Computational approaches for analyzing variation in the transcriptome and methylome of *Zea mays*

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Roman Vladimir Briskine

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Chad L. Myers

December, 2013

Acknowledgements

In my years of graduate study, I had the privilege to work with many talented and insightful people whose help and support were instrumental to my success as a budding scientist. I would like to express my sincere appreciation and gratitude to all of them. First of all, I thank my adviser, Chad Myers, who was always around to answer my questions and provide helpful suggestions to get me on the right track. His guidance was essential for me to finally discover 'the big picture' and to figure out how 'to tell a story'. Also, I would like to thank Nathan Springer, whose suggestions were always very precise and astute. My discussions with Nathan were always engaging and motivating. I am grateful to Rui Kuang and Dan Boley for great comments during my oral examinations. I also thank Rui for substituting for one of the committee members at my oral exam on short notice and for chairing my final defense committee.

Also, I deeply appreciate the outstanding mentorship I received from Nevin Young and Kevin Silverstein. Nevin's guidance was crucial for my understanding how computer scientists can efficiently communicate with biologists. He also introduced me to plant genomics, a fascinating field in which I plan to work after graduation. I am grateful to Kevin for his constant faith in me. His advice often went beyond science into the dark realm of social interaction, and I would not be nearly as capable today without his help. He gave me many opportunities to expand my skills as bioinformatician, which I now carry into my career. In addition, I would like to thank Peter Tiffin and Mike Sadowsky for scientific discussions and Jane Glazebrook for introducing me to genetics. My former adviser, Arindam Banerjee, guided me through my Master's project. Without Arindam's support, it would have been difficult to stay in graduate school long enough to obtain a Master's degree. I am also grateful to Leslie Wilson, Roberta Roth, and Craig Johnson for writing keen recommendation letters which got me into the graduate school in the first place.

I am grateful to my collaborators who also worked on the projects described in this dissertation. It was a great pleasure to work with Ruth Swanson-Wagner, Steven Eichten, Robert Schaefer, Amanda Waters, Qing Li, Peter Hermanson, Evan Starr, and Patrick West. Without fruitful collaboration with other institutions, this research

would not have been possible. In particular, I would like to thank Rajandeep Sekhon, Shawn Kaeppler, Natalia de Leon, Robin Buell, Candice Hirsch, Jeffrey Ross-Ibarra, Matthew Hufford, Matthew Vaughn, Jawon Song, and Irina Makarevitch for sharing their data, results, and ideas. I also thank Elizabeth Koch, Scott Simpkins, Robert Schaefer, Carles Pons, and Kathryn Briskina for proofreading and commenting on this manuscript.

My graduate experience would not be the same without stimulating and intellectual environment formed by the other students and post-docs. Therefore, I would like to extend my gratitude to all students and post-docs with whom I interacted while working on my degree. In particular, I thank Antoine Branca, Arvind Bharti, Benjamin VanderSluis, Brendan Epstein, Carles Pons, Elizabeth Koch, Jeremy Bellay, Jeremy Yoder, John Stanton-Geddes, Joseph Guhlin, Justin Nelson, Masayuki Sugawara, Peng Zhou, Raamesh Deshpande, Rachel Hillmer, Scott Simpkins, Timothy Kunau, Timothy Paape, Wen Wang, Yungil Kim.

Finally, in graduate school and in life, I am thankful to my parents, Liudmyla and Vladimir, brothers Igor and Vladimir, our grandparents Ivan and Liubov, Lev and Olena, my wife Kathryn, my in-laws Leslie and Mary Morris and the Andersons who reminded me of the larger purpose behind my research.

Dedication

To my grandfather, Ivan Chetveryk, who had the insatiable desire to learn.

Abstract

While Mendelian genetic approaches to crop improvement have been successful in the past, effective modern breeding programs are becoming increasingly dependent on accurate information about gene functionality and regulatory mechanisms. Recent advances in sequencing technologies have produced the complete genomes of many organisms, but the annotation of predicted genes still lags behind. Since domesticated varieties tend to be phenotypically divergent from their ancestral species, the examination of domestication effects on their transcriptomes can provide beneficial insights into the function of genes targeted during domestication.

This dissertation focuses on computational approaches for comparative analysis of gene expression, which is a valuable resource for gene annotation. We begin with the analysis of two co-expression networks built on expression data from maize and its wild ancestor, teosinte. We reveal biologically significant differences between the two networks and propose a novel method to identify genes with altered expression covariation between the two species. We show that our approach is more sensitive than existing methods and illustrate its complementarity to differential expression or genome sequence analysis. The approach is also applied to study differences between networks derived from RNA-seq and microarray gene expression data, where we identified and resolved issues with comparing and combining co-expression networks derived from the two data types.

In the second part of the dissertation, we describe a pipeline for the identification of differentially methylated regions in maize and teosinte. Application of this approach to a diverse set of maize lines suggests the presence of purely epigenetic alleles and confirms the prevalence of the negative relationship between DNA methylation and the expression levels of nearby genes.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Bioinformatics lies at the intersection of computer science and biology. Its main goal is the development of computational tools and algorithms for the storage, processing, and analysis of various biological data. In particular, substantial efforts are directed towards characterization of gene functions and elucidation of cellular processes. Owing to the rapid development of sequencing technologies, complete genomic sequences along with many predicted genes are already available for a multitude of species. However, the annotation of predicted genes is far from being complete even in model organisms. Gene expression analysis plays an important role in the gene annotation process. By measuring the differences in gene expression across various tissues or accessions, one can identify genes whose regulation is responsible for phenotypic changes. However, differentially expressed genes cannot reveal the whole picture. Some genes that do not necessarily exhibit differential expression may still have radically different co-expression patterns. The analysis of the co-expression relationships among the genes is also crucial to understanding biological processes that occur inside the cell. In addition, heritable regulation of gene expression is not limited to changes in DNA sequence. Various chromatin modifications such as DNA methylation or histone tail phosphorylation may also modulate the expression of nearby genes. This dissertation focuses on computational approaches to the analyses of expression and DNA methylation data in an important agricultural

crop *Zea mays*.

This introductory chapter provides a brief overview of the biological concepts that appear throughout the rest of the dissertation. We will also frame the major analytical problems that the dissertation tries to address, describe the overall flow of the dissertation, and provide a brief summary of each chapter.

1.1.1 Zea mays

Maize, *Zea mays* ssp. *mays*, is a major agricultural crop whose yearly production by weight (875 million tonnes in 2012) exceeds all other cereals, including rice (718 million tonnes), wheat (675 million tonnes), and barley (132 million tonnes) (FAO, 2013). Maize has a wide range of uses from staple food and animal feed to biofuel production. While innovative breeding has enabled the increase of maize yields in some geographical locations (Hafner, 2003; Troyer, 2006), the rise in demand for maize and other crops currently exceeds the combined production increase from improvement in the yield per hectare, the expansion of the planting area, and the increase in harvest frequency (Ray et al., 2012). Therefore, any research that facilitates the development of agricultural methods to augment the production by increasing tolerance to unfavorable environmental conditions, strengthening pest resistance, or reducing space requirements would be highly beneficial.

The maize genome is relatively large, totaling 2.3 gigabases (Schnable et al., 2009), which is considerably bigger than the 0.16 gigabase genome of the model plant *Arabidopsis thaliana* but smaller than the genomes of some other crops such as common wheat *Triticum aestivum* that approaches 17 gigabases (Bennett and Leitch, 2012). The genome of maize is also highly complex. As did many other plant species, the ancestor of maize underwent several whole-genome duplication events (Schnable et al., 2009) with the most recent one taking place about 5-12 million years ago (Swigoňová et al., 2004). Another factor that affects genome size is the activity of transposable elements, short DNA sequences that can reposition themselves within the genome after being cut or copied. Transposable elements (TEs), or transposons, constitute a large portion of the maize genome (SanMiguel et al., 1996) and have contributed to its growth in the last ~3 million years (Schnable et al., 2009). Transposons are also likely to be responsible for maize's rapid rate of

genome evolution, an effect of transposons that has been observed in many other organisms (Kazazian, 2004).

Domestication studies make a considerable contribution to crop development in particular by identifying the genes that affect desirable traits. Due to the great agricultural importance of maize, the process of its domestication has been extensively researched. Cytogenetic (Doebley, 2004), molecular genetic (Matsuoka et al., 2002), and archeological (Piperno et al., 2009) data suggest that maize domestication occurred in a single event approximately 9,000 years ago in Balsas River basin of southwestern Mexico and its sole wild ancestor was *Zea mays* ssp. *parviglumis* commonly referred to as teosinte. The name "teosinte" may also apply to a number of other species from the genus *Zea* including *Zea diploperennis*, *Zea perennis*, *Zea luxurians*, *Zea nicaraguensis*, *Zea mays* ssp. *mexicana*, and *Zea mays* ssp. *huehuetenangensis*. Unless stated otherwise, we will use the term "teosinte" to refer exclusively to *Zea mays* ssp. *parviglumis*. In addition, we will occasionally distinguish between the traditional maize landraces, i.e. lines that have seen little or no improvement after domestication, and the improved hybrid lines that have experienced significant enhancements due to extensive breeding efforts.

In evolutionary terms, several thousand years is a very short time. Yet, the morphological differences between maize and teosinte are quite striking. For instance, teosinte generally produces multiple stalks yielding several ears with 5-7 grains encased in hard shells while cultivated maize spawns a single large stalk that bears a single ear with several hundred shell-less kernels (Figure 1.1). Even though domestication normally introduces various genomic changes, cultivated varieties belong to the same species and possess considerable genomic similarity to their wild progenitors. Indeed, comparative hybridization analysis of *Zea mays* confirmed genomic similarity between maize and teosinte despite their remarkable phenotypic variations (Swanson-Wagner et al., 2010). Moreover, crosses between maize and teosinte result in viable hybrids (Doebley and Stec, 1993) further supporting their genomic similarity.

In the maize domestication literature, factors that cause phenotypic differences between maize and teosinte are often divided into two groups (e.g. Yamasaki et al.,

Figure 1.1. Ears of teosinte *Zea mays* ssp *mexicana* (left) and maize *Zea mays* ssp *mays* (right) exemplify highly divergent phenotypes of the two subspecies. Yet, their cross-hybridization (center) indicates that they remain relatively similar genetically. Photo by John Doebley; retrieved from `http://teosinte.wisc.edu/images.html` on 07/12/2011.

2005; Buckler et al., 2006; Hufford et al., 2012). The first includes the factors influencing the traits that made some of the plants more desirable during the domestication event. The second group encompasses the selection targets that were affected during more gradual but extensive breeding efforts in subsequent years. What constitutes those factors and how can they be identified?

One way to approach those questions is through DNA sequence analysis (e.g. Vigouroux et al., 2002; Wright et al., 2005; Doebley et al., 2006; Hufford et al., 2012). Modifications in DNA sequence and structure are highly heritable and may influence protein structures and gene expression levels. However, the actual functional effects of sequence changes may be difficult to determine especially when a single modification impacts several seemingly unrelated traits (pleiotropy.) Therefore, the application of additional methods such as differential expression analysis can provide unique insights into domestication and improvement process. Moreover, alterations

in gene expression may be heritable even without any underlying genetic changes, the phenomenon known as epigenetic variation. We will describe genomic variation, expression variation, and epigenetic variation in more detail later in this chapter. Other types of analysis such as proteomics and metabolomics are beyond the scope of this dissertation.

1.1.2 Genomic Variation

The genome is a collection of an organism's hereditary information encoded in its DNA or, for some viruses, in its RNA (Pevsner, 2009). Genomic variation can be broadly characterized by the length of the affected DNA sequence (Scherer et al., 2007). Structural variation describes the alterations to DNA segments of 1 kbp or more but smaller than whole chromosomal changes and includes all types of segment deletion, insertion, duplication, translocation, and inversion. Another category, sequence variation, is comprised of single base changes in DNA sequence known as single nucleotide polymorphisms (SNP.) Finally, some researchers classify alterations between 2 bp and 1 kbp in length as structural variation (Scherer et al., 2007) while others consider them a separate category referred to as insertion-deletion (indel) polymorphisms (IDP) (Springer et al., 2009) or simply indels (Feuk et al., 2006).

Maize is highly diverse in terms of both structural and sequence variation (Buckler et al., 2006; Messing and Dooner, 2006; Springer et al., 2009; Swanson-Wagner et al., 2010; Hufford et al., 2012). Comparative genomic hybridization (CGH) studies identified several hundred copy number variants (genes that have more copies in some genotypes) and a few thousand presence-absence variants (genes that are present only in some genotypes) (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010). The majority (70%) of structural variants were found in both maize and teosinte lines (Swanson-Wagner et al., 2010) indicating that they arose before the domestication event. Based on the genic SNP data, two maize lines may be more different than humans are from chimpanzees (Buckler et al., 2006). Whole genome resequencing of 75 wild and domesticated maize lines uncovered over 20 million high-quality SNPs and found a higher retention rate of nucleotide diversity between the wild ancestor and domesticated lines (Hufford et al., 2012) than reported for rice (Caicedo et al., 2007) or soybean (Lam et al., 2010).

On the gene level, we expect to see lower nucleotide diversity in genes under selection pressure compared to neutral genes. This holds for genes that have been targeted during the maize improvement process. However, the population bottleneck that resulted from the maize domestication also severely reduced genetic diversity (Wang et al., 1999; Buckler et al., 2001) making distinction between domestication and improvement gene candidates more challenging (Vigouroux et al., 2002). In addition, genomic regions around a targeted gene tend to be selected as well, the phenomenon known as linkage disequilibrium (LD) (Reich et al., 2001). Thus, completely neutral genes which reside in these regions may appear under selection. This effect is partially mitigated in maize because it is an outcrossing species with a large population size, low LD, and high levels of recombination (Remington et al., 2001; Fu et al., 2002; Vigouroux et al., 2002).

A large number of maize studies analyzed genomic variation to identify domestication and improvement gene candidates. One of the early popular approaches was Quantitative trait locus (QTL) mapping, which leveraged segregation properties of the offspring (Doebley and Stec, 1993; Doebley et al., 1994; Briggs et al., 2007). The method consists of crossing two homozygous parents, followed by selfing the obtained heterozygous first generation (F_1) progeny, and screening the second generation (F_2) progeny for the markers unique to each parent (Tanksley, 1993). The region around the marker that is statistically associated with the trait is called a QTL. However, the identified regions are often very large and may contain numerous genes. The analysis is further complicated by the tendency of genes to have low diversity under strong selection pressure. Even though some improvement genes were identified by analyzing QTLs in crosses between improved lines and landraces (Yamasaki et al., 2005), QTL analysis overall helped to characterize only a limited number of genes in crop species (Yamasaki et al., 2007).

To identify domestication and improvement gene candidates, association studies similarly rely on segregation properties to find significant correlations between genotypic and phenotypic variants. Genome-wide association studies (GWAS) identified, for example, genomic loci linked to geographic adaptation (Gore et al., 2009) and leaf architecture traits (Tian et al., 2011). Both features have agricultural importance and have been extensively targeted during breeding to expand the crop range and

increase the plant density respectively. Crop species tend to have highly structured populations, which has to be taken into consideration to avoid spurious correlations in association analysis (Rafalski, 2002). To address this issue, Thornsberry et al. (2001) estimated the population structure of maize by applying a Bayesian analysis to 141 simple sequence repeat (SSR) loci. SSRs are very short (1-6 bp) repetitive DNA sequences that can be used as genetic markers due to their polymorphic nature (Pevsner, 2009). The estimated population structure combined with association analysis allowed the authors to identify *Dwarf8* flowering time gene as a selection target (Thornsberry et al., 2001).

Vigouroux et al. (2002) exploited polymorphic properties of SSRs from genic regions to uncover several improvement gene candidates. The authors looked for deviations from the equilibrium model and adjusted for the loss of diversity that happened during domestication. Their subsequent study compared the diversity of SSRs between maize and teosinte to find that the diversity of dinucleotide repeats almost recovered after the domestication bottleneck while longer SSRs still exhibited deficit in maize relative to teosinte. However, the study found no relationship between SSR diversity and proximity to domestication QTLs (Vigouroux et al., 2005).

Wright et al. (2005) applied population genetics methods to the SNP data from 14 maize and 16 teosinte lines to measure the effects of domestication and improvement on maize genome. The authors estimated that between 2% and 4% of approximately 1200 screened genes were potential targets of selection. Yamasaki et al. (2005) employed a similar approach on a different set of approximately 1000 genes but screened 35 candidates in additional improved and landrace lines to confirm six domestication and nine improvement candidates.

In a much larger effort, Hufford et al. (2012) resequenced the whole genomes of 35 improved maize lines, 23 landraces, and 17 wild relatives. Using the cross-population composite likelihood ratio (XP-CLR) method that screens contiguously linked loci for extreme differentiation in allele frequency (Chen et al., 2010), the researchers identified 484 domestication and 695 improvement regions. They reported that some of the regions were devoid of any annotated genes and potentially contained some special regulatory sequences, while other regions spanned multiple

annotated sequences (1,764 and 1,506 genes in domestication and improvement regions respectively.) Domestication and improvement gene candidates were selected based on the proximity to the site with the highest likelihood ratio, yielding 468 domestication and 571 improvement gene candidates (Hufford et al., 2012).

Sequence-based methods have been very beneficial for finding domestication and improvement targets of selection. However, they do not provide the full picture of domestication and improvement effects on maize. First of all, sequence-based methods can only measure the upstream effects. Polymorphisms in the gene body rarely modify the expression level of that gene but the altered structure of the gene may affect the expression of many other genes downstream in the pathway and those changes will be completely invisible to the sequence-based methods. Moreover, we do not have complete knowledge of all the pathways in maize, so it will be very hard to uncover all consequences of a particular polymorphism. Second, some causal polymorphisms reside in promoter and enhancer regions on either side of the gene body (in-*cis* regulation.) Since those regions vary in length, it may not be clear which nearby gene the polymorphism controls, especially when dealing with gene-rich loci. Furthermore, a polymorphism may occasionally affect a gene located very far away or even on a different chromosome (in-*trans* regulation) and the sequence-based methods are unable to identify such relationships. Third, a trait may be influenced by several polymorphisms that are individually optional and the sequence-based methods may not have sufficient statistical power to detect such factors.

1.1.3 Gene Expression

Gene expression analysis addresses many of the aforementioned issues with sequence-based methods by providing essential information about gene relationships and domestication effects on them. Gene expression occurs when a DNA sequence is transcribed into RNA. The collection of data that describes expression levels of all RNA molecules (messenger RNA, ribosomal RNA, transfer RNA, and non-coding RNA) as measured in a specific tissue at a certain time is called a transcriptome (Pevsner, 2009). Early methods such as Northern blotting and reverse transcription polymerase chain reaction (RT-PCR) measured expression on gene by gene basis and

required considerable efforts. With the invention of high-throughput approaches such as serial analysis of gene expression (SAGE) (Velculescu et al., 1995), reconstruction of a whole transcriptome finally became feasible (Velculescu et al., 1997). Another high-throughput method, microarray analysis, started to gain popularity approximately at the same time (Schena et al., 1995). Despite some drawbacks, microarrays required considerably less mRNA than SAGE and surpassed SAGE in efficiency, which made them a preferred choice for experiments involving multiple biological samples (Ye et al., 2002).

A microarray chip consists of a solid substrate with complementary DNA (cDNA) probes densely attached in a grid fashion. These oligonucleotide probes usually represent a portion of a known gene. For one-channel detection, purified RNA extracted from a tissue sample is reverse transcribed into cDNA, labeled with a fluorescent dye and hybridized to the chip. After washing the chip to remove the sequences that failed to hybridized, chip scanning is performed to measure the fluorescence that would be commensurate to the amount of cDNA bound to the probe. Microarrays can also be used for two-channel detection whereby two samples (e.g. case versus control) are labeled with two different dyes and hybridized to the same chip simultaneously. In this case, the relative intensities of each dye can be used to determine whether genes are up or down regulated. (For a detailed review of microarray technology see Pevsner (2009).)

While the initial design of a microarray chip is relatively expensive, the subsequent chip construction and expression profiling is very cheap and enables rapid measurement of the expression levels for thousands of genes at a time. Microarray analysis made a substantial and valuable contribution to our understanding of gene expression in many organisms. However, it also possesses multiple weaknesses. First of all, sequences for microarray probes are derived from the known gene models. Hence, it would be impossible to measure expression for genes and potentially rare alleles that have not been identified at the time of the chip design. Structural variation and extensive SNPs in non-reference lines may also affect hybridization efficiency and confound the measurements. Even though the probes are constructed to be unique and as disparate from other genomic sequences as possible, some level

of cross-hybridization is practically unavoidable. This introduces potentially considerable background noise and restricts the ability to measure expression of highly similar (homologous) genes. Finally, microarrays suffer from signal saturation preventing the accurate evaluation of highly expressed genes.

Until recently, microarray analysis was the main method for measuring gene expression. However, recent improvements in the "next-generation" sequencing, including RNA sequencing (RNA-seq,) shattered microarray hegemony and may soon even relegate microarrays to a niche technology. RNA-seq involves several major steps. As with microarrays, RNA isolated from a tissue sample needs to be purified to remove contaminants. To create the library, sequencing adapters are ligated to one (single-end sequencing) or both (paired-end sequencing) ends of each fragment. The library is then sequenced from one or both ends using a high-throughput deep DNA sequencing platforms (sometimes called next generation sequencing or NGS platforms) such as Illumina HiSeq or Applied BioSystems SOLiD. The obtained reads are either aligned to an existing reference genome or assembled de-novo into a transcriptome. Expression levels for each gene are customarily reported in Reads Per Kilobase of transcript per Million mapped reads (RPKM) for single-end sequencing and Fragments Per Kilobase of transcript per Million mapped fragments (FPKM) for paired-end sequencing.

Unlike microarrays, RNA-seq is not limited to known gene models and measures the levels for all RNA transcripts. However, the ability to detect rare transcripts depends on the depth of sequencing, i.e. the total number of bases to be read. If the sequencing is not very deep, highly expressed genes may prevent the rare transcripts' sequences from being captured. Given adequate sequencing depth, RNA-seq offers considerably better dynamic range than microarrays. It can also better distinguish between highly homologous sequences, albeit mapping parameters may need to be adjusted to prevent reads from mapping to multiple regions. Furthermore, RNA-seq allows improving the quality of the transcriptome by realigning the reads to an updated version of the reference genome once it becomes available.

While RNA-seq is more flexible and cost efficient enough to replace microarrays in most applications, it would still be highly beneficial to leverage the existing wealth of microarray expression data when performing gene expression analysis. Even though

it is hard to compare microarray and RNA-seq results directly and quantitatively, methods that can operate on the data combined from the two platforms would be very useful. For instance, the construction of co-expression networks from the mixed microarray and RNA-seq data is of particular interest and will be mentioned again later in this dissertation.

To complement and expand the findings of the DNA sequence-based studies about the effects of maize domestication, it is necessary to analyze and contrast gene expression data for maize and teosinte lines. There are two main approaches to find genes whose expression has been altered by domestication and subsequent improvement. One way is to search for genes that are differentially expressed in maize and teosinte. Many statistical methods have been developed for both microarray (Grant et al., 2007) and RNA-seq (Garber et al., 2011) data sets. However, these methods cannot detect genes whose expression variation within a species is higher than between the species.

Alternatively, one can investigate how a gene's expression co-varies relative to the other genes within the same species. This can be achieved by calculating similarity (e.g. the Pearson correlation coefficient) between expression profiles of each gene pair. A symmetric matrix that contains all such similarity measures is called a co-expression network. Genes that were targeted during domestication and improvement are likely to have different relationships in maize and teosinte co-expression networks and can be identified by comparing the co-expression profiles of the gene in the two networks. This method is complementary to the differential expression analysis because a gene's co-expression profile may change even when its expression levels remain the same.

While the expression of genes ultimately determines phenotype, gene expression levels depend on many factors that go beyond DNA sequence variation. In particular, there are many types of chromatin (the collection of DNA and proteins within a cell's nucleus) modifications that influence gene expression without altering DNA sequence. Thus, for complete understanding of molecular machinery behind phenotypic alterations it is necessary to expand the research beyond DNA sequence and expression variation. In the next section, we will review one of these mechanisms called DNA methylation.

1.1.4 Epigenetics and Methylation

In the past, the word 'epigenetics' had two different albeit related meanings. Originally, the term was introduced by Conrad Waddington to indicate how the interaction among genes during development determines the phenotype (Waddington, 1957). However, the term was later adopted to signify the study of heritable functional modifications that cannot be entirely attributed to the variations in DNA sequence (Russo et al., 1996; Bird, 2007). At present, molecular biologists typically use the second definition of the word (Haig, 2004; Springer, 2013) and that meaning will be adopted throughout the rest of the dissertation.

There are several molecular mechanisms behind epigenetic inheritance. Structural inheritance includes the inheritance of a specific spatial structure that conforms to the template structure present in the mother cell (Jablonka and Raz, 2009). For example, prions are proteins with alternative heritable conformations that propagate from a mother cell to a daughter cell (Jablonka and Raz, 2009). While prions may cause highly infectious diseases by converting host proteins into the virulent prion conformation, there are also examples of beneficial prion variants (Rando and Verstrepen, 2007). Another mechanism involves non-coding RNAs whose regulation extends from dosage compensation to gene silencing via post transcriptional or post translational modifications (Goldberg et al., 2007). Finally, chromatin modifications comprise a large group of molecular mechanisms that include histone tail modifications (Kouzarides, 2007), histone variants (Law and Cheung, 2013), and DNA methylation (Goldberg et al., 2007). While chromatin modifications are often referred to as 'epigenetic marks', it is important to note that some of them may not be heritable (Springer, 2013).

Unlike certain histone modifications, DNA methylation was shown to be highly heritable (Bird, 2002; Springer, 2013). DNA methylation entails the addition of a methyl group to a cytosine base. In mammals, DNA methylation happens predominantly in symmetric cytosine-guanine (CG) contexts and affects around 70%-80% of all dinucleotides (Bird, 2002). Plants can have DNA methylation in CG, CHG, and even asymmetric CHH contexts (Henderson and Jacobsen, 2007) where H stands for one of adenine, cytosine, or thymine nucleotides. However, methylation frequency in plants is context-dependent (Cokus et al., 2008) and varies from species to species

(Zemach et al., 2010; Feng et al., 2010).

DNA methylation is an ancient process (Zemach et al., 2010) that performs multiple important functions. Bacteria use DNA methylation to mark their own genomes, thus protecting them from restriction enzymes that degrade exogenous DNA from infectious bacteriophages (Bestor, 1990). In mammals, DNA methylation plays an important regulatory role during development (Lister et al., 2009; Hawkins et al., 2010). Yet, the evidence for methylation's involvement in developmental regulation in plants is quite limited (Eichten et al., 2013b). Both plants and mammals extensively employ DNA methylation for defending their genomes against transposable elements (Bird, 2002; Zhang et al., 2006; Henderson and Jacobsen, 2007), gene silencing (Suzuki and Bird, 2008), as well as for reducing transcriptional noise in intergenic regions (Bird, 1995).

DNA methylation occurs in both genic and intergenic regions. The evidence for the negative correlation between DNA methylation in promoter regions and gene expression is substantial (Zhang et al., 2006; Henderson and Jacobsen, 2007; Zilberman et al., 2007). Although positive correlation between methylation and gene expression has been reported as well, they appear to be less common (Bell et al., 2011; Natt et al., 2012; van Eijk et al., 2012). It is possible that these associations either are spurious or exact in-*trans* control over other regulatory genes that in turn influence the expression of the nearby genes. The negative correlation between methylation and expression represents an intriguing alternative to null mutations (gene knock-outs) that are commonly employed in plant research and breeding. Unlike the null mutations that generally lead to the loss of function, methylation alterations, if successful, may enable gain-of-function modifications (Springer, 2013). The exact reasons for gene body methylation are currently unknown (Zhang et al., 2006; Law and Jacobsen, 2010) but it was suggested that CG methylation in gene bodies might silence cryptic promoters to prevent unintended gene regulation (Tran et al., 2005; Zilberman et al., 2007) or to guard genic sequences from transposon insertions (Regulski et al., 2013).

Loci that exhibit variable DNA methylation in different genotypes are referred to as epialleles. Multiple studies reported relationships between epialleles and local (*cis*) or distant (*trans*) genomic variation (Natt et al., 2012; Chodavarapu et al.,

2012; Schmitz et al., 2013b). Based on the type of such relationship, Richards (2006) proposed to classify epialleles into three categories. Methylation state of *obligatory* epialleles is completely determined by local DNA variation. In *facilitated* epialleles, DNA variation influences the epiallele's methylation but the effect is rather probabilistic than stable. Finally, *pure* epialleles are entirely independent from any DNA variation in *cis* or in *trans*. The distinction is noteworthy because the state of obligatory epialleles can be predicted from DNA sequence without methylation profiling. Moreover, these epialleles are much more stable and less likely to change their state in subsequent generations compared to pure and faciliated epialleles.

Several techniques have been developed for methylation profiling (reviewed in Laird, 2010). Three of them are capable of producing methylation measurement on a genome scale. The enzymatic approach employs methylation-sensitive restriction enzymes that cut at unmethylated sites. Patterns exhibited by the obtained fragments allow determination of methylation states of individual cytosine residues (Allegrucci et al., 2007). Another approach, affinity enrichment, relies on antibodies that attach to methylated sites causing subsequent immunoprecipitation that separates methylated fragments from unmethylated. The fragments can be hybridized to a microarray (ChIP-chip) or sequenced (ChIP-seq) to obtain the whole methylome (Eichten et al., 2011; Taiwo et al., 2012). Finally, bisulfite conversion is based on sodium bisulfite treatment of denatured genomic DNA that converts unmethylated cytosine residues to uracil. While special microarray techniques were attempted for bisulfite-treated DNA, hybridization remains challenging (Laird, 2010). As a result, bisulfite conversion is predominantly used in conjunction with sequencing as in MethylC (Lister et al., 2008) or BS-seq (Cokus et al., 2008) methods.

Each of the three approaches have certain strengths and weaknesses (Laird, 2010). For example, enzymatic and affinity enrichment approaches can only detect methylation in the CG context. Moreover, they tend to suffer from relatively low resolution even when coupled with sequencing. Bisulfite conversion may experience incomplete conversion bias but covers all sequence contexts and achieves single-base resolution. While the latter may be important in particular cases, the methylation state of a single cytosine lacks stability and is susceptible to spontaneous mutations even between consecutive generations (Becker et al., 2011; Schmitz et al., 2011).

Although several reports suggested associations between methylation state of a single cytosine residue and phenotype (Xu et al., 2007; Moser et al., 2008), the majority of methylome studies concentrate on the analysis of differentially methylated regions (DMRs) on a scale from a few hundred to several thousand base pairs (Bock, 2012). Methylation levels across DMRs are much more stable than methylation state of individual cytosines and the frequency of polymorphisms on DMR level is comparable to the SNP frequency in genomic DNA (Becker et al., 2011; Schmitz et al., 2011). Therefore, DMRs are likely to be more informative than single-base methylation polymorphisms making DMR identification an important step in methylome analysis.

1.2 Dissertation Focus

There are three major factors that control gene expression: DNA sequence variation, epigenetic variation, and the environment. DNA sequence variation includes structural variation and single nucleotide polymorphisms, both of which can influence gene expression quantitatively or qualitatively. Epigenetic variation such as histone modifications and DNA methylation is responsible for heritable effects on gene expression that cannot be fully explained by DNA sequence variation. Finally, the environmental factors such as temperature, soil salinity, and the presence of pathogens can modify expression levels of many genes simultaneously. While all of these factors contribute to gene expression variation, it is gene expression that ultimately determines the phenotype and the knowledge of phenotype controlling mechanisms will help us improve domesticated species.

Rapid development of DNA sequencing, expression profiling, and other molecular screening technologies has brought us to a point where we are able to generate large amounts of data faster than we can thoroughly analyze them. This dissertation focuses on efficient computational methods and pipelines for the analyses of large expression and DNA methylation data sets. We already mentioned the importance of maize as an agricultural crop for which the demand grows faster than its production yield. To reduce that gap, it is essential to have a better understanding of gene functions and molecular pathways responsible for agronomically important traits.

Using statistical and computational tools, we investigate gene relationships in expression data to identify the differences that have appeared during the process of evolution or domestication. In particular, we construct and compare gene co-expression networks that record similarity among gene expression profiles. Finding the differences between those networks is especially challenging when the underlying expression data sets are already dissimilar because of extraneous factors such as measurement noise or experimental design. We address this challenge by comparing the magnitude of changes to their null expectation. This approach allows us to identify the most likely candidate genes that have been targeted during domestication and improvement. The ultimate goal is to compile a small list of candidate genes for subsequent experimental testing.

In addition to gene co-expression, this dissertation covers certain aspects of epigenetic variation analysis. Lately, epigenetic variation has drawn considerable attention from the scientific community because it may partially explain the 'missing heritability', i.e. heritable variation that appears independent from any DNA sequence variation. DNA methylation is an epigenetic mark extensively used by plants for gene silencing, guarding their DNA from transposable elements and reducing transcriptional noise. We develop a pipeline for the identification of differentially methylated regions in maize and analyze the relationship between the methylation in those regions and the expression levels of neighboring genes.

1.3 Dissertation Organization

We begin the dissertation in Chapter 2 with the overview of co-expression network analysis and how it was applied to identify global differences between maize and teosinte co-expression networks. In Chapter 3, the focus switches to gene-level differences that are often measured by Expression Conservation (EC) score. We introduce a novel method, Altered Expression Conservation (AEC), that makes expression conservation analysis more sensitive to genes with less extreme signatures of alteration. Chapter 4 concentrates on the generalization of the AEC method. It discusses the effects of sample size and evolutionary distance on expression conservation analysis and describes an alternative application of the method to genetic interaction

networks in yeast. Chapter 5 compares maize co-expression networks derived individually from two different expression profiling platforms. It also explores the possibility to combine the heterogeneous data from those platforms while constructing a co-expression network. Chapter 6 investigates DNA methylation and its influence on gene expression, presents a pipeline for the identification of differentially methylated regions (DMRs) and examines relationships between DMRs and local sequence variation. Finally, Chapter 7 provides conclusions from this work and outlines future research directions.

Chapter 2

Global Analysis of Co-expression Networks

2.1 Chapter Overview

The central "dogma" of molecular biology states that information flows from DNA via RNA into proteins. While many notable exceptions to this "dogma" as well as additional mechanisms of information control have been uncovered (Shapiro, 2009), this information flow is usually a starting point for most types of molecular analysis. Due to the rapid development of high-throughput technologies, the analysis of the information flow previously limited to several genes at a time has expanded to the genome level (Henikoff, 2002) whereby whole genomes and transcriptomes are routinely constructed and compared to infer biological associations. At the same time, expression analysis evolved from simple identification of differential expression between two samples to include more intricate co-expression relationships that capture even more biologically relevant information.

We begin this dissertation with a review of the methods for co-expression network analysis, which has become an important basis for research on gene expression. We explain how co-expression networks can be constructed and compared with each other. Finally, we describe how we used co-expression network analysis to uncover transcriptome rewiring that occurred during the domestication of maize.

Portions of the research appearing in this chapter were published in Swanson-Wagner et al. (2012). The work described in this chapter includes contributions from Ruth Swanson-Wagner, Robert Schaefer, Matthew Hufford, Jeffrey Ross-Ibarra, Chad Myers, Peter Tiffin, and Nathan Springer. Ruth performed all the wet lab work, processed the raw microarray data, and participated in the interpretation of the results. Peter conducted the differential expression analysis. Robert contributed analytical tools for the network visualization. Matthew and Jeffrey offered the list of genes from regions affected by domestication according to the independent sequence-based analysis and assisted with interpretation of the results. Chad, Peter, and Nathan designed and supervised the project.

2.2 Review of Co-expression Network Analysis

Recent technological advances have made genome sequencing considerably faster and cheaper. However, a large portion of predicted genes remain uncharacterized and their functional relationships are yet to be discovered. Many network-based prediction methods have been developed to establish gene functions (reviewd by Sharan et al., 2007) including those relying on metabolic networks (Christian et al., 2009; Henry et al., 2010), genetic interactions (Wong et al., 2004; Costanzo et al., 2010), protein-protein interactions (Deng et al., 2004; Tsuda et al., 2005), and co-expression networks (Carter et al., 2004; Zhang and Horvath, 2005) among many others. Due to the prevalence of expression data, methods based on co-expression networks are particularly appealing. Many formal methods have been developed for the analysis of co-expression networks. For instance, Huttenhower et al. (2006) designed a scalable framework that combines co-expression networks from multiple microarray experiments using a Bayesian approach. The method was then successfully applied to predict functional relationships in *Saccharomyces cerevisiae*.

Co-expression networks also represent major building blocks for functional networks that combine them with a multitude of other heterogeneous data such as protein-protein interactions, genetic interactions, and multiprotein complexes (Huttenhower et al., 2009; Lee et al., 2010). In particular, Lee et al. (2010) employed

orthology to incorporate yeast, fly, worm, and human co-expression data into *Arabidopsis thaliana* functional network. According to the study, the data from other distantly related species has adequate predictive power even when it is used in isolation. Adding such data to the composite network improved the overall accuracy (Lee et al., 2010).

Co-expression networks play crucial role in the studies of comparative genomics. For example, using a probabilistic method based on order statistics, Stuart et al. (2003) ranked genes in co-expression networks to find conserved genetic modules in human, fly, worm, and yeast, while Jordan et al. (2004) explored co-expression networks to analyze conservation and co-evolution of human genes. Several comparative genomics methods have been devised to measure the levels of expression conservation and divergence between two species (e.g. Ihmels et al., 2005; Zhang and Horvath, 2005; Dutilh et al., 2006; Tirosh and Barkai, 2007; Essien et al., 2008).

While the construction of co-expression networks is straightforward, it becomes computationally intensive for bigger genomes especially when a large number of samples are involved. For example, an unthresholded network encompassing all 32,540 genes from maize filtered gene set v4a.53 (Schnable et al., 2009) will define over half a billion edges prompting special computational and storage approaches such as preliminary clustering of the data to reduce the number of computed edges (Zhang and Horvath, 2005) and a storage format with memory mapping to improve data access speeds (Huttenhower et al., 2009). In this section, we review various approaches available for co-expression network construction. We also cover the methods for co-expression network comparison as well as techniques for visualizing the differences between co-expression networks at the global level.

2.2.1 Co-expression Network Construction

When analyzing genome-wide expression data, it is often beneficial to group genes based on their expression patterns, also termed expression profiles (Eisen et al., 1998). Strictly speaking, an expression profile of a gene is a vector that contains the gene's expression levels measured in different tissues, under varying environmental conditions, or in different genotypes. Diverse measures have been used to calculate similarity between expression profiles such as Euclidean distance (Tornow and

Mewes, 2003), mutual information (Butte and Kohane, 2000), and Pearson correlation coefficient (Eisen et al., 1998). Similarities between all possible pairs in a set of genes can be combined into a symmetric matrix called a co-expression network. In such a network, nodes represent genes while similarity scores indicate edge weights.

When building co-expression networks, microarray data is usually *log*-transformed to reduce the impact of large values on correlation coefficients. Since the dynamic range of RNA-seq platform is much wider, the inverse hyperbolic sine (asinh) transformation (Burbidge et al., 1988) may be more appropriate for RNA-seq data sets. The *asinh* transformation compresses large values considerably more than the small ones and unlike *log* transformation, it can be applied directly to values below 1.

Suppose we have an expression data set that contains measurements for M genes over N experimental conditions. The data forms an $M \times N$ matrix E where each value E_{gc} denotes the expression level of gene g under experimental condition c. Thus, each row E_g represents the expression profile of gene g. To build a co-expression network represented by a matrix R, the Pearson correlation coefficient is calculated between each pair of gene expression profiles, i.e.

$$R_{ij} = PCC(E_i, E_j)$$

for $i, j = 1..M$ and $i \neq j$. Thus, R_{ij} is a square matrix $M \times M$ where each value represents an edge weight in the co-expression network and measures similarity between expression profiles of two genes.

If we have two data sets E^a and E^b encompassing the same gene set but under different conditions, direct comparison between them may be difficult when the conditions are not paired. In particular, the dimensions of E^a and E^b may be different. However, since E^b contains the same M genes as E^a, the derived co-expression network R^b will have the same dimensions as R^a. Yet, the distribution of values in R^a and R^b may be different depending on various factors such as unequal number of samples in each case. Hence, a value from R^a cannot be compared directly to the corresponding value in R^b. To enable the direct comparison, (Huttenhower et al., 2006) recommends applying Fisher's z-transformation and normalization to both matrices R^a and R^b. For each element r in R^a or R^b, the Fisher transformation is defined

as

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

The transformation guarantees that the distribution of z values is approximately normal and the variance is approximately constant for all values of r. By subtracting the mean and dividing by the standard deviation, the normalized distribution will have a mean of 0 and a standard deviation of 1, i.e. the values in the transformed matrix have a normal distribution $N(0,1)$. Meanwhile, a single value shows how many standard deviations the edge weight is from the mean making cross-network comparison possible.

Some studies advocate transforming similarity matrix R into a binary adjacency matrix A by applying a hard threshold (Butte and Kohane, 2000; Carter et al., 2004) or selecting *K* nearest neighbors (Tornow and Mewes, 2003). However, hard thresholding generally causes the loss of information and sensitivity (Carter et al., 2004). Under the WGCNA framework developed by Zhang and Horvath (2005) and later implemented as an R package by Langfelder and Horvath (2008), the authors suggested to apply soft thresholding while keeping the resulting edge weights. They proposed two thresholding functions, namely the sigmoid function

$$A_{ij} = \frac{1}{1 + exp(-\alpha(r_{ij} - \gamma))}$$

with parameters α and γ and the power adjacency function

$$A_{ij} = |r_{ij}|^{\beta}$$

with parameter β. Both functions yield similar results when the parameters are tuned using the same criteria that relies on the goodness of fit to the scale free network model Zhang and Horvath (2005). Even though such soft thresholding drives some edge weights very close to 0, a certain hard threshold is still necessary to determine the neighbors. However, the framework is designed to avoid hard thresholding in all other types of the network analysis.

The WGCNA framework was successfully used for expression analysis in different contexts such as identification of molecular targets for cancer treatments (Horvath et al., 2006), investigation of gene essentiality and network modularity in yeast (Carlson et al., 2006), and finding associations between expression and phenotype in plants (Weston et al., 2008). The R package, however, never generates a complete co-expression network due to computational difficulties. The main output of the pipeline is a set of modules that include only highly-connected genes. If a data set contains a large number of genes, for instance more than 6000, the application splits it into smaller subsets via hierarchical clustering, finds gene modules in each subset, and merges the modules based on a certain similarity criteria (Langfelder and Horvath, 2008). Thus, the R package does not allow for the direct comparison of unthresholded co-expression networks at the global level and it does not consider potentially important expression profile comparisons between genes from different subsets.

2.2.2 Co-expression Network Comparison

Unlike WGCNA, our approach involves calculation of all pair-wise correlations between full expression profiles as we want to incorporate all available information. Since the values in the network will have a standard normal distribution after Fisher transformation and normalization, we can compare a network R^a to another network R^b that has the same set of genes using the Pearson correlation coefficient,

$$\rho_{R^a R^b} = \sum_{i=1}^{M} \sum_{j=i+1}^{M} \frac{(r_{ij}^a - \mu_{R^a})(r_{ij}^b - \mu_{R^b})}{\sigma_{R^a} \sigma_{R^b}}$$

where M is the number of genes in each network, r_{ij}^a and r_{ij}^b are j-th elements in i-th rows of the co-expression networks R^a and R^b respectively. The mean and standard deviation for each network are calculated based on the values in the upper triangle, i.e. all values r_{ij} for $i = 1, .., M$ and $j = i + 1, .., M$.

We can test the significance of the obtained correlation coefficient via bootstrapping. For similarity analysis, each round of bootstrapping would entail the permutation of gene labels. The obtained random networks will likely appear more different

than the actual networks because each comparison between co-expression profiles will be performed on two random genes. When analyzing the differences, the goal of bootstrapping is to generate random networks that are more similar to each other than the actual networks. Suppose our co-expression networks R^a and R^b were derived from the expression data sets E^a ($N \times M^a$) and E^b ($N \times M^b$) respectively. We can combine E^a and E^b into a single data set E ($N \times M^a + M^b$) and sample the conditions (columns) randomly without replacement to produce two data sets of the same size as the actual ones, i.e. S^1 ($N \times M^a$) and S^2 ($N \times M^b$). Each data set S^1 and S^2 will contain a random mix of columns from both E^a and E^b. Since the columns in E^a and E^b are likely to be more similar within each data set than between the data sets, S^1 and S^2 are likely to be more similar than E^a and E^b. Any variation between S^1 and S^2 would exist mainly due to randomness and the difference in the number of conditions. Using the distribution of correlation coefficients between the random networks, we can derive an empirical p-value for the correlation coefficient between R^a and R^b.

The differences between two co-expression networks can be visualized through the joint edge weight distribution. Two-dimensional binning of the edge weight pairs is performed in such a way that each bin provides a count of edges whose weight falls within a certain range in one network and a potentially different range in the other (see Figure 2.6a for an example.) Since some of these counts could be very high, it may be necessary to apply a log transformation, so that bins with small values can be discerned from empty bins on a graph. The significance of this divergence can be examined on a differential joint edge weight distribution plot that exhibits the difference between the actual joint edge weight distribution and its null expectation derived from bootstrapping by averaging the corresponding counts. If we represent the difference as a ratio of the actual count to its null expectation, after log transformation the sign would indicate whether the actual count is smaller than expected by chance. Since both actual and expected counts may be 0, instead of the regular logarithmic function it is necessary to use

$$f(x) = \frac{\log(\theta x + 1)}{\theta}$$

where θ is often set to 1. To achieve less compression of the counts, one can also use other values of $\theta < 1$.

2.3 Using Co-expression Networks for Maize Domestication Analysis

In Chapter 1, we introduced the topic of maize domestication from its wild ancestor teosinte. While the domestication process introduces certain genomic changes, domesticated varieties generally belong to the same species and possess considerable genomic similarity to their wild progenitors. Comparative hybridization analysis confirmed genomic similarity between maize and teosinte despite their remarkable phenotypic variations (Swanson-Wagner et al., 2010). In the case of domestication, differences convey more information about genome evolution than conservation. Characterization of the changes that conditioned those phenotypic differences will elucidate the genetic architecture of complex traits (Doebley et al., 2006) and response to selection (Purugganan and Fuller, 2009) while contributing the resources for maize breeding efforts (Gross and Olsen, 2010).

Previous studies used various techniques such as the analysis of microsatellite diversity (Vigouroux et al., 2002), selective molecular genetic scans (Yamasaki et al., 2005; Wright et al., 2005), and full genome sequencing (Hufford et al., 2012) to uncover a multitude of genomic regions targeted during maize domestication. Several genes such as *tb1* (Doebley et al., 1997), *ba1* (Gallavotti et al., 2004), and *zfl2* (Bomblies and Doebley, 2006), appeared to have a direct influence on the phenotypic differences between maize and teosinte. Other studies have identified genes with putative regulatory effects contributing to the variation (Doebley et al., 2006; Yamasaki et al., 2007). Recently, Hufford et al. (2012) also found regions that experienced selective pressure during domestication yet lacked any annotated genes. These findings support the hypothesis that changes in gene regulation contribute significantly to the evolution of maize (Doebley et al., 2006; Zhao et al., 2008).

To explore gene expression changes conditioned by maize domestication, we analyzed the transcriptomes of 38 maize and 24 teosinte lines. We engaged co-expression network analysis to detect rewiring of the regulatory relationships in the

transcriptome due to domestication. We show that the differences between co-expression networks of maize and teosinte are biologically significant and that co-expression analysis represents an effective tool for finding regulatory differences in closely related species.

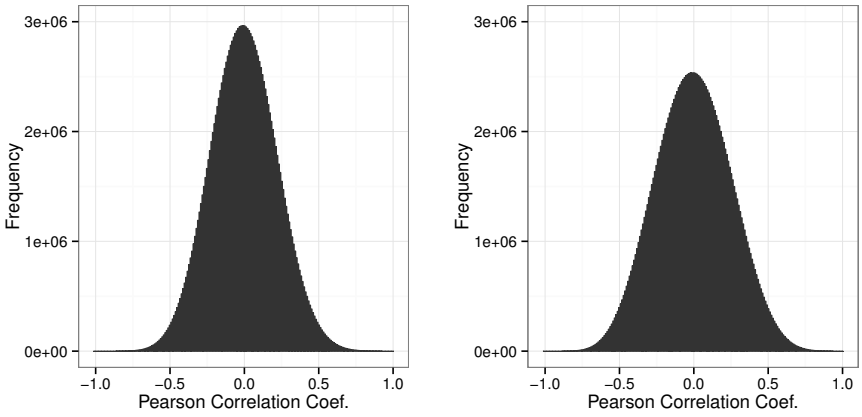
2.3.1 Maize Expression Data and Co-expression Networks

We used a NimbleGen (Roche NimbleGen) microarray containing probes for 32,540 genes annotated in the reference *Zea mays* genome version 4a.53 (Schnable et al., 2009) to profile 38 diverse maize inbred lines, 7 teosinte inbred lines, and 17 teosinte individuals sampled from wild-collected, outcrossing populations. All samples were collected from 8-day old seedlings to reduce the expression changes due to developmental differences among the lines. Probe signal intensities were spatially corrected and processed with NimbleScan to produce robust multichip average (RMA) normalized (Irizarry et al., 2003) gene expression values. Whenever possible, the gene expression values were averaged across technical and biological replicates of the same genotype. Genes with detectable signal significantly higher than the background noise in fewer than three genotypes were excluded from further consideration leaving the data set with 19,792 genes.

Polymorphisms in genomic sequences of non-reference varieties may affect the hybridization of the corresponding cDNA. The effect is independent from the variation in gene expression. To avoid this hybridization bias, we removed 26,937 probes with poor genomic hybridization in at least three genotypes based on the results from the Comparative Genomic Hybridization (CGH) analysis by Swanson-Wagner et al. (2010). The raw data from the remaining 46,167 probes were RMA-normalized again to generate the final expression data set with 18,242 genes (1-4 probes per gene) that was used for subsequent analyses.

We separated the expression data into maize ($18,242 \times 38$) and teosinte ($18,242 \times 24$) data sets. By calculating Pearson correlation coefficient between each pair of genes, we constructed two co-expression networks each represented by a $18,242 \times 18,242$ matrix. Before any transformation and normalization, the distribution of correlation coefficients proved to be different between the networks (Figure 2.1.) While mean values were around zero in both cases ($\mu_{maize} = 0.0012$ and $\mu_{teosinte} = 0.0022$),

standard deviation was higher in the teosinte network ($\sigma_{maize} = 0.2210$ and $\sigma_{teosinte} = 0.2508$). To enable direct comparison between the networks, we applied Fisher transformation and normalization to the edge weights Huttenhower et al. (2006) as described in Section 2.2.



(a) Maize **(b)** Teosinte

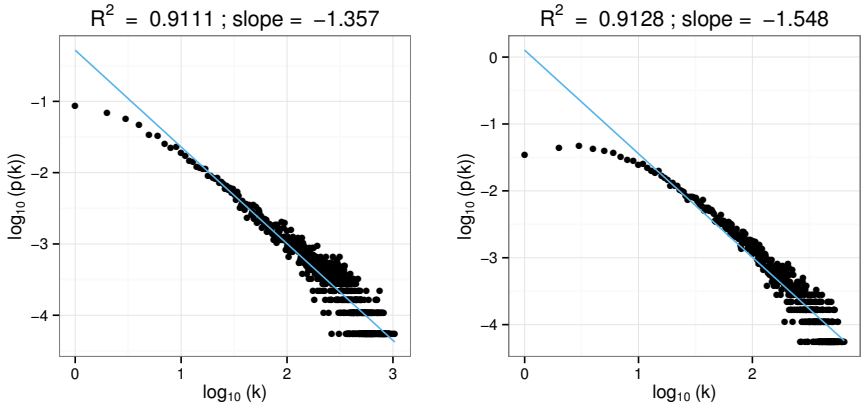
Figure 2.1. Distribution of Pearson correlation coefficients in maize and teosinte co-expression networks. Before Fisher transformation and normalization, (a) maize and (b) teosinte co-expression networks had different edge weight distributions ($\mu_{maize} = 0.0012$ and $\sigma_{maize} = 0.2210$ vs. $\mu_{teosinte} = 0.0022$ and $\sigma_{teosinte} = 0.2508$). The difference could be due to unequal number of samples in each expression data set or higher variability among teosinte lines.

2.3.2 Topology Comparison

Most biological networks have been found to be approximately scale free (Barabási and Oltvai, 2004). According to this property, the connectivity (degree) k of each node approximately follows power law distribution, i.e. connectivity frequency $p(k)$ is proportional to $k^{-\gamma}$ (Barabási and Albert, 1999). To determine whether a network is scale-free, Zhang and Horvath (2005) suggested using the squared correlation coefficient R^2 between $log(p(k))$ and $log(k)$. Whenever $R^2 > 0.8$, there exists a linear relationship between $log(p(k))$ and $log(k)$, which is expected when $p(k) \sim k^{-\gamma}$. In addition, the authors recommend that the slope of the fitted linear model should be $a \approx -1$.

Network	Threshold	R^2	Slope
Maize	2.5	0.8593	-1.2416
Maize	3.0	0.9111	-1.3568
Maize	3.5	0.9162	-1.3908
Maize	4.0	0.9003	-1.4275
Teosinte	2.5	0.7265	-1.2396
Teosinte	3.0	0.9128	-1.5477
Teosinte	3.5	0.9281	-1.6285
Teosinte	4.0	0.9357	-1.6446

Table 2.1. Scale-free topology criteria at different thresholds in maize and teosinte co-expression networks. In a scale-free network, the squared correlation coefficient between log-transformed connectivity and connectivity frequency should be close to 1 while the slope of the linear regression line should be around -1. Based on these two criteria, the threshold $r \geq 3.0$ seems reasonable. While higher threshold would yield higher squared correlation coefficient for teosinte network, it would also increase the slope and make the adjacency matrix much sparser.



(a) Maize **(b)** Teosinte

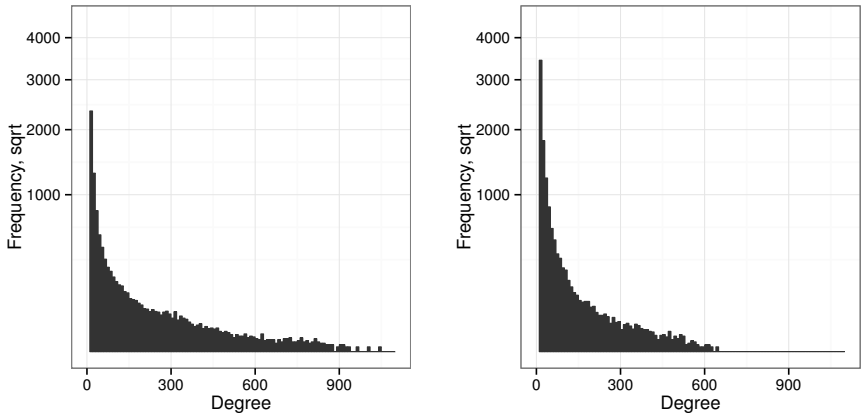
Figure 2.2. Scale-free topology in maize and teosinte co-expression networks. With the hard strict threshold $r \geq 3.0$, both maize and teosinte co-expression networks are approximately scale free. The relationship between $log(p(k))$ and $log(k)$ is almost linear with squared correlation above 0.8 and the slope of the regression line is close to -1.

To determine the connectivity, we applied a strict threshold ($r \geq 3.0$) and counted the remaining edges for each node. According to the criteria recommended by Zhang

and Horvath (2005), both networks possess approximately scale-free topology (Table 2.1; Figure 2.2.) However, at this threshold maize network had distinctly more high degree nodes than teosinte (Figures 2.3a and 2.3b). It indicates that on average, gene co-expression profiles are less similar to each other in the teosinte network. To check whether these differences have a biological basis, we compared the actual distributions to the null expectation derived from the bootstrapping analysis ($N = 1000$, see Section 2.2 for details.) We reasoned that random distribution of genotypes across the subsets would obscure biological signals. Therefore, the variation between the bootstrapped networks would be mainly due to the group size difference. The null expectation for the degree distribution indeed exhibits the same tendency (Figures 2.3c and 2.3d) as observed in the actual co-expression networks (Figures 2.3a and 2.3b), i.e. the networks based on the large groups on average have more hubs than the networks based on the small groups.

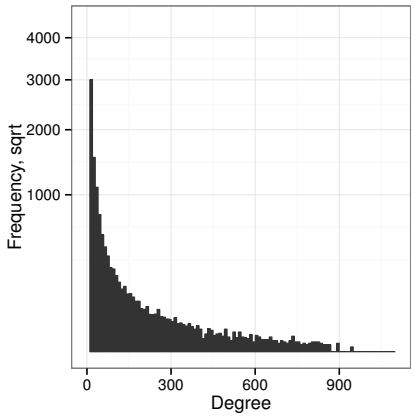
Although the degree distribution in maize and teosinte resembles the null expectation, hubs in one network may not correspond to the hubs in the other. To check this possibility, we calculated the correlation between gene degree in maize and teosinte networks and found it to be lower ($\rho = 0.4976$, Figure 2.4a) than that between typical random networks ($\rho \sim 0.6500$, Figure 2.4b). In other words, a gene may be a hub in one network but possess low degree in the other. For example, alpha-6-tubuline *tua6* (GRMZM2G083243) represents a hub in the maize network as it has 732 edges with the weight of at least 3.0. However, in the tesointe network that gene has only 41 edges. On the contrary, receptor protein kinase *cr4* (GRMZM2G051637) possess 527 edges in teosinte but only 89 in maize. Since the difference cannot be explained by unequal group size, it is likely to have a biological basis.

For further validation, we analyzed the differences at the edge level. Overall, each co-expression network contains over 160 million undirected edges but due to strict thresholding the degree analysis only encompasses about 1 million and 0.85 million edges in maize and teosinte networks respectively. To evaluate all information encapsulated by the networks, we calculated Pearson correlation coefficient between all edge weights in one network and those in the other. While the global correlation appears quite low ($\rho = 0.3038$), it does not necessarily confirm the presence of biologically significant differences. The networks may contain excessive random noise

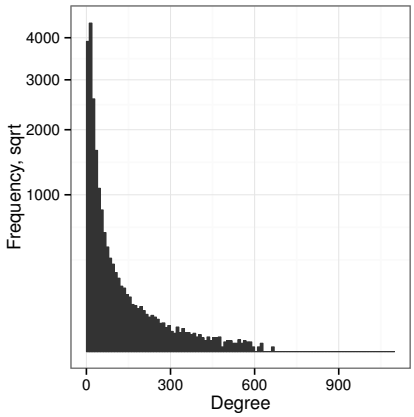


(a) Maize

(b) Teosinte



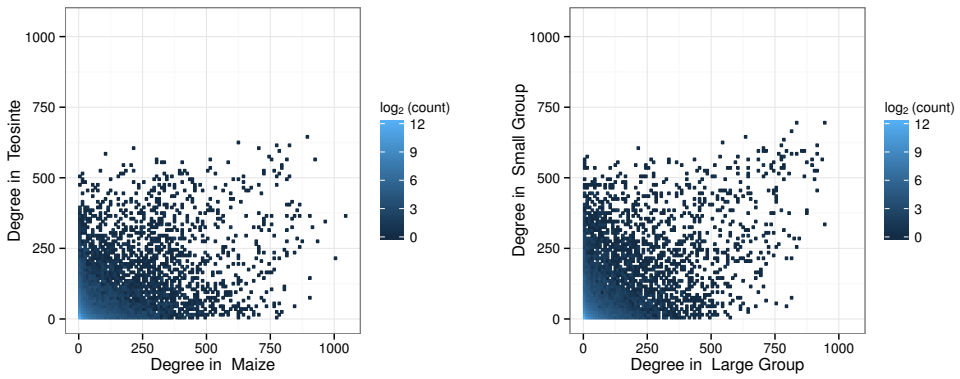
(c) Large Group



(d) Small Group

Figure 2.3. Gene degree distribution at a strict threshold ($r \geq 3.0$) in (a) maize, (b) teosinte, (c) - (d) random co-expression networks with 38 and 24 lines respectively. The distributions for the random networks were averaged across 1000 rounds of bootstrapping.

that can considerably reduce the correlation. To determine that the result is significantly low, we calculated global correlation between co-expression networks derived from bootstrapped expression data ($N = 1000$, see Section 2.2 for details.) The majority (98.6%) had correlation coefficient above 0.3038 (Figure 2.5). Therefore, the network differences are not random and are suggestive of gene rewiring in maize during domestication and improvement.



(a) Maize vs Teosinte **(b)** Typical Random Networks

Figure 2.4. Degree correlation in maize and teosinte co-expression networks. (a) The correlation between node degree in maize and teosinte is relatively low ($\rho = 0.4976$). There are many genes that are hubs in one network but have low degree in the other. (b) Typical pair of random networks ($\rho = 0.6738$). In general, random networks have higher degree correlation ($\rho \sim 0.6500$) than the actual networks.

2.3.3 Differences in Individual Co-expression Relationships

We observed the global lack of similarity between the maize and teosinte co-expression networks. To discover the edges responsible for this result, we explored the joint distribution of their weights. The distribution reveals that moderate to high conservation is very common because most edge weight pairs concentrate near the diagonal (Figure 2.6a). Nevertheless, it is clear that the outliers are present as well.

Since random noise might explain some of the shifts from the diagonal, we compared the actual joint distribution of the edge weights to its null expectation derived from the bootstrapping analysis (Figure 2.6b). The number of conserved edges with low weights matches the expectation as indicated by a large black region in the center of the map. However, low density areas (blue) along the diagonal farther from the center point to decreased conservation of edges with high weights. In other words, genes with highly correlated or anti-correlated expression profiles get modified more often than expected by chance. Moreover, the large high density regions (red) lie predominantly in the II and IV quadrants. Hence, the changes usually reverse the relationship, i.e. correlated profiles become anti-correlated and vice versa. Overall, the joint distribution of maize and teosinte edge weights is substantially different

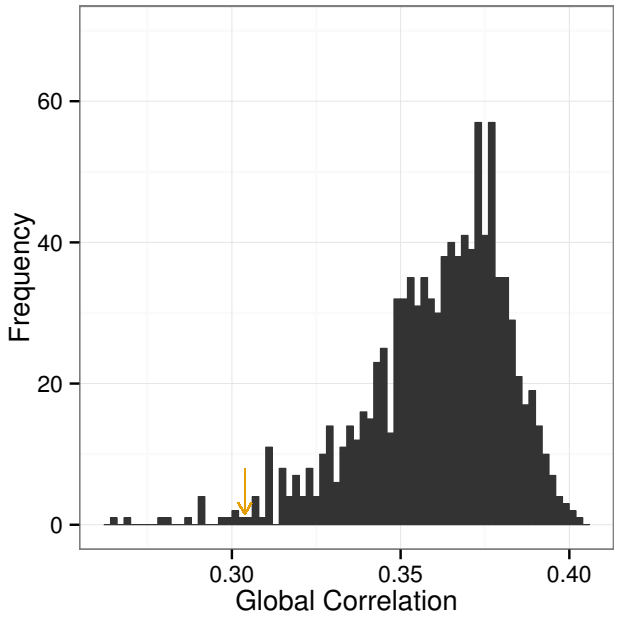
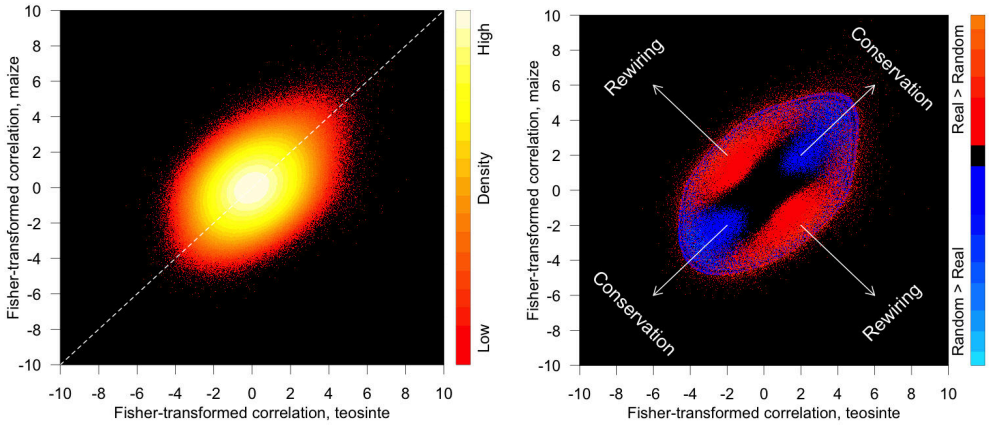


Figure 2.5. Global correlation between maize and teosinte co-expression networks. Distribution of global correlation coefficients is derived from co-expression networks built on 1000 pairs of random networks. Orange arrow indicates the global correlation coefficient between the actual maize and teosinte co-expression networks ($\rho = 0.3038$). Only 14 random network pairs have lower correlation.

from its null expectation, which points to the presence of gene rewiring.

2.4 Conclusions

Co-expression networks play important role in the transcriptome-wide studies of gene expression. We reviewed the major steps of co-expression network analysis including network construction, thresholding, and comparison. Despite the recent technological advances, co-expression network analysis remains a moderately hard computational task that requires creative approaches to algorithm design and data storage. We employed co-expression networks to study the effects of domestication and improvement on maize transcriptome. We used bootstrapping analysis to show that the co-expression networks of maize and its wild ancestor teosinte had biologically significant differences. This result is consistent with the hypothesis that



(a) Joint Edge Weight Distribution

(b) Differential Joint Edge Weight Distribution

Figure 2.6. Joint edge weight distribution in maize and teosinte co-expression networks. (a) Heatmap built for the actual maize and teosinte co-expression networks. X and Y axes display edge weights in teosinte and maize networks respectively. Color scale indicates how many corresponding edges have a certain combination of weights. If maize and teosinte co-expression networks were almost identical, the colored shape would have been a thin oval stretched along the diagonal. (b) Differential heatmap exhibits the difference between the actual joint distribution of edge weights and its null expectation. Blue regions on the diagonal and red regions away from the diagonal indicate that gene rewiring takes place.

regulatory changes substantially influence the evolution of maize.

Chapter 3

Expression Analysis of Individual Genes

3.1 Chapter Overview

Chapter 2 provided a survey of methods for co-expression network construction and comparison. The global differences between two networks are driven by changes in co-expression patterns of individual genes. This chapter presents an expression conservation (EC) score that can be used to specifically identify rewired genes. Previous studies relied on a hard threshold for EC score to select the genes of interest. However, in some cases even a high EC score may be significantly lower than expected by chance. Therefore, to detect rewired genes we developed a new approach that compares EC score to the gene's null distribution. Using the maize domestication data introduced in Chapter 2, we show that our method is more sensitive than the standard approach.

Portions of the research appearing in this chapter were published in Swanson-Wagner et al. (2012). This work includes the contributions from Ruth Swanson-Wagner, Robert Schaefer, Matthew Hufford, Jeffrey Ross-Ibarra, Chad Myers, Peter Tiffin, and Nathan Springer. Ruth performed all the wet lab work, processed the raw microarray data, and participated in the result interpretation. Peter conducted the differential expression analysis. Robert contributed analytical tools for the network

visualization. Matthew and Jeffrey offered the list of genes from regions affected by domestication according to the independent sequence-based analysis and assisted with interpretation of the results. Chad, Peter, and Nathan designed and supervised the project.

3.2 Gene Expression Analysis

Transcriptomes can be compared based on their similarities or differences. The former approach is generally used to elucidate functional characteristics of genes, while the latter aims to discover the effects of adaptation on the transcriptome (Tirosh and Barkai, 2007). To detect expression changes at the gene level, a variety of differential expression (DE) methods have been proposed for both microarrays (reviewed in Grant et al., 2007) and RNA-seq (reviewed in Garber et al., 2011). DE methods can identify genes whose expression levels statistically differ between two data sets but exhibit little variation within each data set. However, these methods can only be applied when the data sets have either matching or very similar set of conditions, albeit the data may come from different species. DE methods are also likely to miss the differences whenever a gene shows considerable variation within a data set. To complement the differential expression methods, techniques that analyze gene expression co-variation (or simply co-expression) have been developed and successfully applied across many species of bacteria, plants, and animals (inter alia Bergmann et al., 2003; Stuart et al., 2003; Ihmels et al., 2005; Oldham et al., 2006).

3.2.1 Expression Conservation Score

Under the WGCNA framework (Zhang and Horvath, 2005) discussed earlier in Chapter 2, both similarities and differences between species on gene level are determined based on either gene connectivity (Oldham et al., 2006) or module alignment (Ficklin and Feltus, 2011). The same approaches are used for comparing transcriptomes of a single species under different conditions (Horvath et al., 2006; Carlson et al., 2006; Weston et al., 2008). While the change in module membership is a good indicator of gene rewiring, it can only help in detecting genes with extreme changes in co-expression because the alterations have to be sufficiently large to remove the gene

from the respective module. Gene connectivity metrics are even less sensitive as rewiring may replace gene connections without changing their number.

The alternative framework (Ihmels et al., 2005; Dutilh et al., 2006; Tirosh and Barkai, 2007; Essien et al., 2008) relies on the comparison of gene's co-expression profiles. Usually, the similarity between the profiles is estimated by Pearson correlation coefficient. This value is known as the expression context conservation (Dutilh et al., 2006) or simply expression conservation (EC) score (Tirosh and Barkai, 2007). We will use the second term, EC score, throughout this chapter. If a gene has a high EC score, its expression profile is similar to the same set of expression profiles (neighbors) in both data sets. If EC score is low, the gene's expression profile has certain neighbors in one data set but different neighbors in the other. Because a co-expression profile describes the gene's relationship with its neighbors in terms of its expression pattern, genes with high EC score possess similar relationships with their neighbors in both organisms and, as a result, are likely to retain the same functionality.

Because co-expression networks are not thresholded before EC score calculation, all available information is leveraged to make inferences about the conservation or divergence of gene co-expression. Nevertheless, there are two notable exceptions. Since the difference between co-expression profiles is not necessarily caused by the difference in expression levels, Tirosh and Barkai (2007) argued that higher weight should be given to correlations between genes with conserved expression while calculating EC score. Therefore, the authors suggested an iterative approach in which the score is repeatedly adjusted by giving low weights to genes with low EC score and recalculating the weighted correlation coefficient until convergence. When comparing several co-expression network analysis methods, Wang et al. (2010) used only nodes from a conserved co-expression network to calculate an EC score for each gene. In other words, each gene's EC score depended only on the genes that had conserved co-expression profiles in two species. While this particular modification had only a minor effect on the overall distribution of EC scores, the results of the iterative algorithm developed by Tirosh and Barkai showed more substantial differences compared to the regular approach (Wang et al., 2010). Due to the absence of a gold standard or a specific comparison metric in the study by Wang et al., it is hard to

determine whether any of the algorithms work better than the others.

3.2.2 Measuring EC Score Significance

To find significant genes with either conserved or diverged expression, the existing co-expression studies have predominantly relied upon the raw EC scores (Dutilh et al., 2006; Tirosh and Barkai, 2007; Essien et al., 2008; Guan et al., 2013). The approach appears reasonable for discovering genes with conserved expression because many factors such as the background noise or minor differences in experimental conditions would only amplify the biologically significant variation. These factors are very unlikely to make a gene's co-expression profiles more similar. For the same reason, the raw score is not appropriate for the identification of genes with diverged co-expression as the extent of biologically conditioned variation would not be clear. Overall, this approach would only identify genes with extreme changes in co-expression relationships.

One alternative would be to evaluate each gene's significance by comparing its EC score to a global null expectation. The null distribution can be derived, for instance, from bootstrapping by averaging each gene's EC score across all bootstraps. Using the distribution of these averaged EC scores, one can select rewired genes based on an empirical p-value or z-score. The former is calculated as a ratio between the number of averaged EC scores from the bootstrapping analysis smaller than the gene's EC score to the total number of bootstrap samples. The latter is simply $z = (EC - \mu)/\sigma$ where μ and σ are the mean and standard deviation of the null distribution.

This alternative approach would still be biased towards outliers because the amplitude of possible expression changes may vary from gene to gene. To address the possible variations in amplitude, we propose to identify divergent genes by comparing each gene's EC score to its own null expectation. The parameters of each gene's null distribution can be derived from the bootstrapping analysis based on all available bootstrapped EC scores for that gene. Genes with low z-scores are likely to be rewired even if their actual EC scores are relatively high. Thus, our selection is no longer limited to genes with extreme EC scores. In addition, a negative EC score does not necessarily indicate that the gene is rewired if the score falls within the gene's

null distribution. Our method allows us to filter out such genes potentially increasing the specificity of the approach as well.

3.2.3 Method Validation and Comparison

Since co-expression can sometimes reflect indirect biological relationships, the results of EC studies can be difficult to validate directly with experiments. Previous studies engaged various indirect approaches to verify their results. In particular, Tirosh and Barkai (2007) showed that the genes with high EC score were enriched for essential genes, which they offered as a proof of their method's validity. Essien et al. (2008) contrasted the distribution of interspecies and intraspecies EC scores to make sure that interspecies scores are generally lower. In the absence of gold standards, method comparison also presents a challenge. While comparing computational models that assess expression conservation, Wang et al. (2010) simply reported the extent of agreement among the models. While informative, the results of agreement among the models are generally insufficient for the identification of the most successful method.

In our maize domestication study (Swanson-Wagner et al., 2012), we compared the different EC methods by calculating the enrichment of the identified rewired genes for domestication and improvement genes uncovered by an independent sequence-based study. This is also an indirect validation approach because sequence-based and expression-based methods are rather complementary. However, certain overlap between the studies is expected because some DNA alterations lead to gene expression changes and, consequently, to co-expression modifications. We will discuss the details of this comparison in the next section.

3.3 Variations of Gene Expression between Teosinte and Maize

In Chapter 2 we identified biologically significant differences between maize and teosinte co-expression networks. The global correlation between the two networks is lower than expected by chance (Figure 2.5) and the changes can be clearly seen

on the differential joint edge weight distribution plot (Figure 2.6). Here, we investigate what drives these changes at the gene level, as our ultimate goal is to discover the genes whose expression has been affected by domestication and subsequent improvement process.

3.3.1 Using EC Score to Find Rewired Genes

To measure EC between maize and teosinte, we took the standard approach of calculating Pearson correlation coefficient between co-expression profiles of each gene in the maize and teosinte networks. Thus, our approach is most similar to Dutilh et al. (2006) except that we apply the Fisher transformation and normalization to the edge weights in co-expression networks before computing the correlation as recommended by Huttenhower et al. (2006). By averaging bootstrapped EC scores of each gene, we derived a null distribution and discovered that the actual EC score distribution is quite different from what is expected by chance (Figure 3.1). It has lower mean and higher standard deviation ($\mu = 0.2603$ and $\sigma = 0.1986$) than the null expectation ($\mu_{null} = 0.3138$ and $\sigma_{null} = 0.1815$). The shift of the actual distribution to the left compared to the null expectation and the existence of the left tail suggest the presence of rewired genes.

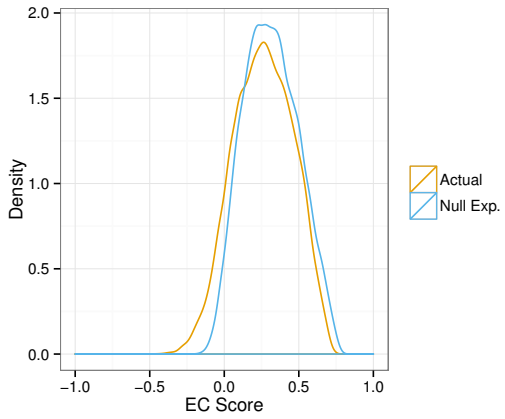


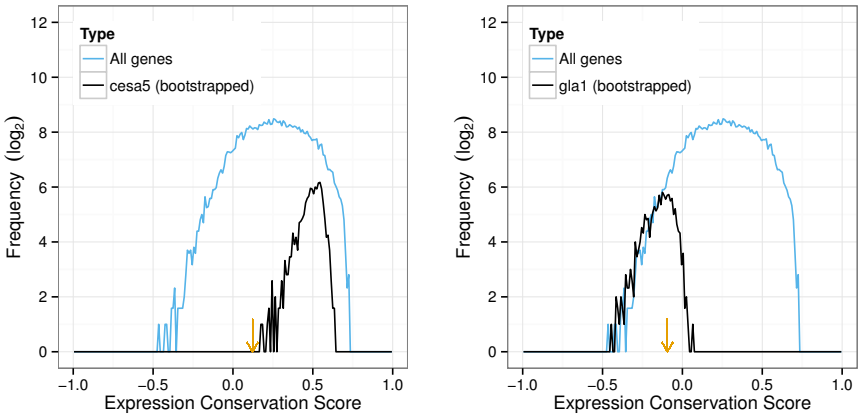
Figure 3.1. Distribution of EC scores between teosinte and maize. The orange line depicts the distribution of the actual EC scores while the blue line shows the null expectation derived from 1000 rounds of bootstrapping. Overall, EC scores between teosinte and maize are lower than expected by chance.

Previous studies either applied a hard threshold to EC scores in order to select the genes of interest or relied on the raw EC score to make inferences about gene rewiring (Dutilh et al., 2006; Tirosh and Barkai, 2007; Essien et al., 2008; Guan et al., 2013). As we are looking for genes rewired by domestication, it would be similar to selecting the left tail of the EC score distribution. The threshold can be chosen arbitrarily (e.g. $EC < 0$) or based on a global null expectation (the region to the left of the red curve on Figure 3.1.) Either approach would likely miss genes with relatively high EC scores that are still below their respective null expectations. For example, the EC score for *cesa5* (cellulose synthase) is greater than 0 but appears outside of the gene's null distribution (Figure 3.2a.) To capture such genes, we derived a z-score from each gene's EC score null distribution. A strict threshold of $z \leq -3.0$ yielded a list of 1,115 rewired genes. Henceforth, we will refer to this method as Altered Expression Conservation (AEC). Our approach is more sensitive as it includes the genes with relatively high EC score like *cesa5*. Yet, it is also more specific because it ignores the genes with low EC score when such a score is likely to occur by chance. For instance, the EC score for *gla1* (dihydrofolate synthase) is low (-0.0936) but the negative score is expected by chance and, therefore, we did not include *gla1* in our list.

3.3.2 Enrichment in Domestication and Improvement Genes Identified by Sequence-Based Analysis

Several studies analyzed large gene groups for genetic diversity present in teosinte and maize in order to uncover the genes targeted during domestication and improvement (Vigouroux et al., 2002; Wright et al., 2005; Yamasaki et al., 2005; Briggs et al., 2007). Since the evidence points to a single domestication event (Doebley, 2004; Matsuoka et al., 2002; Piperno et al., 2009), these genes are often categorized into two overlapping groups based on the time of modification. Genes that diverged during the initial domestication event fall into domestication category, while the genes modified after the domestication event belong to the improvement category.

More recently, Hufford et al. (2012) screened the whole genome for genetic diversity in 75 lines of maize and its wild ancestors including teosinte. The authors used a likelihood method to identify 484 and 695 genomic loci with extreme allele frequency



(a) *cesa5* **(b)** *gla1*

Figure 3.2. Null distribution of EC scores for two genes in maize and teosinte. The blue curves show the global distribution of EC scores. The black curve indicate the null distribution of EC scores for *cesa5* (left) and *gla1* (right) derived from the bootstrapping analysis ($N = 1000$). The actual EC score of the corresponding gene is indicated by the orange arrow. (a) The actual EC score for *cesa5* is relatively high but significantly lower than expected by chance for the gene. (b) The actual score for *gla1* is relatively low but not significantly different from what is expected by chance for the gene.

differentiation caused by selection pressure during domestication and improvement respectively. These domestication and improvement loci contained 1,764 and 1,506 genes, respectively, yielding 3,040 unique genes that were potentially the targets of selection.

Due to linkage disequilibrium effects, the identification of exact targets was problematic. The authors identified the most probable candidates by selecting the genes closest to the locus with the highest likelihood for selection during domestication or improvement. However, changes to a gene's DNA sequence do not necessarily trigger the changes in its expression. Moreover, the actual target of selection may be a regulatory region rather than a gene and this region may regulate either a nearby gene (*cis*-acting regulation) or a gene elsewhere in the genome (*trans*-acting regulation.) Finally, some of the regulatory changes may come from the variants linked to the target of selection, which complicates the analysis even further.

On the other hand, both sequence-based and expression-based methods can find genes with the direct impact on the differences between species. However,

sequence-based methods can also detect the upstream variations that have causal effects on the regulatory process while expression-based methods can also identify the targets downstream of causal changes. As such, the methods are complementary, but a certain overlap is still expected. For instance, a locus identified by a sequence-based method may contain a *cis*-acting regulatory region that significantly affects the expression of the nearby gene. If the regulatory region also has the highest likelihood for selection and the regulated gene is closer to the regulatory region than any other gene, both methods should be able to detect it. Thus, a statistically significant intersection of the results would indicate that both methods retrieve biologically relevant information.

We found that gene lists produced by our method are significantly ($p < 0.05$) enriched for genes in domestication or improvement regions identified by Hufford et al. (2012) at several z-score cutoffs (Table 3.1). The significant overlap with the results of an independent sequence-based method corroborates the robustness of our approach and confirms the potential of using co-expression analyses for studying evolution. Interestingly, the enrichment for domestication genes is stronger than that for improvement genes. In fact, the enrichment for the improvement genes is only significant at $z \leq -2.5$. It is possible that the domestication had a greater impact on the maize transcriptome than the gradual improvements which occurred afterwards. However, further analysis is needed to confirm this hypothesis.

z-score	Total	Domestication		Improvement		Dom. or Imp.	
Cutoff	Genes	Genes	p-value	Genes	p-value	Genes	p-value
-2.5	1841	120	0.0476	105	0.0185	213	0.0023
-3.0	1115	81	0.0106	64	0.0522	135	0.0031
-3.5	650	52	0.0068	36	0.1681	80	0.0138
-4.0	405	31	0.0514	23	0.1969	49	0.0582

Table 3.1. Enrichment of AEC genes for selection targets identified by sequence-based analysis (Hufford et al., 2012) at various z-score cutoffs. Overlap p-values are based on Fisher's exact test.

One can argue that genes with high EC scores should be excluded regardless of their z-score because high EC score implies the lack of co-expression changes. We selected genes with a combination of z-score threshold $z \leq -3.0$ and a range of EC

score thresholds (Table 3.2). In this case, the individual enrichment for genes in the domestication and improvement sets was not statistically significant. While it becomes significant for the combined gene set at certain percentile cutoffs, it remains very likely that the additional EC score threshold removes some genes affected by domestication. Therefore, we argue that z-score threshold is sufficient for the selection of rewired genes.

z	EC Score		Total	Domestication		Improvement		Dom. or Imp.	
Cutoff	%ile	Cutoff	Cnt	Cnt	p-value	Cnt	p-value	Cnt	p-value
-3.0	5	-0.0667	312	22	0.1640	14	0.6035	33	0.3156
-3.0	10	-0.0005	464	32	0.1372	29	0.0701	58	0.0245
-3.0	15	0.0442	551	37	0.1520	33	0.0883	66	0.0387
-3.0	20	0.0830	627	41	0.1784	35	0.1607	72	0.0681
-3.0	25	0.1147	690	46	0.1323	39	0.1287	80	0.0478
-3.0	30	0.1476	747	50	0.1154	40	0.2089	84	0.0770

Table 3.2. Effects of the additional EC score threshold. In addition to z-score threshold $z \leq -3.0$, an EC score cutoff was applied to select rewired genes. The obtained gene sets were no longer enriched in domestication and improvement genes individually. Thus, the addition of EC score cutoff appears to be detrimental to the selection of rewired genes.

3.3.3 Comparison to Other EC Methods

As described previously, our method relies on a z-score threshold to select rewired genes while the alternative EC methods depend exclusively on an EC score threshold. The correlation between z-score and EC score is relatively low ($r = 0.33$). There are particularly many cases where a low z-score corresponds to a high EC score and a few cases in which the opposite is true (Figure 3.3). About 10% of genes (1,834) have negative EC scores, almost twice as many as in the AEC set. The number of genes matching the size of the AEC set can be obtained at the threshold of $EC \leq -0.0474$ but this set is considerably different than AEC (Figure 3.3). The two sets have less than a third (353) of their genes in common. Moreover, the genes selected using the EC threshold are not enriched for either domestication, improvement, or combined gene sets (Fisher's exact test $p = 0.9829$, $p = 0.3643$, and $p = 0.8328$ respectively) from the sequence-based study by Hufford et al. (2012).

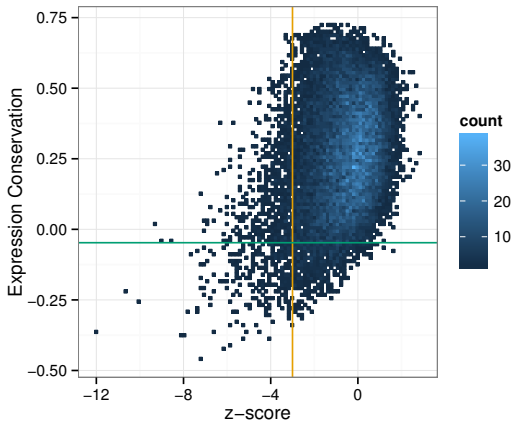


Figure 3.3. Distribution of z-scores in maize and teosinte. The z-scores were derived from random distributions of EC scores for each individual gene. Orange line shows the z-score threshold we used to select 1,115 rewired genes ($z \leq -3.0$; area to the left of the orange line.) To select the same number of genes using EC threshold, it needs to be set to -0.0474 (area below the green line.) The intersection between the two methods is only 353 genes.

To make sure that the threshold selection does not affect the enrichment, we created gene sets for a variety EC score thresholds that corresponded to 1st through 7th percentiles of the overall EC score distribution. We did not discover a statistically significant enrichment in any of the target types (Table 3.3). In fact, some of the gene sets (4th through 7th percentiles) were under-enriched for domestication genes. If the previously stated hypothesis regarding the higher impact of the domestication process on the transcriptome is correct, the under-enrichment indicates that the EC-based filtering method is surprisingly ineffective at detecting the genes with the largest co-expression changes.

3.3.4 Intersection Between AEC and DE Genes

We utilized the Cyber T application (Baldi and Long, 2001) to identify differentially expressed (DE) genes. The application is built upon a Bayesian probabilistic framework that derives the posterior probability of differential expression after modeling the distribution of a gene's log-expression values based on empirical variance and the local background variance of the neighboring genes. Using a conservative cutoff $P_{posterior} > 0.999$, we chose 612 genes that exhibited significantly different levels

EC Score		Total	Domestication		Improvement		Dom. or Imp.	
%ile	Cutoff	Genes	Genes	p-value	Genes	p-value	Genes	p-value
1	-0.1935	183	13	0.2321	12	0.1510	24	0.0755
2	-0.1445	365	18	0.7526	22	0.1358	38	0.3356
3	-0.1110	548	25	0.8866	26	0.5010	49	0.7375
4	-0.0859	730	31	0.9632	33	0.6090	62	0.8756
5	-0.0667	912	37	0.9885	41	0.6307	75	0.9429
6	-0.0489	1096	46	0.9883	53	0.4229	95	0.8846
7	-0.0366	1278	51	0.9976	62	0.4028	109	0.9299

Table 3.3. Enrichment of genes selected based on EC score only. There is no statistically significant enrichment in domestication and improvement genes when selection is based exclusively on EC scores. The number of genes in the 6th percentile is approximately the same as with $z \leq -3.0$ threshold.

of expression in maize compared to teosinte. Almost half of these genes exhibited two-fold or greater difference in expression and the majority of them (58.3%) tended to have higher expression in maize (Table 3.4). Hierarchical clustering of the expression levels of these genes showed that in some cases a subpopulation of maize lines could be more similar to teosinte than to other maize subpopulations (Figure 3.4).

Gene List	Genes	% 2 fold-change	% up-regulated in maize
DE	612	47	58.3
AEC	1115	16	57.1
AEC and DE	276	51	63.4

Table 3.4. Genes differentially expressed between maize and teosinte.

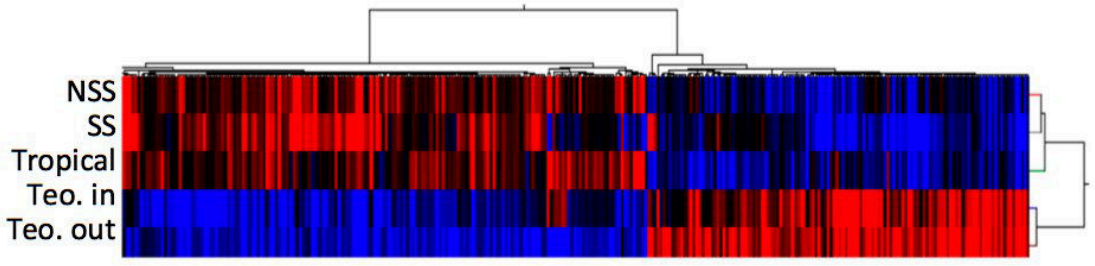


Figure 3.4. Hierarchical clustering of 612 differentially expressed genes. Genotypes were grouped into five subpopulations based on their origin. NSS - nonstiff stalk, SS - stiff stalk, Teo.in - teosinte inbred, Teo.out - teosinte outbred.

Differential expression of a gene between maize and teosinte does not necessarily imply that its regulatory relationships have been altered as well. Conversely, genes with altered co-expression profiles may have similar expression levels in both species. For maize and teosinte, the DE and AEC lists intersect in 276 genes, which is more than expected by chance ($p < 0.05$; Figure 3.6). Despite the intersection, DE and AEC genes possess contrasting characteristics and captured different aspects of expression changes between maize and teosinte (Figure 3.5). Therefore, the co-expression conservation and differential expression methods are complementary to each other. We conducted gene ontology (GO) analysis to find evidence that genes with lower expression in maize compared to teosinte were enriched in amino acid salvage, cellular respiration, and sulfur amino acids biosynthetic processes. Even though the majority of DE genes appear in regions syntenic with rice and sorghum, the genes with reduced expression in maize were mostly maize- or grass-specific (results not shown.)

To test whether the genes with altered expression between maize and teosinte were associated with developmental or anatomical differences between the taxa, we compared the lists of DE and AEC genes with gene clusters from a developmental co-expression network constructed with the expression data from 60 different tissues or stages of B73 (Sekhon et al., 2011). We did not uncover any statistically significant enrichment of DE and AEC genes in developmental co-expression clusters, indicating that the developmental or morphological differences between maize and teosinte cannot fully explain differential expression or altered expression conservation.

Both DE and AEC gene lists are significantly enriched ($p < 3e-3$ and $p < 3e-5$ respectively; Figure 3.6; Table 3.5) for genes found in the regions reported to be selection targets during maize domestication or improvement (Hufford et al., 2012). Selection targets are also overrepresented among genes that are both DE and AEC as well as among DE only genes. However, the enrichment for AEC-only genes is not significant (Table 3.5). This possibly supports the conclusion that AEC-only genes reflect the downstream effects of transcriptome rewiring while the causal targets of domestication are located elsewhere. Interestingly, various AEC and DE gene lists are more frequently enriched for genes from putative domestication regions than from putative improvement regions (Table 3.5). This may indicate that the effects

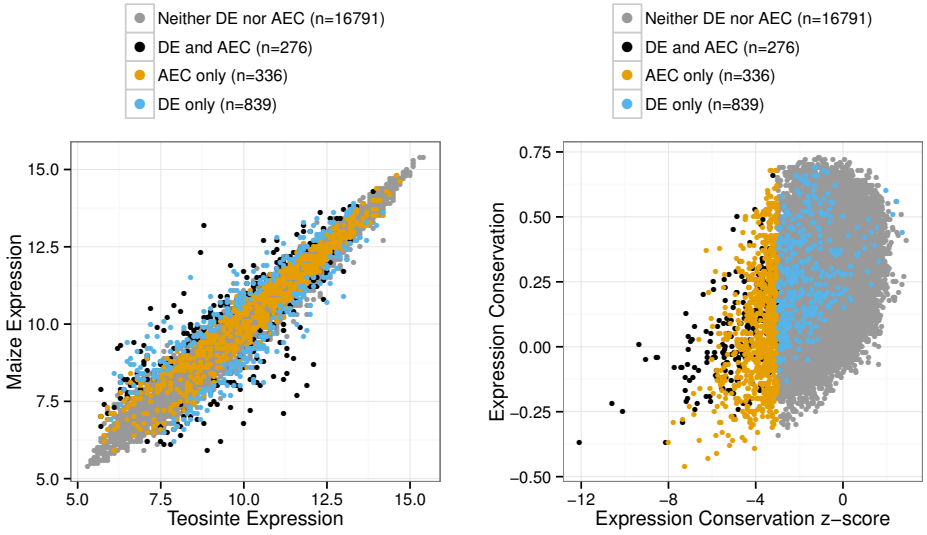


Figure 3.5. Intersection between differentially expressed genes and genes with altered expression conservation. The color of the symbols denotes whether the genes have differential expression (DE; blue), altered expression conservation (AEC; orange), both DE and AEC (black), or neither DE nor AEC (gray). (a) Mean expression levels in teosinte (X axis) are plotted against mean expression levels in maize (Y axis) for all genes. Many AEC genes (near the center of the distribution) do not display significant differential expression in maize and teosinte. (b) EC z-scores (X axis) are plotted against EC score (Y axis) for all genes. Many DE genes (upper right) do not show evidence for low expression conservation.

of domestication and improvement on transcriptome were quite different, quantitatively and qualitatively.

	Total	Domestication		Improvement		Dom. or Imp.	
Gene List	Genes	Genes	p-value	Genes	p-value	Genes	p-value
AEC all	1115	81	0.01057	64	0.05224	135	0.00312
DE all	612	56	0.00023	42	0.00863	90	0.00003
AEC only	839	54	0.16780	40	0.47600	89	0.18370
DE only	336	29	0.01490	18	0.31167	44	0.02299
AEC and DE	276	27	0.00376	24	0.00271	46	0.00017

Table 3.5. Enrichment of AEC and DE genes for selection targets identified by sequence-based analysis.

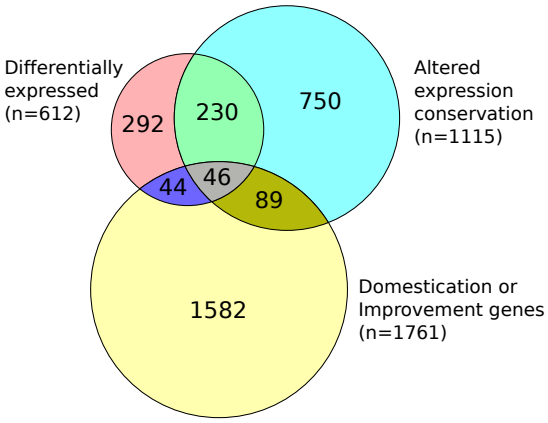


Figure 3.6. Overlap between AEC genes, DE genes, and the genes that occur in genomic regions that have evidence for selective sweeps during maize domestication or improvement.

Because of the observed significant enrichment, we performed additional analyses of the genes from the target regions that are present in either DE-only list (44 genes) or both DE and AEC lists (46 genes.) Annotations from the closest gene matches in *Arabidopsis thaliana* and the domain analysis show that 13 of the 90 genes are related to transcription or chromatin factors. The majority of genes in DE and both DE and AEC lists tend to have higher expression in maize with 35 of 46 and 27 of 44 up-regulated genes respectively. However, their connectivity in the teosinte co-expression network appears to be slightly higher compared to maize (Table 3.6). In fact, there are many genes with considerably different connectivity in the two networks, but this difference does not correlate with the direction of expression changes. Despite the presence of variability in connectivity between the co-expression networks, we have not detected any consistent pattern.

Since many of the domestication and improvement regions uncovered by Hufford et al. (2012) contained multiple genes, the authors designated the genes closest to the point with highest likelihood for selection as the most likely candidate targets. Many of the AEC and DE genes from the selected regions (47 out of 135 and 28 out of 90 respectively) coincide with the most probable selection target identified through population genetic analysis. The other AEC and DE genes that were not closest to the likelihood peak and were not identified as the most likely selection targets might

still represent noteworthy candidates.

Gene list	Dom/Imp genes	% up-reg. in maize	Enrich. signif.	% higher conn. in maize	% cand.
AEC and DE ($n = 276$)	46	76	0.0002	41.3	39.1
DE only ($n = 336$)	44	61	0.0230	40.9	22.7
AEC only ($n = 839$)	89	54	0.1837	57.3	32.6

Table 3.6. Genes in selected regions with evidence for DE or AEC. Cand., candidate; conn., connectivity; dom, domestication; enrich. signif., enrichment significance; imp, improvement; up-reg., up-regulated.

3.4 Conclusions

We have introduced a more sensitive method for the identification of rewired genes in co-expression networks. Previous studies applied a hard threshold to EC scores to determine the genes of interest. However, this approach ignores the genes that have relatively high EC score which is statistically lower than expected by chance. Our method derives a separate null distribution for each gene using bootstrapping analysis and checks whether the gene's EC score is a statistical outlier within that distribution in order to determine if the gene is rewired. Using the expression data from diverse maize and teosinte lines, we have shown that our method identifies a considerably different list of genes compared to the standard approach. Moreover, unlike the genes found by the standard approach, our list of rewired genes is significantly enriched for domestication and improvement selection targets suggested by an independent sequence-based approach. Since the genes with altered EC score do not necessarily have differential expression, AEC and DE methods produce complementary results. In addition, both methods are complementary to the sequence-based methods. While all three methods may identify the genes with direct effects on phenotype, AEC and DE methods also detect the downstream targets of genetic changes while the sequence-based methods identify variants further upstream.

Chapter 4

Characterization and Generalization of Expression Conservation Framework

4.1 Chapter Overview

In Chapter 3, we introduced Altered Expression Conservation (AEC), a novel method to identify genes whose co-expression profiles significantly differ between two networks. The method was successfully applied to find maize genes targeted during domestication and subsequent improvement. The goal of this chapter is to further characterize AEC and to discuss its generalization to other contexts. First of all, we will revisit the applicability of the alternative approaches and provide the results of a comparison between our AEC method and an existing iterative technique for calculation of expression conservation. Since evolutionary distance between the compared organisms and sample size of the underlying expression data may have pronounced effects on co-expression networks, the chapter also describes the results from the application of AEC to a pair of considerably more distant species and examines how the reduction of sample size degrades the power to discern the differences between two co-expression networks. Finally, the chapter demonstrates how AEC method can be used in a completely different context of genetic interactions.

4.2 Methods for Co-expression Network Comparison

There are two major frameworks to analyze the differences between co-expression networks: the Weighted Gene Co-expression Network Analysis (WGCNA) framework (Zhang and Horvath, 2005; Langfelder and Horvath, 2008) and the framework based on Expression Conservation (EC) methods (Dutilh et al., 2006; Tirosh and Barkai, 2007; Essien et al., 2008). Both frameworks rely on the same principles to construct co-expression networks but differ substantially in their approach to network computation and comparison. We discussed both frameworks in Chapters 2 and 3. Therefore, this section will only summarize the major points that were previously made regarding these frameworks.

4.2.1 WGCNA

WGCNA is a popular framework developed by Zhang and Horvath (2005) and later implemented as an R package by Langfelder and Horvath (2008). The framework found its use in several studies including the identification of gene targets for cancer therapies (Horvath et al., 2006), analysis of gene essentiality and network modularity in yeast (Carlson et al., 2006), and investigation of relationships between expression and phenotypic variation in plants (Weston et al., 2008).

For building co-expression networks, the R package provides a choice of three similarity measures (Langfelder and Horvath, 2008). As many prior studies (Eisen et al., 1998; Bergmann et al., 2003; Carter et al., 2004), it uses Pearson correlation coefficient by default. The alternative measures include the biweight midcorrelation (Wilcox, 2012) and the rank-based Spearman correlation. However, the network construction method uses hierarchical clustering to split larger input data sets into smaller subsets and computes expression profile similarities only within each subset. Such an approach substantially reduces computational needs. Moreover, since similarity between any two genes from different subsets is assumed to be 0, it also eases storage requirements due to sparsity of the output. On the other hand, this approach ignores potentially important relationships among genes from different subsets and complicates the comparison of co-expression networks at the global level.

The major contribution of WGCNA is the ability to generate a weighted adjacency

matrix by using a soft thresholding method such as the sigmoid function or the power adjacency function. To find genes' neighbors in the weighted adjacency network, it would still be necessary to use a hard threshold (Zhang and Horvath, 2005) that may cause the loss of information and sensitivity (Carter et al., 2004). The authors obviate this problem by finding the neighbors via clustering of the unthresholded adjacency network.

Clusters from two different networks can be subsequently aligned to find the genes with conserved or divergent co-expression profiles (Horvath et al., 2006; Weston et al., 2008; Ficklin and Feltus, 2011). Genes whose module membership varies between the networks are likely to be rewired. However, the method may not be sensitive enough to detect more subtle changes that are not strong enough to alter genes' module memberships. A few other studies contrasted gene connectivity in each network to find rewired genes (Oldham et al., 2006; Carlson et al., 2006). Yet, this approach is likely to be even less sensitive because rewiring may generate the same number of connections with a different set of genes.

4.2.2 Expression Conservation Framework

The alternative framework also allows a range of similarity measures to be used for construction of co-expression networks although most studies employ the Pearson correlation coefficient (Ihmels et al., 2005; Dutilh et al., 2006; Tirosh and Barkai, 2007; Essien et al., 2008; Guan et al., 2013). However, expression profile similarity is computed for all possible pairs regardless of the data set dimensions and the comparison is performed on unthresholded quantitative data that incorporates all available information. The alternative framework was also applied in many different contexts including identification of transcriptional differences between a fungal pathogen *Candida albicans* and a model yeast species *Saccharomyces cerevisiae* (Ihmels et al., 2005), estimation of the power of sequence identity to predict expression conservation of orthologous genes (Dutilh et al., 2006), and regulatory analysis of lifestyle genes in malaria-causing *Plasmodium* species (Essien et al., 2008).

Conservation of a gene's co-expression profiles between two networks can be estimated by the Pearson correlation coefficient. This value was termed the expression conservation (EC) score by Tirosh and Barkai (2007). A high EC score indicates that

gene's expression profile is similar to the same set of genes (neighbors) in both data sets. Suppose the expression pattern of gene A is the same in two data sets while gene B exhibits altered expression patterns between the data sets. If genes A and B are neighbors in the first data set, they would no longer be neighbors in the second data set. As a result, the EC score of gene A would be negatively affected even though its expression pattern is completely conserved. Therefore, Tirosh and Barkai (2007) reasoned that relationships between conserved genes should receive higher weight when calculating EC score. They suggested an iterative method that involved using EC scores from the previous iteration as weights for calculating the correlation coefficient. The method converges when EC scores do not significantly change between iterations. Similarly, a method comparison paper by Wang et al. (2010) investigated the effects of exclusively using the relationships between conserved genes to calculate EC score. Their approach was not iterative as they determined genes' conservation based on its null expectation. Thus, the approach is equivalent to using predetermined binary weights. Compared to the original approach by Dutilh et al. (2006), this algorithm modification resulted only in a small effect on the overall distribution of EC scores while the iterative algorithm caused more pronounced changes (Wang et al., 2010). In the absence of a good gold standard, the study by Wang et al. (2010) only reported comparative results that could not be used to rank the methods.

Previous studies selected conserved or divergent genes based on raw EC scores (Dutilh et al., 2006; Tirosh and Barkai, 2007; Essien et al., 2008; Guan et al., 2013). Since background noise and other measurement variations drive co-expression profiles apart, the raw EC scores represent a good indicator of co-expression conservation. However, it is more difficult to identify rewired genes because the extent of biological variation is not obvious. Low EC scores may be expected by chance for some genes while high EC scores may be well below the null expectation for others. This can potentially result in both high false positive and high false negative rates. In Chapter 3, we proposed Altered Expression Conservation (AEC) method that addresses this problem by comparing EC score to the gene's null expectation.

Chapter 3 also described the application of AEC method to the expression data from a maize domestication study and reported better sensitivity of AEC compared to

the original gene selection method. However, we computed EC score as unweighted correlation coefficient and we did not previously evaluate the performance of our approach against the iterative EC algorithm proposed by Tirosh and Barkai (2007). In addition, the maize domestication data set had relatively few samples. As the number of samples increases, the estimates of expression profile similarity established by Pearson correlation coefficient become more accurate and statistically significant (Reverter and Chan, 2008). It is not clear what effect this would have on EC scores. Moreover, the evolutionary distance between maize and teosinte is very small and most genes have well defined orthologs with highly similar DNA sequences. Two evolutionary distant organisms are likely to have substantial genomic differences but the comparative analysis will be limited to relatively conserved orthologs. As with sample size, the effects of increased evolutionary distance may be quite convoluted. Finally, AEC method may not generalize well to other contexts even when the experimental setup is very similar to comparative expression analysis. The rest of the chapter discusses these problems in more detail.

4.3 Iterative Expression Conservation

To prevent divergent neighboring genes from reducing a conserved gene's EC score, Tirosh and Barkai (2007) developed an iterative method that computes EC score by assigning more weight to relationships with conserved genes. Suppose R^M and R^T are maize and teosinte co-expression networks respectively. Co-expression profiles of gene i in those two networks are denoted by R_i^M and R_i^T for $i = 1..n$. The algorithm starts by calculating unweighted Pearson correlation coefficient between the two co-expression profiles of each gene:

$$EC_0(i) = PCC(R_i^M, R_i^T)$$

For each subsequent iteration, weighted Pearson correlation coefficient is computed using the EC scores from the previous iteration as weights.

$$EC_k(i) = wPCC(R_i^M, R_i^T, EC_{k-1})$$

The weighted Pearson correlation coefficient $wPCC$ is calculated as

$$wPCC(X,Y,w) = \frac{\sum_{i=1}^{n} w_i(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} w_i(X_i - \bar{X})^2 \sum_{i=1}^{n} w_i(Y_i - \bar{Y})^2}}$$

where $w_i = EC_{k-1}(i)$. The procedure converges when the squared sum of differences is below predetermined level ϵ.

$$\sum_{i=1}^{n} [EC_k(i) - EC_{k-1}(i)]^2 < \epsilon$$

Using the same $\epsilon = 0.1$ as in Tirosh and Barkai (2007), we ran the iterative algorithm to contrast genes in the maize and teosinte co-expression networks described in Chapter 2 and compared the output to the results of our AEC method reported in Chapter 3.

The iterative algorithm converged after 9 iterations. EC scores from the two methods were highly correlated ($r_{pearson} = 0.96$ and $r_{spearman} = 0.95$) but their magnitude was reduced considerably (Figure 4.1). While very few EC scores were driven to 0 (3 original EC scores versus 34 iterative EC scores,) the number of negative values rose from 1,834 to 2,476. Thus, in addition to rotation, the whole distribution shifted down (Figure 4.1).

We used the iterative EC scores as weights to derive a null distribution for each gene by calculating the weighted Pearson correlation coefficient between gene's co-expression profiles in the same random network pairs from our earlier bootstrapping analysis (see Chapters 2 and 3.) Based on these null distributions, we calculated z-scores as before. The were considerable differences between the original and iterative z-scores (Figure 4.2a). The application of weights to the Pearson correlation coefficient made co-expression profiles of many genes very similar across the random networks in a pair. Consequently, it increased many z-scores (Figure 4.2a). Compared to 1,115 genes that were originally significant at $z \leq -3$, only 856 were below the cutoff based on the final iterative z-score. However, the changes between individual iterations were relatively minor indicating only a gradual refinement after

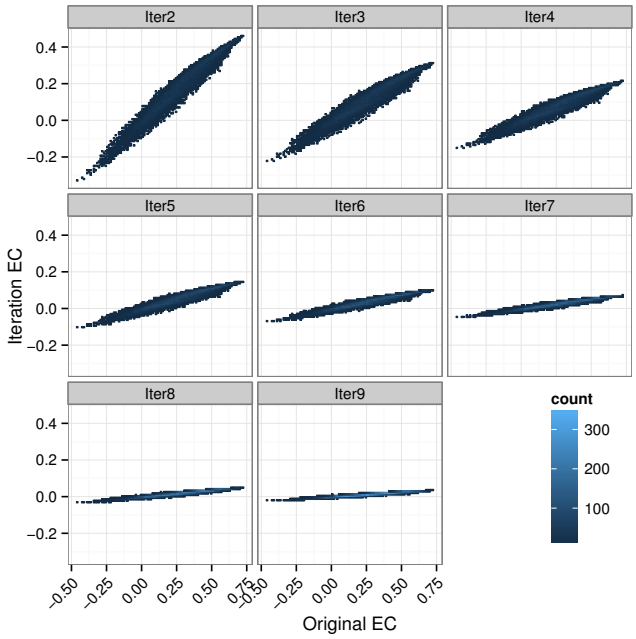
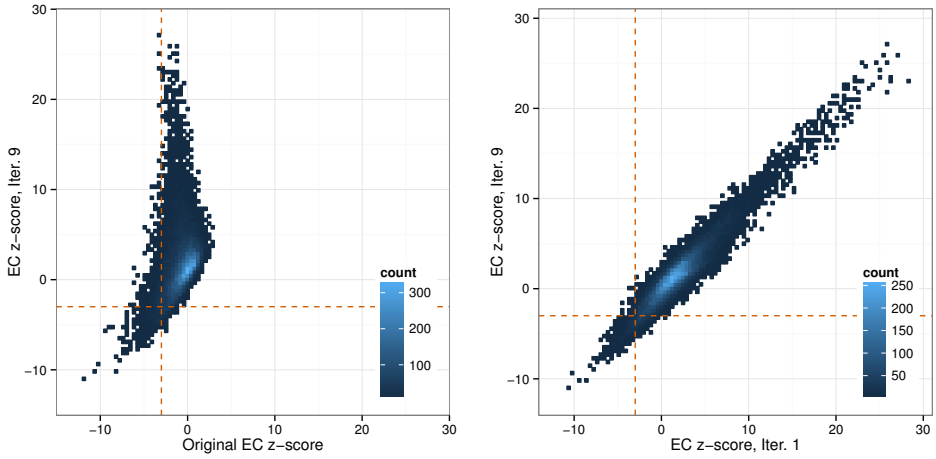


Figure 4.1. Comparison of the original and iterative EC scores. EC scores computed at each iteration (Y axis) are plotted against the original EC scores (X axis.) The changes introduced by the iterative algorithm predominantly affected the magnitude of EC scores leaving their ranking almost the same ($r_{spearman} = 0.95$.)

the first iteration (Figure 4.2b).

We examined the genes with low z-scores ($z \leq -3$) to see whether they are enriched for domestication and improvement genes identified by an independent sequence-based study (Hufford et al., 2012). To make sure that the choice of the convergence criteria for the iterative algorithm did not affect the results, genes were selected and enrichment was computed using EC scores from each iteration. Unlike the original gene list, genes that were identified by the iterative algorithm lacked significant enrichment in the sequence-based domestication and improvement genes (Table 4.1). At each iteration and for each enrichment type, the p-values were below the significance threshold of 0.05.

Since the iterative algorithm potentially improves the accuracy of EC scores, z-scores may be unnecessary for adequate gene selection. Therefore, an EC threshold was determined for each iteration to select the same number of genes as with z-score



(a) Original vs Iteration 9 **(b)** Iteration 1 vs Iteration 9

Figure 4.2. Comparison between (a) the original z-scores and z-scores from the last iteration (b) z-scores from the first and last iterations. Orange line indicate z-score threshold $z \leq -3$. The variation between the original and iterative z-scores is substantial while the changes between iterations are much more subtle.

threshold. The intersection between these lists and domestication and improvement genes was not significant either (Table 4.2). On the contrary, in some instances underenrichment for domestication genes was observed.

	Total	Domestication		Improvement		Dom. or Imp.	
	Genes	Cnt	p-value	Cnt	p-value	Cnt	p-value
Original	1115	81	0.0106	64	0.0522	135	0.0031
Iteration 1	492	34	0.1265	28	0.1661	56	0.1094
Iteration 2	709	43	0.3271	37	0.2689	74	0.2530
Iteration 3	767	48	0.2411	40	0.2604	82	0.1752
Iteration 4	790	48	0.3108	42	0.2156	84	0.1855
Iteration 5	796	48	0.3301	41	0.2839	83	0.2413
Iteration 6	799	47	0.3982	40	0.3523	82	0.2926
Iteration 7	801	48	0.3464	41	0.2978	83	0.2604
Iteration 8	803	48	0.3530	41	0.3035	83	0.2682
Iteration 9	802	48	0.3497	42	0.2455	84	0.2266

Table 4.1. Enrichment of genes with low iterative z-score for sequence-based domestication and improvement genes from Hufford et al. (2012). Genes were selected using $z \leq -3$ threshold.

Thus, for the maize domestication data set the iterative approach to the EC score

	Total Genes	Domestication Cnt	Domestication p-value	Improvement Cnt	Improvement p-value	Dom. or Imp. Cnt	Dom. or Imp. p-value
Original	1096	46	0.9883	53	0.4229	95	0.9299
Iteration 1	492	22	0.8937	24	0.4476	44	0.7278
Iteration 2	709	27	0.9903	31	0.6803	55	0.9678
Iteration 3	768	29	0.9934	35	0.5913	60	0.9691
Iteration 4	791	33	0.9751	37	0.5270	66	0.9115
Iteration 5	796	34	0.9665	35	0.6757	65	0.9378
Iteration 6	801	34	0.9695	37	0.5587	67	0.9094
Iteration 7	801	34	0.9695	36	0.6258	66	0.9285
Iteration 8	826	34	0.9813	37	0.6351	67	0.9475
Iteration 9	806	34	0.9723	37	0.5744	67	0.9184

Table 4.2. Enrichment of genes with low iterative EC score for sequence-based domestication and improvement genes from Hufford et al. (2012). Genes were selected using an EC score threshold to match the number of genes selected with $z \leq -3$ threshold. The actual counts may differ whenever multiple genes have EC score that equals the EC threshold.

calculation did not provide any benefits compared to the original non-iterative approach. It is possible that our AEC approach already mitigates the negative effects of rewired genes on EC scores of their neighbors by calculating the null expectation for each gene individually. Therefore, an additional adjustment would be redundant. It would be interesting to compare iterative and non-iterative methods in another context but without a good gold standard, a precise comparison may be problematic.

4.4 Effects of Evolutionary Distance and Sample Size

So far, we have shown that the original implementation of our AEC method performs better on the maize domestication data set than the alternative approaches. However, the domestication data set has relatively few samples and covers two closely related subspecies. A larger number of samples would increase the accuracy and statistical significance of the expression profile similarity estimates (Reverter and Chan, 2008), which in turn would affect EC score calculations. On the other hand, higher evolutionary divergence may reduce the precision of EC scores due to many factors such as imprecise identification of orthologous relationships. Will the AEC method generalize to other expression data sets?

4.4.1 Yeast Co-expression Networks

Saccharomyces cerevisiae and *Saccharomyces bayanus* are two yeast species that diverged approximately 5-20 million years ago (Kellis et al., 2003). This is a huge distance compared to maize and teosinte especially if one considers that generation time is a year for maize and only a few hours for yeast. Yet, similar to maize and teosinte, *S. cerevisiae* and *S. bayanus* can still form viable hybrids (Kishimoto, 1994). Unlike *S. cerevisiae*, *S. bayanus* had been an obscure, poorly-studied species until a large-scale experiment by Guan et al. (2010) who used the available microarray data from *S. cerevisiae* to determine a minimal set of microarray experiments on *S. bayanus* that would achieve the functional annotation level comparable to *S. cerevisiae*.

The *S. bayanus* co-expression network was constructed from 197 microarray experiments conducted in Guan et al. (2010). However, to avoid potential experimental bias, we used the expression data from 562 *S. cerevisiae* microarray experiments collected in an independent study by Huttenhower et al. (2006) and subsequently refined by Koch et al. (2012). The raw microarray data was processed and normalized in their respective studies. The *S. cerevisiae* and *S. bayanus* data sets contained 6,082 and 5,898 genes respectively. Based on the orthology information from Guan et al. (2010), the data sets encompassed 5,489 orthologous genes and only these genes were used to build the co-expression networks. Otherwise, the process of network construction and comparison followed the procedure described in Chapter 2.

The global correlation between *S. cerevisiae* and *S. bayanus* co-expression networks was much lower than between the maize and teosinte networks (0.1004 and 0.3038 respectively.) On the contrary, the random networks displayed very strong similarity (Figure 4.3). This is driven by similarity of gene co-expression profiles between two networks indicating that the separation between the actual values and the random values would persist on the gene level as well. Indeed, the highest EC z-score was -2.14 and all but two of them were below -3 (Figure 4.4).

The drastic difference in the results may be due to evolutionary distance, sample size, or both. Evolutionary distance is an intrinsic parameter and cannot be altered. However, the effects of the sample size may be measured by subsampling. To this end, we randomly selected 10 times each 30, 40, 62, 200, and 400 microarrays

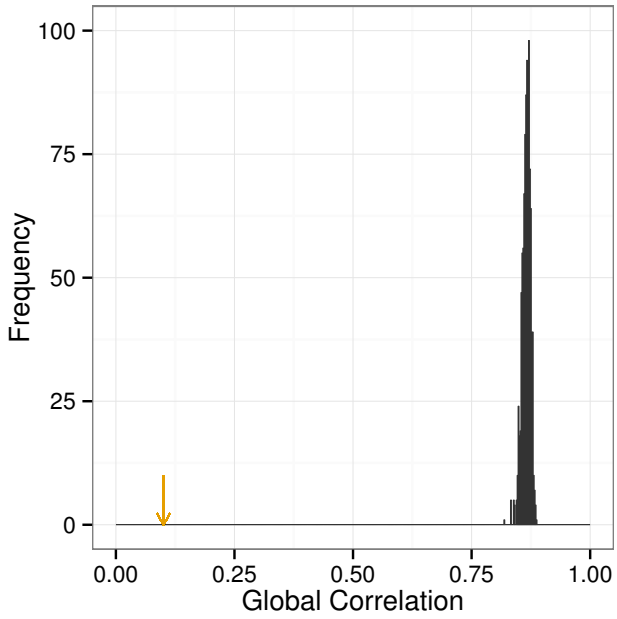


Figure 4.3. Global correlation between *S. cerevisiae* and *S. bayanus* co-expression networks. The null distribution for the global correlation coefficients is derived from 1000 pairs of random networks generated by selecting microarray experiments randomly from either species. Orange arrow indicates the global correlation of the actual co-expression networks ($\rho = 0.1004$). The random networks are much more similar than the actual pair.

while preserving the ratio of *S. cerevisiae* to *S. bayanus* samples. For instance, all subsamples of 30 microarrays contained 22 *S. cerevisiae* and 8 *S. bayanus* microarrays. Within each subsample, we followed our regular pipeline (see Chapters 2 and 3 for details) to construct a co-expression network for each of the two species, calculate the global correlation between them, and derive the null distribution for the global network correlation using bootstrapping analysis. To facilitate the comparison among subsamples, the distance between the actual network correlation ρ and its null distribution was measured within each subsample as

$$z = \frac{\rho - \mu_{null}}{\sigma_{null}}$$

where μ_{null} and σ_{null} were the mean and standard deviation of the subsample's null distribution respectively. The distance is minimal when the value of z approaches

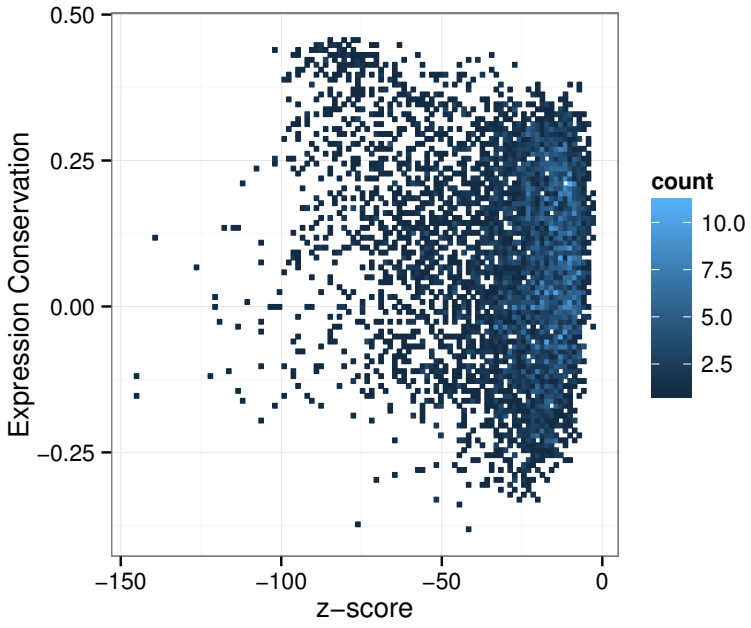


Figure 4.4. Comparison between EC and z-score for all orthologous genes in *S. cerevisiae* and *S. bayanus* co-expression networks. Notably, all z-scores are below zero and there is no correlation between EC scores and z-scores.

zero. As the number of microarrays in a subsample decreases, the random networks become less similar to each other and their distribution shifts closer to the actual network correlation value (Figure 4.5).

4.4.2 Subsampling Maize Domestication Data Set

To see whether the reduction in random network similarity is typical when the sample size gets smaller, we performed similar subsampling analysis on the maize domestication data set. Since the original data set was already small covering only 38 maize and 24 teosinte lines, it was subsampled without preserving the species ratio to produce 20 data sets each of 31/24, 24/24, 20/20, and 15/15 maize/teosinte lines respectively. Co-expression networks were built and analyzed as previously described except the bootstrapping procedure was reduced to 200 rounds due to computational reasons.

Overall, the observed decrease in random network similarity was comparable to

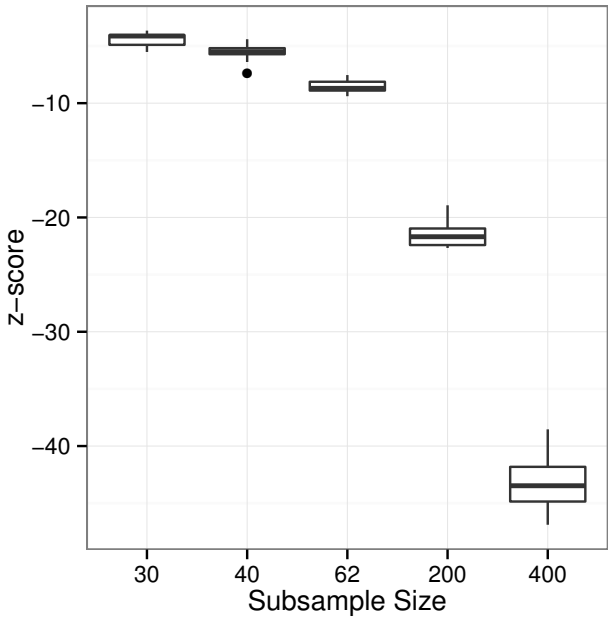
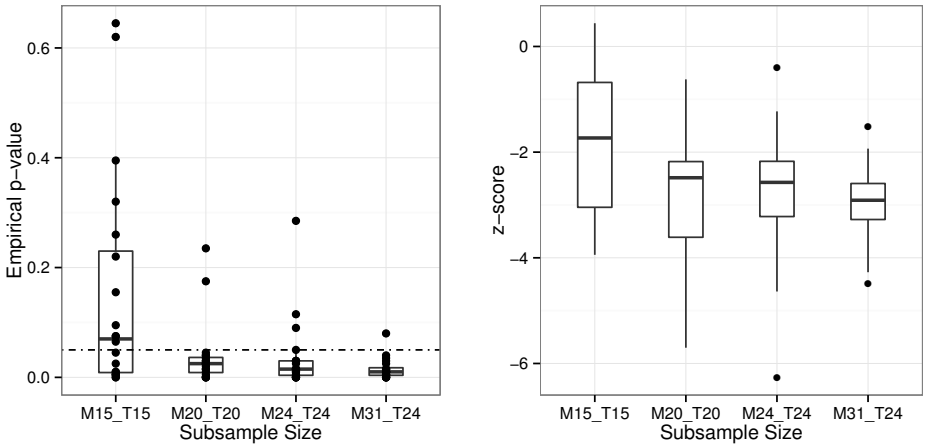


Figure 4.5. Global correlation in subsampled *S. cerevisiae* and *S. bayanus* co-expression networks. The original data set was subsampled to select 10 times each 30, 40, 62, 200, and 400 microarrays. Co-expression networks were constructed and analyzed as described in Chapters 2 and 3. Decreasing the number of samples also reduces the similarity between random network pairs as indicated by the z-score between the actual network correlation and its null distribution.

the one observed in yeast co-expression networks (Figure 4.6a). As the subsample size was getting smaller, the number of random network pairs that were more different than the actual network (denoted by the empirical p-value in Figure 4.6a) was rising. Once the subsample size reached 15/15, the differences between the real network could no longer be distinguished from the differences between the random network pairs in the majority of cases (points above the dash-dot line in Figure 4.6a). This indicates that the minimum number of samples required for co-expression network comparison in closely related species is approximately 20. The few outliers in subsamples of size 20 and larger are likely caused by the reduced number of rounds in bootstrapping. For each subsample, the bootstrapping analysis was limited to 200 rounds compared to 1000 rounds performed with the full maize domestication data set.



(a) **(b)**

Figure 4.6. Global correlation in subsampled maize and teosinte co-expression networks. The original data set was sampled to select 20 times each 31/24, 24/24, 20/20, and 15/15 maize/teosinte lines respectively. Axis X denotes subsampling size. In the labels, M stands for maize and T for teosinte with the subsequent number showing the number of lines chosen. (a) Distribution of empirical p-values for the differences between the co-expression networks. It is calculated as a proportion of random network pair whose global correlation is below the global correlation of the actual networks. The dash-dot line indicates the significance threshold at $p \leq 0.05$. (b) The distance between the actual global correlation and the mean of the null distribution measured as z-score.

4.4.3 Discussion

Both maize and yeast data sets exhibit similar trends. As the number of samples increases, the random co-expression networks become more similar. While the global correlation between the actual networks may also grow, this growth is not as substantial, so the distance between the null distribution and the actual value increases as well. Intuitively, the improvement in the estimate of the actual global correlation is expected because having a larger number of samples raises the accuracy of expression profiles and the statistical significance of co-expression relationships (Reverter and Chan, 2008). The following example illustrates the reasons for the increased similarity between the random networks.

Without loss of generality, consider an expression data set with the equal number of maize and teosinte lines. The data set is used to construct a co-expression network for each species. Suppose a rewired gene has a single neighbor A in the maize

co-expression network but a different lone neighbor B in the teosinte co-expression network. Since the original data set is sampled randomly, the expected ratio of maize and teosinte lines in both random subsets is also one to one. Thus, both co-expression network have a partial signal from both species and the rewired gene will have neighbors A and B in both of them. However, both neighbors will be approximately half as similar to the rewired gene because the signal from one species is diluted by the signal from another species. It is likely that the signal from teosinte, for instance, would be comparable to noise for the relationship with the neighbor A. Thus, the more samples the expression data set contains the more signal would rise above the noise. The real relationships would be more complicated than this simplified example but overall the signal would split approximately evenly between the random networks and cause them to be similar to each other.

Increasing the number of samples widens the distance between the actual global network correlation and its null distribution resulting in unintended effects on the AEC method. EC z-scores may no longer be quantitatively informative as they all drop well below zero. Nevertheless, the ranking based on EC z-scores may still provide insights as to the extent of gene rewiring. Without a gold standard, the usefulness of the AEC method cannot be accurately evaluated in the yeast data set. It would be helpful to increase the number of maize and teosinte samples until the random networks become very similar to each other and then measure the enrichment in sequence-based domestication and improvement genes. Interestingly, the selection of maize and teosinte lines may be important because the similarity between the actual networks became equivalent to the random network similarity in several larger subsamples (Figure 4.6). However, the presence of the outliers may also be explained by the fewer rounds of bootstrapping performed on the subsampled data (200 for the subsamples versus 1000 for the full data set.) This warrants further research in sample selection.

As expected, large evolutionary distance caused the actual yeast co-expression networks to be less similar than maize and teosinte networks. Even though the distance between the actual global network correlation and its null distribution was also higher than in similarly sized maize subsamples, it is hard to make a definitive conclusion based on just these two cases. It would be beneficial to analyze another data

set with either a different domesticated species or a more recently diverged yeast species.

In summary, we showed that several factors influence the construction and comparative analysis of co-expression networks. Reducing sample size eventually makes random co-expression networks less similar than the actual networks and at least 20 samples are required to construct an adequate co-expression network if the network is to be used for comparative analysis. Selection of lines may possibly matter even when the number of samples is as high as 31 but it becomes less important with a larger number of samples. Evolutionary distance negatively correlates with co-expression network similarity and, as it increases, the gap between the actual global network correlation and the null distribution increases as well.

4.5 Alternative Application of the Expression Conservation Framework

Another interesting question is whether the AEC approach is general enough to work with non-expression data. There are many other types of experiments where quantitative data are generated across a multitude of conditions or genotypes. In particular, expression data share several properties with genetic interaction (GI) networks. In both cases, the data is quantitative and available for many genes simultaneously. Previous studies investigated the conservation of GI profile similarity across two organisms (Roguev et al., 2008; Dixon et al., 2008; Frost et al., 2012). Even though the similarity was measured as a binary relationship, the studies provided a good estimate of functional conservation and rewiring between *S. cerevisiae* and another yeast, *Schizosaccharomyces pombe*. In this section we will look at the conservation of GI profile similarities from a different angle. We will examine how the relationships change when interactions are measured against essential versus non-essential genes.

4.5.1 Genetic Interaction Overview

Gene products interact in a complex fashion forming intricate pathways and obscuring how the regulation of a single gene can affect the phenotype (Hartman et al., 2001). For example, deletion of a single gene may cause no visible effects on a cell indicating that the gene's functionality is buffered by another gene or genes (Figure 4.7). In such cases, deleting an additional gene may provide beneficial information about the functionality of both genes (Tong et al., 2001, 2004). In yeast, a convenient quantitative measure for the results of a gene deletion is fitness (Hartman et al., 2001) which can be derived from the size that a colony has after predetermined period of growth (Costanzo et al., 2010).

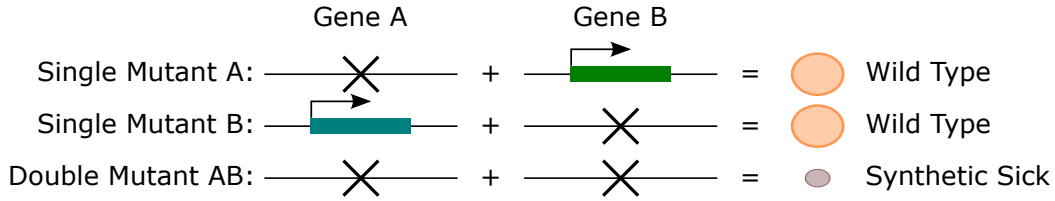


Figure 4.7. Gene loss buffering. Genes A and B have the same functionality. If one of these genes is deleted, the other gene can buffer for the loss of the first gene and the cell remains healthy. However, the deletion of both genes results in a synthetically sick cell.

Suppose the deletion of gene A reduces the fitness from 1 (wild type) to 0.7 while the deletion of gene B results in more severe defects decreasing the fitness to 0.3. If genes A and B are unrelated, we expect the deletion of both genes to have multiplicative effect on fitness, i.e. the expected fitness of the double mutant is 0.21. If the observed double mutant fitness exceeds that value, genes A and B are said to have a positive interaction. If the observed fitness is below 0.21, the interaction is negative (Figure 4.8). Deviations from the expected values for a set of gene pairs can be combined into a GI network where rows and columns represent query and array genes respectively. Each row is also called the GI profile of a gene.

The genome of *S. saccharomyces* includes ~6,000 predicted genes, out of which ~1,100 are essential for the cell survival (Giaever et al., 2002). While an organism with a knocked out essential gene dies under normal conditions, it is still possible

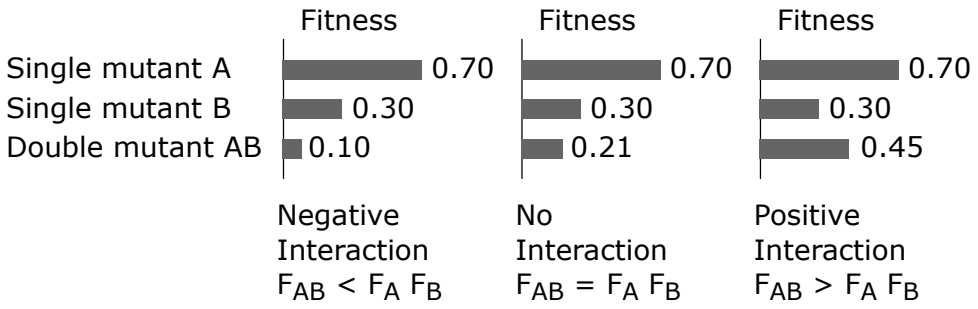


Figure 4.8. Types of synthetic genetic interactions. If the double mutant fitness is lower than expected, the interaction between genes A and B is negative. If it is higher than expected, the interaction is positive.

to measure both single mutant and double mutant fitness for essential genes by using partially functional (conditional or hypomorphic) alleles (Costanzo et al., 2010). Recently published *S. cerevisiae* GI network incorporated ~5.4 million interactions for 1,712 query genes out of which 334 were conditional or hypomorphic alleles (Costanzo et al., 2010). Those query genes were screened against non-essential array genes. While the authors continued expanding the set of query genes, they also started screening query genes against essential array genes (unpublished data). This extended data set at the time of writing contained 2,135 combined essential and non-essential query genes screened against 3,906 non-essential genes (CxN data set) and 1,688 combined essential and non-essential query genes screened against 795 essential query genes (CxE data set.)

Considering the difficulty of essential gene screening, it would be beneficial to know whether the CxE provides any additional information about gene relationships because this would help to refine screening strategies. If CxE only reinforces the relationships established in CxN, most efforts should be initially directed to the screening against non-essential genes. If, however, some genes have different neighbors in CxN and CxE, it would be important to begin screening essential genes at an early stage.

4.5.2 Genetic Interaction Profile Similarity Networks

To examine the differences between gene neighborhoods in CxN and CxE, we used these data sets to build two GI profile similarity networks. Although the construction of such networks is almost identical to the construction of co-expression networks, several distinctions still exist. First, the GI data distribution differs from the distribution of expression data. Since the values in GI profiles represent the difference between the actual fitness and expected fitness, their absolute value is usually close to 0 and rarely exceeds 1. For instance, CxN values ranged from -1.12 to 1.32. On the other hand, expression data only contains non-negative values that often have much larger magnitude. For example, log-transformed microarray data may range from 4 to 16.

Second, GI profiles are more likely to have missing values due to experimental difficulties with double mutant construction. The expression data analyzed in this dissertation did not have any missing values, albeit missing values can appear when combining expression data from independent studies that used varying microarray designs. When Pearson correlation coefficient is calculated between two profiles, the pairs of measurements with one or both missing values are ignored. Therefore, to ensure comparable significance of correlation coefficients between GI profiles, all query genes with more than 20% of missing values were removed from both CxN and CxE. The filtering removed over 40% of rows leaving 982 query genes in common between CxN and CxE.

Third, a GI data set is likely to have considerably more columns than an expression data set. The latter is usually limited to at most a few hundred genotypes or a few dozen conditions. The former may have well over a thousand array genes. In particular, CxN and CxE contained 3,906 and 795 array genes respectively while the relatively large *S. cerevisiae* and *S. bayanus* expression data sets encompassed only 562 and 197 microarrays.

Finally, GI data sets may contain profiles for several alleles of the same gene. CxN and CxE each include 982 rows but cover only 814 unique genes. The presence of multiple alleles per gene does not affect the network construction but it can provide additional directions for research. In principle, expression data may also contain

multiple alleles per gene if several genotypes are profiled across various experimental conditions. However, it is less clear how to integrate such expression data into a single data set. For example, averaging the data for a gene's allele shared across the genotypes may break co-expression relationships when a transcription factor for the gene has divergent alleles.

Apart from the filtering of missing values, our network construction and analysis pipeline remained the same. On the global level, the two profile similarity networks exhibited relatively high correlation ($\rho = 0.4510$). This is higher than the correlation between maize and teosinte co-expression networks ($\rho = 0.3038$) discussed in Chapter 2 as well as the correlation between *S. cerevisiae* and *S. bayanus* co-expression networks discussed earlier in this chapter. However, it is much lower than the correlation between maize co-expression networks based on microarray and RNA-seq data ($\rho = 0.75$) that will be described in Chapter 5.

As expected, due to the large number of array genes (columns) the random profile similarity networks generated by bootstrapping ($N = 1000$) the combined GI data set (CxN + CxE) were much more similar than the actual networks (Figure 4.9). For each allele in the data set, we computed Profile Conservation score analogous to EC score and used the results of the bootstrapping analysis to derive the corresponding z-scores. Out of 982 alleles, 19 exhibited negative Profile Conservation score while 51 more had the Profile Conservation score below 0.1 (Figure 4.10). Due to high similarity of the random networks, all z-scores are negative and their values are unlikely to be informative. However, the ranking based on z-scores may still be useful for the identification of genes with significantly altered profiles. The correlation between Profile Conservation scores and z-scores is fairly high ($\rho = 0.5512$) but there are many examples where the scores disagree (Figure 4.10). However, without a well-defined gold standard, it is hard to evaluate their relative performance. Nevertheless, the lack of conservation between essential and non-essential profile similarity networks demonstrates that CxE provides novel and potentially important information. Therefore, the laborious screening against essential genes may be necessary to reconstruct the full picture of genetic interactions in an organism.

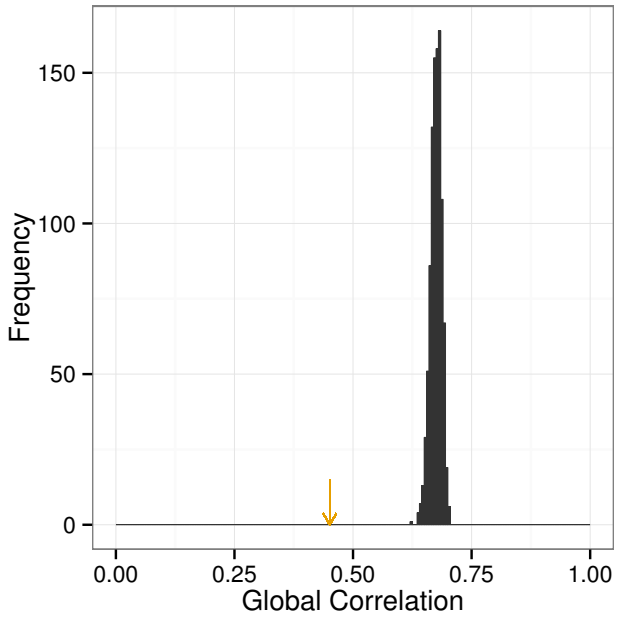


Figure 4.9. Global correlation between essential and non-essential profile similarity networks. Distribution of global network correlation coefficients was obtained from profile similarity networks constructed from 1000 pairs of random GI networks. Orange arrow indicates the global correlation between the actual essential and non-essential profile similarity networks ($\rho = 0.4510$). As expected, due to the large number of array genes (columns) the random networks are much more similar than the actual networks.

4.6 Conclusions

The analysis presented in this chapter suggests that the AEC method works best when the expression data sets are fairly similar, e.g. they are obtained from the same tissue of a closely related species under similar conditions. We showed that the iterative approach to expression conservation does not always perform better than the original non-iterative approach. In particular, the AEC method with non-iterative calculations outperformed both iterative EC and iterative AEC methods when applied to the maize domestication data set. We also demonstrated that for the maize domestication data set at least 20 samples would have been required to detect the differences between the co-expression networks. This number may vary for other data sets and the bootstrapping analysis may be necessary to determine whether

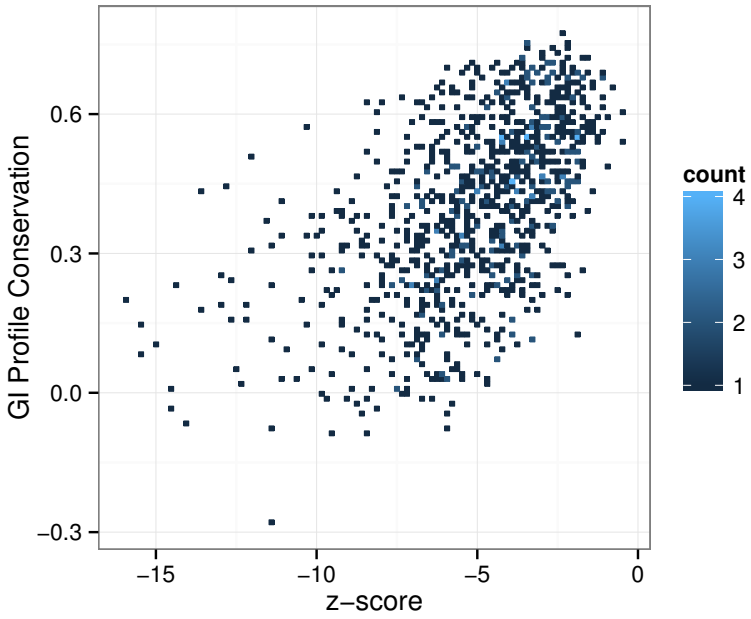


Figure 4.10. Comparison between Profile Conservation and z-score calculated for the essential and non-essential profile similarity networks. Despite relatively high correlation ($\rho = 0.5512$), the Profile Conservation score and z-score appear to measure different aspects of GI conservation.

the number of samples is sufficient. When we employed our AEC method to compare the co-expression networks of two yeast species, the actual co-expression networks were much more different than the random networks built from the bootstrapped expression data. The difference can be partially explained by the substantial evolutionary distance between the species but it could also be due to the large sample size of the yeast data set. We observed a similar trend with the GI profile similarity networks. In both cases, our method produced the results that clearly diverged from the EC and Profile Conservation methods respectively. Due to the lack of a well-defined gold standard, we could not evaluate the relative performance of each method in those two cases. Thus, further research is needed to determine the effects of evolutionary distance and sample size on the aforementioned methods.

Chapter 5

Comparative Evaluation of Transcriptomes Reconstructed from Microarray and RNA Sequencing Data

5.1 Chapter Overview

In previous chapters we performed various analyses on co-expression networks constructed from microarray data. Apart from some niche applications, RNA sequencing has largely replaced the microarray platform in gene expression studies. However, a large body of microarray expression data has accumulated over many years of its active use and it may be beneficial to include these data alongside the RNA sequencing data into co-expression analyses. In this chapter, we compare the expression data generated by the two platforms from the same samples taken from diverse maize tissues. We find that co-expression networks built individually from microarray and RNA-seq data highly correlate with each other. Using bootstrapping analysis, we show that co-expression networks constructed from a mix of microarray and RNA-seq data are very similar to the networks built on the data from a single platform.

The analyses presented in this chapter were published in Sekhon et al. (2013)

with contributions from Rajandeep Sekhon, Candice Hirsch, Chad Myers, Nathan Springer, Robin Buell, Natalia de Leon, and Shawn Kaeppler. Rajandeep, Natalia, and Shawn conceived and designed the experiments. Rajandeep and Candice subsequently conducted the experiments, processed the raw expression data, performed principal component analysis, and ran hierarchical clustering of the expression data. Natalia assisted with biological interpretation of the data. Chad helped with technical aspects of co-expression analysis. Shawn, Nathan, and Robin supervised the project.

5.2 Transcriptome Analysis: from Microarrays to RNA Sequencing

The systems biology approach to elucidating mechanisms behind phenotypic changes involves the integration of large-scale genomic, transcriptomic, proteomic, and metabolomic information. Rapidly evolving technologies have already yielded a plethora of genomic and transcriptomic data for many species prompting additional efforts to analyze temporal and spatial variation in gene expression that occur during development of an organism. These efforts will help us better understand functions of individual genes as well as various properties of expression networks. In plants, developmental expression data have been published for several species including Arabidopsis (*Arabidopsis thaliana*) (Schmid et al., 2005), barley (*Hordeum vulgare*) (Druka et al., 2006), maize (*Zea mays*) (Sekhon et al., 2011), medicago (*Medicago truncatula*) (Benedito et al., 2008), rice (*Oryza sativa*) (Jiao et al., 2009), and soybean (*Glycine max*) (Libault et al., 2010).

Up until recently, microarray technologies dominated the field of genome-wide transcription analysis. There have been considerable advances that migrated the maize expression studies from spotted cDNA amplicons (Lee et al., 2002) to spotted oligonucleotide arrays (Rensink and Buell, 2005) and subsequently to *in situ* DNA synthesis approaches such as Affymetrix (Kirst et al., 2006). The release of the complete maize reference genome (Schnable et al., 2009) further advanced transcriptional analysis efforts and facilitated the release of a genome-wide expression atlas covering various developmental stages and organs of maize (Sekhon et al., 2011).

Despite technological and data processing improvements, microarrays possess

an inherent set of weaknesses. In particular, the design of microarray probes is limited to the gene models defined at the time of array construction. Therefore, a potentially large amount of expression data covering unannotated gene models will be unavailable. Since the microarray technology relies on DNA-DNA hybridization, the estimates for highly homologous genes may be inaccurate. Moreover, structural variation and even single nucleotide polymorphisms may also affect hybridization efficiency in non-reference lines and as a result complicate the accurate assessment of gene expression levels. Finally, microarrays tend to have a considerable background noise and suffer from signal saturation. Both factors limit the expression dynamic range hindering the comparison of very highly or lowly expressed genes.

The improvements in "next generation" deep-sequencing technologies enabled the development of RNA sequencing, termed RNA-seq (Wang et al., 2009), which challenged the leading role of microarray technology in transcriptome analyses. The main advantage of RNA-seq is the ability to improve the quality of transcriptome by reprocessing the raw data whenever genomic sequence or annotation are updated. It also allows discerning highly homologous genes by adjusting the stringency of mapping parameters. Several plant transcriptomes produced with RNA-seq have been published recently including arabidopsis (Filichkin et al., 2010), rice (Zhang et al., 2010; Lu et al., 2010), and soybean (Libault et al., 2010; Severin et al., 2010). In maize, the technology was used to construct comprehensive transcriptomes for leaf (Li et al., 2010) and inflorescence (Eveland et al., 2010).

In this study, we employed RNA-seq to profile a subset of samples from the previous microarray-based study (Sekhon et al., 2011) to improve the accuracy of the expression data for homologous genes and to expand the coverage to the whole genome. We contrasted the differences between the genome-wide expression estimates produced by microarray and RNA-seq technologies in order to determine their overall efficiency. To measure the effects of technology selection on co-expression analysis, we compared co-expression networks individually built on microarray and RNA-seq data. We released the data online to complement the existing microarray-based expression atlas.

5.3 Tools for Transcriptome Comparison

5.3.1 Estimation of gene expression levels

The samples for RNA-seq were obtained from the remaining subset of RNA material collected for the microarray-based maize gene atlas (Sekhon et al., 2011). All tissue samples were retrieved from reference inbred line B73 plants grown at the West Madison Agricultural Research Station (Verona, WI) during summer 2008. The details regarding conditions and sampling methods are provided in Sekhon et al. (2011). This study encompasses 18 tissues from 6 organs (Table 5.1).

#	Tissue Name	Plant Ontol. Term	Plant Ontol. Tissue Descr.
1	24H Germinating Seed	PO:0009001	Fruit (Kernel)
2	6DAS GH Primary Root	PO:0020127	Primary root
3	V3 Stem and SAM	PO:0020148	Soot apical meristem
		PO:0020142	Stem internode
4	V5 Tip of stage-2 Leaf	PO:0025142	Leaf tip
		PO:0009025	Vascular leaf
5	V9 Immature Leaves	PO:0009025	Vascular leaf
6	16DAP Endosperm	PO:0009089	Endosperm
7	16DAP Embryo	PO:0009009	Plant embryo
8	V9 Eighth Leaf	PO:0009025	Vascular leaf
9	V9 Eleventh Leaf	PO:0009025	Vascular leaf
10	V9 Thirteenth Leaf	PO:0009025	Vascular leaf
11	VT Thirteenth Leaf	PO:0009025	Vascular leaf
12	R2 Thirteenth Leaf	PO:0009025	Vascular leaf
13	10DAP Whole seed	PO:0009001	Fruit
14	12DAP Whole seed	PO:0009001	Fruit
15	12DAP Endopsperm	PO:0009089	Endosperm
16	14DAP Whole seed	PO:0009001	Fruit
17	14DAP Endopsperm	PO:0009089	Endosperm
18	16DAP Whole seed	PO:0009001	Fruit

Table 5.1. List of tissues included in RNA-seq gene atlas. The following abbreviations are used in tissue names: H, hours; DAS, days after sowing; GH, greenhouse; V, vegetative; DAP, days after pollination; VT, vegetative tasseling; R, reproductive.

Approximately 5 μg of total RNA was used to isolate mRNA. Fragmented mRNA

was converted to cDNA and PCR amplified according to the Illumina RNA-seq protocol (Illumina, Inc. Sand Diego, CA). Sequence reads were generated by the Illumina Genome Analyzer II (San Diego, CA) and Illumina HiSeq 2000 (San Diego, CA) at the University of Wisconsin Biotechnology Center (Madison, WI). Multiplexing was performed on a portion of the samples using Illumina barcodes. The length of the generated single-end reads varied between 35 and 101 bp (Table 5.2). All presented data passed the quality control based on the Illumina purity filter and distribution of base quality scores at each cycle. Sequences were uploaded to the Sequence Read Archive at the National center for Biotechnology Information (accession number SRP010680).

Sequence reads for each tissue were mapped to B73 reference genome pseudomolecules v1 and v2 (Schnable et al., 2009) using Bowtie version 0.12.7 (Langmead et al., 2009) and the splice site aware aligner TopHat version 1.2.0 (Trapnell et al., 2009). Default values were accepted for all parameters except minimum and maximum intron length, which were set to 5 bp and 60,000 bp respectively. Read mapping was performed without gene model annotation. Normalized gene expression values in fragments per kilobase pair of exon model per million fragments mapped (FPKM) were calculated with Cufflinks version 0.9.3 (Trapnell et al., 2010) using the maximum intron length of 60,000 bp and the quartile normalization option. The bias detection and correction algorithms were configured to use the 4a.53 and 5b.60 annotations (`http://ftp.maizesequence.org/`) for v1 and v2 pseudomolecules respectively. All other parameters were set to their default values. For all the analyses, FPKM values were averaged across three replicates.

5.3.2 Microarray and RNA-seq correlations

Since the microarray probes used to generate the previously published atlas (Sekhon et al., 2011) were based on the 4a.53 annotation of the reference genome, we also used that version for the processing of RNA-seq reads in order to enable the direct comparison between the data sets. The longest peptide was chosen for genes with multiple transcript annotations. Due to the limitations of the microarray design, it only covered 22,151 gene models available from the 4a.53 version. Correlations between microarray and RNA-seq expression profiles were calculated for all those

22,151 genes as well as for 17,811 genes determined to be expressed in both data sets. In the microarray data set, an expressed gene was required to have an average expression value of at least 200 in one of the 60 tissues surveyed in the earlier study (Sekhon et al., 2011). In the RNA-seq data set, expressed genes needed to have both FPKM 95% confidence interval boundaries above zero (Hansey et al., 2012) in at least one tissue. Confidence intervals were determined by Cufflinks (Trapnell et al., 2010). Out of 19,744 genes expressed according to the microarray data, 1,933 were not found expressed in the RNA-seq data set yielding 17,811 unique transcripts for this analysis. Before calculating the correlations, all values were log_2 transformed. Since some FPKM values in the RNA-seq data set were less than 1, they were replaced with 1 to avoid taking the logarithm of very small values.

5.3.3 Hierarchical clustering

Hierarchical clustering was performed using the unweighted pair-group method with complete linkage approach and Pearson's correlation as a similarity measure in the Spotfire DSFG package (`http://spotfire.tibco.com/`).

5.3.4 Tissue Specificity

We estimated tissue specificity by calculating Shannon entropy (Shannon, 1948; Schug et al., 2005; Zhang et al., 2007) for genes with detectable expression levels on both microarray and RNA-seq platforms. The entropy was calculated as

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i$$

where p_i is a relative abundance of the gene's transcript in tissue i.

5.3.5 Co-expression network analysis

Genes that did not have detectable expression levels in either data set were removed leaving 19,328 genes from the filtered gene set (FGS) for further analyses. For the microarray data set, genes with the average expression value exceeding 200 in at

least one of the tissues were considered expressed. In case of RNA-seq, the average expression value of a gene had to be greater than 0 FPKM in at least one of the 18 tissues. Due to the differences in dynamic ranges of the two platforms, we applied log_2 transformation to the microarray expression data and inverse hyperbolic sine ($asinh$) transformation to the RNA-seq data. The latter compresses larger values more than smaller values and works well for the values below 1 (Figure 5.1). Individual co-expression networks were generated based on the transformed data sets by calculating Pearson correlation coefficient for each pair of gene expression profiles using a custom C++ application that relied on Sleipnir library (Huttenhower et al., 2008). Fisher transformation and normalization were applied to the values in both co-expression networks as recommended by Huttenhower et al. (2006).

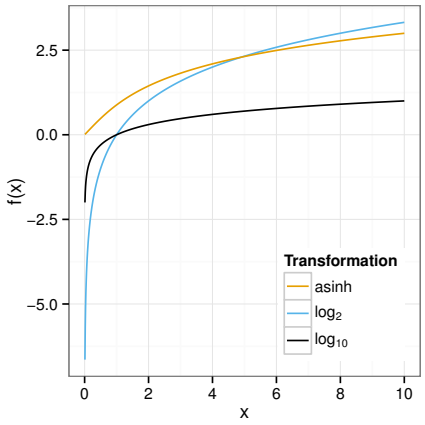


Figure 5.1. Comparison of various transformation functions. Inverse hyperbolic sine transformation ($asinh$) compresses large values more than small values. Moreover, values below 1 remain positive after the transformation.

Expression conservation analysis was performed in a similar fashion as described in Chapter 3. Briefly, expression conservation (EC) scores were obtained by calculating Pearson correlation coefficient between co-expression profiles of a gene in two networks. The significance of EC score was determined based on the gene's null expectation derived from the bootstrapping analysis that involved generation of 1,000 random co-expression network pairs. However, unlike the previous analysis the random networks were generated by selecting a mixture of RNA-seq and microarray

profiles for the 18 tissue samples to make sure that each random network was derived from the same set of tissues as the original networks.

5.4 Comparison of Expression Data

Since we used exactly the same samples to measure expression on both microarray and RNA-seq platforms, the samples are perfectly paired and we can compare expression levels directly for each gene using some correlation measure. Unless the measure is rank-based, the expression data needs to be log-transformed. Since the dynamic range of the two platforms is very different, large values in RNA-seq data may cause spurious relationships.

5.4.1 Overview of samples and quality assessment

To enable direct comparison between microarray and RNA-seq technologies, we selected a subset of the remnant total RNA from the earlier microarray-based study (Sekhon et al., 2011) and profiled it using RNA-seq. The subset contained 18 diverse tissues that represented distinct stages of maize plant development. A complete list of tissues analyzed in this study along with their plant ontology terms and descriptions are given in Table 5.1. For each tissue sample, we sequenced three biological replicates each containing pooled total RNA from three randomly selected plants. The RNA sequencing produced between 5 and 28 million single-end (35-101 bp) reads per tissue (Table 5.2). The reads were averaged across all three biological replicates. Between 55.8% and 88.8% of the obtained reads mapped to the B73 filtered gene set transcripts version 5b.60 (`http://ftp.maizesequence.org`). Whenever multiple transcripts were available for a gene, we used the longest peptide possible. Expression values were calculated in units of fragments per kilobase of exon model per million fragments mapped (FPKM). Despite the variation in read numbers, read length, and percentages of reads mapped across the tissues and biological replicates, the results produced by the RNA-seq technology were highly reproducible. All biological replicates from a single tissue highly correlated with each other having the average Pearson's correlation coefficient of 0.971 ± 0.004 while 83% of the correlations were over 0.950 (results not shown.)

#	Tissue name	Average reads per replicate	Read length (nt)	Platform
1	24H Germinating Seed	23,982,734	35-76	GAII
2	6DAS GH Primary Root	26,816,316	35-76	GAII
3	V3 Stem and SAM	26,601,631	35-76	GAII
4	V5 Tip of stage-2 Leaf	27,955,538	35-76	GAII
5	V9 Immature Leaves	5,025,002	35	GAII
6	16DAP Endosperm	5,423,343	35	GAII
7	16DAP Embryo	5,590,257	35	GAII
8	V9 Eighth Leaf	8,416,273	75	GAII
9	V9 Eleventh Leaf	6,079,889	75	GAII
10	V9 Thirteenth Leaf	6,997,373	75	GAII
11	VT Thirteenth Leaf	6,609,967	75	GAII
12	R2 Thirteenth Leaf	7,936,757	75	GAII
13	10DAP Whole seed	9,665,262	101	HiSeq
14	12DAP Whole seed	5,911,421	101	HiSeq
15	12DAP Endopsperm	9,624,958	101	HiSeq
16	14DAP Whole seed	9,956,128	101	HiSeq
17	14DAP Endopsperm	11,624,855	101	HiSeq
18	16DAP Whole seed	9,063,914	101	HiSeq

Table 5.2. Average number of reads, read length, and other details of RNA sequencing.

5.4.2 Global gene expression trends

The FGS version 5b.60 that we employed for gene-based analyses lists 39,429 genes but excludes transposons, pseudogenes, contaminants, and other low-confidence annotations. We considered the genes to be expressed in the RNA-seq data set whenever their FPKM 95% confidence interval, as reported by Cufflinks (Trapnell et al., 2010), was entirely above zero (Hansey et al., 2012). Using this criterion, we detected 29,447 (74.7%) FGS genes that were transcribed in at least one tissue. Almost a fifth (18.3%) of the non-transcribed genes represented *ab initio* genes predicted by Fgenesh (Salamov and Solovyev, 2000) and accounted for 60.1% of all *ab initio* genes in the FGS version 5b.60. Another fifth (22.3%) corresponded to very short transcripts with sizes below 500 bp and covered 67.6% of such genes. This is in sharp contrast to 0.2% of non-expressed genes that encode transcripts larger

than 5 kb. Finally, the majority (84.4%) of the non-transcribed genes lacked functional annotation in the current version of maize genome. That group accounted for 42.6% of genes without annotation. Overall, even though some genes may not be expressed in the profiled tissues, many other non-expressed genes are likely to be poorly annotated. Some of the non-expressed genes possibly encode rare transcripts that were missed due to low depth of sequencing. This was partially confirmed by the fact that the two tissues with the lowest number of reads, V9 Immature Leaves and 16DAP Endosperm (Table 5.2), also had the highest number of non-expressed genes (Figure 5.2).

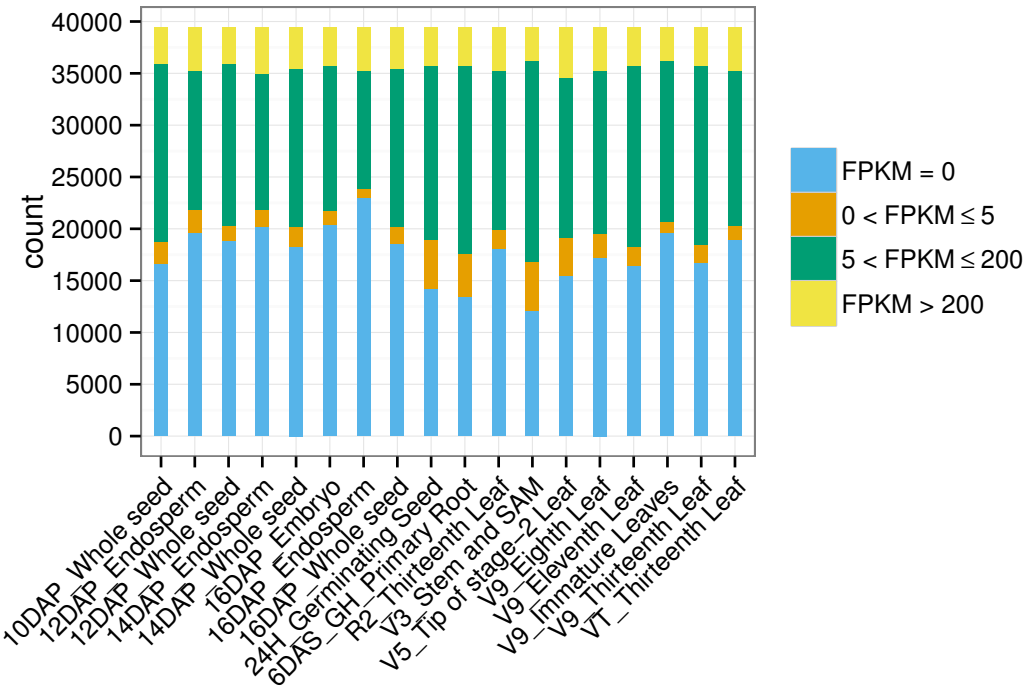


Figure 5.2. Distribution of genes based on magnitude of expression in 18 maize tissues. A gene was considered expressed if the whole FPKM 95% confidence interval was above 0. For each tissue, the expressed genes were further divided into low ($0 < \text{FPKM} \leq 5$), medium ($5 < \text{FPKM} \leq 200$), and high (FPKM > 200) expression.

Hierarchical clustering of the transcriptomes constructed from RNA-seq data revealed coherent grouping of the tissues based on their biological identity (Figure 5.3). Both meristematic tissues, 6 Days After Sowing (DAS) primary root (contains root

apical meristem) and V3 stem and shoot apical meristem (SAM), formed a distinct cluster indicating commonality between their transcriptomes. Germinating seed and embryo appear separate from other tissues consistent with their specialized biological function. The clustering also exhibits trascriptional differences among different developmental stages of the same organ. For example, Sylvester et al. (1990) suggested the division of maize leaf development into at least three stages of cell division and growth (I, II, and III) followed by a fully matured state with distinct morphological and anatomical differences at each stage. V9 immature leaf displays intensive blade and ligule growth that is characteristic to stage II (Sylvester et al., 1990) and its transcriptome is distinctly different from the fully mature V9 eights leaf (Figure 5.3).

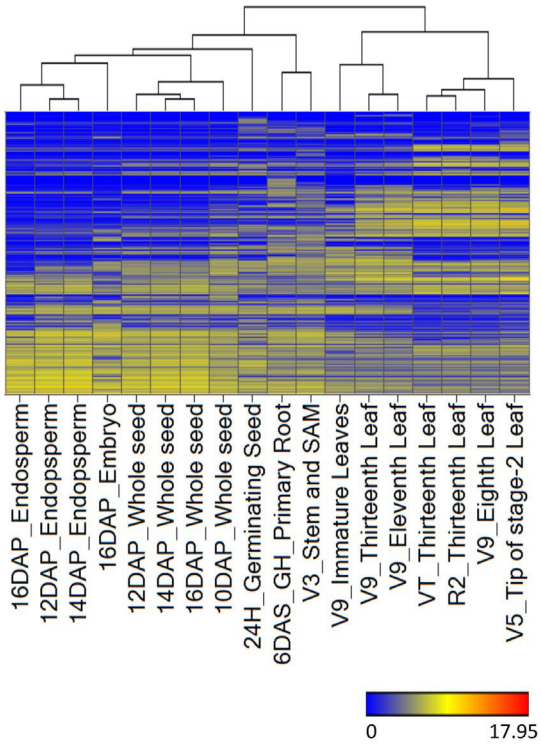


Figure 5.3. Heat map showing hierarchical clustering of tissues based on global gene expression. Clustering was based on log_2 transformed Fragments Per Kilobase Exon model per Million mapped fragments (FPKM) values of 29,038 genes considered expressed in at least one tissue based on the FPKM 95% confidence interval being greater than zero. Red, yellow, and blue colors indicate high, medium, and low levels of gene log_2 transformed expression respectively.

Despite the relatively low depth of sequencing in some tissues (Table 5.2), biologically consistent clustering of the tissues and very high correlation between the replicates indicate that the depth is sufficient for making inferences about the transcriptome.

5.4.3 RNA-seq and microarrays produce very similar global expression trends

Since both RNA-seq and microarray data sets used in this study were generated from the same RNA samples, the variance due to growing conditions, tissue handling, and RNA extraction had been eliminated. However, the microarray data set was based on the version 4a.53 annotation and covered only 22,151 genes (Sekhon et al., 2011). For consistency, we aligned RNA-seq reads to the maize pseudomolecules v1 and utilized the version 4a.53 annotation to create RNA-seq data set for direct comparison with the microarray data set. We calculated the Pearson correlation coefficient between the corresponding tissues in the two data sets using all 22,151 common genes as well as only 17,811 genes expressed in both tissues (see Section 5.3.2 for details.) We observed significant correlation ($p < 0.001$) between gene expression estimates for the eighteen tissues from RNA-seq and microarray data sets with the individual coefficients ranging between 0.70 and 0.83 (Table 5.3). These correlation estimates appeared in line with previous reports (Marioni et al., 2008; Mortazavi et al., 2008; Davidson et al., 2011) and there were only minor differences between correlation coefficients calculated with the expressed genes and those calculated with all common genes (Table 5.3). Overall, these results indicate that RNA-seq and microarray technologies produce highly correlated transcriptomes.

5.4.4 RNA-seq based gene atlas provides better breadth of coverage of the transcriptome compared to the microarray-derived atlas

Based on the criteria described in Subsection 5.3.2, we identified approximately the same percentage ($\approx 82\%$) of genes as being expressed in RNA-seq or microarray

#	Tissue	Pearson Correlation Coefficient Expressed Genes	All Common Genes
1	24H Germinating Seed	0.73	0.71
2	6DAS GH Primary Root	0.71	0.72
3	V3 Stem and SAM	0.67	0.71
4	V5 Tip of stage-2 Leaf	0.77	0.78
5	V9 Immature Leaves	0.75	0.77
6	V9 Thirteenth Leaf	0.71	0.74
7	V9 Eleventh Leaf	0.69	0.72
8	V9 Eighth Leaf	0.72	0.75
9	VT Thirteenth Leaf	0.75	0.76
10	R2 Thirteenth Leaf	0.75	0.77
11	10DAP Whole seed	0.75	0.76
12	12DAP Whole seed	0.79	0.79
13	14DAP Whole seed	0.80	0.78
14	16DAP Whole seed	0.80	0.81
15	12DAP Endosperm	0.83	0.81
16	14DAP Endosperm	0.83	0.82
17	16DAP Endosperm	0.81	0.81
18	16DAP Embryo	0.79	0.79

Table 5.3. Correlation between RNA-seq and microarray expression values. The third column shows Pearson correlation coefficient between genes that are considered expression on both platforms in at least one tissue. The fourth column lists the correlation using all genes common to both platforms.

data sets. However, only 22,153 out of 32,535 gene models (version 4a.53) had representative probes on the custom NimbleGen microarray used in the previous study (Sekhon et al., 2011). Thus, in absolute terms RNA-seq provided more comprehensive coverage than microarray with 26,711 and 18,382 genes expressed in at least one tissue respectively.

Expressed genes detected by RNA-seq also encompassed a higher number of the classical maize genes that had been overrepresented in maize genetics literature due to easily recognizable mutant phenotypes (Schnable and Freeling, 2011). Out of 464 described classical genes, RNA-seq and microarray analysis identified as expressed 427 and 390 respectively. The expression patterns of these genes were consistent with the expected trends. For example, *brown midrib3* (*bm3*) that encodes

caffeic acid O-methyltransferase enzyme involved in the lignin biosynthetic pathway (Vignols et al., 1995) is preferentially expressed in developing leaves where active lignification takes place. An APETALA2-like gene *glossy15* controls juvenile to adult vegetative phase change (Moose and Sisco, 1996) and it was detected only in shoot apical meristem at vegetative-3 stage. A developmental expression gradient peaking in 14DAP endosperm was clearly visible for DMT101 whose closest Arabidopsis homolog MET1 is responsible for DNA methylation (Kankel et al., 2003). DNA hypermethylation accompanies the imprinting process that typically occurs in endosperm (Gehring et al., 2004). Finally, the expresion of a *Myb* transcription factor *purple plant1* was primarily detected in leaves where it was reported to control anthocyanin synthesis (Cone et al., 1993).

Shannon entropy (Shannon, 1948) is often used to estimate the tissue specificity of gene expression across samples (Schug et al., 2005; Zhang et al., 2007). We assessed the tissue specificity of gene expression in both platforms and found more examples of tissue-specific patterns in RNA-seq data than in microarray data (Mann-Whitney U test, $p < 0.01$; Figure 5.4). Thus, RNA-seq provided enhanced coverage of transcriptome with more tissue-specific patterns.

5.4.5 Resolution of expression of paralogs by RNA-seq and microarray

To determine the ability of the two technologies to discern the expression of paralogs, we analyzed the expression of paralogous gene pairs from the two sub-genomes of maize that appeared after a whole-genome duplication event (Schnable et al., 2012). Overall, 2,434 paralogous gene pairs were present in both RNA-seq and microarray data sets. In the RNA-seq data set, paralogs exhibited higher correlation coefficients than a set of randomly selected gene pairs (Figure 5.5). However, the microarray data set possessed more highly correlated (over 2 standard deviations above the random gene average) paralog pairs (41.2%) than the RNA-seq data set (31.1%). This higher correlation between paralogous pairs in microarray data can be explained by the tendency of genes with similar sequences to cross-hybridize on microarrays.

To illustrate these results, we provide a more detailed analysis of two paralogous genes, *Brittle-2* (*Bt2*) and *Agpslzm/L2*, that both encode a small subunit of

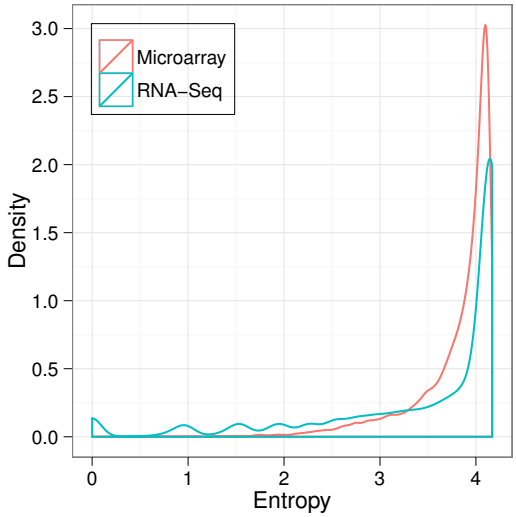


Figure 5.4. Tissue specificity for RNA-seq and microarray platforms was estimated as Shannon entropy for each gene's expression profile. Distribution of the entropy values is shown for both microarray and RNA-seq data sets. Lower entropy indicates higher tissue specificity. Tissue-specific expression patterns are more prevalent in the RNA-seq data set (Mann-Whitney U test, $p < 0.01$) indicating higher sensitivity of the platform to the expression differences between genes.

ADP-glucose pyrophosphorylase (AGP) as reported by Bae et al. (1990) and Prioul et al. (1994) respectively. The duplication likely occurred during tetraploidization of maize genome (Rösti and Denyer, 2007). Despite possessing high nucleotide similarity (84%) between their mRNA sequences, the genes are tissue specific. *Bt2* encodes a cytosolic small subunit with high expression in the endosperm while *Agpslzm/L2* encodes a plastidial small subunit with preferential expression in leaves (Rösti and Denyer, 2007; Hannah et al., 2001). The microarray data set reports substantial expression for *Bt2* in the endosperm and whole seed tissues but detectable levels are also present in other tissues as well (Figure 5.6a). In particular, the expression was unexpectedly high in the embryo sample and potentially caused by cross-hybridization of *Agp2*, the third gene that encodes a plastidial AGP small subunit and is normally activated in embryo (Giroux and Hannah, 1994). However, based on RNA-seq data *Bt2* appears expressed exclusively in the seed samples. Similarly, the leaf specific *Agpslzm/L2* has moderate expression in the seed tissues according

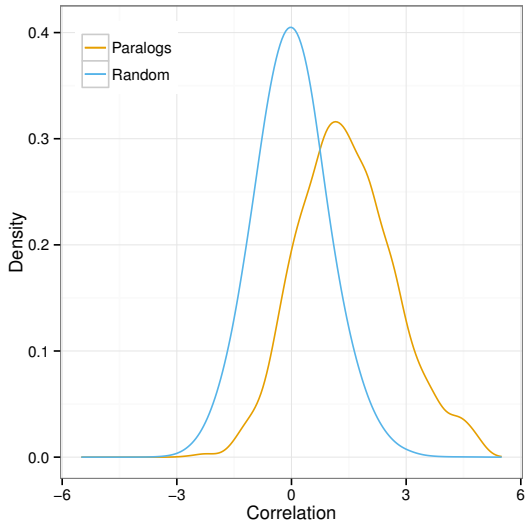
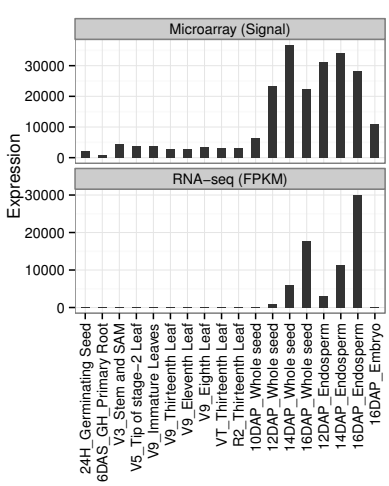
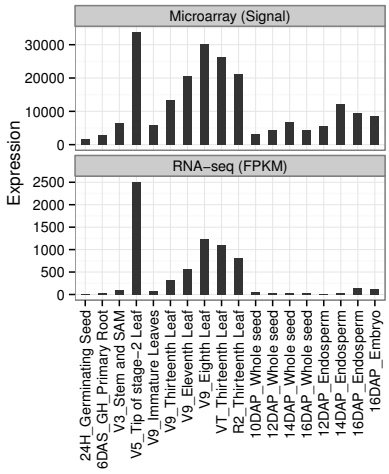


Figure 5.5. Density estimates for the distribution of the correlation coefficients of paralogous genes in the RNA-seq co-expression network. Pearson correlation coefficients were calculated for each of the 2,434 paralogous gene pairs that were expressed in at least one tissue of both RNA-seq and microarray data sets. The density plot illustrates the values for these correlation coefficients relative to a set of randomly selected genes.

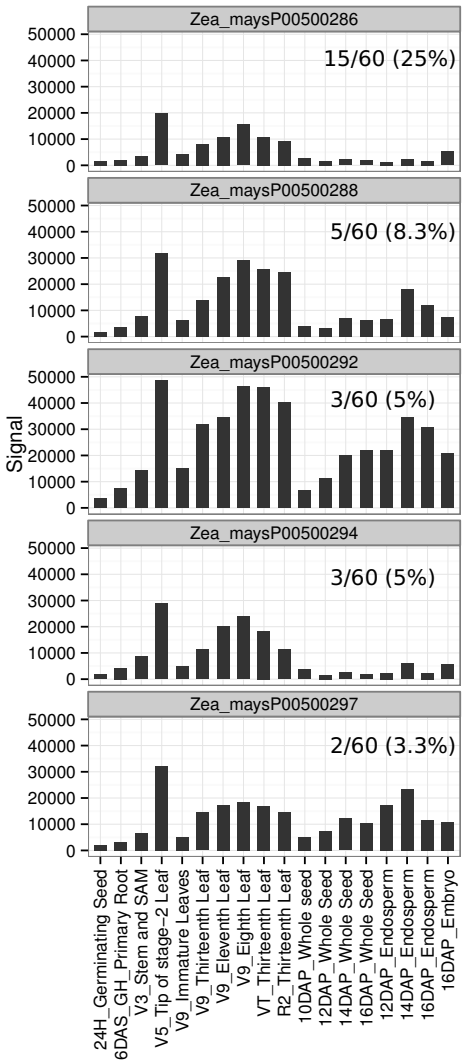
to the microarray results, but RNA-seq reports detectable expression only in mature leaves where starch accumulation is expected (Figure 5.6b). To make sure that cross-hybridization indeed causes spurious correlation in microarray data, we reviewed the expression levels reported by individual probes that represented *Agpslzm/L2* gene. We found that the *Agpslzm/L2* 60-nucleotide probes that had 2-3 mismatches with *Bt2* in fact contributed considerable spurious signal in seed tissues. Only the probes with 5 or more mismatches had seed specific expression (Figure 5.6c). Thus, cross-hybridization has a great negative impact on the ability of microarrays to detect differences in expression levels between paralogs. Consequently, RNA-seq offers better resolution for genes with similar sequences.



(a) *Brittle-2*



(b) *Agpslzm/L*



(c) Individual Probes

Figure 5.6. Comparative performance of RNA-seq and microarray to discern expression of paralogous genes. (a) Expression patterns of endosperm specific *Brittle-2* (*Bt2*) gene. (b) Expression patterns of leaf specific *Agpslzm/L* gene. (c) Expression patterns of five individual probes representing *Agpslzm/L* gene. The insets of each graph display the number of mismatches in each of the 60-mer probes from the paralogous *Bt2*.

5.5 Comparison of Co-expression Networks

The direct comparison between expression data from microarray and RNA-seq technologies allowed us to assess the extent of differences between expression levels as measured by each platform. The comparison of co-expression networks built individually based on the data from each platform will tell us whether a gene retains the same neighbors in terms of expression profile similarity. Since the co-expression profile of a gene consists of correlation coefficients between the gene's expression profile and expression profiles of all other genes, the small variations in expression data between the platforms will compound making the differences between co-expression networks much more pronounced than the differences between expression data.

5.5.1 Similarities and differences in RNA-seq and microarray co-expression networks

To assess how the profiling platform affected network properties, we generated two co-expression networks based on the RNA-seq and microarray transcriptome profiles from 18 samples. The profiles encompassed 19,328 FGS genes that demonstrated detectable expression in the microarray data set and simultaneously registered mapped reads in at least one RNA-seq sample. The two expression profiling platforms have different dynamic ranges, which can complicate comparisons of the data. Therefore, we applied log_2 transformation to the microarray data but transformed the RNA-seq data using an inverse hyperbolic sine function, which allowed for greater compression of the large values present in RNA-seq data. Both co-expression networks were built from the microarray and RNA-seq data averaged across biological replicates in 18 tissues.

We observed a moderately high global correlation between the two networks ($R = 0.75$) rising from a multitude of identical or nearly identical co-expression relationships (Figure 5.7a). Yet, this correlation did not reach the levels we observed between the co-expression networks constructed from two biological replicates of microarray data ($R = 0.86$) or RNA-seq data ($R = 0.90$). The result can be partially explained by a large group of gene pairs with extremely high correlation ($R \approx 1$) in the RNA-seq network that manifested a wide range of values in the microarray

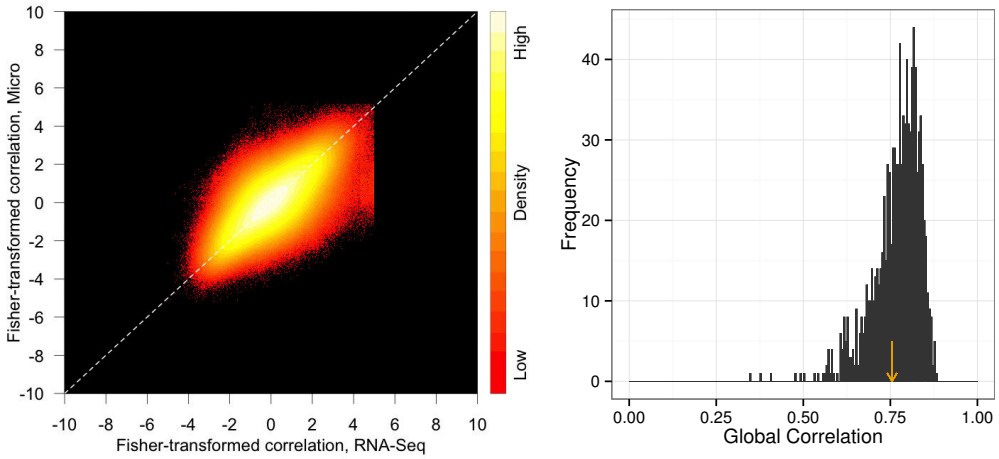


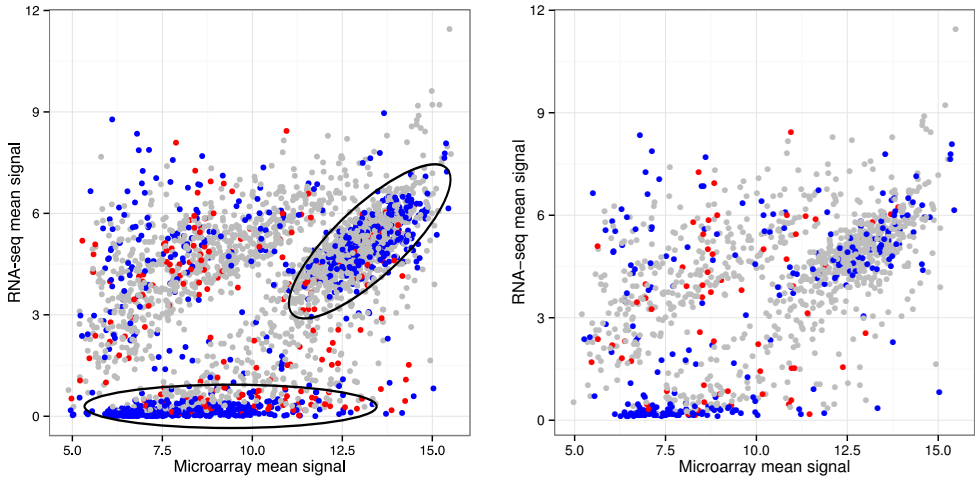
Figure 5.7. Comparison of RNA-seq and microarray co-expression networks. (a) The density of Fisher-transformed and normalized edge weights is shown for both the microarray (y-axis) and RNA-seq (x-axis) co-expression networks. (b) The frequency of network correlation coefficient values for a series of 1000 random co-expression network pairs is plotted relative to the observed value (orange arrow). The random co-expression networks were generated by selecting a mixture of RNA-seq and microarray data for each of the two networks.

network (the region below and to the right of the dashed line in Figure 5.7a). The top 1,000 genes that most frequently appeared in those pairs exhibited significantly lower mean expression in the RNA-seq data set compared to the rest of the expressed genes (Mann-Whitney U test, $p < 1e-55$). Mean expression may be low when a gene is not actually expressed but erroneously registered some reads in a few tissues. Such a gene would strongly correlate with other genes exhibiting a similar expression pattern. Thus, many co-expression relationships formed by the genes with near perfect correlation are likely to be false positives. We repeated the calculations after removing 841 genes that failed to generate FPKM > 5 in at least one tissue and found the global correlation between RNA-seq and microarray co-expression networks to be slightly higher at $R = 0.78$ (not shown). Therefore, additional validation or filtering may be necessary to confirm significant co-expression relationships among genes with very low expression as reported by RNA-seq. The problem does not manifest itself in microarray-based networks likely because the background noise dilutes the low intensity signal considerably reducing spurious correlations.

Further, we evaluated the effects of combining the data from RNA-seq and microarray platforms on the construction of co-expression networks by analyzing 1,000 network pairs generated from a random mix of RNA-seq and microarray data. Each network was based on the expression data that came from the same 18 tissues. However, to compute the first network, the expression data for each tissue was randomly chosen from either platform while the second network was built from the remaining samples. The global correlation between the RNA-seq and microarray co-expression networks fell within the distribution of global correlation coefficients between the networks in each pair (Figure 5.7b). The result suggests that a mixture of RNA-seq and microarray expression profiles can still yield robust co-expression networks.

We explored the similarity between co-expression profiles of individual genes by assessing expression conservation (EC) scores. An EC score indicates whether a gene maintains co-expression relationships with the same neighbors in two different networks. High EC score of a gene signifies that the gene's expression profile is similar to the same group of genes in both networks. We computed EC scores for all genes present in the RNA-seq and microarray co-expression networks. These scores appeared reasonably high for the majority of genes (82.6%) indicating that each of those genes possess roughly the same neighbors in both networks. However, we also found 3,354 genes with significantly low EC scores ($p < 0.01$).

We examined the genes with significantly low EC scores to characterize the factors that potentially lead to variations in co-expression patterns between the two platforms. While the genes with retained duplicates in the two sub-genomes are significantly overrepresented among the genes with low EC scores, we did not find significant enrichment for genes from one of the two sub-genomes (Schnable et al., 2011). Based on the mean expression levels, genes with significantly low EC scores formed two large clusters with unique features (Figure 5.8a). Genes from one of the clusters exhibit very high mean expression levels on both platforms but the individual expression levels in some tissues fall at or near the limit of the microarray's dynamic range. Hence, RNA-seq platform can potentially differentiate those genes better by providing more precise expression measurements due to the improved dynamic range of the platform. The other group of genes exhibits remarkably low mean expression in RNA-seq platform but a wide range of values in microarray. For most of



(a) FPKM > 0 in at least one tissue **(b)** FPKM > 5 in at least one tissue

Figure 5.8. Comparison of expression profiles for individual genes in RNA-seq and microarray co-expression networks based on expression conservation. The color coding indicates relative connectivity: red genes have more connections in the microarray network, blue genes have more connections in the RNA-seq network and grey indicates similar connectivity in both networks. The mean expression level in microarray samples (x-axis; log_2 transformed) and RNA-seq samples (y-axis; inverse hyperbolic sine transformed) were compared. RNA-seq expression data was filtered with a lenient filter (a) and a more strict filter (b) producing 3,354 and 1,221 genes with significantly low EC score ($p < 0.01$) respectively.

these genes, median expression across RNA-seq samples is zero implying that these genes are not expressed in over half of the tissues. In addition, these genes tend to have increased connectivity (the number of highly correlated neighbors) in the RNA-seq network, which induces the decline of EC scores. We applied a more stringent criterion requiring FPKM > 5 in at least one tissue to remove 751 genes with low expression in RNA-seq platform. After rebuilding the co-expression networks, we observed a noticeable reduction in the number genes with low EC scores (Figure 5.8b). This result further supports our recommendation to exercise caution when performing co-expression analysis of genes with low expression levels in RNA-seq platform.

5.6 Conclusions

We analyzed RNA-seq expression data obtained from a subset of samples that were previously used to compile a maize atlas of microarray expression data (Sekhon et al., 2011). We showed that RNA-seq and microarray expression data are highly correlated. However, RNA-seq can better distinguish the genes with high sequence similarity. The differences between the data sets can be partially explained by the better dynamic range of the RNA-seq platform. The global correlation between the co-expression networks generated separately from microarray and RNA-seq data was lower than the correlation between the co-expression networks built from two replicates of RNA-seq or microarray data. Nevertheless, the majority of genes exhibited high expression conservation between the microarray and RNA-seq co-expression networks. Genes with significantly low expression conservation score were enriched in genes with very low mean expression in the RNA-seq data set and a range of values in microarray data set. The result suggests that close attention is required when analyzing genes with low expression in RNA-seq platform.

Chapter 6

Differential DNA Methylation Analysis

6.1 Chapter Overview

In the previous chapters, we primarily focused on the analyses of expression data. One of the factors that controls gene expression is DNA methylation. Despite being an epigenetic mark, the presence of DNA methylation may depend on local or remote genetic variation. Thus, it may be possible to predict methylation levels in some regions from the DNA sequencing data. In addition, DNA methylation controlled by genetic factors tends to be more stable across generations than purely epigenetic methylation. In this chapter, we introduce a pipeline for the identification of differentially methylated regions (DMRs) in a group of lines from a single species. We use this pipeline to uncover DMRs among 51 diverse maize and teosinte lines. Some of these DMRs are significantly associated with local SNPs, which introduces a possibility to predict methylation levels in those regions. Other DMRs seem independent from local genetic variation but may still be influenced by remote genetic features. We also find that methylation in some of the discovered DMRs strongly influences the expression of nearby genes.

Portions of the work presented in this chapter were published in (Eichten et al., 2013a). The chapter includes contributions from Steven Eichten, Jawon Song, Qing Li,

Ruth Swanson-Wagner, Peter Hermanson, Amanda Waters, Evan Starr, Patrick West, Peter Tiffin, Chad Myers, Matthew Vaughn, and Nathan Springer. Steven, Matthew, and Nathan conceived and designed the experiments. Steven, Qing, Ruth, Peter H., Amanda, Evan, and Patrick subsequently conducted the experiments. Jawon and Matthew contributed the association analysis between DMR methylation levels and local SNPs. Steven, Qing, Peter T., and Nathan worked on the biological interpretation of the results. Nathan and Chad supervised the project.

6.2 Role of DNA Methylation in Plants

DNA methylation is an epigenetic regulatory mechanism that plays an important role in genomic imprinting, suppression of transposons, and regulation of gene expression (Bird, 2002). Many DNA methylation variants, known as epialleles, have been identified including *HPT* in *Arabidopsis thaliana* (Mittelsten Scheid et al., 2003), *Cnr* in tomato (Manning et al., 2006), and *CmWIP1* in melon (Martin et al., 2009) among many others. Since epigenetic knowledge may potentially improve phenotype prediction and breeding methods, there is a growing interest to study epigenetic variation in plants (Springer, 2013).

Variation in DNA methylation among genotypes has been assessed on the genome level in several plant species including *Oryza sativa* (rice) (He et al., 2010) and *Glycine max* (soybean) (Schmitz et al., 2013a). It has been shown that variation in methylation of a single cytosine residue occur much more frequently than single nucleotide polymorphisms (SNPs) but changes in the methylation pattern of a region have frequency comparable to that of a SNP (Becker et al., 2011; Schmitz et al., 2011). Even though statistically significant associations between methylation state of a single cytosine residue and phenotype have been reported previously (Xu et al., 2007; Moser et al., 2008), the majority of methylome-wide studies focus on the analysis of differentially methylated regions (DMRs) whose size ranges from a few hundred to several thousand base pairs (Bock, 2012).

Stability of methylation levels varies from region to region. In some cases, DNA methylation levels may occasionally change after just a few generations (Regulski et al., 2013). Methylation linked to genomic variation is expected to be considerably

more stable. In *Arabidopsis*, DMRs have been shown to associate frequently with local (*cis*) and occasionally distant (*trans*) SNPs Schmitz et al. (2013b). Research in maize identified the presence of purely epigenetic variation in DNA methylation (Eichten et al., 2011) as well as spreading of DNA methylation around transposon insertion sites (Eichten et al., 2012). Due to inherent associations between genetic and epigenetic variation, it is necessary to analyze both phenomena jointly at each particular locus in order to completely decipher their effects on phenotype.

In particular, Richards (2006) suggests classifying DNA methylation into three groups depending on the influence of the genetic factors. *Obligatory* epialleles include the cases when a particular genetic alteration such as structural variation or a transposon insertion strictly leads to the changes in DNA methylation. The state of *facilitated* epialleles is determined by genotype in a probabilistic manner. Finally, the chromatin state in *pure* epialleles is independent from genetic alterations. There are two main reasons why this classification of epialleles is important. SNPs in linkage disequilibrium (LD) with the methylation causing structural variation can potentially be used to predict the state of the obligatory epialleles. Such prediction would not be possible for facilitated and pure epialleles. In addition, facilitated and pure epialleles are likely to show reduced stability with higher chances to revert to the original state in subsequent generations. Obligatory alleles, on the contrary, would tend to maintain their state throughout many generations because genomic features that determine the state are likely to be inherited.

Maize is a diverse organism (Buckler et al., 2006; Chia et al., 2012) with a large number of transposable elements interspersed with genes (Rabinowicz and Bennetzen, 2006; Schnable et al., 2009) and, therefore, represents an opportune model for studying epigenetic variation. We developed a pipeline for the identification and characterization of DMRs and employed it to analyze epigenetic variation in 51 divergent maize inbred lines. We identified several thousand DMRs and validated them with MethylC-Seq (Lister et al., 2008) data. Some of these DMRs were strongly associated with local genetic variation. We also uncovered over 300 genes whose expression significantly correlated with methylation level of a neighboring DMR.

6.3 Pipeline for the Identification of DMRs

The methylation state of a single position tends to be relatively unstable and prone to spontaneous mutations (Becker et al., 2011; Schmitz et al., 2011), which may be explained by lower effectiveness of DNA methyltransferases responsible for the methylation maintenance (Genereux et al., 2005). While there have been reports associating individual differentially methylated positions (DMPs), also known as SMPs, with phenotype (Xu et al., 2007; Moser et al., 2008), the majority of genome-wide methylation studies focus on the analysis of differentially methylated regions (DMRs) that may extend from several hundred to several thousand base pairs (Bock, 2012). Within a region, methylation patterns appear more stable and the rate of methylation variation is comparable to the rate of genetic mutations (Becker et al., 2011; Schmitz et al., 2011). Therefore, DMR identification is considered to be an integral part of genome-wide methylation analysis (Bock, 2012; Hansen et al., 2012).

Several previous studies defined DMRs by grouping consecutive SMPs found based on Fisher's exact test (Becker et al., 2011; Lister et al., 2011) or on a consensus among biological replicates with sufficient coverage (Schmitz et al., 2011; Hodges et al., 2011). Other approaches include more advanced algorithms such as QDMR that relies on Shannon entropy (Zhang et al., 2011) and BSmooth that leverages the information about biological variability (Hansen et al., 2012). However, these algorithms generally target case-control studies where all available samples can be divided into two groups. When a study encompasses multiple accessions, samples would still form two groups but the membership in these groups would vary from region to region. In such cases, it may be advantageous to use alternative segmentation algorithms such as those employed for the detection of copy number variation (CNV) in DNA.

A common method for CNV detection is array-based comparative genomic hybridization (array CGH) that involves measuring the intensity difference between sample and reference DNA by hybridizing them to a microarray (Pinkel et al., 1998; Pollack et al., 1999). The output format of the array CGH analysis would be similar to contrasting methylation levels in each accession against a pre-selected reference

accession. Thus, an array CGH segmentation algorithm such as DNAcopy (Venkatraman and Olshen, 2007) can be used to detect contiguous regions with significant methylation differences between an accession and the reference. These segments can be subsequently summarized across accessions to form candidate DMRs.

Rather than performing segmentation on one accession at a time, several methods allow for simultaneous segmentation of multiple accessions. These methods engage a variety of techniques such as dynamic programming (Picard et al., 2005), wavelet decomposition and thresholding (Ben-Yaacov and Eldar, 2008), and sparse group selection on fused lasso components (Tian et al., 2012) among many others. When using these methods, the summarization step is no longer necessary but their accuracy relative to DNAcopy in the DNA methylation context is unknown. Comparative evaluation of the segmentation methods is beyond the scope of this work.

Within each DMR, we expect accessions to form two unambiguous groups, one with normal methylation levels and another one with deviating methylation levels that could constitute either hypermethylation or hypomethylation. For instance, MeDIP-chip analysis usually reports a log-transformed ratio between hybridization efficiency of DNA enriched by immunoprecipitation and unenriched input DNA. Thus, methylation signal values in two groups from an ideal DMR candidate should have means with opposite signs and small variances. In addition, the difference between the means should be as large as possible. The last two criteria are incorporated into Fisher's criterion for Linear Discriminant Analysis (Fisher, 1936) defined as

$$J = \frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2}$$

where μ_1 and μ_2 are means and s_1^2 and s_2^2 are variances of each group respectively. Since the number of accessions would be relatively small in most cases, we can use Fisher's criterion to test each possible threshold between two neighboring points within a DMR to find the best separation. In a sense, this approach is similar to the application of the Expectation Maximization (EM) algorithm (Dempster et al., 1977) to this problem, except instead of maximization step we perform an exhaustive search of the threshold space. DMR candidates for which Fisher's criterion exceeds certain threshold may be selected for further analyses.

The accuracy of Fisher's criterion may be compromised when one of the groups has fewer than three members. In that case, the variance estimation for the smaller group may lack precision. Therefore, it may be necessary to re-examine DMR candidates that failed Fisher's criterion test and select those with one or two distinct outliers. These outliers would form one of the two groups we are seeking in a DMR. While automatic detection of outliers remains a challenging problem, several statistical tests are available for the task (reviewed in Hodge and Austin, 2004). With maize methylation data, we employed Grubbs' test (Grubbs, 1969) for its simplicity and ability to work with single dimensional data. The statistic for one-sided Grubbs' test is given by either

$$G = \frac{X_{max} - \bar{X}}{s}$$

or

$$G = \frac{\bar{X} - X_{min}}{s}$$

to test the maximum or the minimum value respectively; $\bar{X}$ and s denote the mean and standard deviation of the sample. The hypothesis that the value is not an outlier should be rejected at significance level α if

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/N,N-2}}{N-2+t^2_{\alpha/N,N-2}}}$$

where $t^2_{\alpha/N,N-2}$ is a critical value of the t distribution with significance level α/N and $N-2$ degrees of freedom. The p-value can be determined by expressing t in terms of G and N, i.e.

$$\hat{t} = \sqrt{\frac{G^2 N(N-2)}{(N-1)^2 - G^2 N}}$$

that leads to

$$p = N(1 - T(\hat{t}))$$

where T is the t distribution density function for $N-2$ degrees of freedom.

Once DMRs have been identified, we can perform association analyses to detect relationships between methylation and expression as well as genetic variation. Since outlier tests often possess weak statistical power, it may be advantageous to keep

DMRs identified with Fisher's criterion separate from DMRs uncovered by an outlier test. Because the relationship between methylation and expression may not be linear, a rank based approach such as Mann-Whitney U test or Kendall's τ would be most appropriate for finding associations between the two. To find association between DMRs and genetic variation, a two-tailed *t* test between a SNP call and methylation values in the DMR can be performed. Due to large number of tests required in both cases, it is imperative to correct for multiple testing.

6.4 Identification of DMRs in Maize Populations

This section describes each step of the proposed pipeline in more detail using methylation, expression, and genomic variation data from a diverse group of maize accessions.

6.4.1 Methylation and Expression Profiling

Biological Materials

Randomized block design was used to grow three replicates for each of the 20 NAM parental genotypes and one replicate for the additional 31 maize and teosinte genotypes from the Maize HapMap2 project (Chia et al., 2012). Only one replicate per genotype was allowed within a block. Each replicate included four seedlings grown in a single pot positioned randomly within the block. Controlled conditions with 15 h of light, 9 h of darkness, and daily watering were maintained for all blocks at the University of Minnesota Agricultural Research station, Saint Paul, MN. Samples from the third leaf (L3) were collected from each plant after 18 days of growth. The samples were either pooled with other plants in the same pot (replicate) or kept independently and immediately frozen in liquid nitrogen. In accordance to the CTAB procedure (Doyle and Doyle, 1987), DNA was isolated from the frozen samples as described in Eichten et al. (2011). Trizol (Invitrogen) was used for RNA isolation according to the manufacturer's protocol.

meDIP-chip Analysis

Methylated DNA immunoprecipitation (meDIP) analysis followed the pipeline described in Eichten et al. (2011). Briefly, methylated DNA was separated from 400 ng of sonicated DNA via immunoprecipitation using an anti-5-methylcytosine monoclonal antibody from the Methylated DNA IP Kit (Zymo Research.) Negative control included B73 sonicated DNA treated with water rather than monoclonal antibody. The Whole Genome Amplification kit (Sigma-Aldrich) was used for whole-genome amplification of 50 to 100 ng of immunoprecipitated DNA and the same amount of sonicated DNA (input control) per replicate. Following the array manufacturer's protocol (Roche NimbleGen Methylation User Guide v7.0), 3 μg amplified immunoprecipitated DNA from each sample was labeled with Cy5 from the Dual-Color Labeling Kit (Roche NimbleGen). Sonicated DNA from the input control samples was labeled with Cy3. Depending on the sample set, the samples were hybridized to the custom 2.1M, 1.4M, or 270K probe arrays for 16 to 72 hours at 42°. The 1.4M chip was designed to have probes at approximately 200 bp intervals covering the regions with low copy-number levels. The 2.1M chip had higher probe density but also contained all the probes from the 1.4M chip. The 270K chip had a subset of probes from the 1.4M chip and was used only for validation. Slides were washed and scanned as prescribed by NimbleGen's protocols for the GenePix4000B (2.1M platform) and NimbleGen MS200 (1.4M and 270K platforms) scanners. Using NimbleScan software (Roche NimbleGen), the images were aligned and quantified to generate pairs of raw intensity readings for each probe on the array.

The raw array data from 2.1M and 1.4M platforms were imported into R statistical environment for further processing with Bioconductor libraries (Gentleman et al., 2004). The data generated on the 2.1M platform were reduced to match the 1.4M platform. Samples with a single replicate were cloned two times to allow for the use of the same normalization methods with all samples. Analytical weights were set to zero for all non-maize and vendor-supplied control probes. Array-specific effects were mitigated by variance-stabilizing normalization. Limma package (Smyth, 2004) was used to derive hybridization coefficient estimates by fitting fixed linear model while accounting for dye and sample effects. Methylation values (signal) were reported as log_2-transformed ratio between the hybridization efficiency of DNA enriched by

immunoprecipitation and unenriched DNA, $log_2(IP/input)$.

Bisulfite Sequencing

Following the procedure outlined in Schmitz et al. (2011), whole-genome bisulfite sequencing was performed on 14d old whole-seedling DNA isolated from B73 and Mo17 inbread lines. The plants were grown independently from the seedlings used in meDIP analysis. Following the manufacturer's protocol, 500 ng fragmented DNA with ligated TruSeq-methylated adapters was subjected to bisulfite conversion using the MethylCode bisulfite conversion kit (Life Technologies). After conversion, the DNA was partitioned into four reactions, amplified using Pfu Turbo Cx DNA polymerase (Agilent) for four cycles, and pooled. Paired end sequencing was performed on the HiSequation 2000 (Illumina) for 100 cycles. Poor quality and incompletely converted reads were discarded. Good quality reads were aligned to the B73 reference genome v2 (Schnable et al., 2009) using the Bismark aligner v0.7.2 (Krueger and Andrews, 2011) with two maximum mismatches (parameter "-n 2") and the seed length of fifty (parameter "-l 50"). Default parameters were used to identify the positions of methylated cytosines in the aligned reads using Bismark methylation extractor. Methylation level within each DMR was estimated as average methylation from intersecting 100 bp windows using BEDTools (Quinlan and Hall, 2010).

RNA-seq and Expression Analysis

RNA was isolated from the same seedling L3 leaf samples used for meDIP profiling of 50 maize genotypes. The libraries were constructed at the University of Minnesota Genomics Center in compliance with the TruSeq library creation protocol (Illumina). The sequencing was performed on the HiSequation 2000 to yield 8 to 24 million reads per replicate. Poor quality reads were removed with CASAVA software package (Illumina). Reads were mapped to the B73 reference genome v2 (Schnable et al., 2009) and transcript abundance was estimated with TopHat (Trapnell et al., 2009) using default parameters. The number of reads per kilobase of exon per million fragments mapped (RPKM) were calculated with "BAM to Counts" application in the iPlant Discovery Environment (`http://www.iplantcollaborative.org`) using the version

5b.60 of the maize genome working gene set (`http://ftp.maizesequence.org/`).

6.4.2 Segmentation and Summarization

Segment discovery was performed on a set of 20 genotypes that had three replicates profiled on the 2.1M array. After removing the probes with poor comparative genomic hybridization (CGH) between B73 line and other HapMap (HM) lines reported by Swanson-Wagner et al. (2010), we ran DNAcopy algorithm Venkatraman and Olshen (2007) individually on each set of the contrast values (B73 - HM line) to identify multi-probe segments exhibiting similar patterns of differential methylation between B73 and the other genotypes. We excluded the segments that displayed less than two-fold difference between B73 and the other genotypes, i.e. log_2 values between -1 and 1, to obtain 14,230 segments across the 20 lines used for the initial discovery. Some of the segments were extremely long, well exceeding the DMR lengths reported previously in arabidopsis (Schmitz et al., 2013b) and maize (Eichten et al., 2011). Therefore, we split a segment whenever the neighboring probes were separated by more than 700 bp. The boundaries of the parts were determined by probe positions. If any of the resulting parts had only one probe left, the part was discarded. The splitting procedure increased the number of segments to 18,936 while reducing the maximum length from 401,591 bp to 7,264 bp and keeping the minimum length at 168 bp.

We summarized the segments across the genotypes to obtain segment coverage for each position. Each summarized segment consists of ranges indicating how many genotypes have a segment that spans the range. Two neighboring ranges always have different genotype counts, otherwise they would have been merged into a single range. Summarization yielded 12,650 ranges of variable length. For each range, we calculated the ratio between the range's genotype count and the segment's maximum genotype count. Low ratio would indicate the lack of support for differential methylation within the range. Therefore, we discarded all the ranges with the ratio below 0.6 leaving 10,893 ranges.

After filtering, consecutive ranges were combined into 9,899 segments spanning two or more probes. The majority of segments were between 200 and 2,000 bp long with a few of them reaching over 5,000 bp (Figure 6.1). This is comparable to

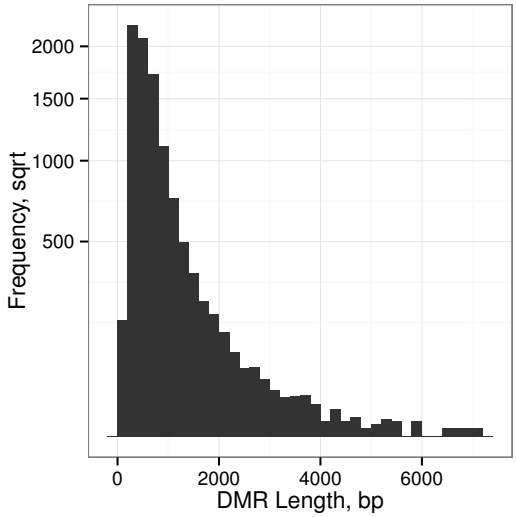


Figure 6.1. Distribution of segment size.

the size of DMRs previously reported in maize (Schmitz et al., 2013b,a). In some cases, a range in the middle of a segment had a ratio below the threshold causing the formation of 10 segments that contained only a single probe. Those segments were excluded from further analyses.

6.4.3 Classification of Genotypes Within Each Segment

Common Variants

In addition to 20 genotypes used for segment discovery, we profiled a single replicate of 31 HM lines to improve our estimates of the frequency of DNA methylation variation. For each of the 51 genotypes, we calculated the mean signal in each segment by averaging the segment's probe values. We used the distribution of those mean signal values to classify the genotypes within each segment. Since we expected the genotypes to form two disparate groups with one of them being either hypomethylated or hypermethylated compared to the other, we tried to select the segments with clear separation between two classes based on Fisher's criterion for

Linear Discriminant Analysis (Fisher, 1936) defined as

$$J = \frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2}$$

where μ_1 and μ_2 are means and s_1^2 and s_2^2 are variances of each class respectively. Since the criterion calculation is very simple and we only have 51 genotypes, we can test each possible separation by placing a threshold midway between two neighboring data points. For the actual classification, we choose the threshold that yields the highest value of J. After applying that threshold to split the data points into two groups, the mean and standard deviation are estimated for each of the groups to define the corresponding distributions. Finally, we calculate the density for each data point under both distributions. A genotype is attributed to the class whose distribution yields the higher density (see Figure 6.2a for an example).

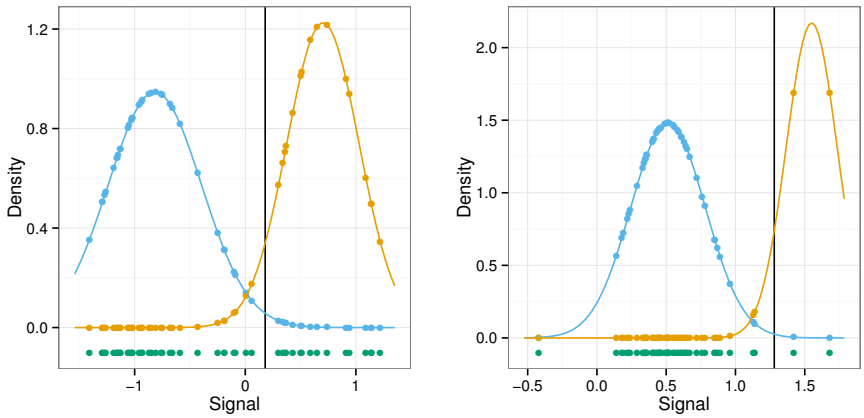


Figure 6.2. Genotype classification in sample DMRs. Black vertical line indicates the optimal threshold selected by maximizing Fisher's criterion for Linear Discriminant Analysis. Parameters for the blue and orange distributions are based on the signal of points that lie to the left or to the right of the threshold, respectively. Green points indicate the signal of each genotype and are displayed to help visualize the difference in signal between genotype. Density under each distribution was calculated for all genotypes (orange and blue points). (a) Sample DMR with a good separation between classes, i.e. the linear discriminant is high, the distribution means have the opposite signs, and the distance between the distribution means is sufficiently large. (b) Maximization of Fisher's criterion causes suboptimial solution in the presence of an outlier. A better placement for the threshold would be around 0.

In most cases, the threshold criterion and the density criterion produce identical classification. However, some data points have non-zero density under both distributions. If the densities are very similar, one can argue that their classification is ambiguous (two points immediately to the left of the threshold in Figure 6.2a). To filter such points, we calculate log odds ratio for each genotype in each segment.

$$L = \log \frac{\rho_1}{\rho_2}$$

where ρ_1 and ρ_2 indicate the density of the genotype's value under each distribution. A point was considered ambiguous and, therefore, unclassified whenever $\rho_1 > 0$, $\rho_2 > 0$, and $|L| < 3$. Since these ambiguous points affected the value of J, it was recalculated based on new distribution parameters derived from the filtered groups.

There are several possible complications with this method. First, the variance cannot be estimated accurately for groups with fewer than 3 genotypes (see Figure 6.2b for an example). In those cases, we estimated the variance as the smaller of the following two values: the variance of three smallest values and the variance of the three largest values. However, when the data points are clustered closely together on one side of the distribution and there is an outlier on the other side, the threshold separating the outlier from the rest of the data points produces very high and possibly biased J value. To avoid this problem, we drop one extreme value on each side of the value distribution in a segment before searching for the optimal threshold. After finding the optimal threshold, we calculate the density for each data point including the extreme values omitted before. Then, we filter the values and recalculate J as described above. While this approach did not resolve the problem completely, it considerably reduced the number of segments where outliers potentially caused problems. We will describe the approach to capture DMRs with outliers later in this section.

We used an arbitrary threshold of $J_{adj} > 8$ as a criterion for selecting common DMRs with at least three genotypes in each class from the set of summarized segments. To improve the selection, we applied two additional criteria. First, we required that at least one class mean in each segment fell within the [-0.8; 0.8] segment. This criterion has biological significance as we expect to have one "normal" class and

a class that deviates from "normal". Second, the absolute difference between the means should be at least 1. While the distance between the distribution is already incorporated into the Fisher's criterion, we wanted to strengthen it even further by imposing the minimum distance between the means. Overall, 1,966 segments satisfied all the criteria of a common DMR.

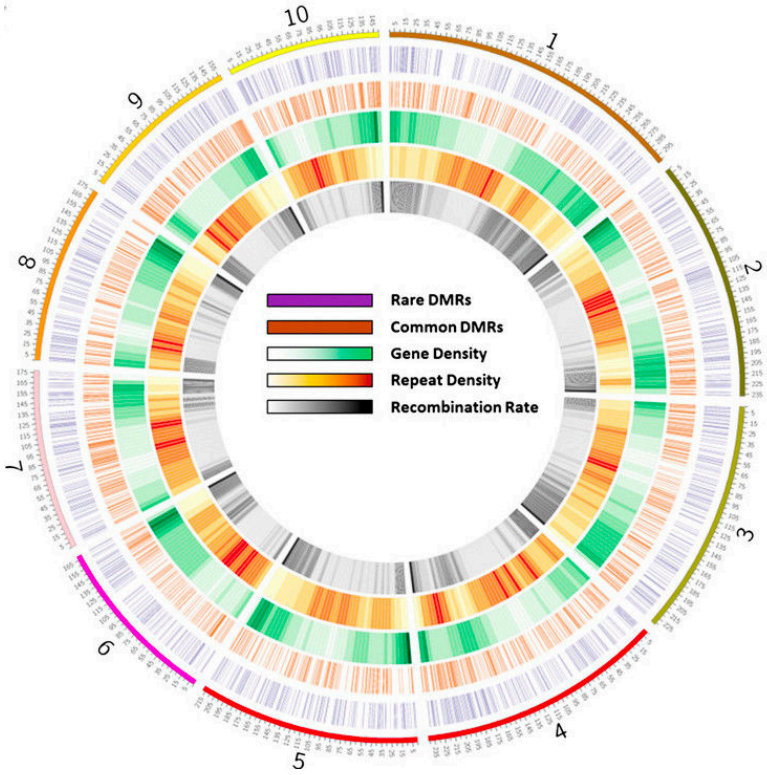


Figure 6.3. Distribution of DMRs across maize genome.

Rare Variants

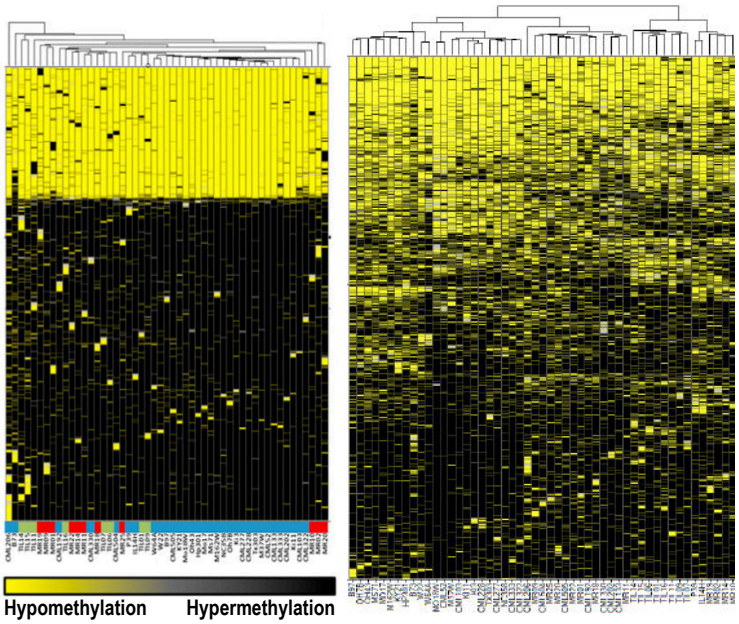
To identify the rare variants, we used the same criteria as for the common variants except one of the classes should have had at most 2 genotypes. In total, 1,213 segments satisfied the criteria. However, we removed an extreme point from each side of the distribution before searching for an optimal threshold with Fisher's criterion. As a result, we potentially missed the segments with single outliers. To be as inclusive as possible, we selected additional rare variants based on the results of Grubbs'

test for outliers (Grubbs, 1969). Since the test assumes normal data distribution, we first removed segments that failed to pass the Anderson-Darling normality test at relatively permissive 0.01 significance level. The same significance level $\alpha < 0.01$ was used for the Grubbs' test yielding 541 additional segments. Along with rare variants identified through Fisher's criterion, we obtained 1,754 rare DMRs.

6.4.4 DMR Characterization and Validation

The distribution of the common and rare DMRs across the maize genome did not reveal any specific pattern (Figure 6.3). However, the majority of them (over 78%) populate low-copy intergenic regions and only 798 out of 3,720 DMRs intersect with annotated genes. Hierarchical clustering of the genotype methylation levels within rare (Figure 6.4a) and common (Figure 6.4b) DMRs did not reveal any genotypes with unique DNA methylation profiles. However, the relationships among the genotypes were essentially similar to the relationships determined from the SNP data in the Maize Hapmap2 study (Chia et al., 2012). Interestingly, hypomethylation is significantly more prevalent ($p < 0.001$) as the minority state in rare DMRs than hypermethylation (Figure 6.4a). It suggests that for a small number of lines it is easier to achieve persistent methylation loss than persistent methylation gain.

To validate the obtained DMRs, we applied bisulfite treatment to the samples from independently grown B73 and Mo17 lines and resequenced them using methylIC-seq pipeline (Lister et al., 2008). In addition to validation, methylIC-seq analysis allowed us to investigate the sequence context of differentially methylated cytosines. Sequencing reads covered more than 80% of the region in 248 out of 878 DMRs with differing methylation state between B73 and Mo17. Among those 248 DMRs, the majority (91%) displayed considerable divergence (over 50%) in CG and/or CHG methylation (Figure 6.5), thus confirming 92% and 89% of common and rare DMRs respectively. Since the proportions of the validated common and rare DMRs are very close, the approaches we have taken for DMR identification appear to be reasonable. Further analysis showed that a large percent (84%) of confirmed DMRs manifested differences in both CG and CHG contexts while CG-only and CHG-only divergence was observed in 9% and 7% respectively. Methylation in CHH context was very rare, less than 10% of all sequences, and we did not detect any substantial variation between



(a) Rare DMRs **(b)** Common DMRs

Figure 6.4. Hierarchical clustering of methylation levels across (a) rare and (b) common DMRs. None of the genotypes exhibit a unique methylation pattern across the DMRs.

the two genotypes in that context.

6.5 Associations between Genomic Variations and DMRs

Even though DNA methylation is considered to be an epigenetic mark, it has been shown that genetic changes may lead to alteration of methylation patterns (Law and Jacobsen, 2010; Hollister et al., 2011; Eichten et al., 2012). To test whether the variations in DNA methylation levels of individual genotypes associate with DNA sequence polymorphisms, we performed a local association scan of the DMR enclosing loci using 56 million SNPs reported by the HapMap2 study (Chia et al., 2012). Because our data set consisted of the limited number of genotypes (51), a genome-wide scan would lack statistical power. Therefore, we constrained our search to the SNPs located either inside the DMR or within 1kb of its boundary. We also omitted the rare DMRs from this analysis because the rare state of fewer than three genotypes would

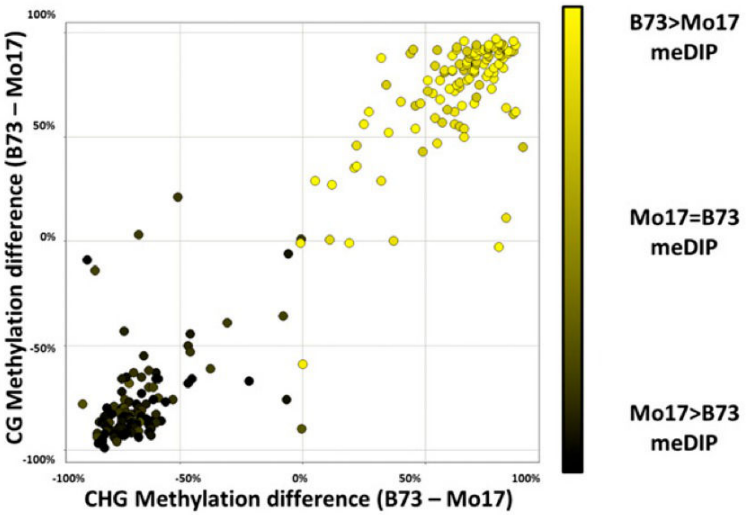


Figure 6.5. DMR validation by methylIC-seq analyses of the samples from independently grown B73 and Mo17 lines. Among the identified common and rare DMRs, 248 simultaneously showed methylation state differences between the two lines and had at least 80% of the length covered by reads from methylIC-seq. Relative methylation differences in CHG and CG contexts are shown on the X and Y axes respectively. The color gradient encodes the difference between B73 and Mo17 based on the meDIP array data. The meDIP array predictions coincide with substantial differences in both CG and CHG contexts for the majority of DMRs. For a small number of the DMRs, the differences are present only in one of the contexts or are absent entirely.

constrain the statistical power as well. In addition, the 36 DMRs without any SNPs in the neighboring regions among the 51 genotypes were not tested.

We evaluated the association between methylation signal, $log_2(IP/input)$, averaged across a DMR and each unambiguous local SNP call using a two-tailed *t* test. To control the false discovery rate, we randomly selected 100 regions with 1000 SNPs each throughout the maize genome and examined them for random associations with methylation levels in each DMR. Overall, about a half (1,003 out of 1,966) of the common DMRs were significantly associated with local SNPs. We found highly significant SNPs lying inside as well as outside of DMRs. There could be several potential reasons why the other 963 DMRs did not exhibit a significant association with local

genetic stats. First, our selection criteria may be too strict for some of the associations. It is also possible that the causative variation located outside of the tested locus act on methylation levels in-*trans*. Finally, some of those DMRs may be purely epigenetic without any genetic influences.

The presence of significant associations between methylation levels within DMRs and local SNPs indicates that it may be possible to predict methylation state in those regions based on the local allelic state. To investigate this possibility, we profiled 12 additional inbred lines from the maize HapMap2 project (Chia et al., 2012) on a smaller meDIP microarray chip with 270k probes designed exclusively for surveying DMRs. In particular, the chip included all necessary probes to measure methylation levels within 535 DMRs where the most significantly associated SNP exhibited variation across the 12 additional inbreds. For multiple DMRs, the minor allele was present in either one (205 DMRs) or two (111 DMRs) of the 12 inbreds. For the remaining 219 DMRs, the minor allele was detected in three or more inbreds. We compared the allelic state against DNA methylation levels in each of those 219 DMRs. We were able to predict accurately the average methylation levels in the 12 inbreds for the majority (77%) of the DMRs. Thus, DNA sequence information can be a sufficient and reliable predictor of methylation patterns in DMRs exhibiting significant association with genetic variation.

6.6 Associations between DMRs and Gene Expression

Since DNA methylation can cause phenotypic variation by affecting gene expression, we examined the functional effects of the DMRs on adjacent genes. For each rare and common DMR, we located one closest gene on each side of the DMR using the B73 reference genome annotation v2 (Schnable et al., 2009). While two genes were found in most cases (2,925 out of 3,720), some DMRs positioned near chromosome ends had only a single nearest gene. Using 10,000 randomly selected regions as controls, we grouped all the regions by their location relative to the nearest gene. Both common and rare DMRs appeared within 5 kb of the nearest gene much more often than expected by chance but they were less likely to intersect or lie within a gene's coding sequence than the regions from the control group (Figure 6.6). The frequencies

of DMRs to appear near the 5’ end (upstream) or 3’ end (downstream) of a gene were roughly equal indicating that methylation variation can potentially occur in low-copy regions around genes. We examined 2,375 genes located within 10 kb of a DMR using the annotation from Schnable et al. (2012) to find that they were equally likely to come from either sub-genome formed by the whole-genome duplication and that the proportion of the inserted and syntenic genes did not deviate from the background levels (Figure 6.7c). The DMRs also had roughly even distribution across the genome with approximately equal proportion of DMRs lying in high-recombination, gene-rich chromosome arms as in low-recombination, gene-poor central loci (Figure 6.7d).

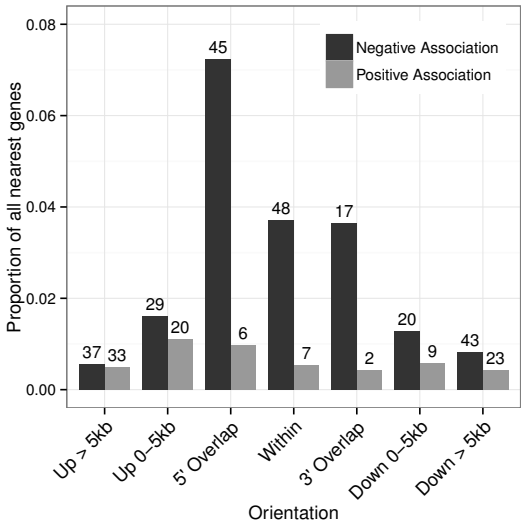


Figure 6.6. DMR location relative to the associated gene. DMRs that are negatively associated with the expression of the adjacent gene tend to overlap with the gene sequence. Positively associated DMRs have almost uniform distribution of the positions relative to the associated gene.

To evaluate the impact of DMRs on the adjacent genes, we measured the transcript abundance for those genes in all genotypes by performing RNA-seq analysis of the tissue samples used for methylation profiling. We assessed the correlation between methylation levels within a DMR and expression levels of the nearby genes using the Mann-Whitney U test and Kendall’s τ. The first test is more appropriate for the cases where the data is not interval scaled, i.e. methylation acts similar to an

on/off switch. The second test can detect quantitative relationship between methylation level and expression. However, the Mann-Whitney U test cannot be applied to rare DMRs due to small number of genotypes in one of the groups. Therefore, we replaced it with a z-score test whereby we estimated the mean and standard deviation for the larger group and calculated the likelihood of the small group's values to come from the same distribution. In all cases we controlled for false discovery rate (FDR) at $q < 0.05$ using Storey's procedure (Storey, 2002).

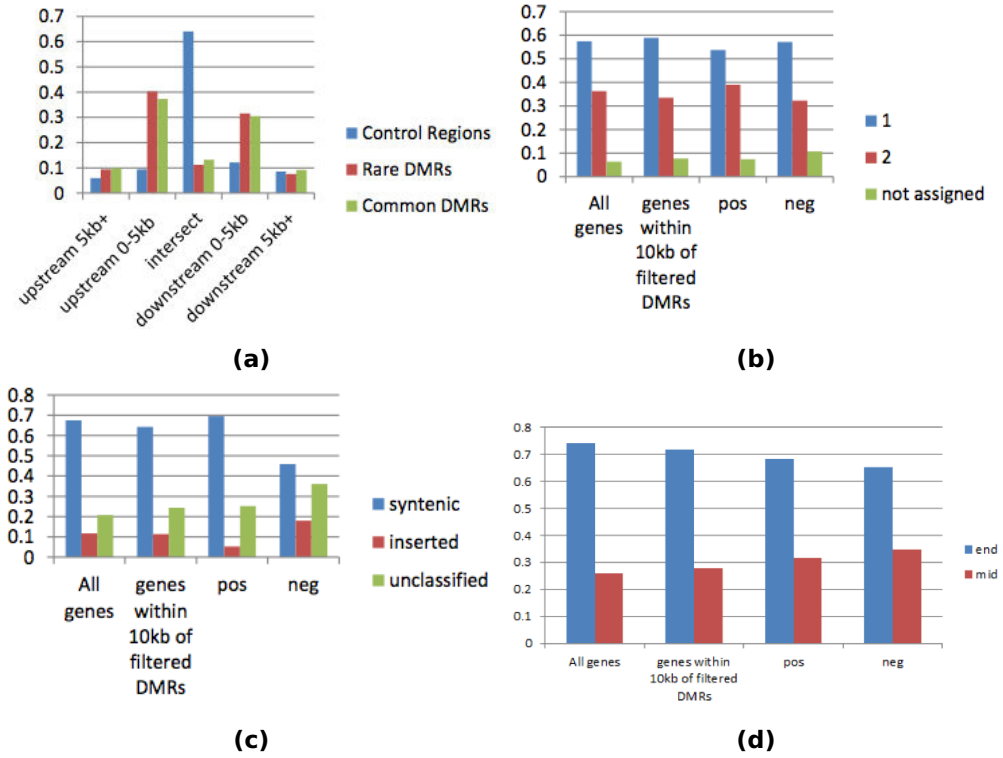
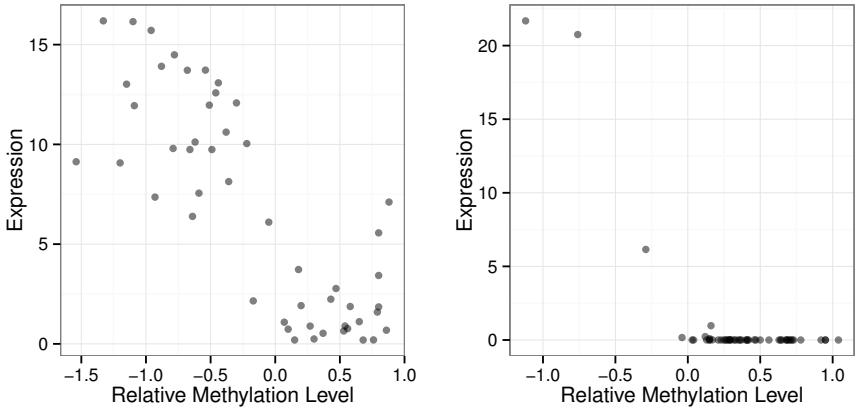


Figure 6.7. Characterization of the DMR-gene relationships. (a) Orientation of the rare and common DMRs relative to the adjacent genes compared to 1,000 randomly selected control regions. Pos and Neg labels indicate the genes with significant positive or negative association between their expression and methylation levels in the adjacent DMR. (b) Frequency of genes from the maize subgenomes among various groups of genes. (c) Frequency of inserted genes and genes from syntenic regions among the various groups of genes. (d) Frequency of genes located in the gene-rich, high-recombination chromosome arms (end) and gene-poor, low-recombination pericentromeric regions (mid) among the various groups of genes.

For the 1,966 common DMRs, the union between the Mann-Whitney U test and Kendall's τ test yielded 277 genes with significant correlation between expression

and DNA methylation. For the 1,754 rare DMRs, the union between z-score and Kendall's τ tests uncovered 111 significant associations. We observed the cases where relative methylation level quantitatively affected gene's expression (Figure 6.8a) as well as the cases where DNA methylation more resembled a switch (Figure 6.8b). DNA methylation generally represses the expression of the nearby genes. As expected, we uncovered predominantly negative associations between expression and methylation levels in both common (70%) and rare (73%) DMRs (Figure 6.6). DMRs with significant negative correlations were enriched for locations near (less than 5 kb) or overlapping gene boundaries (Figure 6.6). Especially striking is the enrichment for locations that overlap with the transcription start site. On the contrary, DMRs with significant positive correlations were almost equally likely to appear near genes as distant (more than 5 kb) from genes. Genes with significant negative correlations exhibited slightly fewer syntenic relationships but were more likely to lack homologs in other grasses and were enriched for inserted sequences (Figure 6.7c). Despite the relatively low proportion of genes conserved in grasses among the genes with significant negative correlations (45%), we can still conclude that at least some of the negatively correlated genes are not erroneously annotated transposons.



(a) DMR 5061 - GRMZM2G103182 **(b)** DMR 7997 - GRMZM5G856812

Figure 6.8. Association between methylation and expression for sample DMRs. (a) Sample common DMR that displays quantitative relationship between methylation and expression of a nearby gene. (b) Sample rare DMR for which the association between methylation and expression resembles an on/off switch.

6.7 Conclusions

We described an approach to locate differentially methylated regions among multiple genotypes of a single organism. We successfully applied this approach to identify over 3,500 DMRs in 51 maize and teosinte lines. We validated the results using methylIC-seq data from two of those lines, B73 and Mo17. Some of the discovered epialleles exhibited strong correlation with local SNPs that could potentially be used to predict methylation levels within the associated DMRs. Those SNPs were likely to be in LD with other genetic features that actually influence the local methylation levels. Other epialleles lacked any associations with local SNPs. Some of those epialleles might be purely epigenetic, which would likely affect their heritability. We also found several hundred genes whose expression was regulated by the neighboring DMRs. As expected, the associations between methylation levels in those DMRs and the expression of the regulated genes were predominantly negative. Some of the regulated genes had homologs in other grasses indicating that they were not misannotated transposons.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In the past, plant and animal breeding relied on phenotypic observations and pure chance to obtain better lines and varieties. Farmers were selecting the largest seeds from the healthiest looking individuals for planting next season. Breeders were choosing two varieties with divergent but desirable traits for hybridization hoping to produce an individual with a union of the parental traits. The process was successful over time but very unreliable and slow. Nowadays, breeders can leverage a cornucopia of genomic and transcriptomic data to target specific genes and pathways that influence the traits of interest. On one hand, the amount of data available these days is astonishing. On the other hand, extracting accurate information from that data is hard and we are still far from complete understanding of biological processes. This dissertation focused on computational techniques and pipelines that could assist with deriving valuable information from expression and epigenetic data. We have also applied these techniques to diverse maize data sets and we hope that our results will be useful to biologists and breeders who work on further improvement of maize.

First, we reviewed co-expression network analysis on the global level and demonstrated its usefulness on maize expression data. Our research uncovered major biologically significant differences between co-expression networks of maize and its wild ancestor teosinte and highlighted the effects of domestication on the maize

transcriptome. We continued with co-expression analysis of individual genes. In particular, we introduced a more sensitive method (Altered Expression Conservation, AEC) to measure conservation of gene's co-expression profiles between two networks. While offering better performance than the standard approach, the method is complementary to differential expression methods and sequence-based methods that could also be used to identify rewired genes.

The maize domestication data set covered closely related lines of the same species and contained relatively few samples. To make sure that the AEC method had wider applicability, we used it to compare co-expression networks of two distant yeast species. To extend it even further, the method was applied to analyze genetic interaction networks in yeast. In both cases, the results produced by AEC method were clearly different from the results returned by the standard method. However, the comparative evaluation of the two methods was hampered by the lack of a well-defined gold standard. We concluded that the analysis of genome sequence variation in the maize domestication context provides the clearest benchmark to date in this area, and this benchmark indicates superior performance of our approach.

The AEC method was also useful for the comparison of co-expression networks derived separately from microarray and RNA-seq data. We discovered that genes with significantly low expression conservation often exhibited low mean expression in the RNA-seq data compared to a range of values in the microarray data set. The most likely explanation is the erroneously aligned reads that cause some genes to highly correlate with each other. This result also suggests that the genes with low expression in RNA-seq platform should be analyzed with caution. The analysis of these two networks also demonstrated that microarray and RNA-seq data can be combined when constructing co-expression networks. This finding is consequential because the plethora of existing microarray data can be used to complement RNA-seq data in future co-expression studies.

It is often beneficial to analyze different types of data in conjunction with one another. For example, domestication genes identified by an independent DNA sequence analysis were valuable for the validation of our AEC method. Expression analysis can be complemented by the analysis of epigenetic marks that in turn may

form relationships with DNA sequence polymorphisms. In the last chapter, we developed a pipeline for the identification of differentially methylated regions (epialleles) and employed it to examine the methylomes of multiple maize lines. In some cases, we discovered strong linkages between epialleles and local SNPs indicating that their state could be predicted from DNA sequence. In other cases, epialleles appeared purely epigenetic and, as a consequence, they were likely to lack stability. We also identified several hundred instances where DNA methylation regulated expression of a nearby gene. This knowledge is important because it may allow controlling phenotype via change to DNA methylation.

7.2 Future Work

7.2.1 Further Investigation of Sample Size Effects

There is abundant evidence that co-expression analysis offers consistent and informative results. For instance, in Chapter 5 we showed that co-expression relationships were largely conserved between the networks derived individually from microarray and RNA-seq data. However, the evidence is predominantly indirect and the relative performance of various EC methods is much harder to measure. Throughout the dissertation, we took full advantage of the expected overlap between the sequence-based and expression-based predictions of genes related to domestication and improvement. However, due to insufficient size of the maize domestication data set, we could only investigate the effects of large sample size in other species or in an entirely different context, neither of which provided a good benchmark (Chapter 4). Moreover, there were other differences that could have also influenced the results in each case. To evaluate the effects of increasing sample size on the method performance, it would be helpful to expand the maize domestication data set by adding expression data from other lines. This would keep other confounding factors fixed and it would help to determine how informative the ranking produced by AEC is and whether a sufficiently large sample size would improve the accuracy of EC scores enough to make AEC redundant. In addition, the investigation of alternative similarity measures may identify an approach that would be less susceptible to the changes

in sample size.

7.2.2 Application of AEC in Other Contexts

In principle, the AEC method can be applied in any context where differential expression analysis is appropriate. In particular, it can be used to find genes with co-expression patterns that vary significantly across tissues or conditions. The comparison of tissue-specific co-expression networks may be helpful for the characterization of genes that control plant development and tissue differentiation. The analysis of condition-specific co-expression networks can elucidate the organization of various pathways that activate only under certain conditions such as drought, extreme temperature change, or bacterial infection.

The AEC method is likely to work well in contexts that are similar to domestication, i.e. whenever genotypes of a species can be divided into two groups based on their phenotype or behavior. For example, several species of bacteria participate in nitrogen or phosphorus fixation by forming symbiotic relationships with plants. Strains that form symbiotic relationships can be contrasted against the non-symbiotic strains to improve the understanding of molecular processes fundamental to symbiosis.

Another example of a context that is similar to domestication would be subgenome comparison. The most recent whole genome duplication occurred in maize several million years ago and well before the domestication event. The majority of gene duplicates have functionally diverged since then. The rewired genes that were identified by our AEC method (Chapter 3) did not exhibit any significant enrichment in genes from any particular subgenome (the results were not reported.) However, a promising future direction would be to examine expression conservation between two subgenomes of a polyploid which only recently, within a few hundred years, underwent polyploidization. Recently formed polyploids tend to lose duplicated genes fairly rapidly but the extent of co-expression differences between two different parental genomes is currently unknown.

7.2.3 Development of Integrative Methods for Identification of Selection Targets

In Chapter 3, we contrasted the results from the application of AEC, Differential Expression, and sequence-based methods to maize. Each method produced a fairly large list of candidate genes but the intersection was much smaller. While the benefit of a smaller list may seem counter intuitive, the genes in the combined list are more likely to be the targets of domestication and, therefore, they may be better candidates for additional more focused research. It may be beneficial to design a method that automatically combines the results from the three approaches to make better predictions.

References

Allegrucci, C., Wu, Y.-Z., Thurston, A., Denning, C. N., Priddle, H., Mummery, C. L., Ward-van Oostwaard, D., Andrews, P. W., Stojkovic, M., Smith, N., Parkin, T., Jones, M. E., Warren, G., Yu, L., Brena, R. M., Plass, C., and Young, L. E. (2007). Restriction landmark genome scanning identifies culture-induced DNA methylation instability in the human embryonic stem cell epigenome. *Hum Mol Genet*, 16(10):1253–1268.

Bae, J. M., Giroux, M., and Hannah, L. (1990). Cloning and molecular characterization of the *brittle-2* gene of maize. *Maydica*, 35(4):317–322.

Baldi, P. and Long, A. D. (2001). A bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Becker, C., Hagmann, J., Muller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D. (2011). Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature*, 480(7376):245–249.

Bell, J., Pai, A., Pickrell, J., Gaffney, D., Pique-Regi, R., Degner, J., Gilad, Y., and Pritchard, J. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol*, 12(1):R10.

Beló, A., Beatty, M., Hondred, D., Fengler, K., Li, B., and Rafalski, A. (2010). Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor Appl Genet*, 120(2):355–367.

Ben-Yaacov, E. and Eldar, Y. C. (2008). A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, 24(16):i139–i145.

Benedito, V. A., Torres-Jerez, I., Murray, J. D., Andriankaja, A., Allen, S., Kakar, K., Wandrey, M., Verdier, J., Zuber, H., Ott, T., Moreau, S., Niebel, A., Frickey, T., Weiller, G., He, J., Dai, X., Zhao, P. X., Tang, Y., and Udvardi, M. K. (2008). A gene expression atlas of the model legume *Medicago truncatula*. *Plant J*, 55(3):504–513.

Bennett, M. and Leitch, I. (2012). Plant DNA c-values database (release 6.0, dec. 2012).

Bergmann, S., Ihmels, J., and Barkai, N. (2003). Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, 2(1):e9.

Bestor, T. H. (1990). DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philos T Roy Soc B*, 326(1235):179–187.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Gene Dev*, 16(1):6–21.

Bird, A. (2007). Perceptions of epigenetics. *Nature*, 447(7143):396–398.

Bird, A. P. (1995). Gene number, noise reduction and biological complexity. *Trends Genet*, 11(3):94 – 100.

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nat Rev Genet*, 13(10):705–719.

Bomblies, K. and Doebley, J. F. (2006). Pleiotropic effects of the duplicate maize *FLORICAULA/LEAFY* genes *zfl1* and *zfl2* on traits under selection during maize domestication. *Genetics*, 172(1):519–531.

Briggs, W. H., McMullen, M. D., Gaut, B. S., and Doebley, J. (2007). Linkage mapping of domestication loci in a large maize–teosinte backcross resource. *Genetics*, 177(3):1915–1928.

Buckler, E. S., Gaut, B. S., and McMullen, M. D. (2006). Molecular and functional diversity of maize. *Curr Opin Plant Biol*, 9(2):172–176.

Buckler, E. S., Thornsberry, J. M., and Kresovich, S. (2001). Molecular diversity, structure and domestication of grasses. *Genet Res*, 77(03):213–218.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *J Am Stat Assoc*, 83(401):123–127.

Butte, A. J. and Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 5:418–429.

Caicedo, A. L., Williamson, S. H., Hernandez, R. D., Boyko, A., Fledel-Alon, A., York, T. L., Polato, N. R., Olsen, K. M., Nielsen, R., McCouch, S. R., Bustamante, C. D., and Purugganan, M. D. (2007). Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet*, 3(9):e163.

Carlson, M., Zhang, B., Fang, Z., Mischel, P., Horvath, S., and Nelson, S. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, 7(1):40.

Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242 –2250.

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res*, 20(3):393–402.

Chia, J.-M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., Gaut, B., Geller, L., Glaubitz, J. C., Gore, M., Guill, K. E., Holland, J., Hufford, M. B., Lai, J., Li, M., Liu, X., Lu, Y., McCombie, R., Nelson, R., Poland, J., Prasanna, B. M., Pyhajarvi, T., Rong, T., Sekhon, R. S., Sun, Q., Tenaillon, M. I., Tian, F., Wang, J., Xu, X., Zhang, Z., Kaeppler, S. M., Ross-Ibarra, J., McMullen, M. D., Buckler, E. S., Zhang, G., Xu, Y., and Ware, D. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet*, 44(7):803–807.

Chodavarapu, R. K., Feng, S., Ding, B., Simon, S. A., Lopez, D., Jia, Y., Wang, G.-L., Meyers, B. C., Jacobsen, S. E., and Pellegrini, M. (2012). Transcriptome and methylome interactions in rice hybrids. *P Natl Acad Sci USA*, 109(30):12040–12045.

Christian, N., May, P., Kempa, S., Handorf, T., and Ebenhoh, O. (2009). An integrative approach towards completing genome-scale metabolic networks. *Mol BioSyst*, 5(12):1889–1903.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184):215–219.

Cone, K. C., Cocciolone, S. M., Burr, F. A., and Burr, B. (1993). Maize anthocyanin regulatory gene *pl* is a duplicate of *c1* that functions in the plant. *Plant Cell*, 5(12):1795–1805.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S., Prinz, J., St. Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis,

B.-J., Shuteriqi, E., Tong, A. H. Y., van Dyk, N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B., Pal, C., Roth, F. P., Giaever, G., Nislow, C., Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A.-C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., and Boone, C. (2010). The genetic landscape of a cell. *Science*, 327(5964):425–431.

Davidson, R. M., Hansey, C. N., Gowda, M., Childs, K. L., Lin, H., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaeppler, S. M., Jiang, N., and Buell, C. R. (2011). Utility of RNA sequencing for analysis of maize reproductive transcriptomes. *Plant Gen*, 4(3):191–203.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Roy Statist Soc Ser B*, 39(1):1–38.

Deng, M., Tu, Z., Sun, F., and Chen, T. (2004). Mapping gene ontology to proteins based on protein–protein interaction data. *Bioinformatics*, 20(6):895 –902.

Dixon, S. J., Fedyshyn, Y., Koh, J. L. Y., Prasad, T. S. K., Chahwan, C., Chua, G., Toufighi, K., Baryshnikova, A., Hayles, J., Hoe, K.-L., Kim, D.-U., Park, H.-O., Myers, C. L., Pandey, A., Durocher, D., Andrews, B. J., and Boone, C. (2008). Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *P Natl Acad Sci USA*, 105(43):16653–16658.

Doebley, J. (2004). The genetics of maize evolution. *Annu Rev Genet*, 38(1):37 – C–4.

Doebley, J., Bacigalupo, A., and Stec, A. (1994). Inheritance of kernel weight in two maize-teosinte hybrid populations: implications for crop evolution. *J Hered*, 85(3):191–195.

Doebley, J. and Stec, A. (1993). Inheritance of the morphological differences between maize and teosinte: comparison of results for two f2 populations. *Genetics*, 134(2):559–570.

Doebley, J., Stec, A., and Hubbard, L. (1997). The evolution of apical dominance in maize. *Nature*, 386(6624):485–488.

Doebley, J. F., Gaut, B. S., and Smith, B. D. (2006). The molecular genetics of crop domestication. *Cell*, 127(7):1309 – 1321.

Doyle, J. J. and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Photochem Bull*, 19:11–15.

Druka, A., Muehlbauer, G., Druka, I., Caldo, R., Baumann, U., Rostoks, N., Schreiber, A., Wise, R., Close, T., Kleinhofs, A., Graner, A., Schulman, A., Langridge, P., Sato, K., Hayes, P., McNicol, J., Marshall, D., and Waugh, R. (2006). An atlas of gene

expression from seed to seed through barley development. *Funct Integr Genomic*, 6(3):202–211.

Dutilh, B., Huynen, M., and Snel, B. (2006). A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. *BMC Genomics*, 7(1):10.

Eichten, S. R., Briskine, R., Song, J., Li, Q., Swanson-Wagner, R., Hermanson, P. J., Waters, A. J., Starr, E., West, P. T., Tiffin, P., Myers, C. L., Vaughn, M. W., and Springer, N. M. (2013a). Epigenetic and genetic influences on DNA methylation variation in maize populations. *Plant Cell*.

Eichten, S. R., Ellis, N. A., Makarevitch, I., Yeh, C.-T., Gent, J. I., Guo, L., McGinnis, K. M., Zhang, X., Schnable, P. S., Vaughn, M. W., Dawe, R. K., and Springer, N. M. (2012). Spreading of heterochromatin is limited to specific families of maize retrotransposons. *PLoS Genet*, 8(12):e1003127.

Eichten, S. R., Swanson-Wagner, R. A., Schnable, J. C., Waters, A. J., Hermanson, P. J., Liu, S., Yeh, C.-T., Jia, Y., Gendler, K., Freeling, M., Schnable, P. S., Vaughn, M. W., and Springer, N. M. (2011). Heritable epigenetic variation among maize inbreds. *PLoS Genet*, 7(11):e1002372.

Eichten, S. R., Vaughn, M. W., Hermanson, P. J., and Springer, N. M. (2013b). Variation in DNA methylation patterns is more common among maize inbreds than among tissues. *Plant Gen*, 6(2):–.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *P Natl Acad Sci USA*, 95(25):14863 –14868.

Essien, K., Hannenhalli, S., and Stoeckert, J. (2008). Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to plasmodium phenotypic diversity. *PLoS ONE*, 3(9):e3122.

Eveland, A. L., Satoh-Nagasawa, N., Goldshmidt, A., Meyer, S., Beatty, M., Sakai, H., Ware, D., and Jackson, D. (2010). Digital gene expression signatures for maize development. *Plant Physiol*, 154(3):1024–1039.

FAO (2013). *FAO statistical yearbook 2013: world food and agriculture.* FAO statistical yearbook. Food and Agriculture Organization of the United Nations.

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., Ukomadu, C., Sadler, K. C., Pradhan, S., Pellegrini, M., and Jacobsen, S. E. (2010). Conservation and divergence of methylation patterning in plants and animals. *P Natl Acad Sci USA*, 107(19):8689 –8694.

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat Rev Genet*, 7(2):85–97.

Ficklin, S. P. and Feltus, F. A. (2011). Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. *Plant Physiol*, 156(3):1244 –1256.

Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., Wong, W.-K., and Mockler, T. C. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res*, 20(1):45–58.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann Eugenic*, 7(2):179–188.

Frost, A., Elgort, M., Brandman, O., Ives, C., Collins, S., Miller-Vedam, L., Weibezahn, J., Hein, M., Poser, I., Mann, M., Hyman, A., and Weissman, J. (2012). Functional repurposing revealed by comparing *S. pombe* and *S. cerevisiae* genetic interactions. *Cell*, 149(6):1339–1352.

Fu, H., Zheng, Z., and Dooner, H. K. (2002). Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proceedings of the National Academy of Sciences*, 99(2):1082–1087.

Gallavotti, A., Zhao, Q., Kyozuka, J., Meeley, R. B., Ritter, M. K., Doebley, J. F., Enrico Pe, M., and Schmidt, R. J. (2004). The role of barren *stalk1* in the architecture of maize. *Nature*, 432(7017):630–635.

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth*, 8(6):469–477.

Gehring, M., Choi, Y., and Fischer, R. L. (2004). Imprinting and seed development. *Plant Cell*, 16(suppl 1):S203–S213.

Genereux, D. P., Miner, B. E., Bergstrom, C. T., and Laird, C. D. (2005). A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns. *P Natl Acad Sci USA*, 102(16):5802–5807.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., and Johnston, M. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896):387–391.

Giroux, M. and Hannah, L. (1994). ADP-glucose pyrophosphorylase in *shrunken-2* and *brittle-2* mutants of maize. *Mol Gen Genet*, 243(4):400–408.

Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, 128(4):635 – 638.

Gore, M. A., Chia, J.-M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. A., McMullen, M. D., Grills, G. S., Ross-Ibarra, J., Ware, D. H., and Buckler, E. S. (2009). A first-generation haplotype map of maize. *Science*, 326(5956):1115 –1117.

Grant, G. R., Manduchi, E., and Stoeckert, C. J. (2007). Analysis and management of microarray gene expression data. In *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc.

Gross, B. L. and Olsen, K. M. (2010). Genetic perspectives on crop domestication. *Trends Plant Sci*, 15(9):529 – 537.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Guan, Y., Dunham, M., Caudy, A., and Troyanskaya, O. (2010). Systematic planning of genome-scale experiments in poorly studied species. *PLoS Comput Biol*, 6(3):e1000698.

Guan, Y., Dunham, M., Troyanskaya, O., and Caudy, A. (2013). Comparative gene expression between two yeast species. *BMC Genomics*, 14(1):33.

Hafner, S. (2003). Trends in maize, rice, and wheat yields for 188 nations over the past 40 years: a prevalence of linear growth. *Agr Ecosyst Environ*, 97(1–3):275–283.

Haig, D. (2004). The (dual) origin of epigenetics. *Cold Spring Harb Symp Quant Biol*, 69:67–70.

Hannah, L. C., Shaw, J. R., Giroux, M. J., Reyss, A., Prioul, J.-L., Bae, J.-M., and Lee, J.-Y. (2001). Maize genes encoding the small subunit of ADP-glucose pyrophosphorylase. *Plant Physiol*, 127(1):173–183.

Hansen, K., Langmead, B., and Irizarry, R. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, 13(10):R83.

Hansey, C. N., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2012). Maize (*Zea mays L.*) genome diversity as revealed by RNA-sequencing. *PLoS ONE*, 7(3):e33071.

Hartman, J. L., Garvik, B., and Hartwell, L. (2001). Principles for the buffering of genetic variation. *Science*, 291(5506):1001–1004.

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat Rev Genet*, 11(7):476–486.

He, G., Zhu, X., Elling, A. A., Chen, L., Wang, X., Guo, L., Liang, M., He, H., Zhang, H., Chen, F., Qi, Y., Chen, R., and Deng, X.-W. (2010). Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell*, 22(1):17–33.

Henderson, I. R. and Jacobsen, S. E. (2007). Epigenetic inheritance in plants. *Nature*, 447(7143):418–424.

Henikoff, S. (2002). Beyond the central dogma. *Bioinformatics*, 18(2):223–225.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotech*, 28(9):977–982.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artif Intell Rev*, 22(2):85–126.

Hodges, E., Molaro, A., Santos, C. O. D., Thekkat, P., Song, Q., Uren, P. J., Park, J., Butler, J., Rafii, S., McCombie, W. R., Smith, A. D., and Hannon, G. J. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell*, 44(1):17 – 28.

Hollister, J. D., Smith, L. M., Guo, Y.-L., Ott, F., Weigel, D., and Gaut, B. S. (2011). Transposable elements and small RNAs contribute to gene expression divergence

between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *P Natl Acad Sci USA*, 108(6):2322–2327.

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., Laurance, M. F., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, A. C., Liau, L. M., Wu, H., Geschwind, D. H., Febbo, P. G., Kornblum, H. I., Cloughesy, T. F., Nelson, S. F., and Mischel, P. S. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *P Natl Acad Sci USA*, 103(46):17402–17407.

Hufford, M. B., Xu, X., van Heerwaarden, J., Pyhajarvi, T., Chia, J.-M., Cartwright, R. A., Elshire, R. J., Glaubitz, J. C., Guill, K. E., Kaeppler, S. M., Lai, J., Morrell, P. L., Shannon, L. M., Song, C., Springer, N. M., Swanson-Wagner, R. A., Tiffin, P., Wang, J., Zhang, G., Doebley, J., McMullen, M. D., Ware, D., Buckler, E. S., Yang, S., and Ross-Ibarra, J. (2012). Comparative population genomics of maize domestication and improvement. *Nat Genet*, 44(7):808–811.

Huttenhower, C., Haley, E. M., Hibbs, M. A., Dumeaux, V., Barrett, D. R., Coller, H. A., and Troyanskaya, O. G. (2009). Exploring the human genome with functional maps. *Genome Res*, 19(6):1093 –1106.

Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O. G. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 22(23):2890 –2897.

Huttenhower, C., Schroeder, M., Chikina, M. D., and Troyanskaya, O. G. (2008). The sleipnir library for computational functional genomics. *Bioinformatics*, 24(13):1559 –1561.

Ihmels, J., Bergmann, S., Berman, J., and Barkai, N. (2005). Comparative gene expression analysis by a differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genet*, 1(3):e39.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Jablonka, E. and Raz, G. (2009). Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. *Q Rev Biol*, 84(2):131–176.

Jiao, Y., Lori Tausta, S., Gandotra, N., Sun, N., Liu, T., Clay, N. K., Ceserani, T., Chen, M., Ma, L., Holford, M., Zhang, H.-y., Zhao, H., Deng, X.-W., and Nelson, T. (2009). A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nat Genet*, 41(2):258–263.

Jordan, I. K., Mariño-Ramírez, L., Wolf, Y. I., and Koonin, E. V. (2004). Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol*, 21(11):2058 –2070.

Kankel, M. W., Ramsey, D. E., Stokes, T. L., Flowers, S. K., Haag, J. R., Jeddeloh, J. A., Riddle, N. C., Verbsky, M. L., and Richards, E. J. (2003). Arabidopsis MET1 cytosine methyltransferase mutants. *Genetics*, 163(3):1109–1122.

Kazazian, H. H. (2004). Mobile elements: drivers of genome evolution. *Science*, 303(5664):1626–1632.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254.

Kirst, M., Caldo, R., Casati, P., Tanimoto, G., Walbot, V., Wise, R. P., and Buckler, E. S. (2006). Genetic diversity contribution to errors in short oligonucleotide microarray analysis. *Plant Biotechnol J*, 4(5):489–498.

Kishimoto, M. (1994). Fermentation characteristics of hybrids between the cryophilic wine yeast saccharomyces bayanus and the mesophilic wine yeast saccharomyces cerevisiae. *J Ferment Bioeng*, 77(4):432 – 435.

Koch, E., Costanzo, M., Bellay, J., Deshpande, R., Chatfield-Reed, K., Chua, G., D'Urso, G., Andrews, B., Boone, C., and Myers, C. (2012). Conserved rules govern genetic interaction degree across species. *Genome Biol*, 13(7):R57.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693 – 705.

Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572.

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet*, 11(3):191–203.

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., Li, M.-W., He, W., Qin, N., Wang, B., Li, J., Jian, M., Wang, J., Shao, G., Wang, J., Sun, S. S.-M., and Zhang, G. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet*, 42(12):1053–1059.

Langfelder, P. and Horvath, S. (2008). WGCNA: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25.

Law, C. and Cheung, P. (2013). Histone variants and transcription regulation. In Kundu, T. K., editor, *Epigenetics: Development and Disease*, volume 61 of *Subcellular Biochemistry*, pages 319–341. Springer Netherlands.

Law, J. A. and Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet*, 11(3):204–220.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., and Rhee, S. Y. (2010). Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat Biotech*, 28(2):149–156.

Lee, J.-M., Williams, M., Tingey, S., and Rafalski, A. (2002). DNA array profiling of gene expression changes during maize embryo development. *Funct Integr Genomic*, 2(1-2):13–27.

Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S. L., Kebrom, T. H., Provart, N., Patel, R., Myers, C. R., Reidel, E. J., Turgeon, R., Liu, P., Sun, Q., Nelson, T., and Brutnell, T. P. (2010). The developmental dynamics of the maize leaf transcriptome. *Nat Genet*, 42(12):1060–1067.

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., Franklin, L. D., He, J., Xu, D., May, G., and Stacey, G. (2010). An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *Plant J*, 63(1):86–99.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133(3):523–536.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322.

Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., Downes, M., Yu, R., Stewart, R., Ren, B., Thomson, J. A., Evans, R. M., and Ecker, J. R. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471(7336):68–73.

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., Feng, Q., Zhao, Y., Guo, Y., Li, W., Huang, X., and Han, B. (2010). Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res*, 20(9):1238–1249.

Manning, K., Tor, M., Poole, M., Hong, Y., Thompson, A. J., King, G. J., Giovannoni, J. J., and Seymour, G. B. (2006). A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat Genet*, 38(8):948–952.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9):1509–1517.

Martin, A., Troadec, C., Boualem, A., Rajab, M., Fernandez, R., Morin, H., Pitrat, M., Dogimont, C., and Bendahmane, A. (2009). A transposon-induced epigenetic change leads to sex determination in melon. *Nature*, 461(7267):1135–1138.

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez G., J., Buckler, E., and Doebley, J. (2002). A single domestication for maize shown by multilocus microsatellite genotyping. *P Natl Acad Sci USA*, 99(9):6080 –6084.

Messing, J. and Dooner, H. K. (2006). Organization and variability of the maize genome. *Curr Opin Plant Biol*, 9(2):157–163.

Mittelsten Scheid, O., Afsar, K., and Paszkowski, J. (2003). Formation of stable epialleles and their paramutation-like interaction in tetraploid *Arabidopsis thaliana*. *Nat Genet*, 34(4):450–454.

Moose, S. P. and Sisco, P. H. (1996). Glossy15, an APETALA2-like gene from maize that regulates leaf epidermal cell identity. *Gene Dev*, 10(23):3018–3027.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7):621–628.

Moser, D., Ekawardhani, S., Kumsta, R., Palmason, H., Bock, C., Athanassiadou, Z., Lesch, K.-P., and Meyer, J. (2008). Functional analysis of a potassium-chloride co-transporter 3 (SLC12A6) promoter polymorphism leading to an additional DNA methylation site. *Neuropsychopharmacol*, 34(2):458–467.

Natt, D., Rubin, C.-J., Wright, D., Johnsson, M., Belteky, J., Andersson, L., and Jensen, P. (2012). Heritable genome-wide variation of gene expression and promoter methylation between wild and domesticated chickens. *BMC Genomics*, 13(1):59.

Oldham, M. C., Horvath, S., and Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *P Natl Acad Sci USA*, 103(47):17973 –17978.

Pevsner, J. (2009). *Bioinformatics and functional genomics*. John Wiley & Sons, Inc., 2nd edition.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(1):27.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B.-m., Gray, J. W., and Albertson, D. G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, 20(2):207–211.

Piperno, D. R., Ranere, A. J., Holst, I., Iriarte, J., and Dickau, R. (2009). Starch grain and phytolith evidence for early ninth millennium B.P. maize from the central balsas river valley, mexico. *P Natl Acad Sci USA*, 106(13):5019 –5024.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*, 23(1):41–46.

Prioul, J. L., Jeannette, E., Reyss, A., Gregory, N., Giroux, M., Hannah, L. C., and Causse, M. (1994). Expression of ADP-glucose pyrophosphorylase in maize (*Zea mays l.*) grain and source leaf during grain filling. *Plant Physiol*, 104(1):179–187.

Purugganan, M. D. and Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature*, 457(7231):843–848.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rabinowicz, P. D. and Bennetzen, J. L. (2006). The maize genome as a model for efficient sequence analysis of large plant genomes. *Curr Opin Plant Biol*, 9(2):149–156.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol*, 5(2):94 – 100.

Rando, O. J. and Verstrepen, K. J. (2007). Timescales of genetic and epigenetic inheritance. *Cell*, 128(4):655 – 668.

Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., and Foley, J. A. (2012). Recent patterns of crop yield growth and stagnation. *Nat Commun*, 3:1293.

Regulski, M., Lu, Z., Kendall, J., Donoghue, M. T., Reinders, J., Llaca, V., Deschamps, S., Smith, A., Levy, D., McCombie, W. R., Tingey, S., Rafalski, A., Hicks, J., Ware, D., and Martienssen, R. (2013). The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res*.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *P Natl Acad Sci USA*, 98(20):11479–11484.

Rensink, W. A. and Buell, C. R. (2005). Microarray expression profiling resources for plant genomics. *Trends Plant Sci*, 10(12):603 – 609.

Reverter, A. and Chan, E. K. F. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, 24(21):2491 –2497.

Richards, E. J. (2006). Inherited epigenetic variation - revisiting soft inheritance. *Nat Rev Genet*, 7(5):395–401.

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S. R., Qu, H., Shales, M., Park, H.-O., Hayles, J., Hoe, K.-L., Kim, D.-U., Ideker, T., Grewal, S. I., Weissman, J. S., and Krogan, N. J. (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, 322(5900):405 –410.

Russo, V. E., Martienssen, R. A., and Riggs, A. D. (1996). *Epigenetic mechanisms of gene regulation.* Cold Spring Harbor Laboratory Press, New York.

Rösti, S. and Denyer, K. (2007). Two paralogous genes encoding small subunits of ADP-glucose pyrophosphorylase in maize, *Bt2* and *L2*, replace the single alternatively spliced gene found in other cereal species. *J Mol Evol*, 65(3):316–327.

Salamov, A. A. and Solovyev, V. V. (2000). Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res*, 10(4):516–522.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z., and Bennetzen, J. L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science*, 274(5288):765–768.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., Hurles, M. E., and Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nat Genet*.

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J. U. (2005). A gene expression map of *Arabidopsis thaliana* development. *Nat Genet*, 37(5):501–506.

Schmitz, R. J., He, Y., Valdés-López, O., Khan, S. M., Joshi, T., Urich, M. A., Nery, J. R., Diers, B., Xu, D., Stacey, G., and Ecker, J. R. (2013a). Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res*.

Schmitz, R. J., Schultz, M. D., Lewsey, M. G., O'Malley, R. C., Urich, M. A., Libiger, O., Schork, N. J., and Ecker, J. R. (2011). Transgenerational epigenetic instability is a source of novel methylation variants. *Science*, 334(6054):369–373.

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R. B., Chen, H., Schork, N. J., and Ecker, J. R. (2013b). Patterns of population epigenomic diversity. *Nature*, 495(7440):193–198.

Schnable, J. C. and Freeling, M. (2011). Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PLoS ONE*, 6(3):e17855.

Schnable, J. C., Freeling, M., and Lyons, E. (2012). Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol Evol*, 4(3):265–277.

Schnable, J. C., Springer, N. M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *P Natl Acad Sci USA*, 108(10):4069–4074.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C.,

Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. (2009). The b73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112 –1115.

Schug, J., Schuller, W.-P., Kappen, C., Salbaum, J. M., Bucan, M., and Stoeckert, C. (2005). Promoter features related to tissue specificity as measured by shannon entropy. *Genome Biol*, 6(4):R33.

Sekhon, R. S., Briskine, R., Hirsch, C. N., Myers, C. L., Springer, N. M., Buell, C. R., de Leon, N., and Kaeppler, S. M. (2013). Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. *PLoS ONE*, 8(4):e61005.

Sekhon, R. S., Lin, H., Childs, K. L., Hansey, C. N., Buell, C. R., de Leon, N., and Kaeppler, S. M. (2011). Genome-wide atlas of transcription during maize development. *Plant J*, 66(4):553–563.

Severin, A., Woody, J., Bolon, Y.-T., Joseph, B., Diers, B., Farmer, A., Muehlbauer, G., Nelson, R., Grant, D., Specht, J., Graham, M., Cannon, S., May, G., Vance, C., and Shoemaker, R. (2010). RNA-seq atlas of *Glycine max*: A guide to the soybean transcriptome. *BMC Plant Biol*, 10(1):160.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst Tech J*, 27(3):379–423.

Shapiro, J. A. (2009). Revisiting the central dogma in the 21st century. *Ann NY Acad Sci*, 1178(1):6–28.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol Syst Biol*, 3.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1).

Springer, N. M. (2013). Epigenetics and crop improvement. *Trends in Genetics*, 29(4):241 – 247.

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A. L., Barbazuk, W. B., Jeddeloh, J. A., Nettleton, D., and Schnable, P. S. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet*, 5(11):e1000734.

Storey, J. D. (2002). A direct approach to false discovery rates. *J Roy Statist Soc Ser B*, 64(3):479–498.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249 –255.

Suzuki, M. M. and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*, 9(6):465–476.

Swanson-Wagner, R., Briskine, R., Schaefer, R., Hufford, M. B., Ross-Ibarra, J., Myers, C. L., Tiffin, P., and Springer, N. M. (2012). Reshaping of the maize transcriptome by domestication. *P Natl Acad Sci USA*, 109(29):11878 –11883.

Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D., and Springer, N. M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res*, 20(12):1689 –1699.

Swigoňová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J. L., and Messing, J. (2004). Close split of sorghum and maize genome progenitors. *Genome Res*, 14(10a):1916–1923.

Sylvester, A., Cande, W., and Freeling, M. (1990). Division and differentiation during normal and *liguleless-1* maize leaf development. *Development*, 110(3):985–1000.

Taiwo, O., Wilson, G. A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., Beck, S., and Butcher, L. M. (2012). Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc*, 7(4):617–636.

Tanksley, S. D. (1993). Mapping polygenes. *Annu Rev Genet*, 27(1):205–233.

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet*, 28(3):286–289.

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T. R., McMullen, M. D., Holland, J. B., and Buckler, E. S. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet*, 43(2):159–162.

Tian, Z., Zhang, H., and Kuang, R. (2012). Sparse group selection on fused lasso components for identifying group-specific DNA copy number variations. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 665–674.

Tirosh, I. and Barkai, N. (2007). Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol*, 8(4):R50.

Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C. W. V., Bussey, H., Andrews, B., Tyers, M., and Boone, C. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368.

Tong, A. H. Y., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D. S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J. N., Lu, H., Menard, P., Munyana, C., Parsons, A. B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A.-M., Shapiro, J., Sheikh, B., Suter, B., Wong, S. L., Zhang, L. V., Zhu, H., Burd, C. G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F. P., Brown, G. W., Andrews, B., Bussey, H., and Boone, C. (2004). Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813.

Tornow, S. and Mewes, H. W. (2003). Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res*, 31(21):6283–6289.

Tran, R. K., Henikoff, J. G., Zilberman, D., Ditt, R. F., Jacobsen, S. E., and Henikoff, S. (2005). DNA methylation profiling identifies CG methylation clusters in arabidopsis genes. *Curr Biol*, 15(2):154 – 159.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 28(5):511–515.

Troyer, A. F. (2006). Adaptedness and heterosis in corn and mule hybrids. *Crop Sci*, 46(2):528–543.

Tsuda, K., Shin, H., and Schölkopf, B. (2005). Fast protein classification with multiple networks. *Bioinformatics*, 21(suppl 2):ii59–ii65.

van Eijk, K., de Jong, S., Boks, M., Langeveld, T., Colas, F., Veldink, J., de Kovel, C., Janson, E., Strengman, E., Langfelder, P., Kahn, R., van den Berg, L., Horvath, S.,

and Ophoff, R. (2012). Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, 13(1):636.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235):484–487.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Jr, D. E. B., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell*, 88(2):243 – 251.

Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663.

Vignols, F., Rigau, J., Torres, M. A., Capellades, M., and Puigdomènech, P. (1995). The brown midrib3 (*bm3*) mutation in maize occurs in the gene encoding caffeic acid o-methyltransferase. *Plant Cell*, 7(4):407–416.

Vigouroux, Y., McMullen, M., Hittinger, C. T., Houchins, K., Schulz, L., Kresovich, S., Matsuoka, Y., and Doebley, J. (2002). Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *P Natl Acad Sci USA*, 99(15):9650–9655.

Vigouroux, Y., Mitchell, S., Matsuoka, Y., Hamblin, M., Kresovich, S., Smith, J. S. C., Jaqueth, J., Smith, O. S., and Doebley, J. (2005). An analysis of genetic diversity across the maize genome using microsatellites. *Genetics*, 169(3):1617–1630.

Waddington, C. H. (1957). *The strategy of the genes*. London: George Allen & Unwin, Ltd.

Wang, R.-L., Stec, A., Hey, J., Lukens, L., and Doebley, J. (1999). The limits of selection during maize domestication. *Nature*, 398(6724):236–239.

Wang, Y., Robbins, K. R., and Rekaya, R. (2010). Comparison of computational models for assessing conservation of gene expression across species. *PLoS ONE*, 5(10):e13239.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.

Weston, D., Gunter, L., Rogers, A., and Wullschleger, S. (2008). Connecting genes, coexpression modules, and molecular signatures to environmental stress phenotypes in plants. *BMC Syst Biol*, 2(1):16.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Wong, S. L., Zhang, L. V., Tong, A. H. Y., Li, Z., Goldberg, D. S., King, O. D., Lesage, G., Vidal, M., Andrews, B., Bussey, H., Boone, C., and Roth, F. P. (2004). Combining biological networks to predict genetic interactions. *P Natl Acad Sci USA*, 101(44):15682 –15687.

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., and Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science*, 308(5726):1310 –1314.

Xu, J., Pope, S. D., Jazirehi, A. R., Attema, J. L., Papathanasiou, P., Watts, J. A., Zaret, K. S., Weissman, I. L., and Smale, S. T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *P Natl Acad Sci USA*, 104(30):12377–12382.

Yamasaki, M., Tenaillon, M. I., Vroh Bi, I., Schroeder, S. G., Sanchez-Villeda, H., Doebley, J. F., Gaut, B. S., and McMullen, M. D. (2005). A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell*, 17(11):2859–2872.

Yamasaki, M., Wright, S. I., and McMullen, M. D. (2007). Genomic screening for artificial selection during domestication and improvement in maize. *Ann Bot*, 100(5):967–973.

Ye, S. Q., LaVoie, T., Usher, D. C., and Zhang, L. Q. (2002). Microarray, SAGE and their applications to cardiovascular diseases. *Cell Res*, 12(2):105–115.

Zemach, A., McDaniel, I. E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 328(5980):916–919.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene coexpression network analysis. *Stat Appl Genet Mol Biol*, 4(1):Article 17.

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., Zhuang, R., Lu, Z., He, Z., Fang, X., Chen, L., Tian, W., Tao, Y., Kristiansen, K., Zhang, X., Li, S., Yang, H., Wang, J., and Wang, J. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res*, 20(5):646–654.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y. V., Pellegrini, M., Goodrich, J., and Jacobsen, S. E. (2007). Whole-genome analysis of histone h3 lysine 27 trimethylation in arabidopsis. *PLoS Biol*, 5(5):e129.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., and Ecker, J. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*, 126(6):1189–1201.

Zhang, Y., Liu, H., Lv, J., Xiao, X., Zhu, J., Liu, X., Su, J., Li, X., Wu, Q., Wang, F., and Cui, Y. (2011). QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res*, 39(9):e58–e58.

Zhao, Q., Thuillet, A.-C., Uhlmann, N. K., Weber, A., Rafalski, J. A., Allen, S. M., Tingey, S., and Doebley, J. (2008). The role of regulatory genes during maize domestication: evidence from nucleotide polymorphism and gene expression. *Genetics*, 178(4):2133–2143.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet*, 39(1):61–69.