A next generation of studies: Heterosis and inbreeding depression

Shreyartha Mukherjee
Iowa State University

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

Part of the Agriculture Commons, Bioinformatics Commons, and the Statistics and Probability Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Graduate College at Digital Repository @ Iowa State University. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Digital Repository @ Iowa State University. For more information, please contact digirep@iastate.edu.
A next generation of studies: Heterosis and inbreeding depression

by

Shreyartha Mukherjee

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

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
William Beavis, Co-Major Professor
Julie Dickerson, Co-Major Professor
Paul Scott
Peng Liu
Thomas Lubberstedt

Iowa State University
Ames, Iowa
2013

Copyright © Shreyartha Mukherjee, 2013. All rights reserved.
DEDICATION

I dedicate this dissertation to my father Siddhartha, my mother Alpana, my sister Shreeja and my friend Chandni without whose support I would not have been able to complete this work.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEDICATION</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>vi</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vii</td>
</tr>
<tr>
<td>CHAPTER 1</td>
<td>1</td>
</tr>
<tr>
<td>INTRODUCTION AND LITERATURE REVIEW</td>
<td>1</td>
</tr>
<tr>
<td>Heterosis and Production Agriculture</td>
<td>1</td>
</tr>
<tr>
<td>Heterotic Biological Systems</td>
<td>4</td>
</tr>
<tr>
<td>Maize as the Model to Study Heterosis</td>
<td>7</td>
</tr>
<tr>
<td>A Next Generation of Heterosis Investigations</td>
<td>11</td>
</tr>
<tr>
<td>Connecting grain yield and transcript levels</td>
<td>12</td>
</tr>
<tr>
<td>Organization of thesis</td>
<td>13</td>
</tr>
<tr>
<td>Difference in transcript counts (differential gene expression)</td>
<td>13</td>
</tr>
<tr>
<td>Abundance of sequence variants (Allele-specific expression)</td>
<td>14</td>
</tr>
<tr>
<td>Abundance of splice variants (alternatively spliced isoform expression)</td>
<td>14</td>
</tr>
<tr>
<td>Statement of Relevance</td>
<td>15</td>
</tr>
<tr>
<td>Author affiliations</td>
<td>16</td>
</tr>
<tr>
<td>References</td>
<td>17</td>
</tr>
<tr>
<td>CHAPTER 2</td>
<td>26</td>
</tr>
<tr>
<td>TRANSCRIPTOMIC ANALYSIS OF FIELD GROWN ELITE MAIZE INBREDS AND THEIR HYBRIDS</td>
<td>26</td>
</tr>
<tr>
<td>Abstract</td>
<td>26</td>
</tr>
<tr>
<td>Introduction</td>
<td>27</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>32</td>
</tr>
<tr>
<td>Experimental Design</td>
<td>32</td>
</tr>
<tr>
<td>Tissue sampling</td>
<td>33</td>
</tr>
<tr>
<td>RNA Extraction</td>
<td>34</td>
</tr>
<tr>
<td>Sequencing and Bioinformatic Analyses</td>
<td>34</td>
</tr>
<tr>
<td>Normalization of expression data and assessment of over-dispersion</td>
<td>35</td>
</tr>
<tr>
<td>Modeling and analyses of gene expression counts</td>
<td>36</td>
</tr>
<tr>
<td>Assessment of gene expression models</td>
<td>37</td>
</tr>
<tr>
<td>Modeling and analyses of ear traits</td>
<td>40</td>
</tr>
<tr>
<td>Data Analyses</td>
<td>41</td>
</tr>
<tr>
<td>Implementation of Data Analyses</td>
<td>41</td>
</tr>
<tr>
<td>Results</td>
<td>42</td>
</tr>
<tr>
<td>Model Implementation</td>
<td>42</td>
</tr>
<tr>
<td>Goodness of Fit</td>
<td>42</td>
</tr>
<tr>
<td>Comparisons of Expression in Hybrids vs Mid Parent values</td>
<td>43</td>
</tr>
<tr>
<td>Genetic Distance and Differential Expression</td>
<td>47</td>
</tr>
<tr>
<td>Functional Analysis of Gene Enrichment Categories for</td>
<td>48</td>
</tr>
<tr>
<td>Differentially Expressed Genes</td>
<td></td>
</tr>
<tr>
<td>Evaluation of whole ear phenotypes for Mid-parent and</td>
<td>53</td>
</tr>
<tr>
<td>High-parent heterosis</td>
<td></td>
</tr>
<tr>
<td>Discussions</td>
<td>55</td>
</tr>
</tbody>
</table>
Mapping reads to AgpV2 ................................................................. 55
Normalization and modeling of biological variability .................. 55
Association between genetic diversity and transcriptional variation .............................................................................. 56
Modes of gene action .................................................................. 56
Functional annotation of differentially expressed genes .......... 58
References ................................................................................... 60

CHAPTER 3 ...................................................................................... 66

ALLELE-SPECIFIC EXPRESSION IS CONSERVED IN MAIZE HYBRIDS .......................................................................... 66
Abstract ....................................................................................... 66
Introduction .................................................................................. 67
Materials and Methods ................................................................. 73
Genotypic Materials .................................................................... 73
Field Plot Design .......................................................................... 74
Tissue sampling ........................................................................... 75
RNA Extraction ............................................................................ 75
Sequencing and Bioinformatics Analyses ...................................... 76
Normalization and modeling ......................................................... 77
SNP calling and variant detection ............................................... 78
Assessment of allele specific counts ........................................... 78
Assessment of differential expression ........................................ 80
Modes of gene action .................................................................. 80
Cis and trans regulation from allele specific counts ................. 81
Results ......................................................................................... 82
Global differential expression analysis ....................................... 82
Modes of gene action .................................................................. 83
Allele specific expression ............................................................ 84
Mono-allelic expression in hybrids ............................................. 85
Assessment of cis and trans regulatory divergence ................. 86
from allele specific expression ................................................ 88
Associating regulatory divergence and global modes of gene action .............................................................................. 88
Discussions .................................................................................. 89
References ................................................................................... 94

CHAPTER 4 ...................................................................................... 103

DETECTION AND QUANTIFICATION OF SPLICE VARIANTS IN...
MAIZE INBREDS AND HYBRIDS FROM RNA-SEQ DATA .................................................................................. 103
Abstract ....................................................................................... 103
Introduction .................................................................................. 104
Methods ....................................................................................... 111
Normalization and modeling biological variability .................. 111
Gene Model Flattening ................................................................. 112
Isoform specific expression with Cufflinks .............................. 113
Methods for classification of AS events .................................... 114
Intron retention ............................................................................. 114
Alternative 3’ and 5’ splice sites ................................................. 116
Exon skipping .............................................................................. 117
Mutually exclusive exons ........................................................... 118
Results ......................................................................................... 119
ACKNOWLEDGEMENTS

I would like to thank my major professors, Dr. William Beavis and Dr. Julie Dickerson, and my committee members, Dr. Paul Scott, Dr. Peng Liu, and Dr. Thomas Lubberstedt, for their supervision and support throughout the course of this research.

In addition, I would also like to thank my friends, colleagues, the department faculty and staff for making my time at Iowa State University a wonderful experience. Finally, thanks to my family for their constant encouragement, love and support.
ABSTRACT

Heterosis or hybrid vigor is a phenomenon in which the F1 progeny is superior in characteristics than its inbred parents. Inbreeding depression refers to the reduced vigor in the offspring from matings involving related parents. The world’s most important grain crop, Zea mays exhibits significant heterosis and inbreeding depression, and despite billions of dollars of sales and investments in hybrid maize, the underlying molecular and genetic mechanisms responsible for inbreeding depression and heterosis still remain unknown. We used next generation sequencing technologies and a sample of the most elite publicly available germplasm to explore the role of gene expression in hybrids vs. inbreds. A comparison of transcript counts from inbred parents and their hybrids revealed additive gene action to be the most prevalent among differentially expressed genes. Gene expression that deviated from additivity showed high parent dominance, suggesting that the alleles from the high parent affected the expression in the hybrid more than the alleles from the low parent. Most differentially expressed genes were inconsistent among families, however at a higher level of organization these sets of genes belonged to the same metabolic pathways and were up-regulated in different parents and hybrids, indicating that organisms often utilize compensatory/complementary genetic networks to perform the same task. Allele-specific expression analysis for all inbred-hybrid combinations showed that a majority of alleles that were preferentially expressed in the hybrids also were expressed differentially between its inbred parents. Cis-trans statistical tests revealed that although most alleles exhibit conserved expression, cis-regulation was found to affect alleles more than trans-regulation. Cis regulation was found to be strongly correlated with additive gene action identified from the
transcript expression study. In order to explain lowered protein metabolism seen in hybrids, we developed novel software pipelines to investigate alternative splicing patterns in inbreds and hybrids. Results showed that the hybrid preferentially produces parental splice forms, but for a small fraction of genes, the hybrids produce splice forms that are not observed in the parental set. We also found 88% of the genes expressed just a single isoform, while other isoforms are expressed trivially. The exact reason for preferential expression of a single isoform remains to be determined, we hypothesize that the protein structure coded by the most prevalent splice isoform is the most stable.

A direct extension of our work can help identify mis-folded proteins in the hybrids and the alleles that code them and eliminate these alleles through breeding strategies.
CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Heterosis and Production Agriculture

The development of hybrid maize (Zea mays L) (Duvick, 2001) is the most significant milestone in farming history (Reif et. al., 2003). In the nineteenth century, it was common for farmers to visually inspect their cornfields and save ears from individual desireable plants for planting the next year’s fields. This method resulted in heterogenous phenotypes with yields of about 30 bu/ac (Duvick, 1985). While Darwin and other naturalists of the 19th century recognized heterosis, or hybrid vigor, it was not until the beginning of twentieth century for heterosis to be exploited through planned development of hybrid varieties. The earliest recorded experiments on heterosis were conducted by (Shull, 1908; Shull, 1911) at Cold Spring Harbour in NY (Reif et. al., 2003) and (East, 1908a) at Connecticut State College. They observed that when plants were self-pollinated, the performance of their progeny, in terms of growth and grain yield, deteriorated. However, when two unrelated inbred lines were crossed the growth and yield characteristics of the F1 hybrid progeny exceeded that of the best parent. Hybrid maize was almost immediately accepted by farmers because hybrid phenotypes were homogeneous, provided better yields per acre and it was more convenient to let a supplier manage collection, storage and packaging of seed for the new crop. From the farmer’s perspective the only negative attribute was the large inbreeding depression exhibited in the subsequent F2 generation, which discouraged farmers with limited cash, from saving seed to plant the the next year’s fields.
Heterosis is evident in nature as well (Mitton et. al., 1998; Hansson and Westerberg, 2002). In coniferous trees, when randomly sampled, allelic frequencies are in Hardy-Weinberg equilibrium, but subsets of the most mature, oldest and largest trees express an overt evidence of
It is apparent that the fittest trees were established through a natural selection process and these are heterozygous, but the underlying molecular mechanisms responsible for the selective advantage have not been identified.

In the seminal heterosis conference held in 1952 at Iowa State University, Shull defined heterosis as “the interpretation of increased vigour, size, fruitfulness, speed of development, resistance to disease and to insect pests, or to climatic rigors of any kind manifested by crossbred organisms as compared with corresponding inbreds, as the specific result of unlikeness in the constitution of the uniting parental gametes” (Reif et al., 2005). In simple objective terms, heterosis may be defined as the difference in characteristics between the hybrid and the average of its two parents (Schnell, 1961). The term panmictic midparent heterosis (PMPH) was proposed by (Lamkey and Edwards, 1999) as expression for the difference between the characteristics of a hybrid population and the mean of its two parental populations in Hardy-Weinberg equilibrium. It is a function of net dominant effect and is computed as the square of the difference in allelic frequencies between the two populations (Falconer and Mackay, 1996). This indicates that, with increasing genetic distance, an increase in PMPH is expected (Reif et al., 2005). Studies on single cross hybrids and their parental inbreds fit within this broader definition.

Early suggestions for grouping breeding germplasms based on the degree of heterotic performance as measured by lower specific combining ability (SCA) relative to general combining ability (GCA) ratio has been advantageous because it emphasized heterotic breeding based on GCA (Reif et al., 2005). The concept of heterotic breeding, termed reciprocal recurrent selection (RRS) was pioneered by (Comstock et al., 1949) to help select maize recombinant lines in order to simplify germplasm management and organization (Reif et al., 2005). Lines
with superior testcross performance were passed through repeated selective cross breeding to create desired breeding pools based on heterotic responses (Lu and Bernardo, 2001). As a result, all maize currently grown for intensive production agriculture are hybrids derived from crosses between pairs of homozygous inbred parents. The inbred parents are derived from breeding pools that are reproductively isolated from each other. In North America the two primary maize breeding pools are referred to as Stiff Stalk (SS) and Non-Stiff Stalk (NSS).

Heterotic Biological Systems

The phenomenon of heterosis is well documented, but its underlying mechanisms are less understood, despite testing of proposed hypotheses for over a century. Three classical hypotheses, namely, dominance (Bruce, 1910; Keeble, 1910; Jones, 1917; Collins, 1921) over dominance (East, 1908b; Shull, 1908; Shull, 1911; Shull, 1946) and epistasis (Minvielle; Schnell and Cockerham, 1992) have been promoted as explanations for heterosis.

Figure 2. Two possible explanations of the genetic basis of heterosis viz. dominance (A) and over-dominance (B). (Birchler et. al., 2003)
The dominance hypothesis first proposed by (Davenport 1908), states that the functional allele for a gene is contributed by only one dominant parent, i.e., heterosis is due to dominant alleles from either parent cancelling the genetic effects of deleterious recessive alleles contributed by the other parent in the heterozygous hybrid. This happens in a genome-wide fashion and as a result, more favorable genes are expressed in the hybrid than either of the inbred parents. On the other hand, over-dominance theory states that the combination of two alleles at a given locus produces a genetic effect superior to that of either of the homozygous combinations of those alleles at that locus. So the two alleles complement each other and there is overexpression of a particular set of genes in the heterozygote. The presence of loci with overdominant phenotypic effect has been demonstrated in maize (Hollick and Chandler, 1998). The dominance and overdominance debate is further complicated by evidence of a third phenomenon, epistasis, which is defined as interactions between genes at two or more loci, influencing the phenotypic expression of a trait. The anthocyanin biosynthesis pathway in maize provides an excellent example of transcriptional epistasis (Springer and Stupar, 2007).

There are suggestions that a combination of the dominance, over-dominance and epistasis models may be responsible because over the years, geneticists have struggled to experimentally validate any to the exclusion of the others (Fehr 1987). There have been two experimental approaches to discriminate among the models (Hallauer and Miranda). First degree statistical analyses of data from generation means or diallele analyses have been utilized to quantify the net effect of dominance and epistatic interactions of heterotic loci in expression of heterosis. However, while demonstration of heterosis was possible qualitatively, a quantitative comparison of dominance at individual loci may cancel each other, corrupting the net statistical result.
Second degree statistics have been used to calculate variance components contributed by additive, dominance and epistatic effects from covariances of relatives. Such quantitative genetic studies have indicated that overdominance and epistasis are not major contributors to heterotic expression in maize (Hinze and Lamkey, 2003; Mihaljevic et. al., 2005). However, approaches based on calculations of variance components also confound contributions of the various genetic effects. Significant proportions of the net dominance effects as well as net epistatic effects can be included in estimates of additive variance (Cheverud and Routman, 1996). Despite the weaknesses of these ‘net effect’ biometric approaches, at the end of the 20th Century the dominance hypothesis emerged as the preferred model for explaining heterosis (Crow, 1998; Coors and Pandey, 1999). However, it was widely recognized as inadequate because it could not explain progressive heterosis or rapid rate of inbreeding depression in tetraploids as well as absence of a decline in degree of heterosis over 50 years of genetic improvements (Duvick, 2001; Birchler et. al., 2003).

With the development of high throughput molecular technologies there has been a renewed interest in heterosis and inbreeding depression. The various ‘omics’ technologies provide the ability to move from ‘net effect’ models to models of gene expression at individual loci throughout the entire genome. (Hugemon et. al., 1967) proposed the concept of metabolic balance as the cause of heterosis. It is further suggested that inbred lines are impaired by an imbalance in metabolic enzymes. Supra or suboptimal enzyme activities fail to complement the process of biosynthesis and eventually undermine their metabolic flux, while hybrids between complementary inbreds, through reciprocal summation, are likely to have a more balanced metabolic flux (Baer and Schrader, 1985; Hageman et. al., 1988). Theoretically this concept extends the metabolic flux theory proposed by (Novotný et. al., 1987; Kacser et. al., 1995) but to
date metabolomic evidence to support such a proposition is insufficient (Reif et. al., 2005). (Fu and Dooner, 2002) found that functional genes are often absent in maize lines, and lines lacking different genes can still complement one another in the F1 hybrid supporting the dominance hypothesis for heterotic expression (Reif et. al., 2005).

DNA methylation activity has been found to be related to gene expression (Tsaftaris, 1995). Methylation is less in hybrids than their parental inbreds, new elite inbreds show less DNA methylation than older inbreds, which might explain why more advanced inbreds exhibit less inbreeding depression than older inbreds (Tsaftaris et. al., 2000).

At the organelle level, increased efficiency in mitochondrial oxidative photophosphorylation of ATP has been claimed by some as another mechanism for hybrid vigor (McDaniel and Sarkissian, 1968; Srivastava, 1981). Early biometric studies on nuclear – cytoplasmic interactions tried to find evidence to support this conjecture with minimal success (Beavis et. al., 1987; Asmussen et. al., 1989) and there are no molecular data to support the concept of nuclear-cytoplasmic heterotic complementation (Reif, Hallauer et al. 2005).

Maize as the Model to Study Heterosis
Virtually all annual crops exhibit some degree of heterosis. Some, such as Sorghum, Sunflower and Rice require cytoplasmic male sterility to produce sufficient quantities of hybrid seed to take advantage of the phenomenon. Chinese researchers (Zhang et. al., 1994; Yu et. al., 1997; Hua et. al., 2003; Zhang and Borevitz, 2009) have been pursuing an understanding of heterosis in hybrid rice because in addition to yield per hectare, elite hybrids display greater resistance to both biotic and abiotic stress relative to inbred parents across a wide range of spatial and temporal environments (Goff and Zhang, 2013). In addition to creating useful genetic resources to study
heterosis in rice, a number of advanced technological resource platforms including extensive cloning and resequencing for high throughput molecular analyses have been developed (Huang et. al., 2012; Jiang et. al., 2012).

It has been apparent for a century that the phenomenon of heterosis in maize is extremely large for grain yield. Yields in the hybrids can be as much as three times that of the highest inbred parents. Also, maize is easy to self or cross pollinate, and produce large numbers of both selfed and hybrid progeny for replicated field plot evaluations. Maize also exhibits enormous genetic variability approximately an order of magnitude greater than in humans (Bhattramakki et. al., 2000; Sunyaev et. al., 2000; Buckler Iv and Thornsberry, 2002; Ching et. al., 2002). This diversity includes Single Nucleotide Polymorphisms (SNPs), insertions, deletions, copy number variants, and repetitive elements that often span several kilobases of DNA fragments (Fu and Dooner, 2002; Flint-Garcia et. al., 2003; Wright et. al., 2005; Troyer, 2006).
In North America, maize, has been used as a model for heterotic research for the past century. Over three decades ago, B73xMo17 was the most widely grown hybrid in the United States. Because this was a hybrid derived from two publicly available inbreds, academic researchers have tended to use B73, Mo17 and the hybrid progeny as the genetic materials for studies on heterosis. For example, seedling tissues from B73, Mo17, and the F1 hybrid or a sample of recombinant inbred lines derived from this hybrid, sampled under growth chamber conditions and assayed using microarrays have been used in most gene expression studies on heterosis in maize (Reif et. al., 2005).

In the late 1990’s, the development of robust molecular markers allowed mapping of quantitative trait loci (QTL) providing a tool for studying the underlying genetic architectures of complex genetic traits. In a pathbreaking study, (Stuber et. al., 1992) mapped QTL and its impact on heterosis for grain yield with a Design III using F3 lines from the elite hybrid B73 x Mo17. They identified 11 QTL for grain yield, most of them displaying over dominance or pseudo-over dominance. However, a reanalysis of the same data using a different biometric model demonstrated dominance and epistasis between linked loci indicating a model dependent on variation of the findings (Stuber et. al., 1992). Fine mapping of one major QTL on chromosome 5 dissected this region into two smaller QTL’s in repulsion phase linkage, which also demonstrated dominance (Graham et. al., 1997;Reif et. al., 2005). In another Design III study in maize, (Lu et. al., 2003) proposed that, QTL for grain yield in maize exhibit true overdominance, and secondly, QTL for grain yield also show partial to complete dominance but their specific effects could not be identified due to linkage disequilibrium. Due to the use of large numbers of markers in QTL studies, fine mapping has become an essential adjunct to
understand the underlying genetic architecture for complex quantitatively measured phenotypes (Jansen et. al., 2003).

Continued improvements to sequencing technologies have enabled high resolution analysis of the structural diversity of the B73 maize genome (Buckler Iv and Thornsberry, 2002; Messing and Dooner, 2006; Schnable et. al., 2009). BAC based studies revealed a high degree of SNP and short indel diversities within intraspecific loci. (Fu and Dooner, 2002). Other maize loci sequenced in multiple inbred lines have also shown structural diversity (Song and Messing, 2003; Schnable et. al., 2012).

(Ohtsu et. al., 2007) and (Brooks III et. al., 2009) utilized laser capture microdissection of shoot apical meristems from seedlings of the maize inbred line B73 and found expression of many regulatory genes involving transcription, chromosome remodeling, and gene silencing. Subsequent studies on genome wide transcript profiles of B73, Mo17 and their reciprocal hybrid seedlings produced in a growth chamber (Swanson-Wagner et. al., 2006), or in a green house (Stupar and Springer, 2006) have suggested that the majority of transcript levels and transcript expressions show additive action. These studies pointed to an association between heterosis and transcript expression but could not establish a clear connection between heterotic phenotypes and gene expression. Moreover, these expression studies relied on microarray technology which does not provide a quantitative assessment of transcript expression or give us a measure of differential expression of transcript isoforms.
A Next Generation of Heterosis Investigations

Of late, there has been an increasing demand for technologies that produce quick, low-cost and more precise genome information than microarrays. This challenge has been partially addressed with next-generation sequencing (NGS) technologies. The low-cost production of large sequence datasets is the primary improvement over conventional microarray methods. RNA-seq, also known as "Whole Transcriptome Shotgun Sequencing" ("WTSS") uses high-throughput sequencing to provide researchers with a method to count transcripts experimentally. Additionally, the sequence data provides useful information about allele specific expression, abundance and variety of gene isoforms and alternative splicing. With the additional information it should be possible to infer underlying molecular mechanisms responsible for heterosis, provided the samples are taken from appropriate genotypes and tissues.

Figure 4. Investigating the connection between gene expression and the phenotype by studying RNA-transcript levels.
Connecting grain yield and transcript levels

In an effort to better understand genetic factors responsible for grain yield (a heterotic trait) researchers have partitioned the trait into component traits related to ear morphology such as ear length, number of kernels per ear and kernel mass. Ear morphology traits are likely to be established early in development of lateral branches. (Vollbrecht et. al., 2005; Bortiri et. al., 2006; Satoh-Nagasawa et. al., 2006) demonstrated that the *ramosa*1, *ramosa*2, and *ramosa*3 mutants were expressed during lateral branch initiation in the spikelet-pair meristem at the apex of the developing ear shoot. Mutations at the *ramosa* loci result in severe distortion in lateral branch development. Hence, we thought it reasonable to explore whether wild type genotypic variation in global genetic expression of spikelet-pair meristems, are associated with phenotypic expression of row number and kernels per row, in the mature ears grown under field conditions.

The objectives of our research were (1) to assay genome-wide gene transcript expression in the ear shoot apical meristem by means of deep-sequencing technology from tissue sampled under field conditions in a generation means design comprised of genetic materials derived from multiple, highly adapted inbred lines; (2) estimate transcript expression genetic and cis-/trans-regulatory effects; (3) ascertain the degree to which these effects accounted for genotypic variation in gene expression in the ear shoot apical meristem; and (4) determine if genotypic variation in transcript expression in the ear shoot meristem was associated with genotypic variation in mature ear phenotypes.

Dr. Richard Johnson developed a set of lines with increasing levels heterozygosity (or inbreeding) from a diallel involving four homozygous lines, two from each of the primary maize breeding populations. The Johnson materials consist of LH1, LH123HT, PHG39, and PHG84 and their respective F1, F2, and reciprocal backcross families. Thus, nested within each of the six crosses comprising the diallel of all four inbreds were six generations with various levels of
inbreeding: P1, P2, F1, F2, BCP1 and BCP2. These materials were grown under replicated field conditions for purposes of investigating possible relationships between phenotype and transcripts across levels of inbreeding within and among six maize families representing different levels of heterotic relationships.

Organization of thesis

Herein we report bioinformatic analysis to investigate the various types of differential gene expression between the inbred parents and their hybrids using a subset of the Johnson panel.

<table>
<thead>
<tr>
<th>Gene Expression levels</th>
<th>Comparisons</th>
<th>To understand</th>
<th>Thesis Chapters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference in transcript counts</td>
<td>compare F1 expression to mid-parent expression</td>
<td>mode of gene-action</td>
<td>Chapter 2</td>
</tr>
<tr>
<td>Alleles-specific expression in F1 hybrids</td>
<td>compare expression of parental alleles in F1</td>
<td>types of allele regulation</td>
<td>Chapter 3</td>
</tr>
<tr>
<td>Differential isoform expression</td>
<td>compare abundance of isoforms in inbreds and F1</td>
<td>types of alternative splicing</td>
<td>Chapter 4</td>
</tr>
</tbody>
</table>

Table 1. Thesis organization

We investigated three different manifestations of gene expression.

**Difference in transcript counts (differential gene expression)**

In an attempt to identify candidate genes responsible for hybrid vigor, there has been a number of gene expression studies of hybrids vs their inbred parents for maize (Kollipara et. al., 2002; Guo et. al., 2004; Auger et. al., 2005; Bao et. al., 2005; Hoecker et. al., 2005; Guo et. al., 2006; Huang et. al., 2006; Swanson-Wagner et. al., 2006; Meyer and Scholten, 2007; Springer and
Maize studies have suggested thousands of differentially expressed genes between the parents and the hybrids. Some of these studies report widespread additive gene action (Li et. al., 2010) while other studies have shown a high volume of non-additive gene action (Stupar et. al., 2007; Stupar et. al., 2008; Chaudhary et. al., 2009). We investigated if specific groups of genes are differentially expressed across families and inbreeding generations from replicates of plants grown under field conditions.

**Abundance of sequence variants (Allele-specific expression)**

Recent studies have shown that preferential expression of alleles is widespread in mammals. Non-imprinted autosomal genes exhibit allelic imbalance at the transcript level in mouse hybrids (Cowles et. al., 2002) and humans (Yan et. al., 2002) and such expression produces proteins associated with diseases. Hence a solid understanding of classification and functional annotation of allele-specifically expressed genes is vital to recognize the extent of functionally important regulatory variation. This will help us identify candidate haplotypes and the correlation between their genetic sequences and heterotic traits. In our study, we developed a novel approach to study allele-specific expression and identified alleles that are preferentially expressed across genetic backgrounds and levels of inbreeding.

**Abundance of splice variants (alternatively spliced isoform expression)**

Alternative splicing (or differential splicing) is a process by which the exons and introns of a gene are preferentially retained or discarded during transcription, providing the gene with the capability of producing more than one protein. Humans use an estimated 100,000 proteins derived from no more than 25,000 genes. Young (Li et. al., 2011) obtained RNA-Seq data from
cultured human B-cells and identified 879 million 50-bp reads comprising 44 Gb of sequence. They identified 20,766 genes and 67,453 of their alternatively spliced transcripts. (Wang and Sigworth, 2009) looked at a span of read matches at either side of an exon junction to determine splice junctions. In their study of the human transcriptome, (Pan et al., 2008) reported discovery of over 4000 novel splice junctions. Defining the location of genes and the precise nature of gene products remains a fundamental challenge in genomics. High throughput tandem mass spectrometry based proteomics provides an important new source of information to help define both the location of transcription units and the reading frame of protein translation.

Herein we report reads mapping to the splice junctions in the B73 reference maize genome (AgpV2) and reported over-representation of some splice junctions. We identified alternatively spliced isoforms and investigated whether isoforms are preferentially expressed. We categorized splicing events into the five broad categories and looked at splicing patterns of inbreds and hybrids.

Statement of Relevance

Maize breeders employ ever increasing resources in developing new inbred lines and evaluating their potential in hybrids. Evaluation of hybrids requires field trials that are cost and time-intensive. Thus, while genetic improvements continue to increase over time, they are decreasing relative to resource expenditures. The phenomenon of heterosis has been the subject of intense research and application in the field of agro-genetics in the last one hundred years and has made remarkable progress over the past decade through use of molecular technologies resulting in the potential for predictive crop breeding. This has far reaching impact on global food production for a burgeoning human population. With the recent advancements of technology related to genomics, transcriptomics, proteomics and metabolomics, investigations
on understanding the molecular basis of heterosis are being revisited and new hypotheses are emerging to explain the phenomenon. The results of our study has significant consequences, and the potential to aid breeding methods in improving hybrid performance efficiently.

Author affiliations

Mukherjee S\(^{4,5}\)*, Johnson GR\(^1\), Scott PS\(^2\), Moran Lauter, A\(^2\), Farmer AD\(^3\), May GD\(^6\), Lamkey KR\(^4\) and Beavis WD\(^{4,5}\)

1 Department of Crop Sciences
   University of Illinois, Urbana, Illinois 61801, USA
2 Corn Insects and Crop Genetics Research Unit,
   USDA-ARS, Ames, IA 50011, USA
3 National Center for Genome Resources,
   2935 Rodeo Park Dr E, Santa Fe, NM 87505, USA
4 Department of Agronomy
   Iowa State University, Ames, Iowa 50011, USA
5 Bioinformatics and Computational Biology Program
   Iowa State University, Ames, Iowa 50011, USA
6 Pioneer Hi-Bred International, Inc.,
   Johnston, IA 50131-0184, USA
*E-mail: shreymuk@iastate.edu (primary author on all chapter manuscripts)
References


G. Collins (1921). TEOSINTE IN MEXICO The closest wild relative of Maize is Teosinte—a forage plant hitherto known only as an annual. A perennial form discovered in Southern Mexico should prove of value to the breeder. Journal of Heredity. 12: 339-350.


F. Minvielle Dominance is not necessary for heterosis: a two-locus model.


maize based on morphological investigations of heterosis for plant height. Plant molecular biology. 63: 21-34.


CHAPTER 2

TRANSCRIPTOMIC ANALYSIS OF FIELD GROWN ELITE MAIZE INBREDS AND THEIR HYBRIDS

A paper to be submitted to The Plant Genome journal

Mukherjee S, Johnson GR, Scott PS, Moran Lauter, A, Farmer AD, May GD, Lamkey KR and Beavis WD

Abstract

Heterosis or hybrid vigor is a phenomenon in which the F1 progeny exhibits values for traits that exceed the values for the same traits in its inbred parents. Maize is one of the most important commercial crops and is an ideal system to study heterosis. In this manuscript, we evaluated transcript expression profiles and whole-plant traits from six different inbred-hybrid families grown under replicated field conditions. We illustrate the use of appropriate normalization and modeling to avoid misinterpretation of results of RNA-seq data obtained from field plots. Differential expression analysis revealed that nearly 80% of differentially expressed genes between inbreds and hybrids in all families exhibit additive expression patterns, while 90% of the whole plant traits in the F1 exhibit phenotypic values that exceed either parent. Of the 20% of non-additively expressed genes, 84% of values in the F1 progeny are near the high parent (dominance and partial dominance). Genetic distance between inbreds was found to be strongly associated with transcriptional variation, suggesting that genetic diversity between inbred parents is associated with gene expression in single cross hybrids. Separate gene-sets were differentially expressed among heterotic families and non-heterotic families. These same gene-sets were over-represented with common functional annotations, suggesting that different sets of genes are
differentially expressed between inbred parents and their progeny, but the differential expression is specific to the family.

Introduction

Inbreeding is a term used to describe matings between relatives. Inbreeding depression refers to the reduced vigor in the offspring of matings between relatives. Increased vigor of progeny from unrelated inbred parents is referred to as either hybrid vigor or heterosis. In many plant species it is possible to produce viable offspring by crossing an individual with itself for dozens of generations. In most grain crops, (rice, soybean, wheat, etc.) this extreme form of inbreeding exhibits very little inbreeding depression. However, the world’s most important grain crop, *Zea mays* exhibits significant inbreeding depression in progeny from self-pollinations. Commercial seed producers have utilized this phenomenon to promote sales of hybrids, thus assuring continual investments for genetic improvement of hybrids. Despite billions of dollars of annual sales and investments in research and development of hybrid maize, the underlying molecular and genetic mechanisms responsible for inbreeding depression (heterosis) still remain unknown.

Soon after rediscovery of Mendel’s publication, three hypotheses of heterosis and inbreeding depression were posed. The dominance hypothesis, proposed by Davenport (1908), Bruce (1910), and (Keeble, 1910), is that favorable dominant alleles at loci responsible for the trait cancel the deleterious recessive alleles at the same loci in the heterozygous F1. Because each parent contributes a subset of the favorable alleles, the performance of the F1 will be heterotic relative to its inbred parents. According to the dominance hypothesis, it should be possible for breeders to select for favorable dominant alleles through breeding strategies and produce inbred lines equal in performance to the most superior hybrids (Budak et. al., 2002).
However, due to the enormous number of loci contributing to overall performance of a trait, breeding superior inbreds has been challenging (Tsaftaris et. al., 2005). (Birchler et. al., 2010) indicated that a simplified dominance model would mean that the number of deleterious recessives in inbreds will increase with greater phylogenetic distance. A corollary of the model is that different parents have different sets of loci with deleterious alleles and their hybrid progeny have more favorable alleles at complimentary loci than either of their parents (Fu and Dooner, 2002). A second theory, known as the over-dominance hypothesis states that the combination of two alleles at a given locus produces a superior effect relative to either of the homozygous combinations of alleles at the locus. This means there are particular sets of heterozygous loci that confer superior performance on the phenotypes of the hybrids relative to their inbred parents. Conceptually, this represents a synergy between the alleles. Over-dominance also has been referred to as “single-gene heterosis”, “super-dominance”, and “stimulation of divergent alleles”.

A third hypothesis for heterosis combines additive and dominant gene action with genetic background. (Wright et. al., 2003) proposed that variation of traits depend on a network of genes and replacing a single allele can affect all members of the genetic network. This was demonstrated by (Fasoulas and Allard), who showed that epistasis, i.e., interactions among alleles at multiple loci, were responsible for phenotypic variability in crosses between near isogenic barley lines. Note that all three hypotheses attempt to explain heterosis based on genetics, i.e., allelic interactions responsible for regulating the structural and functional phenotypic characteristics of progeny from crosses of related or unrelated parents.

Gene expression can be considered as a measurable phenotype. If the expression values in the hybrid are near the average values (mid-parent) of the two inbred parents, it represents additive gene expression. If the expression values in the hybrid are closer to either of
the parents than mid-parent value, it is called dominant or partially dominant gene expression. 
And if the value is more extreme than either inbred parent, then it will be referred to as over-
dominant gene expression. More formally, consider two inbred parents with alleles \( P_1 \) (low) and 
\( P_2 \) (high) at a single locus. Let the genotypic values for low parent be \( P_1P_1=z \), high parent be 
\( P_2P_2=z+2a \) and for the hybrid be \( P_1P_2 = z+a+d \). The mid-parent value can be calculated as the 
mean of parental genetic values 
\[
M=\frac{1}{2}(z + (z+2a)) \\
M=z+a
\]

\[\text{OD}]=\text{Over-dominance, } \text{LPD=}\text{Low Parent Dominance, } \text{HPD=}\text{High Parent Dominance }, A=\text{Additive}\]

Figure 1. Illustration of different genetic effects

The value ‘\( a \)’ can be referred to as the additive effect at the locus, and can be estimated 
from the difference of the Mid Parent value and the low parent value. As the magnitude of \( d \) 
becomes larger, the hybrid expression value moves away from the mid-parent and towards either 
of the parents depending on the sign of \( d \). The dominant effect, \( d \), at the locus can be estimated as 
the difference between genotypic values of the hybrid and the Mid-parent: 
\[
a=M-P1 \text{ and } d=F1-M.
\]

The value of \( d \) determines the level of dominance in the hybrid. If \( a=|d| \) or \( |d/a|=1 \), the 
genotypic values of the hybrid and one of the parents become equal and this is referred to 
complete dominance. When the genotypic value of the hybrid is equal to the mid-parent value i.e 
\( d=0 \) and \( d/a=0 \), no dominance exists. If the hybrid is between the mid-parent value and one of
either of the homozygotes parents $0<d<a$ or $d/a<1$, gene expression is referred to as partial dominance. Lastly, if the genotypic value in the hybrid is significantly outside the parental range, it is called over-dominance i.e. $|d/a|>1$. Thus if we plot the distribution of $d/a$ for all genes in a hybrid, we can estimate allelic interactions throughout the genome.

Molecular studies of heterosis have employed genome-wide transcript analysis. Early gene expression studies reported higher quantities of mRNA and protein in the hybrid compared to their inbred parents (Romagnoli et. al., 1990; Leonardi et. al., 1991; Tsaftaris, 1995; Tsaftaris et. al. (1999)) and suggested higher gene expression could be a possible explanation. In maize, most of these studies used microarray technologies to quantify gene expression patterns of apical meristems in seedlings from B73 X Mo17, its parents, or a sample of recombinant inbred lines from this cross, sampled under controlled environments. Expression profiles of B73, Mo17 and their hybrids apical meristems in seedlings grown in controlled chambers (Swanson-Wagner et. al., 2006) or seedlings, immature ears and embryos produced in the greenhouse (Stupar and Springer, 2006) have suggested that the majority of gene expression differences between inbreds and hybrids are explained with additive genetic models; i.e., the gene expression in the hybrid was intermediate to expression differences between inbred parents. These expression studies relied on the microarray technology which provides a relative measure of transcript expression based on colorimetric intensities, but does not provide a discrete quantitative measure of differential expression of transcript isoforms.

Because of the inadequacies of microarray technologies, there has been a need for technologies that produce quick, low-cost and more precise quantitative expression information. The challenge has been partially addressed with next-generation sequencing (NGS) technologies. RNA-seq uses high-throughput sequencing technologies to provide counts of transcripts as well
as references to their distinct sequences. Importantly, sequence-based data has the potential to provide information about allele specific expression, abundance and variety of gene isoforms and alternative splicing. Thus, it should be possible to combine sequence data and transcript counts to infer underlying molecular mechanisms responsible for inbreeding depression, or heterosis, if the samples are taken from appropriate genotypes and tissues grown under field conditions.

A critical aspect of interpreting results is to properly model and normalize the data. It is important to note that data generated by NGS have their own technical and systemic biases, thus normalization methods developed for microarrays are not applicable for these data. There can be four primary types of variability associated with RNA-seq. These are i) within-lane, ii) between lanes, iii) within sample and iv) between samples variability. Appropriate normalization methods must be applied to account for these sources of variability based on the nature of the biological objective and the experimental design. While conducting the research reported herein, several unique approaches were developed to remove bias associated with RNA-seq (Baggerly et. al., 2004; Robinson and Smyth, 2007a; Robinson and Smyth, 2008a; Anders and Huber, 2010; Bullard et. al., 2010; McCarthy et. al., 2012).

The objective of this manuscript is to report on application of appropriate normalization techniques for RNA-seq data obtained from field plots and the resulting inferences of associations between transcript levels and heterosis among lines having varying levels of genetic diversity and heterotic phenotypes. Our approach is novel in that we use a replicated field-based design to accommodate a diallel mating design consisting of parents and F1 progeny. This enabled the construction of quantitative genetic models to evaluate transcript levels and whole plant phenotypes on a plot mean basis. Transcript expression was assayed using next generation mRNA sequencing (RNA-seq) technology (Marioni et. al., 2008; Bashir et. al.,
2010; Montgomery et al., 2010) and enabled by alignment of sequence reads to a version of the maize genome reference sequence.

Materials and Methods

Experimental Design

Maize exhibits significant inbreeding depression resulting in differences among generations in plant size and vigor of development. Consequently, in order to avoid biasing differences among entries belonging to different generations with inter-plot competition effects, a replicated split-split field plot design (Kempthorne, 1952), was utilized. Whole plot entries were classified by generation: Parental and $F_1$ where the parental inbred entries and $F_1$ hybrids were placed into different whole plots. Each whole plot was divided into two sampling split plots: one to sample ear shoot meristem tissues for transcript profiling and a second to sample mature ears for whole plant phenotyping. Within the $F_1$ plot were six split-split plot entries, each representing a different genetic background, i.e., family. Within the split plots of the parental inbred whole plots were split-split plot entries representing each of the four parental inbreds. To facilitate planting and sampling in the field, the four parental inbred whole plots were augmented with two additional inbred line entries so that all whole plots were apportioned to the same land area. Assignments of generations to whole plots, sampling types to split plots, and entries to split-split plot entries were done randomly with three replicates.

Each split-split plot consisted of two rows, each 5.2 m long, spaced 0.76 m apart. At 60,000 plants/ha, each row consisted of 24 plants. Adjacent split-split plots also were spaced 0.76 m apart. Each whole plot was bordered by a two-row plot containing an entry from a generation consistent with the whole plot generation.
The experiment was planted on May 12, 2009, on the Iowa State University Agronomy Farm on well-drained Clarion silt loam soil. Nitrogen fertilizer at a rate of 92 kg/ha had been applied prior to planting. Pre-planting and two post-emergence herbicide treatments were applied to control weeds. Plots were overplanted and thinned to a stand equivalent to 60,000 plants/ha. The growing season at the experiment site in 2009 was relatively cool with ample rainfall (http://mesonet.agron.iastate.edu/climodat/index.phtml?station=IA0200&report=03).

**Tissue sampling**

Beginning 14 July tissues were sampled on consecutive days through 21 July. To achieve relative temporal constancy of gene expression during sampling, all sampling was done between 10:00 a.m. and noon, CDT. A plant was judged suitable for sampling if the developing tassel was detectable in the whorl of upper leaves. The stalk of a selected plant was cleaved just above ground level, and the hewed plant was carried to the edge of the field for dissection. There, leaves were stripped from the stalk and a section containing the uppermost ear leaf node was sectioned. The sectioned piece was then split longitudinally by a cut transverse to the node, and stalk tissue around the ear shoot meristems was carefully removed. The ear shoot meristems were removed from the stalks and those between 0.8 cm and 1.2 cm in length were placed on the stage of a dissecting microscope where a sterile scalpel was used to separate the apical 1 mm section of the ear shoot meristem. The apical section (weighing about 2 mg) was placed in a microfuge tube on frozen CO₂, transferred to liquid nitrogen. Sampling of a given split-split plot ceased as soon as ten shoot tips had been successfully processed. The samples were taken to the laboratory and stored in a freezer maintained at a temperature of –80⁰ C.
RNA Extraction

Before processing, all ten shoot tips from a given plot were bulked. The bulked samples were randomly assigned to four batches of approximately equal size. Total RNA was extracted using Qiagen’s RNeasy Plant Mini kit (#74904) according to manufacturer’s instructions, adhering to a method based on guanidine thiocyanate extraction with Buffer RLT, followed by purification using supplied spin columns. Tissue was homogenized in a micro-centrifuge tube using a Kontes microtube pestle. RNA was eluted by addition of 50 µl of water, incubated for 10 minutes at room temperature, and then centrifuged. Each bulk sample yielded approximately 10 µg of RNA, at a concentration of 200 ng/µl. RNA was successfully extracted from 83 out of the 84 samples. Samples were randomized for sequencing-by-synthesis.

Sequencing and Bioinformatic Analyses

Polyadenylated RNA was isolated from total RNA using oligo-dT25 magnetic beads (Dynabeads; Invitrogen, Carlsbad, CA). The captured RNA was denatured and annealed to random hexamer primers and reverse transcribed. After end repair and A-tailing, adapters complementary to sequencing primers were ligated to cDNA fragments. The resultant cDNA libraries were size selected by agarose gel electrophoresis, then amplified by PCR for 15 cycles. Sample libraries were qualitatively and quantitatively assessed by Nanodrop ND-1000 (Thermo Scientific, Waltham, MA) UV/Vis spectroscopy and DNA BioAnalyzer 2100 microfluidics (Agilent, Santa Clara, CA). The samples from the cDNA libraries were loaded on an Illumina® single-end flow cell using the Illumina® Cluster Station (Illumina®, Inc., San Diego, CA). 36 bp reads were obtained from an Illumina® Genome Analyzer II where sequenced bases were determined by the Illumina® base-calling pipeline consisting of analysis, base-calling, quality filtering and assignment of per base quality scores.
Genomic alignments of the resulting 36mer reads were performed using GSNAP (Wu and Nacu, 2010) against the maizesequence.org AgpV2 version of the B73 maize genome. The alignments, using GSNAP’s splicing detection mode, identified introns of up to 10Kbp in length, two base mismatches or a single indel. A read was characterized as unique if it aligned with a single location (locus) in the genome. A total of 38,178 unique alignments were available for subsequent data analyses. An alignment was assigned to a gene if the alignment overlapped the genomic locus span with at least 95% identity and over 90% of the transcript length. Annotations of the maize sequence.org AgpV2 alignments are listed in Gene Ontology Annotation (GO Annotation: http://www.ebi.ac.uk/GOA/). Genomic alignments were then converted to expression counts using the AgpV2 annotations.

**Normalization of expression data and assessment of over-dispersion**

Herein, we evaluated relative changes in expression profiles across homozygous inbred parents and heterozygous hybrids for all families. We only need to account for sample specific factors; technical factors unrelated to the experimental design influenced all the samples equally. For example, GC content and gene length are known to be important contributors and directly proportional to the read count of a gene, but these factors do not change from sample to sample and hence need not be corrected for analysis of differential expression. All mRNA samples were sequenced using one lane each and thus the most obvious step was to equalize the library sizes assuming equal sequencing depth. (Bullard et. al., 2010) and (Robinson et. al., 2011) proposed that genes with very high number of transcripts can represent a substantial proportion of the library. Thus counts from other genes in that library will be underestimated. Unless this bias is accounted for, some genes will appear to be down-regulated for the sample and will give rise to false detection of significant differences among biological samples. We used the TMM
(trimmed mean of M values) method proposed and implemented in the package edgeR to adjust for this bias.

**Modeling and analyses of gene expression counts**

Previous studies (Robinson and Smyth, 2007b; Baggerly et. al., 2008; Robinson and Smyth, 2008a; Anders and Huber, 2010; Bullard et. al., 2010; McCarthy et. al., 2012) have established the use of negative binomial distributions to model RNA-seq counts that account for the over-dispersion problem due to biological variation among replicates. It is possible to model appropriate distributional assumptions with edgeR (version 3.2.4).

Sample specific variability of count data detected by the RNA-seq technology could be based on a Poisson process. However, the variance of transcript counts appeared to be over-dispersed relative to the observed average counts among biological replicates. Thus, a single parameter Poisson model is inappropriate for these data. Alternatively, the negative binomial in which there is a dispersion parameter included in the model appears to be more appropriate for these data. (Lu et. al., 2003; Baggerly and Coombes, 2009; Zhou et. al., 2011) have suggested the use of Generalized linear models with the read counts treated as samples from an over-dispersed negative binomial distribution. Generalized linear models require appropriate modeling of the mean-variance relationship. Herein, we used a log linear model (McCarthy et. al., 2012) for each gene per family, to estimate the read counts, i.e.,

\[ \log(\mu_{gr}) = \mu + \rho_r + \beta_g + \epsilon_{gr}, \quad (1a); \]

where \( \rho_r \) specifies replicate \( r \) (one of three biological replicates), \( \beta_g \) is the effect of the \( g^{th} \) split-split plot, which specifies family member \( g \) (parent u, parent v, or the \( F_1 \) progeny) and \( \epsilon_{gr} \) denotes the residual variability. Assuming a negative binomial distribution for the read counts, the link
function is log, allowing the variability of read counts to be linearly related to the logarithm of its predicted value.

QQ plots were used to evaluate goodness of fit for each of the families. QQ plots provide a visual plot of the model predicted values (quantiles) against the observed values (quantiles). The closer the model predicted values are to the observed values, the closer the data points will be to the diagonal.

**Assessment of gene expression models**

To visualize global genetic expression, we plotted count differences between each inbred and their hybrid as a Parental-F1 bi-plot (Swanson-Wagner et. al., 2009) (Figure 2). Thus, on the X axis, we plot the difference of Parent1 and the F1 and on the Y axis we plot the difference of Parent2 and the F1. Genes (indicated by points on this plot) can be categorized according to the different counts between each of the inbred parents and the hybrid and enable a global assessment of gene expression in terms of models that describe genetic effects.
Figure 2. Parental-F1 Bi-plot for visualizing modes of gene action. The difference of the hybrid gene expression values with that of either of its parents is plotted against each other. Quadrants are labeled according to the mode of gene action determined by the location of the hybrid expression value with respect to its parents. Labels for Parent 1 are plotted in blue, Parent 2 in red and the F1 in purple.

**Complete dominance:** If the expression of a gene in the hybrid is not significantly different from its expression in the dominant (high) parent but significantly different from its expression in the recessive (low) parent, then the gene action is categorized as complete dominance. Because of the significant (and negative) difference between only one parent and the hybrid, these genes are plotted close to the negative X and negative Y axes.
**Complete suppression:** If the expression of a gene in the hybrid is not significantly different from its expression in the recessive (low) parent but significantly different from its expression in the dominant (high) parent, then this type of dominant gene action is categorized as complete suppression. Because of positive difference between only one parent and the hybrid, these genes are plotted close to the positive X and positive Y axes.

**Over and under dominance:** If the expression values of the genes are significantly different in the hybrid than the two parents and the expression values are similar between the parents, then the gene action is categorized as over (under) expressed. These genes are plotted close to the 45 degree diagonal (first and third quadrants), points in the first quadrant indicate that the inbred parents were over-expressed relative to the hybrid while points in the third quadrant indicate over expression in the hybrid relative to the parents (heterosis).

**Additive:** If the expression value of a gene in the hybrid lies between that of its parents and the difference in expression value between the parents is significant, then the gene action is categorized as additive. These would be indicated by points lying close to the 45 degree diagonal in the second and fourth quadrants. For example the point (25,0) would indicate the difference in expression in Parent1 and the F1 is 25 while the gene expression is no different between the F1 and Parent2. Similarly a point (-5,-7) would indicate that expression in the F1 is higher than both parents and hence the expression at this locus exhibits over-dominance.

**Dominance by additive ratio:** Because genotypic variability can be partitioned into dominance and additive components an estimate of dominance by additive ratio can be used to evaluate prevalent gene action for all expressed genes. Estimates of $d/a$ are obtained from the difference between expression in the hybrid and the mid-parent divided by the difference between the expression in the hybrid and the high parent. Thus, a distribution of estimates of $d/a$ can be used
to evaluate gene action on a global scale. A distribution to the right or left of zero indicates the prevalence of dominance or suppression respectively.

**Modeling and analyses of ear traits**

After the plants had fully senesced, ten ears were taken at random from plants internal to the row-end plants of each split-split plot in the ear sampling split plot blocks. The ears were placed in a forced-air dryer and dried to less than 5% moisture at a temperature of about 40°C. The dried ears were evaluated for a number of traits (Table 1).

**Table 1. Description of ear phenotypes.**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
<th>Description</th>
<th>Unit of Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>cobl</td>
<td>Cob length butt</td>
<td>Length butt to tip</td>
<td>centimeters</td>
</tr>
<tr>
<td>cdb</td>
<td>Cob diameter butt</td>
<td>Diameter at butt</td>
<td>centimeters</td>
</tr>
<tr>
<td>cdm</td>
<td>Cob diameter mid</td>
<td>Diameter at mid-ear</td>
<td>centimeters</td>
</tr>
<tr>
<td>cdt</td>
<td>Cob diameter tip</td>
<td>Diameter at tip</td>
<td>centimeters</td>
</tr>
<tr>
<td>cwt</td>
<td>Cob weight butt</td>
<td>Weight of cob</td>
<td>grams</td>
</tr>
<tr>
<td>earl</td>
<td>Ear length butt</td>
<td>Length of cob bearing</td>
<td>centimeters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>kernels</td>
<td></td>
</tr>
<tr>
<td>edb</td>
<td>Ear diameter butt</td>
<td>Diameter at butt</td>
<td>centimeters</td>
</tr>
<tr>
<td>edm</td>
<td>Ear diameter mid</td>
<td>Diameter at mid-ear</td>
<td>centimeters</td>
</tr>
<tr>
<td>edt</td>
<td>Ear diameter tip</td>
<td>Diameter at tip</td>
<td>centimeters</td>
</tr>
<tr>
<td>ewt</td>
<td>Ear weight butt</td>
<td>Weight of grain + cob</td>
<td>grams</td>
</tr>
<tr>
<td>kdb</td>
<td>Kernel depth butt</td>
<td>edb - cdb</td>
<td>centimeters</td>
</tr>
<tr>
<td>kdm</td>
<td>Kernel depth mid</td>
<td>edm - cdm</td>
<td>centimeters</td>
</tr>
<tr>
<td>kdt</td>
<td>Kernel depth tip</td>
<td>edt - cdt</td>
<td>centimeters</td>
</tr>
<tr>
<td>kda</td>
<td>Kernel depth</td>
<td>$(kdb+kdm+kdt)/3$</td>
<td>centimeters</td>
</tr>
</tbody>
</table>
Data Analyses

Ear trait analyses were based on a plot-mean basis, with 10 plants sampled per plot. Analyses were conducted on each family using a General Linear Model:

\[ Y_{rg} = \mu + \rho_r + \beta_g + \epsilon_{rg} \quad (1b) \]

where \( \mu \) is the experimental mean; \( \rho_r \) = effect of the \( r^{th} \) rep; \( \beta_g \) = effect of the \( g^{th} \) split-split plot which specifies family member \( g \) (i.e., parent u, parent v or their F1 progeny) and \( \epsilon_{rg} \) is the residual variability not accounted for by the model. Note that the phenotypic traits were analyzed using a General Linear Model equivalent to the Generalized Linear Model (1a) analyses of the transcript counts.

Implementation of Data Analyses

The estimation of all effects in models (1a) and (1b) was based on a fixed effects model utilizing R. Generalized and General linear models were implemented with the \textit{lm} function in R following example code (http://data.princeton.edu/R/linearModels.html). Note that since only 83 out of a possible 84 mRNA samples were obtained, the analyses of transcript counts is slightly unbalanced.
All significance tests for gene expression counts were based on False Discovery Rate (FDR). (Benjamini and Hochberg, 1995; Benjamini, 2010), statistical method to control false positives in multiple testing. FDR corrects for the expected proportion of incorrectly rejected null-hypotheses and renders more power to detect significant results. In context of this manuscript, FDR (Benjamini and Hochberg method) was used to control false positives when we compared transcript counts for ~38,000 genes among the three family members.

Results

Model Implementation

We observed that variability among replicates affected (FDR < 0.005) only a negligible percentage of genes (0.06%) per family. Also, rep effects were not significant in the analyses of phenotypic traits. Thus replicate effects in models 1a and 1b were removed and the data reanalyzed with the more powerful reduced models.

Goodness of Fit

QQ plots (Figure 3) were used to evaluate how well the models addressed the residual variability among generations for each of the families. By modeling the data generation as a negative binomial process with a unique dispersion parameter for each gene there was little evidence for lack of fit, with the exception of a few (~20) outliers (p-value<0.05) for each family (Figure 3). We also generated QQ-plots to evaluate normal, Poisson and Negative Binomial with equal dispersion, all these models had a very high proportion of outliers (~10,000, data not shown).
Comparisons of Expression in Hybrids vs Mid Parent values

A gene transcript was considered differentially expressed if there was a significant difference (FDR < 0.005) between either of the parents and the hybrid. The number of genes detected to be differentially expressed varied across families. The non-heterotic stiff stalk family LH1-PHG39 had the least number of differentially expressed genes among all the families. Most
differentially expressed genes had hybrid expression values that were no different than the mid-parent. This indicated prevalence of additive action among differentially expressed genes. Among the non-additive genes, most were found to be up-regulated in the hybrid relative to the mid-parent.

<table>
<thead>
<tr>
<th>Regulation/Family</th>
<th>LH1-LH123HT</th>
<th>LH1-PHG39</th>
<th>LH1-PHG84</th>
<th>LH123-PHG39</th>
<th>LH123HT-PHG84</th>
<th>PHG39-PHG84</th>
</tr>
</thead>
<tbody>
<tr>
<td>down-regulated</td>
<td>160</td>
<td>74</td>
<td>147</td>
<td>133</td>
<td>108</td>
<td>150</td>
</tr>
<tr>
<td>midParent</td>
<td>3261</td>
<td>2295</td>
<td>2983</td>
<td>3666</td>
<td>3419</td>
<td>3776</td>
</tr>
<tr>
<td>up-regualted</td>
<td>1033</td>
<td>529</td>
<td>520</td>
<td>888</td>
<td>825</td>
<td>658</td>
</tr>
</tbody>
</table>

Table 2. Classification of differentially expressed genes (FDR < 0.005) into up-regulated, down-regulated and mid-parent categories, by family. Mid-Parent represents additive group and up and down-regulated genes represent the non-additive groups. The two non-heterotic families are shaded in green.
Figure 4. Parental-F1 Bi-plots for all 6 inbred-hybrid combinations. Significant genes (FDR < 0.005) plotted in red, extremely significant genes (FDR < 0.000005) plotted in green.
In Figure 4, Red and green points indicate differentially expressed genes at levels with FDR < 0.005 and 0.000005 respectively. Approximately 5000 genes for each of the families showed differential expression (FDR<0.005). In all families, most genes were plotted in the second and fourth quadrants. The more significant differentially expressed genes (plotted in green) also were found in the 2nd and 4th quadrants, but there was evidence of a tendency for these genes to exhibit dominance and partial dominance.

Figure 5. Dominance/Additive plot. Distribution of d/a values were plotted for all differentially expressed genes for the six families, genes having d/a values close to 0 are additive, those significantly away from 0 are dominant.
Based on plots of estimated $d/a$, it is apparent that most of the genes had dominance by additive ratio between 0 and 1 (Figure 5). Also most genes exhibiting additive gene action tend toward partial dominance. These results indicate that the genetic model that best describes global differential expression is partially dominant.

Figure 6. Number of differentially expressed genes for each family plotted against the Nei’s genetic distance between the inbred parents. Non heterotic combinations (SSxSS and NSSxNSS) are plotted in blue, heterotic combinations are plotted in red.

**Genetic Distance and Differential Expression**

The relationship among parental genetic diversity, based on genotypic differences, was assessed using Nei’s distance (Stupar et. al., 2008) between pairs of inbred parents. This was estimated by first normalizing the total number of mismatching loci between the parents relative to the total number of loci (obtained from sequence data with coverage of at least 2 reads
covering each base). Differentially expressed genes among genotypes within each family were identified (FDR< 0.005). The number of differentially expressed genes was variable among the families and it was observed that there was a very strong correlation between the genetic diversity of parents and the number of differentially expressed genes in the corresponding family. This result is similar to a prior study by (Stupar et. al., 2008) who used microarray technology to assess differential expression relative to Nei’s distance. Spots representing crosses between stiff stalk (SS) and non-stiff stalk (NSS) groups are shown in red, and spots representing crosses within either group are shown in blue (Figure 6). The two stiff stalk inbred parents LH1 and PHG39 had the least genetic distance (0.056) i.e., they were genetically most similar among all possible pairs. Also, the LH1-PHG39 family had the least number of differentially expressed genes (2898 genes showed significant difference between either or both parents and the hybrid ) among all groups.

Functional Analysis of Gene Enrichment Categories for Differentially Expressed Genes

Comparisons of annotated lists of differentially expressed genes among the six families revealed unique sets of differentially expressed genes for each family (Table 3). Within the one comparison of non-heterotic families, LH1xPHG39 and LH123HTxPHG84, only about 25.6% (1859) of the differentially expressed genes were consistent for both families. When non-heterotic families were compared with the heterotic families (8 comparisons, shaded in green Table 2) only about 34.4% of the differentially expressed genes were consistent between any paired comparison. Within the six heterotic pairs of comparisons between pairs of families (shaded in pink), about 38% of the differentially expressed genes were consistent and when combined across all four families only 1274 differentially expressed were consistent. There were
only 121 differentially expressed genes that were consistent across all six families (data not shown).

Table 3. Each cell in the table represents the number (upper diagonal) and frequency (percentage, lower diagonal) of differentially expressed genes that were common between pairs of families. The diagonal cells show total number of differentially expressed genes per family (column). Comparisons within non-heterotic families are in blue, within heterotic families are in pink, and between heterotic and non-heterotic families are in green.

Although the differentially expressed genes within families showed little consistency among families, we hypothesized that they had common functions. There are a number of tools available for analysis for enrichment or depletion of functional categories. We used the AgriGo (Du et. al., 2011) tool to look at our gene lists for over-representation of Gene Ontology (GO) functional categories. Plant GO Slim annotations were used for functional analysis.
The non-heterotic families had 1859 genes that were differentially expressed. We found a large number of GO functional annotations (28 GO terms) that were significantly over-represented (compared to AgpV2 background annotation) with our list. To mention a few, GO terms associated with transcription (Biological Process), lipid metabolic process (Biological Process), catalytic activity (Molecular Function), intracellular membrane-bound organelle (Cellular component) were enriched in our gene set for the non-heterotic families.
The heterotic families had 1274 genes that were differentially expressed detected (FDR<0.005). Among others, GO functional annotations associated with cellular metabolic process (Biological Process), translation (Biological Process), transporter activity (Molecular Function), Intracellular non-membrane-bounded organelle (Cellular Component) were significantly enriched in our gene set for the heterotic families.
Figure 9. Example of Differentially Expressed genes from heterotic LH1xPHG84 non-heterotic LH123HT-PHG84 families mapped to the biological processes pathway (GO: Biological Process). Boxes are colored according to degree of over-representation in the list of DE genes (white-no representation to deep red-strong representation).

Differentially Expressed genes from each family were mapped onto the Gene Ontology biological process pathway (GO:P). Although gene lists were over-represented with functional annotations, genes were dispersed over the network, i.e., no particular pathway consistently was over-represented across all families.
Evaluation of whole ear phenotypes for Mid-parent and High-parent heterosis

Table 4.
Significance of Mid-parent Heterosis.

<table>
<thead>
<tr>
<th></th>
<th>LH1- PHG39 (SSxSS)</th>
<th>LH123- PHG39 (NSS x NSS)</th>
<th>LH1- PHG84</th>
<th>LH1- PHG84</th>
<th>PHG39- PHG84</th>
</tr>
</thead>
<tbody>
<tr>
<td>cobl</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>cdb</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>cdm</td>
<td>x</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>cdt</td>
<td>x</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>cwt</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>earl</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>edb</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>edm</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>edt</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>ewt</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>kdb</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>kdm</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>kdt</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>kda</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>nrow</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>kpr</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>wpk</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>gw</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

* denotes significant at FDR < 0.005

x denotes FDR >= 0.005
We also evaluated significant positive differences between the F1 and the High-parent. Most ear traits had significant positive differences (results summarized in Table 5).

Table 5. Results for High-parent heterosis

<table>
<thead>
<tr>
<th></th>
<th>LH1-PHG39 (SSxSS)</th>
<th>LH123-PHG39 (NSS x NSS)</th>
<th>LH123-PHG84</th>
<th>LH1-LH123</th>
<th>LH1-PHG84</th>
<th>PHG39-PHG84</th>
</tr>
</thead>
<tbody>
<tr>
<td>cobl</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>cdb</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>cdm</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>*</td>
<td>* x</td>
<td></td>
</tr>
<tr>
<td>cdt</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>x</td>
<td>* x</td>
<td></td>
</tr>
<tr>
<td>cwt</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>earl</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>edb</td>
<td>x</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>edm</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>edt</td>
<td>x</td>
<td>x</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>ewt</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>kdb</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>kdm</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>kdt</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>kda</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>nrow</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>kpr</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>wpk</td>
<td>*</td>
<td>*</td>
<td>x</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>gw</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

* denotes significant at FDR < 0.005  
 x denotes FDR >=0.005

Analyses reveal that ear-trait phenotypes are best modeled as dominant to over-dominant.
Mapping reads to AgpV2

We used maizesequence.org AgpV2 version of the B73 maize genome (AGPV2) for mapping the reads. 61% of the quality filtered reads could be mapped to the reference genome, 58% were uniquely mapped. It is likely that differences between the B73 reference and our elite inbred lines resulted in a large percentage of reads (39%) to be discarded. Also, a sizeable fraction of reads mapping to the intronic and intergenic regions confirm that there are many novel transcribed regions and splice forms that the B73 reference genome does not describe. Maize is reported to have enormous genetic diversity (Whitt et. al., 2002; Tenaillon et. al., 2004; Wang et. al., 2009) and our mapping indicates that there is need for multiple reference genomes for maize.

Normalization and modeling of biological variability

Normalizing and modeling RNA-seq data appropriately are two of the most important steps for analysis and interpretation of results. Initially, when we assumed the mapped read counts to be normally distributed (data not shown) we identified ~13000 genes per family (FDR<0.05) to be differentially expressed between the inbred parents and the hybrid. After correcting for sample specific biases and modeling counts as being sampled from a negative-binomial distribution, we identified about 7000 genes to be differentially expressed (FDR <0.05) per family. Specifically, we modeled an unequal dispersion parameter for each gene. We also found that assuming equal dispersion (equal mean and variance suggested by the Poisson distribution) for all genes was inappropriate for modeling our data. Based on the goodness of fit analysis (QQ plots), most genes fit well to the over dispersed negative binomial model while there were a much larger proportion of outliers when other models were evaluated (Results not...
shown). Our results are consistent with other studies which modeled RNA-seq data using an over-dispersed negative binomial distribution. (Smyth, 2004; Robinson and Smyth, 2008a; Robinson and Smyth, 2008b; Anders and Huber, 2010; Bullard et. al., 2010; Hardcastle and Kelly, 2010; McCarthy et. al., 2012). Using proper normalization techniques to remove unwanted biases and modeling the data using appropriate statistical distribution and mean-variance relationship was important for detecting repeatable underlying biological signals for samples taken from the replicated field trials.

**Association between genetic diversity and transcriptional variation**

A larger number of genes were differentially expressed in hybrids derived from distantly related inbreds. Consistent with SNP based sequence analysis by (Meyer and Scholten, 2007) and (Stupar et. al., 2008), we observed a strong association between transcriptional variation (number of differentially expressed genes) within families and genetic diversity (genetic distance measured by Nei’s distance) between the two inbred parents of the family. (Messing and Dooner, 2006) reported high level of polymorphism differences between inbreds can be partly responsible for differences in gene expression. (Stupar et. al., 2007) hypothesized that a high number of polymorphisms in the intergenic regions might regulate transcriptional variation. We are not able to test either of these hypotheses with these data.

**Modes of gene action**

Genetically distant parents produced more differentially expressed genes, both additive and non-additive. Most differentially expressed genes between the inbred parents and the hybrid showed additive gene action in all six families. Proportions of additive and non-additive genes were fairly consistent across all families indicating there is no association between additive and non-additive profiles with the heterotic pools responsible for producing these hybrids. Non-
additive genes in the hybrid exhibited expression between the mid-parent and either of the parents can be categorized as partially dominant. There was no common genetic trend to separate heterotic combinations from non-heterotic combinations, but there was a trend of some many genes toward dominant and partially dominant gene action.

In contrast to the expression counts, the predominant form of gene action for the ear phenotypes was best modeled as over-dominant (Table 4 and Table 5). The lack of consistency between gene action models in whole plant phenotypes and differentially expressed transcripts was also observed by (Stupar et. al., 2008). Clearly, the underlying genetic mechanisms for heterosis in maize are not due to a simple association between transcript levels and phenotypes in the developing ears.

A previous study by (Swanson-Wagner et. al., 2006) showed that additive gene action was the most common mode of gene action among differentially expressed genes between inbred parents in the B73xMo17 hybrid. (Stupar et. al., 2008) confirmed these results in a similar study using six genetically diverse inbreds and their hybrids. Our results confirmed these previous studies with similar global results using six families derived from elite germplasm. On the other hand, there are reports of a prevalence of non-additive expression patterns for differentially expressed genes in hybrids (Gibson and Dworkin, 2004;Ranz et. al., 2004;Auger et. al., 2005;Vuylsteke et. al., 2005;Huang et. al., 2006). Differences in modeling gene action among reports are difficult to reconcile because the methods and species are confounded.

Exhibition of partial dominance meant that the allele from the high-parent had a greater influence on the hybrid than the allele from the low-parent. One of the various advantages of RNA-seq technology is that it makes available sequence information along with transcript counts, so it is possible to distinguish between the two parental alleles in the hybrid. By
evaluating read coverage at polymorphic sites, it is possible to investigate if an allele from one
parent is preferentially expressed over the other parental allele in the hybrid. (Wittkopp et. al.,
2004; McManus et. al., 2010) Studying allele-specific expression is beyond the scope of this
manuscript and is reported in a subsequent manuscript.

**Functional annotation of differentially expressed genes**

Most prior studies of heterosis and gene expression in maize have been based on B73 and
Mo17, both of which were retired from breeding pools at least 25 years ago. Thus, we were
interested in whether inferences from these initial studies were relevant to advanced elite
hybrids. As noted above, our inferences from modeling gene action at a global level are
consistent with prior studies. On the other hand the sets of annotated genes that are differentially
expressed in the ear shoot meristems were not consistent among families. In particular, the
comparison of the two non-heterotic families showed the least consistent set of differentially
expressed genes. This may reflect a distinctive pattern of differential gene expression between
inbreds and their progeny within the two breeding pools. But, this hypothesis, based on a single
observation, clearly needs to be evaluated with a much larger sample. While there was greater
consistency of DE among families representing heterotic crosses, the level of consistency is not
much greater than if the sets of genes were randomly drawn from a pool of independent
differentially expressed genes. While somewhat disappointing, these results are consistent with
reports (Stupar et. al., 2007; Stupar et. al., 2008; Li et. al., 2009) where DE genes in inbreds and
hybrids were reported to vary depending on the nature of the experiment and genetic
background. Further, a review of these studies by (Goff et. al., 2011) indicated that the sets of
differentially expressed genes were not associated with any particular biochemical pathways or
functions. Our observations were consistent with these previous studies, although each of the
gene-sets detected to be DE for the six families were over-represented with functional GO annotations, genes were dispersed in the GO biological process network and no pathway was particularly enriched across all six families.

Another notable observation was that down-regulation of a gene in a biochemical pathway doesn’t always imply down-regulation of the entire pathway; alternate genes belonging to the same pathway can potentially be up-regulated and perform the same functions. Thus different genotypes can maintain regulation of biochemical pathways using different genetic networks. A hypothesis to test would be the hybrids, having a choice of more alleles than the inbreds, spend minimum energy to maintain and regulate biochemical processes, whereas inbreds spend much more energy towards metabolism and are thus stunted in growth and other characteristics.
References


O. Kempthorne (1952). The design and analysis of experiments.


CHAPTER 3

ALLELE-SPECIFIC EXPRESSION IS CONSERVED IN MAIZE HYBRIDS

A paper to be submitted to The Plant Genome journal

Mukherjee S*, Johnson GR, Scott PS, Farmer AD, May GD, Lamkey KR and Beavis WD

Abstract

Heterosis in plants refers to the phenomenon in which mating between diverse varieties of homozygous inbred parents produces homogeneous heterozygous offspring that exhibit greater biomass, fertility and growth rates than in the parents. Long-standing theories attempt to explain the basis of heterosis as either due to the sum of dominant alleles across multiple loci or over-dominant interactions between alleles within genetic loci. Next generation sequencing technologies enable us to identify parental alleles and trace their expression in the hybrids. In this manuscript, we quantified whole transcript and allele-specific expression differences between four unrelated maize inbred lines and their six hybrid progeny. A survey of cis and trans regulatory elements revealed that most alleles are expressed in conserved fashion, where there is no expression difference between the parents as well as between the parental alleles in the hybrid. For genes that are differentially expressed among inbreds and hybrids Cis-regulation was found to have a greater effect than trans-regulation. Also for most genes which showed preferential expression in the hybrid of one parental allele, were also differentially expressed between the parents indicating that there are specific sets of genes that show allele-specific expression in parents and their hybrid progeny.
Introduction

Advances in technologies for evaluating gene expression over the last decade have given experimental biologists powerful tools to investigate molecular mechanisms underlying phenotypic variability. Sequence variants, variability of quantities of expressed genes, timing of expressed genes, and alternatively spliced transcripts are among the many molecular characteristics that have been associated with phenotypic diversity (Schadt et. al., 2003; Carroll et. al., 2005; Hoekstra and Coyne, 2007; Wray et. al., 2007; Stern and Orgogozo, 2008).

Phenotypic variability between maize homozygous lines and their heterozygous F1 offspring is enormous and are referred to as either inbreeding depression or heterosis. Two primary hypotheses for inbreeding depression (heterosis) have been promoted: 1. The dominance hypothesis proposed by (Davenport, 1908; Bruce, 1910), and (Keeble, 1910) states that a functional allele is contributed by one parent. It assumes that an allele from one parent dominates the effect of deleterious recessive alleles contributed by the other parent in the heterozygous hybrid. In the hybrid favorable alleles are contributed by both inbred parents at multiple complimentary loci. 2. The over-dominance hypothesis states that the combination of two alleles at a given locus produces an effect that is superior to either of the homozygotes at loci responsible for expression of the trait. As a result both alleles at a locus contribute to an overexpression of a measurable trait in the heterozygote. Although these hypotheses were proposed over 100 years ago, geneticists have struggled to experimentally verify either (Fehr, 1987).
Several microarray based studies compared total gene expression differences between inbreds and hybrids of maize. Early gene expression studies reported higher volumes of mRNA and protein in the hybrid relative to the inbred parents (Romagnoli et. al., 1990; Leonardi et. al., 1991; Tsaftaris, 1995; Tsaftaris et. al., 1999) and suggested higher gene expression in the hybrid could be a possible explanation. Most of these studies sampled mRNA from apical meristems of seedlings of B73 x Mo17, its parents, or a sample of recombinant inbred lines derived from this cross. All of these studies grew the plants in controlled environments. Expression profiles of seedlings, immature ears and embryos of B73, Mo17 and their F1 grown in controlled chambers (Swanson-Wagner et. al., 2006) or in a greenhouse (Stupar and Springer, 2006) have suggested that the majority of expression differences between inbreds and hybrids can be explained with simple additive models i.e., the gene expression in the hybrid was intermediate to both differences between inbred parents.

Mukherjee et al (unpublished) evaluated gene expression in developing ear meristems of six families using RNA-seq. This technology provides a reproducible quantitative measure of transcript numbers from each sample. Each maize family consisted of two homozygous inbred parents, their F1 hybrid and were grown under replicated field conditions. In any given family there were about 30,000 expressed genes, of which about 5,000 were differentially expressed (FDR < 0.005) among the family members. The vast majority of these could be modeled as additive to partially dominant, i.e., the number of transcripts in the parents were different from each other and the number of transcripts from the F1 were intermediate to the number of transcripts from each of the parents. Deviations from the calculated mid-parent transcript counts tended to be closer to the high parent. This additive to partial dominance of F1 transcript counts
suggested that transcript expression in the F1 was specific for one of the parental alleles at any given genetic locus.

Several studies have looked into expression of alleles in homozygous inbred parents without showing allele specific affects from hybridization (Wittkopp et. al., 2004; Stupar and Springer, 2006; Springer and Stupar, 2007a; Chang et. al., 2008; Gruber and Long, 2008; Wittkopp et. al., 2008a; Guo et. al., 2009; Sung et. al., 2009; Zhang and Borevitz, 2009). Studies of allele-specific expression in hybrids have tended to focus on a limited number of genetic loci in plants (Adams, 2007; Guo et. al., 2009; Zhang and Borevitz, 2009; Chodavarapu et. al., 2012; Zhai et. al., 2013) and mammals (Cowles et. al., 2002; Yan et. al., 2002). These studies have shown gene expression at heterozygous loci is not equal between parental alleles. These studies either pooled data from replicates or did not use any biological replication, thus the significance of biological variability associated with transcript expression was not reported. The importance of considering over-dispersion from biological replicates has been discussed at length in reviews by (Dillies et. al., 2012; Kvam et. al., 2012) and simulations have shown that more false positives are detected if mean-variance relationship is improperly modeled and if biological replications are not used. When gene expression is detected to be variable across two conditions, it may be caused by random variability or condition-dependent or both. It is essential to consider within-condition variability before making inferences on differential expression across conditions.

Maize provides an ideal system to study allele-specific expression because modern maize hybrids are obtained from crosses involving homozygous inbred lines derived from two mostly isolated breeding pools. These features enable studies with multiple combinations of inbred lines and biological replicates of inbreds and hybrids grown under field conditions (Mukherjee et. al., unpublished). Further, the diploid maize genome has a very high proportion of single nucleotide
polymorphisms (SNPs) and insertion/deletions (InDels), approximately an order of magnitude higher than that in humans (Bhattaramakki et. al., 2000; Sunyaev et. al., 2000; Buckler Iv and Thornsberry, 2002; Ching et. al., 2002). These polymorphisms often span several kilobases of DNA fragments (Fu and Dooner, 2002). By coupling this biological system with the advent of next generation sequencing technologies, such as RNA-seq, also known as "Whole Transcriptome Shotgun Sequencing" ("WTSS"), should enable quantification of expression levels of individual alleles throughout the genome. RNA-Seq data exhibits more repeatability, sensitivity, and specificity of genomic features (gene, exon, transcript etc.) than is available from microarrays even if the sample size is small (Wang et. al., 2011). (Hansen et. al., 2010) has suggested that technical variability and error in measurements can be reduced with advancement of sequencing technology but the significance of biological variability can’t be undervalued.

Not only is it possible to quantify allele specific expression of transcripts using WTSS and maize families consisting of inbreds and their hybrid progeny, it is possible to assess the regulation of the alleles. Transcript expression is believed to be regulated by cis-acting elements (on the same DNA molecule) or by trans-acting elements (that bind to the cis-elements) or both (Tautz, 2000; Wittkopp et. al., 2004; Williams et. al., 2007) suggested a testing strategy to distinguish between alleles controlled by cis and trans regulatory elements by using the relative abundance of allelic expression in the inbred parents and their hybrids. This assessment is based on the assumption that trans-acting regulators in a hybrid will affect both parental alleles equally, so relative abundance of parental alleles in the hybrid can provide an estimate of the cis regulatory variation. Trans regulatory variation can be estimated from a difference in the ratio of parental allelic variation in the parents and the hybrid. (Table 1)
<table>
<thead>
<tr>
<th>Rules of classification (FDR&lt;0.5%)</th>
<th>Cis - Trans Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cis-only</td>
</tr>
<tr>
<td>P=(P_u-P_v)</td>
<td>Significant</td>
</tr>
<tr>
<td>H=(F1P_u-F1P_v)</td>
<td>Significant</td>
</tr>
<tr>
<td>T=P-H</td>
<td>Not Significant</td>
</tr>
</tbody>
</table>

Table 1. Rules for categorization of allele-specific expression into cis-trans classifications. Columns are categories and rows are statistical hypotheses (McManus et. al., 2010).

_Cis_-acting regulatory variation was found to be the predominant form of parental allelic expression difference in a replicated greenhouse experiment (Guo et. al., 2009) consisting of pooled tissue samples of seedlings and immature ears from two commercial maize lines. Likewise, a study by (Stupar and Springer, 2006) that used pooled tissue samples from seedlings and immature ears of B73xMo17 family found that differentially expressed genes were shown to have primarily _cis_ acting. On the other hand, it has been reported that non-additive expression patterns for differentially expressed genes in the hybrid Arabidopsis (Vuylsteke et. al., 2005), Drosophila (Gibson, 1996;Ranz et. al., 2004), maize (Auger, et al, 2005) and rice (Huang and Khatib, 2010). Differential expression patterns in genetically similar parental alleles were reported to be regulated by trans-regulatory changes in Drosophila (Wittkopp et. al., 2004;Stern and Orgogozo, 2008;Wittkopp et. al., 2008a;Wittkopp et. al., 2008b), in Arabidopsis (Zhang and Borevitz, 2009) whereas differential expression patterns in genetically diverse parental alleles
were shown to be controlled by *cis*-regulation in *Drosophila* (Wittkopp, et al., 2008). (McManus et. al., 2010) reported that a vast majority of alleles in the hybrid of two closely related *Drosophila* species were regulated by *trans* regulatory divergence whereas *cis* regulation affected genes with additive expression pattern more than genes with non-additive expression patterns. In the two maize studies both *cis* (Guo et. al., 2004;Stupar and Springer, 2006) and *cis+trans* (Auger et al, 2005) allele specific expression was detected, but it is difficult to compare the studies because the experiments were conducted with different genetic backgrounds.

Herein, we report on investigations of allele specific expression and regulation in developing ear meristems of plants grown under replicated field conditions based on transcriptomic profiles of four inbred parents and their hybrids. We collected mRNA samples from three biological replicates for each genotype and sequenced using the samples using WTSS with one lane per sample. The quality filtered reads were mapped to the B73 version 2 reference genome (AGPv2), and the resultant transcript counts were screened through rigorous normalization and appropriate modeling before further analysis (Mukherjee et al). Since we were interested in determining if transcript counts and allele-specific counts varied across genotypes, it was important to remove possible WTSS technology-related biases through appropriate normalization (adjusting sequencing depth, scaling RNA-composition etc.). For differential expression analysis, it is also critical to consider the biological variability among replicates, as transcript variability can be an aggregation of biological variability, genotype-dependent variability and sequencing error. (Bullard et. al., 2010;Anders et. al., 2012;McCarthy et. al., 2012) have shown that the mean-variance relationship of RNA-seq counts can be best modeled as over-dispersed negative binomially distributed. We used TMM (Trimmed Mean of M values) normalization and negative binomial models implemented in edgeR (version 3.2.4) (McCarthy et. al., 2012) to
normalize and model data for subsequent analysis. Allele-specific counts were generated for parents and the hybrids based on SNP level counts and we used statistical strategies (Wittkopp et. al., 2004; McManus et. al., 2010) to further classify allele-specific expression into various cis-trans categories.

Materials and Methods

Genotypic Materials

Genotypic materials consisted of four homozygous inbred lines (LH1, LH123HT, PHG39, and PHG84) and their respective F1 hybrids. Thus, nested within each of the six crosses comprising the diallel of all four inbreds were three generations: P_u, P_v, F1.

The inbred lines were released recently from plant variety protection [http://sun.arsgrin.gov:8080/npgspub/xsql/pub/availpvp.xsql?in_crop=Corn,+field](http://sun.arsgrin.gov:8080/npgspub/xsql/pub/availpvp.xsql?in_crop=Corn,+field) and are progenitors of lines that are parents of commercial North American hybrids (Mikel and Dudley, 2006). The four inbreds were chosen because they are elite, unrelated, and have similar rates of growth and development in the Central Corn-Belt. In particular, growing degree units (GDUs) to anthesis are nearly equal among the lines. Seeds representing all generations of all six families were produced in Hawai‘i during the winter of 2008-2009.

The inbred lines LH1 and PHG39 belong to the stiff stalk family while LH123 and PHG84 are members of the non-stiff-stalk breeding pool. The hybrids derived from crosses between stiff stalk lines or between non-stiff stalk lines are considered to represent non-heterotic crosses. Any cross between a stiff stalk and a non-stiff stalk is considered to be heterotic. The concept of developing inbred lines within heterotic groups has emerged from 70 years of hybrid maize breeding.
Field Plot Design

Maize exhibits significant inbreeding depression resulting in differences among generations in plant size and vigor (Lamkey and Edwards, 1999). Consequently, in order to avoid biasing differences among entries belonging to different generations with inter-plot competition effects, a replicated split split plot field design (Kempthorne, 1952) was utilized. Whole plot entries were classified by generation: $P_u$, $P_v$, F1 where the F1 and parental inbred entries were placed into different whole plots. Each whole plot was divided into two sampling split plots: one to sample ear shoot meristem tissues and a second to sample mature ears. Within the split plots of the F1 whole plots were six split-split plot entries, each representing a different genetic hybrid. Within the split plots of the parental inbred whole plot were four split-split plot entries representing each of the parental inbreds as well as two ‘filler’ inbreds to assure that the amount of land area devoted to each whole plot was the same for inbreds and hybrids. Assignments of generations to whole plots, sampling types to split plots, and entries to split-split plot entries were done randomly. Each split-split plot consisted of two 5.2 m rows, spaced 0.76 m apart. At 60,000 plants/ha, each row consisted of 24 plants. Adjacent split-split plots also were spaced 0.76 m apart. Each whole plot was bordered by a two-row plot containing an entry from a generation consistent with the whole plot generation.

The experiment was planted on May 12, 2009, at the Iowa State University Agronomy Farm on well-drained Clarion silt loam soil. Nitrogen fertilizer at a rate of 92 kg/ha had been applied prior to planting. Pre-planting and two post-emergence herbicide treatments were applied to control weeds. Plots were overplanted and thinned to a stand equivalent to 60,000 plants/ha. The growing season at the experiment site in 2009 was relatively cool with ample rainfall (http://mesonet.agron.iastate.edu/climodat/index.phtml?station=IA0200&report=03).
Tissue sampling

Tissue sampling commenced on 14 July and continued on consecutive days through 21 July. To achieve relative temporal constancy of gene expression during sampling, all sampling was done between the hours of 1000 and 1200. A plant was judged suitable for sampling if the developing tassel was detectable in the whorl of upper leaves. The stalk of a selected plant was cleaved just above ground level, and the hewed plant was carried to the edge of the field for dissection. There, leaves were stripped from the stalk and a section containing the uppermost ear leaf node was sectioned. The sectioned piece was then split longitudinally by a cut transverse to the node, and stalk tissue around the ear shoot meristems was carefully removed. The ear shoot meristems were removed from the stalks and those between 0.8 cm and 1.2 cm in length were placed on the stage of a dissecting microscope where a sterilized scalpel was used to separate the apical 1mm section of the ear shoot meristem. The apical section (weighing about 2 mg) was placed in a microfuge tube on frozen CO₂, transferred to liquid nitrogen, and the date of sampling recorded. Sampling of a given split-split plot ceased as soon as ten shoot tips had been successfully processed. The samples were taken to the laboratory and stored in a freezer maintained at a temperature of –80°C. Sampling dates were converted to growing degree days for use as a possible covariate to control for environmental variation within split-split plots.

RNA Extraction

Before processing, all ten shoot tips of ear meristems from a given plot were bulked. The bulked samples were randomly assigned to four batches of approximately equal size. Total RNA was extracted using Qiagen’s RNeasy Plant Mini kit (#74904) according to manufacturer’s instructions, adhering to a method based on guanidine thiocyanate extraction with Buffer RLT, followed by purification using supplied spin columns. Tissue was homogenized in a micro-
centrifuge tube using a Kontes microtube pestle. RNA was eluted by addition of 50 µl of water, incubated for 10 minutes at room temperature, and then centrifuged. Each bulk sample yielded approximately 10 µg of RNA, at a concentration of 200 ng/µl. RNA was successfully extracted from 83 out of the 84 samples. Samples were randomized for sequencing-by-synthesis.

**Sequencing and Bioinformatics Analyses**

Polyadenylated RNA was isolated from total RNA using oligo-dT25 magnetic beads (Dynabeads; Invitrogen, Carlsbad, CA). The captured RNA was denatured and annealed to random hexamer primers and reverse transcribed. After end repair and A-tailing, adapters complementary to sequencing primers were ligated to cDNA fragments. The resultant cDNA libraries were size selected by agarose gel electrophoresis, then amplified by PCR for 15 cycles. Sample libraries were qualitatively and quantitatively assessed by Nanodrop ND-1000 (Thermo Scientific, Waltham, MA) UV/Vis spectroscopy and DNA BioAnalyzer 2100 microfluidics (Agilent, Santa Clara, CA). The samples from the cDNA libraries were loaded on an Illumina® single-end flow cell using the Illumina® Cluster Station (Illumina®, Inc., San Diego, CA). Thirty-six bp reads were obtained from an Illumina® Genome Analyzer II where sequenced bases were determined by the Illumina® base-calling pipeline consisting of analysis, base-calling, quality filtering and assignment of per base quality scores.

Genomic alignments of the resulting 36mer reads were performed using GSNAP (Wu and Nacu, 2010) against the maizesequence.org AgpV2 version of the B73 maize genome. The alignments, using GSNAP’s splicing detection mode, identified introns of up to 10Kbp in length, two base mismatches or a single indel. A read was characterized as unique if it aligned with a single location (locus) in the genome. A total of 38,178 unique alignments were available for subsequent data analyses. An alignment was assigned to a gene if the alignment overlapped the
genomic locus span with at least 95% identity and over 90% of the transcript length. Annotations of the maize sequence.org AgpV2 alignments are listed in Gene Ontology Annotation (http://www.ebi.ac.uk/GOA/). Genomic alignments were then converted to expression counts using the AgpV2 annotations.

**Normalization and modeling**

Data generated by sequencing technologies have their own technical and systemic biases. Normalization methods and distributional assumptions of microarray data are not directly applicable to Whole Transcriptome Shotgun Sequencing (WTSS) data, and a number of recent methods have been published to remove bias associated with RNA-seq (Baggerly et. al., 2004; Smyth, 2004; Robinson and Smyth, 2007; Robinson and Smyth, 2008; Anders and Huber, 2010; Bullard et. al., 2010; McCarthy et. al., 2012). Since the key objective of the manuscript is to assess transcript count differences among samples, we can ignore biases that affect all samples equally. To address sample specific biases for data analysis, we used TMM (Trimmed mean of M values) normalization to remove RNA-composition bias (Smyth, 2004; Robinson and Smyth, 2007; Robinson and Smyth, 2008; Robinson et. al., 2010; McCarthy et. al., 2012). Since all RNA samples were sequenced using one lane each, sequencing depth of all the samples were adjusted based on totals per lane.

Biological variation among replicates indicated that variance of transcript counts were over-dispersed relative to the mean of biological replicates (data available at http://gfspopgen.agron.iastate.edu/resources.html), thus an additional over-dispersion parameter was included for the analyses. (Baggerly et. al., 2004; Lu et. al., 2010; Zhou et. al., 2011) have suggested the use of Generalized linear models with the read counts treated as sampled from over-dispersed Binomial distribution. Generalized Linear Models require appropriate modeling
of the mean-variance relationship (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989), and they can be used to fit a log linear model for each transcribed allele to estimate the read counts. edgeR uses the quantile-adjusted conditional maximum likelihood (qCML) method to estimate dispersions for pairwise comparisons and Cox-Reid profile-adjusted likelihood (CR) method for multi-factor comparisons. Differential expression of genes between and among samples is assessed using appropriate likelihood ratio tests.

**SNP calling and variant detection**

Reads for each sample were aligned to the ZmB73 reference version AgpV2 using GSNAP with splicing allowed and a maximum mismatch level of 6% (Miller et al., 2008). Uniquely aligned reads were used to call variants, with read data from all samples representing replicates of the same cross pooled for the purposes of establishing genotype information; although these are also kept separate for expression quantification. SNP and indel annotations were based on uniquely aligned reads for SNP variants with an average PHRED quality score of \( \geq 20 \) and at least 20% of the sequence aligned with the reference position for the allelic variant. These criteria seem to be appropriate in the context of expression data where coverage can be extremely variable and preference for allelic expression is possible, particularly in the hybrid members of a maize family. Further, in this dataset, since we had the data for the parental lines as well as for the hybrid progeny, variant calls could be checked with respect to their pedigrees in cases of lower confidence calls.

**Assessment of allele specific counts**

All SNP sites were listed for each of the four inbred parents. A minimum coverage of 2 reads per SNP and a frequency cut-off of 100% was applied to make sure that the detected SNP was not a sequencing error. In terms of alignment, all considered reads from the sample aligned
to the particular base on the reference genome (SNP site) had the same variant nucleotide. SNPs were summarized in a gene-wise fashion according to the AgpV2 annotations of the B73 reference.

Figure 1. Frequency distribution of number of SNPs per gene

The filtered SNPs were divided into their respective families and pairwise comparisons were made between all pairs of parents. The resulting SNPs had unique variants that could be ascribed to each parent in a pair. Subsequently, haplotypes were formed for each pair of parents. Coverage was determined for each parent of each family at the haplotype sites. Since each homozygous parent has a single allele, the counts were summarized in a gene-wise fashion to
obtain allele-specific reads for the parents. Reads from the hybrids were classified as belonging to each parental haplotype by matching the nucleotide of the F1-read at the haplotype sites to the parental nucleotide at the same site. As was done with parents, counts were summarized in a gene-wise fashion to obtain allele specific counts for the hybrids. SNP level counts for each parent were obtained for comparisons with the F1 allele specific counts (Bell et. al., 2013). Allele specific counts for each family were normalized and modeled using total transcript counts (see normalization and modeling section).

Assessment of differential expression

Normalized and predicted transcript counts were used to assess global expression patterns. For allele specific expression patterns, coverage of SNP sites were considered only if the two parental alleles could be distinguished. TMM normalization implemented in edgeR was used to remove technical bias and a negative binomial distribution with an estimated dispersion parameter was used for both whole transcript and allele specific transcript counts. Pairwise comparisons between samples were done using a likelihood ratio test with multi-factor comparisons and assessment of differential expression. To correct for false positives arising from multiple testing, an FDR of 0.005 was imposed, following similar analysis by (McManus et. al., 2010; Bell et. al., 2013). All of the above methods are implemented in edgeR (v 3.2.4).

Modes of gene action

Mid parent values were estimated for whole transcript counts and allele specific SNP level counts as

\[ MP = (P_u + P_v)/2 \]

Differentially expressed genes were assessed for each of the 6 families. A gene was considered as differentially expressed if there was a significant difference (FDR < 0.005) between either or
both the parents and the hybrid. A gene was classified as additive if the expression in the hybrid was statistically no different from the estimated mid parent value. All other genes were classified as non-additive. A similar analysis was done with allele-specific counts.

**Cis and trans regulation from allele specific counts**

In the hybrid, *trans*-acting factors affect alleles from both parents equivalently. So the ratio of expression of the two parental alleles in a hybrid is a direct indicator to *cis*-regulated activity. The difference between the ratio of parental alleles in the parents to that in the hybrid gives an estimate of the *trans*-acting factors. Likelihood ratio test (McCarthy et al., 2012) was used to evaluate significant expression differences. Genes were further classified into *cis*-only, *trans*-only, *cis*+*trans*, *cis*\*trans, compensatory, conserved and ambiguous according to the classification methods used in (McManus et al., 2010; Bell et al., 2013) (see Table 1 in Introduction). The sign of the log fold change between the groups \( \log(P_u/P_v) \), \( \log(F1_{Pu}/F1_{Pv}) \), \( \log(P_u/P_v)-\log(F1_{Pu}/F1_{Pv}) \) were used as an indicator of the direction of differential expression, e.g., \( P_u > P_v \) is indicated by a positive fold change. Genes were classified as *cis*-only if there was evidence of differential allelic expression in both the parental and the hybrid datasets within a family but no significant *trans* effect (difference in allelic imbalance in parents and hybrid). *Trans*-only regulated genes had significant allelic differences in the parents, but insignificant difference in the hybrids. *Cis*+*trans* affected genes had all three significant (allelic imbalance in parents, differences in hybrid and difference between both) and also the difference in the allelic imbalance in the parents and hybrid groups had the same direction i.e the parent with higher expression than either parent also had higher allelic expression in the hybrid. *Cis*+*trans* acting regulators act complimentary to affect gene expression. *Cis* x *trans* represents a group meeting the same expression criteria as the *cis*+*trans* group but with opposite signs. *Cis* and *Trans* acting
factors favor the expression of the reciprocal allele. The compensatory group consisted of genes where the effect of cis and trans regulation cancelled each other and there is no difference in expression in the parents, so the differences in hybrid and both groups is significant but in the parents allelic expression occurs at similar levels. Genes classified as conserved regulation consists of alleles expressed at similar values among all three members of the family (insignificant differential expression). Genes following all other combinations are classified as ambiguous.

Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total genes</td>
<td>31779</td>
<td>31894</td>
<td>31818</td>
<td>31633</td>
<td>31440</td>
<td>31687</td>
</tr>
<tr>
<td>Number of DE genes (FDR&lt;0.005)</td>
<td>4454</td>
<td>2898</td>
<td>3650</td>
<td>4687</td>
<td>4352</td>
<td>4584</td>
</tr>
<tr>
<td>% DE genes</td>
<td>14.02%</td>
<td>9.09%</td>
<td>11.47%</td>
<td>14.82%</td>
<td>13.84%</td>
<td>14.47%</td>
</tr>
<tr>
<td>Mean % DE genes</td>
<td>12.95%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Number of differentially expressed genes with per family (FDR<0.005) based on whole transcript counts.

Global differential expression analysis

Reads from samples were mapped uniquely to 38,178 genes. Analysis revealed that, when averaged across families, about 13% of the genes were expressed differentially and the alleles could be distinguished by parental origin (FDR < 0.005).

The non-heterotic stiff stalk family LH1-PHG39 had the least proportion of differentially expressed genes (9%) among all the families and LH123HT-PHG39 had the most (14.8%).
families with LH123HT as one of the inbred parents had the largest proportion differentially expressed genes (14.2%) and the families with the stiff stalk inbred parent LH1 had the least number of differentially expressed genes (11.5%). Transgressive gene expression (over and under dominant) were identified for each family. These genes in the hybrid had expression values significantly more than the high parent (over-dominance) or significantly lower than the low parent (under-dominance). Only 6% of differentially expressed genes exhibited transgressive gene expression. Of the transgressive genes, 95% exhibited under-dominance.

Modes of gene action

Differentially expressed genes for each of the families were evaluated for hybrid expression relative to the mid-parent value. Differentially expressed genes which had expression values not significantly different from the mid-parent were classified as additive, all other genes were categorized as non-additive. Significant comparisons for all families revealed on an average 79% of differentially expressed genes were estimated as additive (Mukherjee et. al., unpublished). Of the non-additive genes, 85% exhibited hybrid expression significantly greater than the mid-parent value indicating at least partial-dominance for the high-parent allele. Only 15% of the non-additive genes showed expression significantly below the mid-parent. The trend of most differentially expressed genes showing additive mode of action and most non-additive genes having expression patterns higher than the mid-parent value was prevalent for all families, irrespective of their heterotic classification.
Allele specific expression

<table>
<thead>
<tr>
<th>Families</th>
<th>#Significant P=P \textsubscript{u}-P \textsubscript{v}</th>
<th>#Significant F=F1P \textsubscript{u}-F1P \textsubscript{v}</th>
<th>#Common genes between P&amp;F</th>
<th>Percentage of F that were common with P</th>
<th>#total genes Used for DE analysis</th>
<th>Percentage significant P relative to all genes</th>
<th>Percentage significant F relative to all genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHG39-PHG84</td>
<td>794</td>
<td>320</td>
<td>229</td>
<td>71.56%</td>
<td>12614</td>
<td>6.29%</td>
<td>2.54%</td>
</tr>
<tr>
<td>LH123HT-PHG84</td>
<td>1352</td>
<td>641</td>
<td>457</td>
<td>71.29%</td>
<td>13452</td>
<td>10.05%</td>
<td>4.77%</td>
</tr>
<tr>
<td>LH123HT-PHG39</td>
<td>1390</td>
<td>793</td>
<td>565</td>
<td>71.25%</td>
<td>13127</td>
<td>10.59%</td>
<td>6.04%</td>
</tr>
<tr>
<td>LH1-PHG39</td>
<td>569</td>
<td>214</td>
<td>145</td>
<td>67.76%</td>
<td>10107</td>
<td>5.63%</td>
<td>2.12%</td>
</tr>
<tr>
<td>LH1-PHG84</td>
<td>751</td>
<td>230</td>
<td>170</td>
<td>73.91%</td>
<td>12417</td>
<td>6.05%</td>
<td>1.85%</td>
</tr>
<tr>
<td>LH1-LH123HT</td>
<td>1322</td>
<td>720</td>
<td>494</td>
<td>68.61%</td>
<td>13111</td>
<td>10.08%</td>
<td>5.49%</td>
</tr>
<tr>
<td><strong>Column Means</strong></td>
<td><strong>1029.67</strong></td>
<td><strong>486.33</strong></td>
<td><strong>343.33</strong></td>
<td><strong>70.73%</strong></td>
<td><strong>12471.3</strong></td>
<td><strong>8.12%</strong></td>
<td><strong>3.8%</strong></td>
</tr>
</tbody>
</table>

Table 3. Number and percentage of intra-specific (P\textsubscript{u}-P\textsubscript{v}) and allele-specific (F\textsubscript{1}P\textsubscript{u}-F\textsubscript{1}P\textsubscript{v}) differentially expressed genes per family (FDR<0.005), based on SNP counts. Note that all numbers reported in Table 3 are based on number of differentially expressed genes.

Allele specific expression in the hybrid was analyzed using SNP coverage across the genome for the parents and the hybrid. 18866 (49%) genes spanning 103270 SNPs were used for this analysis. The gene set was considerably reduced relative to the total number of differentially expressed genes, because not all alleles contained SNPs enabled us to distinguish between the parental alleles. SNP level counts were summarized into transcript level allele specific counts by summing coverages relative to the reference genome, B73 AgpV2 annotations. The number of SNPs evaluated per gene varied from 1 to 61, with an average of 5.4 SNPs per gene. Approximately 12500 genes could be evaluated for allele specific expression in each hybrid.
that can be distinguished based on SNPs as opposed to whole transcript coverage in Table 1. Since SNPs were used to differentiate between the parental alleles, SNP level counts were used to determine allele specificity in both the hybrid and the parents to make fair comparisons. In the parents (P in Table 3), 8.1% (~1030 out of 12471) of the genes had significant differences between the two parents, in the hybrids (F) the significant allelic imbalance was assessed for approximately 4% (~486 out of 12471) of the genes. Interestingly there was an overlap of 70% between genes showing allele specific expression in the parents and the corresponding hybrid. So a majority of genes that showed differential expression between the parental alleles in the hybrid also showed differential expression between its parents. The stiff stalk cross LH1xPHG39 showed the least number of genes (5.6%) with SNP distinguishable differential expression between the parents. When only the alleles in the hybrids were compared, the alleles from PHG84 was most often (65%) the high parent, the allele from LH1 was the low parent in most cases (63%). When allelic imbalance in the parents was assessed, we got the same results with PHG84 allele being the high allele most often (58%) and LH1 being the low allele on most cases (52%).

Mono-allelic expression in hybrids

In this set of genes, the allele count for any one parent is zero and the allele count for the other parent is significantly more than zero (exact test results from edgeR).
Table 4. Number of genes for each hybrid where each of the allele counts is zero, significantly less than the other parental allele count.

It was interesting that alleles from LH123HT had a zero count most often (111) on average per cross, while alleles from the parent PHG84 had the least number of zero counts (38.33) in all of its hybrid progeny. In all cases of significant mono-allelic expression in the hybrid, 42% involved the LH123 allele as the suppressed allele while in only 14% was the allele from PHG84 were undetectable.

**Assessment of cis and trans regulatory divergence from allele specific expression**

*Cis* and *trans* regulation was inferred by comparing combinations of allelic imbalance in the parents and the hybrid. Recall that only transcripts in the hybrids with confirmed SNPs between the parents were used for this analysis because the allele specific data in the hybrids can only be summarized using SNPs (Bell et. al., 2013). *Cis*-regulatory variation was reported by comparing allelic imbalance in the hybrids since *trans* regualtory variation in the hybrid affects both alleles equally. *Trans*-regulatory variation was assessed as a difference of allelic imbalance in the parents and hybrids.
Table 5. Percentage of genes classified according to cis and trans regulation. Percentages are based on expression of distinguishable SNP alleles.

Cis-regulation accounted for 3.8% of the total genes on an average for each family while trans-regulation was detected in 0.15% of the genes per family. The trans genes were further classified as 'cis + trans', 'cis * trans' and 'compensatory' according to rules suggested by previous studies of regulatory evolution (Fontanillas et. al., 2010), (McManus et. al., 2010) with compensatory genes assuming a meagre 0.12%.

A vast majority of genes (91%) for all families were classified as ‘conserved’, i.e., for the set of expressed genes for which we can distinguish the parental alleles, there is no significant evidence for differential expression among parents and hybrid progeny. (see introduction Table 1 for rules of classification) Cis-only regulation accounted for 2.6% of genes across all families. All other classifications including ‘trans-only’ had a negligible percentage of genes for all the families.
Associating regulatory divergence and global modes of gene action

Cis regulatory variation has been reported in many studies to be associated with additivity in the F1 hybrid maize (Guo et. al., 2004; Stupar et. al., 2007; Stupar et. al., 2008). Other species have also shown that cis regulatory elements played a role in additive hybrid expression (Cheung and Spielman, 2002; Pastinen and Hudson, 2004; Wittkopp et. al., 2004). Several other studies have reported the prevalence of trans-regulated expression in the hybrid (Gibson, 1996; Ranz et. al., 2004; Auger et. al., 2005; Vuylsteke et. al., 2005; Huang et. al., 2006).

Based on sequence counts, we extracted all the genes that were declared to have additive expression pattern in the hybrid \( (F1-(P_u+P_v)/2=0) \) and compared the relative contributions of cis and trans effects. The same was done for the non-additive genes. Violin plots were used to visualize the relative distribution of cis and trans contributions for additive as well as non-additive genes.

![Figure 2. Violin plots showing the relative contribution of cis and trans regulation on differentially expressed genes showing additive mode of gene action](image)
Figure 3. Violin plots showing the relative contribution of cis and trans regulation on differentially expressed genes showing non-additive mode of gene action.

Note that the cis median was higher than the trans median for additive genes and for non-additive genes as well (Figures 2 and 3) indicating that cis-regulation had a greater impact than trans regulation for both models of gene action.

**Discussions**

RNA-seq technologies not only provide a more accurate measure of mRNA expression than microarrays but also offer useful information about sequences of alleles. Because we evaluated homozygous inbred and heterozygous hybrids representing six families, it is possible to trace the alleles from inbred parents in the hybrids. Distinguishing parental alleles is possible by aligning parental reads to the reference genome and identifying divergent mutations and single base differences (SNPs). Quantifying allelic expression was done by summarizing coverage of the hybrid at the SNP sites and matching the read base back to the parental
haplotype. This approach can be implemented with a system like maize where homozygous
inbreds and their direct heterozygous hybrid progeny can be easily developed for a diploid
system where it is easy to trace parental alleles.

We used a negative binomial model and appropriate statistical tests to measure
expression differences between the inbred parents and their hybrid progeny. 13% of the genes
were detected to be differentially expressed for each family. These are similar proportions to a
previous study by (Stupar et. al., 2008) that was based on use of microarrays to study gene
expression in multiple inbred-hybrid combinations in seedling meristems. Use of biological
replicates make our estimates more accurate and our models more robust than some of the
previous studies which report their results based on single libraries per genotype (Wittkopp et.
al., 2004;McManus et. al., 2010;Bell et. al., 2013). Also we used appropriate normalization
methods to remove technical bias from our data and used a conservative FDR of 0.005 for strict
error control, as in (McManus et. al., 2010;Bell et. al., 2013).

Most differential expression (80% of DE genes) was estimated to be additive i.e. the
hybrid expression values were not significantly different than the mid-parent value, \((P_u+P_v)/2\).
These results are similar to previous studies that reported prevalence of additive expression in
most expressed genes of hybrids formed from inbred parents (Swanson-Wagner et. al., 2006),
(Guo et. al., 2006), (Stupar and Springer, 2006). We note that some studies have reported a
predominance of non-additive expression in hybrids in Arabidopsis, Drosophila, maize and rice
(Rieseberg et. al., 2003;Auger et. al., 2005;Landry et. al., 2005;Meyer and Scholten,
2007;Użarowska et. al., 2007;Meyer et. al., 2008;Fontanillas et. al., 2010). Some studies in
sunflower and various other species also report transgressive (outside parental range) expression
of genes in the hybrid (Rieseberg et. al., 1999;Arnold, 2004;Rieseberg et. al., 2007). We found
only 6% of differentially expressed genes exhibited significant transgressive expression. Of the transgressively expressed genes, a vast majority (95%) showed expression significantly below the low parent (under-dominance) while only 5% exhibited expression levels significantly above the high parent (over-dominance). Our results of transgressive expression are similar to previous studies in maize, Senecio and rice (Hegarty et. al., 2006; Swanson-Wagner et. al., 2006; He et. al., 2010) where only a small percentage of differentially expressed genes were transgressively expressed. Studies in Drosophila (Wittkopp et. al., 2004) have reported higher (30%) levels of transgressive expression in the hybrid. These transgressive genes are believed to be responsible for the hybrids' ability to adapt to rapid drastic changes in environment (Rieseberg et. al., 2003).

Significant allele specific expression was detected in 8.26% genes between the parents and in 4% genes (between parental alleles) in the hybrid. Interestingly a majority (70%) of genes that show differential expression of alleles in the hybrid also show differential expression between the parents. This suggests that the presence of a particular set of genes dependent on genetic background that are involved in allele specific expression. Mono-allelic expression was observed in all of the families, such evidence of allele-specific silencing has been reported previously by (Swanson-Wagner et. al., 2006) and (Adams, 2007).

Trans-acting regulators should affect both alleles in the hybrid equally, thus allelic imbalance in the hybrid is a signature of cis-regulatory variation. Significant allelic imbalance (FDR < 0.005) was found in about 3% of the genes in the hybrid indicating cis regulation of allelic expression. The parental alleles are affected by both cis and trans regulatory factors, so significant trans regulation can be determined by comparing the allelic imbalance in the parents and that in the hybrids. Significant trans regulation was determined in a mere 0.15% of the differentially expressed genes. A vast majority of genes showed conserved expression (no
differential allelic imbalance between parents or between parents and hybrids). Failure to detect differential expression might be attributed to low coverage at SNP level. Compared to about ~10% differentially expressed genes on an average between the parents based on whole transcript counts (Mukherjee et. al. unpublished), 8.1% genes were detected as differentially expressed between the parents based on SNP level counts (Table 3).

Many of these genes might be differentially expressed between the inbreds and the hybrids if we look at counts for the whole transcript, but low SNP level coverage prevents us from detecting these differences. All statistical models and tests published to date for RNA-seq data analysis are driven by high counts per sample and their power to detect reduces considerably when coverage is very low.

Where detectable, Cis-regulation had a greater influence than trans regulation across all families. This is consistent with previous studies in maize (Springer and Stupar, 2007b), yeast (Tirosh, 2009) and poplar (Adams, 2007). Other studies involving yeast (Sung et. al., 2009; Emerson et. al., 2010), Drosophila (Wittkopp et. al., 2004; Wittkopp et. al., 2008a; Graze et. al., 2009; Fontanillas et. al., 2010), Arabidopsis (Zhang and Borevitz, 2009), Cirsium arvense (Bell et. al., 2013) have detected trans-regulation as the primary cause of allele specific expression. Of the genes that showed evidence of both cis- and trans regulation, 94% of the genes (in Table 5) were classified as ‘cis*trans’ and ‘compensatory’, types of regulation in which cis and trans regulators antagonistically to favor expression of the opposite allele. In the genes classified as ‘compensatory’, the effect of cis and trans regulation cancel each other resulting in no parental differential expression. Evidence of such expression have been reported for Drosophila (Landry et. al., 2005) and yeast. Birchlet et al (2011) and (Birchler et. al., 2010) have interpreted such expression as evolutionary mechanisms to sustain expression levels. In the
remaining 6% (Table 5) of the genes classified as ‘cis+trans’, cis and trans regulation act complimentarily towards expression or suppression of alleles.

Finally we looked at the relative contributions of cis-and trans-regulation on additive and non-additive genes detected from global whole transcript expression differences. We classified all genes for which the expression levels in the hybrid were not significantly different from the mid-parent expression value and categorized them as additive. A majority of genes (80%) were modeled as additive with the remaining 20% as dominant or transgressive. The relative contributions of cis- and trans- regulation were determined from the difference of hybrid allelic expression and the allelic imbalance difference between the parents and the hybrid. A violin plot showed the median cis percentage was more than the median trans percentage for both additive and non-additive genes, ie., cis regulation was associated with both additive and non-additive genes more than trans-regulation. These results were consistent with the hypothesis that cis-regulation primarily causes additive gene expression in the hybrids. (Lemos et. al., 2008; Stern and Orgogozo, 2008).
References


O. Kempthorne (1952). The design and analysis of experiments.


maize based on morphological investigations of heterosis for plant height. *Plant molecular biology*. **63**: 21-34.


CHAPTER 4

DETECTION AND QUANTIFICATION OF SPLICE VARIANTS IN MAIZE INBREDS AND HYBRIDS FROM RNA-SEQ DATA

A paper to be submitted to PLOS ONE

Mukherjee S*, Johnson GR, Scott PS, Farmer AD, May GD, Lamkey KR and Beavis WD

Abstract

Heterosis or hybrid vigor refers to improved performance of the homogeneous heterozygous F1 progeny over its homozygous inbred parents. It has been a focus of intense research over 100 years, but the molecular basis of heterosis still remains poorly understood. Differential gene expression studies have not been able to conclusively prove why the hybrids grow bigger and faster than its parents. Reduced protein metabolism suggests that the hybrids avoid spending energy degrading misfolded proteins by producing the most stable protein isoforms. These protein structures are generated from spliced mRNA isoforms and investigating differential splicing between inbreds and hybrids may provide an alternative molecular mechanisms to explain heterosis. In this manuscript, we report methods to detect and quantify splicing in four unrelated homozygous maize inbred lines and their six F1 hybrids based on RNA-seq data. Consistent with previous studies in maize and other plants, we identified 3` alternative acceptor site as the most common form of splicing among inbreds and hybrids, while the mutually exclusive exon was the most rare. Unique transcripts were identified by a couple of methods. Similar numbers of unique transcripts were detected in both inbreds and hybrids, but 10% of the genes in the hybrid showed evidence of splice forms not seen in either of its parents.
Isoform expression was quantified for multi-exon genes and we found that a vast majority of the genes preferentially expressed a single isoform in both inbreds and their F1 progeny.

**Introduction**

Hybrids from genetically distinct inbred parents often show superiority in performance relative to the inbred parents. The improved performance of hybrids due to heterosis has been exploited by plant breeders over the years and paved the way for development high-performing hybrid varieties in numerous crop species, including maize (Duvick, 2001). The molecular basis of heterosis is still not well understood. Dominance, over-dominance and epistasis theories have been hypothesized but none have been validated for multi-genic heterotic traits like flowering time, grain yield, vegetative growth, resistance to biotic and abiotic stress etc. Several gene expression studies between homozygous inbreds and their F1 hybrid progeny have tried to decipher the molecular basis of heterosis based on differential gene expression but there have been no consistent results among these studies. While whole plant phenotypic traits exhibit over-dominance, global expression of gene transcripts exhibit additive to partial dominance, i.e., the F1 hybrids produce quantities of transcripts that are intermediate to the levels of their respective inbred parents (Mukherjee et. al. unpublished)(Swanson-Wagner et. al., 2006). Further, sets of differentially expressed genes in developing ear meristems of maize are not consistent among maize families consisting of parents and F1 hybrids (Mukherjee et. al. unpublished). Moreover, differentially expressed genes from these studies have not revealed over-representation of particular biochemical pathways to be responsible for heterosis (Mukherjee et. al. unpublished). (Goff, 2011) has suggested that non-additive distinctions between inbreds and hybrids most likely occur at the level of protein metabolism because there is evidence fewer protein metabolism genes being differentially expressed in the hybrids than inbreds. (Goff, 2011)
proposed a model to describe the mechanisms for increased growth and reduced protein metabolism in the hybrids with respect to inbreds. According to the Goff hypothesis, hybrids have an abundance of allelic transcripts which they can choose to express as proteins. Homozygous inbreds on the other hand have fewer choices and sometimes generate unstable isoforms that need to be degraded by protein metabolism. As protein metabolism is energy intensive, it would further impede growth of the inbreds. In addition to the availability of two distinct alleles from each parent in the hybrid, alternative splicing among the exons from each allele could generate an abundant pool of translatable mRNA. Herein, we report an evaluation of relative abundance of different transcripts in six maize families consisting of homozygous inbred parents and their heterozygous hybrid progeny. Although alternative splicing is considered a critical component of isoform expression, there are very few studies that have looked at splice variants for multiple genes in a families consisting of inbreds and hybrids (Scascitelli et. al., 2010), and none have investigated global splicing patterns across inbreds and hybrids using the RNA-seq technology.

Alternative splicing (AS) is a molecular mechanism that results in a variety of protein isoforms from a limited number of alleles. Shortly after the concept of introns and exons were first introduced (Berget et. al., 1977;Chow and Broker, 1978) in adenovirus, it was postulated that exons within a gene can combine selectively and form different mRNA products from the same gene (Gilbert, 1978). This provides more plasticity to the transcriptome and proteome, thereby providing the organism flexibility to react to different environmental signals (Graveley, 2001;Black, 2003;Lareau et. al., 2004;Stamm et. al., 2005a). In Drosophila, it is reported that the SCAM gene can generate 38016 alternatively spliced isoforms, more than the reported number of genes (Celotto and Graveley, 2002). Thus, alternative splicing serves as a critical piece in the
transcriptional, post-transcriptional and post-translational biological networks in determining the phenotype (Reddy, 2007). Promoter activity regulates transcript expression but alternative splicing can affect the structure of the transcripts thus playing a key role in determining localization, binding properties of the coded proteins, enzymatic potential and stability of the protein structures. Transcript levels can be regulated by different spliced protein isoforms which are subsequently degraded in the nonsense-mediated decay (NMD) pathway (Chang et. al., 2007). Alternative protein structures not only play regulatory roles in gene expression, but also truncated translations may give rise to unstable structural configurations that are likely unfavorable for the cell (Stamm et. al., 2005b).

Initially, alternative splicing was identified in a small frequency of genes in humans (0.05, (Sharp, 1994)) and plants (0.01 in Arabidopsis (Zhu et. al., 2003)). However, with the advancement of technology and cognizance of the importance of AS events, recent studies have revealed that AS has been identified for most genes in both animals and plants. The number of publications on AS increased from 16 in 1985 to 1073 in 1998 and over a 1000 publications per year thereafter (Stamm et. al., 2005b). In humans more than 95% of multi-exonic genes are alternatively spliced (Pan et. al., 2008). Recent resequencing of plant genomes has shown that about 80% of coding transcripts have embedded non-coding introns (Alexandrov et. al., 2006). Studies in Arabidopsis (Arabidopsis thaliana) and rice (Oryza sativa) respectively have shown 42% and 33% of genes with imbedded introns exhibit alternative splicing (Wang and Dooner, 2006;Wang et. al., 2008a), (Filichkin et. al., 2007;Barbazuk et. al., 2008). Most recent RNA-seq data from Arabidopsis shows that the figure may be as much as 61% for plants under normal conditions and the frequency might increase across various development stages and tissues for plants that need to adapt to extreme environments (Marquez et. al., 2012). With the availability
of RNA-Seq data from different stages of development, diverse tissues and varied environments, it is likely that we will identify novel splice variants and an increased number of AS events (Yoshimura et. al., 2002; Reddy, 2007). Such information will provide us with a better understanding of the molecular mechanisms underlying alternative splicing.

Five basic models of alternative splicing have been reported (Black, 2003; Matlin et. al., 2005; Pan et. al., 2008).

Figure 1. Five major types of splicing events adapted from (Cartegni et. al., 2002)

1. Exon skipping or cassette exon: in this case, an exon may be spliced out or retained in the primary transcript. This is the most common mode in mammals. (Gao et. al., 2008)

2. Mutually exclusive exons: One of two exons is retained in mRNAs after splicing, but not both.
3. Alternative donor site: An alternative 5’ splice junction (donor site) is used, changing the 3’ boundary of the upstream exon.

4. Alternative acceptor site: An alternative 3’ splice junction (acceptor site) is used, changing the 5’ boundary of the downstream exon.

5. Intron retention: A sequence may be spliced out as an intron or simply retained. This is distinguished from exon skipping because the retained sequence is not flanked by introns. If the retained intron is in the coding region, the intron must encode amino acids in frame with the neighboring exons, or a stop codon or a shift in the reading frame will cause the protein to be non-functional. This is the rarest mode of splicing in mammals (Gao et. al., 2008).

Relative frequencies of types of alternative splicing events are not consistent between plants and animals. In animals, exon skipping is the prevalent (58% in humans) form of splicing whereas the frequency in Arabidopsis was reported to be 8%. In Arabidopsis the majority of splice variants (56%) are represented by intron retention, whereas in humans this occurs about in about 5% of alternative splice events (Iida et. al., 2004; Ner-Gaon et. al., 2004; Baek and Green, 2005; Wang and Brendel, 2006a; Filichkin et. al., 2007; Labadorf et. al., 2010). Alternative 3’ and 5’ splice sites occur at a similar frequency (25%) in plants and animals (Barbazuk et. al., 2008).

In the last decade next generation sequencing technologies and deep sequencing of transcriptomes have enabled discoveries in many plant species. A wealth of information is available from sequencing reads that align to various regions of the genome, but sophisticated analytic methods need to be employed for correct interpretations of the data. Alternative splicing can be detected primarily by studying reads with split alignments, where a portion of the read aligns to a particular region of the genome, and the other portion aligns to another region of the
reference genome. The split alignment in such reads provides proof of presence of an intron between the two aligned regions of the read which was possibly spliced out. But there may be several challenges to accurately detect alternative splicing events from RNA-seq data. For example, short reads often map to multiple regions of the genome, especially in plant genomes that have evolved from ancient polyploidy events. These make it hard to uniquely identify the genomic region responsible for the transcript. In a split alignment of a short read, parts of the same read are mapped to different regions of the genome, making the split parts of the read more susceptible to multiple mapping sites (Rogers et. al., 2012). Several algorithms have been developed to predict splice sites. (De Bona et. al., 2008). Candidate regions for mapping these genes can be eliminated by using cutoff intron lengths and considering only sites that are limited by canonical GT-AG or GC-AG splice-site dimers (Rogers et. al., 2012).

Several methods have been published for detection and quantification of alternative transcript regulation from RNA-seq data. Contrary to microarray based approaches, sequencing based methods can detect novel transcribed regions and previously unknown splice forms (Wang et. al. 2009). Because transcripts of the same gene use shared DNA, it is not possible to deterministically assign reads that map to shared regions in the genome. Several probabilistic methods have been developed to quantify isoform expression, Cufflinks (Trapnell et. al., 2009;Trapnell et. al., 2010a;Trapnell et. al., 2012) being the most popular Some of these methods are capable of doing de-novo transcript assembly with and without information from a reference genome, e.g., Velvet (Zerbino and Birney, 2008), Scripture (Guttman et. al., 2010), Tophat (Trapnell et. al., 2010b;Trapnell et. al., 2012), SplicingCompass (Aschoff et. al., 2013). (Martin and Wang, 2011) and (Garber et. al., 2011) have reviewed limitations of the existing methods and have suggested that computational methods used to study transcript counts and
differential splicing from RNA-seq data make naive assumptions. Cufflinks and other methods have been successful so far but ambiguity still remains in the context of assigning reads to particular transcripts (Birney et. al., 2007). De-Bruijn graphs have been used to predict splice events even without completed assemblies (Sacomoto et. al., 2012), and Splice-Grapher (Rogers et. al., 2012) have detected a very high frequency of unresolved splicing events. Also detection and quantification of isoform expression is variable depending on the length of the reads and sequencing depth (Aschoff et. al., 2013). Additional methods such as DEXSeq (Anders and Huber, 2010a; Anders et. al., 2012) and edgeR (Robinson et. al., 2010; McCarthy et. al., 2012) SplicingCompass (Aschoff et. al., 2013) use a simplistic exon-bin method to quantify and compare differential exon usage across samples. However, inferences produced by binning methods clearly are not exclusively indicative of alternative splicing.

The purpose of this manuscript is to report detection and quantification of isoform expression in unrelated homozygous inbred maize lines and their heterozygous hybrids progeny. Initially we discuss methods that can be used to identify and categorize splicing events and visualize them (with SpliceGrapher). Then we apply these methods to explain heterotic behavior in hybrids by investigating the potential to differentiate splicing patterns in hybrids from their inbred parents. We look at results from the two popular software packages DEXSeq and Cufflinks, and discuss the advantages and disadvantages of using either of these approaches. Finally we use normalized Cufflinks results to look at relative isoform abundances across multi-exon genes in inbreds and hybrids.
Methods

Normalization and modeling biological variability

General conclusions about isoform regulation require biological replications and thereafter appropriately modelled biological variability. Prior studies of alternate splicing in animals (Wang et. al., 2008b; Griffith et. al., 2010; Kratz et. al., 2010; Trapnell et. al., 2010a) combined replicates or used single samples to look at alternate isoforms in the transcriptome. These approaches do not address the importance of biological replication. (Blekhman et. al., 2010) suggested the use of biological replicates for large number of samples and at least three publications (Robinson and Smyth, 2007a; Anders and Huber, 2010b; Bullard et. al., 2010; Robinson et. al., 2010; McCarthy et. al., 2012) have shown that sampling biological variability is needed to assure appropriate inferences by minimizing false positives. Most of these manuscripts have shown that replicated RNA-seq read counts mapping to particular genomic features have greater variability than simple averages of transcript counts, thus a dispersion parameter is required to define the extra deviation. Log linear models with negative binomial distributional assumptions have been suggested to model these counts (Robinson et. al., 2010; McCarthy et. al., 2012), (Mukherjee et. al, unpublished).

We use TMM (Trimmed Mean of M values) normalization method implemented in edgeR (Robinson and Smyth, 2007a; Robinson and Smyth, 2008a; Robinson et. al., 2010; McCarthy et. al., 2012) and negative binomial model implemented in DEXSeq (Anders et. al., 2012) with Cox-Reid dispersion estimates. In DEXSeq, the alternative model to be tested is

\[ \text{sample} + \text{exon} + \text{condition} \times I(\text{exon} == \text{exonID}) \]
Compared to the null model,

\[ sample + exon + condition \] .................................................................(2)

Likelihood ratio test comparing the null model to the alternative model is used to detect significant exon-condition interactions with 1 degree of freedom per test. Estimating dispersions for 38,178 genes with 38,7489 exonic bins is computationally expensive, so we parallelized the process using 6 CPU cores (code available through GFS group website http://gfspopgen.agron.iastate.edu/index.html).

**Gene Model Flattening**

(Bullard et. al., 2010) have demonstrated the use of exon-condition interactions to infer differential isoform regulation from microarrays. These methods were further extended by (Blekhman et. al., 2010) for RNA-seq data. More sophisticated methods have been published recently (Robinson and Smyth, 2007b; Robinson and Smyth, 2008b; Anders and Huber, 2010c; Oshlack et. al., 2010; Robinson et. al., 2010; Robinson et. al., 2011). DEXSeq (Anders et. al., 2012) has implemented the use of flattened gene models to assess exonic counts.
The various transcripts of a gene can use different exon combinations and thus have different boundaries and lengths. Some of these transcripts will have common and overlapping regions. Transcript specific mapping becomes complex when transcripts share common genomic regions. DEXSeq takes into account the different genomic features that span different transcripts and creates exon count bins. In the event that splice junction of an exon of a transcript is extended towards the 3’ end, and in another transcript, the same exon starts from it designated start position, the exon is divided into 2 bins, shared and non-shared. Reads are mapped to each bin and treated as separate exons. Reads that span multiple bins are assigned to all.

**Isoform specific expression with Cufflinks**

Because multiple transcripts of the same gene can share DNA, we cannot deterministically measure the expression of individual transcripts with RNA-seq data. Cufflinks (Trapnell et. al., 2010a; Trapnell et. al., 2012) assign reads to transcripts in a probabilistic fashion by considering the shortest path between exons in a read-overlap graph. Moreover, transcripts of the same gene can be of different lengths, thus the number of reads mapping to transcripts must
be normalized to detect their relative abundance. We used the Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) normalization method to remove length bias (Mortazavi et. al., 2008).

**Methods for classification of AS events**

In RNA-seq mapping, we try to map the transcripts to the genome and as a result we can find evidence of different types of splicing events. We use in-house suite of scripts to identify and categorize the different types of alternative splicing events that match with the reference annotation. We can also identify novel splicing events and transcribed regions not described by the B73 annotation, but we have not considered the novel events for downstream analysis. Visualization of splicing events and isoforms were done using the software SpliceGrapher (Rogers (Rogers et. al., 2012).

**Intron retention**

Detection of intron retention across a gene is probably the most difficult among all types of alternative splicing events to identify. There can be two scenarios for intron retention that can be detected. The first and the simplest case is when reads span the entire length of an intron and boundaries are determined by flanking exon limits as described by the gene model. In the second case, the gene model describes a large exon, while there is evidence of split alignment reads within the large exon. If we assume that the intronic region was retained in the larger exon and was transcribed in its constitutive form, the split reads suggest presence of novel splice junctions.
Figure 3. Model for a constitutive intron retention event, here the transcript described by the gene model uses an exon B. In an alternatively spliced isoform as seen from evidence, we see two smaller exons D and E and an intronic stretch between the two. Note that the acceptor site of D must match with the acceptor of B, and the donor site of E must be the same as the donor site as B.

Novel intron retention events can be unambiguously detected if the boundaries of the two flanking exons and the larger merged exon (which retains the intron) can be determined by combining the gene models and the alignment reads. In cases where we are not able assign definitive boundaries to novel exons that show intron retention, it is best categorized as ambiguous.
Alternative 3’ and 5’ splice sites

Detection of alternative 3’ and 5’ splice sites require evidence reads to show splice junctions extending the 3’ end of the acceptor exon or the 5’ end of the donor exon. As a rule of thumb, boundaries of the larger exon need to be definitive to be categorized as a 3’/5’ alternative splicing event. Evidence of novel exons and their extension into 3’ and 5’ ends also can be detected.

**Figure 4. Model for alternative 3’ and 5’ alternate sites.** Here the transcript described by the gene model uses an intron between exons A and C. In alternatively spliced forms as seen from evidence, we see two transcripts, both novels. One of the transcripts use exons D and E and the other transcript uses exons F and G. Note that the acceptor 3’ site of F and the 5’ donor site of G are extended into the flanking introns with respect to the 3’ site of D and 5’ site of E.

It should be noted that these methods do not attempt to predict donor or acceptor sites by looking for known donor and acceptor motifs in the corresponding introns. Boundaries are resolved by the evidence of split reads. These methods can be further sophisticated by incorporation of prediction algorithms. (Cui et. al., 2013; Giulietti et. al., 2013; Li et. al., 2013; Wang and Landsverk, 2013; Zhang et. al., 2013).
Exon skipping

Exon skipping is detected when an exon is a part of one transcript while it is spliced out of its isoform. These events are fairly easy to detect with RNA-seq reads. Split alignments within an intron where both acceptor and donor junction can be unambiguously resolved provide evidence of exon skipping events, along with evidence of splice junctions at the 5’ of donor and 3’ of acceptor sites. The following criteria must be met ---- there can’t be any transcript that uses both exons B and D.

Figure 5. Model for constitutive and novel exon skipping. The above figure shows the gene model describes three transcripts, the first one with exon combination A-B-C while the second transcript excludes the exon B and forms a transcript A-C, and the third one . In the data alignments, there is evidence of both A-B-C and A-C, along with evidence of a novel transcript A-D-C using a novel exon D.
Mutually exclusive exons

Mutually exclusive exons are detected when one transcript has an exon while another transcript has a different exon but not the first one. For simplicity, we only consider presence or absence of consecutive exons in the transcripts. Novel exons can be detected if their boundaries can be resolved by evidence of split alignments.

Figure 5. Model for novel mutually exclusive exons. In the above figure, the gene model describes three transcripts, A-B-C-D, A-C-D and A-B-D. The gene model doesn’t describe transcripts A-C-D and A-B-D, so these are novel. We can see the novel alternatively spliced forms either use B or C but not both. All three transcripts must be present to categorize this event as mutually exclusive splicing.
Results

Mapping sequence data to AGPV2

61% of the quality filtered reads could be mapped to the reference genome, 58% were uniquely mapped. 65% of the mapped reads consisting of exonic (63 %) and splice junctions (2%) were within annotated genes., Of the remaining mapped reads 15% were of introns while 20% mapped to intergenic regions. It is likely that the large fraction of reads (39%) that could not be mapped to the B73 reference were due to genomic differences between B73 and the elite inbred lines used in this study. Also, the sizeable fraction of reads mapping to the intronic and intergenic regions suggests that there are many novel transcribed regions and splice forms that the B73 reference genome does not describe. Maize is reported to have enormous genetic diversity and our results indicate a need for multiple reference genomes to accurately capture transcription and translation.

Figure 7. Percentage of reads mapping to different genomic features

54,542 (out of 110130) genes have just a single exon bin relative to the B73 v2 annotations).
Figure 8. Frequency distribution of number of exons per gene, X axis shows the number of exons per gene, Y axis shows the number of genes having a certain number of exons.

The gene GRMZM2G074906 had the most number of exon bins (70) with 13 transcript annotations.

**Exon-generation interactions for each inbred-hybrid combination**

78% of the exonic bin expression had significant interaction terms (FDR < 0.05). The exons detected to have significant interaction terms spanned across 64% of the genes across all families. Individual families also had similar percentages for rejecting the null hypothesis of no interaction.
Figure 9. Comparison of model-fitted expression values between LH1xLH123HT hybrid and its inbred parent LH123HT for exonic bins in gene AC149829.3_FG003.

The interpretation of significant interaction is straightforward when only a small fraction of exons of a gene show significant interactions. As in the case with AC149829.3_FG003, only the expression in the third exonic bin changes significantly between hybrid LH1xLH123HT and the parent LH123HT. This can be interpreted as alternative splicing occurring with all transcripts involving the third exon E003 in the F1 were expressed more than transcripts of the LH123HT-Parent. There were no significant exonic bin changes between F1 and its other inbred parent LH1.

Interpretation becomes complicated when most of the exons of a gene are differentially expressed because these methods can’t distinguish between overall expressional changes across all three family members. MMSeq (Turro et. al., 2011) and Cufflinks (Trapnell et. al.,
have devised approaches to assign reads to isoforms and look at isoform abundance rather than differential exon usage. Cufflinks provides the ability to compare isoform expression between samples with Cuffdiff functionality (Trapnell et. al., 2010a).

Assessment of alternative splice isoforms

We wanted to capture all splicing events for every inbred and hybrid, so we combined replicates for this analysis. We calculated frequencies of alternatively spliced genes as the proportion of all genes that have evidence of expressing multiple transcripts divided by number of genes that had at least one splice junction (Xi et al., 2013). We found on an average 9093 genes, from all four inbreds and six hybrids, had evidence from all three biological replicates of multiple annotated isoforms. The inbred parent, LH123HT had the smallest proportion of genes (34.6%) with evidence of alternative transcripts. The hybrid LH123HTxPHG39-F1 had the highest proportion (43.3%). Annotation of the B73 reference genome was used for detection of splice forms, we have not considered novel and ambiguous splicing events as a part of our analysis. To confidently state that we detected evidence of a transcript for a gene, we considered at least 2 reads to map to a genomic region unique to the transcript. Since reads that mapped to shared genomic regions can’t unambiguously ascertain which transcript produced the read, they were not included in the analysis. 15,310 genes (62.3%) expressed a single annotated isoform, although 3674 (24%) of them had evidence of novel exons or other splicing events that are not described in the B73 annotations.
Types of AS events

Figure 10. Frequency of types of splicing in inbreds and hybrids

We found about 47,874 splicing events per inbred or hybrid. Alternative 3’ acceptor sites were detected to be the most common type of splicing (33%). These events were detected by reads that extended the acceptor junction of an exon into the adjoining intron. The estimated proportion of intron retention isoforms was 28.7%, 5’ alternative donor 27.9%, exon skipping 8% and mutually exclusive exons 1.4%. Splicing was categorized as exon skipping when the event was confounded with mutually exclusive exons.

Splicing in inbreds and hybrids

The number of unique transcripts (31,902) spanned 24,593 genes in the hybrids was slightly higher than those detected in the inbred parents (31,858). We assigned the transcripts to genes and compared the number of unique transcripts per gene in the hybrids with that in their corresponding parents. About 936 genes (~9%) of the genes in the hybrid showed evidence of
unique annotated transcripts which were not detected in its corresponding parents. We used Splicegrapher (Rogers et. al., 2012) to visualize the AS events. To illustrate, consider figure 10, where evidence of different transcripts of the gene GRMZM2G074906 are compared in the hybrid and its corresponding parents, the hybrid shows an exon skipping event (green) not detected in its parents. This is a special case where the longer transcript in parent 1 (LH1) uses a subset of exons used by the longer transcript of parent 2 (LH123HT).  In this case the alternative

A  LH1-LH123 (F1 Hybrid)

B  LH1-Parent
Figure 11. Illustration of evidence for all transcripts in the LH1xLH123HT-F1, LH1-Parent and LH123HT-Parent. An exon skipping event is reported in the F1 (indicated by green), but no pattern of splicing inheritance is observed.

Figure 12. GRZM2G158729 gene showing an intron retention event (blue outline) where reads span across intron between two consecutive exons.
Abundance of isoforms

We wanted to test the hypothesis that there is a single isoform that is preferentially expressed by the organism. We found that for all multi-exonic genes (2681) with evidence of at least 4 transcripts, 88% (median % FPKM for prevalent transcript), of the reads would map to single major isoform.

![Figure 13. Plot of ratio of the FPKM value of the most abundant transcript (blue) and the median FPKM for all transcripts of a gene.](image)

We plotted the expression of the major transcript (blue) with the median expression of all transcripts (red) for genes showing at least 4 isoforms. In the blue-red ratio, blue consistently dominates for a vast majority of genes, indicating preferential expression of a single isoform. These results were consistent across all four inbreds and their six hybrids, suggesting that preferential isoform expression is consistent across all maize germplasm and may not be correlated with inbreeding levels.
Discussions

Modeling biological variability

We used a negative binomial model to account for the extra variation that the Poisson model cannot describe (Mukherjee et. al., unpublished). The method of detecting differential exon usage among the various genotypes within a family tests for significant exon-genotype interactions. Biological variability in expression of a genomic feature (gene, exon, transcript etc.) is described by the variation of mapped read counts among biological replicates. Some previous methods (Wang et. al., 2008a; Graveley et. al., 2010) used Fishers Exact test with contingency tables to compare differential exon usage between two conditions. (Katz et. al., 2010) with their tool MISO suggested an alternative way of using contingency tables, but still lacked consideration of sample to sample variability. (Anders and Huber, 2010b; Anders et. al., 2012) showed that tests that do not consider biological variation among replicates can create a vast number of false positives in detecting differential expression. Another approach to model transcript counts was to apply transformations to mimic a normal distribution and analyze results in a manner similar to data generated by microarray (AC't Hoen et. al., 2008; Cloonan et. al., 2008; Langmead et. al., 2009). The problem with such an approach is magnified with smaller counts, which even after transformation are distributed far from normality. Also normality does not take into account the mean-variance relationship that is often a feature of count data, resulting in biases statistical inferences. Discrete models have been historically used to model count data (Smyth and Verbyla, 1996; Smyth, 2004; Robinson and Smyth, 2007b; Robinson and Smyth, 2008a; Anders and Huber, 2010b; Bullard et. al., 2010; Robinson et. al., 2010; Anders et. al., 2012; McCarthy et. al., 2012). Mukherjee et al (unpublished) showed that transcript counts
that are modeled as a Poisson process will underrepresent variability as the mean number of counts increases.

**Bin based counts**

Alternative transcription based on significant exon-condition interactions can be easily detected if it is possible to detect differential usage exons that are unique to the isoforms. However, interpretations are complicated if the global expression pattern of the gene changes from condition to condition, all or most exons will show differential exon usage which does not necessarily infer expression of alternative transcripts. (Anders et. al., 2012; McCarthy et. al., 2012) have implemented analytical tools DEXSeq and edgeR respectively to test differential exon usage. Other methods by (Jiang and Wong, 2009; Trapnell et. al., 2010c; Turro et. al., 2011) directly assigns reads to the different isoforms and detect alternative transcript expression in terms of differential expression in isoform counts. The problem with mapping reads to isoforms is that isoforms often use shared DNA, thus assigning reads generated by WTSS technologies are associated with a level of uncertainty. A per-exon analysis can exclude exons shared exons and identify expression levels of exons that are unique to isoforms. Also per-isoform analysis can be restricted by the lack of functional evidence of proteins coded by the different isoforms. Given these caveats our results suggest the need to look at alternative methods for individual transcript level estimation.

**Global frequency of splicing**

Alternative splicing is extensive in eukaryotes and provides the organisms with increased plasticity and diversity of genome and proteome (Graveley et. al., 2010; Kalsotra and Cooper, 2011). We found that at least 38.7% of the genes expressed more than one transcript and had at
least one splice junction in at least one member of all six families. Our results were lower than that reported in the reference annotation (42.6%, (Wang et. al., 2009)), B73 endosperm tissue analysis (50.7%, (Lu et. al., 2013), and leaf transcriptome (56.4%, (Li et. al., 2010a)). Out of 62.3% the genes which showed evidence of just one transcript (according to annotation), 24% also showed evidence of novel exon boundaries, indicating possible isoforms not annotated for B73 reference. (Lu et. al., 2013) reported tissue specific isoform expression and found greater isoform abundance in embryo than endosperm. Different rates of splicing from different tissues and under different conditions indicate that splicing is tissue sensitive but rates don’t change drastically among different genetic backgrounds for the same tissue. The consistency of proportion of splicing events in our data-sets follows these results.

**Categorization of various splicing events**

3’ alternative acceptor site was the most common observed splicing (33%), followed closely by intron retention (28.7%) and 5' alternative donor site (27.9%), exon skipping (8%) and mutually skipped exons (1.4%). Our percentages were consistent with previous studies in plants, although most previous studies reported intron retention to be the most prevalent form of splicing. (Black, 2003; Wang and Brendel, 2006b; Barbazuk et. al., 2008; Filichkin et. al., 2010; Li et. al., 2010b; Zhang et. al., 2010; Marquez et. al., 2012; Lu et. al., 2013). To our knowledge, our study is the first to classify alternative splicing events and their frequency in maize. The estimated frequencies of various isoforms in plants are in sharp contrast to estimated frequencies in animals where exon skipping is reported to be the most common form of splicing (42%) while intron retention is the most rare (9%) (Barbazuk et. al., 2008; Sultan et. al., 2008; Wang et. al., 2008a; Daines et. al., 2011; McManus and Graveley, 2011).
Calculating Isoform abundance

To our knowledge this is the first study on global alternative splicing events in families consisting of inbred parents and their hybrid progeny. We used Cufflinks (Trapnell et. al., 2010a) to obtain isoform specific expression for the ten genotypes. The number of transcripts detected was slightly higher in the hybrids than the inbreds. But the number of transcripts that we detected is restricted by the AGPV2 annotations, and novel splicing events and transcribed regions were excluded from our analysis. Many of these transcripts might have premature stop codons and the protein products maybe broken down in the Nonsense-mediated decay (NMD) pathway (Cali and Anderson, 1998; González et. al., 2001; Ishigaki et. al., 2001; Isshiki et. al., 2001; Frischmeyer et. al., 2002; Conti and Izaurralde, 2005; Chang et. al., 2007; Neu-Yilik and Kulozik, 2008). We tested the hypothesis that the organisms preferentially express one transcript abundantly, although it remains to be tested if the major transcript codes for the most stable protein. Consistent with the hypothesis suggested by (Goff, 2011), we found that most genes preferentially produce a single transcript; some other splice forms are trivially expressed and might be attributed to ‘housekeeping’ and regulation activities (Greer et. al., 2010). We also found a small fraction of genes (9%) in the hybrid progeny that expressed novel transcripts not detected in its parents. This was similar to a previous Arabidopsis study. We suggest that these result justify ab initio analyses of protein structure for these novel splice forms, and investigate whether the hybrid benefits from expression of a splice form that its parents are incapable of producing. Another obvious and unanswered question is what happens to the other isoforms that are expressed minimally? mRNAs are reported to have a varied range of half-life (3-90 mins in yeast) depending on the function they perform. The mRNAs with shorter half-life include many
transcription factors and are believed to be involved in regulation while those with longer life span participate in central metabolic processes (Wang et. al., 2002).

**Protein metabolism**

Heterosis in the hybrid may be attributed to lower rates of protein metabolism than its inbred parents. (Mitton and Koehn, 1975; Hedgecock et. al., 1996; Bao et. al., 2005; Goff, 2011; Goff and Zhang, 2013) Although we could not directly infer differences in splice forms between inbreds and hybrids, we suggest future research on protein structural differences from their gene products using *ab initio* methods involving decoys (Mukherjee et. al. unpublished). Since we observed that both inbreds and hybrids preferentially express a single isoform, we can test the hypothesis that the prevalent isoform codes for the most stable protein structure among the many isoforms. If the (Goff, 2011) hypothesis is true, then the frequency of mis-folded proteins in the inbreds should be larger than in hybrids. It will be interesting to determine which predicted protein isoforms have the lowest energy states. From an applied perspective, the ability to computationally identify alleles coding for unstable proteins (Adzhubei et. al., 2010) will enable elimination of unstable alleles from the hybrid through breeding strategies. This will save breeders a lot of time and expensive field trials and enable them to create enhanced hybrids.
References


HTSeq-count.


CHAPTER 5

SUMMARY AND CONCLUSIONS

Since Darwin’s publication (1876), heterosis has been the focus of research in plant breeding and has been exploited extensively by commercial breeders. Superior characteristics in the highly heterozygous hybrid plants are assumed to be not only due to the genetic diversity of its parental alleles but also have been attributed to global changes in gene expression patterns between the hybrids and the inbreds. Fully inbred maize lines and their hybrids are fairly easy to produce, and its diploid genome exhibits enormous genetic variability, thus making it an ideal model system to study heterosis and inbreeding depression.

This dissertation represents a part of a larger study that was initiated by Dr. Richard (Dick) Johnson in 2008 to explore possible associations between gene expression, evaluated as genetic transcripts, and whole plant phenotypic expression of ear traits. The original mating design consisted of six families consisting of parental inbreds, their F1 progeny, F2 progeny and reciprocal backcrossed progeny. The six families were developed through a diallel crossing scheme among four elite inbred lines. The four elite inbreds consisted of two representatives from each of the two primary breeding pools that are used by commercial breeding companies to develop commercial hybrids. These breeding pools, referred to as Stiff Stalks (SS) and non-Stiff Stalks (NSS) produce hybrids that exhibit heterotic responses in traits of economic importance, especially those that are associated with grain production, for example cob length, cob weight, grain fill, ear weight, grain weight per ear etc. In his analysis, Dr. Johnson reported that genotypic variation in kernels per row was highly
correlated with cob length, cob weight, grain fill, ear weight, and grain weight per ear and only non-additive sources of variation were significant genetic components of these traits. Although the phenotypes were modeled as over-dominant, additive genetic variance (equated to cis-regulatory) was the most prevalent mode of gene action observed. One of the main purposes of this study was to see if using new elite inbreds in field conditions produced similar results to previous greenhouse studies with the historic B73xMo17 cross.

Over the past decade, there has been an ever increasing demand for technologies that produce quick, low-cost and more precise genome information than microarray. This challenge has been partially addressed with next-generation sequencing (NGS) technologies. The low-cost production of large sequence datasets is the primary improvement over conventional microarray methods. RNA-seq, also known as "Whole Transcriptome Shotgun Sequencing" ("WTSS") uses high-throughput sequencing technologies to provide researchers with a method to count transcripts experimentally. RNA-seq provides the ability to simultaneously assess transcript sequences and their expression levels. At an exploratory level of scientific investigation such information from maize inbreds and their hybrids provides quantitative measures of allelic and isoformic regulation in inbreds and hybrids. For purposes of this dissertation, we decided to focus on six aspects of transcriptome data analyses in the replicated samples from the four inbred parents and their six hybrids. These aspects include alignment to a reference, quantification of genomic features, normalization, modeling, allele-specific expression and alternative splicing. The overall goal of the analyses was to test the hypothesis that genetic mechanisms responsible for heterotic phenotypic responses of ear traits are revealed in transcripts of developing ear meristems. This
hypothesis was developed from observations of maize mutants in developing ears. Toward this goal, we addressed the following specific questions:

1. Is it possible to identify specific sets of genes that consistently exhibit differential expression between inbreds and hybrids within heterotic and non heterotic families?

2. Are parental alleles preferentially expressed in their hybrid progeny? What are the types of allele regulation in the hybrid?

3. To what extent is alternative splicing responsible for gene expression in inbreds and hybrids? Can we differentiate between patterns of isoform regulation in the hybrid and its inbred parents?

As our genetic materials were grown in the field in a replicated experimental design, inferences from this study provides greater power and larger scope of inference.

As a species, *Zea mays* consists of dozens of breeding populations (races) and is known to exhibit enormous genetic diversity both within and between breeding populations. Genetic diversity within a breeding population is orders of magnitude greater than diversity across the entire human population (citation). It has a highly polymorphic diploid genome with significant differences in copy number variation (CNV) and presence/absence variation (PAV) among different members within breeding populations This probably is the basis for the ability of Maize to rapidly adapt to environments throughout the world. At the same time it represents a challenge for transcriptomic studies such as reported herein. Although we were able to map a large proportion of sequencing reads to the B73 reference genome, we discarded a considerable proportion of reads from RNA-seq because they didn’t meet our mapping criteria, often failing to map to any part of the most advanced instantiation of the
reference genome. As a consequence, we have to recognize that all of our quantitative results are likely biased. It also emphasizes the need for multiple reference genomes in species that show great genetic diversity.

Before we addressed the specific objectives, it was important to remove the technical biases that are inherent with next-generation-sequencing data. Our samples were sequenced using one lane for each of the samples, so it was important to equalize the library sizes for purposes of comparison among replicates and among inbreds and hybrids. We decided that library scaling is too simplistic to compare gene expression across samples. Rather effective library sizes for each sample need to be computed for each sample to adjust for the RNA-composition bias. Next, the normalized data had to be modeled by fitting mean-variance relationships with appropriate distributional assumptions. Consistent with recent studies, we found the negative binomial distribution to be the most appropriate fit for our data, considering over-dispersion in transcript counts. We also found that the usual approach of treating data as normally distributed (e.g., Johnson et al, unpublished) is unsuitable for discrete RNA-seq transcript counts because assumptions of normally distributed errors in the transcript counts produced a large number of false positives during differential expression analysis.

To address the question of consistency among differentially expressed genes in hybrids and their inbred parents, we first quantified differentially expressed genes between either of the parents and the hybrid. We found different sets of genes that were consistently detected as being differentially expressed for the heterotic and non-heterotic inbred-hybrid combinations, indicating that regulation of underlying genetic networks is determined by the heterotic nature of family. Differentially expressed genes from both these groups (heterotic
and non-heterotic) were enriched with common functions related to translation, oxidoreductase activity, aromatic compound biosynthetic process etc. The largest fraction of differentially expressed genes in all six families exhibited additive gene action with respect to the mid-parent value. The remaining differentially expressed genes exhibited partial dominance toward the high parent expression counts. This is inconsistent with heterotic differential expression of whole plant phenotypic traits for the same set of inbreds and hybrids (Johnson et al, unpublished), although it does not provide unambiguous evidence in support of the dominance hypothesis of heterosis. Most differentially expressed genes were inconsistent among families, however at a higher level of organization these sets of genes belonged to the same metabolic pathways and were up-regulated in different parents and hybrids, indicating that organisms often utilize compensatory/complementary genetic networks to perform the same task. An extension of our differential expression analysis would be to look at differential expression of these data from the perspective of individual metabolic pathways to determine how gene regulation varies between inbreds and hybrids. A hypothesis to test is whether the hybrids are more efficient in metabolism than inbreds, and use minimum overhead to perform metabolic functions.

Evidence of high parent dominance in non-additive genes suggested that the alleles from the high parent affected the expression in the hybrid more than the alleles from the low parent. This motivated us to investigate if parental alleles were preferentially expressed in the hybrid. Based on single nucleotide changes, we formed haplotypes for parental alleles and quantitatively evaluated their relative expression levels in the hybrid. We found that a majority of genes that expressed specific alleles in the hybrids were also expressed differentially between its inbred parents. We also identified mono-allelic expression or allelic
specific silencing events for a small fraction of genes in the hybrids. *Cis-trans* statistical tests revealed that most alleles exhibit conserved expression, i.e., we couldn’t detect any significant expression differences between the parents, between the parental alleles in the hybrid and between the parents and the hybrids. These tests were based on small sample sizes from SNP level coverage, so it likely that the statistical tests did not detect as many differentially expressed genes as there actually are. Technology has advanced since we sequenced these materials; perhaps using longer reads and more sequencing depth for the same materials will help us eliminate false negatives. Consistent with previous studies, *cis-*regulation was found to affect alleles more than trans-regulation and *cis* regulation was strongly correlated with additive mode of gene action. In conjunction with *cis* and *trans* acting genetic changes, the interaction of DNA- methylation machineries of the two parental genomes is also believed to play a major role in regulating epigenetic changes. A future study of these materials should include evaluation of epigenetic analyses including methylation, miRNA expression and siRNA distribution patterns in the inbreds and hybrids.

Greater performance in hybrids than inbreds has been hypothesized to be due to lower energy metabolism rates and decreased protein degradation in hybrids than its parents. As a result, the hybrids save energy from primary metabolism and translate it into biomass and vigor. Homozygous inbreds have a single allele of a gene and are thus limited in their ability to yield alternatively spliced transcripts. Heterozygous hybrids on the other hand, receive alleles from both parents and can potentially produce a larger number of splice variants. Our analysis revealed that the hybrid preferentially produces parental splice forms, but for a small fraction of genes, the hybrids produce splice forms that are not observed in the parental set. The reason hybrids to create new splice forms is not explained, although it can be
hypothesized that the novel splice forms code for more stable and functionally efficient protein structures. The new splice forms may also result from “genomic shock”. It is important to note, however that despite the presence of multiple splice isoforms, 88% of the genes preferentially expressed just a single isoform. Why one isoform was preferentially expressed remains to be determined. We hypothesize that the protein structure coded by the most prevalent splice isoform is the most stable. If this hypothesis is true, the alleles and isoforms that code for the unstable proteins will be down-regulated in the hybrids. Progressive polyploids have exhibited greater levels of growth and heterosis than diploids, indicating that more choice of alleles and isoforms provides the hybrid with flexibility to avoid abnormal translations. With the advancement in development of protein-structure prediction algorithms, it is now possible to computationally investigate the functional potency and stability of proteins based on sequences of the RNA transcripts. If we can detect abundance of transcripts in the hybrid which code for unstable and functionally inefficient protein structures, it is possible to eliminate these deleterious alleles from the hybrid through breeding strategies. This could save time and resources dedicated to hybrid improvement through trial and error.

Concluding Remarks

The manuscripts comprising this dissertation are unique relative to contemporary publications on differential gene expression between inbreds and hybrids:

1. The genetic material consisted of advanced unrelated lines. Previous inferences from similar studies in maize were primarily done with B73, Mo17 and their reciprocal
hybrids which were retired from the most advanced commercial breeding pools at least 25 years ago.

2. The plants were grown in replicated field conditions as compared to previous studies that used unreplicated biological samples grown under controlled conditions. The use of biological replications in our study provided measures of biological precision and provides quantifiable measures of confidence for our inferences.

3. We used RNA-seq as the technology platform instead of microarrays. This next generation sequencing technology is more repeatable and sensitive than microarrays, and it provides a more accurate estimate of transcript levels than measures based on spectral intensities that microarrays provide.

4. We normalized and modeled the data accurately. Previous studies on heterosis have not emphasized the importance of biological replication, nor removal of systemic biases associated with the technology.

5. Combining transcript sequence and expression levels, enabled us to compare genetic changes, allelic expression and alternative splicing in hybrids and inbreds. To our knowledge, no other study has been published so far that has looked into global patterns of alternative splicing in an inbred-hybrid system.

6. Our approach and results provide preliminary evidence for the recent (Goff, 2011) hypothesis to explain heterosis.