Dissecting complex phenotypes via multiple transcriptome-based GWAS

Hung-Ying Lin
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd
Part of the Bioinformatics Commons, and the Genetics Commons

Recommended Citation
https://lib.dr.iastate.edu/etd/17243

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Dissecting complex phenotypes via multiple transcriptome-based GWAS

by

Hung-Ying Lin

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

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Patrick S Schnable, Major Professor
Dan Nettleton
Jianming Yu
Peng Liu
Carolyn J Lawrence-Dill

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2018

Copyright © Hung-Ying Lin, 2018. All rights reserved.
TABLE OF CONTENTS

ACKNOWLEDGEMENT ................................................................................. iv
ABSTRACT ...................................................................................................... v
CHAPTER 1. GENERAL INTRODUCTION ................................................... 1
   Introduction ............................................................................................... 1
   Research Goals ....................................................................................... 3
   Dissertation Organization ....................................................................... 4
   References ............................................................................................... 5

CHAPTER 2. SUBSTANTIAL CONTRIBUTION OF GENETIC VARIATION IN THE
EXPRESSION OF TRANSCRIPTION FACTORS TO PHENOTYPIC VARIATION
REVEALED BY eRD-GWAS ................................................................. 7
   Abstract ................................................................................................... 7
   Background .............................................................................................. 8
   Results ..................................................................................................... 10
   Discussion .............................................................................................. 18
   Conclusions ............................................................................................ 24
   Methods ................................................................................................... 24
   References ............................................................................................... 31
   Tables ....................................................................................................... 38
   Figures ..................................................................................................... 39
   Supplemental Figures ............................................................................ 42
   Supplemental Tables ............................................................................... 50

CHAPTER 3. THE IDENTIFICATION OF GENES ASSOCIATED WITH
PHENOTYPIC VARIATIONS VIA GLOBAL ANALYSES OF ALTERNATIVE
SPICING EVENTS (AS-GWAS) .......................................................... 56
   Abstract ................................................................................................. 56
   Background ............................................................................................. 57
   Results ..................................................................................................... 60
   Discussion ............................................................................................... 72
   Conclusions ............................................................................................ 74
   Methods ................................................................................................... 75
   References ............................................................................................... 80
   Tables ....................................................................................................... 86
   Figures ..................................................................................................... 89
   Supplemental Figures ............................................................................ 94
   Supplemental Tables ............................................................................... 97

CHAPTER 4. HP-GWAS: ASSOCIATION OF GENIC HAPLOTYPES WITH
PHENOTYPES ..................................................................................... 100
   Abstract ................................................................................................. 100
   Introduction ............................................................................................ 101
   Results ..................................................................................................... 103
   Discussion .............................................................................................. 109
   Methods ................................................................................................... 112
CHAPTER 5. GENERAL CONCLUSION ................................................................. 133
ACKNOWLEDGEMENT

First, I would like to thank my advisor Dr. Patrick S. Schnable for providing me with great opportunities to learn the state of art knowledge in his lab. Next, I like to thank all the colleagues in the lab for their scientific and personal life support. I’m also grateful to my POS committee: Drs. Dan Nettleton, Peng Liu, Jianming Yu, Carolyn Lawrence-Dill for their invaluable suggestion on my research. Alvis Hu and Qiang Liu gave me great help in both research and life. My sincere thanks also go to Dr. Rohan Fernando and Dr. Dorian Garrick for all the great discussions in their lab. Finally, I would also like to thank my parents, Cassy, and Freya for giving me great support and fulfill me spiritually throughout my life.
ABSTRACT

Genome-Wide Association Study (GWAS) have been widely used to detect the QTLs based on Linkage Disequilibrium (LD) relationships between SNPs and QTLs. However, in conventional GWAS both false positive and false negative results cause serious concerns. In this research, we developed three different transcriptome-based GWAS approaches which are complementary to conventional SNP-based GWAS. The ability to identify trait-associated genes in these three different methods are supported by cross-validation, transposon knockout mutants, and the analysis of a gene regulatory networks. In summary, we provide novel methods of detecting trait associated loci to further understand the complex gene regulatory systems which will benefit plants, animals, and disease treatment development in the future.
CHAPTER 1. GENERAL INTRODUCTION

Introduction

Next generation sequencing (NGS) Era

The first plant genome, Arabidopsis was published in 20002, even before the human genome was published (2003)1. Maize genome with 2.3 GB in size and 80% repetitive contents was also sequenced and published in 20093. With the help of reference genomes, the genomic research has come to a new era. However, use of conventional sequencing methods to generate reference genomes time consuming and expensive. Thus, the next generation sequencing (NGS) had been invented to provide a cost-effective and high-throughput solution. Within the past decade, NGS has dramatically increased the sequencing capacity by 100–1,000 factor. Not only capacity but the cost dropped even sharper than Moore's law4. Now the genomics community is towards a $1,000 human genome era5.

Genome Wide Association Studies (GWAS)

NGS generates a tremendous amount of sequence data. Therefore the next challenge is how to find meaningful biological interpretation among sequences. Genome-wide association studies (GWAS) have been developed to identify the associations between variants and phenotypes that can be used to detect the genetic loci of traits. There are two main approaches to GWAS; mixed linear models (MLM) and Bayesian-based approaches. MLM is a common approach which overcomes the population structure and the relatedness among individuals by fitting fixed and random effects. There are few MLM GWAS studies in rice6, maize7, and other species8.9. However, most of the MLM strategies estimate single marker and derive covariance to control population structure that reduces the statistical
power and increase the false positive rate\textsuperscript{10}. The other approach of GWAS is Bayesian-based approaches. Instead of estimating single marker effect, Bayesian-based GWAS applies multiple variable linear regression framework combined with Markov Chain Monte Carlo (MCMC) sampling to infer the posterior distribution for parameters\textsuperscript{11}. The common methods are ridge-regression BLUP, BayesA and BayesB\textsuperscript{12}. BayesB assumes a certain proportion of markers without effect on phenotype and markers follow identical and independent univariate t-distribution\textsuperscript{12}. Similar to BayesB, BayesC inherits most of BayesB assumptions except all the markers follow a common variance\textsuperscript{13}. Overall, Bayesian-based GWAS have been successfully applied and widely used into different species\textsuperscript{14,15,16}.

**Using different types of data to predict phenotypes**

Applying different types of data to make precise predictions on phenotype have been proposed recently\textsuperscript{15}. Instead of sequencing DNA, RNA-Seq measures cDNA after reverse transcription RNA to determine the general expression pattern within the sample. RNA-Seq provides new opportunities to understand the regulatory mechanisms among transcriptome level. There are several applications, For example: Using gene expression data associated with the single nucleotide polymorphism (SNP) data can discover the QTLs effect on gene expression [expression QTL (eQTL)]\textsuperscript{6}. Aligning RNA-Seq reads to reference genome can also detect alternative splicing (AS). Other applications are Transcriptome co-expression network analysis, gene regulatory network analysis, and identifying differentially expressed genes among samples. Overall RNA-Seq provides an opportunity to further study the genetic mechanism within complex traits and fill the gaps between variations on DNA level and final phenotypes.
Research Goals

Creating Novel Bayesian-based framework to associate gene expression with phenotype

Integrating RNA-Seq data to better understand and predict complex phenotypic traits is important in livestock, agriculture and human health. We plan to create a Bayesian-based method for genome-wide association studies (GWAS) in which the transcript accumulation of RNA-Seq are used as explanatory variables (eRD-GWAS). The results of eRD-GWAS will be validated by other analyses of RNA co-expression networks, protein–protein interaction networks, gene regulatory networks and eQTL analyses. The goal of this project is to develop a robust pipeline to identify genetic regulatory loci among transcriptome levels.

Identifying Genome wide alternative splicing effect on phenotype

Many studies indicate that AS also involves in plant growth and development, circadian control, hormone responses, and flowering. Plants respond and change phenotypes by using different types of AS isoforms and isoform abundance. Although many studies have shown that specific AS isoforms affect the phenotypes, the genome wide evidence that support association of AS patterns with phenotypes are still lacking. We plan to apply both MLM and Bayesian-based GWAS methods to create a novel method to associate genome-wide AS patterns with phenotype and recognized this process as “Alternative Splicing Genome Wide Association Study (AS_GWAS)”. The aims of this project are 1. Identifying the trait associated AS events. 2. Constructing a gene regulatory network (GRN) for the trait associated AS detected by AS_GWAS. Overall this project will provide the first evidence to support the genome-wide AS pattern contribution on phenotypes.
Creating Gene-based Haplotypes associate phenotypes and crop improvement process

Conventional GWAS approaches only focus on the association between single variants and phenotype. In theory, aggregating single variant effects into a gene level could increase the statistical power of GWAS and also decrease the false positive rate\(^{28}\). So we plan 1. Creating a haplotype based GWAS (Hp-GWAS) method that utilizes the genic haplotype which integrates all SNPs within the genic coding region and directly associates with phenotypes. 2. Generating transposon induced knock-out mutants to support the Hp-GWAS results.

Dissertation Organization

This Dissertation provides the introduction (Chapter 1), three journal articles (Chapter 2 -4) and final general conclusions (Chapter 5). Chapter 2 is a journal paper about the novel Bayesian-based method that use gene expression pattern directly as the explanatory variables to associate with phenotype. This paper has been published in Genome Biology 2017, and my major contributions include experiment design, analyses, and manuscript writing under Dr. Schnable’s guidance. Dr. Jinliang Yang provided the original idea. We applied both MLM and Bayesian-based framework to associate genome-wide AS patterns with phenotypes and investigate the regulatory networks within AS. The results have been described in chapter 3. Dr. Lakshmi Attigala assisted with manuscript writing. Chapter 4 describes a novel haplotype-based GWAS method which is an accurate and complementary approach to conventional GWAS. My contributions to this paper include manuscript writing, experiment design, and analyses under Dr. Schnable’s guidance. Dr. Lakshmi Attigala provided manuscript writing support.
References


CHAPTER 2. SUBSTANTIAL CONTRIBUTION OF GENETIC VARIATION IN
THE EXPRESSION OF TRANSCRIPTION FACTORS TO PHENOTYPIC
VARIATION REVEALED BY ERD-GWAS

Hung-ying Lin\textsuperscript{1,2}, Qiang Liu\textsuperscript{1,2}, Xiao Li\textsuperscript{2,3,4}, Jinliang Yang\textsuperscript{1,2,5}, Sanzhen Liu\textsuperscript{1,2,6}, Yinlian Huang\textsuperscript{7,8}, Michael J. Scanlon\textsuperscript{9}, Dan Nettleton\textsuperscript{10} and Patrick S. Schnable\textsuperscript{1,2,3,7,*}

1. Department of Agronomy, Iowa State University, Ames, IA 50011-3650
2. Interdepartmental Genetics and Genomics Graduate Program, Iowa State University, Ames, IA 50011-3650
3. Department of Genetics, Developmental and Cellular Biology, Iowa State University, Ames, IA 50011-3650
4. Current address: The Broad Institute of MIT and Harvard, 75 Ames Street, Cambridge, MA, 02142-1403
5. Current address: Department of Plant Sciences, University of California, Davis, CA 95616-5270 and Department of Agronomy and Horticulture, University of Nebraska, Lincoln, Nebraska 68583-0660
6. Current address: Department of Plant Pathology, Kansas State University, Manhattan KS, 66506-5502
7. Department of Plant Genetics & Breeding, China Agricultural University; Beijing, China 100193
8. Current address: DATA Biotechnology Beijing Co. Ltd., Beijing, China, 102206
9. Plant Biology Section, Cornell University, Ithaca, New York 14850, USA
10. Department of Statistics, Iowa State University; Ames, IA 50011-1210
* Corresponding Author: Patrick S. Schnable
2035 B Roy J Carver Co-Lab, Iowa State University, Ames, IA, 50011-3650 USA

Abstract

Background: There are significant limitations in existing methods for the genome-wide identification of genes whose expression patterns affect traits.

Results: The transcriptomes of five tissues from 27 genetically diverse maize inbred lines were deeply sequenced to identify genes exhibiting high and low levels of expression variation across tissues or genotypes. Transcription factors are enriched among genes with the most variation in expression across tissues, as well as among genes with higher-than-median levels of variation in expression across genotypes. In contrast, transcription factors are depleted among genes whose expression is either highly stable or highly variable across
genotypes. We developed a Bayesian-based method for genome-wide association study (GWAS) in which RNA-seq-based measures of transcript accumulation are used as explanatory variables (eRD-GWAS). The ability of eRD-GWAS to identify true associations between gene expression variation and phenotype diversity were supported by the enrichment of eRD-GWAS genes within specific nodes of RNA co-expression networks, protein-protein interaction networks, and gene regulatory networks. Genes associated with 13 traits were identified using eRD-GWAS on a panel of 369 maize inbred lines. Predicted functions of many of the resulting trait-associated genes are consistent with the analyzed traits. Importantly, transcription factors are significantly enriched among trait-associated genes identified with eRD-GWAS.

**Conclusions:** eRD-GWAS is a powerful tool for associating genes with traits and is complementary to SNP-based GWAS. The results of our eRD-GWAS are consistent with the hypothesis that genetic variation in TFs expression contributes substantially to phenotypic diversity.

**Keywords:** Transcription factors, Gene expression, GWAS, phenotypes, traits, association studies

**Background**

Many projects are underway to identify loci that contribute to traits, and the methods to do so remain under development [1]. Most commonly, associations are sought between genetic variants (e.g., SNPs) and variation in trait values via Genome-Wide Association Studies (GWAS). Typical approaches to GWAS exploit linkage disequilibrium (LD) between genetic variants such as SNPs and loci that directly affect traits of interest. There are two
main approaches for identifying such associations, mixed linear models (MLM) and Bayesian-based approaches.

MLM solutions have been developed to overcome the confounding effects of population structure and the relatedness among individuals, and provide increased computational efficiency and statistical power [2-5]. Typical MLM solutions estimate effects based on single markers and require the use of covariances to account for population structure. However these approaches for controlling for population structure also decrease statistical power [6]. In contrast, Bayesian methods apply multiple variable regression models combined with prior distributions and Markov Chain Monte Carlo (MCMC) sampling to generate posterior distributions [7-9]. Meuwissen et al. (2001) first proposed the methods of ridge-regression BLUP, BayesA and BayesB. BayesB assumes marker effects have identical and independent univariate t-distributions and assume that a designated portion of markers have no effect. BayesC is similar to BayesB, but marker effects are assumed to have a common variance [10].

Genes with regulatory functions often exhibit high levels of expression variation across species [11, 12] as compared to metabolism-related genes [13]. Several studies have revealed that among primates transcription factors can evolve rapidly in response to selection [14-16]. Within species, genes exhibit different levels of variation in expression among individuals and alterations in the regulation of the expression of transcription factors can contribute to novel phenotypes [17], such as branching in maize [18] or pelvic loss in threespined stickleback fish [19].

Because variation in the regulation of gene expression contributes to phenotypic diversity [20], efforts have been made to identify genetic variants associated with variation in
transcript accumulation, i.e., eQTL analyses [21]. Genetic variants detected via eQTL analysis can act in cis or in trans. The relative contributions of cis- and trans-acting eQTL on phenotypic variation are unknown. Cis-variation is often considered a key mechanism in creating phenotypic novelty [22] and contributes to adaptive evolution [23-26]. Indeed, cis-effects have played a major role on gene expression during the domestication of maize [22]. It is worth noting, however, that due to limitations in statistical power it is typically more difficult to detect trans-acting eQTL than cis-acting eQTL [27]. Even so, many trans-eQTLs have been identified in maize [27, 28] and other species [29, 30].

Maize is one of the most genetically and phenotypically diverse species [31] and has a rich collection of genetic resources [32], making it an important model system. Because maize exhibits high levels of SNP diversity and low LD, it exhibits high statistical power and resolution in GWAS [33]. We used this model species to test the role of variation in the expression of transcription factors (TFs) and more generally variation in transcript accumulation on phenotypic variation. Following deep RNA-seq analysis of multiple tissues from diverse inbred lines we established that TFs are depleted among genes that exhibit high levels of expression variation across genotypes. Next, we developed a Bayesian-based statistical method for using RNA-seq measurements of transcript accumulation as the explanatory variables in GWAS and thereby directly demonstrated an association between variation in transcript accumulation of TFs and phenotypic variation for a diverse collection of traits.

Results

RNA-seq was conducted on mRNA extracted from multiple maize organs (seedling shoot apex, immature unpollinated ears, immature tassels, seedling shoots and roots)
collected from the 27 inbred founders of the NAM population. 6 billion raw 101-bp reads were generated, trimmed, filtered and aligned to the B73 reference genome (Methods). 2.9 billion non-stacking uniquely aligned reads were used to assay transcript accumulation levels (Table S1).

Identification of genes that are variably or stably expressed across tissues

To identify genes that exhibit extreme levels of variation in transcript accumulation across tissues, a series of model selection procedures was performed. Ultimately, we selected negative binomial distributions to model the distributions of read counts for genes, and the scaled log of over-dispersion parameters of quasi-negative binomial Generalized Linear Models to minimize the correlation between expression variation and expression levels (Methods). Henceforth, the scaled log10 over-dispersion parameters will be termed “variation in gene expression”. Of 39,656 high-confidence “Filtered-Gene Set” (FGS) genes, 29,609 genes have sufficient levels of transcript accumulation (Methods) to be used in subsequent analyses. The distribution of variation of gene expression across tissues was a left-skewed distribution (Fig. S1a). We defined the upper and lower 2.5% percentiles of this distribution as Tissue Variable Expression (T-VE) genes (N=741 genes) and Tissue Stable Expression (T-SE) genes (N=741; Fig. S1a). TFs as a group were enriched among T-VE genes (P-value = 0.008**) and homeobox (P-value = 0.03*) and MADS box families of TFs (P-value =5 × 10^{-5}**) were specifically enriched among T-VE genes (Methods). In contrast, TFs were depleted among T-SE genes (P-value = 0.0005**; Fig. S1b and Table S2c).

Identification of genes that are variably or stably expressed across genotypes

A similar approach was used to identify Genotype Variable Expression and Genotype Stable Expression (G-VE & G-SE) genes. The distribution of variation in gene expression
across genotypes demonstrated a left-skewed distribution (Fig. S2a). Although TFs were enriched among genes that exhibited higher than median levels of variation in gene expression (P-value = $3 \times 10^{-5**}$), TFs were underrepresented among both G-VE and G-SE genes (P-values = 0.002* and = 0.046*, respectively; Fig. S2b). Specifically, although 46 TFs would be expected among the G-VE by chance, only 22 were observed (Table S3). Similar results were obtained when the G-VE and G-SE genes were defined as being the upper and lower 5% and 10% of all genes.

*Arabidopsis thaliana* RNA-seq data generated by Kawakatsu *et al.* [34] (N=727 genotypes; 2016) were analyzed using similar approaches. Consistent with our maize results, TFs were depleted among G-VE (P-value = $1.0 \times 10^{-5**}$) and G-SE (P-value = 0.004**) genes in Arabidopsis (Fig. S3). As was observed for maize, Arabidopsis TFs were enriched among those genes that exhibit higher than median levels of expression variation across genotypes (P-value = $4 \times 10^{-8**}$).

**Correlation of variation in maize gene expression across genotypes and tissues**

A linear trend was observed between tissue-wise and genotype-wise variation in gene expression ($r^2 = 0.64$, P-value ~ $2 \times 10^{-16**}$ Fig. S4). Based on whether a gene demonstrated stable or variable variation of gene expression across tissues and genotypes, maize genes could be classified into 9 categories (Table S2). The 520 T-VE genes that are neither G-VE nor G-SE are significantly enriched in TFs overall (P-value = $9 \times 10^{-5**}$) and enriched in several specific TF families, including Homeobox/HOX (P-values = 0.02*), MADS (P-values = $2.6 \times 10^{-5**}$) and Squamosa promoter binding proteins/SPBs (P-values = 0.03*; Table S2e). Generally, HOX genes function in organ identity [35] and SBPs function in phase change [36]. In contrast, the 330 genes classified as being both T-SE and G-SE are depleted for TFs (P-value = 0.006**; Table S3).
Expression Read Depth Genome Wide Association Study (eRD-GWAS)

Based on the findings that TFs exhibited moderate variation in expression across genotypes, we were interested in testing the contribution of variation in transcript accumulation levels of TFs to phenotypic diversity. To directly test this association, we developed a Bayesian-based statistical approach for using transcript accumulation as the explanatory variable during GWAS.

Typically, GWAS is conducted using SNP genotypes as explanatory variables. We reasoned that using transcript accumulation as an explanatory variable for GWAS would have certain advantages in that gene expression levels potentially integrate the effects from multiple loci that contribute to phenotype variation. To the extent that these hidden multiple locus effects poorly explained by single genotyped SNPs, eRD-GWAS may better explain variation in trait values. eRD-GWAS also has the potential to integrate the effects of epigenetic variation that contributes to variation in gene expression and other traits. To test the hypothesis that variation in transcript accumulation can explain diversity in trait values that is missed by traditional GWAS, we analyzed a set of lines which had been both genotyped and phenotyped and for which RNA-seq data were available.

The SAM (Shoot Apical Meristem) diversity panel consists 369 diverse inbred lines, including commercially relevant inbreds with expired plant variety protection (PVP) [37]. We have genotyped this panel with 1.28M SNPs [37]. In addition, we conducted RNA-seq on apex tissue (which includes the shoot apical meristem or SAM) from each of these inbreds [37]. Using these RNA-seq data we calculated RPKM values for each of the 39,656 FGS genes in the maize genome for each of the inbreds in the SAM panel.

Each of the inbreds in the SAM diversity panel had previously been phenotyped for multiple traits related to the shoot apical meristem [38], viz., volume, height, parabola radius,
arc length and SAM surface area [37] and a variety of other traits including the mean node number [37], ear height and days to anthesis (DTA)[39]. During the current study we phenotyped these inbreds for five additional traits, viz., stalk circumference, stalk cross-sectional area, maximum and minimum stalk diameter and number of nodes with brace roots. These traits exhibit varying degrees of correlation (Fig. S5), some of which have been reported previously [40]. To test the hypothesis that eRD-GWAS can identify loci that contribute to variation in traits that are not identified by traditional SNP-based GWAS, we analyzed all five SAM-related and eight other traits using both SNP genotypes and RPKM values as explanatory variables (Methods). Typically, GWAS software that relies on MLMs is designed to use SNPs as the explanatory variables. We elected to use a BayesB-based approach to conduct eRD-GWAS in which RPKM values (expression data) served as the explanatory variables. Our rationale for selecting a Bayesian approach to GWAS is described in the methods.

The BayesB model is a widely used in genomic selection. Instead of predicting phenotype, we used model frequency (the frequency with which a gene was included in a model) as a measure of the strength of the relationship between that gene’s expression pattern and the phenotype of interest. To validate the BayesB approach we repeated Leiboff’s SNP-based analysis of SAM volume using a MLM approach and in parallel conducted a SNP-based GWAS for SAM volume using a BayesB approach (Methods). As expected the results we obtained from our SNP-based GWAS using the MLM approach (Fig. 1) were very similar to those of Leiboff et al. (2015). The upper and middle panels of Figure 1 provide results from the SNP-based MLM GWAS and the SNP-based BayesB GWAS. The 14 significant signals that overlap between the two approaches are indicated by vertical dashed lines on
chromosomes 1, 2, 6, 7, 9 and 10. 9 of these 14 SNPs that were detected via both approaches are located in or near genes that have been shown previously to be associated with SAM volume [37]. If we consider SNPs present in the same genomic regions (to account for LD), 19 of the 54 SNPs detected by SNP_MLM were present in 30 kb windows centered on SNPs detected by SNP_BayesB. Similarly, 15 of 53 SNPs detected by SNP_BayesB were present in 30 kb windows centered on SNPs detected by SNP_MLM. These results established that the BayesB approach identified a significant subset of those SNPs identified by MLM GWAS, but that the BayesB approach also identified signals not identified by the MLM approach.

Based on these results we used BayesB-based eRD-GWAS to identify genes whose variation in transcript accumulation is associated with diversity in SAM volume. Approximately 500 genes (lower panel of Fig. 1) exceed the arbitrarily selected model frequency cutoff of 0.02 in the eRD-GWAS. If we search for candidate genes, GRMZM2G140721 is detected by both the SNP-based BayesB and eRD-GWAS. GRMZM2G140721 is a predicted transcriptional factor in Arabidopsis, rice and maize. In total, 120 genes identified via eRD-GWAS (i.e., eRD genes) were not located within 30 kb windows centered on the chromosomal positions of SNPs identified via either SNP-based GWAS approach (MLM or Bayesian). Even so, some of these genes detected via eRD-GWAS but not by SNP-based GWAS have previously been demonstrated to affect the morphology of the SAM. For example, ZEA CENTRORADIALIS4 (ZCN4) functions in the maintenance of indeterminate shoot meristem, thereby affecting the transition to an inflorescence meristem [41] and BRANCH ANGLE DEFECTIVE 1 (BAD1) [41] is a TCP class II gene that is expressed in inflorescence meristems and lateral organs where it functions to promote cell proliferation.
All GWAS provide lists of genes that are hypothesized to be associated with traits of interest. To assay the accuracy of the gene/trait associations from eRD-GWAS we performed a series of analyses including, 10-fold cross-validation, eQTL analyses of eRD-GWAS genes, tests for the enrichment of eRD-GWAS genes within specific nodes of RNA co-expression networks, protein-protein interaction networks, and gene regulatory networks.

**10-fold cross-validation**

10-fold cross validation is a technique used for assessing the accuracy of prediction models [42]. Our 10-fold cross-validation analyses of the results of eRD-GWAS (Methods) yielded accuracies of 0.41-0.76, indicating that eRD-GWAS accurately detects associations between variation in transcript accumulation and multiple traits (Table S4). Based on comparisons to similar cross-validation analyses conducted using results from SNP BayesB, the accuracies of the two approaches are similar for multiple traits (Table S4).

**eQTL for eRD-GWAS-detected genes (eRD genes)**

If eRD-GWAS is accurately identifying genes that contribute to variation in a trait, we would expect that eQTL that act in trans to regulate the expression of eRD genes may also be associated with variation in that trait. Hence, we conducted an eQTL analysis using an MLM approach (Methods) for the five eRD genes associated with the DTA trait that had the highest model frequencies. The resulting eQTL were compared to the eRD genes and also to the genes associated with the DTA trait via BayesB GWAS (Table S5). Hypergeometric analyses (Methods) established that the eQTL were enriched in genes associated with variation in the DTA trait. To ensure this phenomenon was robust across traits, we used the same strategy on multiple traits (Table S6). The results were consistent with our hypothesis,
i.e., that the eQTL associated with specific eRD genes are enriched overall in eRD genes. This result provides support for the view that eRD-GWAS accurately identifies genes whose expression is associated with variation in trait values.

**eRD-GWAS enriched in an RNA Co-expression network**

To enhance the power of this analysis we first constructed an RNA co-expression network using WGCNA [43] using the RNA-seq data from the SAM diversity panel. We then determined GO terms that were enriched among the genes within specific modules of the co-expression network (Table 1). The modules that were enriched for eRD genes associated with the DTA trait were also enriched for a variety of GO categories. The “honeydew" module was enriched for the GO category “maintenance of floral meristem identity” which would appear to be relevant to the DTA trait. Other modules were enriched for categories that the literature reported may relevant to the DTA trait, such as “metal ion transport”, “response to nitrate” and “NAD(P) metabolic” [44-46].

**eRD-GWAS in Protein-Protein interaction network (PPIN)**

Protein-protein interaction networks (PPIN) can be used to identify proteins (and genes) that contribute to phenotypes and thereby help elucidate complex genetic mechanisms [47]. We downloaded maize PPIN data from the maize PPIM [48], clustered proteins into network communities and then tested whether eRD genes were enriched in network communities. As was the case for the enrichments tests within the RNA co-expression network, eRD genes were significantly enriched (Methods) GO categories associated with the DTA trait among three of the 12 network communities that contained more than 1 eRD gene (Table 2 and Fig. 2). This finding provides further evidence that eRD-GWAS can identify biologically relevant gene/trait associations.
**eRD genes in Gene Regulatory Network**

Unlike co-expression network, a gene regulatory networks (GRN) is composed of directed edges that indicate biological relationships between pairs of nodes. For example, regulators are predicted to activate or suppress downstream genes. We examined the characteristics of our eRD genes within maize GRNs constructed using RNA-seq (23 tissues) or proteomic (33 tissues) data [49], eRD genes were enriched among regulators in both the RNA- and protein-based GRNs (Fig. 3 a-c). Sets of eRD-GWAS genes selected using model frequency cut-offs larger than 0.03 have enrichment test p-values smaller than 0.05, indicating that the targets of eRD-GWAS regulators are themselves enriched in eRD-GWAS genes (Fig. 3 a-c). These results indicate that eRD-GWAS can identify both GRN regulators and their downstream targets.

**TFs are enriched among trait-associated genes from eRD-GWAS**

As discussed earlier, TFs are enriched among genes that exhibit a higher than median level of variation in gene expression across genotypes. To test the hypothesis that the variation in expression of TFs affects phenotype, we conducted enrichment tests for TFs among eRD genes associated with 13 phenotypes using various model frequency cutoffs (Fig. 3 d-g and Fig. S6). For 11/13 traits, as the stringency of model frequency cutoffs was increased, the enrichment of TFs among the eRD-GWAS genes also increased. This result demonstrates the importance of variation in the expression of TFs on phenotypic variation.

**Discussion**

We were interested in comparing the variation in expression of TFs across tissues and genotypes to that of other genes. Using an RNA-seq data set derived from five tissues and 27
genotypes, we identified genes that exhibit low and high levels of expression variation across tissues (T-VE and T-SE genes) and genotypes (G-SE and G-VE). T-VE genes are enriched in TFs, and specifically enriched for Homeobox, MADS and Squamosa promoter binding proteins. In contrast, T-SE genes are depleted for TFs.

In contrast to what was observed across multiple tissues, TFs were depleted among the G-VE and G-SE genes of both maize and Arabidopsis. Even so, in both species, TFs were enriched among those genes that exhibit higher than median levels of variation in gene expression. Interesting, even though there is positive correlation between maize genes that exhibit high levels of expression variation across genotypes and tissues, TFs are not enriched among G-VE genes that are also T-VE. Based on these findings we hypothesize that extreme variation in expression of TFs across genotypes is constrained by selection against the extreme phenotypic variation that would be expected to arise via the action of TFs with extreme expression levels upon multiple downstream target genes. Similarly, because the NAM founders exhibit substantial phenotypic diversity, the depletion of TFs among the G-SE genes is consistent with a role of TF in contributing to phenotypic diversity.

**Overview of eRD-GWAS**

To test the hypothesis that variation in the expression of TFs (and other genes) across genotypes contributes to phenotypic variation, we developed eRD-GWAS, a statistical method that permits gene expression level to be tested as an explanatory variable during GWAS.

Using eRD-GWAS we detected several hundreds of trait-associated genes for each of multiple traits included in this study. The results of 10-fold cross validation indicated that the predicted phenotypes based on genes detected via eRD-GWAS are highly correlated with
empirically measured phenotypes. In addition, many trait-associated genes have annotations consistent with their presumed roles in regulating the associated traits (Table S6). Hence, we concluded that eRD-GWAS pipeline can successfully identify associations between variation in gene expression and diversity in phenotype. eQTL analyses of eRD-GWAS genes, tests for the enrichment of eRD-GWAS genes within specific nodes of RNA co-expression networks, protein-protein interaction networks, and gene regulatory networks provided further support for this conclusion.

**Challenges Associated with GWAS**

GWAS strategies identify genes that putatively contribute to variation in phenotypes. However, false positive (FP) results remain a challenge in GWAS [50]. The use of other types of genomic data in combination with SNP data has the potential to decrease biases and increase the power to detect true associations in GWAS. For example, efforts have been made to make use of eQTL results to increase the accuracy of GWAS [51, 52]. Although including eQTL results has the potential to decrease the rate of false positive associations, this approach can also result in elevated rates of false negative calls [50].

An alternative approach which we employed in this study, is to use gene expression levels directly as explanatory variables for GWAS. This approach substantially reduces the multiple testing problem by using as explanatory variables expression data from ~40,000 maize genes vs. millions of available SNPs. This reduction in the number of explanatory variables also reduces the computational cost of eRD-GWAS as compared to traditional SNP-based approaches.

Another group has shown that RNA expression patterns can predict human disease [20]. However, their statistical framework was intolerant of missing data which required that
transcriptomic data be imputed based on SNP data. This imputation would be expected to decrease accuracy. Further, their approach is limited to binary phenotypes (e.g., healthy vs. diseased). Jin et al. [53] also attempted to associate phenotypes with expression patterns. For a given gene, they classified lines as either being expressed or not based on RNA-seq data. Lines having intermediate levels of expression were treated as missing data. The conversion of continuous gene expression data into a binary classification scheme would be expected to decrease statistical power [54]. Because the data of Jin et al. (2016) were analyzed using an MLM approach, the limitations discussed in the Introduction apply. In contrast to the method of Jin et al. (2016), eRD-GWAS does not require that lines with intermediate expression levels be treated as missing data. In addition, our statistical framework is not limited to binary phenotypes as is the case for Gamazon et al. (2015). This is important because most important traits exhibit quantitative variation.

Because eRD-GWAS directly associates candidate genes with phenotypes, it eliminates the need to hunt for causative genes within windows surrounding trait-associated SNPs. One potential concern with eRD-GWAS is whether LD creates false trait associations between the expression of a gene that is simply linked to the causative gene. The Bayesian framework employed by eRD-GWAS functions to distinguish the effects of LD loci, our data suggest that this is in fact true not only for SNPs but also for expression data. For example, even though the expression patterns of various alleles of ZmMADS69 are correlated with the expression patterns of other genes within the adjacent 1-Mb window (as well as genes across the genome), eRD-GWAS still could detect ZmMADS69 as the gene with the highest model frequency for flowering time (Fig. S7).
Before using expression data to conduct eRD-GWAS, it is necessary to align RNA-seq reads to a reference genome. The substantial amount of SNPs [37] and structural polymorphism among maize haplotypes [55] may result in alignment biases that distort RPKM values and hence the power of eRD-GWAS. Although this bias did not interfere with our ability to detect trait-associated loci, the use of new alignment approaches that better control for polymorphisms [56] may provide additional power to eRD-GWAS.

This study included a direct comparison between the use of SNPs and expression data as explanatory variables within a common statistical framework. Our results establish that the two types of explanatory variables provide different association signals, such that some signals are detected by only one type of explanatory variable. This result argues that eRD-GWAS is complementary with SNP-based GWAS.

The Bayesian approach requires the selection of a model frequency cut-off which unlike the q-value associated with MLMs is in some sense arbitrary. If our selected model frequency cut-off (0.02) had been too relaxed it is unlikely that strong statistical evidence for module-specific enrichment within the co-expression and protein-protein interaction networks would have been observed. Nor would we have been likely to observe a statistically significant enrichment of eQTL for eRD genes among the eRD genes. If the selected model frequency cut-off were more stringent (i.e., if a larger model frequency) fewer genes would have been called as being associated with a given phenotype. This relationship was explored in Figure 3 and Fig. S6 which demonstrated that the enrichment of TFs among the eRD genes for multiple traits is robust across a wide range of model frequency cut-offs, but that the enrichment p-value can become more significant at increasingly stringent model frequency cut-offs. This finding is consistent with the hypothesis that a more stringent cut-off would
result in a higher proportion of true positives, although presumably at the cost of more false negatives.

Transcription factors contribute significantly to phenotypic variation

Variation in gene expression contributes to phenotypic variation [57] upon which natural and artificial selection can act. The mechanisms that regulate variation in gene expression can act in cis (e.g., transcription binding sites) or trans (e.g., transcription factors). It has, for example, been shown that variants located upstream of maize genes are enriched in GWAS analyses of multiple morphological traits [58]. Similarly, GWAS signals are enriched near human TF binding sites [59]. These findings are at least consistent with the hypothesis that variation in transcription factor binding sites contributes to phenotypic variation.

It is also likely that variation in the expression of TFs per se can contributes to phenotypic variation, and indeed specific cases of this type have been identified [18, 60]. Previous case studies have revealed roles for TFs in phenotypic evolution [61, 62] In addition, genome-wide comparative genomics studies among primates have demonstrated that genes responsible for directional/diversifying selection are often TFs [11, 12, 63, 64]. As a step towards testing the hypothesis that TFs contribute substantially to phenotypic variation in maize, we demonstrated that TFs exhibit elevated levels of variation in expression across genotypes. More directly, using our newly developed eRD-GWAS method we established that genes associated with phenotypic variation for multiple traits are enriched in TFs, demonstrating that variation in the expression of TFs contributes substantially to phenotypic diversity in maize.
Conclusions

TFs are enriched among genes with the most variation in expression across tissues and among genes with higher than median levels of variation in expression across genotypes. To better understand the relationship between variation in gene expression on phenotypes, we developed eRD-GWAS, which identifies associations between variation in gene expression and variation in phenotypes or traits. The enrichment of TFs among trait-associated genes identified via eRD-GWAS highlights the impact of variation in expression on phenotypes. eRD-GWAS is complementary with SNP-based GWAS.

Methods

Tissue collection, library preparation and RNA sequencing

Maize shoot apex, immature, unpollinated ears, immature tassels, and seedling shoots and roots of 27 NAM founders were collected for RNA extraction (Table S1). There exists a universal dilemma of sampling tissues from genotypes with different maturities. One must either sample from a common environment (same harvest date) and accept variation in developmental stage at harvest, or harvest at a common developmental stage and accept the risk of differences in micro-environment at harvest. For the NAM RNA-Seq experiment we elected to use the second approach.

Ear and tassel were harvested from greenhouse-grown plants with the exception of Ms71 ears which were harvested from field-grown plants. Immature ear tips were harvested ~68 days after planting (depending on the maturity rate of each line). At this stage ear ranged from 0.5 to 3 inches; only the top 1/3 ~1/5 of each ear were collected. Tassels were harvested prior to tassel emergence, i.e., ~60 days after planting. Three healthy plants were sampled and pooled per genotype prior to homogenization in liquid nitrogen and RNA extraction.
Maize shoot apexes were collected by pooling 3-6 14-day old seedlings from each NAM founder. Seedling were grown by planting 10 kernels of each line in germination paper which was rolled and placed in a tall plastic beaker filled with approximately 3 inches of tap water. Beakers were covered with “Cling-wrap” and placed in a dark 28°C incubator, for approximately 4-5 days, when shoots emerged from the germination paper. 2-3 inches of the shoot and root were cut and frozen in liquid nitrogen for immediate homogenization and extraction. Samples from three plants of each inbred were pooled for homogenization. For the SAM diversity panel, all plants were grown and sampled under Leiboff et al. (2015) study.

All RNA extractions were performed with the Qiagen RNeasy kit, according to manufacture protocol. RNA was eluted twice with 30 ul RNase free water. Indexed RNA-seq libraries were prepared using the Illumina protocol outlined in “TruSeq RNA Sample Preparation Guide” (Part# 15008136 Rev. A, November 2010). Maize shoot apex RNA was sequenced with Illumina Genome Analyzer II instrument while ear, root, shoot and tassel RNA were sequenced with Illumina HiSeq 2000 instrument.

**RNA-seq reads: Processing, alignment and SNP calling**

The quality trimming, alignment to the B73 reference genome and SNP calling were as described by Leiboff et al. (2015) [37]. Alignment coordinates of confidently (uniquely) mapped reads within the same chromosomal regions were compared for potential read stacks caused by PCR artifacts during sequencing. If a stack consisting of 2 or more reads with identical start and end positions were detected, only a single read with best alignment score (least number of mismatches and least number of ambiguous bases) was selected for variant detection. If the distance from the left base pair to right base pair was more than 12,000 bp,
the reads/read pairs were further removed. Reads with non-canonical splice sites were also removed.

**Discovery and annotation of expression variable/stable genes**

Read counts are discrete and usually exhibit correlation between mean and variance [65]. Proper models, techniques and summary statistics are essential to evaluate expression variation. To reduce ascertainment bias between expression level and expression variability, Pearson correlations were computed between expression level and each of several summary statistics (Fig. S8), including over-dispersion parameter of the Poisson model [66], mean coefficient of variance based on the Poisson model, deviance of the Negative Binomial model [67] and over-dispersion parameter of quasi-Negative Binomial model [66]. The R packages edgeR (version 3.14.0) [68] and QuasiSeq (version 1.0-8) [66] were used to estimate dispersion parameters and over-dispersion parameters of quasi-Negative Binomial GLMs (some graphical display used ggplot2, version 2.2.1 [69]). Full models were fitted when comparing Poisson, Negative Binomial and quasi-Negative Binomial GLMs, as follows:

\[
\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + o_{ijk}
\]

where \(\lambda_{ijk}\) is mean fragment count for genotype i, tissue j and observation k, \(\mu\) is an intercept parameter, \(\alpha_i\) is an effect of genotype i, \(\beta_j\) is an effect of tissue j and \(o_{ijk}\) is the normalization offset for genotype i, tissue j and observation k.

Of the four measures of variation discussed above, the over-dispersion parameter of quasi-Negative Binomial model, which measures the deviation of a gene’s read counts from the best-fitting Negative Binomial distribution, had the smallest correlation with expression
level, and was thus used to measure expression variability (Fig. S8). The over-dispersion parameter $\Phi$ of Quasi-Negative Binomial GLMs is

$$
\Phi = \frac{\text{Var}(Y)}{\kappa E(Y)^2 + E(Y)}
$$

where $Y$ is fragment count for a gene, $\text{Var}(Y)$ and $E(Y)$ are the variance and expectation of $Y$ respectively, and $\kappa$ is the dispersion parameter of a negative binomial GLM. Tissue-wise over-dispersion parameters were estimated with genotype as the only factor in the model, while genotype-wise over-dispersion parameters were estimated treating tissue as the only factor in the model. A total of 29,609 genes with mean read counts $\geq 5$ and numbers of samples with zero read counts $\leq 2$ were included in the analysis. Z-score normalization was performed against log transformed over-dispersion parameter estimates, where

$$
Z = \frac{\log(\hat{\Phi}) - \bar{E}(\log(\Phi))}{\sqrt{\text{Var}(\log(\Phi))}}
$$

upper and lower 0.025 quantiles of transformed normalized distributions were used to define highly variable and highly stable genes. MAPMAN annotation of maize filter gene sets (5b.61) was used to perform functional enrichment tests [70]. Fisher-exact test was performed with Benjamini-Hochberg method controlling false discovery rates (FDRs).

**Collection of Phenotypic data**

Phenotypic trait data were collected from a panel of 369 diverse inbreds designated as the “SAM Panel” [37]. Data were collected from 3 plants per location in two fields grown in Ames, Iowa during the summer of 2014. Prior to data collection leaf sheaths and brace roots (if present) were removed. Measured traits included maximum and minimum stalk diameters, stalk circumference, stalk cross sectional area, total node number, and number of nodes with
brace roots (Table S7). Additional data from the SAM Panel (or members of it) were obtained from the literature. For example, several traits associated with the SAM, including SAM height, radius, surface area, volume and arc length from P1 notch to apex were obtained from [37]. Ear height and DTA data were obtained from [39]. Phenotypic regression and phenotypic density distributions were conducted using the R “corrgram” package version 1.10 [71].

Mixed Linear Models GWAS

GAPIT version 3.35 [72] was used for MLM GWAS. The model implemented in GAPIT was

\[ y = W\nu + X\beta + Z\upsilon + e \]

where \( y \) is the phenotype value, \( \nu \) and \( \beta \) are unknown fixed effect vectors, and \( \upsilon \) is a vector of random effects that follows a multivariable normal distribution with a null mean and a covariance matrix of \( G \). \( G = K\sigma^2_\delta \) where \( K \) is the kinship matrix [2]. \( e \) follows a normal distribution with null mean and \( \sigma^2_e I \) variance. In general \( W, X \) and \( Z \) are the matrices containing principal component scores that account for population structure, known covariates and SNP genotypes, respectively. In our case, \( W \) contains scores for the first three principal components, \( X \) was not used because we had no known covariates to adjust for, and \( Z \) had data on 1.28M SNPs. Manhattan plots were generated from our in-house R scripts based on the p-value from GAPIT result. The cutoff was arbitrarily set on \( 10^{-7} \). Other settings followed GAPIT’s default.
Bayesian-based GWAS

We selected a Bayesian approach for exploring the relationship between gene expression and phenotype, rather than a MLM approach for two major reasons. First, the multivariate Bayesian framework internally controls for the effects of other genes by testing whether the inclusion of a given marker (i.e., the expression level of a given gene) can explain more genetic variance in each MCMC (Markov chain Monte Carlo) iteration. Although it may be possible to fit all the markers (i.e., gene expression levels of all genes) simultaneously by iterating an MLM approach this would be time consuming. In contrast, this feature is “baked into” the Bayesian approach. Equally important, population structure can be controlled automatically via Bayesian approaches that include multiple genes in each MCMC (Markov chain Monte Carlo) iteration [73]. In contrast, population structure information is required to control false-positives as covariances in MLM, which can decrease statistical power.

Multiple genomic selection models were constructed employing different values of $\pi$ (the proportion of SNPs, assumed to have no effect on phenotype). The accuracies of these various models were evaluated using 10-fold cross validation and heritability. We selected for each phenotype a value of $\pi$ that yielded the maximum accuracy based on 10-fold cross validation that has a heritability that is not so high as to raise concerns of over-fitting. This had the effect of thinning the number of predictors, resulting in a more limited number of descriptors, similar to the output of GWAS. Our approach differs from MLM GWAS in that rather than using a p-value to reflect the strength of the relationship between a marker and a phenotype, we used the model frequency (the frequency with which a gene was included in a model) to reflect the strength of the relationship between that gene’s expression and the phenotype of interest.
The Bayesian-based GWAS was constructed using GenSel v4.1 [10] BayesC and BayesB methods. The model in GenSel was
\[ y = X\beta + Z\upsilon + e \]
where the \( X, Z, \beta \) and \( \upsilon \) are the same as in the MLM model, \( e \) follows a normal distribution with null mean and covariance matrix \( \sigma^2_e R \) (\( R \) is a diagonal matrix), \( \sigma^2_a \) and \( \sigma^2_e \) have independent inverse Chi-square priors with degree of freedom 4 and scale parameters set to 50% of phenotypic variation as prior. For Bayes B (eRD-GWAS) and BayesC (SNP-based GWAS), the fraction (\( f \)) of markers having no effect was set at 0.9996 and 0.995, respectively. We used a chain length of 41,000 and discarded 1,000 iterations as burn-in run. Significance cut-offs for SNP_BayesB and eRD-GWAS were set as model frequencies of 0.01 and 0.02, respectively. Then we used genetic variance and error variance posteriors from BayesC as priors in BayesB; other settings were as above. The accuracy of Bayesian-based GWAS results were estimated via 10 fold cross-validation.

Cross-validation, enrichment tests, network visualization and GO enrichment

10 fold cross-validations were conducted using the R “cvTools” package version 0.3.2 [74]. Enrichment test P-values were based on hypergeometric distributions. Network visualization was conducted using “MANGO” software version 1.20 [75]. Clustering was conducted using the fastgreedy community method [76]. GO term enrichment analyses were conducted using the GOseq package version 1.20.0 [77]. Functional enrichment tests were based on MAPMAN annotations. The list of TFs used in the enrichment tests were obtained from the “Grassius database” [78]. A list of Arabidopsis thaliana TFs was downloaded from “AGRIS” [79]. P-values for TF enrichment were obtained from single-tailed Fisher tests.
Acknowledgements

Dr. Wei Wu (formerly of the Schnable Lab at Iowa State University) generated the RNA-Seq data from the NAM founders and Cheng-Ting Yeh (Schnable Lab, Iowa State University) conducted bioinformatics processing of these data. Drs. Heng-Cheng “Alvis” Hu, Wei Huang and Kai Ying (all formerly of the Schnable Lab at Iowa State University) provided useful suggestions during the early stages of the gene enrichment studies. Drs. Rohan L Fernando and Dorian Garrick (both of Iowa State University) provided useful input during the development of eRD-GWAS. We thank Drs. Nathan Springer (University of MN) and James Schnable (University of NE-Lincoln) for helpful discussions and suggestions.

References


### Tables

**Table 1.** GO enrichment tests of RNA Co-expression modules containing multiple eRD genes for the DTA trait.

<table>
<thead>
<tr>
<th>Module</th>
<th>GO term enrichment</th>
<th>No. eRD gene (% of eRD genes)</th>
<th>eRD genes log2 odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thistle3</td>
<td>Metal ion transport; Transferring phosphorus-containing groups; ATP binding</td>
<td>20 (54.1%)</td>
<td>9.17**</td>
</tr>
<tr>
<td>Navajowhite2</td>
<td>NAD(P) metabolic</td>
<td>18 (34.6%)</td>
<td>8.03**</td>
</tr>
<tr>
<td>Firebrick4</td>
<td>Nitrate transport; Magnesium ion binding</td>
<td>15 (35.7%)</td>
<td>8.38**</td>
</tr>
<tr>
<td>Palevioletred3</td>
<td>terpene synthase; regulation of transcription; response to nitrate</td>
<td>6 (11.1%)</td>
<td>6.33**</td>
</tr>
<tr>
<td>Honeydew</td>
<td>cell wall organization; maintenance of floral meristem</td>
<td>4 (11.1%)</td>
<td>6.92**</td>
</tr>
</tbody>
</table>

**Note:** ** indicates P-value of enrichment test < 0.01

**Table 2.** GO enrichment among Protein-Protein interaction network communities that contain multiple eRD genes for the DTA trait.

<table>
<thead>
<tr>
<th>Community No.</th>
<th>GO term enrichment</th>
<th>No. eRD gene (% of eRD genes)</th>
<th>eRD genes log2 odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>ATP biosynthesis process; Metal ion transport</td>
<td>8 (7.41%)</td>
<td>4.75**</td>
</tr>
<tr>
<td>6</td>
<td>MADS-gene family; floral meristem maintain</td>
<td>5 (8.93%)</td>
<td>5.96**</td>
</tr>
<tr>
<td>4</td>
<td>Oxidation-reduction process; nitrate assimilation; Steroid 22-alpha hydroxylase activity (BR)</td>
<td>12 (4.67%)</td>
<td>2.58*</td>
</tr>
</tbody>
</table>

**Note:** * indicates P-value of enrichment test < 0.05

**Note:** ** indicates P-value of enrichment test < 0.01
Fig. 1 Manhattan plots of three types of GWAS results. Upper panel reports result from a SNP-based MLM implemented in GAPIT. Only signals with p-values smaller than $1.0 \times 10^{-7}$ are presented. The middle and lower panels report results from the SNP-based BayesB analysis and eRD-GWAS, respectively. The model frequency cutoffs for SNP BayesB and eRD-GWAS are 0.01 and 0.02, respectively (Methods). Overlapping associated SNPs in the upper two panels are indicated by dashed lines. Note not all overlapping SNPs can be distinguished in this plot. Gene IDs of some trait associated genes (Methods) are indicated.
Fig. 2 Visualization of protein-protein interaction network that contains eRD genes. Highlighted Communities that contain more than one eRD gene and in which eRD genes are statistically enriched are highlighted.
Fig. 3 Enrichment testing for eRD genes. a-c) enrichment of "regulators" among eRD genes associated with the DTA trait at various model frequency cutoffs. d-g) show the enrichment of TFs among eRD genes for various traits at various model frequencies. The number of eRD genes above indicated model frequencies cutoffs are shown in each plot. The red dashed lines indicate p-values of 0.05.
Figure S1. Identification and functional enrichment of T-VE and T-SE genes. a) Histogram of expression variation across tissues; dashed line corresponds to upper and lower 2.5% percentiles of the distribution. b) Functional enrichment of significant categories within T-VE and T-SE genes.
Figure S2. Identification and functional enrichment of G-VE and G-SE genes. a) Histogram of expression variation across tissues; dashed line corresponds to upper and lower 2.5% percentiles of the distribution. b) Functional enrichment of significant categories within G-VE and G-SE genes.
Figure S3. Identification and functional enrichment of G-VE and G-SE genes in Arabidopsis.

A) Histogram of expression variation across tissues; dashed line corresponds to upper and lower 2.5% percentiles of the distribution. B) Functional enrichment of significant categories within G-VE and G-SE genes.
Figure S4. Scatterplot of variation in expression across genotypes and tissues. Colors designate expression levels, with blue the lowest and red the highest; the red dashed line is the diagonal, black dashed lines corresponded to upper and lower 2.5% percentiles for variation in expression.
Figure S5. Correlations among 13 traits.
Figure S6. Enrichment testing for eRD genes among 9 traits. This figure shows the enrichments of TFs among eRD genes for various traits at various model frequencies. The number of eRD genes above indicated model frequencies cutoffs are indicated within each plot. The red dashed lines indicate p-values of 0.05.
Figure S7. The effect of expression LD on eRD-GWAS. a) A heatmap showing correlation values among RPKM values for 112 genes. The red rectangle is a 1 Mb window centered on the physical position of the ZmMADS69 gene. The remainder of the figure contains 100 randomly chosen genes. b) Relationship between model frequencies from eRD-GWAS for the DTA trait and the correlation between RPKM values of genes and the DTA trait within the 1 Mb window.
Figure S8. Selection of parameters for measuring expression variation. Pairwise correlations between expression levels (FPKM), the over-dispersion parameter of Poisson (poisson), Coefficient of Variation (cv), dispersion of Negative Binomial models (nb) and over-dispersion parameters of Negative Binomial models (qnb) across genotypes (left panel) and tissues (right panel).
## Supplemental Tables

Table S1. Summary of RNA-seq processing.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Tissue</th>
<th>Read Length</th>
<th># of Raw PE. Frags.</th>
<th># after trimming PE. Frags.</th>
<th># of aligned frags PE. Frags.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sing.</td>
<td>Sing.</td>
<td>Sing.</td>
</tr>
<tr>
<td>B73</td>
<td>apex</td>
<td>76</td>
<td>0</td>
<td>30,674,152</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>ear</td>
<td>101</td>
<td>29,645,539</td>
<td>0</td>
<td>26,588,836</td>
</tr>
<tr>
<td></td>
<td>root</td>
<td>101</td>
<td>26,346,272</td>
<td>0</td>
<td>23,714,272</td>
</tr>
<tr>
<td></td>
<td>shoot</td>
<td>101</td>
<td>24,696,725</td>
<td>0</td>
<td>21,926,968</td>
</tr>
<tr>
<td></td>
<td>tassel</td>
<td>101</td>
<td>23,607,132</td>
<td>0</td>
<td>21,026,158</td>
</tr>
<tr>
<td>B97</td>
<td>apex</td>
<td>76</td>
<td>0</td>
<td>27,299,903</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>ear</td>
<td>101</td>
<td>18,239,276</td>
<td>0</td>
<td>16,555,436</td>
</tr>
<tr>
<td></td>
<td>root</td>
<td>101</td>
<td>25,488,003</td>
<td>0</td>
<td>18,003,838</td>
</tr>
<tr>
<td></td>
<td>shoot</td>
<td>101</td>
<td>26,744,051</td>
<td>0</td>
<td>21,527,498</td>
</tr>
<tr>
<td></td>
<td>tassel</td>
<td>101</td>
<td>29,354,763</td>
<td>0</td>
<td>26,373,780</td>
</tr>
<tr>
<td>CML103</td>
<td>apex</td>
<td>76</td>
<td>0</td>
<td>21,199,050</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>ear</td>
<td>101</td>
<td>26,358,175</td>
<td>0</td>
<td>23,498,541</td>
</tr>
<tr>
<td></td>
<td>root</td>
<td>101</td>
<td>25,743,306</td>
<td>0</td>
<td>19,837,508</td>
</tr>
<tr>
<td></td>
<td>shoot</td>
<td>101</td>
<td>23,598,599</td>
<td>0</td>
<td>20,271,349</td>
</tr>
<tr>
<td></td>
<td>tassel</td>
<td>101</td>
<td>26,650,791</td>
<td>0</td>
<td>23,319,536</td>
</tr>
<tr>
<td>CML228</td>
<td>apex</td>
<td>76</td>
<td>0</td>
<td>13,533,836</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>ear</td>
<td>101</td>
<td>27,020,600</td>
<td>0</td>
<td>23,323,929</td>
</tr>
<tr>
<td></td>
<td>root</td>
<td>101</td>
<td>27,358,041</td>
<td>0</td>
<td>23,569,439</td>
</tr>
<tr>
<td></td>
<td>shoot</td>
<td>101</td>
<td>26,301,690</td>
<td>0</td>
<td>23,574,287</td>
</tr>
<tr>
<td></td>
<td>tassel</td>
<td>101</td>
<td>24,590,006</td>
<td>0</td>
<td>21,935,993</td>
</tr>
<tr>
<td>CML247</td>
<td>apex</td>
<td>76</td>
<td>0</td>
<td>29,305,872</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>ear</td>
<td>101</td>
<td>29,489,145</td>
<td>0</td>
<td>27,258,889</td>
</tr>
<tr>
<td></td>
<td>root</td>
<td>101</td>
<td>25,658,069</td>
<td>0</td>
<td>21,826,964</td>
</tr>
</tbody>
</table>
Table S2. Functional enrichment and Go term enrichment tests. a. Functional enrichment tests of genes that exhibit extreme expression variation across genotypes. b. Go term enrichment tests of genes that exhibit extreme expression variation across genotypes. c. Functional enrichment tests of genes that exhibit extreme expression variation across tissues. d. Go term enrichment tests of genes that exhibit extreme expression variation across tissues. e. Functional enrichment tests of genes that exhibit extreme expression variation across genotypes and tissues. f. Go term enrichment tests of genes that exhibit extreme expression variation across genotypes and tissues.

<table>
<thead>
<tr>
<th>G-VE Layer</th>
<th>p-value (BH adjusted)</th>
<th>log2 odds ratio</th>
<th>cat function</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.00160</td>
<td>1.18</td>
<td>cell wall</td>
</tr>
<tr>
<td></td>
<td>0.00160</td>
<td>0.80</td>
<td>misc</td>
</tr>
<tr>
<td></td>
<td>0.00237</td>
<td>-0.64</td>
<td>protein</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>2.51</td>
<td>PS</td>
</tr>
<tr>
<td></td>
<td>0.00388</td>
<td>-0.63</td>
<td>RNA</td>
</tr>
<tr>
<td></td>
<td>0.00235</td>
<td>1.24</td>
<td>secondary metabolism</td>
</tr>
<tr>
<td>L2</td>
<td>0.04216</td>
<td>2.07</td>
<td>cell wall.cell wall proteins</td>
</tr>
<tr>
<td></td>
<td>0.04474</td>
<td>1.87</td>
<td>cell wall.modification</td>
</tr>
<tr>
<td></td>
<td>0.04714</td>
<td>4.07</td>
<td>lipid metabolism.lipid transfer proteins etc</td>
</tr>
<tr>
<td></td>
<td>0.00361</td>
<td>2.59</td>
<td>misc.glutathione S transferases</td>
</tr>
<tr>
<td></td>
<td>0.04216</td>
<td>2.33</td>
<td>misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein</td>
</tr>
<tr>
<td></td>
<td>0.04216</td>
<td>-0.94</td>
<td>protein.degradation</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>2.87</td>
<td>PS.lightreaction</td>
</tr>
<tr>
<td></td>
<td>0.00213</td>
<td>2.43</td>
<td>secondary metabolism.flavonoids</td>
</tr>
<tr>
<td>L3</td>
<td>0.00000</td>
<td>4.28</td>
<td>PS.lightreaction.photosystem I</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>3.83</td>
<td>PS.lightreaction.photosystem II</td>
</tr>
<tr>
<td>L4</td>
<td>0.03206</td>
<td>-1.56</td>
<td>protein.degradation.ubiquitin.E3</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>5.49</td>
<td>PS.lightreaction.photosystem II.LHC-II</td>
</tr>
<tr>
<td></td>
<td>0.00686</td>
<td>4.91</td>
<td>PS.lightreaction.photosystem I.LHC-I</td>
</tr>
<tr>
<td></td>
<td>0.00899</td>
<td>3.76</td>
<td>PS.lightreaction.photosystem I.PSI polypeptide subunits</td>
</tr>
</tbody>
</table>
Table S3. Enrichment test for TFs among 9 different gene categories.

<table>
<thead>
<tr>
<th>Category</th>
<th>T-VE (2.5%)</th>
<th>T-Not-Extreme</th>
<th>T-SE (2.5%)</th>
<th>Expected</th>
<th>Observed</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-VE (2.5%)</td>
<td>Expected: 13.1</td>
<td>Observed: 7</td>
<td>P-value: 0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-Not-Extreme</td>
<td>Expected: 28.2</td>
<td>Observed: 52</td>
<td>P-value: 9 x 10^{-5}**</td>
<td>Expected: 23.8</td>
<td>Observed: 23</td>
<td>P-value: 1</td>
</tr>
<tr>
<td>G-SE (2.5%)</td>
<td>*NA</td>
<td></td>
<td></td>
<td>Expected: 41.4</td>
<td>Observed: 59</td>
<td>P-value: 0.008*</td>
</tr>
</tbody>
</table>

*There are no genes among in these categories.

Table S4: Accuracy of associations from SNP-BayesB and eRD-GWAS as estimated via 10 fold cross-validation.

<table>
<thead>
<tr>
<th>Trait</th>
<th>SNP-BayesB</th>
<th>eRD-GWAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BraceRoot</td>
<td>0.55</td>
<td>0.41</td>
</tr>
<tr>
<td>dim1</td>
<td>0.57</td>
<td>0.44</td>
</tr>
<tr>
<td>dim2</td>
<td>0.62</td>
<td>0.49</td>
</tr>
<tr>
<td>StalkArea</td>
<td>0.60</td>
<td>0.48</td>
</tr>
<tr>
<td>peri</td>
<td>0.61</td>
<td>0.49</td>
</tr>
<tr>
<td>DTA</td>
<td>0.86</td>
<td>0.76</td>
</tr>
<tr>
<td>SqrtEHT</td>
<td>0.69</td>
<td>0.56</td>
</tr>
<tr>
<td>SAM volume</td>
<td>0.54</td>
<td>0.51</td>
</tr>
<tr>
<td>Height</td>
<td>0.54</td>
<td>0.56</td>
</tr>
<tr>
<td>Radius</td>
<td>0.48</td>
<td>0.46</td>
</tr>
<tr>
<td>area</td>
<td>0.57</td>
<td>0.48</td>
</tr>
<tr>
<td>arc-length</td>
<td>0.59</td>
<td>0.49</td>
</tr>
<tr>
<td>diameter</td>
<td>0.49</td>
<td>0.45</td>
</tr>
<tr>
<td>mean-nodes</td>
<td>0.76</td>
<td>0.58</td>
</tr>
</tbody>
</table>
Table S5. Characterization of 5 genes associated with DTA identified via eRD_GWAS with highest model frequencies. (XLSX 41KB)

<table>
<thead>
<tr>
<th>Top5 DTA eRD genes</th>
<th>Model Frequency</th>
<th>Maize</th>
<th>Rice</th>
<th>Arabidopsis</th>
<th>Annotation</th>
<th>P-value of eichment test</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G173865</td>
<td>0.666</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRMZM2G171650</td>
<td>0.632</td>
<td>MADS69</td>
<td>OsMADS65</td>
<td>K-box region and MADS-box transcription factor</td>
<td>2 (0.0022)</td>
<td>0</td>
</tr>
<tr>
<td>GRMZM2G147716</td>
<td>0.599</td>
<td>MADS67</td>
<td>MADS-box family gene with MIKCc type-box</td>
<td>AGAMOUS-like 8</td>
<td>2 (0.0482)</td>
<td>0</td>
</tr>
<tr>
<td>GRMZM2G430526</td>
<td>0.598</td>
<td></td>
<td></td>
<td></td>
<td>Electron transfer flavoprotein subunit alpha mitochondrial precursor</td>
<td>2 (0.0458)</td>
</tr>
</tbody>
</table>
Table S6. eQTL analyses for 13 traits eRD genes comparison with different GWAS results.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>SNP eRD</th>
<th>SNP BayesB</th>
<th>SNP MLM</th>
<th>Rice Annotation</th>
<th>Maize Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G035405</td>
<td>X</td>
<td></td>
<td></td>
<td>auxin response factor, putative</td>
<td>Auxin response factor 18</td>
</tr>
<tr>
<td>GRMZM2G148810</td>
<td>X</td>
<td></td>
<td></td>
<td>PHD finger protein, putative</td>
<td>Alfin-like-transcription factor 1 (alf1)</td>
</tr>
<tr>
<td>GRMZM2G111696</td>
<td>X X X</td>
<td></td>
<td></td>
<td>Voz1</td>
<td>Braceroot</td>
</tr>
<tr>
<td>GRMZM2G114427</td>
<td>X</td>
<td></td>
<td></td>
<td>cytokinin dehydrogenase precursor</td>
<td>Cytokinin dehydrogenase 8</td>
</tr>
<tr>
<td>GRMZM2G076345</td>
<td>X</td>
<td></td>
<td></td>
<td>OsSAUR33</td>
<td>SAUR33-auxin-responsive</td>
</tr>
<tr>
<td>GRMZM2G540538</td>
<td>X</td>
<td></td>
<td></td>
<td>kinase, pkB family</td>
<td>Adenosine kinase 2</td>
</tr>
<tr>
<td>GRMZM2G129761</td>
<td>X</td>
<td></td>
<td></td>
<td>OsIAA17</td>
<td>AuxIAA-transcription factor 28 (iaa28)</td>
</tr>
<tr>
<td>GRMZM2G143433</td>
<td>X X</td>
<td></td>
<td></td>
<td>MOC1</td>
<td>Gras17</td>
</tr>
<tr>
<td>GRMZM2G306945</td>
<td>X X</td>
<td></td>
<td></td>
<td>succinate dehydrogenase flavoprotein subunit</td>
<td>Dim2</td>
</tr>
<tr>
<td>GRMZM2G475263</td>
<td>X X</td>
<td></td>
<td></td>
<td>auxin response factor 1</td>
<td>Dim2</td>
</tr>
<tr>
<td>GRMZM2G146644</td>
<td>X</td>
<td></td>
<td></td>
<td>BES1/BZR1 homolog protein, putative</td>
<td>Dim2</td>
</tr>
<tr>
<td>GRMZM2G070034</td>
<td>X</td>
<td></td>
<td></td>
<td>OsMADS56</td>
<td>Mads76</td>
</tr>
<tr>
<td>GRMZM2G171650</td>
<td>X X</td>
<td></td>
<td></td>
<td>OsMADS65</td>
<td>MADS69</td>
</tr>
<tr>
<td>GRMZM2G146644</td>
<td>X</td>
<td></td>
<td></td>
<td>cytokinin dehydrogenase precursor</td>
<td>cytokinin oxidase1 (cko1)</td>
</tr>
<tr>
<td>GRMZM2G093186</td>
<td>X</td>
<td></td>
<td></td>
<td>ras-related protein</td>
<td>Ras protein RGP2</td>
</tr>
<tr>
<td>GRMZM2G864065</td>
<td>X</td>
<td></td>
<td></td>
<td>chaperone protein dnaJ, putative</td>
<td>Putative dnaJ chaperone</td>
</tr>
<tr>
<td>GRMZM2G057243</td>
<td>X</td>
<td></td>
<td></td>
<td>carotenoid cleavage dioxygenase</td>
<td>White cap1</td>
</tr>
<tr>
<td>GRMZM2G430052</td>
<td>X</td>
<td></td>
<td></td>
<td>auxin responsive protein, putative</td>
<td>DTA</td>
</tr>
<tr>
<td>GRMZM2G076796</td>
<td>X</td>
<td></td>
<td></td>
<td>OsGrx-S17</td>
<td>DTA</td>
</tr>
<tr>
<td>GRMZM2G147716</td>
<td>X</td>
<td></td>
<td></td>
<td>OsMADS18</td>
<td>Mads67</td>
</tr>
<tr>
<td>GRMZM2G355233</td>
<td>X</td>
<td></td>
<td></td>
<td>OsWAK92</td>
<td>Height</td>
</tr>
<tr>
<td>GRMZM2G161472</td>
<td>X</td>
<td></td>
<td></td>
<td>cytochrome P450, putative, expressed</td>
<td>kaurene oxidase2</td>
</tr>
<tr>
<td>GRMZM2G146192</td>
<td>X</td>
<td></td>
<td></td>
<td>glycosyl hydrolase family 3 protein, putative, expressed</td>
<td>Auxin-induced beta-glucosidase</td>
</tr>
</tbody>
</table>
Table S7. Phenotypes analyzed in this study

<table>
<thead>
<tr>
<th>Phenotypes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dim1</td>
<td>Thickest diameter of stalk</td>
</tr>
<tr>
<td>Dim2</td>
<td>Thinnest diameter of stalk</td>
</tr>
<tr>
<td>Bracerooot</td>
<td>Number of nodes with brace roots</td>
</tr>
<tr>
<td>StalkArea</td>
<td>Cross section area of stalk</td>
</tr>
<tr>
<td>Peri</td>
<td>Circumference of stalk</td>
</tr>
<tr>
<td>Mean-nodes</td>
<td>Number of nodes</td>
</tr>
<tr>
<td>Height</td>
<td>Height of SAM from apex to P1 notch¹</td>
</tr>
<tr>
<td>Radius</td>
<td>Radius of meristem at P1 notch¹</td>
</tr>
<tr>
<td>Area</td>
<td>Surface area under SAM parabola¹</td>
</tr>
<tr>
<td>Volume</td>
<td>Volume of paraboloid¹</td>
</tr>
<tr>
<td>Arc-length</td>
<td>Estimated length from P1 notch to apex¹</td>
</tr>
<tr>
<td>SqrtEHT</td>
<td>Square root of Ear Height²</td>
</tr>
<tr>
<td>DTA</td>
<td>Days to anthesis²</td>
</tr>
</tbody>
</table>

¹Leiboff et al. 2015
²Peiffer et al. 2014
CHAPTER 3. THE IDENTIFICATION OF GENES ASSOCIATED WITH PHENOTYPIC VARIATIONS VIA GLOBAL ANALYSES OF ALTERNATIVE SPLICING EVENTS (AS-GWAS)

Hung-ying Lin$^{1,2}$, Qiang Liu$^{1,2}$, Lakshmi Attigala$^{1,2}$ and Patrick S. Schnable$^{1,2}$

1. Department of Agronomy, Iowa State University, Ames, IA 50011-3650
2. Interdepartmental Genetics and Genomics Graduate Program, Iowa State University, Ames, IA 50011-3650
* Corresponding Author: Patrick S. Schnable
2035 B Roy J Carver Co-Lab, Iowa State University, Ames, IA, 50011-3650 USA

Abstract

Background: Alternative splicing (AS) has widespread effects on gene function and hence phenotypes. However, so far, it has not been possible to efficiently conduct genome-wide scans to identify such associations.

Results: We created a novel approach to GWAS (AS-GWAS) that uses AS events as the explanatory variables for phenotypic response variables. The accuracy of AS-GWAS is supported by its ability to identify biologically relevant associations, cross-validation, transposon knockout mutants, and the analysis of a gene regulatory network. Further, the gene-phenotype associations detected by AS-GWAS are complementary to those detected via conventional GWAS and eRD-GWAS.

Conclusions: This ability to associate genome-wide AS patterns with traits provides direct evidence on how AS contributes on phenotype. This study also illustrates opportunities to use
AS-GWAS in combination with genome-wide statistical analyses to explore the complex regulatory networks that influence phenotypic variation.

**Keywords:** Alternative Splicing, GWAS, Phenotypes, Traits, Gene regulatory network, Splicing QTL

**Background**

Alternative splicing (AS) is a regulatory mechanism observed in eukaryotes and in some bacteria and archaea [1, 2] that increases transcriptome and proteome diversity. It has been estimated that 95%, 61% and 70% of multi-exon genes of humans, *Arabidopsis* and maize undergo alternative splicing, respectively [3, 4]. There are 5 types of AS events: intron retention (IR), exon skipping (ES), mutually exclusive exons (ME), alternative 5’ donor site (Alt5) and alternative 3’ acceptor site (Alt3). There are substantial differences in the patterns of AS among species. For example, the most common type of AS in humans (35%) is ES [5], while in plants the most common form is IR [4].

Proteins such as heterogeneous nuclear ribonucleoproteins (hnRNPs) and serine/arginine-rich (SR) proteins that function in processing pre-mRNAs and thereby generate AS isoforms have been identified via such analyses [6]. Nonsense mediated decay (NMD) is a mechanism that identifies and degrades not only mRNAs that contain premature termination codons but also some AS isoforms [7]. AS and canonical isoforms of a given
gene can also be differentially regulated by microRNAs [8]. Thus, both NMD and microRNAs can contribute to variation in the abundance of AS isoforms Genetic variation in these regulatory processes would therefore be expected to influence ratios of AS and canonical isoforms. And indeed, splicing QTLs (sQTL) [4, 8], identified in a manner similar to the identification of eQTL, regulate the differential accumulation of various AS isoforms.

Just as AS events can be regulated by sQTL, AS events can comprehensively contribute to phenotypic variation [4, 9-11]. In this study we report a new method for conducting GWAS that exploits these relationships to identify genes associated with phenotypic variation. Two general GWAS approaches have been used to identify associations between explanatory variables and phenotypes: mixed linear models (MLM) and Bayesian-based approaches [12]. Typically, MLM solutions estimate effects based on single markers and partition the variance into covariance, to account for population structure [13]. In contrast, Bayesian frameworks apply multiple variable regression with prior distributions and Markov Chain Monte Carlo (MCMC) to infer posterior distributions [12].

The next generation sequencing (NGS) revolution has made it possible to easily generate millions of genetic markers (typically SNPs) that have been used as explanatory variables in GWAS [14]. We (Lin et al. [15]) recently reported a novel Bayesian approach (eRD-GWAS) that uses transcript accumulation as explanatory variables in GWAS. The ability to link variation in transcript accumulation to traits or phenotypes demonstrated that explanatory variables in GWAS need not be restricted to variation in DNA sequence or
structure. Importantly, we found that SNP GWAS and eRD-GWAS identify complementary sets of trait associations.

Network analysis involves the construction of relationships between nodes to assist in identifying hub genes with a goal of better understanding regulatory mechanisms [16]. RNA co-expression networks and Gene Regulatory Networks (GRN) are two methods of network analysis widely used in studying complex biology phenomena [17]. A co-expression network is an undirected graph that connects nodes (genes) by significant similarities in expression [18]. Unlike co-expression networks, a GRN represents directed biological relationships between two nodes/genes [19]. Zhang et al. (2016) [20] used co-expression networks to study powdery mildew disease in wheat, while Walley et al. (2016) [21] used both a co-expression network and several GRNs to enhance our understanding of the regulation of RNA and protein accumulation in maize. We have previously used both types of networks to show that genes whose expression patterns are associated with phenotypic variation via eRD-GWAS, cluster in modules enriched for gene categories relevant to the associated phenotypes [15].

In the current study, we report a novel method (Alternative Splicing-Genome Wide Association Studies; AS-GWAS) that uses AS events as explanatory variables in GWAS. First, we identified AS events genome-wide in a maize diversity panel. Using these AS events in an AS-GWAS we identified large numbers of associations, demonstrating that AS has genome-wide impacts on phenotype in a model plant. Our ability to associate genome-wide patterns of AS with specific phenotypes expand our understanding of the complex
genetic regulation of phenotypes. Finally, our analyses of trait-associated AS events offer a strategy for the exploration of complex biological regulatory pathways.

**Results**

**Identification of Alternative Splicing Events**

Our analyses were based on existing RNA-seq data generated using an Illumina sequencing platform. Inferring an entire isoform based on short sequencing reads can generate biases due to mis-alignment in the face of isoform complexity [22]. Thus, rather than quantifying the accumulation of an entire isoform, we focused on the accumulation of each of potentially multiple AS events in a given gene (e.g., the retention of each of several introns). AS events were identified from existing RNA-seq data generated from shoot apices collected from a maize diversity panel (SAM panel, N=369 inbred genotypes) [23]. 253,634 distinct AS events were identified across all genotypes. Consistent with prior studies in plants [24] the most common forms of AS events observed were Alt5 and IR which, in combination represented more than 60% of all AS events (Table 1).

Particular AS events were detected in varying percentages of the lines that comprised the SAM panel (Additional file 1). Although many AS events exhibited low minor allele frequencies (MAF), 38,894 AS events, affecting 11,996 genes had MAF greater than 7.5%. This MAF cut-off provided the highest accuracy in 10-fold cross validation (Methods) Of these 38,894 AS events, 15,533 were detected in the B73 inbred. 70% (10,572/15,533) of
these could be aligned to full-length PacBio cDNAs from B73 [3], even though the Illumina-based RNA-Seq reads used to discover the AS events and the PacBio cDNA reads used for validation were generated from different tissues (with different expression patterns) and had been sequenced via different technologies [23]. Hence, these alignment results support the overall accuracy of our identification of AS events.

**Principal Component Analysis (PCA) based on AS events**

To test whether overall patterns of AS differed among subpopulations of our diversity panel, PCA was conducted using both SNPs and patterns of AS. For the SNP-based PCA we used a previously reported set of 1.28M SNPs generated from the SAM panel [23]. The AS PCA was based on the presence or absence of each of 38,894 AS events in each of the inbreds of the SAM panel (Methods). The first and second components of both PCAs explain similar amounts of phenotypic variance. The first components of the SNP and AS PCA explain 5.9% and 5.4% of the variance, while the second components explain 3.3% and 2.4% variance, respectively. Additionally, both PCAs were reasonably successful at classifying members of the SAM panel into previously defined subpopulations based on known ancestral history [25](Fig. 1). These results support the view that global patterns of AS are under genetic control.
Alternative Splicing-Genome Wide Association Studies (AS-GWAS)

Given that patterns of AS are under genetic control and because we had previously been able to use gene expression patterns to identify associations between genes and traits/phenotypes (i.e., eRD-GWAS) that are complementary to those associations detected via SNP GWAS, we designed a similar approach (AS-GWAS), based on patterns of AS rather than SNPs or gene expression. Thus, binary patterns of AS were used as the explanatory variable for the trait/phenotype, shoot apical meristem (SAM) volume, which was previously collected for our diversity panel [23]. These analyses were conducted using both a MLM model and a Bayesian framework (Methods).

Using a 0.05 Bonferroni adjusted p-value for the MLM AS-GWAS and an arbitrary model frequency cut-off of 0.05 for the Bayesian-based AS-GWAS, 53 and 124 AS events affecting 131 genes were found to be associated with SAM volume, respectively. Only four AS events were detected as being associated with SAM volume by both the MLM and Bayesian methods (Fig. 2a, b and Table 2). The annotations of some of the genes identified via AS-GWAS are consistent with our understanding of the regulation of SAM volume (Table 2). For example, auxin is known to affect the development of the SAM [26] and several Auxin Response Factor (ARF) proteins are involved in SAM maintenance [27]. Consistently, an IR event in GRMZM2G006042, which encodes the ARF transcription factor 28 is associated with an increase in SAM volume.

The rough sheath2 (GRMZM2G403620; RS2) interacting protein (GRMZM2G079823; ZmRIK), functions in SAM tissue during leaf primordia initiation [28].
RS2- and AS1-encoded proteins form a complex with the ZmRIK protein to repress knox genes. The finding that a loss of function mutation of ZmRIK derepresses knox genes in developing leaves, thereby initiating organogenesis, highlights the role of ZmRIK in SAM determinacy [29]. The presence of an IR event in the ZmRIK gene that results in the inclusion of an additional 1,512bp to the predicted ZmRIK protein-protein interaction domain (Fig. 2c) is associated with increased SAM volume.

GRMZM5G867798 is a maize homolog of the rice Brassinosteroid (BR) Insensitive 1 receptor kinase (BAK1) gene. BR is a hormone involved in many aspects of plant development and reproduction such as growth, cell division, flowering, including SAM organ boundary formation [30]. BAK1 is a BR co-receptor located on the cell membrane [30].

Only a few overlapping signals were found between MLM and BayesC (Fig. 2). One example, is GRMZM5G825707, which encodes maize Aux/IAA transcription factor 6. Auxin is another hormone that functions in SAM development and the Aux/IAA transcription factor can control downstream auxin inducible genes [27]. IAA proteins control the expression of auxin transporter genes (PINs) and thereby regulate SAM development [31]. Another example of overlap is AC211474.3-FG0006, which is a homolog of the rice growth regulator protein gene [32].

Several growth regulator proteins have been reported to affect SAM maintenance by repressing knox genes [32]. For example, overexpression mutants of AtGRF4, AtRGF5, and AtGRF6 and OsRGF10 cause developmental abnormalities as a consequence of repression of
knox in the SAM [32]. Because independent perturbations in multiple genes associated with SAM volume via AS-GWAS have been previously shown to affect SAM volume via other mechanisms we conclude that AS-GWAS does not only associate specific AS events with phenotype, but also provide a novel strategy to detect genes that more generally affect variation in the phenotype of interest.

**Confirmation via reverse-genetics**

The ZmAgo104 (GRMZM2G141818) gene was associated with SAM volume via AS-GWAS. ZmAgo104 has been reported to function in repressing germ cell fate in somatic tissues via a small RNA pathway [33]. To functionally test this association, a Mu transposon knockout mutant of this gene was isolated (Method). The Mu transposon was inserted into the first intron (Figure 3a). We generated a population segregating 1:1 for plants homozygous and heterozygous for the Mu insertion mutant (Methods). Plants homozygous for the Mu insertion mutant allele accumulate lower concentrations of ZmAg104 transcripts than do heterozygous (non-mutant) control siblings (Figure 3d). Additionally, consistent with the results from AS-GWAS, plants homozygous for the Mu insertion mutants have larger SAM volumes than do heterozygous (non-mutant) control siblings (Figure 3b,c; P-value = 0.032). This result demonstrates that the ability of AS-GWAS ont only detects trait-assoicated AS event, but also identifies genes associated with traits of interest.
Accuracy of AS-GWAS as Determined via 10-fold Cross-validation

As discussed above a number of specific gene/trait associations detected via AS-GWAS make biological sense. Another approach to evaluate whether AS affects phenotype on a global scale is to use AS events as the explanatory variables in genomic prediction models. We compared the ability of SNP-based and AS-based genomic prediction models to predict phenotype. Bayes-based genomic prediction models based on SNPs and AS events achieved accuracies of 0.58 and 0.54, respectively (Additional file 2). Accuracies were measured as the mean of correlation coefficient of each method via 10-fold cross validation. Similar results were obtained using five methods of genomic prediction (GBLUP, LASSO, Ridge R, BayesA, and BayesB) (Additional file 2). The finding that the accuracies of SNP-based and AS-based genomic predictions were similar provides strong support for the accuracy of AS-GWAS and a global role of AS in contributing to phenotypic variation in a plant species.

Comparisons of GWAS Methods

Traditional GWAS associates SNPs with variation in phenotype. In contrast, eRD-GWAS associates gene expression patterns with phenotype [15]. While, both GWAS approaches identify biologically relevant gene/trait associations, there is little overlap between the signals each detects [15]. Similarly, none of the 143 significant genes detected by Bayesian-based SNP GWAS or Bayesian-based AS-GWAS as being associated with SAM volume were detected by both methods (Fig. 4). Further, the linear correlation of the
model frequencies of genes as determined the two methods is non-significant ($r^2 = 0.0031; p$-value = 0.47; Additional file 3a). The lack of overlap between associations detected via SNP GWAS and AS-GWAS demonstrates that these methods are complementary.

Because only an expressed gene can experience AS, it is possible that gene expression and AS might be confounded. We therefore determined the overlap between significant associations detected via Bayesian-based eRD-GWAS and Bayesian-based AS-GWAS. Only 3 of 226 genes associated with SAM volume were detected by both GWAS methods (Fig. 4). Further, the linear correlation of the model frequencies of genes as determined via eRD-GWAS and AS-GWAS is non-significant ($r^2 = 0.0096; p$-value = 0.32; Additional file 3b). This lack of overlap demonstrates that eRD-GWAS and AS-GWAS are also complementary.

**sQTL Analyses of Trait-associated AS Events**

Using approaches described above we identified genes in which AS events were associated with SAM volume. To determine whether these trait-associated AS events are functionally associated with phenotypes rather than simply being genetically linked to causative genetic variation as a function of linkage disequilibrium (LD), we identified splicing QTL (sQTL) for the 173 trait-associated AS events in 131 genes detected via MLM-(N=23) and Bayesian-based AS-GWAS (N=104) or both (N=4). An sQTL analysis was conducted using the previously discussed set of 1.28M SNPs. This analysis is similar to an
eQTL analysis except instead of using gene expression as the response variables we used the ratio of AS reads to the overall read depth of a given gene in a given inbred as the response variable (Methods).

In total, 1,623 sQTLs (SNPs) were identified and at least one sQTL was identified for ~80% (139/173) of the trait-associated AS events (Fig. 5a). Because this analysis was conducted using a diversity panel it offers substantially higher mapping resolution of sQTL than our previously reported analysis based on a bi-parental population [4]. Candidate genes identified within 30kb windows centered on the SNP (Methods) associated with each AS event (i.e., sQTL) include those that function in RNA splicing, meristem maintenance, cell division, and development (Table 3). In addition, several GO categories associated with SAM development and RNA processing were enriched among sQTL candidate genes (e.g., GO:0019827 stem cell maintenance; GO:0010451 floral meristem growth; GO:1903507 negative regulation of nucleic acid-templated transcription and GO: 0005686 U2 snRNP; Additional file 4). Significantly, none of the detected sQTL was located within 1 Mb of the gene whose AS it regulated.

As a control, we conducted a similar analysis for all 38,894 AS events that affect 11,996 genes; as was the case for the AS events associated with SAM volume, few of these AS events are regulated exclusively in cis (Fig. 5b). This finding that the AS events associated with SAM volume are themselves mostly regulated in trans which is consistent with Chen et al. (2018) reported [11] but disagrees with the hypothesis that the gene/trait
associations detected via AS-GWAS simply reflect LD with causative genetic variation. Instead, this result provides strong evidence that AS has a causative (though not necessarily direct) role in determining SAM volume and therefore provides additional evidence that AS plays a global role in regulating plant phenotypes.

**One mechanism by which AS affects phenotype**

The analyses described above associated AS in specific genes with a particular phenotype (SAM volume). We hypothesized that these AS events affect SAM volume via changes in gene expression. To test this hypothesis, we used an approach similar to the previously described sQTL analysis. The expression levels of four genes known to affect SAM volume and development (Kn1 GRMZM2G159431, fea2 GRMZM2G104925, LAX2 GRMZM2G129413 and ZmBAK1 GRMZM2G145720,) [27, 34], were used as response variables, and global AS patterns as the independent variables to identify AS events associated with the expression levels of these four genes. *Kn1* and *fea2* contribute to SAM indeterminacy [34]. *LAX2* is an auxin transport which also functions in SAM development via its effects on auxin transport [27]. *ZmBAK1* is a brassinosteroids (BR) co-receptor that affects SAM volume [23].

In total, 8 AS events in six genes were associated with variation in expression in one or more of the four SAM-related genes. Each of these six genes were previously associated with SAM volume via AS-GWAS (Table 4). For example, an ES event in the *ZmCKI-like*
gene (GRMZM2G156035) is associated both with a statistically significant increase in the expression of ZmBAK1 in the SAM panel (p-value =1.03 e-13) and associated with a decrease in SAM volume. ZmCKI-like is a casein kinase I-like protein gene (ZmCKI-like). In rice casein kinase I (CKI) is a putative transcription factor highly induced by BRs and involved in cellular division [35].

So how does AS in the ZmCKI-like gene affect expression of ZmBAK1? It has been established that microRNAs can differentially regulate AS and canonical isoforms of a given gene [8]. According to the psRNATarget database [36] the portion of the ZmCKI-like gene affected by the ES AS event contains a “pab_miR3707” miRNA binding site. The sQTL analysis for the ES AS event in ZmCKI-like identified a single sQTL SNP that is only 4kb away from, the phosphatase 2C isoform gamma gene (ZmPP2C; GRMZM2G360455). In plants PP2C functions in signal transduction and RNA splicing [37]. Based on these results we propose a model in which ZmPP2C triggers the ES AS event in the ZmCKI-like gene which allows the ZmCKI-like gene to escape from miRNA-induced repression, thereby inducing higher ZmBAK1 expression.

**AS-GRN**

Based on the finding (Fig. 4) that genes associated with SAM volume identified via eRD-GWAS and AS-GWAS exhibit little overlap, we hypothesized that gene expression and AS are under independent regulation. To test this hypothesis, we constructed a gene
regulatory network (GRN). Co-expression networks cluster genes based on the correlations between their expression levels. In contrast, GRNs represent directed relationships between upstream regulators and downstream targets.

We used a machine learning-based approach to construct a GRN based on the intron retention ratio (IRR; Methods) for each of the 173 trait-associated AS events and a measure of transcript accumulation (RPKM) for each of the 131 corresponding genes in each of 369 genotypes in the SAM panel. A total of 2,009 previously defined transcription factors [38] were considered as potential regulators of the 173 AS events in the 131 genes. The resulting GRN included 22,970 edges and 1,622 nodes (1,336 regulators, 172 AS events and 114 genes) (Additional file 5). Of the 1,336 regulators, 503 affect only gene expression and 1,119 affect both AS events and gene expression. Only ~2.5% of the regulators of specific trait-associated AS events also contribute to the expression levels of the genes within which these AS events occur. This indicates, that as hypothesized, the regulatory networks for AS events and gene expression are distinct. By comparing sQTLs between canonical gene expression and alternative splicing isoform, Chen et al. (2018) also made some conclusion that gene expression and alternative splicing under different regulatory mechanism [11].

As discussed above, there is substantial evidence for the role of ZmRIK in SAM development [28] and this gene was identified via AS-GWAS (Fig. 2a). The GRNs for the ZmRIK intron retention (ZmRIK-IR) event and the expression of ZmRIK include 23 and 91 regulators, respectively (Additional file 6); only one of these regulators is in common.
Several of the regulators of ZmRIK-IR, such as *Rough sheath 1* (*RS1*; GRMZM2G028041) and *Rolled leaf 1* (*Rld1*; GRMZM2G109987) genes are known to affect SAM volume [39, 40]. The Arabidopsis gene, *ULTRAPETALA1* gene acts as a negative regulator of cell accumulation in inflorescence and floral meristems [41]. In the GRN, the maize homolog of this gene (*UTL1* GRMZM2G082745) was identified as a regulator of ZmRIK-IR. Importantly, none of these genes associated with SAM volume via AS-GWAS and GRN had been detected via SNP GWAS or eRD-GWAS. Hence, the combination of AS-GWAS and GRN has the potential to enhance our understanding of the complex regulation of traits.

The 10% of the 378 regulators that affect AS events with the most connections to AS events are presented in Additional file 7. The regulator with the most connections is the transcription factor (TF) C3H36 (GRMZM2G031827), a homolog of the rice splicing factor U2AF, which binds the intron branch point and 3′ AG boundary and further recruits U2 snRNP to induce splicing of immature mRNAs [42]. The top 37 regulators also include 6 MADS box family TFs, 5 No Apical Meristem proteins (NAC protein family members) and 2 Squamosa promoter binding proteins (SPBs); these enrichments were statistically significant (p-value = 2.3e-08**, p-value= 5.9e-0.7** and p-value= 4.9e-0.4**, respectively). Among the top 37 regulators are three genes known to function directly in SAM or apical inflorescence meristem regulation (*Bif4* GRMZM5G864847, *Lg3* GRMZM2G087741, and *Gn1* GRMZM2G452178) [26, 43].
Discussion

Over the past decade, SNP-based GWAS has been conducted in many species and for many phenotypes [44, 45]. One of the major challenges with interpreting the results of these analyses has been the existence of false-positive SNP-trait associations [46]. Combining eQTL analysis with GWAS can decrease biases and increase the statistical power of GWAS [47, 48]. More recently, Lin et al. (2017) used transcript accumulations directly as explanatory variables for conducting GWAS (eRD-GWAS) and found gene/trait associations that were complementary to those detected via comparable to conventional SNP-based GWAS. Hence, expression data have been used both indirectly and directly to enhance the identification of gene/trait associations. Although these analyses have mostly relied on transcript count data without distinguishing among isoforms, one recent study used gene expression and alternative splicing data to prioritize candidate genes for human disease detected via SNP-based GWAS [49].

In this study we used AS events directly as the explanatory variables to conduct GWAS via a method we have termed AS-GWAS. AS-GWAS identifies sets of gene/trait associations that are complementary to those detected via both SNP-based GWAS and eRD-GWAS.

AS-GWAS utilizes a binary presence/absence scoring of AS events to identify trait associations. We also tested using IRR (or PSI) values as explanatory variables but based on cross-validation binary scoring provided higher accuracy. We used both MLM and BayesC
statistical models for AS-GWAS. In MLM we used a 0.05 Bonferroni adjusted p-value as our statistical cut-off. For BayesC we arbitrarily selected a model frequency of 0.05. The appropriateness of these cut-offs is supported by the fact that they yielded gene/trait associations consistent with the literature and the Mu transposon inserted mutant phenotype. The *ago104-05518* mutants phenotype support *Ago104* functions in SAM development. Few of Ago family proteins have been reported also regulate the SAM tissue growth [50]. All information implies the microRNAs play role in SAM development controlling.

One of the causes of false-positive associations from SNP-based GWAS has been the identification of candidate genes within LD blocks surrounding trait-associated SNPs. Because we demonstrated that most AS events are regulated *in trans*, the gene which contains a trait-associated AS event is likely to be the correct candidate gene, as opposed to a nearby gene within an LD block. Because 78% of eQTL act *in trans* [51] a similar argument can be made for eRD-GWAS. Although the identification of candidate genes underlying sQTL and eQTL detected via conventional SNP-based GWAS may be affected by LD, this does not influence the accuracy of associations of candidate genes with phenotypes detected via AS-GWAS and eRD-GWAS.

We have described one mechanism by which AS may be able to influence phenotype. Specifically, we hypothesize that AS of the ZmCKI-like gene enables it to escape miRNA-induced repression, thereby inducing higher expression of *ZmBAKI* which is known to affect SAM volume. One can envision other potential mechanisms by which AS events influence
phenotype. Protein-protein interaction networks (PPINs) are based on patterns of physical interactions between pairs of proteins [52]. AS events generate novels isoforms which can be translated into distinct proteins. In humans, different AS isoforms exhibit substantial differences in their PPIN profiles [53]. Indeed, the differences in PPINs among different isoforms encoded by the same gene can be as large as the differences between proteins encoded by different genes [54]. Because our analyses of a GRN demonstrate that the accumulations of alternatively spliced and canonical transcripts are regulated via different processes we hypothesize that AS events generate diverse protein isoforms that can differentially participate in PPINs, thereby differentially influencing phenotypes.

Our identification of an AS event within the ZmRIK gene as being associated with SAM volume and our subsequent identification of specific transcription factors linked to ZmRIK in a GRN that are themselves associated with SAM volume, even though they had not been detected via SNP GWAS or eRD-GWAS illustrates opportunities to use AS-GWAS in combination with GRNs and other statistical genome-wide analyses to explore the complex regulatory networks that influence phenotypic variation (Fig. 6).

**Conclusions**

Certain AS isoforms have been proven to function in certain phenotypes. To globally understand the relationship between AS and traits we developed the AS-GWAS pipeline to identify some trait-associated AS events. The associations detected via AS-GWAS are complementary to those detected via conventional GWAS and eRD-GWAS, and the signals
are also supported by a Mu transposon knockout mutant, GRN and sQTL. This study provides a novel strategy to detect trait-associated genes and combining it with other genomic analyses will expand the scope of understanding certain traits.

**Methods**

**Calling AS events**

RNA-Seq reads generated from shoot apices by Leiboff et al. (2015) [23] were aligned to Maize B73 RefGen-v2 genome after trimming low quality bases as described in Lin et al., (2017) [15] using STAR RNA aligner version 2.4 with 2-pass mapping strategy [55]. Alignment results were used to call AS events using Spladder software at the highest confidence level [56]. Other in-house scripts were implemented for generating required files and graphs by R, Python and Perl environments. The scripts can be downloaded from (URL to be provided upon acceptance of the manuscript).

**Mixed Linear Model (MLM) GWAS**

MLM analyses were conducted under GAPIT version 3.35 [57]. The Settlement of MLM Under Progressively Exclusive Relationship (SUPER) GWAS model was selected [58] and default settings were used. SUPER uses a common MLM framework as follows:

\[ y = X\beta + Zu + e \]

Where \( y \) is the phenotype, \( \beta \) is the unknown fixed effect vectors, and \( u \) is a random vector
with a covariance matrix of G. \( G = K \sigma_a^2 \) where K is the kinship matrix. SUPER GWAS model functions in three steps. First, it sorts SNPs by the effect on phenotype and then selects influential bins to construct a kinship matrix. Finally, it tests SNPs for effects under the kinship from the second step.

**Bayesian-based GWAS**

The current study used BayesC [59] as the Bayesian approach to perform the

Bayesian based GWAS as follow:

\[
y = X\beta + Z\upsilon + e
\]

where \( X, Z, \beta, \) and \( \upsilon \) are the same as in the MLM model. The error term \( e \) follows a normal distribution with 0 mean and covariance matrix \( \sigma_e^2 R \) (where \( R \) is a diagonal matrix). The \( \sigma_a^2 \) and \( \sigma_e^2 \) follow the independent inverse Chi-square priors with 4 degrees of freedom. For the test run, the priors of \( \sigma_a^2 \) and \( \sigma_e^2 \) were 50% of the phenotypic variation. The posteriors of the test run were taken as the priors for the true run. 41,000 iterations were used as chain length and the first 1,000 treated as burn-in. For BayesC, the fraction of markers having no effect was set to 0.99. Bayes-based GWAS does not report p-values. Instead, we employ an arbitrary model frequency cut-off (0.05). The model frequency is the proportion of MCMC iterations in which a given AS event was included in the model.
**UniformMu transposon mutagenesis, qRT-PCR and phenotyping**

A reverse genetic transposon knockout approach was used to confirm the association between *ZmAgo104* and SAM volume that had been detected by AS-GWAS. Stocks carrying a *Mu*-insertion allele of *ZmAgo104* (*ago104-05518*) were ordered from the Uniform*Mu* project [60]. Populations segregating for mutants (homozygous for *ago104-05518*) and non-mutants (heterozygous for *ago104-05518*) were generated via controlled pollinations. Genotypes were determined using the *Mu*-specific primer (MuTIR: AGAGAAGCCAACGCCAWCGCCTCYATTTCGTC) and the gene-specific primers flanking the *Mu* insertion, UFMu17-79-F (AGCACTGCAAAATGGATGCG) and UFMu17-79-R (CCGTTCGAGAGACGAGCTAC). 34 PCR cycles’ condition as follow: Denaturing, annealing and extension conditions are 95°C 1min, 61°C 30s and 72 °C 1min respectively. SAM phenotyping was performed as described by Leiboff et al. (2015) [23]. RNAs were extracted from 14 Days seedlings Stem and SAM by RNeasy Plant mini kit (Qiagen, Hilden, Germany) from four mutant and four wildtype samples; each sample contained three seedlings. Ubiquitin (GRMZM2G409726) was used as reference gene to calculate relative expression level.

**Genomic Selection and Cross-validation**

Genomic Best Linear Unbiased Prediction (GBLUP), Bayesian Least absolute shrinkage ans selection operator (LASSO) and Ridge regression genomic selection for 1.28M
SNPs and AS events were conducted under R package BLR [61]. BayesA, BayesB, and BayesC were conducted by GenSel [59]. Ten-fold cross validation was used to evaluate the accuracy of the different methods. 90% of whole dataset divide into training data for training the model and 10% of dataset treated as testing data to determine the accuracy of the trained model in one of ten rounds validation. The samples were divided into training and test dataset by using the R package “cvTools” [62]. Then the prediction accuracy was calculated via mean correlation coefficient between estimated phenotype and observed phenotype within the testing data among 10 times validation.

Construction of AS-GRN

GRNs were generated using GENIE3 software [63] and only the top 5% high confident interactions were retained. For expression GRNs, only genes with average RPKM values > 1 across the whole diversity panel were included and that RPKM value was used to construct the gene expression GRN. For AS-GRN, 173 AS events identified via AS-GWAS were selected and the intron retention ratio (IRR) or percentage splice in (PSI) [4] values that depend on the type of the AS form were used to construct the AS-GRN. Networks were visualized by MANGO software version 1.20 [64].

Principal Component Analysis (PCA) and sQTL analyses

1.28M SNPs and 38,894 AS event patterns were used to conduct PCA and were calculated by GAPIT package [57]. Inbred classification data were obtained from Romay et
al. (2013) [25]. The IRR or PSI (depend on the type of AS) of AS event values were used as the response variable and SNPs were used as the independent variable for R package MatrixEQTL [65] to detect sQTL. Minor allele frequency cutoff was set to 0.05 (860,999/1,279,930 SNPs passed filtering) and the significant cutoff set to 0.05 (Bonferroni adjusted p-value =5.81 e-8). For sQTL candidate gene searching, we selected 30 kb as window to search the candidate genes. The LD decay ($r^2$) in 30kb window roughly closed to 0.1 based on previous study [25].

**AS event In Silico Validation**

PacBio reads were downloaded for six different tissues of Maize B73 [3] and the AS event sequence was used as the query to align PacBio reads with blastn 2.2.28 [66] by following the criteria that the aligned region should cover more than 97% of query with at least 99% sequence identity.

**Comparisons of Different GWAS Strategies**

For a given AS event, we collected SNPs within 30 kb window and extracted the model frequency from BayesC, then the highest model frequency among the collected SNPs were used to compare with the AS event model frequency from AS-GWAS. We also compared the model frequency from eRD-GWAS [15] for canonical gene expression with a given AS event from AS-GWAS.
Acknowledgements

We thank Drs. Rohan Fernando, Dorian Garrick, Yanhai Yin and Heng-Cheng “Alvis” Hu (all of Iowa State University) for helpful discussions.

References


60. McCarty DR, Meeley RB. Transposon resources for forward and reverse genetics in maize. In: Handbook of maize. Springer; 2009: 561-84


## Tables

### Table 1. Numbers of different types of alternative splicing events.

<table>
<thead>
<tr>
<th></th>
<th>Intron retention (IR)</th>
<th>Alternative 3' (Alt3)</th>
<th>Alternative 5' (Alt5)</th>
<th>Exon skipping (ES)</th>
<th>Mutually exclusive (ME)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number Events (%)</strong></td>
<td>75,396 (29.7%)</td>
<td>81,834 (32.3%)</td>
<td>48,304 (19.0%)</td>
<td>42,500 (16.8%)</td>
<td>5,600 (2.2%)</td>
<td>253,634</td>
</tr>
</tbody>
</table>

### Table 2. Genes associated with SAM volume via MLM-based and Bayes-based AS-GWAS and their annotations from maize (MaizeGDB; http://www.maizegdb.org/) and their associated *Arabidopsis* and rice homologs selected from among a total of 173 associations, representing 131 genes.

<table>
<thead>
<tr>
<th>GWAS Type</th>
<th>Gene ID</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLM IR</td>
<td>GRMZM2G006042</td>
<td>ARF-transcription factor 2</td>
</tr>
<tr>
<td></td>
<td>GRMZM2G079823</td>
<td>Maize RIK (RIK) RS2-interacting KH protein</td>
</tr>
<tr>
<td>Alt3</td>
<td>GRMZM2G053117</td>
<td>Phospholipid sterol acyl transferase 1</td>
</tr>
<tr>
<td></td>
<td>GRMZM5G825707</td>
<td>Iaa6</td>
</tr>
<tr>
<td>Alt5</td>
<td>AC211474.3_FG006</td>
<td>O-fucosyltransferase family protein</td>
</tr>
<tr>
<td>IR; Alt5</td>
<td>GRMZM2G311165</td>
<td>Thiamine pyrophosphate dependent pyruvate decarboxylase family</td>
</tr>
<tr>
<td>BayseC ES</td>
<td>GRMZM5G867798</td>
<td>Leucine-rich repeat transmembrane protein kinase</td>
</tr>
<tr>
<td>Alt3</td>
<td>GRMZM2G127386</td>
<td>SPIRAL1-like2</td>
</tr>
<tr>
<td>Alt5</td>
<td>GRMZM2G141818</td>
<td>Argonaute family protein</td>
</tr>
<tr>
<td></td>
<td>GRMZM2G079850</td>
<td>Ribosomal L18p/L5e family protein</td>
</tr>
</tbody>
</table>
Table 3. Selected candidate genes from a set of 1,401 sQTL affecting 173 AS events in 131 genes associated with SAM volume via AS-GWAS

<table>
<thead>
<tr>
<th>Gene</th>
<th>chr</th>
<th>start</th>
<th>Annotation</th>
<th>Maize</th>
<th>Arabidopsis</th>
<th>Rice</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G152172</td>
<td>1</td>
<td>8,458,456</td>
<td>BZ9</td>
<td>BES1/BZR1 homolog 1</td>
<td>BES1/BZR1 homolog protein expressed protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G079823</td>
<td>1</td>
<td>26,197,451</td>
<td>RIK</td>
<td>RS2-interacting KH protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRMZM2G017087</td>
<td>1</td>
<td>271,340,805</td>
<td>Kn1</td>
<td>KNOTTED-like from Arabidopsis thaliana response regulator 6</td>
<td>Homeobox domain containing protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G040736</td>
<td>2</td>
<td>2,081,848</td>
<td>Crr1</td>
<td>OsRR6 type-A response regulator</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRMZM2G080295</td>
<td>2</td>
<td>195,649,882</td>
<td>Srs2</td>
<td>Lateral root primordium (LRP) protein-related</td>
<td>LRP1</td>
<td></td>
</tr>
<tr>
<td>AC194970.5_FG002</td>
<td>2</td>
<td>207,226,557</td>
<td>BZR7</td>
<td>Brassinosteroid signalling positive regulator (BZR1) family protein</td>
<td>BES1/BZR1 homolog protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G025014</td>
<td>2</td>
<td>232,641,434</td>
<td>C3H4</td>
<td>Zinc finger C-x8-C-x5-C-x3-H type family protein</td>
<td>splicing factor U2AF</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G068217</td>
<td>3</td>
<td>2,874,829</td>
<td>Ein2</td>
<td>NRAMP metal ion transporter family protein</td>
<td>ethylene-insensitive protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G088741</td>
<td>3</td>
<td>53,881,358</td>
<td>Lg3</td>
<td>KNOTTED1-like homeobox gene 6</td>
<td>Homeobox domain containing protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G162119</td>
<td>3</td>
<td>130,373,684</td>
<td>C3H4</td>
<td>U2 snRNP auxiliary factor small subunit, putative</td>
<td>splicing factor U2AF</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G116204</td>
<td>3</td>
<td>133,888,888</td>
<td>Abp1</td>
<td>endoplasmic reticulum auxin binding protein 1</td>
<td>auxin-binding protein 4 precursor</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G015592</td>
<td>3</td>
<td>229,657,774</td>
<td>-</td>
<td>RED family protein</td>
<td>splicing factor, putative, expressed</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G154169</td>
<td>4</td>
<td>5,375,612</td>
<td>Gf2</td>
<td>GRF1-interacting factor 3</td>
<td>GRF-interacting factor 2</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G360455</td>
<td>5</td>
<td>91,798,187</td>
<td>-</td>
<td>Protein phosphatase 2C family protein</td>
<td>protein phosphatase 2C</td>
<td></td>
</tr>
<tr>
<td>AC200124.3_FG005</td>
<td>6</td>
<td>62,747,476</td>
<td>Gras66</td>
<td>GRAS family transcription factor</td>
<td>DELLA protein SLR1</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G044194</td>
<td>7</td>
<td>3,022,266</td>
<td>Psh1</td>
<td>phytosulfokine 4 precursor</td>
<td>phytosulfokines precursor</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G038066</td>
<td>7</td>
<td>164,998,583</td>
<td>Alfin3</td>
<td>alfin-like 5</td>
<td>PHD finger protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G053503</td>
<td>8</td>
<td>35,561,736</td>
<td>Erf1</td>
<td>-</td>
<td>AP2 domain containing protein</td>
<td></td>
</tr>
<tr>
<td>AC232238.2_FG004</td>
<td>8</td>
<td>166,791,835</td>
<td>Irs1</td>
<td>bZIP transcription factor family protein</td>
<td>transcription factor</td>
<td></td>
</tr>
<tr>
<td>AC188001.3_FG009</td>
<td>8</td>
<td>169,908,908</td>
<td>-</td>
<td>splicing factor Prp18 family protein</td>
<td>pre-m-RNA-splicing factor</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G550865</td>
<td>9</td>
<td>141,986,753</td>
<td>-</td>
<td>TOPLESS-related 1</td>
<td>lissencephaly type-1-like homology motif</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G180150</td>
<td>10</td>
<td>2,030,427</td>
<td>Brk3</td>
<td>transcription activators</td>
<td>NAP1</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G443509</td>
<td>10</td>
<td>113,161,531</td>
<td>-</td>
<td>Protein phosphatase 2C family protein</td>
<td>protein phosphatase 2C</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G392101</td>
<td>10</td>
<td>147,844,785</td>
<td>Crr2</td>
<td>response regulator 6</td>
<td>OsRR6 type-A response regulator</td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Trait-associated AS events affecting the expression of four target genes involved in SAM development and the maize genes in which they occur and associated *Arabidopsis* and rice homologs.

<table>
<thead>
<tr>
<th>Affected</th>
<th>AS Type</th>
<th>GeneID</th>
<th>Effect On SAM</th>
<th>Effect On Maize</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZmBAK1</td>
<td>IR</td>
<td>GRMZM5G898597</td>
<td>Decrease</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td>GRMZM2G156035</td>
<td>Decrease</td>
<td>Casein kinase I isoform delta-like</td>
<td>casein kinase I-like 3</td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td>GRMZM2G473533</td>
<td>Increase</td>
<td>Delta(14)-sterol reductase</td>
<td>Ergosterol biosynthesis ERG4/ERG24 family</td>
</tr>
<tr>
<td>IR</td>
<td></td>
<td>GRMZM2G045171</td>
<td>Decrease</td>
<td>sucrose synthase</td>
<td>sucrose synthase 6</td>
</tr>
<tr>
<td>Kn1</td>
<td>Alt3</td>
<td>GRMZM2G056849</td>
<td>Increase</td>
<td>-</td>
<td>RNI-like superfamily protein</td>
</tr>
<tr>
<td>IR</td>
<td></td>
<td>GRMZM5G898597</td>
<td>Decrease</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LAX2</td>
<td>Alt3</td>
<td>GRMZM2G056849</td>
<td>Increase</td>
<td>-</td>
<td>RNI-like superfamily protein</td>
</tr>
<tr>
<td>fea2</td>
<td>ES</td>
<td>GRMZM2G473533</td>
<td>Decrease</td>
<td>Delta(14)-sterol reductase</td>
<td>Ergosterol biosynthesis ERG4/ERG24 family</td>
</tr>
</tbody>
</table>
Fig.1 a) Two-dimensional PCA analysis of SAM panel (N=369 inbreds) constructed using 1.28M high-quality SNPs with a minimum MAF of 5%. b) Two-dimensional PCA constructed using 38,894 AS events with a minimum MAF of 7.5%. Colors indicate different subgroups of inbreds from Romay et al. 2013. Gray dots indicate samples with that lack subgroup information.
Fig. 2  

a) Associations detected via MLM-based and Bayes-based AS-GWAS. Upper panel presents significant associations for MLM-based AS-GWAS. Arrows indicate selected annotated associations. Lower panel presents associations for Bayes-based AS-GWAS having model frequencies greater than 0.05. Vertical dashed lines indicate overlapping associations from the two AS-GWAS strategies. 

b) Venn diagram of associations detected via MLM-based AS-GWAS and Bayes-based AS-GWAS. 

c) Intron retention at the ZmRIK gene. 

I: Canonical form of ZmRIK mRNA and predicted protein functional domains; 

II: The ZmRIK_IR mRNA and predicted protein domains.
Fig.3 a) Structure of the ZmAgo104 gene showing the ALT5’ alternative splicing event associated with SAM volume and the position of the Mu transposon inserted into the first intron in the ago104-05518 allele. b) Micrographs of SAMs from heterozygous non-mutant (Sample14 mu+/+) and homozygous mutant (Sample31 mu/mu) siblings. c) SAM volumes of heterozygous non-mutant and homozygous mutant siblings. d) ZmAgo104 transcript accumulations in heterozygous non-mutant and homozygous mutant siblings.
Fig. 4 Number of trait-associate genes from each of three AS-GWAS strategies: BayesC-based SNP-GWAS, BayesC-based eRD-GWAS and BayesC-based AS-GWAS.

Fig. 5 a) Summary of sQTL analyses for 139 trait-associated AS events. Red arrows show selected candidate genes. b) Venn diagram of numbers of AS events controlled by cis-sQTL, trans-sQTL or both.
Fig. 6 Summary of analyses used between SNPs, expression levels and AS events to phenotype. a) SNP variants. b) gene expression levels. c) alternative splicing. d) phenotype (SAM-volume). SNPs can be explanatory variables for eQTL, sQTL and phenotypes. Gene expression levels and AS events can be explanatory variables for eQTL, sQTL, GRNs (Gene Regulatory Networks) and phenotypes.
Supplemental Figures

Figure S1. Minor allele frequencies (MAF) of 253,634 alternative splicing events. Red dashed line indicates the 7.5% MAF cutoff for AS-GWAS.

Figure S2. Ten-fold cross validation accuracies of different genomic selection methods based on high density SNPs and alternative splicing events. Blue and red bars indicate the accuracies from high density SNPs and AS events using different methods of genomic selection.
Figure S3. Comparison of AS-GWAS, eRD-GWAS and conventional SNP-based GWAS, all conducted using BayesC. Red dashed lines indicate cut-offs. Blue lines indicate the regression line; a) Comparison of model frequencies between AS-GWAS and SNP-based GWAS; b) Comparison of model frequencies between eRD-GWAS and AS-GWAS.
Figure S4. Gene Regulatory Networks for all trait-associated AS events. Black dots are regulators. Red and blue dots are genes and trait-associated AS events, respectively.
## Supplemental Tables

Table S1. Gene Ontology enrichment tests for candidate genes of 1,409 sQTL.

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Category</th>
<th>Over represented p-value</th>
<th>q-value</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>GO:0070727</td>
<td>0.0034</td>
<td>0.0393</td>
<td>cellular macromolecule localization</td>
</tr>
<tr>
<td>GO:1903507</td>
<td>0.0082</td>
<td>0.0481</td>
<td></td>
<td>negative regulation of nucleic acid-templated transcription</td>
</tr>
<tr>
<td>GO:0009738</td>
<td>0.0123</td>
<td>0.0481</td>
<td></td>
<td>abscisic acid-activated signaling pathway</td>
</tr>
<tr>
<td>GO:0009867</td>
<td>0.0160</td>
<td>0.0481</td>
<td></td>
<td>jasmonic acid mediated signaling pathway</td>
</tr>
<tr>
<td>GO:0035264</td>
<td>0.0226</td>
<td>0.0481</td>
<td></td>
<td>multicellular organism growth</td>
</tr>
<tr>
<td>GO:0051493</td>
<td>0.0229</td>
<td>0.0481</td>
<td></td>
<td>regulation of cytoskeleton organization</td>
</tr>
<tr>
<td>GO:0000184</td>
<td>0.0238</td>
<td>0.0481</td>
<td></td>
<td>nuclear-transcribed mRNA catabolic process, nonsense-mediated</td>
</tr>
<tr>
<td>GO:0048831</td>
<td>0.0241</td>
<td>0.0481</td>
<td></td>
<td>regulation of shoot system development</td>
</tr>
<tr>
<td>GO:0009944</td>
<td>0.0267</td>
<td>0.0481</td>
<td></td>
<td>polarity specification of adaxial/abaxial axis</td>
</tr>
<tr>
<td>GO:0040008</td>
<td>0.0352</td>
<td>0.0481</td>
<td></td>
<td>regulation of growth</td>
</tr>
<tr>
<td>GO:0019827</td>
<td>0.0361</td>
<td>0.0481</td>
<td></td>
<td>stem cell maintenance</td>
</tr>
<tr>
<td>GO:0010451</td>
<td>0.0368</td>
<td>0.0481</td>
<td></td>
<td>floral meristem growth</td>
</tr>
<tr>
<td>GO:0048437</td>
<td>0.0430</td>
<td>0.0481</td>
<td></td>
<td>floral organ development</td>
</tr>
<tr>
<td>GO:0031054</td>
<td>0.0438</td>
<td>0.0482</td>
<td></td>
<td>pre-miRNA processing</td>
</tr>
<tr>
<td>GO:0000463</td>
<td>0.0480</td>
<td>0.0492</td>
<td></td>
<td>maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)</td>
</tr>
<tr>
<td>CC</td>
<td>GO:0005657</td>
<td>0.0052</td>
<td>0.0449</td>
<td>replication fork</td>
</tr>
<tr>
<td>GO:0044732</td>
<td>0.0059</td>
<td>0.0481</td>
<td></td>
<td>mitotic spindle pole body</td>
</tr>
<tr>
<td>GO:0015629</td>
<td>0.0158</td>
<td>0.0481</td>
<td></td>
<td>actin cytoskeleton</td>
</tr>
<tr>
<td>GO:0005686</td>
<td>0.0166</td>
<td>0.0481</td>
<td></td>
<td>U2 snRNP</td>
</tr>
<tr>
<td>GO:0071006</td>
<td>0.0353</td>
<td>0.0481</td>
<td></td>
<td>U2-type catalytic step 1 spliceosome</td>
</tr>
<tr>
<td>MF</td>
<td>GO:0005200</td>
<td>0.0412</td>
<td>0.0481</td>
<td>structural constituent of cytoskeleton</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>ZmRIK_IR Regulator</th>
<th>Annotation</th>
<th>Maize</th>
<th>Arabidopsis</th>
<th>Rice</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC187891.3_FG006</td>
<td>ovate family protein 6</td>
<td>Ofp39</td>
<td>DUF623 domain containing protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G082745</td>
<td>developmental regulator, ULTRAPETALA</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G117193</td>
<td>myb-related protein 1</td>
<td>Glk6</td>
<td>myb-like DNA-binding domain containing protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G070673</td>
<td>IQ-domain 19</td>
<td>-</td>
<td>IQ calmodulin-binding motif family protein, putative, expressed</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G034840</td>
<td>Transcriptional factor B3 family protein / auxin-responsive factor AUX/IAA-related</td>
<td>Arft4</td>
<td>auxin response factor, putative, expressed</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G132794</td>
<td>GRAS family transcription factor</td>
<td>Gras58</td>
<td>SHR, putative, expressed</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G419239</td>
<td>myb domain protein 7</td>
<td>Myb42</td>
<td>MYB family transcription factor</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G129817</td>
<td>homolog of yeast ADA2 2B transcriptional adaptor</td>
<td>Mybr106</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G109987</td>
<td>Homeobox-leucine zipper family protein / lipid-binding START domain-containing protein</td>
<td>Rolled leaf1</td>
<td>START domain containing protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G030744</td>
<td>basic helix-loop-helix (bHLH) DNA-binding superfamily protein</td>
<td>Bhih168</td>
<td>helix-loop-helix DNA-binding domain containing protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G175543</td>
<td>ethylene responsive element binding factor 4</td>
<td>Ereb19</td>
<td>AP2 domain containing protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G422205</td>
<td>floral homeotic protein (HUA1)</td>
<td>C3h22</td>
<td>zinc finger C-x8-C-x5-C-x3-H type family protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G028041</td>
<td>KNOTTED-like from Arabidopsis thaliana</td>
<td>Rough sheath1</td>
<td>Homeobox domain containing protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM5G816457</td>
<td>WRKY DNA-binding protein 2</td>
<td>Wrky72</td>
<td>WRKY35</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G024530</td>
<td>LJRHL1-like 3</td>
<td>Prf1</td>
<td>PTF1</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G306357*</td>
<td>CCCH-type zinc finger family protein</td>
<td>C3h2</td>
<td>zinc finger C-x8-C-x5-C-x3-H type family protein, expressed</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G016150</td>
<td>bZIP transcription factor family protein</td>
<td>ohp1</td>
<td>bZIP transcription factor domain containing protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G174917</td>
<td>DREB and EAR motif protein 3</td>
<td>Ereb47</td>
<td>AP2 domain containing protein, expressed</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G065896</td>
<td>ZIM-LIKE 2</td>
<td>Zim2</td>
<td>GATA transcription factor 25, putative, expressed</td>
<td></td>
</tr>
</tbody>
</table>
Table S3. Top 10% major regulators in AS event Gene Regulatory Networks (AS-GRN) and their associated Arabidopsis and rice homologs.

<table>
<thead>
<tr>
<th>Gene_ID</th>
<th>No. conn</th>
<th>Maize</th>
<th>Arabidopsis</th>
<th>Rice</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G031827</td>
<td>16</td>
<td>C3H36</td>
<td>Zinc finger C-x8-C-x5-C-x3-H type family protein</td>
<td>splicing factor U2AF</td>
</tr>
<tr>
<td>GRMZM2G422205</td>
<td>9</td>
<td>C3H22</td>
<td>floral homeotic (HUA1)</td>
<td>zinc finger C-x8-C-x5-C-x3-H type family</td>
</tr>
<tr>
<td>GRMZM2G070034</td>
<td>8</td>
<td>MADS76</td>
<td>AGAMOUS-like 20</td>
<td>OsMADS56 - MADS-box family gene with MIKCc type-box</td>
</tr>
<tr>
<td>GRMZM2G018254</td>
<td>8</td>
<td>GRAS48</td>
<td>GRAS family transcription factor</td>
<td>GRAS family transcription factor domain containing protein</td>
</tr>
<tr>
<td>GRMZM5G864847</td>
<td>7</td>
<td>Bif4</td>
<td>phytochrome-associated protein 1</td>
<td>OsIAA16 - Auxin-responsive Aux/IAA gene family member</td>
</tr>
<tr>
<td>GRMZM2G430522</td>
<td>7</td>
<td>NACTF114</td>
<td>NAC (No Apical Meristem) domain transcriptional regulator superfamily protein</td>
<td>CUP-SHAPED COTYLEDON3</td>
</tr>
<tr>
<td>GRMZM2G177229</td>
<td>7</td>
<td>C3H40</td>
<td>Zinc finger C-x8-C-x5-C-x3-H type family protein</td>
<td>splicing factor U2AF</td>
</tr>
<tr>
<td>GRMZM2G171650</td>
<td>7</td>
<td>MADS69</td>
<td>K-box region and MADS-box transcription factor family protein</td>
<td>OsMADS65 - MADS-box family gene with MIKC* type-box</td>
</tr>
<tr>
<td>GRMZM2G163813</td>
<td>7</td>
<td>SBP19</td>
<td>squamosa promoter binding protein-like 2</td>
<td>OsSPL11 - SBP-box gene family member</td>
</tr>
<tr>
<td>GRMZM2G108147</td>
<td>7</td>
<td>DBPTF1</td>
<td>DNA-binding protein phosphatase 1</td>
<td>protein phosphatase 2C, putative</td>
</tr>
<tr>
<td>GRMZM2G085678</td>
<td>7</td>
<td>EREB46</td>
<td>Integrate-type DNA-binding superfamily protein</td>
<td>AP2 domain containing protein</td>
</tr>
<tr>
<td>GRMZM2G069047</td>
<td>7</td>
<td>NACTF115</td>
<td>NAC (No Apical Meristem) domain transcriptional regulator superfamily protein</td>
<td>no apical meristem protein</td>
</tr>
<tr>
<td>GRMZM2G016150</td>
<td>7</td>
<td>Ohp1</td>
<td>bZIP transcription factor family protein</td>
<td>bZIP transcription factor domain containing protein</td>
</tr>
<tr>
<td>GRMZM5G832409</td>
<td>6</td>
<td>Knox5</td>
<td>KNOTTED1-like homeobox gene 6</td>
<td>Homebox domain containing protein</td>
</tr>
<tr>
<td>GRMZM2G431309</td>
<td>6</td>
<td>GRAS75</td>
<td>scarecrow-like 5</td>
<td>gibberellin response modulator protein</td>
</tr>
<tr>
<td>GRMZM2G171395</td>
<td>6</td>
<td>NACTF86</td>
<td>NAC (No Apical Meristem) domain transcriptional regulator superfamily protein</td>
<td>no apical meristem protein</td>
</tr>
<tr>
<td>GRMZM2G132794</td>
<td>6</td>
<td>GRAS58</td>
<td>GRAS family transcription factor</td>
<td>SHR</td>
</tr>
<tr>
<td>GRMZM2G126018</td>
<td>6</td>
<td>SBP23</td>
<td>squamosa promoter binding protein-like 9</td>
<td>OsSPL17 - SBP-box gene family member</td>
</tr>
</tbody>
</table>
CHAPTER 4. HP-GWAS: ASSOCIATION OF GENIC HAPLOTYPES WITH PHENOTYPES

Hung-Ying Lin\textsuperscript{1,2}, Qiang Liu\textsuperscript{1,2}, Xiao Li\textsuperscript{2,3}, Lakshmi Attigala\textsuperscript{1}, & Patrick S. Schnable\textsuperscript{1,2,4*}

1. Department of Agronomy, Iowa State University, Ames, IA 50011-3650
2. Interdepartmental Genetics and Genomics Graduate Program, Iowa State University, Ames, IA 50011-3650
3. Current address: The Broad Institute of MIT and Harvard, 75 Ames Street, Cambridge, MA 02142-1403
4. Department of Genetics, Developmental and Cellular Biology, Iowa State University, Ames, IA 50011-3650
* Corresponding author

Abstract

Gene-based GWAS has the potential to provide higher statistical power and lower bias as compared to SNP-based GWAS. In this study, we report a GWAS pipeline based on genic haplotypes that is. First, we identified genic haplotypes and demonstrated that haplotypes can be associated with tissue-specific effects on gene expression. Next, we developed a haplotype-based GWAS pipeline (Hp-GWAS) that uses genic haplotypes as the explanatory variable to associated genes with phenotypes. Both reverse-genetic analyses and cross-validation support the accuracy of Hp-GWAS. Finally, Hp-GWAS was used to identify associations with adaptation. Hp-GWAS is complementary to both conventional SNP-based GWAS and eRD-GWAS
Introduction

Identifying genetic variants that contribute to particular phenotypes via genome-wide association studies (GWAS) plays a crucial role in biological research [1]. Two main statistical frameworks are used in traditional GWAS: mixed linear models (MLM) and Bayesian-based approaches [2]. Population structure can reduce the accuracy of GWAS. The MLM approach controls for population structure using covariances [3]. This, however results decreased statistical power [4]. In contrast, Bayesian-based GWAS uses multiple variable regression to infer the real genetic effect of each marker and automatically controls for population structure [5]. BayesB and BayesC are two commonly used methods for conducting Bayesian-based GWAS. BayesC assumes all markers follow the same genetic variance, whereas BayesB does not [6].

Rather than associating genomic variants with phenotype. eRD-GWAS uses expression levels as the explanatory variable for GWAS [2]. Gene expression can be regulated by expression QTL (eQTL) that act in cis and/or trans [7]. Cis-acting eQTL are located within or very close to the gene they regulate. In contrast, eQTL that act in trans such as transcription factors (TF) typically regulate the expression of other genes that are not closely linked [8].

In conventional GWAS, a stringent cutoff is often used to control the rate of false positive rates, however, this necessarily concurrently increases the rate of false negatives [9]. Furthermore, different LD patterns across populations make GWAS signals difficult to
reproduce across studies [10]. An alternative approach, gene-based GWAS combines the
effects of single genomic variants reduces the number of explanatory variables and therefore
has the potential to improve accuracy and reproducibility by avoiding the need for overly
conservative cutoffs and limitations associated with LD structure [11]. Some methods infer
final genic p-values based on conventional single SNP GWAS approach by subsetting
represented SNPs [10-13]. However, there is still no standard agreement regarding which
approach to use. Users typically explore different approaches and then select an approach for
their dataset [14].

One approach for gene-based GWAS is based on haplotypes. After identifying
haplotypes they are used as explanatory variables for GWAS. There are two major
approaches to identifying haplotypes: fixed window length and hidden Markov model (HMM)
based [10]. Having an arbitrarily defined window length and different LD structure across the
genome [15] are two of the major drawbacks of the fixed window length approach, while
HMM methods suffer uncertainty biases among the models [13]. Yano et al. (2016) deeply
sequenced 176 rice inbreds to define haplotypes for every gene, thereby avoiding both of
these limitations [16]. In the current study, we used cost-efficient RNA-seq data to infer
genic haplotypes. Next, we developed a novel haplotype-based method for conducting
GWAS (Hp-GWAS). Hp-GWAS is complementary to both SNP-based GWAS and eRD-
GWAS. The accuracy of Hp-GWAS is supported by both reverse genetic analyses of trait-
associated genes and 10-fold cross validation studies.
Results

Tissue-specific haplotype effects on gene expression

To investigate the impacts of haplotype on phenotype, we first studied the contribution of genic haplotypes to gene expression. An algorithm was designed to classify genic haplotypes using exonic SNPs (Methods) derived from RNA-Seq obtained from five tissues from 27 inbred lines (the NAM Founders). Among the 38,553 filtered gene set (FGS) genes of maize having RNA-Seq read support (Methods), at least two haplotypes were detected for 25,459 (66%) of the FGS genes. The haplotype missing rate of a gene was calculated as the ratio between the number of inbreds whose alleles could not be classified into a haplotype and the total number of inbreds analyzed (Fig.1a). The 17,684 genes with a 0% haplotype missing rate were used to estimate the number of haplotypes. Haplotype diversity of a gene was defined as the ratio between the number of haplotypes and the number of inbreds which had been successfully haplotyped (Fig.1b). 9.5% (1,682/17,684) of genes have haplotype diversity > 50% reflecting substantial allelic diversity. The high diversity genes are enriched in leucine-rich repeat signaling receptor kinases and auxin response factor family genes (Tab.S1).

We then asked if genic haplotypes are associated with tissue-specific differences in gene expression. Negative Binomial GLMs were used to evaluate haplotype, tissue, and haplotype-by-tissue effects on gene expression levels. A 5% false discovery rate (FDR) was used to define significant main effects on expression. 21,839 genes with multiple haplotypes,
sufficient average read counts and estimated convergent parameter (Methods) were included in the analysis. Of these genes, 62.4% (13,633) exhibited differential expression across haplotypes, and 90.3% (19,718) exhibited differential expression across tissues. Moreover, 30.6% (6,684) of genes exhibited significant haplotype-by-tissue interactions, indicating tissue-specific effects of haplotype on gene expression (Fig.1c).

**Haplotype genome-wide association study (Hp-GWAS)**

Because variation in haplotype is associated with variation in gene expression and in tissue-specific differences in gene expression, we hypothesized that different haplotypes will have differential influences on phenotypes. To test this hypothesis we used the maize SAM diversity panel (N=369) for which both RNA-seq data and phenotypic data were available [2]. 490,884 exonic SNPs were extracted from a set of 1.28M genome-wide SNPs. These were used to generate haplotypes for this panel using the same procedure as described for the NAM Founders. 27,962 genes had at least two haplotypes.

Typical GWAS uses individual SNPs as the explanatory variable of phenotype. In contrast, the Hp-GWAS Bayesian-based pipeline developed in the current study uses haplotypes as the explanatory variables (Methods). To compare these methods, we conducted three types of GWAS on the same diversity panel: conventional SNP-based GWAS, eRD-GWAS which uses expression levels as the explanatory variable [2], and Hp-GWAS. Each analysis employed a BayesC model [6] with an arbitrarily assigned model frequency cutoff.
of 0.02. The GWAS results for the traits days to anthesis (DTA) and plant height (PHT) and are shown in Fig. 2 and Table 1.

**Confirmation of Gene/Trait Associations Detected via Hp-GWAS signals using Reverse Genetics**

To test the gene/trait associations detected via Hp-GWAS we isolated Mu transposon insertion mutants in genes associated with DTA or PHT (Methods). Hp-GWAS detected associations between the trait DTA and the ZmCCT9 gene (GRMZM2G004483), which contributes to flowering time adaptation [17]. The ZmCCT9 gene encodes a CCT domain protein and is a homolog of the rice Ghd7 gene, which regulates heading date [18]. A mutant allele that contains a Mu insertion in the into the second exon of the ZmCCT9 gene flowers two days earlier than wildtype (p-value = 0.037*; Fig. 3a), a result consistent with the phenotype of a ZmCCT9 CRISPR knock-out mutant [17].

Hp-GWAS also detected an association between a maize homolog of the rice OsBAK1 gene (GRMZM2G122717) and PHT (Fig. 2). BAK1 is a leucine-rich repeat receptor-like kinase that regulates the brassinosteroid (BR) receptor, BRI1 [19]. BRs are involved in many aspects of plant development and reproduction such as growth, cell division and flowering [19]. Transposon insertion mutants of GRMZM2G122717 are taller than heterozygous and wildtype plants (p-value = 0.03*; Fig. 3b). The finding that mutants in two trait-associated genes exhibit the expected phenotypes support the accuracy of Hp-GWAS.
Comparisons Among Explanatory Variables via 10-fold Cross-Validation

To compare the utilities of using individual SNPs, expression levels and haplotypes for phenotypic predictions, we conducted 10-fold cross-validation for each of these explanatory variables for each of the two traits (i.e., DTA and PHT; Fig. 4). Haplotypes exhibited higher correlation coefficients than did exonic SNPs for both traits. Even as compared to genome-wide SNPs, haplotypes still showed greater correlation coefficients than the SNP-BayesC with DTA trait (but not for PHT). Although the prediction correlation coefficients are not significantly different from eRD-GWAS and SNP-BayesC for the PHT trait, the median correlation coefficient of Hp-GWAS was still higher than the other two methods. In summary, the accuracies of predictions based on haplotypes were higher than or equal to those based on individual SNPs.

Comparisons of SNP-based GWAS, eRD-GWAS and Hp-GWAS

Comparisons among the gene/trait associations identified via conventional SNP BayesC, eRD-GWAS, and Hp-GWAS for DTA and PHT are shown in Fig. 5 A, C. Few significant signals were observed in both types of GWAS (Fig. 5 A, C, Supplementary Fig1 a, b). There ere moderate positive correlations between the model frequencies obtained from conventional SNP-based GWAS and Hp-GWAS for both DTA and PHT (R²=0.042 and R²=0.086 respectively). The signals from Hp-GWAS were generally stronger than those detected via conventional SNP-based GWAS, indicating that Hp-GWAS can provide higher statistical power as compared to conventional SNP-based approach. Only a few signals could
be detected by both eRD-GWAS and Hp-GWAS (Supplementary Fig1 c, d). The correlations between the model frequencies obtained from eRD-GWAS and Hp-GWAS are even weaker than between conventional SNP-based GWAS and Hp-GWAS ($r^2=0.0035$ and $r^2=0.0025$ for DTA and PHT respectively; Fig. 5 B, D).

Only one gene (GRMZM2G132854) was detected by all three methods for PHT trait. Thus, the probability of detecting the signals via Hp-GWAS by either eRD-GWAS or conventional SNP-based GWAS is low. Hence, Hp-GWAS is complementary to conventional SNP-based GWAS and eRD-GWAS.

**Geographical GWAS**

Artificial selection on maize has resulted in dramatic genetic [20] and phenotypic changes [21] relative to its wild progenitor, teosinte (Zea mays subspecies parviglumis) [22]. To use Hp-GWAS to identify haplotypes that have experienced strong selection we replaced the phenotype as the response variable with the location (i.e. longitude and latitude) at which inbreds were developed (Methods; Fig. 6). The annotations of genes associated with longitude and latitude are enriched in plant development, hormones response and nutrient transport (Fig. 6). Gene annotation GO categories related to nitrate transport and root structure were enriched in the associations with both longitude and latitude (Table 2; Methods). Genes associated with longitude only are enriched in GO categories for salt response, cell structure and root structure. In contrast, genes associated with latitude only are
enriched in GO categories for starch metabolism, amino acid response and plant hormones response (Table2).

Over the past 10,000 years, maize has undergone dramatic transformations in response to its adaptation to new growing environments [23]. Thus, we hypothesized that genes associated with latitude and longitude might overlap with domestication and improvement genes. Comparisons of a list of maize domestication and improvement genes [21] with location-associated genes identified at different Hp-GWAS model frequency cutoffs (Fig. 7) revealed that latitude-associated genes are significantly enriched among improvement genes. Therefore, we infer that during a centuries of maize breeding process, human selection of maize based on latitude is stronger than longitude.

The ZmPWD-like gene (GRMZM2G040968) was only detected via association with latitude. PWD encodes a water dikinase protein that has a positive effect on transitory starch degradation at night and is required for proper starch metabolism under different light/dark cycles [24]. ZmPWD-like haplotype 8 (ZmPWD-like_H08) was detected by latitude Hp-GWAS with relatively high model frequency (0.179) and the ZmPWD-like gene is also included in the improvement gene list [21] (Fig. 8A). The ZmPWD-like_H08 haplotype clusters at latitudes 20°-30°. These latitudes are lower than the traditional corn-belt latitudes (35°N-40°N) (Fig. 8B). Most maize inbreds contain ZmPWD-like_H08 that was inherited from Mexico and South Africa (Fig. 8C). In summary, these results provide support for the
view that combining Hp-GWAS and geographical information can detect genes which contributed to crop improvement and/or adaptation.

**Discussion**

We developed Hp-GWAS, a novel analysis pipeline that associates genic haplotypes to with phenotype using a Bayesian-based framework. The relative accuracy of using haplotypes as explanatory variables was evaluated via 10-fold cross-validation which showed it to have equal or higher accuracy as compared to single SNPs or expression values. In addition, a reverse genetic approach confirmed that mutants of two genes associated with traits via Hp-GWAS have the expected phenotypes.

**Genic haplotypes provide stronger signal than single SNP**

Most of GWAS studies are conducted using the conventional single SNP approach. In theory, haplotypes include the effects of several SNPs and should therefore provide greater statistical power than single SNPs [25]. In addition, the use of haplotypes reduces the number of explanatory variables, thereby reducing the multiple testing problem. For example, we collected 1.28M genome-wide SNPs, from which only about 100K genic haplotypes were used for GWAS. The phasing haplotypes could be of two main types, fixed window- length or model-based [10]. Each of them provides solution for phasing haplotype but also produces biases. The window-length approach requires arbitrarily defined window length, and the
model-based approach suffers errors from possibility model. In our approach, we directly use RNA-Seq reads alignment to reference and then phasing the genic haplotype directly to avoid the issues discussed.

**Hp in MLM**

Hp-GWAS associates genic haplotype with trait. A similar approach was used by Yano et al. (2016) [16]. They deeply sequenced 176 rice inbred genomes and identified genic haplotypes prior to conducting a MLM-based GWAS. However, MLMs typically estimate genetic effect based on a single marker and require to use covariance to control the population structure both of which decrease statistical power [4]. Our Bayesian approach overcomes these challenges. Furthermore, geographical GWAS would be difficult to conduct using a conventional MLM framework, due to the large proportional variance explained by population structure [23]. Instead of MLM, our Bayesian-based GWAS framework in Hp-GWAS, which is a multiple variable regression and could control the population structure among each MCMC iteration automatically. Further, because the genic haplotypes in the current study were generated from RNA-Seq data, that allows conducting both Hp-GWAS and eRD-GWAS concurrently without extra effort. Because Hp-GWAS is complementary with eRD-GWAS, we believe this combined approach can comprehensively detect trait-associated loci using only inexpensive RNA-Seq data.
Different types of Haplotype

In this study, we used RNA-seq data to call SNP located in exons. However, SNPs within intron can regulate splicing and expression which might also affect phenotype. So we extended our pipeline to call haplotypes bases on all genic SNPs (i.e., both exonic and intronic SNPs), then conducted Hp-GWAS and compared the results to the previous exon only results (Supplementary Fig2). To make a comparison, we use the highest model frequency from the two Hp-GWAS to represent each gene. The overall correlation is 0.524 (p-value < 2.2e-16) for the regression line, indicating that the results are correlated between haplotypes obtained from the two sources of SNPs. Some genes exhibit higher model frequencies when we include intronic SNPs, while others exhibit lower model frequencies. We hypothesis the main effect might not come from the intron region and include more unrelated variants which introduces biases.

Further improvements in Hp-GWAS

In this study, genic SNPs were used to identify haplotypes. One obvious drawback of this strategy is that ignores information from outside genes. Because the sequencing platforms and methods are evolving rapidly, we believe that it will be possible to combine all the information from genome-wide eQTL mapping, genetic regulatory network and SNPs located within promoters or enhancers to reconstruct a “systemic haplotype” in the near future, instead of collecting geneic SNPs naively. The systemic haplotype approach should
include both cis/trans regulation of genes at the same time. Fig. 4 shows that even with the exclusion of the information outside the genic region, the Hp-GWAS still provides similar accuracy with the conventional SNP-based method, which contains genome-wide SNPs. Hence, we hypothesize that a systemic haplotype approach should further improve Hp-GWAS accuracy and statistical power in the future.

Methods

Plant materials, Phenotype, RNA-Seq library and SNP-calling

The RNA-Seq library preparation, raw reads processing, SNP-calling, plant grown conditions, tissue preparation and field experiment design for two RNA-Seq datasets used in this study were conducted [2]. DTA and PHT trait data were downloaded from Peiffer et al [26].

Assignment of Haplotypes

SNPs in the exon region of maize B73 canonical transcripts were used to identify gene haplotypes. An algorithm was developed to handle the missing values in SNP calling, in which the missing values were first treated as a SNP, and then the principle of incompatibility was used to reduce the false-positive rate of haplotypes, i.e., two haplotypes that differed only due to missing values were grouped into one. All of these processes were implemented using in-house Perl scripts.
Negative binomial haplotype tissue model

Negative Binomial GLMs were employed to investigate main effects of haplotype, tissue and haplotype-by-tissue interaction on gene expression among 27 NAM founders and 5 tissues.

\[ \log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \tau_{ij} + o_{ijk} \]

where \( \lambda_{ijk} \) is mean fragment counts for haplotype \( i \), tissue \( j \), and observation \( k \), \( \mu \) is the overall mean effect, \( \alpha_i \) is haplotypes effect, \( \beta_j \) is tissue effect and \( \tau_{ij} \) is haplotype by tissue effect and \( o_{ijk} \) is the offset. Log of 0.75 quantiles of fragment counts distribution was used as the offset. The genmod function of SAS 9.3 was used in parameter estimations of negative binomial GLMs. Type III analysis was performed to test effects of haplotype, tissue and haplotype-by-tissue interactions on gene expression. To avoid non-convergence issues due to zero read counts, only 24,690 genes with mean read counts > 10 and number of samples with positive read counts > 40 were used in the modeling.

Different GWAS (GenSel)

The Bayesian-based GWAS framework was applied under GenSel v4.1 [6]. The basic model of Bayesian-based GWAS is:

\[ y = X\beta + Z\upsilon + e \]

where \( y \), \( \beta \), \( \upsilon \) are the phenotype, unknown fixed effect vectors, and a random vector respectively. The error term, \( e \) follows a normal distribution with 0 mean and covariance
matrix $\sigma_e^2 R$ (where $R$ is a diagonal matrix). Instead of solving the specific genetic effect in MLM, Bayesian-based GWAS applies MCMC (Markov chain Monte Carlo) to lower the unrelated markers and amplified associated signals. The BayesC approach was used in this study [6]. The fractions of markers having no effect were set at 0.9996, 0.995 and 0.996 in SNP-BayesC, eRD-GWAS [2] and Hp-GWAS respectively. Model allele frequency was set at 0.05 for given haplotype. A chain length of 41,000 was used and the first 1,000 iterations were set as burn-in. The significant cutoffs were set at 0.02 for the three types of GWAS. The $\sigma_a^2$ and $\sigma_e^2$ follow the inverse Chi-squared distribution with degree of freedom 4 and 50% of phenotypic variations as prior for the test-runs. The posterior estimations of $\sigma_a^2$ and $\sigma_e^2$ from test-run were used as the priors for the real run. The accuracy of different GWAS results was estimated by 15 times 10-folds cross-validation.

**UniformMu reverse genetics, phenotyping and qRT-PCR**

A transposon-based reverse genetic resource [27] was used to test gene/trait associations detected via Hp-GWAS. Two Mu insertion mutants from the UniformMu resource were phenotyped in Ames, Iowa during the summer of 2018. The boxplots and t-test were generated by an in-house R script [28]. qRT-PCR was conducted using RNA samples from four mutant and four wildtype V1 stem RNA samples. Each sample contained three biological replicates. Maize actin1 (GRMZM2G126010) gene expression was treated as the reference to calculate relative gene expression.
Geographical information and GO-test

The geographical information for SAM parents was obtained from the U.S. National Plant Germplasm System (https://npgsweb.ars-grin.gov/gringlobal/search.aspx). A state’s central point of latitude and longitude was used to represent the location of all inbreds from that state. For those inbreds from out of U.S the central point of the entire country of origin was used as the location. GO enrichment tests for geographical Hp-GWAS signals were performed with GOseq version 1.20.0 [29]. The GO enrichment cutoff was set to 0.05 FDR. The maize GO data were downloaded from the maize-GAMER project [30].

Data availability

27 NAM founders and 5 tissues raw reads are available in NCBI Sequence Reads Archives (SRA) (http://www.ncbi.nlm.nih.gov/sra) under accession numbers SRA050451 (apex) and SRA050790 (ear, root, shoot, and tassel). The accession number for SAM diversity panel RNA-Seq raw reads is SRP055871.

Reference


## Tables

Table 1. Candidate gene list for the three different GWAS methods: SNP-based, eRD-GWAS and Hp-GWAS.

<table>
<thead>
<tr>
<th>GeneID</th>
<th>Trait</th>
<th>SNP</th>
<th>eRD</th>
<th>Hp</th>
<th>Maize annotation</th>
<th>Arabidopsis homologous</th>
<th>Rice homologous</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G179264</td>
<td>DTA</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>ZmZCN8</td>
<td>(FT) PEBP family protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G143525</td>
<td>DTA</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Polynucleotidyl transferase, ribonuclease H-like superfamily protein</td>
<td>exonuclease, putative</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G171650</td>
<td>DTA</td>
<td>X</td>
<td></td>
<td>X</td>
<td>ZmMADS69</td>
<td>AGAMOUS-like 79</td>
<td>OsMADS65</td>
</tr>
<tr>
<td>GRMZM2G147716</td>
<td>DTA</td>
<td>X</td>
<td></td>
<td>X</td>
<td>ZmMADS67</td>
<td>AGAMOUS-like 8</td>
<td>OsMADS18</td>
</tr>
<tr>
<td>GRMZM5G835516</td>
<td>DTA</td>
<td>X</td>
<td></td>
<td>X</td>
<td>(NIK1) NSP-interacting kinase 1</td>
<td>OsBAK1</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G010693</td>
<td>DTA</td>
<td>X</td>
<td></td>
<td>X</td>
<td>NSP-interacting kinase 3</td>
<td>OsBAK1</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G154320</td>
<td>DTA</td>
<td>X</td>
<td>LBD-transcription factor 10</td>
<td>ASYMMETRIC LEAVES 2-like 1</td>
<td>DUF260 domain containing protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRMZM2G072280</td>
<td>DTA</td>
<td>X</td>
<td></td>
<td>X</td>
<td>Photosystem I light harvesting complex gene 2</td>
<td>Chlorophyll A-B binding protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G004483</td>
<td>DTA</td>
<td>X</td>
<td>ZmCCT9</td>
<td></td>
<td>CONSTANS-like 4</td>
<td>CCT motif family protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G132880</td>
<td>DTA</td>
<td>X</td>
<td>ZmZCN6</td>
<td></td>
<td>Centroradialis</td>
<td>RCN1 Centroradialis-like1 homogous to TFL1 gene</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G163798</td>
<td>DTA</td>
<td>X</td>
<td></td>
<td>X</td>
<td>Walls Are Thin 1</td>
<td>auxin-induced protein 5NG4</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G392101</td>
<td>PHT</td>
<td>X</td>
<td></td>
<td></td>
<td>crr2 (cytokinin response regulator2)</td>
<td>response regulator 6 OsRR6 type-A response regulator</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G044175</td>
<td>PHT</td>
<td>X</td>
<td></td>
<td></td>
<td>sphingoid base hydroxylase 2</td>
<td>sphingolipid C4-hydroxylase SUR2</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G130230</td>
<td>PHT</td>
<td>X</td>
<td></td>
<td></td>
<td>nactf39 (NAC-transcription factor 39)</td>
<td>NAC domain containing protein 36 no apical meristem protein</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G131254</td>
<td>PHT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>rhn1 (ras-related protein RHN1)</td>
<td>RAB homolog 1 ras-related protein</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. GO enrichment among Geography Hp-GWAS. A. GO enrichment for longitude Hp-GWAS. B. GO enrichment for latitude Hp-GWAS.

<table>
<thead>
<tr>
<th>Category</th>
<th>Ontology</th>
<th>Over Represented p-value</th>
<th>q-value</th>
<th>No. DE In Category</th>
<th>No. In Category</th>
<th>term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0018996</td>
<td>BP</td>
<td>0.0035</td>
<td>0.0493</td>
<td>4</td>
<td>63</td>
<td>molting cycle, collagen and cuticulin-based cuticle</td>
</tr>
<tr>
<td>GO:0061088</td>
<td>BP</td>
<td>0.0045</td>
<td>0.0493</td>
<td>2</td>
<td>9</td>
<td>regulation of sequestering of zinc ion</td>
</tr>
<tr>
<td>GO:0080055</td>
<td>BP</td>
<td>0.0073</td>
<td>0.0493</td>
<td>2</td>
<td>12</td>
<td>low-affinity nitrate transport</td>
</tr>
<tr>
<td>GO:0009651</td>
<td>BP</td>
<td>0.0085</td>
<td>0.0493</td>
<td>21</td>
<td>1199</td>
<td>response to salt stress</td>
</tr>
<tr>
<td>GO:0016036</td>
<td>BP</td>
<td>0.0086</td>
<td>0.0493</td>
<td>7</td>
<td>223</td>
<td>cellular response to phosphate starvation</td>
</tr>
<tr>
<td>GO:0048765</td>
<td>BP</td>
<td>0.0192</td>
<td>0.0493</td>
<td>8</td>
<td>327</td>
<td>root hair cell differentiation</td>
</tr>
<tr>
<td>GO:0009870</td>
<td>BP</td>
<td>0.0279</td>
<td>0.0493</td>
<td>2</td>
<td>27</td>
<td>defense response signaling pathway, resistance gene-dependent</td>
</tr>
<tr>
<td>GO:0015698</td>
<td>BP</td>
<td>0.0248</td>
<td>0.0493</td>
<td>2</td>
<td>24</td>
<td>inorganic anion transport</td>
</tr>
<tr>
<td>GO:0048513</td>
<td>BP</td>
<td>0.0294</td>
<td>0.0493</td>
<td>31</td>
<td>2334</td>
<td>organ development</td>
</tr>
<tr>
<td>GO:0080054</td>
<td>MF</td>
<td>0.0073</td>
<td>0.0493</td>
<td>2</td>
<td>12</td>
<td>low-affinity nitrate transmembrane transporter activity</td>
</tr>
<tr>
<td>GO:0005876</td>
<td>CC</td>
<td>0.0453</td>
<td>0.0493</td>
<td>3</td>
<td>77</td>
<td>spindle microtubule</td>
</tr>
</tbody>
</table>
Figure 1. Summary statistics of haplotypes and contributions to tissue specific expression. a) Distribution of haplotype missing rate in NAM RNA-Seq dataset. b) Distribution of haplotype diversity of genes. The red dashed line is 0.5 haplotype diversity and the blue dashed line is 0.05 haplotype diversity. c) Venn Diagram of genes with significant haplotypes, tissue and haplotype by tissue interactions.
Figure 2. Manhattan plots of the three types of GWAS results for a) DTA and b) PHT. The upper panel shows the results from a SNP-based BayesC and the middle and lower panels show results from eRD-GWAS and Hp-GWAS, respectively. The horizontal red dashed lines represent 0.02 model frequency cutoff for all three analyses. Gene IDs and annotations are indicated.
Figure 2 continued.
Figure 3. Confirmation of Hp-GWAS results through Mu transposon inserted mutants a) ZmCCT9 and b) ZmBAK1-like. The bold blue arrow indicates the location of the Mu insertion. The small green arrow is the genic SNP location. The red color SNPs indicate the sequence of trait associated haplotype. The boxplots report the phenotype distribution between the whole population except the trait-associated haplotype and the trait-associated haplotype. The bar-plots represent the phenotypes between wild-type and homozygous Mu insertion mutants. The qRT-PCR result demonstrated mu inserted mutant show low ZmBAK1-like expression.
Figure 4. 10-fold cross-validation results of the different methods for a) PHT and b) DTA.

The t-test p-values are also shown for each method.
Figure 5. Differences between Hp-GWAS, eRD-GWAS and conventional SNP BayesC for DTA and PHT traits. The blue line indicates the regression lines between the two axes. The red dashed lines are the cut-offs of the two methods. The model frequency difference A) between Hp-GWAS and SNP BayesC for DTA trait; B) between Hp-GWAS and eRD-GWAS for DTA trait; C) between Hp-GWAS and SNP BayesC for PHT trait; D) between Hp-GWAS and eRD-GWAS for PHT trait.
Figure 6. Manhattan plots of geographical Hp-GWAS for longitude and latitude. The upper and lower panels show longitude Hp-GWAS and latitude Hp-GWAS, respectively. The model frequency cutoff for all three analyses are 0.02 (indicated in red dashed line). Overlapping associated haplotypes are indicated by vertical green dashed lines. Gene IDs and annotations are indicated.
Figure 7. Enrichment testing for geographical Hp-GWAS among improvement and domestication genes. The number of haplotypes above indicated model frequency cutoff on x-axis are shown within each plot. P-value=0.05 are indicated in read dashed lines. (A, B) Enrichment of improvement and domestication genes among longitude Hp-GWAS. (C, D) Enrichment of improvement and domestication genes among latitude Hp-GWAS.
Figure 8. a) Latitude Hp-GWAS signals and improvement region XP-CLR test. b) Latitude distribution of ZmPWD-like_H08 c) Geographical distribution of ZmPWD-like_H08.
Supplemental Figures

Figure S1. Venn diagrams of SNP-based, eRD-GWAS and Hp-GWAS in DTA and PHT traits.

Figure S2. The comparison between Hp-GWAS results in DTA trait from exonic and both exonic and intronic SNPs. The red dash is 45-degree line. The blue line is the correlation line.
Supplemental Tables

Table S1: Functional enrichment among genes with < 5% or > 50% haplotype diversity.

<table>
<thead>
<tr>
<th>Haplotype diversity</th>
<th>Function Layer</th>
<th>Functional Category</th>
<th>Adjusted p-val</th>
<th>Log2 odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5%</td>
<td>L3</td>
<td>Signalling receptor kinases wall associated kinase</td>
<td>1.37E-04</td>
<td>3.69</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>RNA regulation of transcription AS2 lateral Organ boundaries Genes family</td>
<td>2.57E-04</td>
<td>3.15</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>RNA regulation of transcription MADS box transcription factor family</td>
<td>3.05E-04</td>
<td>2.23</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>Secondary metabolism isoprenoids terpenoids</td>
<td>3.36E-02</td>
<td>1.88</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>Stress abiotic unspecified</td>
<td>1.37E-04</td>
<td>1.60</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>RNA regulation of WRKY transcription factor family</td>
<td>8.44E-05</td>
<td>1.60</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>Protein degradation cysteine protease</td>
<td>7.79E-04</td>
<td>1.39</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>RNA regulation of transcription AP2/EREBP, APETAL.A2/Ethylene responsive element binding protein family</td>
<td>8.44E-05</td>
<td>1.35</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>RNA regulation of transcription MYB transcription factor family</td>
<td>1.68E-06</td>
<td>1.31</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>RNA regulation pf transcription putative transcription regulator</td>
<td>3.31E-02</td>
<td>-0.60</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>Amino acid metabolism synthesis branched chain group</td>
<td>3.31E-02</td>
<td>-2.69</td>
</tr>
<tr>
<td>&gt;50%</td>
<td>L3</td>
<td>Signalling receptor kinases leucine rich repeat IV</td>
<td>3.22E-02</td>
<td>5.36</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>Signalling receptor kinases leucine rich repeat III</td>
<td>1.06E-06</td>
<td>2.51</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>RNA regulation of transcription ARF Auxin response factor family</td>
<td>1.54E-02</td>
<td>1.89</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>Protein degradation ubiquitin</td>
<td>3.29E-02</td>
<td>-0.43</td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>Protein synthesis ribosomal protein</td>
<td>1.12E-03</td>
<td>-0.90</td>
</tr>
</tbody>
</table>
Table S2: Fisher’s exact test for overlapped gene number between SNP-based, eRD-GWAS and Hp-GWAS.

<table>
<thead>
<tr>
<th></th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=138</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hp-GWAS (n=300)</td>
<td>SNP-based 2</td>
<td>SNP-based 2</td>
</tr>
<tr>
<td></td>
<td>eRD-GWAS (n=372) 7</td>
<td>eRD-GWAS (n=171) 1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.28</td>
<td>0.18</td>
</tr>
<tr>
<td>Conclusion</td>
<td>not significant</td>
<td>not significant</td>
</tr>
</tbody>
</table>

* indicate the p-value < 0.05
CHAPTER 5. GENERAL CONCLUSION

Identifying genetic loci is the first step to study the regulation of complex traits. Since the central dogma of molecular biology (DNA transcripts into RNA, RNA translates into protein) explained the basic mechanism of the biological system in 1958, DNA has been treated as a template of life. People conducted a large amount of research on DNA level to explain the observed phenotypes. Now, more and more studies support that the variations within different genomic data play important roles in affecting traits. This thesis consists of three methods to associate final phenotypes either based on RNA-Seq data or haplotypes within DNA level both are fundamental for exploring more loci contributed to important agronomy traits.

In Chapter 2, we developed a Bayesian-based method known as expression read depth genome-wide association study (eRD-GWAS), which can use transcript accumulation as the explanatory variable to directly associate the variation among RNA level with phenotypes. Based on the eRD-GWAS results, we found transcription factors are enriched among the significant eRD-GWAS signals. All the eRD genes were tested by RNA co-expression and protein-protein interaction networks and found eRD genes clustered into the modules with GO categories related to phenotypes. Furthermore, the accuracy of eRD-GWAS was also supported by 10-fold cross-validation. These results provide evidence that eRD-GWAS can associate transcriptomic variation to final phenotypes.

We used a different aspect of transcriptome data, alternative splicing (AS), to associate phenotypes and further calculate the genetic regulatory network for AS in Chapter 3. AS is an important genetic mechanism for increasing transcriptome and proteome diversity in all eukaryotes. We create a pipeline called AS-GWAS that applies MLM and Bayesian-
based models to associate the genome-wide alternative splicing patterns with phenotypes. Over 100 trait-associated AS events were detected that provides a genome-scale evidence to show the effect of AS on the phenotype. Genetic regulatory network (GRN) of the significant AS signals show that some essential regulatory genes serve as hub genes within the network. Comparing GRN for AS and gene expression we found that AS utilize different regulation mechanisms from canonical gene expression. In summary, our findings emphasize the importance of AS in phenotypes. Further, our analyses framework provides another strategy to study complex relationships between AS and phenotypes.

Since we have associated gene expressions and AS pattern with phenotypes we asked a different question in Chapter 4. Can we infer variants within DNA from RNA-Seq data and integrate genic variants to further associate with phenotypes? We developed the haplotype genome-wide association study (Hp-GWAS) pipeline that directly applies genic haplotypes as the explanatory variable to associate with phenotypes. Both cross-validation and reverse-genetic transposon knockout mutants supported the accuracy of Hp-GWAS. Based on the comparisons between results from three types of GWAS, we concluded that Hp-GWAS is complementary to eRD-GWAS and SNP-based GWAS. In a practical view, because the genic haplotypes were generated from RNA-Seq data, that allowed us to conduct both Hp-GWAS and eRD-GWAS concurrently without extra effort. In summary, the Hp-GWAS is a novel gene-based method which is accurate and complementary to conventional GWAS approaches.

Overall, these three projects provide novel methods to detect more trait-associated loci and understand the complex gene regulatory systems. eRD-GWAS and AS-GWAS efficiently target the candidates among two aspects of the transcriptome, gene expression and
alternative splicing. Following networks and GRN analyses provide comprehensive information to understand the complex regulatory mechanism. In summary, this thesis provides comprehensive information to understand the complex gene regulatory mechanisms in RNA and DNA level and will benefit plants, animals, medical and disease treatment development studies in the future.