutant from those that make up a significantly greater proportion of the observed reads only because of the absence of many other 24 nt siRNA species in the mop1 mutant, thereby making them proportionately more abundant. For the purposes of this study we therefore carefully define up-regulation to mean that a particular species is significantly more abundant in one sample of reads than in another. It is important to note that this does not necessarily mean that the number of RNA molecules of that particular species increases on a per cell basis.

qRT–PCR validation and data analysis
Primer design for qRT-PCR was conducted as described [54]. RNA samples independent from those used in the RNA-seq experiment were extracted from four biological replications from mop1-1/mop1-1 and Mop1/mop1-1 (3 SAMs pooled within each replicate per genotype) by using the same procedure as the RNA-seq experiment. To prepare the cDNA template, combined oligo(dT) and random hexamers was used to perform reverse transcription reactions at 55°C for 1 hour with SuperScript III. A reverse transcription without SuperScript III was conducted to control for genomic DNA contamination. qRT-PCR was conducted on an Mx4000 multiplex quantitative PCR system (Stratagene). RNA from a human gene (GenBank accession no. AA418251) was spiked into each reaction as an external reference for data normalization. Genotype-specific Ct values for each gene and control were calculated and then the ΔΔCt was computed. For each gene, ΔΔCt across 4 biological replications was used to conduct a t-test in R (www.r-project.org) [55].

ACKNOWLEDGMENTS
We thank the Maize Genome Sequence Project (NSF DBI-0527192) and the Maize Transposon Consortium (NSF DBI-0607123) for sharing genome sequences and annotation prior to publication; other members of the Maize Shoot Meristem Project (NSF DBI-0820610) and Dr. Gernot Presting for thoughtful discussions; and Schnable Lab members Ms. Marianne Smith for wet lab support, Mr. Cheng-Ting Yeh for computational support, and Sanzhen Liu and Dr. An-Ping Hsia for thoughtful discussions.

REFERENCES
TABLE 1
Expression patterns from diverse super-families of TEs

<table>
<thead>
<tr>
<th>TE Class</th>
<th>Super-family</th>
<th>No. sub-families/families</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>I</td>
<td>RIL</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>RIX</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>RLC</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>RLG</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td>RLX</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>608</td>
</tr>
<tr>
<td>II</td>
<td>CACTA</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>hAT</td>
<td>228</td>
</tr>
<tr>
<td></td>
<td>Helitron</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Mariner</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>MULE</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>PIF/Harbinger</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Stowaway</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>Tourist</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>797</td>
</tr>
</tbody>
</table>

*RIL, LINE (L1) retrotransposons; RIX, unknown LINE retrotransposons; RLC, Ty1/Copia, LTR retrotransposons; RLG, Ty3/Gypsy, LTR retrotransposons; RLX, unknown LTR retrotransposons; sub-families for class I TEs; families for class II TEs; Up, number of up-regulated sub-families/families; Down, number of down-regulated sub-families/families; Up+down, total number of differentially expressed sub-families/families.
### TABLE 2
Differentially expressed genes in RdDM pathway

<table>
<thead>
<tr>
<th>Chromdb.ID</th>
<th>Gene Name</th>
<th>log2(FC)</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCL102</td>
<td>dcl3a</td>
<td>1.2</td>
<td>8e-06</td>
</tr>
<tr>
<td>DCL104</td>
<td>dcl3b</td>
<td>-0.5</td>
<td>1e-02</td>
</tr>
<tr>
<td>AGO104</td>
<td>ago4a</td>
<td>-0.7</td>
<td>9e-21</td>
</tr>
<tr>
<td>AGO105</td>
<td>ago4b</td>
<td>-1.7</td>
<td>6e-17</td>
</tr>
<tr>
<td>AGO119</td>
<td>ago4c</td>
<td>-0.5</td>
<td>1e-06</td>
</tr>
<tr>
<td>CHR127</td>
<td>drd1</td>
<td>-1.3</td>
<td>1e-11</td>
</tr>
<tr>
<td>DMT101</td>
<td>met1</td>
<td>-1.0</td>
<td>3e-11</td>
</tr>
<tr>
<td>DMT106</td>
<td>drm1/2</td>
<td>-0.7</td>
<td>3e-13</td>
</tr>
<tr>
<td>DMT102</td>
<td>cmt3a</td>
<td>-1.0</td>
<td>3e-40</td>
</tr>
<tr>
<td>DMT105</td>
<td>cmt3b</td>
<td>-0.7</td>
<td>4e-22</td>
</tr>
<tr>
<td>SDG118</td>
<td>suvh4/kyp</td>
<td>-1.0</td>
<td>4e-16</td>
</tr>
<tr>
<td>HDA108</td>
<td>hda6</td>
<td>-0.7</td>
<td>2e-09</td>
</tr>
<tr>
<td>CHR101</td>
<td>ddm1</td>
<td>-0.5</td>
<td>8e-04</td>
</tr>
<tr>
<td>DNG101</td>
<td>dme1</td>
<td>-1.0</td>
<td>4e-12</td>
</tr>
<tr>
<td>DNG103</td>
<td>dme2</td>
<td>-1.7</td>
<td>9e-26</td>
</tr>
</tbody>
</table>

\(^a\) ID used in Chromdb (http://www.chromdb.org); \(^b\) ID used in general RdDM pathway [5]; \(^c\) log2 transformation of fold change as the relative abundance of transcripts in mutants/non-mutants. Positive value indicates the up regulation and negative value indicates down regulation; \(^d\) The false discovery rate calculated using Benjamini and Hochberg’s procedure [19] for the p-value from Fisher’s exact test.
TABLE 3

Key regulators of SAM developments showing differential expression

<table>
<thead>
<tr>
<th>Gene.ID</th>
<th>log2(FC)</th>
<th>FDR</th>
<th>SwissProt.ID</th>
<th>Protein Name</th>
<th>E-value</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G104925</td>
<td>-0.76</td>
<td>5.00e-04</td>
<td>Q940E8</td>
<td>Fasciated ear2</td>
<td>0</td>
<td>[19]</td>
</tr>
<tr>
<td>GRMZM2G085113</td>
<td>-0.86</td>
<td>1.62e-09</td>
<td>A9XIW7</td>
<td>Terminal ear1-like 2 protein</td>
<td>6e-138</td>
<td>[20]</td>
</tr>
<tr>
<td>GRMZM2G123140</td>
<td>-0.50</td>
<td>1.05e-02</td>
<td>B3GW90</td>
<td>Putative HD-ZIP IV family transcription factor OCL4</td>
<td>0</td>
<td>[21]</td>
</tr>
<tr>
<td>GRMZM2G087741</td>
<td>0.50</td>
<td>6.09e-06</td>
<td>Q9SYT6</td>
<td>Knotted class 1 homeodomain protein liguleless3</td>
<td>0</td>
<td>[22]</td>
</tr>
</tbody>
</table>

* Refer to [http://www.maizesequence.org](http://www.maizesequence.org); ^log₂ transformation of fold change as the relative abundance of transcripts in mutants/non-mutants. Positive values indicate up regulation and negative values indicate down regulation; ³False discovery rate calculated using Benjamini and Hochberg’s procedure [19] for the p-value from Fisher’s exact test; ⁴Protein ID retrieved from SwissProt_Trembl database via blastp (1e-10 as cutoff); ⁵E-value from blastp search.
FIGURE LEGENDS

Figure 1. Comparison of SAM morphologies between mop1 mutants and non-mutants. Safranin O/FastGreen stained image of SAMs from (A) a mop1/mop1 mutant and (B) a non-mutant sibling. See Table S1 for a quantitative analysis.

Figure 2. Validation of eight differentially expressed genes via real-time PCR. Eight differentially expressed genes were chosen from the RdDM pathway (ago4a [Chromdb ID: AGO104], ago4b [Chromdb ID: AGO105], ago4c [Chromdb ID: AGO119], ddm1 [Chromdb ID: CHR101] and met1 [Chromdb ID: DMT101]), transposons (hAT [a member of the ZM_hAT_8 sub-family; http://www.maizesequence.org]) and MULE [a member of the MULE sub-family DTM_Zm33205;http://www.maizesequence.org]) and a regulator of SAM development (liguleless3 [Gene model ID: GRMZM2G087741; http://www.maizesequence.org]) for qRT-PCR validation. Primers used for qRT-PCR are presented in Table S12. Fold change was presented as the relative abundance of transcript in the mop1 mutant/non-mutant. The quantitative fold changes obtained from between RNA-seq and qRT-PCR experiments were significantly correlated (Pearson correlation coefficient was 0.94, $r^2 = 0.88$, p-value = 0.0005). A t-test of equal expression between the mutant and non-mutant using the data from four biological replications of qRT-PCR were conducted (p-value <=0.05, *; p-value <=0.01, **).
Figure 2 Jia et al., 2009
CHAPTER 4. NATURAL ANTISENSE TRANSCRIPTS CO-LOCALIZE WITH OPEN
CHROMATIN MARKS AND SMALL RNAS AND ARE REGULATED BY A COMPLEX
GENETIC ARCHITECTURE IN MAIZE

A paper to be submitted to Genome Research
Yi Jia, Ruth A. Swanson-Wagner, Sanzhen Liu, Xuefeng Zhao, Dan Nettleton, Patrick S. Schnable

ABSTRACT
Differential gene expression is ultimately responsible for any phenotypic variation. Recent studies have shown that genome transcription is regulated by the interplay of multiple factors, such as transcriptional factors, non-coding RNAs, epigenetic modifications, etc. Natural antisense transcripts (NATs) are widespread in various organisms and many of them have been reported to play important roles for gene expression under specific biological contexts. To explore the possible functions of NATs in crops, we used maize as crop model and identified genes accumulating NATs by both bioinformatics and strand-specific RNA-seq approaches. We have conservatively estimated that almost 40% of maize genes have their antisense partners. Interestingly, the NAT accumulation showed strong positive correlation with multiple open chromatin modification marks but negative correlation with close chromatin modification marks, which suggests the possible regulatory role of NATs via epigenetic process. Importantly, expression analysis demonstrated that NATs significantly accumulated between maize inbred lines and consequently caused the significant change of ratio of sense versus antisense transcripts between genotypes. To elucidate the genetic control of antisense transcript accumulation, we have conducted the first NAT expression QTL (NAT-eQTL) experiment to our best knowledge. We have found the antisense transcript accumulation can be regulated by cis- or trans-eQTL and some sense and antisense transcript accumulation can be regulated by the same trans-eQTL. Based on these findings and the fact that NATs can play important regulatory functions, we hypothesize that the complex NAT regulatory networking and the interactions between sense and antisense transcripts in different genotypes may contribute to novel gene expression in the hybrids and consequently heterosis.

INTRODUCTION
Alteration of gene expression patterns is ultimately responsible for any phenotypic variation. Hundreds of microarray and RNA-seq experiments had been published to detect the variations of gene expression between different biological samples of interest. Such gene expression variations can be regulated by genetic or epigenetic factors, which can be classified as either local (or cis-) regulatory elements or distant (or trans-acting) factors (1). Recent advances of genomic technologies are extending our capabilities searching and identifying such functional elements or polymorphisms contributing to the regulation or alteration of gene expression (2-4). For example, chromatin immunoprecipitation-sequencing (CHIP-seq) uses specific antibodies to generate the respective
sequencing libraries for identifying the DNA sequences interacting with the functional proteins in genome-wide (4, 5). Structure variation is the main factor contributing the gene expression alteration between different genotypes. Currently, all types of structure variations, such as SNPs (6), present and absent variation (7), copy number variations (8), can be discovered and typed in an unprecedented efficiency. To identify and associate the local or distant regulators with the variations of gene expression, the expression QTL (eQTL) analysis has been proven an effective approach to associate the expression variation of a specific gene with some chromosome regions or segments (1, 2, 9). Another exciting field for gene expression regulation is the epigenetic modifications and their roles for gene expression. By combining biochemical pretreatment of genomic DNA samples and 2nd generation sequencing or genome tilling array hybridizations, epigenome landscapes for several organisms, such as Arabidopsis, rice, maize and human, have been published (5, 10-14). All of these findings and resources facilitate our understanding of gene expression regulation in genome wide to a much higher level.

RNA molecule was mostly considered as the bridge between the DNA and protein in the central dogma. However, the discovery of regulatory RNAs has largely expanded the scope of RNA function and regulation of gene expression. For example, microRNAs have been found to be the important regulators for gene expression in different developmental stages for plants and animals (15, 16). Small interfering RNAs (siRNAs) are responsible for silencing heterochromatin consisting of transposons and other repetitive elements (17). The searching of novel regulatory RNA species is adding our knowledge about the RNA regulation network by discovering new regulatory RNA molecules. Endogenous natural antisense transcripts (NATs) are one of the large classes of newly identified regulatory RNAs (10, 18, 19).

NATs can be defined as the complementary transcripts to the sense transcripts. There are two types of NATs. A cis-encoded NAT is transcribed from the same genomic locus as the sense transcript but from the opposite strand. In contrast, a trans-encoded NAT is transcribed from a different genomic locus compared to the sense transcript. NATs have been shown to play important roles in siRNA biogenesis (20), epigenetic gene silencing (21), transcriptional interference (22), RNA editing (23) and splicing (24) and are widespread observed in many eukaryotes and implicated in a wide variety of biological pathways. Recent genome-wide analyses indicate that NATs accumulate in human (25, 26), mouse (18, 27), rice (28, 29), Arabidopsis (10, 30, 31). Chen et al. (32) reported computational evidence that up to 22% of human transcripts are members of cis-encoded sense-antisense (S/A) pairs. In mouse, more than 72% of all genome-mapped transcription units overlap with cDNAs that map to the opposite strand (18). In plants, approximately 30% of annotated Arabidopsis genes have been shown via hybridization experiments to accumulate NATs (33). These studies delivered valuable overall insights of NATs in various organisms. However, several questions regarding to NATs remain to be addressed though we have realized the widespread and potential regulation functions of NATs: 1) Current genome-wide NATs studies in plants were conducted in rice [29] and Arabidopsis [33], both of which has a small genomes. How is the NAT
expressed in a more complex plant genome, such as maize? 2) The gene expression of sense and antisense transcripts have been widely investigated and found that diverse expression patterns exists for the sense and antisense pairs, including anti-correlation and co-expression of similar level. However, is there significant alteration of NAT accumulations between different genotypes? Furthermore, is there interaction of sense and antisense pairs between different genotypes? 3) NATs have been demonstrated to play a key role in small RNA biogenesis. Considering the functions of small RNAs in epigenetic regulations of genome stability, is there any correlation or relationship between NATs and epigenetic modifications? 4) Accounting for the importance of NAT functions, how are NATs expression regulated?

To answer these questions, maize is selected as a complex crop model in this study due to several advantages: 1) Maize has a relative complex sequenced genome comparable to human with rich resources of EST, full-length cDNA sequence resources. These resource will make maize a possible system to study how NATs express in the complex plant genome; 2) Maize exhibits high diversity of both gene expression and genotypes (7, 34), which facilitates to study the genotype effect on NAT expression; 3) Maize is a good system for epigenetic regulation study. For example, transposon silencing and paramutation were actively studied in maize genome (17); 4) Maize have been used as a model for crop domestication research model and several important functional genes identification related to the domestication (35); 5) Maize is one important crops and NATs have not been investigated in maize in genome wide.

In this study, we developed one stringent bioinformatics procedure and strand-specific RNA-seq to identify cis- or trans- NATs. Importantly, both bioinformatics and experimental procedures were designed to avoid the possible NAT artifacts reported previously (36, 37). Furthermore, not only the sense transcript but also the antisense transcripts differed significantly between the two inbreds, which is consistent with our previous hypothesis that NATs may contribute to differences in transcript abundance in F1 hybrids relative to their inbred parents (34). Due to the NAT expression controlled genetically, the first NAT-eQTL experiments had been conducted to identify the genomic intervals associated with the variation of NAT expression.

RESULTS
Bioinformatic identification of maize genes that accumulate cis- and trans-NATs
In this study, NATs are defined as strand-specific ESTs that overlap gene models in opposite orientations with overlap lengths >=75 bp and identity >= 95%. Genes that accumulte NATs are defined as NAT genes. Astringent bioinformatics pipeline was developed to identify NATs and NAT genes, (Methods; Figure S1). First, a total of 2,018,337 maize cDNA sequences downloaded from GenBank in early January 2009 were trimmed to remove vector and bacterial sequence contamination; second, strand-specific ESTs were identified using strict criteria (Methods); finally, the identified strand-specific ESTs were aligned to the filtered gene set (Methods). Using this procedure, a total of 836,704 strand-specific ESTs were identified.
and 10,924 (39% of 27,810 assayed genes) NAT genes were identified. Of the 10,924 NAT genes, 7,179 (26% of 27,810 assayed genes) overlap with cis-NATs, 5,384 (19% of 27,810 assayed genes) overlap with trans-NATs and 1,639 (6% of assayed genes) overlap with both cis- and trans-NATs.

**Identification of NAT genes via strand-specific RNA-seq**

To experimentally identify NATs in the maize genome, strand-specific RNA-seq (ssRNA-seq) libraries were constructed from B73 and Mo17 seedlings following a published protocol with modifications (Methods; Figure S2). In total, 11.6 million (M) and 12.7 M RNA-seq reads were generated from the B73 and Mo17 libraries, respectively (Table 1). More than 8 M reads from each library could be uniquely mapped to the reference genome (Methods). Due to the strand-specific nature of these RNA-seq reads, those that mapped to the Filtered Gene Set can be assigned as being either sense or antisense reads by referencing the annotated gene orientation. Among the uniquely mapped reads, about 6 M (~75%) reads from each library could be mapped to the filtered gene set (38) Among the RNA-seq reads that mapped to the Filtered Gene Set, 97% mapped to the sense strand and 3% mapped to antisense strand (Table 1).

A Chi-square test was conducted to test whether the two sets of NAT genes detected via *in silico* searching and experimental identification are significantly correlated. The test results showed two sets of NAT genes are highly correlated and agreed with each other (p value =5.9e-13; Table S1). It must be noted that only those RNA-seq reads that uniquely mapped to the reference genome were included in this analysis. NATs in our study are defined the transcripts overlapped with sense transcript >=75 bp and the identity >=95%. Thus, these criteria force the NATs detected in the RNA-seq experiment to be cis-NATs. So the estimated proportion of genes accumulating NATs is 24% (6,178/27,810). As for the bioinformatics searching method, we still required the strand-specific ESTs uniquely mapped to the reference genome. Due to the longer length of the strand-specific ESTs, it is possible that a sub-portion of a strand-specific ESTs overlaps with a gene but located in different chromosome. Thus, the predicted NAT genes can overlap with either cis- or trans-NATs. It is worth pointing out among the 10,924 NAT genes detected from public ESTs, 1,639 (15%) genes can accumulate both and antisense transcripts. And our stringent bioinformatic pipeline will cause underestimate of NAT genes in maize since we did not consider the ESTs or RNA-seq reads if they are mapped to multiple genomic locations.

**Functional annotation of maize NAT genes**

We used Fisher’s exact test to determine whether specific functional categories are enriched among the NAT genes detected via the bioinformatic pipeline and the ssRNA-seq experiment. Two GO terms, lipoxygenase activity and monoxygenase activity, were enriched in both data sets (Table 2). Interestingly, RNA splicing and spliceosome categories were enriched in the cis/trans- NAT genes, but not enriched among the cis-NAT gene set, suggesting that antisense transcripts can potentially regulate the functions of genes related to RNA splicing and consequently affect the RNA splicing of other genes. In
addition, the transferase activity and light harvesting GO terms enriched in our cis/trans-NAT gene sets were also over-represented in the Arabidopsis genes accumulating trans-NATs (39), suggesting a conservation of NATs across species.

**Expression analysis of sense and antisense via ssRNA-seq and microarray**

To determine whether variation in the accumulation of NATs varies by genotype, the expression levels of NATs identified via ssRNA-seq were compared from B73 and Mo17 using Fisher’s exact test. Using an estimated false discovery rate (FDR) of 5%, ~10% (605/6,178) of genes show significant differences in the accumulation of NATs between B73 and Mo17. For comparison, when the same test was used to analyze sense strand transcripts, more than 40% of genes exhibit significant differences in accumulation between the two maize inbred lines.

To validate and extend these analyses, we also measured the accumulation of sense and antisense transcripts using a custom strand-specific oligonucleotide microarray (Methods). Our microarray experiment included two special features that allowed it to detect NATs. First, to independently detect sense and antisense transcripts the array contained complementary sense and antisense oligos that were printed side by side in two different sectors of the array. Second, a direct RNA labeling technique was used to avoid the production of antisense artifacts (36, 37). Among the 4,608 complementary oligo pairs on the microarray, 2,920 (63%) can be uniquely mapped to the B73refv1 genome (38) using the stringent criteria of 85% identity and 85% coverage to rule out the oligos cross hybridizing to multiple highly similar RNA molecules. Of the 2,920 uniquely mapped oligo pairs, 2,420 (83%) can be mapped to the annotated filtered gene set (38). Of 2,420 oligo pairs, 182 (8%) sense oligos and 50 (2%) antisense oligos can detect significant changes of transcript accumulation (5% FDR cutoff). Importantly, such strand-specific expression pattern detected by microarray was significantly correlated with the strand-specific expression pattern in RNA-seq experiment (p value=6.3e-11).

**Interaction between genotypes and S/A transcript pairs**

Endogenous antisense transcripts have been reported to interact with sense transcripts (20). We have shown that both sense and antisense transcripts can independently accumulate to significantly different levels between B73 and Mo17. To check whether the difference of sense and antisense accumulation level is significantly increase or decrease between genotypes, indicating the interaction between transcription orientation and genotypes, we modeled the observed number of counts for each gene on every strand as a Poisson random variable and conducted a likelihood ratio test to test a null hypothesis of equal ratio of sense and antisense transcript accumulation between the two genotypes (Methods). Among the 5,341 genes with at least 1 read from each genotype on either sense or antisense strand, 465 showed significant differences in the ratio of sense and antisense transcripts between the two maize genotypes. Such differences can arise if there is significant change in the accumulation of sense transcripts, antisense
transcripts or both strands between the two genotypes. To classify all these possibilities, 2,649 genes showing at least one significant change among the three possibilities, were classified into seven groups (Table 3). For examples, of the 2,649 genes, 465 genes exhibited significant differences of sense and antisense accumulation between B73 and Mo17, in which there are 268 (58%) genes showed significant changes of antisense transcript accumulation between the two genotypes. This result suggested that the change of antisense transcript accumulation can regulate the sense transcript accumulation differently between different genotypes and consequently novel gene expression pattern could be produced in hybrid derived from the two genotypes due to such interaction effects.

NATs preferentially align to sequences upstream of transcription start sites and downstream of transcription stop sites

It has previously been reported that NATs can align to different portions of genes, e.g. UTRs, coding regions or introns (20, 25, 40). To determine the distribution of NATs on maize genes, we used a set of high-quality, full-length cDNAs generated from the maize full-length cDNA project (Methods). The strand-specific manner of the RNA-seq reads allowed us to map the reads to specific strands not only within the full-length cDNA but also the upstream and downstream of the full-length cDNAs. Each gene for which a full-length cDNA was available was divided into different sections, including the promoter (1kb upstream from the TSS), 5’ UTR, coding region, introns, 3’ UTR and terminator (1kb downstream of the polyA site). For each gene the numbers of sense and antisense RNA-seq reads that mapped to these different gene sections were counted. sum up together for each section. To check the distribution along the different sections of genes, the total sum counts of reads in each section was normalized to the number of reads per million bases (Mb). Interestingly, the proportion of antisense RNA-seq reads in the non-transcribed regions (i.e., promoter and terminator regions) was much higher than within transcribed regions (5’ UTR, coding region, introns and 3’ UTR) Furthermore, the frequency of NATs in terminators was much higher than in promoters (Figure 1). Some sense RNA-seq reads map to upstream of the TSSs but in lower frequency relative to the gene body regions. These observations demonstrate that there is active transcription upstream of these genes.

Relationship between NATs and epigenetic modifications

Gene expressions can be impacted by different epigenetic modifications, such as H3K36me3, H3K9ac, H3K4me3, H3K27me3 and DNA methylation. Some epigenetic modification marks the status of open chromatin (H3K36me3, H3K9ac, H3K4me3) and others closed chromatin (H3K27me3 and heavy DNA methylation). Wang et al had previously reported the epigenetic landscape in maize B73 shoots and roots (5). In this study, correlation analysis between each type of epigenetic modification and NATs has been conducted across all 32,540 filtered genes. Specifically, we asked the question whether the probability that genes with a given epigenetic modification accumulate NATs is more than by chance. Furthermore, the number of genes with both NAT accumulation and open chromatin modification are significantly higher
than what would be observed by chance (all p value < 2.2e-16). In contrast, the opposite pattern was observed between the NAT accumulation and closed chromatin modification mark H3K27me3. By mapping the NAT reads and open chromatin modification reads to a set of full-length cDNA set, 5' -UTR regions were found to accumulate the NATs and open chromatin modification reads more frequently than the rest of the transcribed regions (Figure 2).

**Relationship between NATs and small RNAs**

NATs have been shown to function in the biogenesis of small RNAs to response to the abiotic stress. Wang et al. previously published small RNA data derived from the seedlings of B73 grown in a very similar environmental conditions (5). To check whether there is correlation between the NATs and small RNAs, we have integrated these small RNA data into this analysis. First, we mapped the 21, 22 and 24nt small RNAs back to the maize reference genome by requiring perfectly match to unique genomic location. Those mapped small RNAs were then projected to the maize filtered gene set for the correlation analysis with NATs. Chi-square test was performed to check if the genes accumulating NATs were more likely to match 21 nt small RNAs. The similar analysis was also conducted for 22 and 24nt small RNA. Interestingly, all such three Chi-square test indicated NATs are highly correlated with small RNAs (all p value < 2.2e-16). Furthermore, 24 nt small RNAs were more distributed more frequently in the upstream and downstream of the transcribed regions, which is similar with the distribution pattern of NATs on the genes. This consistent observation may suggest the potential regulatory functions of NATs were related to 24 nt small RNAs.

**Genetic control of NATs expression**

We have shown that the accumulation of individual NATs differs between the B73 and Mo17 inbreds. This is significant given that NATs can have regulatory roles [20]. We used an eQTL approach to dissect the genetic regulation of these differences in NAT accumulation. Using our strand-specific oligo array the accumulation of strand-specific transcripts was measured in an intermated B73xMo17 population of recombination inbred lines (RILs) and their hybrids obtained by crossing each RIL per se to their parents with two biological replications (Methods). The advantage of this design is to have the opportunity studying the effect of heterozygous and homozygous genotypes on the gene expression in the hybrids. In total, 1064 IDP markers were used for the eQTL mapping analysis (9). Our interest is to map the genomic intervals controlling the antisense transcript accumulation. Thus the 4,840 oligos uniquely mapped onto the gene models were used to measure the expression traits in the eQTL analysis. We used the multiple interval mapping method to identify eQTL controlling the accumulation of NATs (41). Permutation tests were used as described to control the false QTL discovery rate in this genome-wide analysis (41). The eQTL analysis resulted in 815 non-redundant significant eQTL and 4 epistatic QTL controlling sense and antisense expression with 22% estimated FDR across three cross-types. Because all the analyzed oligos can be projected to the filtered gene set, for a given association, the corresponding eQTL can be designated as being either a sense or an antisense eQTL. Of the 815 non-redundant eQTL, 396 (49%) control the
accumulation of sense transcripts and 419 (51%) control the accumulation of antisense transcripts. In addition, eQTL can also be classified using the relative genetic positions of a regulated gene and the associated genetic marker. Consistent with prior work (9) we define trans-eQTL as those cases in which the regulated gene and the significant marker are located on different chromosomes; cis-eQTL as those cases in which the genetic distance between the regulated gene and the significant marker is less than 15 cM; the remainder of the eQTL are defined as “others”. Among the 815 eQTL, 681 (84%) are trans-eQTL (298 sense eQTL versus 383 NAT eQTL) and 41 (5%) are cis-eQTL (37 sense eQTL versus 4 NAT eQTL). Visualization of the eQTL locations across the genome reveals several eQTL hotspots (Figures).

**DISCUSSION**

**Nearly 40% of maize genes accumulate NATs**

NATs have been shown to function in various biological pathways. In summary, bioinformatics and wet-lab experimental approaches can be used to identify NATs and estimate their frequencies in genome wide in diverse mammals and plants. For example, alignments among human polyadenylated transcripts established that between 8%-22% of transcripts are members of S/A pairs in human (26, 32, 42). The analysis of transcription units and EST in mouse concluded that 72% of the transcriptional units accumulate NATs (18). In plants, transcriptomic hybridizations to a whole-genome array established that NATs accumulate for ~30% of annotated Arabidopsis genes (33). Similarly, microarray analyses conducted by Ma et al (43) established that NATs accumulate for 6-14% of maize genes with probes designed using coding sequences, depending on which organ was analyzed. Recently, the possible NAT artifacts introduced during the cDNA synthesis via reverse transcription had been realized in several reports (36, 37). Since this kind of artifact is most likely produced in the late stage of reverse transcription by using the synthesized cDNA as template through folding back structure, the synthesized cDNAs have long tails that cannot align to the genome (36). In this study, both our wet-lab and bioinformatics procedures for NAT discovery have been designed to exclude such artifacts. In the strand-specific RNA experiment, the RNAs were fragmented to 100-120 nt and ligated with the RNA adaptors with unused end blocked followed the cDNA synthesis. Three facts from the RNA-seq procedures make it impossible of NAT artifacts: 1) The fragment of RNA is short, which is relative difficult to form the fold back structure; 2) The both end of RNAs with 5' and 3' RNA adaptors are blocked for further enzymatic reaction; 3) Fixed sequencing primer were used to read the first 75 bp of original RNA molecules (out of >100 nt RNAs) no matter whether there is a tail or not. For microarray hybridization conducted for the maize inbred lines of differential expression and RILs of eQTL analysis, the hybridized probes onto the microarray were all derived from direct RNA labeling without any reverse transcription reactions using the RNA direct labeling technology (44). Our bioinformatics procedures identifying the antisense transcripts are also designed to minimize the possibility by requesting splicing site signals or long coding for strand-specific orientation plus tail smaller than 10 bp.
The bioinformatics searching approach identified 10,924 (39%) out of 27,810 assayed genes that can accumulate NATs, representing the overall proportion of maize genes accumulating NATs. However, this proportion is probably largely underestimated due to several facts. First, though ~0.8 million strand-specific ESTs were identified for NAT discovery, a large proportion of NATs could be still missed in the current maize EST collection due to the incomplete sampling; second, repetitive sequences are almost everywhere in the maize genome (38). Repetitive elements in the genome can be transcribed from different orientations and form the sense and antisense pairs (19) (45). In this study, all the repetitive sequences were discarded for unambiguous analysis; lastly, only strand-specific ESTs were used in the NAT identification. Similarly, we used the strand-specific RNA-seq data to estimate proportion of cis-NAT as 24%, which is also underestimated due to the similar reasons. Overall, we found that in seedling RNA NATs can anneal to almost 40% of maize genes, a fraction that is substantially higher than the proportion (6-14%) reported for maize by Ma et al. (43). This significant difference could be largely due to the high sensitivity of sequencing approach used by us. In addition, the low proportion of NATs detected by Ma et al could be due to the probe design in their experiment. The antisense probes used in Ma et al. were designed based on coding regions of maize sense transcripts. This lower proportion observed in Ma et al. is consistent with our finding that the NATs is more frequently in the 5’ and 3’ end of the genes but lower frequent in the coding regions. Furthermore, we found the NATs were distributed in much more frequent in the upstream and downstream of the transcribed regions (see discussion below). Kapranov et al. (23) identified many short RNAs (<200 nt) that map to the 5’ and 3’ ends of human genes. Based on these findings, they suggested a model for genome organization in which protein-coding genes are at the center of a complex network of overlapping sense and antisense transcription that provides the potential for both cis- and trans-regulatory mechanisms. Our finding is consistent with this model.

**Interplay of DNA, RNA and epigenetic modification around genic regions**

By projecting the uniquely mapped RNA-seq reads to the upstream, transcribed and downstream regions of the annotated maize gene models, we can determine the antisense RNA-seq reads and consequently how the NATs distribute within and flanking the genes. In this study, we have found that the antisense RNA-seq reads were non-randomly distributed within the genes. Within the transcribed regions, the normalized frequency of NAT between intron and exon is similar as the sense RNA-seq reads. However, much more antisense RNA-seq reads were observed in the upstream and downstream of the transcribed regions compared to sense RNA-seq reads only few proportion on these regions. This overall distribution pattern of NATs in the gene flanking regions is similar as the pattern reported in the human antisense transcriptome. This consistent observation suggested the possible universal regulatory role of NATs for the gene expression. However, we found that the relative frequency of NATs in upstream versus downstream show opposite pattern by comparing maize and human, which may suggest the different regulatory behaviors between mammals and plants.
To further explore the relationship of NATs with other regulatory components, the epigenetic landscapes and small RNA sequences in maize of same type of tissues and growth stage was integrated in this study (5). We first conducted the correlation analysis between NATs and different epigenetic modifications and different classes of small RNAs in a gene basis. Interestingly, we found the genes accumulating NATs is highly positively correlated with the genes accumulating the open chromatin modification sequencing reads but negatively correlated with the closed chromatin modification sequencing reads, suggesting that most of genes accumulating NATs are actively transcribed. In addition, we also found strong positive correlation between the genes accumulating NATs and genes matching 21, 22 and 24 nt small RNAs, respectively. All the small RNAs used in the correlation analysis are uniquely mapped to the genome. Furthermore, Wang et al reported the 21 and 24 nt siRNAs of maize show peaks in the upstream and downstream of the transcribed regions (5). This consistent distribution patterns between NATs and small RNAs may indicate the tight relationship of the two types of molecules. In Arabidopsis, the pairs of sense and antisense transcripts can be the templates for the biogenesis of 24 nt small RNAs that is accompanied with production of other small RNAs, such as 21 nt small RNAs (20). Overall, the positive correlation of NATs and small RNAs and open chromatin modification make one tempt to hypothesize that NATs can recruit the open chromatin remodeling complex via small RNAs. This hypothesis is at least supported by two different serial of evidence. First, Janowski et al reported the duplex RNAs (19-21nt) targeting to specific promoter regions can activate gene expression in mammalian cells (46). Second, several studies have reported that the antisense transcripts in the V_H regions of B cell IgH locus of mammalian immune system can recruit remodeling complex to open the chromatin structure for recombination (47, 48). However, the direct evidence of the biogenesis of small RNAs using NATs in the upstream and downstream of transcribed regions needs to be achieved and how the small RNAs possible to recruit the open chromatin remodel complex remain to be elucidated.

**Ratios of sense/antisense transcripts exhibit genotype-dependent patterns**

From the RNA-seq experiment, more than 40% of genes were identified differentially expressed from the sense strand. Because maize inbreds vary in their gene content and organization (31) we hypothesized that they may also vary in their ability to accumulate NATs. In deed, the NAT accumulation levels of 10% (605) genes showed significantly difference between B73 and Mo17. Furthermore, direct comparison of sense and antisense ratio between B73 and Mo17 was performed via likelihood ratio test by treating each observed read count of sense or antisense strand in B73 or Mo17 genotype as a random Poisson variables (Method). In total, the likelihood ratio test can detect 465 genes showing the significant difference of sense and antisense ratio between the two genotypes. In other words, the level of the ratio between sense and antisense transcripts is different on different genotype. Of the 465 genes, 268 (58%) exhibit differential NAT accumulation, indicating the significant contribution of NATs to the complex gene expression patterns.
NAT contributes to a complex regulatory network for gene expression

NATs had been shown widespread in various organisms and play key roles in various biological processes, such as genomic imprinting, stress response of plants, etc. Though its widespread and functional importance, how NATs were exactly transcribed and regulated remain unclear. eQTL have been proven to be an efficient and effective approach to elucidate the overall regulatory network for gene expression. Thus, in this study, we measured the expression levels of 2,420 oligo pairs mapped to the gene models on both strands and treated the transcript accumulation level as traits to map the QTL controlling not only the sense but also the antisense transcript accumulation. To our knowledge, this is the first study for detecting NAT eQTL. Of the 4,840 unique oligos, 590 oligo has at least 1 eQTL associated with the oligo-measured traits. 331 oligos actually detect the antisense transcript accumulation and there are 419 non-redundant eQTL detected with 10% cis-NAT-eQTL. These results of NAT-eQTL are consistent with our RNA-seq experiment in which the NAT exhibited significant differential accumulation between different genotypes. It is not surprised to identify some genomic intervals controlling lots of transcript accumulation, also termed as eQTL hotspots, from sense strand as reported before. Interestingly, we also found several NAT-eQTL hotspots. Though the hotspots for sense and antisense transcript accumulation are generally not overlapped, we found lots of intervals controlling both sense and antisense transcript accumulation, suggesting the possibility of co-regulation of sense and antisense transcript accumulation by some common regulators. However, the identified eQTL pattern for sense and antisense transcript accumulation are different since we found that compared to the sense eQTL, the identified antisense eQTL are more likely to be trans-eQTL (p value 1.2e-12).

Heterosis (or hybrid vigor) is the phenomenon that the offspring of two selected parents exhibit greater fitness than the parents. The molecular mechanism(s) responsible for heterosis are not known. Crosses between the maize inbreds B73 and Mo17 yield a heterotic F1 hybrid. Previously, we (34) and others (49, 50) detected novel gene expression profiles in the B73xMo17 F1 hybrid relative to its inbred parents. Based on the results of the current study, one hypothesis to explain the differential accumulation of transcripts between an F1 and its inbred parents is the differential regulation and accumulation of NATs in the two inbred parents, which when combined in the F1 could differentially regulate the accumulation of sense transcripts and potentially contribute to heterosis.

MATERIALS AND METHODS

Genetic Stocks and Plant Growth

The maize inbreds B73 (Schnable Lab Ac #660) and Mo17 (Schnable Lab Ac #2618) used in this study for RNA-seq and microarray hybridization were derived from self-pollination of stocks originally obtained from Drs. Donald Robertson and Mike Lee, respectively. In the eQTL experiment, 56 RILs were selected from 91 IBM mapping population. For a given RIL, the RIL, B73xRIL and Mo17x RIL lines were produced by selfing RIL, crossing RIL (as male) onto B73 and Mo17, respectively. All maize plants were
grown in SB 300 Universal soil (Sun Gro Horticulture, Bellevue, WA) in growth chamber highly controlled conditions (15 hours of light at 25°C and 9 hours of dark at 20°C). The seedlings were harvested at 14 days after the planting.

**Strand-specific EST identification and bioinformatics searching of NATs**

The maize whole genome pseudomolecule (Release 4a53) was downloaded from www.maizesequence.org and divided into pieces separated by 100 N in each chromosome due to the physical and sequencing gaps. Maize EST sequences were downloaded from GenBank on Sep., 2009 and trimmed with Seqclean (http://compbio.dfci.harvard.edu/tgi/software/) to remove polyA tail and vector contamination. Trimmed ESTs were mapped to maize genome with via GMAP (51). The strand-specific ESTs need to meet the following 3 criteria: 1) Uniquely mapped to genome with at least 95% identity; 2) Both tails of the alignment is no longer than 10 bp; 3) at least one splicing site in the ESTs or more than 100 codons in the open reading frame.

The identified strand-specific ESTs were reverse complemented if the ESTs are on the antisense strand of the original mRNA based on the splicing site signal or codon orientation. The derived strand-specific mRNA sequences were clustered together if such mRNAs were overlapped on the same genomic locus and mapped to the same strands at the given locus. The derived clusters were treated as strand-specific transcriptional units used for in silico NAT prediction in this study. To search the putative pairs of sense and antisense pairs, the strand-specific transcriptional units were used to blast search themselves via ungapped blastn (parameters: -F "m D" -S 2 -g F -r 1 -q -10). The following criteria are required to call the sense and antisense pairs: 1) The overlap length need to be at least 75 bp; 2) The sequence identity in the overlap region is at least 95%; 3) One mRNA sense strand overlaps with the antisense strand of the paired mRNA.

**Strand-specific RNA-seq experiment**

*Constructing and sequencing the strand-specific RNA-seq libraries*

The strand-specific RNA-seq library was constructed by following the previously published procedures (10) with modifications. Total RNA was extracted using Trizol reagent (Invitrogen) followed by 3 rounds of mRNA isolation using oligodex kit (Qiagen) to select the polyA+ mRNA molecule. The isolated mRNA was treated with 2.5U Tobacco Acid Pyrophosphatase for 1.5 h at 37°C in 10 ul solution followed by DNase I (QIagen) incubation at room temperature for 30 minutes. The treated mRNA was purified by phenol:chloroform extraction and ethanol precipitation. The reaction conditions for downstream 3’ and 5’ adaptor ligation and reverse transcription reactions are same as the conditions used in Lister et al (10) except the initial selected size of mRNA from the PAGE gel is from 100-120 nt. The produced cDNA samples were used template for a 18-cycle PCR amplification to enrich the two adaptor ligated cDNAs. The PCR products were run on 1% agarose gel and 170-200 bp gel were recovered by Qiagen gel extraction kit. This library was subjected to Illumina 75-cycle sequencing run.
Alignment of strand-specific RNA-seq reads to genome and gene models
To map the RNA-seq read to the maize reference genome, the maize reference genome (refv1) was firstly indexed via novoindex with a 13 hash length and a step size 2 (-k 13, -s 2; version 2.05.17MT; http://www.novocraft.com). Then the 75 bp RNA-seq reads were aligned to the indexed genome via Novoalign (version 2.05.17MT) using the Needleman-Wunsch algorithm with affine gap penalties (-g 40, -x 3). The alignment with quality score >= 20 was used for further analysis (the probability of correct alignment is >= 99%). Reads were required uniquely mapped to the reference genome with less than or equal to 4 mismatches. Due to possibility of RNA-seq reads jumping across the intron, the unmapped reads by Novoalign were used as input for GMAP, which can map the cDNA to genome with splice site alignment, with parameter (-S -B 2 -K 5000 -t 8). The unique best GMAP alignment was used for further analysis for a given read if the alignment identity >=95% and the coverage >=90%. The mapping results from Novoalign and GMAP were pooled together for each genotype for further analysis.

Identification of genes accumulating NATs via RNA-seq reads
The uniquely mapped reads via Novoalign and GMAP in each genotype were projected onto the filtered maize gene models to determine the transcription orientation relative the annotated gene orientation. Due to the strand-specific property of the RNA-seq library, the achieved RNA-seq reads represent the original mRNA orientation. Thus, for a given RNA-seq read, if the two orientations from the alignment to the genome and projection to the gene models are consistent (both forward or reverse), the RNA-seq read derives from sense transcript. Otherwise, the RNA-seq read derived from antisense transcript.

Sense and antisense expression analysis within and between genotypes
The RNA-seq data generally involves in two factors with 2 levels for each factor. One factor is genotype with B73 or Mo17, and another factor is transcription strand with sense or antisense. With this data structure, 5 possible informative statistical tests can be performed to explore the gene expression pattern on different strand between different genotypes. To investigate whether the sense and antisense show equal expression within genotype, the binomial test were used in R package. To test the equality of sense or antisense accumulation between genotypes, the Fisher’s exact test were used. Finally, to test the equality of ratio of sense versus antisense reads between the two genotypes, likelihood ratio test was used by treating the read count for a given gene on specific strand and genotype as a random Poisson variables.

NAT gene expression analysis between B73 and Mo17 via microarray
Experimental design
Eight biological replications of B73 and Mo17 were grown in a randomized complete block design under highly controlled conditions in growth chambers for microarray experiment.
Microarray printing, labeling, hybridization, scanning and data acquisition

Total RNA was extracted from each genotype in each biological replication followed by mRNA isolation using oligodex kit (Qiagen) and RNA labeling using ULS-RNA labeling kit (Kreatech) by following the manufacture’s instructions. The B73 and Mo17 RNA samples labeled with cy3 or cy5 were pooled together and manually hybridized onto ISU maize oligo array at 42 degree for 16 hours. The microarray slides were washed 3 times using different washing buffers (10 minutes for 1st wash in 2X SSC and 0.1% SDS; 20 minutes for 2nd wash in 0.1X SSC and 0.1% SDS; 5 minutes for 3rd wash in 0.1X SSC solutions). Dry the slides in air, scan the slides and quantify the images according to previous procedures (34). The signal intensities were background corrected followed by lowess normalization to remove the signal-intensity-dependent dye effects for each slide.

Microarray data analysis

A mixed linear statistical model was fitted using the normalized, log-scale signal intensities to identify the strand-specific transcripts accumulating significantly differently among genotypes. The mixed linear model included fixed effects of genotype and dye and random effects related to the other experimental design factors.

Mapping strand-specific oligos to the gene models

To map the oligos to the annotated maize gene models, the plus strand sequence from each complementary oligo pair were firstly used to BLAST against the maize reference genome (B73refv1). The criteria of 95% identity and 90% coverage was used to decide if a given oligo is mapped to the genome. To rule out the cross-hybridization, the criteria of 85% identity and 85% coverage was used to decide if a oligo can be mapped to more than 1 genome locations. By using these two stringent criteria, 2,920 oligo pairs can be uniquely mapped to the maize reference genome. Of those uniquely mapped oligos, 2,420 (83%) can be projected to the filtered gene set. By using the annotated gene orientation, 2,420 oligos were defined to detect the sense transcripts accumulation and another 2,420 oligos to detect the antisense transcripts accumulation.

NATs eQTL experiment

Experimental design of eQTL experiment

The recombination inbred lines (RILs, N=56) from the intermated B73 and Mo17 (IBM) mapping population and two test x RIL crosstypes, B73xRIL and Mo17xRIL, were grown using a split-plot design with RIL group as described previously. For a given RIL, loop design were used for the Cy3 or Cy5 labeled RNA hybridized to the strand-specific microarray (Figure). In total, two biological replications were conducted with dye-swap for hybridization.

Mapping molecular markers and oligos to genetic positions of integrated physical-genetic map
In total, 1064 highly informative markers based on previously published genetic map were selected as described previously. And 2420 oligo pairs can be uniquely mapped to the genome and projected on the filtered gene set. By using the physical positions of genetic markers, a General Additive Model (GAM) function was generated to integrate the physical and genetic maps. The GAM function was used to predict the genetic position on the integrated physical and genetic map. Thus, the genetic positions of the 1,064 molecule markers and 2,420 oligo pairs were predicted and used in the eQTL analysis.

Detection of NATs eQTL detection via multiple interval mapping

The mean of the expression levels from 2 technical and 2 biological replications were used for eQTL mapping via multiple interval mapping by following the published procedures (41). To calculate the false discovery rate, the method reported by Carlborg et al was used (9).

ACKNOWLEDGEMENTS

We thank Cheng-Ting Yeh (Iowa State University), Karthik Viswanathan (Iowa State University), Pengcheng Lu (Iowa State University) for computational support. We thank former members of the Schnable Lab: Drs. Scott J. Emrich (Iowa State University), Yan Fu (Iowa State University), Ling Guo (Iowa State University), and Rhonda Decook (University of Iowa) for thoughtful discussions. This project is supported by the National Research Initiative competitive grant no. 2007-35301-18372 from the USDA National Institute of Food and Agriculture with additional funding from Iowa State University’s Plant Sciences Institute, Hatch Act, and State of Iowa funds.

REFERENCES

### TABLE 1

Mapping RNA-seq reads to specific strands of the Filter Gene Set

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Total reads</th>
<th>Mapped reads</th>
<th>No. NAT genes</th>
<th>Total genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>To genome</td>
<td>To genes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sense</td>
<td>Antisense</td>
<td>Sense</td>
</tr>
<tr>
<td>B73</td>
<td>11,606,941</td>
<td>8,050,114</td>
<td>5,811,524</td>
<td>203,895</td>
</tr>
<tr>
<td>Mo17</td>
<td>12,680,380</td>
<td>8,223,696</td>
<td>5,868,488</td>
<td>192,868</td>
</tr>
</tbody>
</table>

*The number of reads that aligned to a unique location of the B73 reference genome; \(^b\) The number of uniquely mapped reads that can be projected onto sense or antisense strands of gene models; \(^c\) The union of genes with antisense RNA-seq reads; \(^d\) The union of genes captured from two genotypes from both strands.*
<table>
<thead>
<tr>
<th>Approach</th>
<th>GO term</th>
<th>P value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ssRNA-seq</strong> (cis-NATs)</td>
<td>ATPase activity, coupled to transmembrane movement of substances</td>
<td>2e-04</td>
</tr>
<tr>
<td></td>
<td>ATP binding</td>
<td>5e-06</td>
</tr>
<tr>
<td></td>
<td>lipoxygenase activity</td>
<td>2e-06</td>
</tr>
<tr>
<td></td>
<td>monooxygenase activity</td>
<td>4e-04</td>
</tr>
<tr>
<td></td>
<td>protein amino acid phosphorylation</td>
<td>3e-06</td>
</tr>
<tr>
<td></td>
<td>protein serine/threonine kinase activity</td>
<td>3e-04</td>
</tr>
<tr>
<td><strong>Bioinformatics</strong> (cis/trans NATs)</td>
<td>GTPase activity</td>
<td>7e-06</td>
</tr>
<tr>
<td></td>
<td>lipid transport</td>
<td>2e-06</td>
</tr>
<tr>
<td></td>
<td>lipoxygenase activity</td>
<td>8e-05</td>
</tr>
<tr>
<td></td>
<td>monooxygenase activity</td>
<td>2e-05</td>
</tr>
<tr>
<td></td>
<td>metabolic process</td>
<td>3e-04</td>
</tr>
<tr>
<td></td>
<td>photosynthesis, light harvesting</td>
<td>6e-07</td>
</tr>
<tr>
<td></td>
<td>nucleosome assembly</td>
<td>2e-09</td>
</tr>
<tr>
<td></td>
<td>nucleosome</td>
<td>1e-10</td>
</tr>
<tr>
<td></td>
<td>RNA splicing</td>
<td>3e-05</td>
</tr>
<tr>
<td></td>
<td>spliceosome</td>
<td>9e-06</td>
</tr>
<tr>
<td></td>
<td>transferase activity, transferring glycosyl groups</td>
<td>3e-04</td>
</tr>
<tr>
<td></td>
<td>transferase activity, transferring hexosyl groups</td>
<td>6e-05</td>
</tr>
</tbody>
</table>

<sup>a</sup>False discovery rate (FDR) <0.05 was used as significant cutoff and calculated using Benjamini and Hochberg’s procedure [34].
TABLE 3

Contribution of NATs to the interaction between transcription orientation and genotypes

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Sig. Interaction</th>
<th>Sig. NAT</th>
<th>Sig. Sense</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>48</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>149</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>127</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>141</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>465</td>
</tr>
</tbody>
</table>
### TABLE 4
Summary statistics of NATs eQTL analysis

<table>
<thead>
<tr>
<th>Cross-type</th>
<th>No. eQTL</th>
<th>FDR</th>
<th>No. eQTL (strand)</th>
<th>No. eQTL (modes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sense</td>
<td>NAT</td>
</tr>
<tr>
<td>RIL</td>
<td>279</td>
<td>10%</td>
<td>136</td>
<td>143</td>
</tr>
<tr>
<td>B73xRIL</td>
<td>276</td>
<td>42%</td>
<td>132</td>
<td>144</td>
</tr>
<tr>
<td>Mo17xRIL</td>
<td>298</td>
<td>24%</td>
<td>154</td>
<td>144</td>
</tr>
<tr>
<td>Total</td>
<td>815</td>
<td>22%</td>
<td>396</td>
<td>419</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

**Figure 1.** Distribution of NAT RNA-seq reads in flanking and transcribed gene regions. The promoter and terminator regions were defined as 1kb upstream from transcriptional start site and downstream from transcriptional stop site for NAT mapped genes, respectively. The proportion of NAT reads was calculated from the number of reads normalized to 1 million bases on each type of regions.

**Figure 2.** Co-localization of NATs and H3K4me3 reads within transcribed gene regions. The structural annotation for the 5′ and 3′ UTRs, coding and intron regions was retrieved from GenBank records for a set of full-length cDNA used in this analysis (Methods). The proportion of NAT reads was calculated from the number of reads normalized to 1 million bases on each type of regions.

**Figure 3.** Distribution of sense and NAT eQTL across the maize genome. The frequency of sense and NAT-eQTL at each maker locus were plotted across whole maize genetic distance from chromosome 1 to 10.
Figure 1 Jia et al.
Figure 2 Jia et al.
Figure 3 Jia et al.
CHAPTER 5. CELL-LAYER SPECIFIC GENE EXPRESSION IN MAIZE SHOOT APICAL MERISTEMS

A paper to be submitted to PLoS GENETICS

Yi Jia1, Kazuhiro Ohtsu1, Michael J Scanlon2, Dan Nettleton1, Patrick S Schnable1,3

ABSTRACT

All above ground organs of higher plants are ultimately derived from shoot apical meristems (SAMs). The SAM exhibits distinctive structural organization, and monocot SAMs such as maize are comprised of a single cell layered tunica (L1) and a corpus (L2). Although recent research has revealed roles of these cell layers in the SAM, intra- and inter-cell-layer signaling networks involved in organ development remain largely unknown except for a few differentially expressed genes. In this study, various types of transcription factors were identified that exhibited differential expression between L1 and L2 cell layer within maize SAMs using microarray and RNA-seq experiment. In addition, genes related to signal transduction were identified as being highly expressed in the L1 cell layer relative to L2, suggesting the possible novel cellular communication pathways. In total, 5,739 genes (26% of assayed genes) were identified as differentially expressed genes between the two cell layers (false discovery rate < 5%).

Interestingly, although the L1 cell layer only occupies a small proportion of SAMs, the transcript diversity in the L1 is about 3 fold higher than in the L2. Transcriptomic diversity in the L1 is further supported by our fluorescence hybridization in situ and cross-reference analysis showing various genes exhibit highly divergent expression pattern across the L1 cell layer. Furthermore, genes involved in various aspects of chromatin modification are significantly over-represented (p value = 1.2e-14) among the differentially expressed genes, suggesting a role for chromatin modification in cellular differentiation in maize SAMs.

INTRODUCTION

All above ground organs of higher plants are ultimately derived from specialized organogenic structures termed shoot apical meristems (SAMs) [1] which are specified during embryogenesis. SAMs exhibit distinctive structural organization, marked by clonally distinct cell layering [2]. In Arabidopsis, SAMs are composed of three layers: L1 (epidermal), L2 (subepidermal) and L3 (inner layers). L1 and L2 each consist of a single cell layer, together forming the tunica, and undergo principally anticlinal cell divisions. The underlying L3 cells do not show particular restrictions in their division planes and form the corpus. In contrast, the SAMs of monocots, such as maize, usually are comprised of two cell layers, a single cell layered tunica (L1) and a corpus (L2).

In addition to this layered structure, SAMs also display histological zonation associated with different

1 Iowa State University, Ames, IA; 2 Cornell University, Ithaca, NY;
3 To whom correspondence should be addressed (schnable@iastate.edu).
functions during the development. For example, lateral organs arise from the peripheral zone and the self-
regenerating stem cells are located in the central zone that includes both tunica and corpus cells.
Maintaining stem cell identity and organogenesis activities are tightly regulated by a complex regulatory
network involving various transcriptional factors and signal transduction proteins, as well as epigenetic
factors. In Arabidopsis, WUSCHEL/CLAVATA (WUS/CLV) and KNOX define two of the main pathways
involved in the SAM development. The WUS gene encodes a homeodomain transcription factor
responsible for the initiation and maintenance of stem cells during embryogenesis and SAM development,
respectively [3]. The CLV family includes CLV3, small polypeptide produced from stem cells, which
interacts with the CLV1-CLV2 receptor complex to initiate downstream events that repress the expression
of WUS. The WUS/CLV pathway is conserved between mono- and dicot species. The KNOX pathway
regulates meristem cell fate and in the SAM the highly induced KNOX gene suppresses leaf differentiation
[4]. SAM development can also be regulated by chromatin remodeling factors involved in chromatin
assembly and remodeling, histone modification, and other functionally unknown chromatin proteins. Such
chromatin remodeling factors can regulate the spatial or temporal expression of key transcription factors
required for either meristem maintenance or leaf or flower organogenesis [5]. For example, the Arabidopsis
FAS1 and FAS2 genes encode two subunits of chromatin assembly factor 1 (CAF-1) complex, which
suppress the spatial expression of WUS in the SAM; consequently mutants of FAS1 and FAS2 exhibit
aberrant SAM development.

Nevertheless, the intra- and inter-cell-layer signaling networks that enable coordinated organ development
from the SAM remain largely unknown. It is known that some genes are differentially expressed between
these cell layers. Maize “outer cell layer” genes such as ZmOCL1 and ZmOCLA, which belong to the class
IV homeodomain-leucine zipper (HD-ZIP) IV gene family, exhibit L1-specific expression in the SAM
[6,7]. On the other hand transcripts of the KNOTTED1 (KN1) gene, which is important for meristem
indeterminacy, are known to accumulate in the L2 but not in the L1 of the maize SAM [8].

We used laser capture microdissection (LCM) to isolate RNA from the L1 and L2 layers. In an effort to
identify novel genes potentially critical for cell-layer-specific functions the transcriptomes of these two cell
layers were characterized using both microarray and RNA-seq technologies. The L1 transcriptome is much
more diverse than that of the L2. Furthermore, >5,000 genes were differentially expressed between L1 and
L2 in which various types of genes were over-represented including homeobox genes, lipid-related
signaling proteins, chromatin-modification related genes, etc. Possible functions of these genes in clonally
distinct cell layers in the SAM are discussed.
RESULTS

Gene expression analysis in L1 and L2 cell layers of SAM via LCM, microarray and RNA-seq technologies

Gene expression patterns were characterized in the L1 and the L2 of the maize SAMs. L1 and L2 tissues were collected from SAMs in 14-day-old B73 maize seedlings via LCM (Figure 1). L1 and L2 tissues in the SAM proper above the plastochron 0 (P0) level were mainly targeted with occasional inclusion of P0 (Methods). In total, 6 biological replications of RNA samples were extracted from L1 and L2 cell layers followed by RNA amplification and synthesis of double-stranded cDNA according to previous procedures and subjected to hybridization onto 3 different SAM microarrays [9] spotted with meristem-enriched ESTs (Table S1). A pool of RNA sample from L1 of 13 SAMs and a pool of RNA samples from L2 of 13 SAMs were extracted, amplified and subjected to Illumina/Solexa sequencing (Methods). Quality controls were performed for all the amplified RNA samples (Figure S1).

For the RNA-seq experiment, in total, 3.1 million reads from the L1 and 8.6 million reads from the L2 could be uniquely mapped to the Maize Genome Sequencing Project’s (MGSP’s) B73 reference genome [10]. Differentially expressed genes were identified from the microarray and RNA-seq experiments, respectively. The reads from RNA-seq experiment were mapped to the “filtered gene set” annotated by MGSP [10]. Among reads that could be uniquely mapped to the genome, 2.6 million (85%) from L1 and 7.4 million (86%) from L2 aligned to gene models (Table S2). At least one Illumina/Solexa read from at least one of the two cell types aligned to 21,602 of the 32,540 genes in the MGSP’s filtered gene set (Table S2). Of the 21,602 genes assayed in the RNA-seq data, 5,605 (26% of 21,602) could be declared to be differentially expressed between the L1 and L2 cell layers at an estimated 5% False Discovery Rate (FDR).

For microarray experiment, each microarray hybridized with cDNA samples of L1 and L2 RNAs were scanned at multiple intensities [11] followed by differential expression analysis for each type of microarrays across 6 biological replications using LIMMA [12]. Using an estimated FDR of less than 0.05 as significance cut-off, 805 ESTs (2.1% of the 37,662 informative maize ESTs on the microarrays) were found to differentially accumulate between the L1 and the L2.

Significant correlation between RNA-seq and microarray experiment

To compare the results of gene expression measured from microarray and RNA-seq, EST sequences on the microarrays were mapped to the MGSP’s filtered gene set. Of the 37,662 ESTs assayed in this study, 29,460 (78%) can be uniquely mapped to the genome and of these 29,460 ESTs, 22,450 (76%) can be mapped to 9,671 filtered genes. Consequently, the 805 ESTs showing differential expression can be mapped to 370 filtered genes with 198 (54%) being up-regulated in L1 relative to L2. In total, there are 9,053 genes existing in both experiments. Of these, 236 genes were identified as differentially expressed genes in both experiments (Figure S2). To check the consistency between the two technologies, three
different analyses were conducted. First, the observed number of DE genes overlapped in both experiments is significant higher than expected by chance through a Chi-square test (p value=0.0008). Second, among the 236 DE genes detected in both experiments, we also investigated whether the expression ratios (L1/L2) from both experiments showed same direction. In total, 221 out of 236 genes (91%) are consistent with each other, which is much higher than the expected number of genes under the null hypothesis of no correlation (Table S3; p value < 2.2e-16). Lastly, the overall fold changes obtained from the two experimental approaches also generally agreed with each other (Figure S2). Accounting for the high consistency for detecting the differential expression between the microarray and RNA-seq experiments, a union of expression results was established by pooling the results from both experiments (Methods). In total, 5,739 (26%) genes were identified as being differentially expressed between the L1 and L2 cell layers, with about half (2,917) being up-regulated in the L1 relative L2 cell layer (Table S4). By using the InterPro annotation of the filtered gene set from the MGSP, several protein domains were identified that are overrepresented in the differentially expressed gene set (Table 2), including homeobox, lipid-binding START, TATA box binding protein associated factor, pentatricopeptide repeat (PPR) and several histone related ones.

Control genes exhibit SAM cell-layer specific expression pattern
Several previously described genes that accumulate in either the L1 or the L2 were considered as marker genes for the two cell layers. For example, the maize outer cell layer (ZmOCL) genes, the member of HD-ZIP IV gene family, have been showed to be specifically expressed in the whole L1 cell layers. In either or both our microarray and RNA-seq experiments, ZmOCL1, ZmOCL3-5 were shown to be up-regulated in the L1 cell-layer (Table S8). Several genes have been shown previously to be mainly expressed in the L2 rather than the L1, such as the shoot apical meristem maintenance gene knotted-1 [13], ZmWuschel [14], ZmPIN1a and ZmPIN1b, two putative candidates for polar auxin transport and plant architecture determination in maize [15]. Similarly, all such control genes up regulated in the L2 cell layer were confirmed by our RNA-seq experiments or detected as being differentially expressed by our microarray experiment.

Higher diversity in the L1 transcriptome than in the L2 transcriptome
Our foldchange and diversity analysis revealed that the L1 showed transcript higher diversity than the L2. In the RNA-seq experiment, among the 2,838 genes up-regulated in L1 relative to L2, 389 (14%) genes showed 10 times higher transcript accumulation levels in L1 relative to L2 (Figure 2). However, among the 2,767 genes that are up-regulated in the L2 relative to L1, only 99 (3.6%) showed 10 times higher fold changes (Figure 2). Thus, more genes are highly expressed in the L1 than those that are highly expressed in the L2. This type of fold change pattern was also observed in the microarray experiment. Of those showing differential expression, 444 genes (55%) are up regulated in the L1 relative to the L2 and 361 (45%) genes are up regulated in L2 relative to the L1. Furthermore, fold changes varied between the L1 up-
regulated and the L2 up-regulated ESTs (Figure 2). Approximately 6% (22/444) of the L1-up-regulated ESTs exhibited more than a ten-fold change (L1/L2), whereas less than 1% (3/361) of the L2-up-regulated ESTs exhibited more than a ten-fold change (L2/L1).

The observation of higher expression in L1 could be due to the higher diversity of L1 transcriptome than L2. For example, if a gene is expressed in L1 but not in L2, the foldchange of L1/L2 will become to infinitely positive. To test this hypothesis, the number of genes sampled by every 1 million RNA-seq reads was calculated between the L1 and L2 transcriptomes (Table S2). This analysis indicated the diversity of L1 transcriptome is about 3 fold higher than in the L2. This conclusion can be further supported by the cross-reference analysis of domain-specific gene expression analysis within SAMs. In our previous study, we compared gene expression between the P0/P1 leaf primordia and the SAM proper section and identified 962 genes that are differentially expressed in the P0/P1 leaf primordial. By mapping these genes to the filtered gene set used in this study, 334 genes were identified that exhibited differential expression in both experiments. Of these, 124 genes were up-regulated in L1 relative L2 and also up-regulated in P0/P1 relative to the SAM proper. Thus, among the 2,917 genes up regulated in the L1, at least 124 (4.3%) mainly express in the leaf primordial sites. It is known that some other genes, e.g. ZmOCL 1-5, show almost uniform expression across the L1 cell layer. These observations indicate the expression pattern across different genes exhibits large variation and further support the diverse L1 transcriptome in spite of the small proportion of L1 within SAM tissues.

**Lipid genes are highly expressed in the L1 cell layer**

Lipids are hydrophobic or amphiphilic small molecules many of which have structural or functional roles in cell membranes, signal transduction, energy storage, etc. In plants, the wax on epidermal cells mainly consist of very long chain aliphatic lipids that are believed to play a role in defense against pathogens. In Arabidopsis, a list of candidate genes/proteins have been collected in the Arabidopsis Lipid Gene Database [16]. To study the expression of lipid-related genes between the L1 and L2 cell layers, the predicted maize homologs of Arabidopsis lipid genes were searched against the filtered gene set derived from the MGSP (See Methods). Using annotations from Arabidopsis, in total, 87 (1.5% of DE genes) lipid-related genes were identified as being differentially expressed between L1 and L2 cell layers using FDR<5%. The 87 differentially expressed lipid-related genes were classified into 3 different categories: 63 lipid metabolism genes, 6 genes encoding lipid transfer proteins and 18 genes encoding lipid signal proteins (Table S5). Interestingly, of these DE genes, 22 across all three gene categories exhibited 10-fold higher expression in the L1 cell layer than in the L2 cell layer, suggesting the possible important functions of lipids in multiple aspects of SAM development. In addition, we have conducted an over-representation analysis of the functional annotation, independent of the Arabidopsis lipid annotation, and identified the lipid-binding START domain as being highly over-represented among the differentially expressed genes (Table 2).

**Class III and IV HD-ZIP genes are highly differentially expressed between SAM cell layers**
Proteins in HD-ZIP family of transcription factors include both a homeodomain and a leucine zipper motif. The HD-ZIP family is unique to plants and functions in multiple biological processes, including for example, environmental responses, meristem regulation, mediation of hormones action, organ and vascular development. The HD-ZIP family can be classified into four subfamilies (HD-ZIP I-IV) [17]. In this study, maize HD-ZIP genes were searched against the filtered gene set using the characterized Arabidopsis genes in each of the four subfamilies via blastp and their phylogenetic relationships established using MEGA4 (Methods). Among the four subfamilies, HD-ZIP III proteins have been characterized in various developmental events including SAM development. The HD-ZIP IV proteins are generally expressed in the outer cell layer of the plant organs. Within SAMs, HD-ZIP IV genes have been characterized to being specifically expressed in the L1 cell layer. The gene expression pattern of HD-ZIP III and IV were compared to their phylogenetic relationships (Figure 3). Interestingly, five maize HD-ZIP III genes assayed in this study all showed differential expression with different expression patterns, indicating the complex regulatory network for SAM development by HD-ZIP III genes. In contrast, among the 12 maize HD-ZIP IV genes, 9 were assayed in this study and they all 9 exhibited differential expression with extremely high expression in the L1 cell layer relative to L2. Thus, in addition to the previously described zmOCL genes, we have identified 4 novel HD-ZIP IV like genes that are highly expressed in the L1 cell layer.

**Genes involved in chromatin modifications are over-represented in the DE genes**

It is known that several chromatin modifications function in the network that regulates SAM development. However, to our knowledge a global study of how chromatin modification genes are differentially expressed in different cell types within the SAM tissue has not previously been reported. In total, we found 140 genes involved in chromatin modifications that were differentially expressed between L1 and L2 cell layers (Table S6). Of these, 77 (53%) were up-regulated in the L1 cell layer. Compared to the remainder of genes, those involved in chromatin modifications are highly enriched among the DE genes (Table S7; p value =1.2e-14), demonstrating the importance of chromatin modification in the regulation of SAM function. A wide variety of chromatin-associated genes exhibited differential expression between the L1 and L2 cell layers within SAMs, including those affecting various histone modifications, such as histone methylation, acetylation and deacetylation. Similarly to the lipid genes, the independent statistical analysis on functional annotation showed several histone related protein domains are over-represented in the differentially expressed genes.

**Quantitative RT-PCR (qRT-PCR), RT-PCR and in situ hybridization**

qRT-PCR analyses were performed on six genes that were significantly up-regulated in the L1 relative to the L2. These six genes included zyb15, DL, bHLH1, bHLH2, WW1 and WW2 (Table 1). Expression level of WW3, the third maize WW gene identified via BLAST searches (Figure S3), between the L1 and the L2 was also analyzed via qRT-PCR. Each of these six genes is more highly expressed in the L1 than in the L2 (Figure 4). Although two L2 samples did not yield fluorescence above the detection threshold for WW3
(see Legend of Figure 4), the qRT-PCR result supported its higher expression in the L1 than in the L2. Expression of the WW genes was further analyzed in various maize organs via RT-PCR. The organs analyzed were brace roots, mixed tissues (see Legend of Figure S4), husks, silks, tassels, ears, kernels and LCM-collected SAMs [9]. With one exception, transcripts of the WW genes were detected in all analyzed organs (Figure S4).

Transcript accumulation of zyb15, a maize homolog of DL (Table 1) was further analyzed via in situ hybridization. Transcripts of zyb15 were mainly detected on the adaxial side of the P0 and in the P1. In the P0, transcripts accumulated in the L1 and to a lesser extent in the underlying layer (Figure 5A). Transcripts of the DL homolog accumulated in a pattern similar to that of zyb15, but DL homolog transcripts were also detected in the inner part of the P0 and in older leaf primordia such as P2, and transcript accumulation was not adaxialized in the P0 (Figure 5B).

**DISCUSSION**

**More diverse transcriptome exists in L1 than in L2 cell layer within SAMs**

Global gene expression profiling between the L1 and L2 cell layers of SAMs were conducted using both microarray and RNA-seq approaches. In total, 5,724 genes were identified as differentially expressed using these two technologies (Table S4). The comparison between the results of microarray and RNA-seq indicates the detected expression patterns from the two technologies were highly correlated (p value <2.2e-16) and fold changes are similar between the two experimental approaches (Figure S2). Furthermore, expression patterns of the control genes were consistent with published data (Table S8). A number of lipid-related genes are known to exhibit L1 specific-expression in the SAM. These include genes encoding enzymes involved in lipid biosynthesis [18,19] and lipid transfer proteins [20]. In our study, nearly 70% of the significant genes that were annotated as lipid-related genes were up-regulated in the L1 relative to the L2 and among them 22 genes showed 10-fold higher expression in L1 than L2 (Table S5). Transcripts of ZmPIN1a and Zm[21]PIN1b, maize homologs of Arabidopsis PIN1, accumulated in the L2 but not in the L1 in the SAM [15]. Both genes were identified as being up-regulated in the L2. These data indicate that LCM-microarray and LCM-RNA-seq analyses indeed identified genes that are differentially expressed between distinct cell layers in the maize SAM. In addition to the two independent experiments (microarray hybridization and RNA-seq) showing high correlation, additional independent types of experiments, including qRT-PCR (Figure 4) and in situ hybridizations (Figure 5), further validated the expression patterns for various interesting genes.

The transcriptome diversity analysis demonstrated that the L1 is about 3-fold more diverse than that of the L2. L1-up-regulated genes had generally higher fold changes than did the L2-up-regulated genes (Figure 2). This observation raises the possibility that the L1 transcriptome is even more complex than that of the L2 (i.e., if a given gene is expressed in the L1 but not in the L2, the fold change of the transcript accumulation
[L1/L2] is generally very high) despite the fact that the L1 occupies only approximately one quarter of the SAM (Table S1). This observation was supported by the RNA-seq experiment.

**Differential expression of transcriptional factors between L1 and L2 cell layers**

Homeobox genes are transcriptional factors playing crucial roles in various developmental processes and characterized by the inclusion of a homeodomain (HD), a conserved stretch of 60 amino acids that exists in all eukaryotes. In plants, the HD-containing super gene family can be categorized into six families according to the different features of the HD domain and other associated motifs: HD-leucine zipper (HD-ZIP), HD-PHD, HD-WOX, HD-ZF, HD-KNOX and HD-BELL. Various of these families exhibited differential expression between the two cell layers targeted in this study. The several transcription factor genes that exhibit L1-specific transcript accumulation in the SAM belong almost exclusively belong to the HD-ZIP IV gene family [6,21,22,23]. Our study not only confirmed the expression patterns of previously characterized genes (Table S2) but also identified several novel HD-ZIP IV like genes (Figure 3). It will be interesting to characterize these genes via genetic and biochemical approaches in the future.

In addition to the HD-ZIP gene family, we also identified *yabby* family genes and putative basic helix-loop-helix transcription factor (*bHLH*) genes as being preferentially expressed in the L1 relative to the L2 (Table 1, Figure 4). Transcripts of zyb15 were analyzed via in situ hybridization and were detected in the L1 and to a lesser extent in the underlying layer of P0 at the adaxial side but not in the L1 of the SAM proper (Figure 5A). This pattern is quite different from those of the HD-ZIP IV genes whose transcripts accumulate in the L1 of the entire SAM and the young leaf primordia [6]. Instead, the transcript accumulation pattern of zyb15 is similar to that of zyb9, which is involved in the establishment of leaf polarity [24]. Expression of zyb9 was also significantly up-regulated in the L1 relative to the L2 although zyb9 exhibited a lower fold change than zyb15 (Table 1). This fold change difference may be due to the transcript accumulation of zyb9 in three to four cell layers from the L1 of P0 [24]. It is known that continuity of the L1 of the SAM proper to the incipient leaf primordium is essential for the adaxial cell fate in establishing leaf polarity [25]. It would be interesting to study whether zyb15 can coordinate signals from the L1 of the SAM proper to P0 with adaxial cell fate together with the other zyb genes [24]. Another *yabby* family gene, the maize *DL* homolog exhibited a similar but somewhat different expression profile from that of zyb15 (Table 1, Figures 4 and 5). The transcript accumulation pattern in the SAM (Figure 5B) is quite similar to that of the maize *yabby*-like gene described in Zhang et al. [26], which shares a high nucleotide sequence identity (~90% for their predicted coding regions) with the *DL* homolog analyzed in this study. Indeed, the four other maize *yabby*-like genes were L1-up-regulated by 7-32 fold. These genes could play redundant roles in initiation and expansion of lateral organ primordia.

**Expression of signal transduction proteins**
Cell–cell communication is essential for growth and development of multicellular organisms. The role of the outermost L1 cell layer and its derived epidermis in coordinating the growth of the inner-cell layers has long been debated. Savaldi-Goldstein expressed the enzyme for biosynthesis of brassinosteroid in the null mutant and demonstrated that the epidermis both drives and restricts plant shoot growth by sending growth signals — either physical or chemical — to the inner layers [27].

Inter-cell-layer communication, such as the KN1 movement between cell layers in the SAM [28], is important for coordinated organ growth in plants [27]. Given this, it is reasonable that expression of a number of genes encoding signal transduction components and transporters were differentially regulated between the L1 and the L2 in the SAM. Among the genes categorized in signal transduction, genes that were annotated as genes encoding WW domain-containing proteins (WW genes) were substantially up-regulated in the L1 relative to the L2, exhibiting the highest fold changes in the category (Table 1). WW domains are small protein-protein interaction modules that often associate with short proline-rich motifs in proteins that are involved in a variety of signaling pathways [29]. Database searches identified another WW gene (WW3; Figure S3), whose expression was also upregulated in the L1 relative to the L2 (see Legend of Figure 4). The deduced amino acid sequences of these WW genes contained conserved WW domains (Figure S3) but exhibited no significant ($e < 10^{-10}$) similarities with previously characterized genes.

Transcripts of these WW genes were detected in various organs from mature maize plants (Figure S4). Taken together, we hypothesize that these WW genes encode novel signaling components that play an important role in the outer cell layer(s) for coordinated organ growth from the SAM and in other parts of the plant via protein-protein interaction.

MATERIALS AND METHODS

Plant materials and growth conditions
Maize (Zea mays L., inbred B73) plants were grown as described previously [9]. SAMs were harvested 14 days after planting.

Preparation of paraffin sections and collecting L1 and L2 tissues in the SAM
SAMs were fixed, embedded in paraffin, sectioned and tissues collected via LCM as described previously [9]. Only SAM sections in which the L1 was distinguishable were used for tissue collection (Figure 1A). First, the position of P0 was predicted relative to the other young leaf primordia such as P1 in longitudinal SAM sections. The L1 was outlined via the laser cutting from above the predicted P0 position through the top of the SAM until the position with approximately the same altitude of the other end relative to the SAM axis (Figure 1B). After collecting the L1, a straight line was cut using the laser between the two ends of the L1 to outline the L2 (Figure 1D). In some cases, the L1 and the L2 of P0 were presumably included as well due to ambiguous positioning of P0. L1 and L2 tissues were collected separately from the same SAMs within each replication. RNA yield has been previously estimated as ~0.6 pg per cell in the maize SAM [9].
According to this estimation, a quantify of L1 tissue that would be sufficient to yield ~1 ng of RNA was collected for each replication (Table S1).

**RNA extraction and amplification**

Each set of the L1 and the L2 tissues was collected via LCM into the cap of 500-μl microcentrifuge tube filled with 40 μl mineral oil. After tissue collection, 100 μl of the Extraction Buffer of the PicoPure RNA Isolation Kit (Molecular Devices, Mountain View, CA, USA) was added to the mineral oil including the LCM-collected tissue and mixed thoroughly. RNA was extracted according to the kit’s manual with the RNase-free DNase I treatment [9] as if there were no mineral oil. For each replication, the same amount of RNA (~1 ng) from the L1 and the L2 was amplified (Table S1) using RiboAmp HS RNA Amplification Kit according to the manual. Amounts of the aRNA were determined by measuring OD_{260} values via SPECTRAmax PLUS (Molecular Devices, Sunnyvale, CA, USA). The yields of the aRNA samples are presented in Table S1.

**RT-PCR and RNA gel**

Each sample of aRNA was quality checked via RT-PCR using primers specific to the maize *tub6* gene (Genbank accession no. L10633) and RNA gel electrophoresis according to Ohtsu *et al.* [9] with some modifications: the primer pairs used for PCR were as follows: *tub6*-forward; 5’-TTCGTGGAGTGGATCCCCAACA-3’, *tub6*-reverse; 5’-ACGAGGTCGTTCATGTTGCTCT-3’, which are located in the 3’ part of the coding sequence. These primers were annealed to templates at 58°C in the PCR cycle. Because these primers do not span any introns, a PCR reaction using an RT product derived from a reaction without reverse transcriptase was included for each PCR experiment to check for genomic DNA contamination. Three micrograms of each aRNA sample were used for the RNA gel.

Expression of the maize *WW* genes (*WW1*, *WW2* and *WW3*) in various maize organs were analyzed via RT-PCR according to Emrich *et al.* [30] with slight modifications. The specific primer pairs used to amplify transcripts from the *WW* genes were as follows: *WW1*-forward; 5’-CCACTGGCAGAAGCTTGTAGA-3’, *WW1*-reverse; 5’-ACCGACCCTACCCCTTTTTA-3’, *WW2*-forward; 5’-ACTTCACCACGGCGCACAGATT-3’, *WW2*-reverse; 5’-AATTTTCATTCAACCAGCTCA-3’, *WW3*-forward; 5’-CGGGCGTGCTGTGTTATTAC-3’, *WW3*-reverse; 5’-CTGTCATATTCCAGGATACGACC-3’. Positions of these primers are presented in Figure S3A. The primers for the *tub6* genes described above were also used for the analysis.

**Synthesis, labeling and microarray hybridization of cDNA**

Eight micrograms (with a few exceptions; see footnote in Table S1) of aRNA were labeled according to Ohtsu *et al.* [9]. To remove dye-specific effects in the statistical analyses, dyes were swapped as described previously [9]. Six biological replications (Table S1) of the labeled cDNA samples (L1-Cy3/L2-Cy5 or
L1-Cy5/L2-Cy3) were hybridized to each of the three SAM-enriched cDNA microarrays [SAM1.0 (GPL2557), SAM2.0 (GPL2572), and SAM3.0 (GPL3538)] [9] according to Swanson-Wagner et al. [31]. Details regarding the cDNA microarrays are available at http://www.plantgenomics.iastate.edu/maizechip/.

**Microarray analysis**

Each of SAM1.0 arrays was scanned seven times with a ScanArray 5000 (Packard BioScience, which is now PerkinElmer, Wellesley, MA, USA) and each of SAM2.0 and SAM3.0 arrays was scanned nine times with a Pro Scan Array HT (PerkinElmer). Subsequently, three data sets (with low, medium and high laser settings) were selected from each array [31]. The Limma R package was used to identify differentially expressed genes. On a spot-by-spot basis, the scan with the smallest FDR value was selected for subsequent analyses. As a result, this study reports the gene expression patterns of 37,662 “informative” spots from the three microarrays. Microarray data have been deposited in the GEO database (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE9610.

**Illumina/Solexa high-throughput sequencing and data processing**

Thirteen separate collections of L1 and thirteen separate collections of L2 cell layers were used to form the L1 and L2 pools. Collected tissues were used for RNA extraction, RNA amplification and synthesis of double stranded cDNAs according to our previous published procedures. Illumina/Solexa libraries were constructed using these double stranded cDNAs following Illumina/Solexa’s standard protocol for genomic library preparation. The resulting Solexa reads were processed as described previously [32]. For genes assayed in both experiments, the results from the experiment showing the lower FDR were used in the union result set. To generate a union set of statistical test results from microarray and RNA-seq, the genes assayed unique to either microarray or RNA-seq were pooled together directly.

**Phylogenetic analysis of HD-ZIP gene family**

The known protein sequences in HD-ZIP III and IV sub-gene family from Arabidopsis and maize were used as query sequences to search the candidate maize homologs via blastp against maize protein sequences corresponding to the filtered gene set annotated by the MGSP. Those sequences with e-values smaller than 1e-10 and bit scores equal to or greater than 100 were kept for further analyses. The retrieved protein sequences within each sub-gene family were aligned with ClustalW using the default settings [33]. Phylogenetic and molecular evolutionary analyses were conducted on the aligned sequences using MEGA version 4 [34].

**Quantitative RT-PCR**

Quantitative RT-PCR (qRT-PCR) analyses were performed using the aRNA sample pair 1, 2 and 3 (Table S1) according to Skibbe et al. [11] with minor modifications. The sequences of the primers are as follows: 
zyb15-forward: 5'- TTAGAGCGCGCATCAAGTAG-3', zyb15-reverse: 5'-. 
GCAGACATACGCAAACATGG-3’ (both AY313902 and DN208255 include these sequences; Table 2),

DL-forward; 5’- CCGCTGTTGTAGCCCAGTAT-3’, DL-reverse; 5’- GCYTACAATCCAACCATAACG-3’,
bHLH1-forward; 5’- ACCAACAGACAGTTGTA-3’, bHLH1-reverse; 5’- 
GGATGATGTAACCAACAGCCTAC-3’ (both DV492549 and DN211708 include these sequences; Table 1),
bHLH2-forward; 5’- ATGCTGCGAGATTTGGTAGG-3’, bHLH2-reverse; 5’-
ACAAACAGGCTCGATCT-3’ (both DV492799 and DN208982 include these sequences; Table 1).

The primers used for the maize WW genes (WW1, WW2 and WW3) are described in the previous section.

ACKNOWLEDGEMENTS

We thank Ms. Marianne Smith for wet lab support, Dr. Xuefeng Zhao and Mr. Cheng-Ting Yeh for computational support, Dr. Lu Gao, Mr. Sanzhen Liu and Dr. An-Ping Hsia for thoughtful discussions. This project was supported by the National Science Foundation Plant Genome program (award numbers DBI-0321595 and DBI-0527192) and Hatch Act and State of Iowa funds.

REFERENCES

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Annotation</th>
<th>FC</th>
<th>FDR</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G079293</td>
<td><em>Zea mays yabby15 (zyb15)</em></td>
<td>26.4</td>
<td>5e-120</td>
<td>Y</td>
</tr>
<tr>
<td>GRMZM2G074543</td>
<td><em>zyb9</em></td>
<td>55.7</td>
<td>5.6e-39</td>
<td>N</td>
</tr>
<tr>
<td>GRMZM2G088309</td>
<td><em>DROOPING LEAF</em> homolog (DL)</td>
<td>3.4</td>
<td>4.7e-04</td>
<td>Y</td>
</tr>
<tr>
<td>GRMZM2G178182</td>
<td>Putative basic helix-loop-helix transcription factor</td>
<td>25.7</td>
<td>8.6e-37</td>
<td>Y</td>
</tr>
<tr>
<td>GRMZM2G046493</td>
<td><em>bHLH2</em></td>
<td>18.8</td>
<td>1.3e-53</td>
<td>Y</td>
</tr>
<tr>
<td>GRMZM2G067223</td>
<td>WW domain-containing protein gene (WW1)</td>
<td>31.5</td>
<td>1.3e-243</td>
<td>Y</td>
</tr>
<tr>
<td>GRMZM2G035305</td>
<td><em>WW2</em></td>
<td>1.8</td>
<td>1.0e-03</td>
<td>Y</td>
</tr>
<tr>
<td>GRMZM2G124053</td>
<td><em>WW3</em></td>
<td>Inf</td>
<td>4.8e-16</td>
<td>N</td>
</tr>
<tr>
<td>Interpro_ID</td>
<td>No. DE Genes</td>
<td>No. non-DE genes</td>
<td>P value</td>
<td>FDR</td>
</tr>
<tr>
<td>------------</td>
<td>--------------</td>
<td>------------------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>IPR000164</td>
<td>14</td>
<td>5</td>
<td>6.52E-05</td>
<td>3.41E-02</td>
</tr>
<tr>
<td>IPR001356</td>
<td>46</td>
<td>44</td>
<td>1.06E-05</td>
<td>8.28E-03</td>
</tr>
<tr>
<td>IPR002885</td>
<td>47</td>
<td>350</td>
<td>6.16E-16</td>
<td>2.89E-12</td>
</tr>
<tr>
<td>IPR002913</td>
<td>19</td>
<td>6</td>
<td>1.37E-06</td>
<td>1.29E-03</td>
</tr>
<tr>
<td>IPR004823</td>
<td>13</td>
<td>4</td>
<td>6.38E-05</td>
<td>3.41E-02</td>
</tr>
<tr>
<td>IPR005818</td>
<td>15</td>
<td>6</td>
<td>6.36E-05</td>
<td>3.41E-02</td>
</tr>
<tr>
<td>IPR005819</td>
<td>34</td>
<td>17</td>
<td>2.27E-08</td>
<td>2.67E-05</td>
</tr>
<tr>
<td>IPR007125</td>
<td>51</td>
<td>24</td>
<td>2.27E-12</td>
<td>5.33E-09</td>
</tr>
<tr>
<td>IPR009072</td>
<td>57</td>
<td>46</td>
<td>1.99E-08</td>
<td>2.67E-05</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

**Figure 1.** Collecting L1 and L2 tissues of maize SAMs from 14-day-old maize seedlings via LCM. (A) to (E) show the same section. (A) A SAM section before L1 collection. Leaf primordia are numbered according to their relative developmental ages, wherein P0 (0) corresponds to the incipient leaf forming on the flank of the SAM. (B) The laser cutting outlined the L1. (C) After collecting the L1 (Methods). (D) The laser cutting outlined the L2. (E) After collecting the L2. Bar: 100 μm.

**Figure 2.** Density plots and bar plots of foldchange for both array and RNA-seq experiments. Density plot of foldchange from microarray (A) and RNA-seq (C); Bar plot of foldchange from microarray (B) and RNA-seq (D).

**Figure 3:** Phylogenetic trees of the HD-ZIP III and IV sub gene families with the corresponding gene expression patterns. (A) HD-ZIP III sub-family; (B) HD-ZIP IV sub-family.

**Figure 4.** Quantitative RT-PCR (qRT-PCR) analyses of six differentially regulated genes. Fold changes (L1/L2) from qRT-PCR analyses (black bars) and the microarray analysis or RNA-seq (gray bars) are indicated on a logarithmic scale with six differentially regulated ESTs derived from the microarray analysis: *zyb15*, *DL*, *bHLH1*, *bHLH2*, *WW1* and *WW2* (Table 1). Means +SD of the three biological replications (Methods) are shown.

**Figure 5.** In situ hybridizations. Fourteen-day-old maize shoot apices were analyzed via *in situ* hybridization using anti-sense probes prepared from *zyb15* (A), a maize *DL* homolog (B), *bHLH2* (C), *WW1* (D) and *WW2* (E) cDNA clones. Leaf primordia are numbered as described in Figure 1. Bar: 100 μm.
Figure 1 Jia, et al.,
Figure 2 Jia, et al.,
Figure 3 Jia, et al
Figure 4 Jia, et al.,
Figure 5 Jia, et al.,
CHAPTER 6 General Conclusions

SUMMARY AND DISCUSSION

Maize shows extremely high diversity in DNA sequences, gene expression and phenotypes (1-3). The transcriptome studies in Chapter 2-5 demonstrate that gene expression is dependent on genotype, tissue types as well as regulated by epigenetic factors, confirming that gene expression is controlled and regulated by both genetic and epigenetic determinants. Importantly, the identified natural antisense transcripts (NATs) and hundreds of identified NAT-eQTL regulate the expression of NATs increase the complexity of the regulation of gene expression. These observations and conclusions will enhance our understanding of various fundamental biology questions by providing new testable hypothesis.

Multiple factors contribute to heterosis

Heterosis is the phenomenon that the progeny of particular inbred lines shows superior performance compared to both parents (4). Heterosis has been widely exploited by plant breeders and seed companies for ~80 years. However, the genetic basis of heterosis is still not clear. Multiple hypotheses have been proposed to explain heterosis, including the classic “dominance” and “overdominance” models. The paper in Chapter 2 investigated the modes of gene action between a maize hybrid and its inbred parents. Based on the fact that all possible modes of gene actions were observed for gene expression between maize Mo17x B73 and its parents, we hypothesize that multiple factors or mechanisms contribute to heterosis.

In Chapter 4, a genome-wide discovery of natural antisense transcripts was presented. We conservatively estimated almost 40% of maize genes accumulating NATs based on our bioinformatics and RNA-seq experiments. NATs have been demonstrated to accumulate in response to stress environments. Once they accumulate they can form sense and antisense pairs followed by multiple rounds of small RNA biogenesis. Thus, one attractive molecular hypothesis for heterosis is that the actions of small interfering RNAs (siRNAs) in hybrids cause the observed underdominant and overdominant gene action observed in the gene expression study described in Chapter 2. Furthermore, siRNAs are typically produced from transposons and repeats, both of which exhibit extreme diversity across maize inbred lines. In addition, at least two transposons were observed to show differential expression between the two inbred lines (5). Thus, siRNAs are more likely to differentially accumulate in different maize inbreds. This was supported by a recent small RNA study in rice, in which the abundance of distinct small RNAs size classes show differences between the two rice inbred lines with many small RNAs being differentially expressed in the hybrids as compared to the parents (6). Based on these findings and the fact that siRNAs can regulate gene expression by cleaving mRNAs (7) and via transcriptional silencing (8), we hypothesize that the interactions between sense and antisense transcripts in different genotypes may contribute to novel gene expression in the hybrids and also heterosis.
In addition, to the hypothesis involving NATs and siRNA, other epigenetic phenomena may contribute to
heterosis. Recently, Swanson-Wagner et al (9) reported that transcript accumulation in maize hybrids is
dominantly regulated by the paternally transmitted alleles, suggesting the widespread contribution of
imprinting or epigenetic process to the regulation of gene expression in maize hybrids. Finally, it is both
possible and likely that there is interplay between the siRNA and epigenetics mechanisms responsible for
heterosis. This is supported by the facts that siRNAs are important components in many epigenetic
processes (10-13). Another evidence for the interplay of these mechanisms is that the accumulation of
NATs exhibits strong positive correlation with open chromatin modifications and negative correlation with
closed chromatin modifications as revealed in Chapter 4, indicating the possible epigenetic manner of NAT
functions.

**Epigenetic regulation of gene expression in SAMs**

All above ground organs of higher plants are ultimately derived from shoot apical meristems (SAMs) (14).
Maintaining stem cell identity and organogenesis activities are tightly regulated by a complex regulatory
network involving various transcriptional factors and signal transduction proteins, as well as epigenetic
factors (15). In Arabidopsis, WUSCHEL/CLAVATA (WUS/CLV) and KNOX are the two main pathways
involved in the SAM development (16). The results from Chapter 3 indicated that most genes (80%) related
to chromatin modifications were significantly affected by a mutation of epigenetic components in the SAM
tissue, indicating the extreme importance of epigenetic regulation within the SAM. In addition, in Chapter
4, NATs have been found widespread in the maize transcriptome and hundreds of NAT-eQTL were
identified. Accounting for the previous report that NATs can regulate the sense transcription via chromatin
remodeling, the identified NAT-eQTL and the NATs themselves add the complexity of the regulation of
SAM development.

**PERSPECTIVE**

More and more research resources have become available for the maize research community. For example,
the whole maize genome assembly is now available with high quality of gene annotation. The maize
community is for a leader in the application of association studies due to the availability of the NAM
mapping population [2] with the Hapmap [1] to connect phenotypes and genotypes. In addition, the first
large-scale data on the maize epigenome landscape have recently been published. Accounting for the
possibility of interplay between genetic and epigenetic regulation for various different biological processes,
such as heterosis and signal transduction within the SAM, multiple dimensions of data or approaches
should be incorporated together to study a specific biological question or problem. Due to the importance
of epigenetic regulation and relative sparseness of epigenetic data, it would be valuable to generate more
epigenetic resources, such as epiRIL mapping populations [17], to associate specific phenotypic traits with
specific intervals linked by different epigenetic markers.
REFERENCE


## APPENDIX A. SUPPORTING MATERIAL FOR CHAPTER 3

### TABLE S1

*mop1* mutants and non-mutants have distinct SAM morphologies

<table>
<thead>
<tr>
<th>Genotype</th>
<th>SAM ID</th>
<th>Width (μM) (^a)</th>
<th>Height (μM) (^b)</th>
<th>Ratio(Height/Width)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>mop1/mop1</strong></td>
<td>1</td>
<td>171.0</td>
<td>118.4</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>180.1</td>
<td>162.6</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>170.7</td>
<td>145.2</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>161.3</td>
<td>142.7</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>170.8</td>
<td>142.2</td>
<td></td>
<td>0.83</td>
</tr>
<tr>
<td><strong>Mop1/mop1</strong></td>
<td>1</td>
<td>221.3</td>
<td>249.3</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>156.8</td>
<td>171.6</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>157.3</td>
<td>181.3</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>151.4</td>
<td>160.8</td>
<td>1.06</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>171.7</td>
<td>190.8</td>
<td></td>
<td>1.11</td>
</tr>
</tbody>
</table>

\(^a\) The length of a horizontal line drawn between the initiation points of P0 and P1 was defined as the width of the SAM. \(^b\) The length of a vertical line drawn from the top of the SAM to the horizontal line was defined as the height of the SAM. The averaged width and height were calculated for mutant and non-mutants across four different SAMs, respectively. A Welch two-sample t-test demonstrated that the mean ratio (height/width) of mutants (0.83) is significantly smaller than that of non-mutants (1.11) with p-value 0.006.
TABLE S2
List of differentially expressed DNA transposon families (PDF file)

This table is available and published as a supporting material in PLoS Genetics website at:
TABLE S3
List of differentially expressed retrotransposon sub-families

This table is available and published as a supporting material in PLoS Genetics website at:
<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Expression Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene1</td>
<td>1.2</td>
</tr>
<tr>
<td>Gene2</td>
<td>0.8</td>
</tr>
<tr>
<td>Gene3</td>
<td>1.5</td>
</tr>
</tbody>
</table>

This table is available and published as a supporting material in PLoS Genetics website at:
### TABLE S5
Alignment of RNA seq reads to genome and genes

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Total Reads</th>
<th>No. Non-uniquely mapped</th>
<th>No. Uniquely Mapped Reads</th>
<th>Total genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>To Genome</td>
<td>To Genes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>mop1/mop1</td>
<td>4,474,243</td>
<td>950,295 (21%)</td>
<td>2,775,266 (62%)</td>
<td>2,156,241 (78%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2,156,241 (78%)</td>
<td></td>
</tr>
<tr>
<td>Mop1/mop1</td>
<td>5,982,599</td>
<td>1,350,921 (23%)</td>
<td>4,113,045 (69%)</td>
<td>3,248,869 (79%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3,248,869 (79%)</td>
<td></td>
</tr>
</tbody>
</table>

* The number of reads after trimming low quality bases at the 3’ end and removing the reads shorter than 32 bp; *b* The number of reads (percentage) that aligned to more than 1 genomic locations on B73 reference genome allowing a maximum of 2 mismatches including insertion/deletions; *c* The number of reads (percentage of total reads) that aligned to a unique location of the B73 reference genome; *d* The number of uniquely mapped reads (percentage of uniquely mapped to genome) that can be projected onto gene models; *e* The union of genes (percentage of all 32,540 all genes) captured from two genotypes.
TABLE S6
List of differentially expressed genes and related annotation

This table is available and published as a supporting material in PLoS Genetics website at:
TABLE S7
List of RDR2-sensitive and RDR2-resistant 24 nt siRNA

This table is available and published as a supporting material in PLoS Genetics website at:
### TABLE S8
Primer sequences used for qRT-PCR experiment

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sequence for forward primer</th>
<th>Sequence for reverse primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>hAT</td>
<td>AATGCTATGGAGGAGACGA</td>
<td>AGCAATCCATGCTTTGTTGA</td>
</tr>
<tr>
<td>ago4a</td>
<td>GCCCACCACCTATCATGTTCC</td>
<td>TCTCGTCGGCATTTAAACCT</td>
</tr>
<tr>
<td>ago4b</td>
<td>TCGAAGTTCCTTTGGATGACA</td>
<td>TCAATCTTGGAAGCCAGAT</td>
</tr>
<tr>
<td>ago4c</td>
<td>CGCACCCAATCACCTAAAAAT</td>
<td>CCGGGCTTTAAAACATGAGTC</td>
</tr>
<tr>
<td>ddm1</td>
<td>GCCCCTGCAGAAGTAGCTTTT</td>
<td>TGGGTACCGTGAGGAGAT</td>
</tr>
<tr>
<td>liguleless3</td>
<td>GATCTGGCATGTGGAGTGAC</td>
<td>CAACCTCAGTCAGCAGATT</td>
</tr>
<tr>
<td>metl</td>
<td>TGCACCATGAGAAGTTTTCC</td>
<td>CTTAGCTTCATGGTCATGG</td>
</tr>
<tr>
<td>MULE</td>
<td>AGTGCAAAGTGCTGGACTCAA</td>
<td>CATGTTGAGTCTAAAACG</td>
</tr>
</tbody>
</table>

*Details are provided in the legend of Figure 2.*
SUPPORTING FIGURE LEGENDS

Figure S1. Overall expression fold changes of mutant versus non-mutant for differentially expressed DNA TEs. In this analysis all members of each differentially expressed super-family were treated as a group. The percentage of reads that match each super-family among all mapped reads in each genotype was calculated and the fold change was computed as the ratio of the percentage of mutant versus non-mutant for each super-family.

Figure S2. Overall fold changes of mutant versus non-mutant for differentially expressed retrotransposons. In this analysis all members of each differentially expressed family were treated as a group. The percentage of reads that match each family among all mapped reads in each genotype was calculated and the fold change was computed as the ratio of the percentage of mutant versus non-mutant for each family.

Figure S3. Distribution of numbers of mapped reads across tested genes.

Figure S4. RDR2-sensitive and RDR2-resistant 24 nt siRNAs in wild-type and mop1 mutants. The log₂ transformation of read counts in non-mutant (x-axis) versus mop1 mutant (y-axis) for each species of the 4,950 RDR2-resistant 24 nt siRNAs (green dots) and the 33,614 RDR2-sensitive 24 nt siRNAs (red dots) were plotted.

Figure S5. p-value comparison between likelihood ratio test and Fisher’s exact test.
Figure S1 Jia et al., 2009
Figure S2 Jia et al., 2009
Figure S3 Jia et al., 2009
Figure S4 Jia et al., 2009
Figure S5 Jia et al., 2009
APPENDIX B. SUPPORTING MATERIAL FOR CHAPTER 4

TABLE S1
Correlation of NAT gene sets from bioinformatics and RNA-seq approaches

<table>
<thead>
<tr>
<th>NAT genes via ssRNA-seq</th>
<th>NAT genes via Bioinformatics</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>3,625</td>
<td>6,178</td>
</tr>
<tr>
<td>No</td>
<td>14,129</td>
<td>26,362</td>
</tr>
</tbody>
</table>

| Total                   | 17,754                      | 14,786|

Chi-squared = 52, p-value = 5.9e-13
SUPPORTING FIGURE LEGENDS

**Figure S1.** Bioinformatics procedures for identifying putative genes accumulating NATs

**Figure S2.** Experimental procedures for identifying genes accumulating NATs via RNA-seq

**Figure S3.** Plot of genetic positions for eQTL markers and regulated sense and antisense transcripts
Whole maize EST sets (1,782,468)

1,400,341 Aligned to PM (79%)
95% Identity
90% coverage

1,227,590 (88%)

1,171,815 ESTs

836,704 ESTs

335,111 ESTs

415,181 ESTs

836,704 strand-specific ESTs

10,924 genes (39%)

Blast against filtered genes
95% ID and 75bp overlap on antisense strand

Figure S1 Jia et al.
RNA Isolation

mRNA Isolation x3

Bioanalyzer sample check

Decapping (TAP)

DNasI treatment

Fragment RNA

RNA size selection via gel

5' dephosphoration (CIP)

3' adaptor ligation

5' phosphoration (T4 PNK)

5' adaptor ligation

RNA size selection via gel

Reverse transcription

PCR (18 cycles)

library size selection

QC via bioanalyzer and TOPO cloning

Illumina sequencing

Figure S2 Jia et al.
Figure S3 Jia et al.
APPENDIX C. SUPPORTING MATERIAL FOR CHAPTER 5

<table>
<thead>
<tr>
<th>RNA sample pairs&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Cell source</th>
<th>Estimated no. of collected cells&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Estimated amount of RNA used for amplification&lt;sup&gt;c&lt;/sup&gt; (ng)</th>
<th>Amount of amplified RNA (μg)</th>
<th>Microarrays that the RNA samples were used for&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L1</td>
<td>1,700 (5)&lt;sup&gt;e&lt;/sup&gt;</td>
<td>1.0</td>
<td>85</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>3,800 (5)</td>
<td>1.0</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>L1</td>
<td>2,100 (8)</td>
<td>1.3</td>
<td>76</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>7,000 (8)</td>
<td>1.3</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>L1</td>
<td>2,000 (8)</td>
<td>1.2</td>
<td>71</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>6,200 (8)</td>
<td>1.2</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>L1</td>
<td>1,600 (9)</td>
<td>1.0</td>
<td>33</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>5,200 (9)</td>
<td>1.0</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>L1</td>
<td>1,600 (6)</td>
<td>1.0</td>
<td>52</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>4,600 (6)</td>
<td>1.0</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>L1</td>
<td>1,400 (5)</td>
<td>0.8</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>4,200 (5)</td>
<td>0.8</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>L1</td>
<td>1,500 (7)</td>
<td>0.9</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>5,100 (7)</td>
<td>0.9</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>L1</td>
<td>1,800 (8)</td>
<td>1.1</td>
<td>76</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>6,400 (8)</td>
<td>1.1</td>
<td>70</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>These pairs were used for independent biological replications.

<sup>b</sup>Cell numbers were estimated as described previously (Ohtsu et al., 2007a).

<sup>c</sup>All of the RNA that was extracted from the L1 (~1 ng) was used for RNA amplification, whereas only the same amount of RNA extracted from the L2 (~1 ng) was used for RNA amplification within each replication, which was approximately one third of the total amount of RNA that was extracted from the L2.

<sup>d</sup>Six biological replications were used for each of the three microarrays, SAM1.0 (1), SAM2.0 (2) and SAM3.0 (3). Eight micrograms of each amplified RNA (aRNA) sample were labeled for the microarray hybridization except for RNA sample pairs 2 and 3 for SAM1.0 (5 μg of each aRNA sample were labeled).

<sup>e</sup>Numbers in parentheses indicate numbers of the SAMs used for collecting the L1 and the L2 tissues. These numbers were not necessarily in proportion to the numbers of collected cells mainly because in some cases more sections from a single SAM were not used for tissue collection due to ambiguous morphology than in other cases.
### TABLE S2

Alignment of RNA seq reads to genome and genes

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Run</th>
<th>No. Reads&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Total Reads</th>
<th>No. Uniquely Mapped Reads</th>
<th>No. Genes</th>
<th>Complexity&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Total Gene&lt;sup&gt;e&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To Genome&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To Gene&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAM_L1</td>
<td>1</td>
<td>2,062,887</td>
<td>3,871,980</td>
<td>3,092,818 (80%)</td>
<td>2,617,839 (85%)</td>
<td>20,681</td>
<td>7,902</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1,809,393</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAM_L2</td>
<td>1</td>
<td>6,031,517</td>
<td>10,213,960</td>
<td>8,564,638 (84%)</td>
<td>7,377,516 (86%)</td>
<td>19,343</td>
<td>2,622</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4,182,443</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Number of reads after trimming low quality reads; <sup>b</sup>Number (percent) of reads that uniquely mapped to B73 reference genome with ratio (genome/all); <sup>c</sup>Number (percent) of reads that uniquely mapped to genes; <sup>d</sup>Complexity was measured as the number of genes sampled by every one million of RNA-seq in each library; <sup>e</sup>Union of genes sampled via RNA-seq by two tissue types.
TABLE S3
Correlation between microarray and RNA-seq experiments

<table>
<thead>
<tr>
<th></th>
<th>No. DE Genes (Arrays)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Up</td>
<td>Down</td>
</tr>
<tr>
<td>No. DE Genes (Arrays)</td>
<td>Up</td>
<td>112 (60)</td>
</tr>
<tr>
<td></td>
<td>Down</td>
<td>11 (63)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>123</td>
</tr>
</tbody>
</table>

Chi-squared value =177 and p-value < 2e-16. *numbers in parentheses are the expected counts under the null hypothesis of no association.
Table S4
Identified significant genes between L1 and L2

This table is uploaded into the ProQuest/UMI database at
### TABLE S5

Significant genes that were annotated as lipid-related genes

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Lipid metabolism</th>
<th>Lipid signal</th>
<th>Lipid transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of significant genes</td>
<td>63</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Up-regulated in the L1</td>
<td>43 (16)</td>
<td>12 (5)</td>
<td>5 (1)</td>
</tr>
<tr>
<td>Up-regulated in the L2</td>
<td>20 (0)</td>
<td>6 (0)</td>
<td>1 (0)</td>
</tr>
</tbody>
</table>

Numbers of L1-up or L2-up-regulated genes (FDR <0.05) that were annotated as lipid metabolism components, lipid signal or genes encoding lipid transfer proteins are presented. Parentheses indicate the number of genes showing more than 10 fold changes.
TABLE S6
Differentially expressed genes annotated involving in chromatin modifications

This table is uploaded into the ProQuest/UMI database at
TABLE S7
Overpresentation of chromatin-related genes

<table>
<thead>
<tr>
<th>No. Chromatin genes</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>145 (84)</td>
<td>5,594 (5,655)</td>
<td>5,739</td>
</tr>
<tr>
<td>No</td>
<td>180 (241)</td>
<td>16,280 (16,219)</td>
<td>16,460</td>
</tr>
</tbody>
</table>

Chi-squared value =59.6 and p-value=1.2e-14. a numbers in parentheses are the expected counts under the null hypothesis of no association.
### TABLE S8
Control SAM genes showing differential expression between L1 and L2

<table>
<thead>
<tr>
<th>Name</th>
<th>Gene.ID</th>
<th>Pattern</th>
<th>FDR</th>
<th>FC</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZmOCL1</td>
<td>GRMZM2G026643</td>
<td>L1&gt;L2</td>
<td>5.9e-162</td>
<td>3.9</td>
<td>Array</td>
</tr>
<tr>
<td>ZmOCL3</td>
<td>GRMZM2G116658</td>
<td>L1&gt;L2</td>
<td>3.8e-107</td>
<td>32.8</td>
<td>Array, RNA-seq</td>
</tr>
<tr>
<td>ZmOCL4</td>
<td>GRMZM2G123140</td>
<td>L1&gt;L2</td>
<td>2.4e-300</td>
<td>51.9</td>
<td>Array, RNA-seq</td>
</tr>
<tr>
<td>ZmOCL5</td>
<td>GRMZM2G004957</td>
<td>L1&gt;L2</td>
<td>9.5e-61</td>
<td>27.6</td>
<td>RNA-seq</td>
</tr>
<tr>
<td>Knotted-1</td>
<td>GRMZM2G017087</td>
<td>L1&lt;L2</td>
<td>0</td>
<td>11.2</td>
<td>Array, RNA-seq</td>
</tr>
<tr>
<td>Wus</td>
<td>GRMZM2G047448</td>
<td>L1&lt;L2</td>
<td>0.04</td>
<td>8.5</td>
<td>RNA-seq</td>
</tr>
<tr>
<td>ZmPIN1a</td>
<td>GRMZM2G098643</td>
<td>L1&lt;L2</td>
<td>7e-106</td>
<td>4.8</td>
<td>Array, RNA-seq</td>
</tr>
<tr>
<td>ZmPIN1b</td>
<td>GRMZM2G074267</td>
<td>L1&lt;L2</td>
<td>8e-12</td>
<td>3.6</td>
<td>RNA-seq</td>
</tr>
</tbody>
</table>
TABLE S9
Cross-reference analysis of domain-specific gene expression in maize SAMs

<table>
<thead>
<tr>
<th>Domain expression</th>
<th>L1 vs. L2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UP</td>
<td>DOWN</td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>114</td>
</tr>
<tr>
<td>Total</td>
<td>151</td>
<td>183</td>
</tr>
</tbody>
</table>
Supported figure legend

**Figure S1.** Quality check of amplified RNA (aRNA) samples. The results of the aRNA sample from the L1 (RNA sample pair 5, Table S1) are presented. (A) RT-PCR products that were amplified with the specific primers for the maize β-6 tubulin (tub6) gene and RT products derived from reactions with (RT+) and without (RT-) reverse transcriptase were separated on a 1.2% (w/v) agarose gel. The expected size of the RT-PCR product is 236 bp. (B) An image of an aRNA sample in a denaturing agarose gel. Three micrograms of aRNA were loaded.

**Figure S2.** Correlation between microarray and RNA-seq experiments. Overlap of differentially expressed genes (A) and foldchange correlation (B) between the two technologies.

**Figure S3.** Predicted gene structures and deduced amino acid sequences of the maize WW genes. (A) Schematic representations of structures of the maize WW genes. These structures are based on gene predictions on the corresponding MAGIs and the ESTs: MAGIv3.1_25235 and BM078110 for WW1, MAGIv3.1_22781 and 60825, DY401079 and DN228961 for WW2, MAGIv3.1_86529 for WW3. Black boxes represent coding sequences. One intron was predicted at a conserved position (connected by dotted lines) for each of these genes. Intron positions of WW1 and WW2 were supported by the ESTs. Arrows on WW1 and WW2 indicate positions of PolyA starting sites, which were predicted based on terminal polyA (T) sequences of BM078110 and DY401079, respectively. Arrowheads on each of the genes indicate positions of primers that were used for qRT-PCR and RT-PCR analyses (see Methods). (B) Alignment of deduced amino acid sequences of the maize WW genes. Identical and conserved amino acids are boxed with black and grey, respectively. A deduced amino acid sequence (accession no. ABA96334) of one of the rice predicted genes that exhibited significant similarities to the maize WW proteins is also presented. Core amino acid sequences of the WW domains are indicated with bars above the alignment. Important amino acid residues in the domain are highlighted with asterisks.

**Figure S4.** Expression of the maize WW genes in various maize organs. RT-PCR analyses were conducted using RNA extracted from various maize organs and primers designed to amplify transcripts from the maize WW genes (Methods, Figure S2). Primers for the maize tub6 gene were also used with RT+ and RT- reactions (see Figure S1). Each of the primer pairs detected transcripts with the expected size except for the tub6 primers with RT- reactions. The organs analyzed were: 1, 59DAPL brace root; 2, 79DAPL mixed tissues including tassel, husk, silk, ear with ear shank, leaf with ligule and oracle, and node with pith and leaf sheath; 3, 79DAPL husk; 4, 79DAPL silk; 5, 65DAPL tassel; 6, 65DAPL ear; 7, 15DAP kernel; 8, LCM-collected SAM (Ohtsu et al., 2007a). DAPL, days after planting; DAP, days after pollination.

**Figure S5.** Scatter plot of expression levels for DE genes
Expression level was calculated from the log transformed proportion of reads for a given gene within the library. Different colors indicate different levels of foldchange as noted.
Figure S1 Jia, et al.,
Figure S2 Jia, et al.,
Figure S3 Jia, et al.,
Figure S4 Jia, et al.,
Figure S5. Jia, et al
ACKNOWLEDGEMENTS

I would like to first express my special appreciation to my advisor, Dr. Patrick Schnable, for his great guidance and support during my professional and personal development. I also thank Dr. Dan Nettleton for his precious time and helpful suggestions that contributed to my dissertation and my other POS committee members: Drs Hui-Hsien Chou, Thomas Peterson and Steven Whitham for their guidance and helpful suggestions. I am grateful to Dr. Damon Lisch for his expert contributions to the mop1 paper and helpful suggestions for my professional development. I also thank Dr. An-Ping Hsia, Dr. Yan Fu, Dr. Sanzhen Liu, Eddy Yeh, and other members of Schnable Laboratory for thoughtful scientific discussion and help. For their support, I thank my parents, Gengxian Jia and Lingyun Meng. Finally, I am sincerely grateful to my wife, Xiujuan Wang, for her understanding and encouragement, to my daughter, Carolyn, and my son, Daniel, for the joys they bring to me.