2014

# Genome-wide association studies to dissect the genetic architecture of yield-related traits in maize and the genetic basis of heterosis

Jinliang Yang
*Iowa State University*

# Genome-wide association studies to dissect the genetic architecture of yield-related traits in maize and the genetic basis of heterosis

by

**Jinliang Yang**

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Interdepartmental Genetics

Program of Study Committee:
Patrick S Schnable, Major Professor
Dan Nettleton
Erik Vollbrecht
Nick Lauter
Jack Dekkers

Iowa State University

Ames, Iowa

2014



Graduate College
Iowa State University

**TABLE OF CONTENTS**

Page

# CHAPTER 1. GENERAL INTRODUCTION

## Introduction

### Maize genetics in the next generation sequencing (NGS) era

Taking advantage of NGS technologies, large amounts of sequencing data are being generated at an unprecedented rate. According to the records in NCBI database, ~13,000 Gb (gigabases) of maize NGS data have been deposited. This is equivalent to ~6,000 depth of coverage of the maize genome, assuming a maize genome size of 2.3 GB. More than 90% of these data were generated after the completion of the maize B73 reference genome project in 2009 (Schnable et al. 2009).

Various types of maize NGS data have been deposited, such as genomic re-sequencing a group of six elite maize inbred lines with high depth (Lai et al. 2010), transcriptome profiling of 503 maize inbred lines using RNA-seq (Hirsch et al. 2014), comprehensive genotyping of 2,815 maize inbred lines using genotyping-by-sequencing (GBS) (Romay et al. 2013). As the amount of data generated has soared, the interpretation of these data for biological meaning has become a challenge.

One typical application of NGS data is to identify genetic variants after aligning the short reads to the reference genome. The maize HapMap projects (Gore et al. 2009; Chia et al. 2012) sequenced 27 parental lines of the nested association-mapping (NAM) population and 103 historical lines, which yielded 1.6 million and 55 million variants, respectively. These variants were imputed onto the NAM populations to identify thousands of trait-associated variants (TAVs) (Poland et al. 2011; Tian et al. 2011; Cook et al. 2012; Peiffer et al. 2013; Peiffer et al. 2014).

Another application of NGS is to estimate the transcript abundance via RNA-seq. Li *et al.* sequenced ~100 recombinant inbred lines (RILs) from intermated B73 and Mo17 (IBM) population (Li et al. 2013b). The transcript abundance for each gene was treated as a quantitative trait and thousands of expression QTL (eQTL) were identified including some regulation hotspots. Recently, to study the maize oil biosynthesis, 368 maize inbred lines, including the high-oil lines, were sequenced

via RNA-seq (Fu et al. 2013). After analysis, 26 loci were claimed to be associated with the trait of oil concentration and could explain 83% of the phenotypic variation (Li et al. 2013a). In addition, with the detected eQTLs, a large-scale gene regulatory network was established (Fu et al. 2013).

**Challenges of NGS-enabled GWAS approaches**

With the tremendous amount of heterogeneous NGS data, combing the data and making meaningful biological interpretations is challenging. GWAS (Genome-wide Association Study) is a method to identify the genetic control of the quantitative traits. GWAS can be conducted using variants called using NGS data.

Although multiple statistical approaches for conducting GWAS exist, including both single-variant and multi-variant approaches, there is not yet a consensus about which approach performs the best (Bush and Moore 2012). Single-variant analyses compare the phenotypic distributions of alternative genotypes at each polymorphic site independently. In comparison, multi-variant approaches can explicitly account for large effect loci and estimate effects of multiple polymorphic sites simultaneously. Recently, Bayesian-based multi-variant approaches have been used for GWAS (Fan et al. 2011; Fernando and Garrick 2013). These approaches fit a mixed model, where the effects of variants are treated as random, with prior assumptions regarding the distributions of their effects.

GWAS is typically associated with high rates of false discovery (Visscher et al. 2012). In human studies, a second cohort is often used to cross-validate the most significant SNPs discovered in the first cohort, thereby cost effectively reducing the number of false discoveries (Sladek et al. 2007). To our knowledge GWAS experiments in plants have not been subjected to cross-validation.

**Diverse genetic materials for cross-validation a GWAS**

For many plant species, large and diverse collections of germplasm accessions, including wild relatives, landraces, and breeding lines, have been collected and are available for analyses. To utilize these diverse genetic materials for a cross-validation experiment, the kernel row number (KRN) phenotype was selected for

our analyses. KRN is a component trait of yield and also a model trait for genetic studies (Hallauer et al. 2010). It is highly heritable and exhibits little variation in response to environment (Lu et al. 2011). In addition, it is easily scored as an integer, and this scoring can be conducted after completion of the busy pollination season.

USDA Plant Introduction Station maintains a set of elite inbred lines, which are commercial lines that have been subject to IP (Intellectual Property) protection via the Plant Variety Protection (PVP) act. These inbred lines are ideal for a cross-validation experiment. In addition, the germplasm resources information network (GRIN) database contains KRN records of ~7,000 accessions. The KRN variation ranges from 4 rows to more than 30 rows according to the records. Besides, a long-term selection project conducted by Arnel Hallauer and his colleagues at Iowa State University aimed to divergently select long and short ears from a single founder population (BSLE) (Hallauer et al. 2004; Hallauer 2005). Parental lines of BSLE and bulked seeds from cycle 0 and cycle 30 are available to us. All these materials and resources are available to cross-validate the initial findings of GWAS and generate a lower bound for true positive GWAS results.

**Opportunities to understand heterosis via GWAS**

Even with the development of NGS and advancement of statistical approaches such as GWAS, many classical genetic questions remain unexplained. This includes heterosis, which refers to the phenomenon that the progeny of diverse inbred lines exhibit improved phenotypic performance as compared to their inbred parents. Researchers proposed many genetic models, including dominance, overdominance and epistasis to explain heterosis (Birchler et al. 2003; Goff and Zhang 2013). However, heterosis phenomena could not be adequately explained by a single gene or a simple model. Because of the quantitative nature of the heterotic traits, large population sizes and high marker densities are required to better characterize the genetic composition pertaining to heterosis. Recently, with ultra-high-density maps, researchers studying rice hybrids demonstrated that the accumulation of multiple effects, including dominance, overdominance (or pseudo-overdominance), and dominance by dominance interactions could largely explain the genetic basis of

heterosis, although their relative contribution varied with different traits (Zhou et al. 2012).

### Opportunities to enhance the identification of trait-associated variants (ATVs)

Despite the development of QTL mapping (Morton 1955), GWAS (Klein et al. 2005) and the bulk segregant analysis (BSA) (Michelmore et al. 1991), it remains challenging to rapidly and cost-effectively identify SNPs or genes associated with variation in complex traits. QTL mapping and GWAS approaches require large numbers of individuals to be genotyped and phenotyped, which can be expensive for large populations even using recently developed cost-effective genotyping methods (Elshire et al. 2011). The BSA method requires to genotype pools of individuals sorted by phenotype, however, it can only conducted on a bi-parental segregating population. To meet the needs of genotype-phenotype association mapping for researchers studying minor crops, a rapid and cost-effective method needs to be developed.

## Research Goals

### Identification and cross-validation of GWAS loci controlling KRN variation

KRN is a consequence of inflorescence branching, which is determined by the fates and identities of an array of meristems (Barazesh and McSteen 2008). Traditional genetic analyses have identified several genes or metabolic pathways relevant to inflorescence development. For instance, *fasciated ear2* (*fea2*) (Bommert et al. 2013) and *thick tassel dwarf1* (*td1*) (Bommert et al. 2005) caused fasciated ear phenotypes with irregular but higher KRNs. *Fea2* and *td1* are homologs of *Arabidopsis CLV1* and *CLV*2, respectively, both of which belong to the CLV-WUS regulatory pathway that promotes stem cell differentiation (Clark 2001). Even with these findings, the detailed genetic control of KRN still remains unknown, especially for minor effect loci. One of our goals was to characterize the genetic architecture controlling the KRN phenotype. Uncovering additional candidate genes and statistically enriched pathways will expand our understanding of the developmental processes involved in ear development (Chapter 2).

**Comparison of different statistical approaches for conducting GWAS**

We conducted GWAS with three distinct approaches, including single-variant, stepwise regression and Bayesian-based multi-variant approaches. All of these approaches have been widely used in conducting GWAS. However, to the best of our knowledge, there was no such a comparison with cross-validation to directly assess false discovery rates of different approaches. Therefore, the second goal of this research was to compare the performance of different GWAS approaches (Chapter 2).

**Genetic architecture of yield-related traits and insights into heterosis**

Seven yield-related traits for which we collected data exhibited varying degrees of heterosis. With the dissection of the genetic architecture of these traits via GWAS, the mode of inheritance for the trait-associated variants (TAVs) could be estimated. Consequently the third goal of this research was to find patterns of gene actions for the identified TAVs in the seven traits showing different degrees of heterosis. The knowledge gained here will help to better understand heterosis (Chapter 3).

**Development of a rapid and cost effective method to identify TAVs**

Although QTL mapping and GWAS are widely used for the identification of TAVs, it remains challenging to rapidly and cost-effectively identify variants and therefore, genes associated with variation in complex traits, especially for species without well-established genetic mapping populations. To meet this challenge, we developed a novel approach of NGS exome-sequencing using pools of individuals that exhibit extreme phenotypes from a large diversity panel to identify TAVs. This fourth goal of method development is of interest for researchers who study taxa for which large and individually genotyped diversity panels do not exist (Chapter 4).

## Dissertation Organization

This dissertation includes a general introduction (Chapter 1), three journal manuscripts (Chapters 2 to 4) and a section of general conclusions (Chapter 5). The paper in Chapter 2, which compares three statistical approaches for conducting

GWAS and identifies the genetic architecture controlling for KRN trait, has been submitted for publication. I made major contributions include designing and performing the experiments, analyzing data and writing the manuscript under the guidance of Dr. Schnable. Dr. Nettleton and Dr. Dekkers provided technical support and conceptual advice for this work. The paper in Chapter 3, which investigates the modes of inheritance of trait-associated variants for seven yield-related traits and provides insight into heterosis, will soon be submitted for publication. I made major contributions to this experiment that include designing the experiments, supervising the data collection, data investigations, and writing the manuscript under the guidance of Dr. Schnable. Dr. Nettleton provided technical support and conceptual advice for this work. The paper in Chapter 4 reports a new method (termed XP-GWAS,) that uses pools of extreme-phenotype for conducting GWAS, will also be submitted for publication. My contributions to this paper include developing the concept, designing experiments, analyzing data and writing the manuscripts under the guidance of Dr. Schnable. The co-first author, Haiying Jiang assisted with data collection. Dr. Nettleton advised on the data analysis and wrote some custom R scripts for the data analysis.

## References


Barazesh S, McSteen P. 2008. Hormonal control of grass inflorescence development. *Trends Plant Sci* **13**(12): 656-662.

Birchler JA, Auger DL, Riddle NC. 2003. In search of the molecular basis of heterosis. *The Plant cell* **15**(10): 2236-2239.

Bommert P, Lunde C, Nardmann J, Vollbrecht E, Running M, Jackson D, Hake S, Werr W. 2005. thick tassel dwarf1 encodes a putative maize ortholog of the Arabidopsis CLAVATA1 leucine-rich repeat receptor-like kinase. *Development* **132**(6): 1235-1245.

Bommert P, Nagasawa NS, Jackson D. 2013. Quantitative variation in maize kernel row number is controlled by the FASCIATED EAR2 locus. *Nature genetics* **45**(3): 334-337.

Bush WS, Moore JH. 2012. Chapter 11: Genome-wide association studies. *PLoS computational biology* **8**(12): e1002822.

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet* **44**(7): 803-807.

Clark SE. 2001. Cell signalling at the shoot meristem. *Nat Rev Mol Cell Biol* **2**(4): 276-284.

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia SA. 2012. Genetic architecture of maize kernel composition in the

nested association mapping and inbred association panels. *Plant Physiology* **158**(2): 824-834.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**(5): e19379.

Fan B, Onteru SK, Du ZQ, Garrick DJ, Stalder KJ, Rothschild MF. 2011. Genome-Wide Association Study Identifies Loci for Body Composition and Structural Soundness Traits in Pigs. *Plos One* **6**(2).

Fernando RL, Garrick D. 2013. Bayesian methods applied to GWAS. *Methods in molecular biology* **1019**: 237-274.

Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, Zhang J, He C, Du X, Peng Z et al. 2013. RNA sequencing reveals the complex regulatory network in the maize kernel. *Nature communications* **4**: 2832.

Goff SA, Zhang Q. 2013. Heterosis in elite hybrid rice: speculation on the genetic and biochemical mechanisms. *Current opinion in plant biology* **16**(2): 221-227.

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J et al. 2009. A first-generation haplotype map of maize. *Science* **326**(5956): 1115-1117.

Hallauer AR. 2005. Registration of BSLE(M-S)C30 and BSLE(M-L)C30 maize germplasm. *Crop Sci* **45**(5): 2132.

Hallauer AR, Carena MJ, Filho JBM. 2010. Quantitative Genetics in Maize Breeding. *Handb Plant Breed* **6**: 1-663.

Hallauer AR, Ross AJ, Lee M. 2004. Long-term divergent selection for ear length in maize. *Plant Breeding Reviews* **24**(2): 153-168.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Penagaricano F, Lindquist E, Pedraza MA, Barry K et al. 2014. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* **26**(1): 121-135.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**(5720): 385-389.

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M et al. 2010. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genetics* **42**(11): 1027-1030.

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N et al. 2013a. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat Genet* **45**(1): 43-50.

Li L, Petsch K, Shimizu R, Liu S, Xu WW, Ying K, Yu J, Scanlon MJ, Schnable PS, Timmermans MC et al. 2013b. Mendelian and non-Mendelian regulation of gene expression in maize. *PLoS genetics* **9**(1): e1003202.

Lu M, Xie CX, Li XH, Hao ZF, Li MS, Weng JF, Zhang DG, Bai L, Zhang SH. 2011. Mapping of quantitative trait loci for kernel row number in maize across seven environments. *Mol Breeding* **28**(2): 143-152.

Michelmore RW, Paran I, Kesseli RV. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc Natl Acad Sci U S A* **88**(21): 9828-9832.

Morton NE. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet* **7**(3): 277-318.

Peiffer JA, Flint-Garcia SA, Leon Nd, McMullen MD, Kaeppler SM, Buckler ES, de Leon N. 2013. The genetic architecture of maize stalk strength. *PLoS ONE* **8**(6): e67066.

Peiffer JA, Romay MC, Gore MA, Flint-Garcia SA, Zhang ZW, Millard MJ, Gardner CAC, McMullen MD, Holland JB, Bradbury PJ et al. 2014. The genetic architecture of maize height. *Genetics* **196**(4): 1337-1356.

Poland JA, Bradbury PJ, Buckler ES, Nelson RJ. 2011. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proceedings of the National Academy of Sciences of the United States of America* **108**(17): 6893-6898.

Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA et al. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome biology* **14**(6): R55.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science (Washington)* **326**(5956): 1112-1115.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**(7130): 881-885.

Tian F, Bradbury PJ, Brown PJ, Hung HY, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES. 2011. *Genome-wide association study of leaf architecture in the maize nested association mapping population*. Nature Publishing Group, New York.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* **90**(1): 7-24.

Zhou G, Chen Y, Yao W, Zhang C, Xie W, Hua J, Xing Y, Xiao J, Zhang Q. 2012. Genetic composition of yield heterosis in an elite rice hybrid. *Proc Natl Acad Sci U S A* **109**(39): 15847-15852.

# CHAPTER 2. IDENTIFICATION AND CROSS-VALIDATION OF GENETIC VARIANTS ASSOCIATED WITH THE KERNEL ROW NUMBER TRAIT OF MAIZE: A COMPARISON OF GWAS APPROACHES

See supplemental information in Appendix A

Jinliang Yang[1], Cheng-Ting "Eddy" Yeh[1], Rohan L. Fernando[2], Jack C.M. Dekkers[2], Dorian J. Garrick[2], Dan Nettleton[3], Patrick S. Schnable[1,4,*]

[1]Department of Agronomy, [2]Department of Animal Science and Center for Integrated Animal Genomics, [3]Department of Statistics, [4]Center for Plant Genomics.

[*]Author for correspondence

**Abstract**

Advances in next generation sequencing (NGS) technologies and the development of appropriate populations and statistical approaches enable genome-wide dissection of the genetic determinants of traits via genome-wide association studies (GWAS). Although multiple statistical approaches for conducting GWAS are available, there is not yet consensus about which approaches perform best. Kernel row number (KRN) trait data were collected from a set of 6,230 lines derived from four related maize populations. A set of ~13M variants (~2M of which were newly discovered) were projected and/or imputed onto the 6,230 lines. Three distinct approaches to GWAS were compared: 1) single-variant, 2) stepwise regression and 3) Bayesian-based multi-variant model fitting. In combination, these analyses identified associations in 764 100-kb chromosomal bins. A subset of these KRN-associated variants (KAVs) were subjected to cross-validation using three unrelated populations that were not included in the GWAS; approximately 50% of successfully genotyped KAVs were cross-validated in at least one unrelated population. Importantly, ~60% of cross-validated KAVs were identified by only one of the three statistical approaches. This finding demonstrates that the three GWAS approaches are complementary. These identified KAVs have the potential to enhance our understanding of maize domestication and the developmental steps involved in ear development.

## Introduction

Subsequent to the adoption of genome wide association studies (GWAS) (Klein et al. 2005), ~2,000 loci have been identified as being statistically associated with human disease and other quantitative traits (Visscher et al. 2012). Similarly, GWAS has been used to identify hundreds of loci associated with traits in crops such as maize (Brown et al. 2011; Tian et al. 2011), rice (Huang et al. 2010), sorghum (Morris et al. 2013) and barley (Cockram et al. 2010) and in non-crop models such as *Arabidopsis* (Atwell et al. 2010; Meijon et al. 2014).

There are multiple statistical approaches for conducting GWAS, including both single-variant and multi-variant approaches. There is not yet a consensus about which approach performs the best (Bush and Moore 2012). Single-variant analyses compare the phenotypic distributions of alternative genotypes at each polymorphic site independently. They can be conducted without correction for population structure or with correction using techniques such as genomic control (Devlin and Roeder 1999), principle component analysis (Price et al. 2006) or mixed linear models (Yu et al. 2006). Although single-variant analyses are most often used in published literature, they have a number of inherent limitations, such as not being able to distinguish among the contributions of closely linked loci (Yang et al. 2012), and they sometimes overcorrect for inflation caused by polygenic inheritance (Yang et al. 2011). In comparison, multi-variant approaches have already been demonstrated to be superior in classical linkage analyses, where for example, composite interval mapping outperforms simple interval mapping (Zeng 1993). Multi-variant approaches to GWAS can explicitly account for large effect loci and estimate their effects simultaneously. Similarly, it has been suggested that the power of GWAS may be improved by conditioning on major-effect loci (Kang et al. 2010). One challenge to using multi-variant approaches is, however, the substantial computational burden associated with analyzing a large number of polymorphic

sites. As a partial solution, stepwise regression, which selects markers based on forward inclusion and backward elimination, has been proposed (Segura et al. 2012). Because the order of marker inclusion has large effects on model fitting, a robust subsampling-based method was developed (Valdar et al. 2006). As an alternative to stepwise regression, multi-variant Bayesian-based approaches that were initially developed for genomic prediction by simultaneously fitting all genotyped loci across the genome (Meuwissen et al. 2001) have been used for GWAS (Fan et al. 2011; Fernando and Garrick 2013). These approaches fit a mixed model, where the effects of variants are treated as random, with prior assumptions regarding the distributions of their effects. To compare the effectiveness of different approaches for conducting GWAS, we analyzed a single data set using the single-variant, stepwise regression and the Bayesian-based multi-variant approaches.

GWAS is typically associated with high rates of false discovery (Visscher et al. 2012). In human studies, a second cohort is often used to cross-validate the most significant SNPs discovered in the first cohort, thereby cost effectively reducing the number of false discoveries (Sladek et al. 2007). To our knowledge GWAS experiments in plants have not been subjected to cross-validation. For many plant species, large and diverse collections of germplasm accessions, including wild relatives, landraces and breeding lines, have been collected and are available for analysis. Using these types of genetic resources, we conducted cross-validation of polymorphic variants detected by each of the three GWAS approaches. The results of these cross-validation studies provided an opportunity to evaluate the performance of the three approaches.

These analyses were conducted on the kernel row number (KRN) trait, which is both a component of yield and a model trait for genetic studies (Hallauer et al. 2010). It is highly heritable and exhibits little variation in response to environment (Lu et al. 2011). In addition, it is easily scored as an integer, and this scoring can be

conducted after completion of the busy pollination season. To map the genetic architecture controlling KRN, KRN trait data were repeatedly collected from 6,230 individuals in four related populations.

Collectively, the three GWAS approaches identified 231 putative KRN-associated variants (KAVs). A subset of these putative KAVs was subjected to cross-validation tests using three unrelated populations that were not included in the GWAS. Approximately 40% of cross-validated KAVs had been detected by two or three of the GWAS approaches, but ~60% of cross-validated KAVs were identified by only one of the three approaches. In addition, changes in allele frequencies of a subset of the 231 KAVs over evolutionary time and enrichment of KAV-linked genes in developmental pathways known to be relevant to the KRN trait provide evidence for the biological relevance of the identified KAVs.

## Results

### Phenotypic observations of KRN in four related populations

KRN trait data were collected from plants grown at at two locations over four years from 6,230 lines within four related GWAS populations. The first GWAS population consisted of the intermated B73 and Mo17 (IBM) (Lee et al. 2002) and nested association mapping (NAM) (Yu et al. 2008) RILs, which were developed from crosses of 25 inbreds by the common B73 inbred. The second and third GWAS populations were obtained by crossing a subset of the IBM and NAM RILs by B73 or by Mo17. The final GWAS population was a partial diallel of the 27 inbred founders of the IBM and NAM RILs. Additional KRN trait data were extracted from published data collected from NAM RILs grown in eight environments (Brown et al. 2011). The resulting KRN data were analyzed using a mixed model to estimate the phenotype of each of the 6,230 lines in the four GWAS populations (Table S1). In this combined analysis, KRN phenotype values ranged from 9.1 to 23.6, with a mean of 14.9 rows,

whereas the B73 inbred had an above average KRN phenotype of 17.1 rows. Density plots of the four GWAS populations exhibited the expected bell-shaped distributions (Figure 1A).

**KRN exhibits little heterosis or reciprocal effects**

One of the four GWAS populations (i.e., the IBM and NAM RILs) was developed by crossing the B73 inbred to each of 26 diverse inbreds and developing ~200 RILs per cross (Lee et al. 2002; McMullen et al. 2009). The median KRN phenotype computed for each of the 26 sets of RILs (Figure 1B) was significantly correlated ($r$ = 0.74, Pearson's correlation test $P$ value < 0.01) with the KRN of the corresponding non-B73 parent. As expected based on prior research (Srdic et al. 2007; Toledo et al. 2011), we observed little heterosis for KRN in the other three GWAS populations (Figure S1). Hence, we conclude that KRN is, in general, mostly controlled by additive gene effects rather than dominant gene effects or epistasis. Further, there is no evidence (Student's t-test, $P$ value = 0.59) for the existence of reciprocal effects for this trait based on comparisons of reciprocal crosses between B73 and Mo17 (Figure S2).

**Separate and joint QTL studies**

QTL linkage analyses were conducted using the KRN trait data and published genetic maps derived from analysis of the IBM (Liu et al. 2010) and NAM (Buckler et al. 2009) RILs. Separate QTL studies on the 26 individual subpopulations of bi-parental RILs identified a total of 146 QTLs (Figure 2A, Table S2), although many of these presumably represent the same QTLs detected in different subpopulations. For example, a significant QTL in the region chr4:189-237Mb was detected in 88% (23/26) of the subpopulations. In this QTL region, the B73 allele was favorable in all subpopulations for which the QTL was detected. The largest contrast of QTL effect at this locus occurs between B73 and NC350 alleles, wherein the B73 allele increases

KRN by 2.5 rows and accounts for 37% of phenotypic variation in this subpopulation. However, for some QTL regions the B73 allele is favorable in some subpopulations but unfavorable in other subpopulations (e.g., regions of chr1:6-50Mb, chr1:204-232Mb, chr5:12-84Mb and chr10:99-142Mb; Table S3). This phenomenon could be caused by tightly linked QTLs.

A joint QTL linkage analysis of the 25 NAM RIL subpopulations identified 28 QTLs (Figure 2B, Table S4). Consistent with the observation that the average KRN trait value of B73 is higher than the average KRN value, the favorable alleles of 79% (22/28) of identified QTL were provided by B73. In this joint analysis, the QTL detected on chromosome 4 could be resolved into three QTLs. Therefore, the large effects observed in some of the individual subpopulations may be caused by multiple linked QTLs all having the same direction of effects. To distinguish among these possible explanations, higher resolution mapping was required.

**Joint GWAS using three different statistical approaches**

Assuming enough markers are used, GWAS, which utilizes historical recombination events carried by the diverse founders of the IBM and NAM RILs would be expected to increase mapping resolution as compared to even joint QTL studies (Yu et al. 2008). Three sources of genotypic variants: maize HapMap1 (Gore et al. 2009), HapMap2 (Chia et al. 2012) and independently discovered RNA-seq derived variants (Barbazuk et al. 2007) (Li, Yeh and Schnable, unpublished data), were merged and filtered to form a set of ~13 million variants having a call rate of > 0.4 and a minor allele frequency (MAF) of > 0.1. These variants were imputed for three of the GWAS populations (IBM and NAM RILs, B73 x RILs and Mo17 x RILs) and projected onto the $F_1$ hybrids derived from the partial diallel of IBM and NAM founders.

Three GWAS approaches were used to identify KAVs. In each approach, population and subpopulation were included as fixed effects to account for inherent structure in the 6,230 lines included in the GWAS. First, a single-variant approach (Manolio 2010), was used to scan the ~13M variants one-by-one using QTL detected in the joint analysis as covariates. Using an arbitrary cutoff of $-\log_{10}(P) > 20$, this approach identified linked clusters of variants, most of which were located within the 28 QTL intervals that had been identified by the joint QTL analysis (Figure 3C and 3D). To diminish the over-representation of certain regions by significant variants, a thinning procedure was developed that resulted in the identification of 257 KAVs representing 192 100-kb bins (Table S5), which in combination accounted for 51% of the phenotypic variation.

Second, in an attempt to improve mapping resolution, a multi-variant stepwise regression approach was used, which automatically controls for background QTL effects. Using cutoff described in Materials and Methods, 300 variants representing 296 100-kb bins that covered 22 of the 28 QTL intervals detected in the joint analysis were identified; in combination, these variants accounted for 78% of phenotypic variation (Figure 3B, Table S5).

Third, a Bayesian-based approach (Fernando and Garrick 2013) was used to estimate effects of all ~13M variants simultaneously via a mixed model. After applying the variant thinning procedure and cutoffs described in Materials and Methods, a set of 442 variants representing 343 100-kb bins, which together accounted for 74% of the phenotypic variation, was identified (Figure 3A, Table S5). Most promisingly, this approach identified smaller chromosomal intervals than the single-variant approach.

In the separate linkage mapping analyses described above, four QTL regions were detected for which B73 alleles exhibited opposite effects in different subpopulations. After conducting GWAS with the three approaches, individual

variant effects were examined for the 135 KAVs located in these four regions. Two out of four of these QTL regions contained KAVs at which the B73 allele had either positive or negative effects (Table S6), suggesting that the improved resolution afforded by GWAS was better a distinguishing among tightly linked loci than the joint QTL analyses.

**Comparison of KAVs identified by the GWAS approaches**

In combination, the three GWAS approaches identified 764 100-kb bins (Table S5), each of which contained one or more significant variants. Encouragingly, among these 764 bins, 66 (containing 169 variants) were detected by at least two approaches (Figure 3). Only one of these bins was detected by all three approaches. That bin (chr4:229.0-Mb) overlaps the most significant QTL peak detected in the joint QTL study (Figure 2B). To estimate the upper bounds for false positive discovery for each approach and to determine whether the KAVs that were detected by more than one approach are more reliable, a set of 231 KAVs was selected for cross-validation testing. This set of KAVs (Figure 3, Table S7) included the 169 variants in the 66 bins detected by at least two approaches and 62 of the most significant one or two variants selected from 20 bins that had only been detected by one approach (approach-specific variants). Hence, in total KAVs from a total of 126 bins (66 + 20 × 3) were selected for cross-validation.

In combination, the 231 selected KAVs explained 64% of phenotypic variation. Individually, most of the KAVs (83%, 192/231) explained less than 5% of the phenotypic variation, but, 17% (39/231) of the KAVs individually accounted for more than 5% but less than 10% of phenotypic variation (Figure S4). As expected for the reasons described previously, the B73 variant-type was favorable for nearly three-quarters (73%, 168/231) of these KAVs. Other characterizations of these KAVs are presented in Figure S5. Consistent with our previous study (Li et al. 2012), KAVs are substantially enriched for variants located within genes or within

5-kb upstream of genes (Chi-square *P* value < 0.01) and enriched in variants discovered from the RNA-seq data (Chi-square *P* value < 0.01), relative to the ~13M variants used for GWAS. These observations emphasize the value of including genic variants derived from RNA-seq data as a complement to low pass whole genome sequencing (WGS) data such as maize HapMap2 variants.

**KAVs that differentiate maize from teosinte**

Maize was domesticated from its wild ancestor teosinte between 6,000 and 10,000 years ago (Matsuoka et al. 2002). A typical teosinte female inflorescence has two ranks of spikelets (each rank of spikelet is similar to two kernel rows). One of the most significant morphological changes that occurred during domestication was the development of ears that had four (or more) rows of kernels. In contrast to the substantial phenotypic change between teosinte and landraces, there has been little change in KRN values from landraces to the inbreds that were developed by public breeding programs (Figure 4A). If KAVs are indeed associated with the KRN phenotype, it would be reasonable to expect that the frequencies of the favorable alleles of at least some of the KAVs are higher in landraces than in teosinte, but not in improved lines as compared to landraces. To test this hypothesis, we analyzed a small set (N = 108) of teosinte, landrace and improved lines that had been genotyped with HapMap2 variants (Chia et al. 2012). The expected pattern was observed for 7 of the 152 KAVs that were derived from the HapMap2 variants at a false discovery rate (FDR) (Benjamini and Hochberg 1995) < 0.05 (Table S8). Each of these 7 variants exhibited dramatic increases in the frequency of the favorable allele in maize lines relative to teosinte. In addition, for the same reasons we would expect landraces to carry the favorable allele at more KAV loci than the teosinte lines. As shown in Figure 4B, on average, landraces (0.37) had a significantly higher (Monte Carlo simulation *P* value < 0.1) ratio of favorable:unfavorable alleles at KAV loci than did teosinte (0.33). In contrast, this ratio was identical in the improved

lines (0.37) compared to the landraces, which is consistent with the observation that KRN does not differ between landraces and improved lines (Figure 4A).

**Cross-validation of KAVs using three unrelated populations**

To distinguish true positive association signals from false positives, three cross-validation populations that were unrelated to the GWAS populations and to each other were genotyped with the KAVs. PCR-based genotyping-by-multiplexed-amplicon-sequencing (GBMAS) assays were designed for 140/231 (61%) KAVs (Wu, Liu and Schnable, unpublished). A total of 1,102 DNA samples from elite inbred lines (N = 208), extreme KRN accessions from the USDA germplasm collection (N = 606) and individuals from the Iowa Long Ear Synthetic (BSLE, N = 288) were individually genotyped by sequencing all multiplexed amplicons from all 1,102 samples in one HiSeq 2000 lane (Table S9). A variant calling pipeline was used to identify variants that were consistent with those detected in the GWAS analyses and that were segregating in at least one of the three cross-validation populations (Table S10-S12). Informative variants, defined as those which were successfully genotyped, were polymorphic, and had a call rate of $> 0.4$ and a MAF $> 0.05$ were used for cross-validation analyses.

The 208 elite inbred lines were phenotyped for the KRN trait (Table S9). Among these lines 70/140 (50%) of the KAVs were informative. To control for population structure, a set of SNPs that had previously been used to genotype a subset (N=91) of these lines was fitted (Nelson et al. 2008). Using this control, 22/70 (31%) of the informative KAVs could be cross-validated in the set of 91 elite inbreds with an FDR $< 0.05$. Because the elite inbreds are not closely related to the GWAS populations, it is unlikely that uncontrolled population structure could yield false-positive cross-validation assays for KAVs derived from the GWAS populations. Hence, we also conducted a naive analysis using the entire set of elite inbreds (N = 209) without controlling for population structure. In this analysis, 33/70 (47%) of the KAVs,

which included all of the 22 KAVs discussed above, could be cross-validated (Table S9, S10 and S13).

The USDA Plant Introduction station maintains a large collection of maize germplasm. 6,952 of their maize accessions have been phenotyped for the KRN trait. We selected the 225 accessions with the largest KRN values, the 208 accessions with the smallest KRN values, and 173 random accessions to serve as the second cross-validation population (Table S9**)**. The KRN phenotypes in this population are extreme, ranging from 16-30 rows in the high KRN pool to 4-12 rows in the low KRN pool. Because these accessions were maintained via random pollination within accessions, individual accessions are both heterogeneous and heterozygous. We therefore genotyped pools of DNA extracted from up to 12 plants per accession. A model fitted to the estimated allele frequencies was used to test the hypothesis that favorable KAV alleles have higher frequencies in the high KRN pools than in low KRN pools. Among the 56/131 (43%) informative variants, 14/56 (25%) could be cross-validated using the cutoffs described in Materials and Methods (Table S9, S11 and S13).

The BSLE population had been subjected to 30 generations of divergent selection for long ears (LE) and short ears (SE) (Hallauer 2005). During selection, KRN exhibited a negatively correlated response ($r$ = –0.6, Pearson's correlation test $P$ value < 0.05), i.e., longer and shorter ears had smaller and larger KRN trait values, respectively. Genotyping was conducted on the parental lines and bulked seeds from cycle 0 (C0), cycle 30 long ear (C30 LE) and cycle 30 short ear (C30 SE) populations. Of the 51 informative KAVs in the BSLE population, 7/51 (14%) showed significant differences in allele frequency between C30 LE and C30 SE populations using the cutoffs described in Materials and Methods. A simulation procedure that mimicked the selection program was conducted to test whether observed changes in allele frequency were larger than expected by genetic drift or stochastic sampling error.

After simulation, one validated KAV did not pass the cutoff (FDR < 0.05) and was removed. Hence, even after accounting for drift and stochastic sampling errors, 6/51 (12%) KAVs were deemed to have been under divergent selection (Table S9, S12 and S13). Collectively, these loci account for ~40% of the total between-population variance in KRN. Variants that are segregating in BSLE but not in GWAS populations or that were simply not detected as being KAVs in the GWAS populations may explain the remaining ~60% of variation between C30 LE and C30 SE.

In summary, 40/77 (52%) of informative KAVs, which represent 39 100-kb chromosomal bins were cross-validated in at least one population (Figure 3). The cross-validation results from the three GWAS approaches are illustrated in Figure 5. Considering all KAVs detected by each approach, the cross-validation rates were 61% (20/33) for the single-variant approach, 43% (6/14) for the stepwise regression approach and 45% (14/31) for the Bayesian-based approach. Cross validation rates were 67% (10/15) for KAVs detected only by the single-variant approach, 40% (6/15) for those detected only by stepwise regression, 35% (9/26) for KAVs detected only by the Bayesian-based approach, 76% (16/21) for KAVs detected by both single-variant and Bayesian-based approaches, and 11% (1/9) for control variants (Table S10-S13). Although both the regression and Bayesian approaches had lower cross-validation rates than the single variant approach, these results demonstrate that each of the three approaches identified cross-validated KAVs that were not identified by other approaches. Thus, the three GWAS approaches are complementary.

Informative genotyping data were also obtained for 34 KAVs reported in an earlier GWAS (Brown et al. 2011). Using the statistical analyses described above, 26% (9/34) of these KAVs could be cross-validated in at least one of the three unrelated populations (Table S10-S13).

**Functional analyses identified candidate genes within KAV-associated chromosomal bins**

A set of 2,690 KAV-linked genes was defined as those gene models (FGSv2.5b) falling into 500-kb regions flanking the 231 selected KAVs. KRN is a consequence of inflorescence branching. Thus, KRN is determined by the fates and identities of an array of meristems (Barazesh and McSteen 2008). Traditional genetic analyses have identified genes or metabolic pathways relevant to inflorescence development. For instance, *fasciated ear2* (*fea2*) (Bommert et al. 2013) and *thick tassel dwarf1* (*td1*) (Bommert et al. 2005) cause fasciated ear phenotypes with irregular but higher KRNs. *Fea2* and *td1* are homologs of *CLV1* and *CLV*2 of *Arabidopsis,* both of which belong to the CLV-WUS regulatory pathway that promotes stem cell differentiation (Clark 2001). In addition, genes involved in various stages of grass inflorescence development have been identified, including 1) auxin and 2) cytokinin signal transduction (Barazesh and McSteen 2008; Sigmon and Vollbrecht 2010), 3) ramose genes (Bortiri et al. 2006) and 4) other ungrouped genes (McSteen and Hake 2001; Upadyayula et al. 2006; Xu et al. 2011). In total, ~200 evidence-supported genes or their maize homologs were mapped onto maize gene models (FGSv2.5b). A Monte Carlo simulation test indicated that the KAV-linked genes are over-represented (12 genes overlapped, $P$ value < 0.01) among this set of evidence-supported genes. Furthermore, the KAV-linked genes are significantly enriched in members of two metabolic pathways, auxin (10 genes) and cytokinin (2 genes) signal transduction, both of which were known to be involved in inflorescence development (Table S14).

## Discussion

GWAS facilitates dissection of the genetic control of traits. Unfortunately, GWAS findings are often associated with high rates of false discovery (Visscher et al. 2012). With immortalized genotypes and replicated observations, GWAS in plants have the possibility to better control for stochastic factors, such as environmental effects, that

could affect the rate of false discovery in studies conducted on humans or some other species. However, we are not aware of any study that has determined the FDR of a plant-based GWAS.

This study compared the FDRs of three statistical approaches for identifying associations between genetic variants and the KRN trait. Each approach has strengths and weaknesses. Although the single-variant approach (Balding 2006) can control for the effects of other QTL (by treating them as covariates (Kang et al. 2010)), it typically detects linked clusters of trait-associated variants and therefore has difficulty to distinguish tightly linked QTLs. Although the stepwise regression approach (Segura et al. 2012) can identify variants by controlling background effects using a multi-variant model, it can only identify a small set of such variants. Although the Bayesian-based multi-variant approach (Fernando and Garrick 2013) automatically controls for population structure and background QTLs and generates various posterior distributions that can be used for inference, it does not provide formal significance cutoffs.

Cross-validation strategies that exploit the extensive genetic resources of maize were used to estimate maximum rates of false discovery. Overall, at least 52% (40/77) of KAVs could be cross-validated in at least one of three unrelated populations, indicated that the FDR is less than 48%. The true FDR is likely to be less because KAVs could fail to cross-validate in unrelated populations for a variety of reasons, including biological differences in the genetic control of the KRN trait among populations and Type II errors in the cross-validation analyses. Because KRN is mainly controlled by additive effect loci, traits controlled by different modes of inheritance may yield different cross-validation results.

Although the single-variant approach had somewhat higher cross-validation rates than the other two approaches (possibly at least partly because of analytic similarities between the single-variant approach and the cross-validation

experiments), each approach identified cross-validated KAVs that were not detected by the others. Hence, the use of multiple approaches or the development of a statistical method that combines their advantages, promises to enhance the power of GWAS.

The cross-validation rate of KAVs identified in this experiment (40/77 = 53%) is higher than KAVs identified in an earlier KRN GWAS (9/34 = 26%) (Brown et al. 2011). The improved power of our study (which made use of data from Brown et al. 2011, as well as additional data generated as part of our study) could also be due to the inclusion of more genotypes, more phenotypic data and higher marker density. The use of three complementary approaches for identifying KAVs may also have contributed to the higher cross-validation rate.

The cob was an evolutionary innovation that arose during the domestication of maize from teosinte (Goodman 1988). Based on the results of classical genetic experiments conducted in the 1930s, Beadle hypothesized that five major loci differentiate ear morphology traits (including the actual existence of a cob or "ear") of maize and its wild ancestor teosinte (Doebley 2004). Over the last few years, several of these major effect loci whose effects were observed by Beadle have been cloned via transposon tagging and chromosome walking (Dorweiler et al. 1993; Doebley et al. 1997; Wills et al. 2013). Beadle also hypothesized the existence of many small effect genes that affect ear morphology traits such as KRN (Balding 2006). It would be difficult to clone these genes via the approaches used to identify large effect loci. In contrast, the GWAS approaches used in the current study offer access to small effect loci, some of which may have been important during domestication of maize. Indeed, 7/231 KAVs exhibit statistically significant changes in the frequencies of the favorable allele between teosinte and landraces.

In conclusion, this study identified hundreds of KAVs that in combination explain 64% of phenotypic variation for KRN in lines that sample ~60% of the genetic

diversity of maize (Liu et al. 2003). Over 50% of KAVs that were tested could be cross-validated. The KAVs detected in this study can be used to facilitate marker-assisted breeding or transgenic approaches for crop improvement. Further in-depth analyses of KAV-linked genes will enable us to better understand the molecular and developmental processes that control variation in the KRN trait and may eventually be useful in breaking the negative correlation between KRN and ear length (Hallauer et al. 2004), thereby increasing grain yields.

## Materials and Methods

**KRN phenotyping.** KRN phenotypes were collected from several related populations, including recombinant inbred lines (RILs) of intermated B73 and Mo17 (IBM, N = 325 RILs) (Lee et al. 2002) and the nested association mapping (NAM, N = 4,699 RILs) (Yu et al. 2008) populations, a subset of the RILs that were backcrossed to the inbred line B73 (B73 x RILs, N = 692 BC1 lines), a subset of the RILs that were backcrossed to the inbred line Mo17 (Mo17 x RILs, N = 289 BC1 lines) and a partial diallel of the 26 NAM founders plus Mo17 (N = 225 F1 hybrids). Because reciprocal crosses were not considered and some of the crosses were not successful, the diallel population was both a partial and incomplete (225/351 = 64%) diallel. For statistical analyses, the IBM RILs were treated as a subpopulation of the NAM RILs.

During the years 2008-2011, subsets of the above populations were planted in replicated field trials in up to three fields in Ames, IA (summer season) and one field in Molokai, HI (winter season). There were 5-12 plants of the same line grown within each row. KRN counts were collected from mature ears. Phenotypic values were estimated for each line using a mixed linear model implemented in R (R Development Core Team 2010), with fixed effects for lines and random effects for locations, years, plots and blocks. Phenotypic density distributions in this study were estimated and plotted using R with default smoothing parameters.

**QTL analyses.** A two-step composite interval mapping (CIM) (Zeng 1993) method was employed using a suite of programs within QTL cartographer (Silva Lda et al. 2012). First, an automatic stepwise regression procedure was used to sequentially test all SNP markers; the most significant marker (inclusion threshold = 0.05) was kept after each iteration. This procedure was repeated until none of the added SNPs improved the model. In the second step, linkage analyses were conducted at 1-Mb intervals along the chromosome treating previously selected SNPs (other than those within the 1-Mb interval under analysis) as co-variants. A significance threshold was determined by conducting 1,000 permutations and support intervals were defined using a 1.5-LOD drop from QTL peak (Lander and Botstein 1994).

**Variant processing.** A set of 6.2 million genic variants (SNPs and small Indels) was identified via analysis of RNA-seq data from five tissues (shoot apical meristem, ear, tassel, shoot and root; Li, Yeh, and Schnable, unpublished data) on 26 NAM founder lines and Mo17. Another two sets of variants generated from the maize HapMap project were extracted from the Panzea database (www.panzea.org). These three sets of variants were merged using the consensus mode of PLINK (Purcell et al. 2007). The merged variants were further filtered by discarding variants with a call rate of < 0.4 and a MAF of < 0.1 across genotypes. The finalized set consists of 12,966,279 variants on NAM founders, which were used for imputation or projection onto the four related populations.

Genotyping scores for ~1,000 tagging SNPs that had been directly genotyped on the ~5,000 NAM RILs were obtained from the Panzea database. Based on these tagging SNPs and known pedigree information, the ~13 million variants discovered in the NAM founders were imputed onto NAM RILs using customized Perl scripts based on the method of Yu *et al.* (Yu et al. 2008). Because B73 x RIL, Mo17 x RIL and partial diallel populations were composed of hybrids of two known haplotypes, their genotypic data were directly projected from their known parents.

**Three statistical approaches for conducting GWAS.** *Single-variant model***.** Data from the four populations discussed above were used for GWAS. To account for documented stratification effects, the statistical model included fixed effects for population and subpopulation. Additional fixed effects were fitted in the model to control for effects of QTLs on other chromosomes, while all variants on a single chromosome were scanned, resulting in the following model for the *k*th variant:

$$Y_l = u_k + \sum_{i=1}^{4} a_{ik} P_{il} + \sum_{j=1}^{26} b_{jk} S_{jl} + \sum_{m \in Ch(-k)} c_{km} Q_{ml} + d_k VAR_{kl} + e_{kl}$$

where $Y_l$ is the adjusted KRN phenotypic value for line *l* from the mixed linear model analysis; $u_k$ is an intercept parameter; $P_{il}$ is 1 if line *l* is of GWAS population *i* and is 0 otherwise, and $a_{ik}$ is the effect of the *i*th population in the model for variant *k*; $S_{jl}$ is 1 if line *l* is from subpopulation *j* and 0 otherwise, $b_{jk}$ is the effect of subpopulation *j* in the model for variant *k*; $Q_{ml}$ indicates the line *l* genotype of the *m*th QTL detected by the joint linkage analyses, $c_{km}$ is the effect of the *m*th QTL in the model for variant *k*, *Ch(-k)* is the set of QTLs detected by the joint linkage analysis that lie on chromosomes other than the chromosome of variant *k*; $VAR_{kl}$ indicates the genotype of the *k*th variant in line *l*, $d_k$ is the effect of the *k*th variant; and $e_{kl}$ is an error term. This single-variant model was implemented using SNPTEST v2.3.0 (Marchini and Howie 2010).

*Stepwise regression model.* In the stepwise regression test, population and subpopulation effects were fitted first, and then marker effects were added to the model based on their *P* values computed from the marginal F-test. Using this automatic model selection procedure, a maximum of 300 variants was selected using a *P* value cutoff of 0.05.

*Bayesian-based multi-variant model.* A Bayesian-based multi-variant model was constructed using the BayesC option of GenSel v4.1 (Fernando and Garrick 2013).

This model differs from the single-variant model in that it estimates the effects of all variants simultaneously rather than testing them one-at-a-time. Because biases could be introduced by population stratification, known population and subpopulation factors were included in the model as fixed effects. The effects of the variants were fitted as random effects. The following mixed model used was:

$$Y_l = u + \sum_{i=1}^{4} a_i P_{il} + \sum_{j=1}^{26} b_j S_{jl} + \sum_{k}^{\sim 13M} c_k VAR_{kl} + e_l$$

where $VAR_{kl}$ indicates the genotype of the $k$th variant in line $l$ and $c_k$ is the effect of the $k$th variant; other terms in the model are as described in the single-variant model except that neither the $u$, $a_i$, or $b_j$ parameters nor the $e_l$ error terms are specific to the $k$th variant in the multi-variant model.

The BayesC option of GenSel requires that the fraction of markers having no effect (π) be inputted as a prior. In the test runs, several π values (0.9995, 0.9999 and 0.99995) were tried and similar posterior genetic variations accounted for by the markers were observed. In this study, π was set as 0.9999 (i.e., 1 – π, the number of markers with effects, was assumed to be ~1,300). Other prior information such as residual and genotypic variances was estimated using a testing run consisting of 1,000 iterations. The estimated variances were used to seed their respective priors for full training with a chain length of 41,000; the first 1,000 iterations were discarded as a burn-in. The posterior model frequency of a variant, which is the proportion of draws in which that variant was included in the model, was used *in lieu* of a traditional measurement of significance.

**Variant thinning procedure.** A variant thinning procedure was developed to select the most significant variants and to avoid concentration of selected variants in certain regions. For variants located in the 28 QTL intervals from the joint analysis and their 1-Mb flanking regions, the top 10 most significant variants were selected.

For variants located in other regions, significant variants were determined by the following arbitrary thresholds: $-\log_{10}(P) > 20$ for the single-variant approach, posterior model frequency (MF) > 0.02 for the Bayesian-based approach and an inclusion *P* value < 0.05 for the stepwise regression. These significant variants were clustered as groups if none of their pair-wise physical distances exceeded10-Mb. From these clustered groups, no more than 10 most significant variants were selected.

**Monte Carlo simulation**. A Monte Carlo simulation procedure was employed to rule out the possibility that population structure was responsible for observed differences in ratios of favorable:unfavorable alleles of KAVs in the teosintes, landraces and improved lines. The 231 KAVs mapped to 123 100-kb chromosomal bins. To conduct a single run of the Monte Carlo simulation, 231 variants (from the ~13M variants genome wide) were randomly selected from a randomly selected set of 123 chromosomal bins. These variants were mapped to HapMap2 and used to compute the ratio of favorable:unfavorable allele in each population. After randomly assigning one variant type as being the favorable allele, the difference in the average ratios between the two populations was recorded as a test statistic. After 1,000 runs, a *P* value was calculated as the proportion of random statistics exceeding the observed statistics.

**Cross-validation populations.** *Elite inbred lines***.** A total of 220 elite inbred lines, commercial lines that had formerly been subject to IP (Intellectual Property) protection via the plant variation protection act, were obtained from the USDA Plant Introduction Station in Ames, IA (http://www.ars.usda.gov/main/site_main.htm?modecode=36-25-12-00). These lines were planted in three randomized field trials and observed for KRN phenotypes. DNA was isolated from seedling tissues and used to conduct GBMAS (a PCR-based method that exploits Next Generation Sequencing to provide rapid and

cost-effective genotyping results (Wu, Liu and Schnable, unpublished)). After sequence trimming, barcode sorting, and alignment to the reference genome, polymorphic variants were discovered using a variant calling pipeline we developed previously (Li et al. 2012).

The following statistical model was used to test the hypothesis that the favorable KAVs were associated with high KRN in the elite inbred lines:

$$Y_l = u_k + \sum_{i=1}^{3} a_{ik} PC_{il} + d_k VAR_{kl} + e_{kl}$$

where $Y_l$ is the KRN phenotypic value from the mixed linear model analysis; $u_k$ is an intercept parameter; $PC_{il}$ designates the principle components $i$ for line $l$ derived from a random set of SNPs to account for population structure, and $a_{ik}$ is the effect of the $i$th principle component for variant $k$; $VAR_{kl}$ indicates the genotype of the $k$th variant in line $l$ and $d_k$ its effect; and $e_{kl}$ is the residual error. The R add-on package GenABEL (Aulchenko et al. 2007a) was used to conduct the analysis. Significant variants were determined using an false discovery rate (FDR) (Benjamini and Hochberg 1995) cutoff of < 0.05. In addition, the directions of variant effects were compared with the direction of effects for each KAV in the GWAS populations. Variants with conflicting effects were discarded.

*Extreme KRN USDA accessions*. The germplasm resources information network (GRIN) database (http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl) of the USDA contains KRN records of ~7,000 accessions, from which the 225 lines with highest KRN values, the 208 lines with lowest KRN values and 173 random lines were obtained (Table S14). Because of the genetic heterozygosity of the obtained accessions, up to 12 random seeds were germinated and pooled together for each accession for DNA isolation. After genotyping by GBMAS, variants were discovered using an approach that allowed for the calling of heterozygous variants.

The following model was used to test whether the alleles at the *k*th variant are associated with KRN in the selected USDA accessions:

$$Y_l = u_k + \beta_k F_{kl} + e_{kl}$$

where $Y_g$ is the KRN phenotype for line *l*; $u_k$ is an intercept parameter; $F_{kl}$ is the number of favorable alleles at variant *k* in line *l*, $\beta_k$ is the additive effect of the favorable allele at variant *k*; and $e_{kg}$ is the residual error. Significant variants were determined using an FDR cutoff of < 0.05 and variants with conflicting effects compared with KAVs in the GWAS populations were discarded.

*Iowa Long Ear Synthetic (BSLE).* The BSLE population was the product of a long-term selection project conducted by Arnel Hallauer and his colleagues at Iowa State University, whose goal was to divergently select long and short ears from a single founder population (Hallauer et al. 2004; Hallauer 2005). Parental lines of BSLE and bulked seeds from cycle 0 (C0), cycle 30 short ear (C30 SE) and cycle 30 long ear (C30 LE) were obtained from Arnel Hallauer. DNA was isolated individually from seedling tissues of these obtained materials (N = 60 for C0, N = 101 for C30 SE and N = 96 for C30 LE). After genotyping by GBMAS, variants were called as described above. Population allele frequencies were estimated based on the surveyed samples.

The 'qtscore' function of GenABEL (Aulchenko et al. 2007b) was used to conduct a score test of association between a C30 population indicator variable (0 for C30 LE, 1 for C30 SE) and genotype of the *k*th variant. Significant variants were determined using an FDR cutoff of < 0.05 and variants with conflicting effects compared with KAVs in the GWAS populations were discarded.

To rule out the possibility that detected differences in allele frequency were due to genetic drift, a procedure that simulated the selection process without considering directionality was implemented. The simulation was started with the

initial variant allele frequencies from C0; and the same number of alleles (N = 60) was randomly sampled without replacement from the same sampling space (N = 800) as the selection program proceeded. After each cycle of re-sampling, variant allele frequencies were updated. The re-sampling process was conducted for 30 cycles to mimic the 30 generations of selections in the real selection program. The above procedure was repeated 10,000 times and the *P* value was calculated as the probability of the difference between the observed variant allele frequencies in the two subpopulations being larger than the values obtained from the simulation. The *P* values were adjusted using the FDR method to correct for multiple testing.

**Functional analyses of KAV-linked genes.** Auxin and cytokinin biosynthesis and signal transduction related genes, as well as *CLV-WUS* related genes were extracted from The *Arabidopsis* Information Resource (TAIR) database (Poole 2007). Genes related to auxin and cytokinin hormone biosynthesis and signal transduction pathways in maize and rice were downloaded from KEGG database (Kanehisa 2002). In addition, various genes involved in inflorescence development (Barazesh and McSteen 2008) were manually extracted from the literature, including *ramosa* genes (Bortiri et al. 2006) and others (McSteen and Hake 2001; Upadyayula et al. 2006; Xu et al. 2011). These sets of evidence supported genes were blasted against the filtered gene set (FGS_5b) on B73 reference genome (RefGen_v2) with coverage > 50% and identity > 50%. KAV-linked genes were defined as genes located in the 500-kb flanking regions of the identified KAVs.

## Acknowledgments

We gratefully acknowledge Dr. Arnel Hallauer, Dr. Kendall Lamkey and Mr. Paul White of Iowa State University and Dr. Candice Gardner of the USDA's North Central Regional Plant Introduction Station (NCRPIS) for sharing genetic stocks, Drs. Sanzhen Liu (currently Kansas State University) and Mr. Ed Allen (Monsanto) for useful discussions, Dr. Wei Wu (currently LGC Genomics), Dr. Haiying Jiang

(currently Shenyang Agricultural University), Dr. Li Li (currently Northwest Agriculture and Forestry University), Ms. Uyen Pham and Ms. Talissa Sari for technical support, and Ms. Lisa Coffey for the generation and maintenance of genetic stocks. This research was supported by grants from Monsanto and the National Science Foundation (IOS-1027527) to PSS.

## References

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT et al. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**(7298): 627-631.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. 2007a. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**(10): 1294-1296.

-. 2007b. GenABEL: an R library for genorne-wide association analysis. *Bioinformatics* **23**(10): 1294-1296.

Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**(10): 781-791.

Barazesh S, McSteen P. 2008. Hormonal control of grass inflorescence development. *Trends Plant Sci* **13**(12): 656-662.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. 2007. SNP discovery via 454 transcriptome sequencing. *Plant Journal* **51**(5): 910-918.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**(1): 289-300.

Bommert P, Lunde C, Nardmann J, Vollbrecht E, Running M, Jackson D, Hake S, Werr W. 2005. thick tassel dwarf1 encodes a putative maize ortholog of the Arabidopsis CLAVATA1 leucine-rich repeat receptor-like kinase. *Development* **132**(6): 1235-1245.

Bommert P, Nagasawa NS, Jackson D. 2013. Quantitative variation in maize kernel row number is controlled by the FASCIATED EAR2 locus. *Nature genetics* **45**(3): 334-337.

Bortiri E, Chuck G, Vollbrecht E, Rocheford T, Martienssen R, Hake S. 2006. ramosa2 encodes a LATERAL ORGAN BOUNDARY domain protein that determines the fate of stem cells in branch meristems of maize. *Plant Cell* **18**(3): 574-585.

Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S, Holland JB, Flint-Garcia S, McMullen MD, Buckler ES et al. 2011. Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS genetics* **7**(11): e1002383.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC et al. 2009. The genetic architecture of maize flowering time. *Science* **325**(5941): 714-718.

Bush WS, Moore JH. 2012. Chapter 11: Genome-wide association studies. *PLoS computational biology* **8**(12): e1002822.

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nature genetics* **44**(7): 803-807.

Clark SE. 2001. Cell signalling at the shoot meristem. *Nat Rev Mol Cell Biol* **2**(4): 276-284.

Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, Luo Z, Kearsey MJ, Werner P, Harrap D et al. 2010. Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proc Natl Acad Sci U S A* **107**(50): 21611-21616.

Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* **55**(4): 997-1004.

Doebley J. 2004. The genetics of maize evolution. *Annu Rev Genet* **38**: 37-59.

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

Dorweiler J, Stec A, Kermicle J, Doebley J. 1993. Teosinte glume architecture 1: A Genetic Locus Controlling a Key Step in Maize Evolution. *Science* **262**(5131): 233-235.

Fan B, Onteru SK, Du ZQ, Garrick DJ, Stalder KJ, Rothschild MF. 2011. Genome-Wide Association Study Identifies Loci for Body Composition and Structural Soundness Traits in Pigs. *Plos One* **6**(2).

Fernando RL, Garrick D. 2013. Bayesian methods applied to GWAS. *Methods in molecular biology* **1019**: 237-274.

Goodman MM. 1988. The history and evolution of maize. *CRC Critical Reviews in Plant Sciences* **7**(3): 197-220.

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J et al. 2009. A first-generation haplotype map of maize. *Science* **326**(5956): 1115-1117.

Hallauer AR. 2005. Registration of BSLE(M-S)C30 and BSLE(M-L)C30 maize germplasm. *Crop Sci* **45**(5): 2132.

Hallauer AR, Carena MJ, Filho JBM. 2010. Quantitative Genetics in Maize Breeding. *Handb Plant Breed* **6**: 1-663.

Hallauer AR, Ross AJ, Lee M. 2004. Long-term divergent selection for ear length in maize. *Plant Breeding Reviews* **24**(2): 153-168.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z et al. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics* **42**(11): 961-967.

Kanehisa M. 2002. The KEGG database. *Novartis Foundation symposium* **247**: 91-101; discussion 101-103, 119-128, 244-152.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**(4): 348-354.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**(5720): 385-389.

Lander ES, Botstein D. 1994. Mapping Mendelian Factors Underlying Quantitative Traits Using Rflp Linkage Maps (Vol 121, Pg 185, 1989). *Genetics* **136**(2): 705-705.

Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, Hallauer A. 2002. Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant molecular biology* **48**(5-6): 453-461.

Li X, Zhu C, Yeh CT, Wu W, Takacs EM, Petsch KA, Tian F, Bai G, Buckler ES, Muehlbauer GJ et al. 2012. Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genome Res* **22**(12): 2436-2444.

Liu KJ, Goodman M, Muse S, Smith JS, Buckler E, Doebley J. 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* **165**(4): 2117-2128.

Liu S, Chen HD, Makarevitch I, Shirmer R, Emrich SJ, Dietrich CR, Barbazuk WB, Springer NM, Schnable PS. 2010. High-throughput genetic mapping of mutants via quantitative single nucleotide polymorphism typing. *Genetics* **184**(1): 19-26.

Lu M, Xie CX, Li XH, Hao ZF, Li MS, Weng JF, Zhang DG, Bai L, Zhang SH. 2011. Mapping of quantitative trait loci for kernel row number in maize across seven environments. *Mol Breeding* **28**(2): 143-152.

Manolio TA. 2010. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363**(2): 166-176.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nature reviews Genetics* **11**(7): 499-511.

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J. 2002. A single domestication for maize shown by multilocus microsatellite genotyping. *Proc Natl Acad Sci U S A* **99**(9): 6080-6084.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C et al. 2009. Genetic properties of the maize nested association mapping population. *Science* **325**(5941): 737-740.

McSteen P, Hake S. 2001. barren inflorescence2 regulates axillary meristem development in the maize inflorescence. *Development* **128**(15): 2881-2891.

Meijon M, Satbhai SB, Tsuchimatsu T, Busch W. 2014. Genome-wide association study using cellular traits identifies a new regulator of root development in Arabidopsis. *Nature genetics* **46**(1): 77-81.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4): 1819-1829.

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE et al. 2013. Population genomic and

genome-wide association studies of agroclimatic traits in sorghum. *Proc Natl Acad Sci U S A* **110**(2): 453-458.

Nelson PT, Coles ND, Holland JB, Bubeck DM, Smith S, Goodman MM. 2008. Molecular characterization of maize inbreds with expired U S Plant Variety Protection. *Crop Sci* **48**(5): 1673-1685.

Poole RL. 2007. The TAIR database. *Methods in molecular biology* **406**: 179-212.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**(8): 904-909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3): 559-575.

R Development Core Team. 2010. R: A language and environment for statistical computing.

Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q, Nordborg M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics* **44**(7): 825-830.

Sigmon B, Vollbrecht E. 2010. Evidence of selection at the ramosa1locus during maize domestication. *Molecular Ecology* **19**(7): 1296-1311.

Silva Lda C, Wang S, Zeng ZB. 2012. Composite interval mapping and multiple interval mapping: procedures and guidelines for using Windows QTL Cartographer. *Methods in molecular biology* **871**: 75-119.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**(7130): 881-885.

Srdic J, Pajic Z, Drinic-Mladenovic S. 2007. Inheritance of maize grain yield components. *Maydica* **52**(3): 261-264.

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES. 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics* **43**(2): 159-162.

Toledo FH, Ramalho MA, Abreu GB, de Souza JC. 2011. Inheritance of kernel row number, a multicategorical threshold trait of maize ears. *Genetics and molecular research : GMR* **10**(3): 2133-2139.

Upadyayula N, da Silva HS, Bohn MO, Rocheford TR. 2006. Genetic and QTL analysis of maize tassel and ear inflorescence architecture. *Theor Appl Genet* **112**(4): 592-606.

Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J. 2006. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics* **38**(8): 879-887.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* **90**(1): 7-24.

Wills DM, Whipple CJ, Takuno S, Kursel LE, Shannon LM, Ross-Ibarra J, Doebley JF. 2013. From many, one: genetic control of prolificacy during maize domestication. *PLoS genetics* **9**(6): e1003604.

Xu XM, Wang J, Xuan Z, Goldshmidt A, Borrill PG, Hariharan N, Kim JY, Jackson D. 2011. Chaperonins facilitate KNOTTED1 cell-to-cell trafficking and stem cell function. *Science* **333**(6046): 1141-1144.

Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ et al. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics* **44**(4): 369-375, S361-363.

Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'Connell JR, Mangino M et al. 2011. Genomic inflation factors under polygenic inheritance. *European journal of human genetics : EJHG* **19**(7): 807-812.

Yu J, Holland JB, McMullen MD, Buckler ES. 2008. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**(1): 539-551.

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38**(2): 203-208.

Zeng ZB. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc Natl Acad Sci U S A* **90**(23): 10972-10976.

## Figure Legends

**Figure 1. Phenotypic distribution of the KRN trait.** In panel (A), density plots of the four cross type populations. In panel (B), boxplots of the 26 NAM RIL subpopulations. Blue and red dashed lines indicate the mean phenotypic values of B73 (KRN=17.1) and Mo17 (KRN=10.8), respectively.

**Figure 2. Plots of separate and joint QTL linkage analyses.** In panel (A), physical positions of the significant QTLs and their effects by separate analyses on the 26 RIL subpopulations. Blue color indicates the favorable QTL allele is derived from B73 and red color indicates that the non-B73 allele is favorable. In panel (B), joint QTL results using the 25 NAM RIL subpopulations. The red dashed line denotes the significant threshold determined by 1,000 permutations. Under the curve, QTL confidence intervals were plotted using black solid lines.

**Figure 3. Stacked Manhattan plots of joint QTL and GWAS results.** From upper to lower panels are results from the Bayesian-based (A), stepwise regression (B) and single-variant (C) approaches for GWAS and joint QTL mapping (D), respectively. The red dashed line in the QTL plot indicates the 1,000 permutation threshold and black lines show the QTL confidence intervals. Red squares in panel (A), triangles in panel (B) and circles in panel (C) indicate the KAVs selected for further cross-validation.

**Figure 4. KRN phenotype distribution in maize landraces and improved lines and percentage of KAVs that are favorable in three evolutionary groups of Zea.** In panel (A), the mean KRN values for landraces (KRN=13.5) and improved lines (KRN=14.3) are indicated by vertical dashed lines. In panel (B), percentage of favorable KAVs for teosinte lines is significantly different (*P* value < 0.1) than the average value of landraces. However, there is no evidence (*P* value = 0.7) of a difference between landraces and improved lines.

**Figure 5. Cross-validation for KAVs identified from the three different GWAS approaches.** Transformed single-variant *P* values and Bayesian-based posterior model frequencies were extracted and plotted for all the 77 informative KAVs identified by at least one of the three GWAS approaches. KAVs detected only by the single-variant approach are plotted in the lower right quadrant, KAVs detected only by the stepwise regression approach are plotted as non-grey dots in the lower left quadrant, the KAVs detected only by the Bayesian-based approach are plotted in the upper left quadrant, KAVs detected by both the single-variant and Bayesian-based approaches are plotted in the upper right quadrant and control variants are plotted as grey dots in the lower left quadrant. Cross-validated KAVs are marked in red.

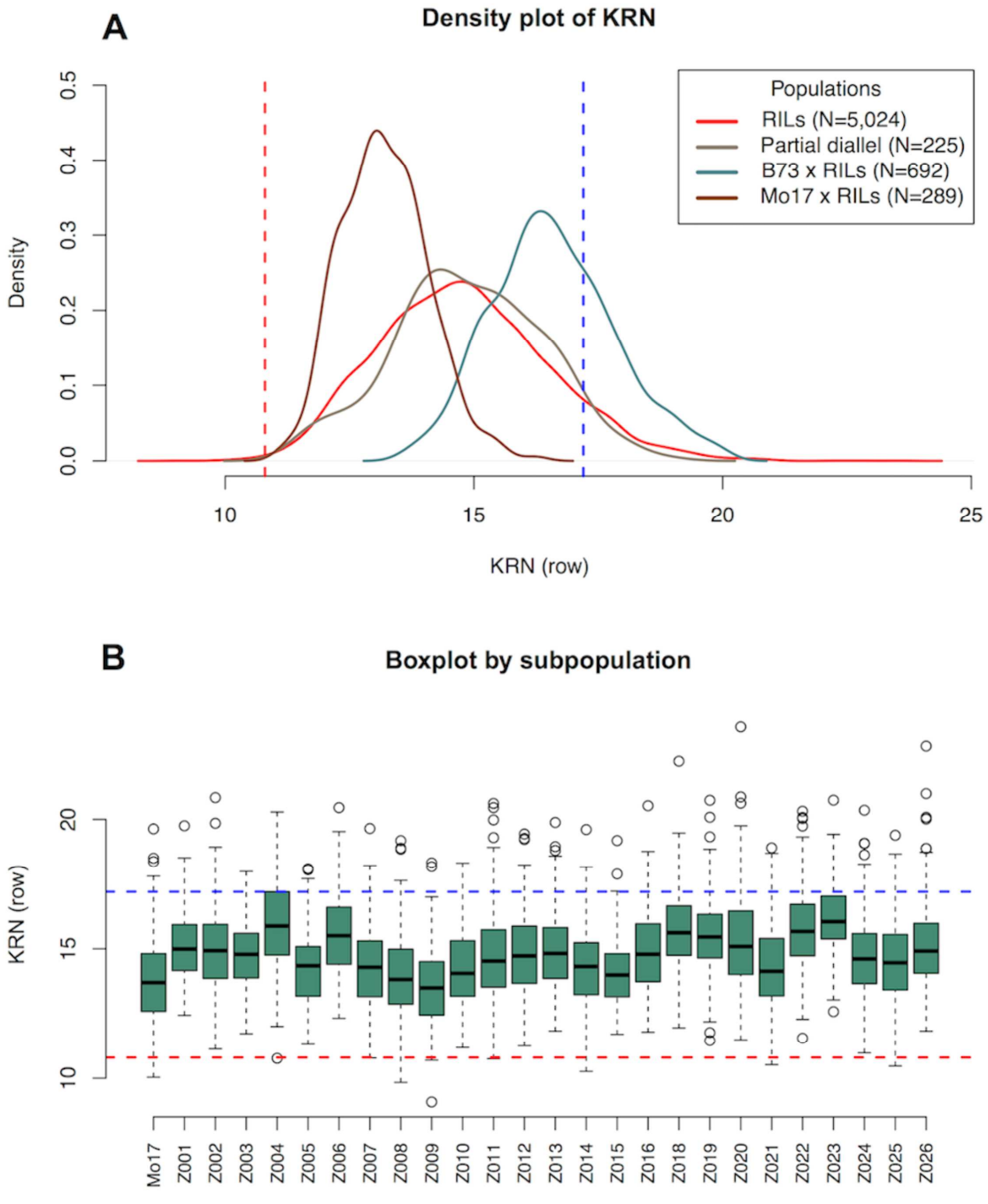
A
Density plot of KRN
Populations
RILs (N=5,024)
Partial diallel (N=225)
B73 x RILs (N=692)
Mo17 x RILs (N=289)
Density
0.0
0.1
0.2
0.3
0.4
0.5
10
15
20
25
KRN (row)
B
Boxplot by subpopulation
KRN (row)
10
15
20
Mo17
Z001
Z002
Z003
Z004
Z005
Z006
Z007
Z008
Z009
Z010
Z011
Z012
Z013
Z014
Z015
Z016
Z018
Z019
Z020
Z021
Z022
Z023
Z024
Z025
Z026

**Figure 1**

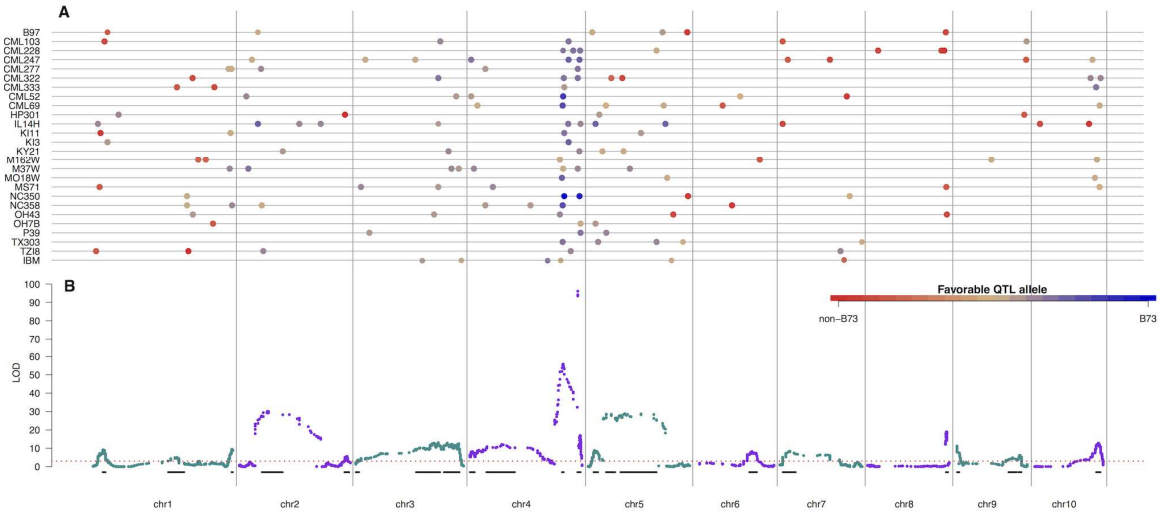
A
B97
CML103
CML228
CML247
CML277
CML322
CML333
CML52
CML69
HP301
IL14H
KI11
KI3
KY21
M162W
M37W
MO18W
MS71
NC350
NC358
OH43
OH7B
P39
TX303
TZI8
IBM
B
LOD
100
90
80
70
60
50
40
30
20
10
0
Favorable QTL allele
non−B73
B73
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10

**Figure 2**

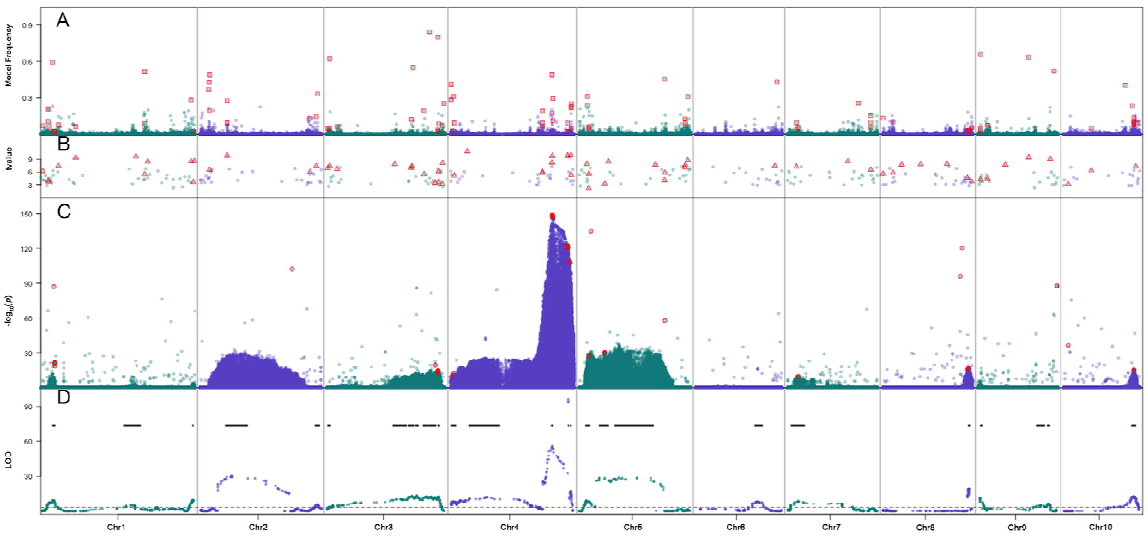


**Figure 3**

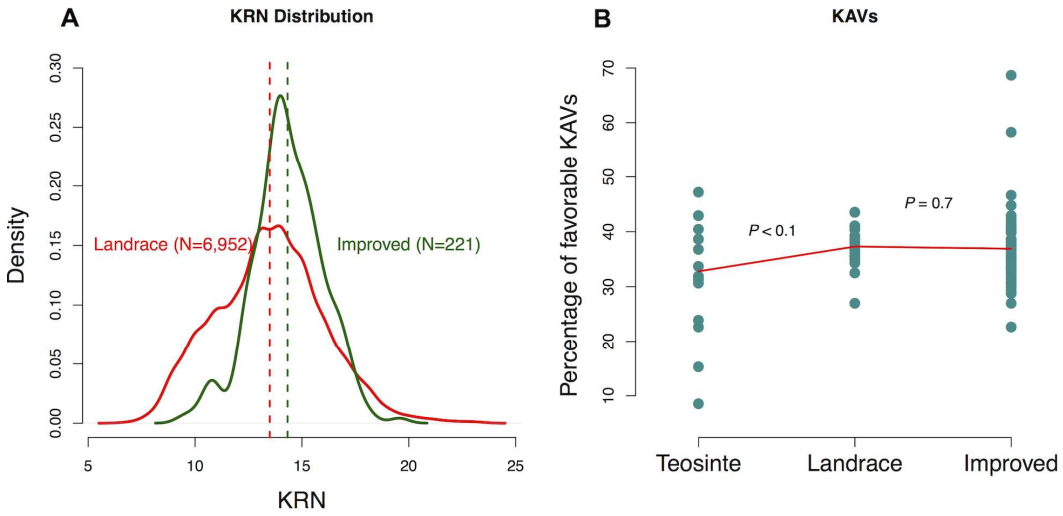
A
KRN Distribution
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Landrace (N=6,952)
Improved (N=221)
5
10
15
20
25
KRN
B
KAVs
Percentage of favorable KAVs
10
20
30
40
50
60
70
P < 0.1
P = 0.7
Teosinte
Landrace
Improved

**Figure 4**

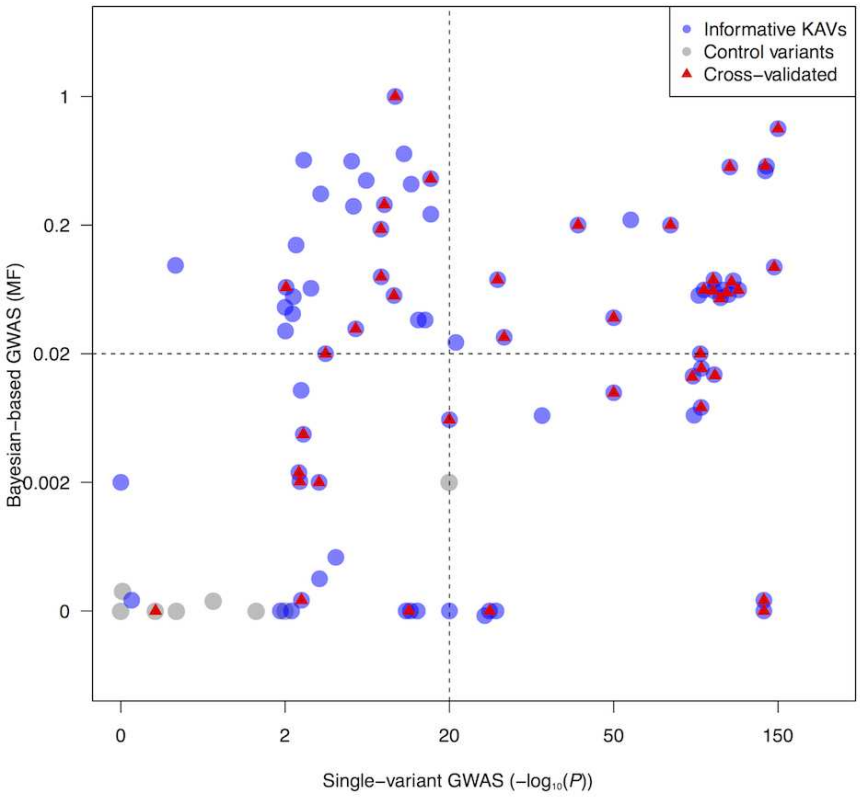
Informative KAVs
Control variants
Cross−validated
Bayesian−based GWAS (MF)
1
0.2
0.02
0.002
0
0
2
20
50
150
Single−variant GWAS (−log10(P))

**Figure 5**

# CHAPTER 3. DOMINANT GENE ACTION ACCOUNTS FOR MUCH OF THE MISSING HERITABILITY IN A GWAS AND PROVIDES INSIGHT INTO HETEROSIS

See supplemental information in Appendix B

Jinliang Yang[1], Li Li[1,2], Haiying Jiang[1,3], Dan Nettleton[4], and Patrick S. Schnable[1,5,*]

[1]Department of Agronomy, [2]Current address: Northwest Agriculture & Forestry University, Xi'an, China, [3]Current address: Shenyang Agricultural University, College of Agronomy, Shenyang, 110161, China, [4]Department of Statistics, [5]Center for Plant Genomics.

[*]Author for correspondence

**Abstract**

The phenomenon of heterosis has been observed for more than a century, but the underlying genetic mechanisms remain elusive. To better understand these genetic mechanisms, seven yield-related traits with varying levels of heterosis were subjected to GWAS using four related populations composed of 6,230 lines for which genotypes were available at ~13M sites. Multi-variant GWAS approaches that considered only additive effects explained 41-72% (Bayesian-based approach) or 43-72% (stepwise regression approach) of the narrow sense heritability of the seven traits. Interestingly, the percentage of phenotypic variance explained by these genome-wide markers was negatively correlated with the level of heterosis. A GWAS model that included both additive and dominant gene action increased 15-45% of the narrow sense heritability from additive only models, thereby accounting for much of the missing heritability. The level of heterosis was positively correlated with the number of trait-associated variants identified via GWAS that exhibited positive dominant gene action and magnitudes of their effects. Consistent with this, the inclusion of dominance into a genomic prediction model increased the accuracy of prediction performance for hybrids. In combination, these findings provide strong support for the view that positive dominant gene action contributes to heterosis.

## Introduction

Heterosis, or hybrid vigor, is the superior performance of hybrid offspring as compared to their parents. This genetic phenomenon, first documented by Darwin, was rediscovered by George Shull and Edward East decades later (Duvick 2001). Starting from early 1930s, the adoption of hybrid maize was followed by constant increases in grain yields in North American (Duvick and Cassman 1999; Tracy and Chandler 2006). As the rising demands for better hybrids, modern breeders are devoted to improving the inbred lines and maximizing the combining ability among these inbred lines. The intensive selection has rewired the evolution processes of this crop species and differentiated the relatively homogeneous starting population into different heterotic groups (van Heerwaarden et al. 2012). Although heterotic groups are consequences of human selection rather than historical or geographical contingencies (Tracy and Chandler 2006), it is still unclear what are the major reasons contribute to heterosis.

Researchers have proposed many genetic models, including dominance, overdominance and epistasis to explain heterosis (Birchler et al. 2003; Goff and Zhang 2013). Under certain circumstances, some of these genetic models might be favored. For example, an overdominant locus with heterozygous loss-of-function allele in tomato elevated yield by up to 60% (Krieger et al. 2010). However, in general, heterosis could not be adequately explained by a single gene or a simple model (Birchler et al. 2003). Because of the quantitative nature of heterotic traits, large population sizes and high-densities of markers are required to characterize the genetic control of heterosis. Recently, with an ultra high-density map, researchers in studying rice hybrids demonstrated that the accumulation of multiple effects, including dominance, overdominance (or pseudo-overdominance), and dominance by dominance interaction, could partly explain the genetic basis of heterosis, although their relative contributions varied by trait (Zhou et al. 2012).

The advent of NGS-enabled GWAS has provided an unprecedented ability to dissect the genetic architecture controlling for the phenotypic traits. However, for most of the studies, only a modest amount of the phenotypic variance could be accounted, which raised an issue of ‘missing heritability’ (Manolio et al. 2009). The

phenomenon of missing heritability is widely observed in GWAS conducted for human diseases and quantitative traits. For example, only about 5% of the phenotypic variance could be accounted for human height, although more than 50 height-associated loci have been identified in a combined population size composing of ~63,000 individuals (Visscher 2008). Yang *et al*. argued that the heritability is not missing but previously failed to be detected due to the presence of a large number of small effect variants (Yang et al. 2010). By adding these small effects loci in the analysis, they were able to explain 45% of phenotypic variance of human height. Even with the great improvement of this approach in explaining missing heritability, there still a proportion of the heritability that remains unexplained. Advanced statistical method with the modeling of non-additive gene actions (Zeng et al. 2013), but excluding GxE interactions that are not part of narrow-sense heritability, may fill in the remaining gap.

In maize, the availability of diverse genetic stocks and the availability of genome-wide analytical tools, make an in-depth investigation of the mode of inheritance of trait-associated variants (TAVs) identified via GWAS possible. Seven yield-related traits including average kernel weight (AKW), cob diameter (CD), cob length (CD), cob weight (CW), kernel count (KC), kernel row number (KRN), and total kernel weight (TKW) were phenotyped on 6,230 lines in four related populations. The yield of maize is determined by several component traits, such as ears per area, KRN per ear, kernel numbers per row and AKW (Stange et al. 2013). The seven phenotypic traits collected in this study were either yield component traits such as KRN, TKW, AKW and KC, or yield related traits having correlations with the yield component traits, such as CD, CW and CL. Consistent with previous observations, heterosis was prevalent and varied in these yield-related traits (Flint-Garcia et al. 2009). However, heterosis exhibited little correlation among different traits using the measurements of both the percentage of high parental heterosis (HPH) and the percentage of mid-parental heterosis (MPH) (Flint-Garcia et al. 2009). In contrast, the seven yield-related traits *per se* are highly correlated and some of the negative correlations would impede further yield improvement if one were to select for a single of these traits.

As a step towards understanding the genetic basis of heterosis and missing heritability, we conducted a GWAS for seven yield-related traits that exhibit varying levels of heterosis. Using a strictly additive genetic model, more than 80% of the phenotypic variance of low-heterosis traits could be explained using genome-wide markers. In contrast, less than 50% of phenotypic variance of high-heterosis traits could be explained in this manner. Inclusion of dominance gene action in the model had little effect on the percentage of variance explained for low-heterosis traits, but substantially increased the percentage of variance that could be explained for high-heterosis traits. These results suggest that the failure to account for non-additive genetic variation in statistic models accounts for much of the missing heritability observed in many GWAS. Further, the finding that the proportion of TAVs for a given trait that exhibited positive dominance gene action was correlated with the levels of heterosis, provides evidence for the role of dominance gene action in heterosis. It has previously been observed that negative correlations impede selection for yield component traits (Hallauer et al. 2004). Finally, we identified pleiotropic QTLs, some of which broke the negative correlations of yield-related traits and that may therefore be appropriate targets for marker assisted selection (MAS).

## Results

### Correlations of seven yield related traits

Using methods and populations described in Chapter 2, phenotypic values were collected from four GWAS populations, including (1) recombinant inbred lines (RILs) from nested association mapping (NAM) (Yu et al. 2008) and intermated B73 and Mo17 (IBM) (Lee et al. 2002) populations, (2) B73 crossed to a subset of those RILs, (3) Mo17 crossed to a subset of the same RILs and (4) a partial diallel population created from NAM founder lines and Mo17. From these populations, phenotypic data for seven yield-related traits (KRN, AKW, CD, CL, CW, KC, and TKW) were collected. Additional phenotypic data were obtained from the Panzea database (www.panzea.org) for one of the populations (NAM RILs). These data were combined and a mixed linear model was fit to estimate the average phenotypic value for each of the 6,230 lines (**Table S1, Figure S1**).

Pairwise correlations of the seven yield-related traits indicated that many exhibited either positive or negative correlations (**Figure S2**), with correlation coefficients (*r*) that ranged from –0.13 (KRN vs. AKW) to 0.90 (KC vs. TKW). Because the other six traits were all positively correlated with TKW, naively, one might assume that selection for any of these six traits could potentially increase TKW. However, the existence of negative correlations among some of the traits, such as KRN vs. CL ($r$ = –0.011), CD vs. CL ($r$ = –0.012) and KRN vs. AKW ($r$ = –0.13), impedes such an approach for increasing yield.

**Identification of pleiotropic QTLs with congruent versus antagonistic effects**

To overcome the obstacles these negative correlations pose for yield improvement, it would be useful to identify pleiotropic QTLs that have same signs of effect for multiple traits (or synergistic effect QTLs) that contribute to TKW. Such ~~synergistic~~ effect QTLs could potentially be identified by conducting separate linkage analyses of each of the 26 subpopulations that comprised one of our GWAS populations. By scanning the 549 QTLs identified in the 26 RIL subpopulations for the seven traits (**Table S2**), 15 sets of QTLs were detected (**Table S3**) that beneficially affect both traits from pairs of negatively correlated traits, such as KRN vs. AKW, KRN vs. CL, and CD vs. CL (**Figure S3-S5**). Most (12/15; 80%) of the favorable alleles at these consistently beneficial effect QTLs were derived from B73.

In addition to these 15 loci, 26 other sets of pleiotropic effect QTLs, which have significant effects on at least two traits, were identified. Of these sets of pleiotropic effect QTLs, 13/26 (50%) were found to control two traits, 9/26 (35%) control three traits and 1/26 (4%) QTL controls four traits.

**Yield-related traits exhibit different levels of heterosis**

With the availability of phenotypic traits of the inbred parents, levels of heterosis for the F1 hybrids in the partial diallel population could be estimated. After calculating the heterosis indexes of percentage of HPH and percentage of MPH, the average levels of heterosis varied for both the indexes of heterosis across seven traits. Results show that the levels of heterosis for the seven yield-related traits varied from low (i.e., very little heterosis) for the CD and KRN traits, to moderate, for

the CW, CL and AKW traits, to high for KC and TKW (**Figure 1**). And the levels of heterosis were weakly correlated after pairwise comparisons, whereas the correlation coefficients (*r*) ranged from –0.013 to 0.9 (mean = 0.49 and median = 0.55) for HPH and ranged from 0.081 to 0.93 (mean = 0.58 and median = 0.59) for MPH. These weak correlations among the levels of heterosis for different traits are consistent with previous observations (Flint-Garcia et al. 2009).

**Levels of heterosis negatively correlated with narrow sense heritability of the traits**

Our previous research demonstrated that different statistical approaches for conducting GWAS were complementary (Yang et al. submitted). Using three complementary approaches, GWAS was conducted separately for the seven yield-related traits using a set of ~13M variants. Collectively, 758 trait-associated variants (TAVs) were identified with the arbitrary thresholds of $-\log_{10}(P\text{-value}) > 50$ for single-variant approach, $-\log_{10}(P\text{-value}) > 10$ for stepwise approach and posterior model frequency > 0.1 for Bayesian-based approaches. These TAVs represented 524 1-Mb bin. As shown in **Figure 2**, approximately 10% of these trait-associated bins (1-Mb) exhibited pleiotropic effects.

The phenotypic variances due to genetic factors (or narrow sense heritability) were obtained for the seven yield-related traits. Using the Bayesian-based multi-variant approach, which can estimate the phenotypic variance explained by all the genome-wide markers simultaneously, 71% and 68% of the phenotypic variances associated with the low-heterosis traits CD and KRN could be explained. In contrast, only 51% and 41% of the phenotypic variances could be explained for the high-heterosis traits, TKW and KC. Overall, a negative correlation was observed between the heritability of traits and the levels of their heterosis (*P* value < 0.05). A similar negative correlation was observed using data from a different multi-variant approach (stepwise regression) (*P* value < 0.05) (**Figure 3A**). Hence, using both types of analyses, high-heterosis traits exhibit high levels of "missing heritability" (i.e., the fraction of phenotypic variance that can not be explained by genome-wide markers).

Both approaches (Bayesian-based multi-variant approach and stepwise regression approach) accounted for only additive effects. We therefore hypothesized that the failure to account for non-additive genetic effects (dominance, overdominance or epistasis) contributes to the missing heritability observed for high-heterosis traits. To test this hypothesis, we conducted GWAS on the partial diallel population (the only population which was segregating for all three possible genotypes at a given locus) using a version of the Bayesian-based multi-variant approach that included dominance effects (Zeng et al. 2013). Using this approach, an average of ~90% of the phenotypic variance could be explained for seven yield-related traits. Importantly, the model greatly improved the proportion of phenotypic variances explained for traits exhibiting high-heterosis and therefore substantially reduced the missing heritability associated with these traits. After partitioning the total variance into additive and dominant variances, both positive and negative correlations were observed with levels of heterosis, respectively (**Figure 3B**). These results provided strong support for the view that the phenomenon of missing heritability observed in GWAS is at least in part a consequence of a failure to account for non-additive genetic effects in statistical models.

**Mode of inheritance for individual TAVs**

Next, an investigation on the mode of inheritance (or gene action) of individual TAVs was conducted. To reduce over-representation of TAVs concentrated in certain regions, a stepwise approach was used to select TAVs for this experiment, each of which was required to have a $-\log_{10}(P\text{-value}) > 5$. Based on the degree of dominance (DD), TAVs were classified into three categories of mode of inheritance: positive dominance (DD > 0.5), negative dominance (DD < –0.5), and additive (–0.5 <= DD <= 0.5). Of these 959 TAVs investigated for mode of inheritance, 532 were characterized as exhibiting positive dominance effects, 211 were characterized as exhibiting negative dominance effects, and 216 were characterized as exhibiting additive effects (**Figure 4**). For all traits, regardless of the levels of heterosis, approximately 20% of the TAVs exhibited an additive mode of inheritance. For low-heterosis traits, such as KRN and CD, approximately equal numbers of TAVs were classified as exhibiting positive and negative dominance. In contrast, among the high-heterosis traits, such as KC and TKW, TAVs exhibting negative dominance

effects were numerically replaced with TAVs exhibiting positive dominance effects. Consequently, the proportions of positive and negative dominance effects TAVs were positively and negatively correlated with the levels of heterosis observed across the traits, respectively (*P* value < 0.05) (**Figure 4**). In addition, the magnitude of the effects of the heterozygous positive-dominance TAVs was higher for high-heterosis traits than for low-heterosis traits. Varying the *h* value cutoff used to classify loci from 0.3 to 0.7 did not alter these general trends.

**Inclusion of dominant gene action terms increases the accuracy of hybrid performance predictions**

In the past, breeders primarily relied upon empirical yield testing to determine which pairs of inbreds should be crossed to produce commercial hybrids. With the advent of doubled haploid technologies (Debuyser and Henry 1986; Smith et al. 2008), it has become possible for breeders to generate so many isolines, such that it is no longer possible to empirically testing all possible hybrid combinations. Therefore, there is great interest in identifying methods that predict hybrid performance of pairs of inbreds. To avoid the necessity of yield testing all possible pairwise combinations of inbreds, various methods of predicting hybrid performance have been reported (de los Campos et al. 2013; Desta and Ortiz 2014). We projected the ~13 million variants from the 27 NAM parental inbreds to obtain the genotypes of the 351 possible hybrids that could be obtained by intercrossing these 27 inbreds. Using the phenotypic data that were available for 221 of these hybrids we constructed two genomic selection (GS) models that could be used to predict hybrid performance. The first model considered only additive gene action, while the second included dominance gene action. The correlations of the predicted phenotypic performance with the observed values were heritability we discussed above. Two cross-validation strategies were used to compare the prediction accuracies for these two models. The first of the strategies was random splitting the partial diallel population into a training set (80%) and a validation set (20%). The second strategy involved separating the population into a training set by removing hybrids derived from certain founders and a validation set containing these removed hybrids.

The prediction accuracies obtained from first strategy ranged from 0.8 to 0.9 (median=0.85) for the seven yield related traits using only the additive model; the inclusion of dominance gene action significantly increased the prediction accuracies from 0.85 to 0.95 (median=0.9) based on a paired t-test (*P* value < 0.05). As expected lower prediction accuracies were obtained for the second strategy, which simulated the introduction of new founder lines to the diallel system (**Figure 5**). The additive model provided a median accuracy of 0.75, while the dominance model provided a median accuracy of 0.78. Notably, a large improvement was observed for the high-heterosis traits, with 10% improvement for KC and 12% for TKW, respectively.

## Discussion

The advent of marker-assisted selection (MAS) has transformed plant breeding programs that previously relied solely upon phenotypic selection (Collard and Mackill 2008). The identification of the genetic architecture controlling phenotypic variations is the first step towards MAS. Because yield is controlled by many loci, most of which have small effects, it is challenging to use this strategy to increase yields. An alternative strategy would be to develop molecular markers that could be used to select yield component traits that typically exhibit greater heritability than does yield *per se* (Robbins and Staub 2009). In this study, GWAS was used to identify TAVs for seven yield-related traits. Even with the availability of many target TAVs for selection, selection for only a single yield component trait may eventually encounter obstacles due to negative correlations among pairs of yield-related traits, such as CL vs. CD. For example, Hallauer conducted a long-term divergent selection project primarily focused on cob length (Hallauer et al. 2004). After 30 generations of selection, he obtained very long ears, but failed to improve the yield because as ears got longer their KRN became smaller. To overcome these obstacles, we identified 15 pleiotropic QTLs that improve both traits among pairs of negatively correlated traits. MAS based on these QTLs would be expected to prove more effective than MAS selection for individual yield component traits.

One of the limitations of MAS technology is that it only targets major effect loci and ignores loci with minor effects. Genomic selection bridges this gap by simultaneously estimating the effects of all the markers across the genome,

regardless of the magnitudes of their effects. Using genome-wide markers, we were able to explain a high proportion of the phenotypic variance for low-heterosis traits, such as KRN and CD. When applying the same model to high-heterosis trait, such as KC and TKW, a large proportion of the phenotypic variance remained unexplained. This phenomenon, i.e., 'missing heritability', has been widely observed in human GWAS (Manolio et al. 2009). In an attempt to uncover basis for the missing heritability, a Bayesian-based multi-variant approach that included dominant gene action was implemented and used to conduct GWAS. Using this model an average of 92% of the phenotypic variance could be accounted for, even for high-heterosis traits (92% for KC and 96% for TKW) that exhibited particularly high levels of missing heritability. Hence, this study suggests that much of the missing heritability may be a consequence of the failure to include non-additive gene action in the statistical models used for GWAS.

The amount of heterosis displayed by a trait was positively correlated with the number of TAVs identified via GWAS that exhibited positive dominant gene action and the magnitudes of their effects for that trait. These findings are consistent with a bi-parental QTL study that found that dominant gene action contributes to heterosis in rice (Zhou et al. 2012). Our population-based findings can explain why different hybrids exhibit different amounts of heterosis. Specifically, hybrids that exhibit high levels of heterosis are likely to be heterozygous at many loci that exhibit positive dominant gene action and homozygous for many loci that exhibit negative dominant gene action. Similarly, our results suggest that it may be possible to convert a trait that exhibits a low level of heterosis, such as KRN into one that exhibits a high level of heterosis by selecting (or creating) parents that are polymorphic for loci that exhibit positive dominant gene action but that are not polymorphic for loci that exhibit negative dominant gene action.

Our population-based view of the genetic control of heterosis is also consistent with the hypothesis that selection that alters allele frequencies in two populations in ways that enhance opportunities for positive dominant gene action and reduce opportunities for negative dominant gene action has the potential to be a major force for the development of heterosis, as may have happened with maize heterotic groups (van Heerwaarden et al. 2012).

As the cost of genotyping has fallen relative to the cost of hybrid yield tests, there has been interest in using genotypic data to predict the performance of the vast number of potential hybrids that could be created by intercrossing the many inbreds that can now be efficiently generated using doubled haploid technology (Forster and Thomas 2005; Desta and Ortiz 2014). Using the knowledge gained about the genetic control of heterosis and missing heritability, we were able to build genomic selection models for heterosis prediction that outperformed by 10% prediction models that did not account for non-additive gene action. Of course because our training data were generated in the same environment in which prediction took place it is expected that the accuracy of prediction will decrease in novel environments. Even so, because genetic gain is a function of selection intensity and prediction models allow for more stringent selection intensity, the use of prediction models that account for non-additive gene action has the potential to increase the rate of genetic gain.

## Materials and Methods

**Phenotypic data collection**. Seven yield-related traits, including AKW, CD, CL, CW, KC, KRN and TKW phenotypes were collected from four related populations as described in the previous study (Yang et al. submitted). These populations were (1) recombinant inbred lines (RILs) of intermated B73 and Mo17 (IBM, N = 325 RILs) (Lee et al. 2002) and the nested association mapping (NAM, N = 4,699 RILs) (Yu et al. 2008) populations, (2) a subset of the RILs that were backcrossed to the inbred line B73 (B73 x RILs, N = 692 BC1 lines), (3) a subset of the RILs that were backcrossed to the inbred line Mo17 (Mo17 x RILs, N = 289 BC1 lines) and (4) a partial diallel of the 26 NAM founders plus Mo17 (N = 221 F1 hybrids). For statistical analyses, the IBM RILs were treated as a subpopulation of the NAM RILs.

During the years 2008-2011, subsets of the above populations were planted in replicated field trials in up to three fields in Ames, IA (summer season) and one field in Molokai, HI (winter season). There were 5-12 plants of the same line grown within each row. KRN counts were collected from mature ears. Phenotypic values were estimated for each line using a mixed linear model implemented in R (R Development Core Team 2010), with fixed effects for lines and random effects for

locations, years, plots and blocks. Phenotypic density distributions in this study were estimated and plotted using R with default smoothing parameters. After calculation, the seven yield-related traits in the four GWAS populations were normally distributed (Anderson-Darling normality tests *P* values ranged from 0.5 to 0.8) and their distributions stratified among different GWAS populations.

**Computing the levels of heterosis**. The partial diallel population was used to estimate the levels of heterosis for the seven yield-related traits. Average levels of heterosis for these hybrids were computed for each trait using percentage of high-parent heterosis (HPH) and percentage of mid-parent heterosis (MPH). The following formulas were employed for the calculations,

$$HPH = \frac{\sum_{i=1}^{n} \frac{F1_i - \max(P1_i, P2_i)}{F1_i}}{n}$$

$$MPH = \frac{\sum_{i=1}^{n} \frac{F1_i - \text{mean}(P1_i, P2_i)}{F1_i}}{n}$$

Where $F1_i$ indicates the phenotypic value of the *j*th hybrid from the partial diallel population; $P1_i$ and $P2_i$ are the phenotypic values of the corresponding parents of the hybrid; *n* indicates the total number of hybrids in the population.

**Identification of pleiotropic effect loci.** To identify the synergistic effect QTLs, separate linkage analyses were conducted in the 26 RIL subpopulations using composite interval mapping (CIM) method (Zeng 1993). After 1000 permutation tests for each trait, QTLs that exceeded the thresholds were identified for each of the seven yield-related traits and the QTL support intervals were determined using a 1.5-LOD drop-down from QTL peaks (Lander and Botstein 1994). For each pair of negatively correlated traits, extensive scanning of the previously identified QTLs was conducted in each of the 26 RIL subpopulations. Finally, loci were identified that exhibited overlapping support intervals for QTLs that had same signs additive effects on otherwise negatively correlated traits.

**Statistical model for GWAS.** The same statistical approaches, including single-variant, stepwise regression and Bayesian-based multi-variant approaches, were

used for conducting GWAS with additive model as reported in the previous study (Yang et al., submitted). While, the following dominance model was used for the Bayesian-based multi-variant approach:

**Bayesian-based additive and dominance models.** The following additive and dominant model was used for the Bayesian-based multi-variant approach:

$$Y_l = u_k + \sum_{i=1}^{4} a_{ik} P_{il} + \sum_{j=1}^{26} b_{jk} S_{jl} + \sum_{k=1}^{\sim 13M} (c_k X_{kl} + d_k W_{kl}) + e_l$$

where $Y_l$ is the adjusted phenotypic value for line $l$ from the mixed linear model analysis; $u_k$ is an intercept parameter; $P_{il}$ is 1 if line $l$ is of GWAS population $i$ and is 0 otherwise, and $a_{ik}$ is the effect of the $i$th population in the model for variant $k$; $S_{jl}$ is 1 if line $l$ is from subpopulation $j$ and 0 otherwise, $b_{jk}$ is the effect of subpopulation $j$ in the model for variant $k$; $X_{kl}$ indicates the additive genotype (coded as -10 for B73-like genotype, 0 for heterozygous genotype and 10 for non-B73 genotype) of the $k$th variant in line $l$ and $c_k$ is the effect of the $k$th variant; $W_{kl}$ indicates the dominance genotype (coded as -10 for homozygous genotype and 0 for heterozygous genotype) of the $k$th variant in line $l$ and $d_k$ is the dominance effect of the $k$th variant; and $e_l$ the error term.

The BayesC option of GenSel was used for the above analysis. Testing runs were conducted before the full experiment to optimize the prior information to feed the model, such as residual and genotypic variances. In the full experimental runs, 41,000 chains of iterations were used with the first 1,000 iterations discarded as a burn-in. After finishing the model training, posterior model frequency of each variant was used *in lieu* of a traditional measurement of significance and the posterior phenotypic variance explained by genome-wide markers were obtained as a measurement of heritability.

**Procedure to determine mode of inheritance.** TAVs selected by stepwise regression approach with –log10(P-value) > 5 were used to reduce the possible over-representation of variants in certain regions. Before the calculation, phenotypic values were normalized and re-scale to make them comparable among the seven yield-related trait. Then, for each TAV, a separate generalized linear model was

constructed, where the population and subpopulation were fitted as co-variants. With the model, two homozygous allele effects and one heterozygous allele effect were estimated; and these values were used to compute the additive and dominance effects. The degree of dominance ($h$) was the calculate as $h = \boldsymbol{d}/|\boldsymbol{a}|$, where $\boldsymbol{d}$ indicates the dominance effect and $\boldsymbol{a}$ indicates the additive effect. Of the computed $h$ value for each TAV, if $h > 0.5$, a positive dominance locus was claimed; if $h < -0.5$, a negative dominance locus was claimed; if $-0.5 <= h <= 0.5$, an additive locus was claimed. To test the sensitivity of the arbitrary cutoff of 0.5, other cutoff from 0.3 to 0.7 were also used in characterization of the TAVs' gene action.

## Acknowledgements

We gratefully acknowledge technical support provided by Ms. Uyen Pham and Ms. Talissa Sari, and thank Ms. Lisa Coffey for generating and maintaining genetic stocks. This research was supported by grants from Monsanto and the National Science Foundation to P.S.S. and colleagues (IOS-1027527).

## References

Birchler JA, Auger DL, Riddle NC. 2003. In search of the molecular basis of heterosis. *The Plant cell* **15**(10): 2236-2239.

Collard BC, Mackill DJ. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **363**(1491): 557-572.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* **193**(2): 327-+.

Debuyser J, Henry Y. 1986. Use of Doubled Haploids in Plant-Breeding. *B Soc Bot Fr-Actual* **133**(4): 51-57.

Desta ZA, Ortiz R. 2014. Genomic selection: genome-wide prediction in plant improvement. *Trends in plant science* **19**(9): 592-601.

Duvick DN. 2001. Biotechnology in the 1930s: the development of hybrid maize. *Nat Rev Genet* **2**(1): 69-74.

Duvick DN, Cassman KG. 1999. Post-green revolution trends in yield potential of temperate maize in the north-central United States. *Crop Science* **39**(6): 1622-1630.

Flint-Garcia SA, Buckler ES, Tiffin P, Ersoz E, Springer NM. 2009. Heterosis is prevalent for multiple traits in diverse maize germplasm. *PloS one* **4**(10): e7433.

Forster BP, Thomas WTB. 2005. Doubled Haploids in Genetics and Plant Breeding. *Pl Bred Re* **25**: 57-88.

Goff SA, Zhang Q. 2013. Heterosis in elite hybrid rice: speculation on the genetic and biochemical mechanisms. *Current opinion in plant biology* **16**(2): 221-227.

Hallauer AR, Ross AJ, Lee M. 2004. Long-term divergent selection for ear length in maize. *Plant Breeding Reviews* **24**(2): 153-168.

Krieger U, Lippman ZB, Zamir D. 2010. The flowering gene SINGLE FLOWER TRUSS drives heterosis for yield in tomato. *Nature genetics* **42**(5): 459-463.

Lander ES, Botstein D. 1994. Mapping Mendelian Factors Underlying Quantitative Traits Using Rflp Linkage Maps (Vol 121, Pg 185, 1989). *Genetics* **136**(2): 705-705.

Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, Hallauer A. 2002. Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant Mol Biol* **48**(5): 453-461.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**(7265): 747-753.

Robbins MD, Staub JE. 2009. Comparative analysis of marker-assisted and phenotypic selection for yield components in cucumber. *Theor Appl Genet* **119**(4): 621-634.

Smith JSC, Hussain T, Jones ES, Graham G, Podlich D, Wall S, Williams M. 2008. Use of doubled haploids in maize breeding: implications for intellectual property protection and genetic diversity in hybrid crops. *Molecular Breeding* **22**(1): 51-59.

Stange M, Schrag TA, Utz HF, Riedelsheimer C, Bauer E, Melchinger AE. 2013. High-density linkage mapping of yield components and epistatic interactions in maize with doubled haploid lines from four crosses. *Molecular Breeding* **32**(3): 533-546.

Tracy W, Chandler M. 2006. The historical and biological basis of the concept of heterotic patterns in corn belt dent maize. *Plant Breeding: The Arnel R Hallauer ...*: 219-233.

van Heerwaarden J, Hufford MB, Ross-Ibarra J. 2012. Historical genomics of North American maize. *Proc Natl Acad Sci U S A* **109**(31): 12420-12425.

Visscher PM. 2008. Sizing up human height variation. *Nature genetics* **40**(5): 489-490.

Yang JA, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42**(7): 565-U131.

Yu JM, Holland JB, McMullen MD, Buckler ES. 2008. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**(1): 539-551.

Zeng J, Toosi A, Fernando RL, Dekkers JCM, Garrick DJ. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genetics, Selection, Evolution* **45**(11): 26 April 2013.

Zeng ZB. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc Natl Acad Sci U S A* **90**(23): 10972-10976.

Zhou G, Chen Y, Yao W, Zhang C, Xie W, Hua J, Xing Y, Xiao J, Zhang Q. 2012. Genetic composition of yield heterosis in an elite rice hybrid. *Proc Natl Acad Sci U S A* **109**(39): 15847-15852.

## Figure Legends

**Figure 1. Levels of heterosis for the seven yield-related traits**. Mean percentage of HPH and percentage of HPH values were plotted using the partial diallel populations.

**Figure 2. The catalog of GWAS results of seven yield-related traits.** The colored dots in the plot represent TAVs in a 1-Mb bin.

**Figure 3. Relationships between the levels of heterosis and narrow sense heritability observed for the seven yield-related traits.** The levels of heterosis are measured using HPH (A) and MPH (B). In the panels, red, blue and black lines indicate heritability accounted by additive, dominant and both of these gene action, respectively.

**Figure 4. Mode of inheritance for individual TAVs.** In panel (**A**), the pie charts show the proportion of the number of TAVs exhibiting positive dominant, negative domiant and additive gene action. In panel (**B**), (**C**) and (**D**), magnitudes of the effects of TAVs with positive dominant, negative dominant and additive gene action. The seven yield-related traits were ordered according to their level of heterosis (HPH).

**Figure 5. Prediction accuracies for genomic selection using two strategies**.

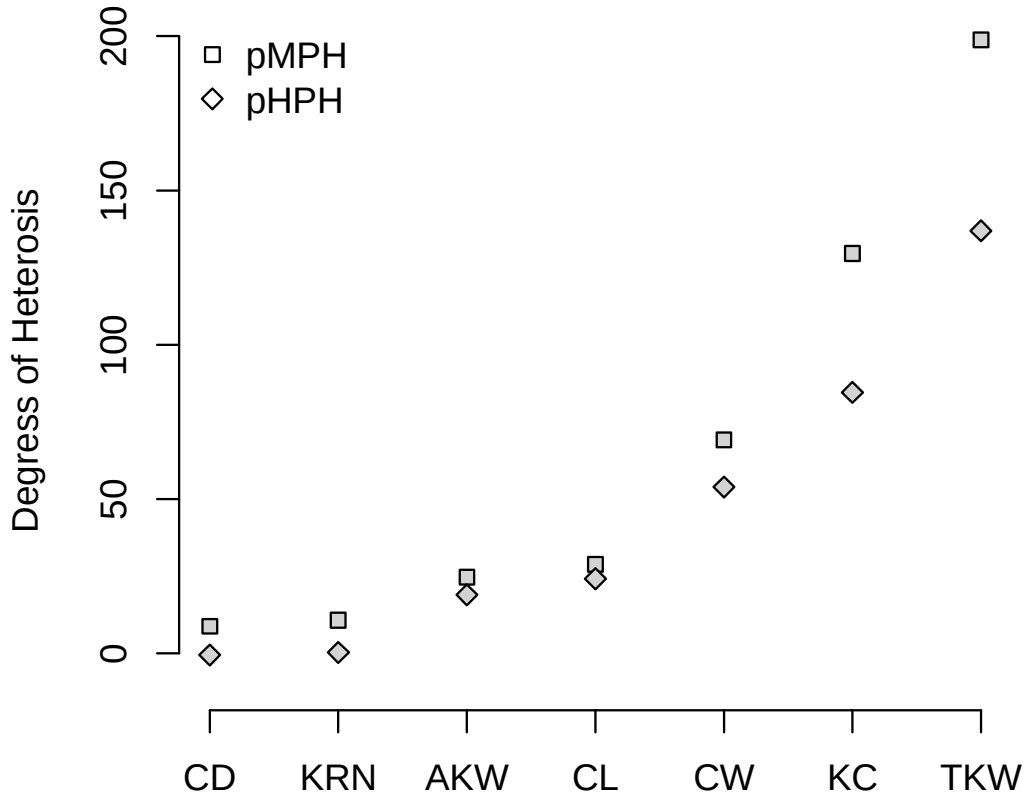
Degress of Heterosis
0
50
100
150
200
pMPH
pHPH
CD
KRN
AKW
CL
CW
KC
TKW

**Figure 1**

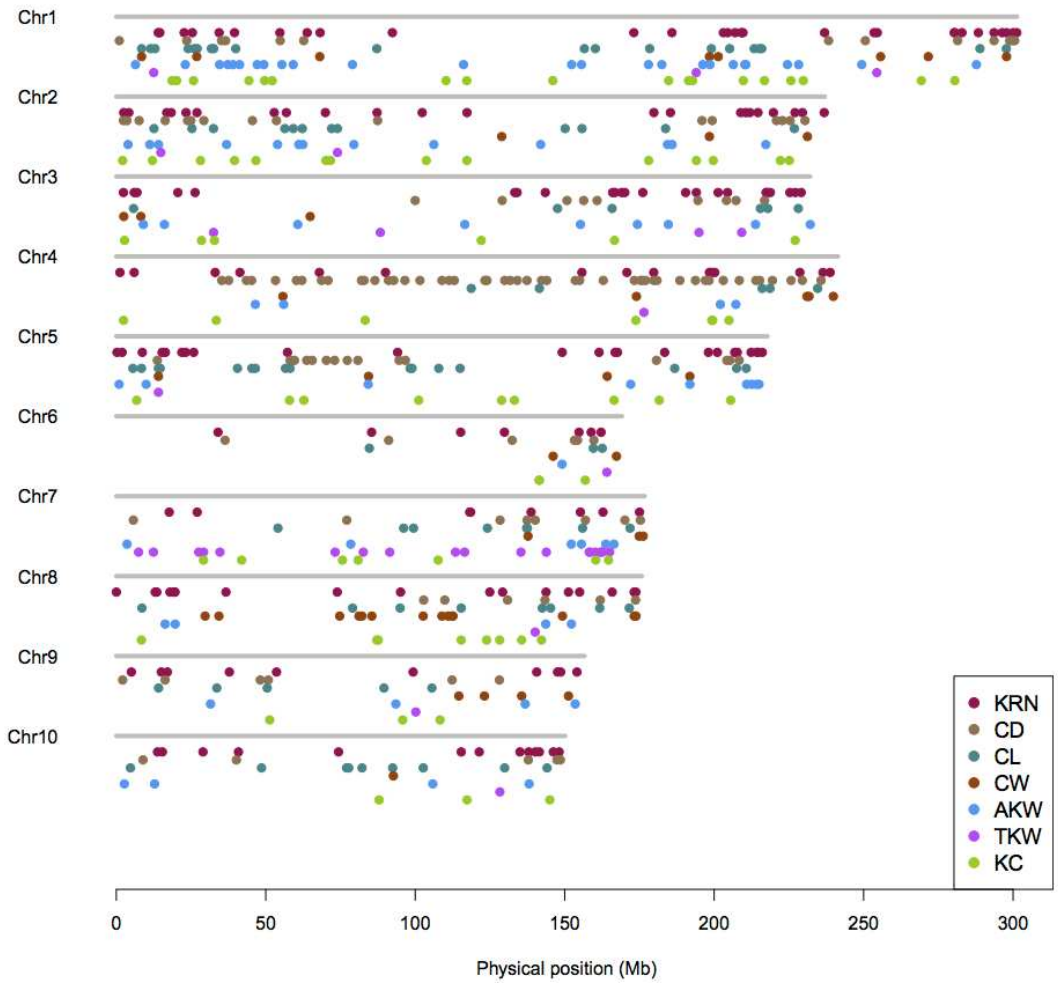
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
KRN
CD
CL
CW
AKW
TKW
KC
0
50
100
150
200
250
300
Physical position (Mb)

**Figure 2**

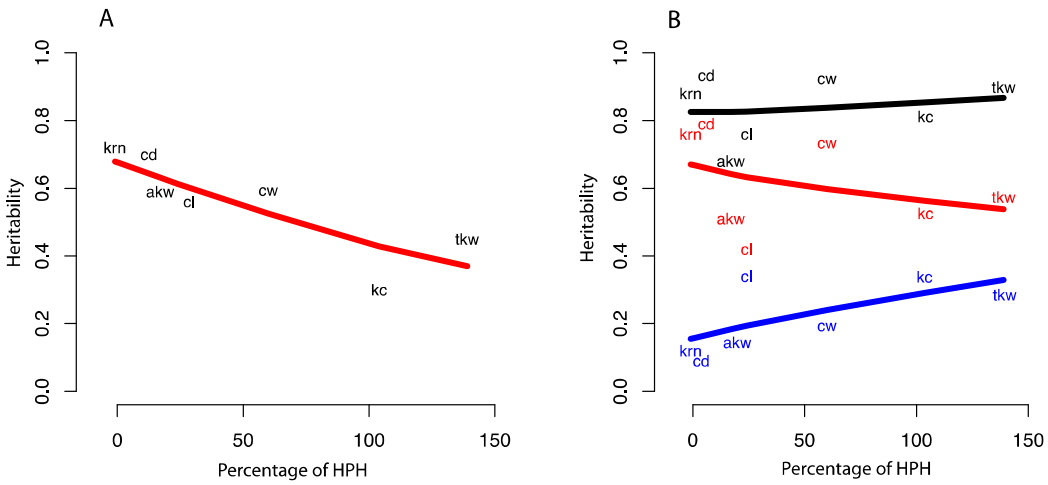
A
Heritability
0.0
0.2
0.4
0.6
0.8
1.0
krn
cd
akw
cl
cw
tkw
kc
0
50
100
150
Percentage of HPH
B
Heritability
0.0
0.2
0.4
0.6
0.8
1.0
cd
krn
cw
tkw
kc
krn
cd
cl
cw
akw
tkw
kc
akw
cl
cl
kc
tkw
cw
krn
akw
cd
0
50
100
150
Percentage of HPH

**Figure 3**

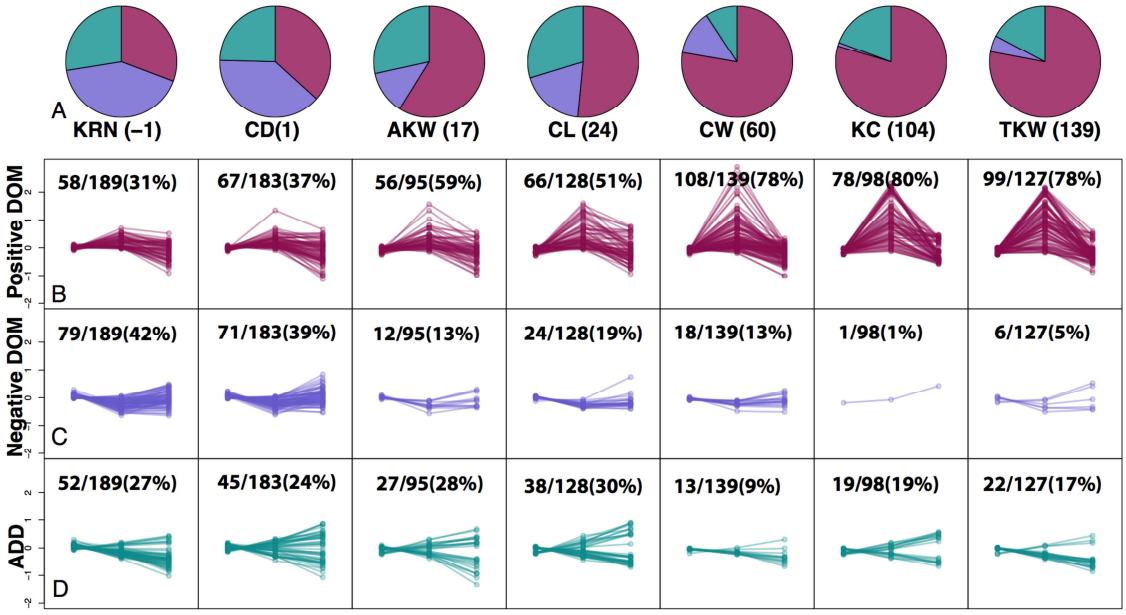
A
KRN (−1)
CD(1)
AKW (17)
CL (24)
CW (60)
KC (104)
TKW (139)
Positive DOM
58/189(31%)
67/183(37%)
56/95(59%)
66/128(51%)
108/139(78%)
78/98(80%)
99/127(78%)
B
Negative DOM
79/189(42%)
71/183(39%)
12/95(13%)
24/128(19%)
18/139(13%)
1/98(1%)
6/127(5%)
C
ADD
52/189(27%)
45/183(24%)
27/95(28%)
38/128(30%)
13/139(9%)
19/98(19%)
22/127(17%)
D

**Figure 4**

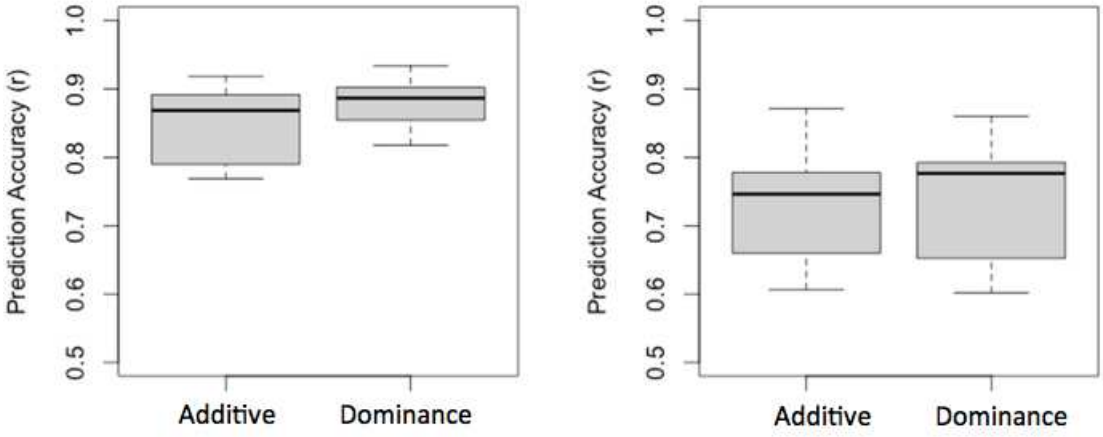
Prediction Accuracy (r)
0.5
0.6
0.7
0.8
0.9
1.0
Additive
Dominance
Prediction Accuracy (r)
0.5
0.6
0.7
0.8
0.9
1.0
Additive
Dominance

**Figure 5**

# CHAPTER4. EXTREME PHENOTYPE-GENOME-WIDE ASSOCIATION STUDY (XP-GWAS): A METHOD FOR IDENTIFYING TRAIT-ASSOCIATED VARIANTS BY SEQUENCING POOLS OF INDIVIDUALS

See supplemental information in Appendix C

Jinliang Yang[1,6], Haiying Jiang[1,2,6], Cheng-Ting "Eddy" Yeh[1], Jeffrey A. Jeddeloh[3], Dan Nettleton[4], and Patrick S. Schnable[1,5,*]

[1]Department of Agronomy, [2]Current address: Shenyang Agricultural University, College of Agronomy, Shenyang, China, 110161,[3]Research Informatics, Roche NimbleGen, Madison, WI 53705, USA, [4]Department of Statistics, [5]Center for Plant Genomics.

[6]These authors contributed equally to this work.

[*]Author for correspondence

**Abstract**

Phenotypic traits of agronomic importance are typically controlled by quantitative trait loci (QTLs). Although approaches for conducting QTL mapping and genome-wide association studies (GWAS) are well developed, it remains challenging to map QTLs to high resolution and GWAS requires high-density genotyping of large numbers of individuals. Here we report a new method for conducting GWAS that does not require the genotyping of large numbers of individuals. Instead XP-GWAS (extreme phenotype GWAS) relies on genotyping pools of individuals from a diversity panel having extreme phenotypes. This analysis generates allele frequencies in the extreme pools, enabling the discovery of associations between genetic variants and traits of interest. This method was tested in maize using the kernel row number (KRN) trait, which was selected to enable comparisons between the results of XP-GWAS and a conventional GWAS. An exome-sequencing strategy was employed to focus sequencing resources on genes and their flanking regions. A total of 0.94 million variants with adequate depth of sequencing coverage were identified; via comparisons among pools, 145 of these variants were identified as being associated with the KRN phenotype. These trait-associated variants were significantly enriched in regions identified by a conventional GWAS. The high resolution of XP-GWAS was demonstrated by resolving linked QTLs and detecting trait-associated variants (TAVs) within a single gene under a QTL peak. XP-GWAS will be of particular valuable for detecting genes or alleles responsible for quantitative variation in species that do not have access to extensive genotyping resources, such as orphan crops or ecological species.

## Introduction

Despite the development of quantitative trait loci (QTL) mapping (Morton 1955) and genome-wide association studies (GWAS) (Klein et al. 2005), to rapidly and cost-effectively identify SNPs or genes associated with variation in complex traits remains challenging. Conventional QTL mapping is typically conducted using newly occurring recombination events in the progeny of bi-parental crosses. Hence, typically at most only two to four alleles are segregating in such crosses, limiting the number of trait-associated loci that can be detected. In addition, the limited number of recombination events usually results in relatively large confidence intervals. GWAS, which employs historical recombination events with diverse parental origins, has the potential to discover a greater fraction of the genetic diversity within a species that contributes to the trait of interest. When conducted on large diversity panels, GWAS has the potential to provide high resolution mapping of trait-associated variants (TAVs). One of the limitations of the existing QTL mapping and GWAS approaches, however, is that they require genotyping large numbers of individuals, which can be expensive for large populations, even using recently developed cost-effective genotyping methods such as genotyping arrays (Steemers et al. 2006; Fu et al. 2010) and genotyping-by-sequencing (Elshire et al. 2011).

An alternative method for the identification of TAVs is bulk segregant analysis (BSA), which involves the genotyping of pools of individuals sorted by phenotype rather than genotyping individuals within a segregating population or a diversity panel (Michelmore et al. 1991). BSA can be conducted using any type of genetic marker that provides a quantitative read-out that is correlated with allele frequencies in the phenotypically distinct pools. New implementations of BSA have recently been reported that exploit advances in genotyping technologies, especially the development of next generation sequencing (NGS). For example, NGS-based BSA methods that rely on whole genome shotgun (WGS) sequencing have been applied to species with small genomes such as *Arabidopsis* (Schneeberger et al. 2009) and yeast (Wenger et al. 2010). Because these methods are not suitable for species with large genomes, we developed Sequenom-based BSA (Liu et al. 2010) and RNA-seq based BSA (BSR-Seq) (Liu et al. 2012) and used these technologies to map or clone

several maize genes whose qualitative mutants that have large effects (Yi et al. 2011; Makarevitch et al. 2012). Similarly, a mapping-by-sequencing strategy based on exome-capture was used to identify a single mutant gene that was segregating in a bi-parental F2 mapping population of barley (Masher et al., 2014). The extension of BSA to quantitative traits was demonstrated in a bi-parental cross of yeast (Ehrenreich et al. 2010).

As is the case with QTL mapping studies, all of these NGS-based BSA studies analyzed bi-parental populations that were segregating for only a fraction of the genetic diversity within a species. We were interested in extending the NGS-based BSA approach to diversity panels to more fully sample the genetic diversity that controls quantitative traits within a species. We were encouraged in this effort by a simulation study that indicated that if a sufficient number of progeny were used, NGS-based BSA could detect even small-effect loci (Ehrenreich et al. 2010). In addition to reducing the number of samples that must be genotyped, a pooling strategy has the potential to enrich for rare alleles and augment allele effects via extreme phenotypic selection. Hence, we elected to sequence pools of individuals that exhibit extreme phenotypes from a large diversity panel that would contain historical recombination events. Hence, this method combines the simplicity of genotyping pools with the superior mapping resolution of GWAS; it was thus termed eXtreme Phenotype-Genome-Wide Association Study (XP-GWAS).

We conducted XP-GWAS for the quantitative trait kernel row number (KRN) using a diversity panel of ~7,000 accessions. This trait was selected to enable comparisons to the results of a conventional GWAS. Approximately 200 lines with the lowest KRNs and a similar number with the highest KRNs were selected from the diversity panel. In addition, a random set of ~200 lines from the diversity panel were used as a control. These three pools were genotyped via an exome-capture and sequencing strategy that provided quantitative allele frequencies. XP-GWAS identified 145 TAVs. These variants are enriched in regions previously detected via traditional GWAS (Brown et al. 2011). We also demonstrated the resolution of XP-GWAS by separating multiple linked QTL and identifying a single candidate gene under a single QTL peak. XP-GWAS leverages BSA's simple experimental design with the high mapping resolution of GWAS, and may be particularly attractive for

researchers studying species for which large, individually genotyped diversity panels do not exist or can not easily be generated, such as orphan crops or ecological species.

## Results

### Identify and pool lines having extreme KRN phenotypes

The North Central Regional Plant Introduction Station (NCRPIS), which is part of the U.S. National Plant Germplasm System (NPGS), maintains more than 10,000 accessions of maize germplasm from across the world, representing the vast diversity of this species (Vigouroux et al. 2008). Phenotypic data, including KRN counts are available for 6,952 of these accessions via the Germplasm Resources Information Network (GRIN) database (http://www.ars-grin.gov/). The KRN trait is approximately normal distributed within this diversity panel with a mean of 13.4 (**Figure 1**). Using these KRN data we established three pools of accessions. The mean and median phenotypic values for the three pools are 8.7/9 (low KRN pool), 13.5/13 (random KRN pool) and 19.7/19 (high KRN pool). Each pool consists of ~200 (selection intensity ~3%) accessions (**Table S1**). The random pool was created in addition to high and low extreme phenotypic pools to reflect background population allele frequencies. The ~600 selected accessions originated from ~60 countries on six continents.

### Exome-sequencing of three XP-GWAS pools

XP-GWAS begins with genotyping the extreme phenotype pools. Genotyping with a pre-defined SNP array will create an inherent ascertainment bias. This bias can be overcome by *de novo* SNP discovery within the pools. This could, for example, be accomplished via whole genome sequencing (WGS) of each pool. However, because of its large genome (~2.3Gb) (Schnable et al. 2009) and high proportion of repetitive DNA (~80%) (Baucom et al. 2009), we elected to focus our sequencing resources on the genic regions of each pool. This was achieved by sequencing the products of an exome-capture experiment (Bashiardes et al. 2005a; Fu et al. 2010).

A solution-based sequence capture library was designed and manufactured by NimbleGen (**Materials and Methods**) to survey the complete B73 exome plus additional sequences which were not used in the current analysis (**Materials and Methods**). Using this 'Zeanome' probe library, sequence captures were conducted on barcoded, fragmented genomic DNA samples from three XP-GWAS pools. The captured DNAs were then sequenced using four lanes from an Illumina HiSeq2000 instrument, generating a total of ~770 million 100 bp paired-end (PE) reads. A custom bioinformatics pipeline (Li et al. 2012) was employed to align the raw reads to the maize B73 reference genome (RefGen_v2) (**Materials and Methods**). After data processing, about 302, 368 and 294 million single-end (SE) reads were uniquely mapped to the reference genome (**Table S2**) for the high, low and random KRN pools, respectively. These uniquely mapped reads were analyzed to evaluate capture performance and to call variants.

The exome of the filtered gene set (FGSv2) of the B73 reference genome was considered our intended target, although the design space included probes designed to other sequences (**Material and Methods**). Approximately 61% (high), 61% (low) and 63% (random) of uniquely mapped reads were captured by probes from the Zeanome library, even though only ~75% of the probes on the array were designed from the B73 reference exome. Average depths of sequencing on the filtered gene set of the three pools were 142X (high KRN), 175X (low KRN) and 145X (random KRN). Approximately 85% (84% for high, 87% for low and 84% for random) of reference genes have greater than 50X depth of coverage (**Figure 2A-2C**). The average percentages of coverage from transcript start to end for reference genes were 99.0% (high), 99.3% (low) and 98.6% (random). Approximately 98% (98% for high, 99% for low and 97% for random) of reference genes have at least 90% coverage (**Figure 2D-2F**). Bait probes can capture adjacent regions (Fu et al. 2010); therefore we anticipated capturing not only exonic regions but also intronic and promoter regions. Indeed, 7% (high), 6% (low) and 7% (random) of the reads were mapped to intronic or 5-kb upstream regions. The results indicated that, even using conservative estimates, the Zeanome Seq-Cap proved to be an efficient method to enrich the intended target with high depth of sequencing and high rate of coverage.

A total of 5.14 million variants including SNPs (N = 4.75 million, 92%) and small indels (N = 0.39 million, 8%) were identified using a custom variant calling pipeline (Li et al. 2012; Liu et al. 2012) (**Materials and Methods**). An adequate depth of read support for a given variant is critical to accurately estimate allele frequencies in XP-GWAS pools. However, increasing the minimum required depth of support before calling a variant dramatically reduces the number of common variants found across the three pools (**Figure S1**). A simulation study indicated that low depth of reads support would cause large inference variation due to random sampling (**Figure S2**, **Material and Methods**). Based on these simulations, we concluded that a minimum depth of support of 50X provides an appropriate balance between maintaining high numbers of variants and minimizing the negative effects of sampling variation. After filtering, using a 50X reads minimum cut-off, 944,549 common variants were retained, including 828,855 SNPs (88%) and 115,694 small indels (12%). These variants were distributed across 87% of the high confidence maize filtered genes (FGSv2); on average, 18 variants were detected for each gene and the most extreme gene (*GRMZM2G047347*) contains 246 polymorphic sites. As anticipated, variants were not limited to exonic regions; only ~41% of variants were located in exons. An additional ~34% of variants were located in introns and ~9% were located within 5-kb upstream of genes and ~10% were located within 5-kb downstream of genes (**Figure S3**). This is relevant because although genes and 5-kb upstream regions comprise only 13% of the genome, variations within these regions account for about 35%-47% of phenotypic variation in maize (Li et al. 2012). The ability of the Zeanome Seq-Cap library to capture both the exome and adjacent regions enabled us to focus sequencing resources thereby enhancing the power of this study to identify associations.

**Identification of extreme phenotype-associated variants**

The primary factor used to create the three phenotyptic pools was the KRN phenotype of accessions. Even though an effort was made to maintain geographic diversity with the pools, population structure or cryptic within-group relatedness was unavoidable. This cryptic population structure could lead to over-dispersion of the Chi-square test statistic, thereby resulting in false discovery. To attenuate the effects of population structure, a genomic control method (Devlin and Roeder 1999)

was implemented to adjust the Chi-square test statistic (**Materials and Methods)**. After implementing this genomic control, the quantile-quantile plot (**Figure S4**) showed that most of the observed data conformed closely to expectation except at the tail, which indicated the population structure was successfully controlled and some association signals were detected.

Using this approach, 145 TAVs were identified at a false discovery rate (FDR) (Benjamini and Hochberg 1995a) of 0.05 (**Figure 3**). These identified TAVs represent 121 1-kb bins distributed across 10 chromosomes. To understand the patterns of differences in allele frequencies amongst the pools, at each TAV site read counts matching the reference allele were divided by total read counts to derive the reference allele frequency (RAF). We noted that the B73 inbred (which provided the reference genome) with an average of 17.6 kernel rows is phenotypically closer to the mean value of the high KRN pool (mean KRN = 19.5) than to the value of low KRN pool (mean KRN = 8.7), and previous studies found that the KRN trait was mostly controlled by additive gene effects (Toledo et al. 2011). Therefore, it was not surprising that 81% (118/145) of RAFs of the TAVs exhibited inheritance pattern of high > random > low as compared with only 1% (2/145) that exhibited the opposite pattern (high < random < low) (**Figure S5 A** and **C**). The remaining TAVs exhibited other patterns (**Figure S5 B** and **D**).

**Comparisons between the results from XP-GWAS and traditional GWAS**

We compared the 145 TAVs to 261 TAVs previously detected via a conventional GWAS (Brown et al. 2011). The two sets of variants were mapped to the same version of the reference genome (AGPv2). Using a bin size of 1-Mb, 17% (25/145) TAVs were overlapped with the variants identified by traditional GWAS. This number of overlapping bins was statistically significant (*P* value < 0.05) based on a simulation test (**Materials and Methods**). The TAVs were also compared with our recently identified 986 TAVs by three complementary statistical approaches using ~13M variants and a lager population (Yang et al. submitted). Using the same 1-Mb bins, 35% (51/145, *P* value < 0.05) of the TAVs overlapped with the TAVs identified via this 2nd conventional GWAS.

**TAVs hit linked QTL regions with high resolution**

Several previous QTL studies detected multiple QTLs, which co-localize on the long arm of chromosome 4 (Beavis et al. 1994; Veldboom et al. 1994; Austin and Lee 1996). A recent conventional study via GWAS (Yang et al., submitted) also detected clusters of TAVs in these regions. In an attempt to resolve these linked QTLs, a chromosome walking experiment was conducted (**Methods**), ultimately mapping one of the KRN QTL to a 2.7-Mb interval defined by a pair of SNPs (**Figure S6**).

XP-GWAS also identified TAVs in this region. Using pairwise comparisons of three independent Chi-square tests (high vs. low, high vs. random and low vs. random) with genomic control, four variants passed an FDR < 0.05 threshold and one variant was supported by two of the independent pairwise tests (high vs. low and low vs. random)(**Figure 4**). These four TAVs were all located in the gene *GRMZM2G039106*, which itself located under the peak of the fine mapped QTL interval (**Figure S6**). The high KRN pool of these variants maintained high RAFs, which is consistent with our original determination that the favorable allele was derived from B73. In addition to identifying TAVs in this region, XP-GWAS identified TAVs in three other chromosomal regions on the long arm of chromosome 4: chr4:185-186M, chr4:186-187M and chr4:200-201M (**Figure S7-S9**). These variants were located in genes *GRMZM2G111928*, *GRMZM2G095141* and *GRMZM2G098557*. Favorable alleles of these loci were all derived from B73, which is also consistent with previous QTL findings.

**Phenotypic validation via replicated field trials**

The XP-GWAS analysis relied upon phenotypic data downloaded from the GRIN database. To test the reproducibility of the phenotypes, replicated field trials (**Materials and Methods**) were conducted using a subset of accessions selected because they exhibited extreme KRN phenotypes in the GRIN database. The correlation (*r*) between the GRIN data and our measurements in the low KRN pool was only 0.27 (N = 16 accessions, *P* value = 0.3) and for high KRN pool was only 0.45 (N = 29 accessions, *P* value < 0.01) (**Figure S10**). The cause of these low correlations is probably due to the fact that the accessions were both highly heterozygous and genetically heterogeneous. Even though the within-pool phenotypic correlations are

relatively low, a high correlation was observed between the two pools ($r$ = 0.96, $P$ value < 0.01). This indicates that the high KRN and low KRN pools could be clearly distinguished even using the phenotypes extracted from the GRIN database.

## Discussion

Conventional GWAS experiments have been used to identify loci associated with important traits in agricultural species. These analyses require that large panels of individuals be genotyped, which is still expensive even given recent advances in genotyping technologies. Using the trait-dependent pooling of extreme phenotypes from a diversity panel, we cost-efficiently identified 145 TAVs with only several days of hands-on time and at modest cost. The resolution of XP-GWAS is comparable with conventional GWAS because both methods employ diversity panels. In this study, we were able to identify variants within a fine-mapped QTL region that is imbedded with a cluster of linked-QTL. Because of the exome-sequencing strategy used in this study, ~90% of the TAVs were located in genes and their 5-kb flanking regions. Although as a consequence of linkage disequilibrium (LD), these genes may not be causative but provide potential candidates for further investigation.

The power of XP-GWAS is affected by many factors, including phenotyping precision, size of pools, selection intensity, marker density, and the depth of sequencing. Our results demonstrate that XP-GWAS can tolerate a degree of inaccuracy for the phenotyping data. For example, the KRN phenotypic data used in this study were collected based on rough observations rather than via a systematic field trial design. Nevertheless, XP-GWAS would be expected to have more power if the underlying phenotypic data are more precisely assayed. In addition, simulated power analyses for BSA found that increasing the bulk size with constant selection intensity (5%) could increase the power to detect small effect QTLs (Ehrenreich et al. 2010). Another simulated study suggested BSA would be more powerful if the selection intensity is higher than 10% with sufficient quantitative genotyping (Magwene et al. 2011). For sequencing-based genotyping, Magwene et al. also suggested increasing the depth of sequencing until the depth is bigger than the bulk size. In addition to using a high depth of sequencing, adequate maker density is also critical. XP-GWAS is built on the hypothesis that QTL allele frequencies are

statistically different in different phenotypic pools, which are reflected by the detectable markers in LD with the underlying QTLs. Therefore, the appropriate marker density is determined by the average extent of the LD blocks. Although LD structure is affected by many factors (Flint-Garcia et al. 2003), in general, LD decays rapidly in out-crossing plant species such as maize (ranged 1-10 kb) (Yan et al. 2009) and more slowly in self-pollinating species such as rice (*japonica* ~150 kb and *indica* ~75kb) (Mather et al. 2007) and *Arabidopsis* (250 kb) (Nordborg et al. 2002). In maize with a genome size of 2.3Gb, more than 200,000 to 2M markers (2.3 Gb/1-10 kb) are required to capture most of the genomic variations assuming an LD of 1-10 kb. The 944,549 markers employed in the current study are within this range. Estimates of the number of markers required for other species can be computed similarly.

The above considerations not only determine the power of the study; they can also inform decisions about the appropriate genotyping technologies to be employed for XP-GWAS given the size of the target species' genome, and available resources. For relative small genomes such as cucumber (243.5 Mb) (Huang et al. 2009) and strawberry (~240 Mb) (Shulaev et al. 2011), WGS could detect not only SNPs and short indels, but also present/absent variations (PAVs) and copy number variations (CNVs). For larger, complex genomes, options of reduced representation genotyping include restriction digestion based methods (Van Tassell et al. 2008), RNA-sequencing (Haseneyer et al. 2011), and targeted sequence capture (Bashiardes et al. 2005b). Some of these methods can be applied to species that lack reference genomes. If an RNA-seq based genotyping approach (Barbazuk et al. 2007) is used, loci that exhibit associations to traits could be interpreted within the context of their expression profiling, thus advancing our biological understanding of complex traits.

This study introduced the Zeanome capture library to the maize genetics toolkit. Using the Zeanome Seq-Cap library, it is possible to focus sequencing resources on the non-repetitive portions of this large and repetitive genome.

Although XP-GWAS has significant advantages as compared to other methods of identifying marker-trait associations, it also has some inherent limitations. For

example, as a consequence of the necessity to pool individuals by phenotypes, a separate XP-GWAS experiment must be conducted for each trait of interest. Furthermore, because inferences rely on allele frequencies in populations, it is probably not possible to estimate individual variant effects and heritability via XP-GWAS. The number of marker-trait associations detected by XP-GWAS (N = 145) in this experiment was somewhat lower than the number obtained via a conventional GWAS (N = 260). However, because the two study populations may have different genetic compositions, the absence of complete overlap between the two experiments may be a consequence of population-specific signals. According to the above analysis, the power of XP-GWAS could be increased with greater depth of sequencing, larger and/or better-designed pools and more precise phenotyping. But in general, the pooling associated with XP-GWAS is not expected to yield better performance than conventional GWAS because pooling introduces stochastic effects and uncertainties. This is counter-balanced by the substantial reduction in genotyping cost in XP-GWAS as compared to conventional GWAS.

Conventional GWAS has the problem of false positive signals caused by population structure; this remains an important issue for XP-GWAS. False associations arise if a set of closely related lines are included in one extreme pool and another set of related lines are present in the other extreme pool. To reduce the effects of population structure, we introduced a random pool. Because this pool is a random sample of the population (i.e., the diversity panel), variant frequencies in this pool were treated as estimates of these frequencies in the population. Secondly, a statistical approach widely used in conventional GWAS was adapted to correct the inflation of the test statistic in XP-GWAS. In this method, a genomic control parameter $\lambda$ was defined as the median (or mean) $\chi^2$ association statistic across genome wide markers divided by its theoretical value under the null distribution. A value of $\lambda=1$ indicates no population structure effect; while $\lambda>1$ indicates the existence of some degree of population structure which should be corrected.

In addition to the above approaches, a careful experimental design is recommended to reduce the degree of population structure. Matched sampling (selecting pairs of extreme samples from the same geographic origin) has potential to reduce population structure. However, accessions from the same geographic

region do not necessarily have similar genetic backgrounds as a consequence of for example, migration. To overcome this, ancestry matching through genotyping was proposed in a human case and control study (Crossett et al. 2010). Type I error could be effectively controlled by this method. But this would require genotyping individual samples within the pools, albeit perhaps with only a small number of markers.

By taking advantage of advances in sequencing technologies and the development of the appropriate statistical approaches, XP-GWAS promises to enhance the rate of genetic gain in crops, e.g., by identifying loci that could be used for marker assisted selection (MAS) and allele mining. XP-GWAS can also be used for the discovery of loci that play important roles in ecologically significant wild species, e.g., genes that confer resistance to stresses associated with climate change. Our initial XP-GWAS was conducted in maize because it was possible to compare our results to those obtained from conventional GWAS. The most appropriate targets for XP-GWAS are probably not the major crops such as maize for which extensive previously genotyped diversity panels exist. Instead, XP-GWAS may be most relevant for minor/orphan crops (Collard and Mackill 2008; Varshney et al. 2012), many of which already have large, phenotypically characterized germplasm collections. These existing phenotypic data can be used for XP-GWAS as an efficient and cost effective method to identify loci that control agronomical significant loci.

## Materials and Methods

**Genetic materials, DNA extraction and phenotyping.** Maize germplasm accessions were obtained from NCRPIS (http://www.ars.usda.gov/main/) based on the KRN records in the GRIN (http://www.ars-grin.gov/) database. These accessions were bulked into three phenotypic pools: the high KRN pool (N=226), low KRN pool (N=208) and random KRN pool (N=173) (**Table S1**). Because these accessions are both heterogeneous and heterozygous, tissue for DNA extraction was sampled from 12 plants per accession and pooled. After pooling tissues from all accessions, DNA was extracted using a CTAB method (Clarke 2009).

To test the accuracy of the phenotypic data, 45 accessions (29 from the high KRN pool and 16 from the low KRN pool) were planted in two different locations. Within

each replication, each accession was planted in a row of 12 plants. KRN phenotypes were collected at harvest. Values of KRN were estimated by fitting a mixed linear model (Gilmour et al. 1997), where genotype was fitted as a fixed effect and location was fitted as a random effect. The mixed model was implemented using an R (http://www.R-project.org/) add-on package 'nlme' (Pinheiro et al. 2013).

**Zeanome capture probe design and Exome-sequencing.** A solution-based Zeanome capture library was designed by Roche Nimblegen. This library contains 186,513 probes designed from the 39,656 maize gene models (FGSv2) that comprise ~60 Mb of non-repetitive sequences. These probes were supplemented with 48,303 probes designed from additional sequences expressed in other maize lines and 13,361 control probes that were not used in the current analysis, but that will be described in a subsequent publication (Alina Ott, Alvis Hu, Wei Wu and Patrick Schnable's unpublished results).

Before exome-capturing, indexed adapters (barcodes) were separately added to the three pooled DNA samples following Illumina's TruSeq DNA sample preparation guide. After DNA quantification using Agilent Bioanalyzer and a High Sensitivity DNA Bioanalyzer kit (5067-4626), 300 μg of DNA from each sample were pooled together. Then sequence captures were conducted following the NimbleGen protocol. The captured DNA was quantified on the Bioanalyzer again and diluted to 10 ng/μL. The prepared library was sequenced on Illumina HiSeq2000 machine using the paired-end 100-cycle protocol.

**Genome alignment and variant calling.** As reported in our previous study (Li et al. 2012), the nucleotides of raw reads were scanned for low quality and quality values lower than the threshold were trimmed using a custom pipeline. Trimmed reads were then aligned to the reference genome using GSNAP (Wu and Nacu 2010) as paired-end fragments. The coordinates of confident and single (unique) alignment that passed our filtering criteria were used for SNP and small indel discovery. Polymorphisms at each potential variant site were carefully examined and putative variants were identified (Li et al. 2012; Liu et al. 2012).

**XP-GWAS with genomic control.** To conduct XP-GWAS, a generalized linear model was fitted for each variant using the binomial link function, which was implemented

using 'glm' function in R (R Development Core Team 2010). In the model, a 2x3 matrix composed of reference allele count and alternative allele count as the row and three phenotypic groups as the column was treated as response matrix; and three integers (1, 2 and 3) indicating low, random and high phenotypic pools were treated as explanatory vector. After model fitting, Chi-square statistics were obtained. Due to the non-independence of samples raised from population structure, the normal Chi-square statistic is overly dispersed relative to a non-central Chi-square distribution. To correct for this (Devlin and Roeder 1999), an inflation factor λ was estimated using an R add-on package 'gap' (Jing 2013). Then the Chi-square test statistics were adjusted by λ and used to derive *P* values. Finally, the FDR method (Benjamini and Hochberg 1995b) was used to adjust multiple hypothesis testing problem.

**Fine mapping QTL located in Chr4:169-180Mb interval.** In an earlier study we identified KRN QTL using 291 intermated B73 and Mo17 recombinant inbred lines (Yang et al. submitted). One large effect QTL (effect = 1.3 rows, heritability = 15%) located in the Chr4:169-179Mb interval was selected for fine mapping. Two SNP markers M13783 and M89103 were designed to define the QTL interval. After genotyping a set of IBM RILs, two RILs (M0024 and M0054) which containing Mo17 fragment on the QTL interval were identified. These two RILs were backcrossed to B73 twice to create the F1BC2 mapping population, where the first backcross were carried out after genotyping. On average, the mapping population contained ~6% Mo17 materials in a B73 background. After screening ~6,100 these F1BC2 plants with the two SNP markers, 262 recombinants were initially identified; and they both selfed and outcrossed to their recurrent parent B73. These identified recombinants were further genotyped using 26 SNPs located within the QTL interval to define their recombination break points. For the backcrossed recombinants, 26 heterozygous families with unique break points were chosen and planted in replicated plot trails with 12 plants x 6 replications to collect the KRN phenotype. For the selfed recombinants, 220 homozygous recombinants families were successfully created by further selfing the identified homozygous plants. These homozygous recombinants were phenotyped using a replicated field design with 7 plants x 8 replications.

**Monte Carlo simulation procedures.** A Monte-Carlo simulation procedure (Rubinstein 1981) was implemented to test the hypothesis that the number of overlapping loci between this study and previous results via traditional GWAS has no difference greater than expected by chance. First, the same number of variants was sampled from a set of 0.94M variants to resemble the TAVs. Then the number of overlap TAVs was recorded as the test statistic. This procedure was repeated 10,000 times, and the number of test statistics larger than the observation value was divided by total number of simulation to derive an empirical Monte-Carlo *P* value (Johnson et al. 2011).

## Acknowledgements

We gratefully acknowledge Dr. Wei Wu, Dr. Heng-Cheng "Alvis" Hu and Ms. Talissa Sari for technical assistance, Ms. Lisa Coffey for germplasm management, and the USDA-ARS North Central Region Plant Introduction Station (NCRPIS) for providing germplasm accessions. This research was supported by a grant from the National Science Foundation to P.S.S. and colleagues (IOS-1027527).

## References

Austin DF, Lee M. 1996. Comparative mapping in F-2:3 and F-6:7 generations of quantitative trait loci for grain yield and yield components in maize. *Theor Appl Genet* **92**(7): 817-826.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. 2007. SNP discovery via 454 transcriptome sequencing. *The Plant journal : for cell and molecular biology* **51**(5): 910-918.

Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. 2005a. Direct genomic selection. *Nature methods* **2**(1): 63-69.

-. 2005b. Direct genomic selection. *Nature methods* **2**(1): 63-69.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, Sanmiguel PJ, Bennetzen JL. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet* **5**(11): e1000732.

Beavis WD, Smith OS, Grant D, Fincher R. 1994. Identification of Quantitative Trait Loci Using a Small Sample of Topcrossed and F4 Progeny from Maize. *Crop Sci* **34**(4): 882-896.

Benjamini Y, Hochberg Y. 1995a. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**(1): 289-300.

Benjamini Y, Hochberg Y. 1995b. controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B ….*

Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S, Holland JB, Flint-Garcia S, McMullen MD, Buckler ES et al. 2011. Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet* **7**(11): e1002383.

Clarke JD. 2009. Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harbor protocols* **2009**(3): pdb prot5177.

Collard BC, Mackill DJ. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci* **363**(1491): 557-572.

Crossett A, Kent BP, Klei L, Ringquist S, Trucco M, Roeder K, Devlin B. 2010. Using ancestry matching to combine family-based and unrelated samples for genome-wide association studies. *Stat Med* **29**(28): 2932-2945.

Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* **55**(4): 997-1004.

Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, Shapiro JA, Gresham D, Caudy AA, Kruglyak L. 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**(7291): 1039-1042.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**(5): e19379.

Flint-Garcia SA, Thornsberry JM, Buckler ESt. 2003. Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* **54**: 357-374.

Fu Y, Springer NM, Gerhardt DJ, Ying K, Yeh CT, Wu W, Swanson-Wagner R, D'Ascenzo M, Millard T, Freeberg L et al. 2010. Repeat subtraction-mediated sequence capture from a complex genome. *The Plant journal : for cell and molecular biology* **62**(5): 898-909.

Gilmour AR, Cullis BR, Verbyla AP. 1997. Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* **2**(3): 269-293.

Haseneyer G, Schmutzer T, Seidel M, Zhou R, Mascher M, Schon CC, Taudien S, Scholz U, Stein N, Mayer KF et al. 2011. From RNA-seq to large-scale genotyping - genomics resources for rye (Secale cereale L.). *BMC Plant Biol* **11**: 131.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P et al. 2009. The genome of the cucumber, Cucumis sativus L. *Nature genetics* **41**(12): 1275-1281.

Jing HZ. 2013. gap: Genetic Analysis Package.

Johnson C, Drgon T, Walther D, Uhl GR. 2011. Genomic regions identified by overlapping clusters of nominally-positive SNPs from genome-wide studies of alcohol and illegal substance dependence. *PloS one* **6**(7): e19210.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**(5720): 385-389.

Li X, Zhu C, Yeh CT, Wu W, Takacs EM, Petsch KA, Tian F, Bai G, Buckler ES, Muehlbauer GJ et al. 2012. Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genome Res* **22**(12): 2436-2444.

Liu S, Chen HD, Makarevitch I, Shirmer R, Emrich SJ, Dietrich CR, Barbazuk WB, Springer NM, Schnable PS. 2010. High-throughput genetic mapping of mutants via quantitative single nucleotide polymorphism typing. *Genetics* **184**(1): 19-26.

Liu S, Yeh CT, Tang HM, Nettleton D, Schnable PS. 2012. Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PloS one* **7**(5): e36406.

Magwene PM, Willis JH, Kelly JK. 2011. The statistics of bulk segregant analysis using next generation sequencing. *PLoS Comput Biol* **7**(11): e1002255.

Makarevitch I, Thompson A, Muehlbauer GJ, Springer NM. 2012. Brd1 gene in maize encodes a brassinosteroid C-6 oxidase. *PloS one* **7**(1): e30798.

Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD. 2007. The extent of linkage disequilibrium in rice (Oryza sativa L.). *Genetics* **177**(4): 2223-2232.

Michelmore RW, Paran I, Kesseli RV. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc Natl Acad Sci U S A* **88**(21): 9828-9832.

Morton NE. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet* **7**(3): 277-318.

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ et al. 2002. The extent of linkage disequilibrium in Arabidopsis thaliana. *Nature genetics* **30**(2): 190-193.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. 2013. nlme: Linear and Nonlinear Mixed Effects Models.

R Development Core Team. 2010. R: A language and environment for statistical computing.

Rubinstein RY. 1981. *Simulation and the Monte Carlo method*. Wiley, New York.

Schnable PS Ware D Fulton RS Stein JC Wei F Pasternak S Liang C Zhang J Fulton L Graves TA et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**(5956): 1112-1115.

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen JE, Weigel D, Andersen SU. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature methods* **6**(8): 550-551.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP et al. 2011. The genome of woodland strawberry (Fragaria vesca). *Nature genetics* **43**(2): 109-116.

Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. 2006. Whole-genome genotyping with the single-base extension assay. *Nature methods* **3**(1): 31-33.

Toledo FHRB, Ramalho MAP, Abreu GB, de Souza JC. 2011. Inheritance of kernel row number, a multicategorical threshold trait of maize ears. *Genet Mol Res* **10**(3): 2133-2139.

Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature methods* **5**(3): 247-252.

Varshney RK, Ribaut JM, Buckler ES, Tuberosa R, Rafalski JA, Langridge P. 2012. Can genomics boost productivity of orphan crops? *Nat Biotechnol* **30**(12): 1172-1176.

Veldboom LR, Lee M, Woodman WL. 1994. Molecular Marker-Facilitated Studies in an Elite Maize Population .1. Linkage Analysis and Determination of Qtl for Morphological Traits. *Theor Appl Genet* **88**(1): 7-16.

Vigouroux Y, Glaubitz JC, Matsuoka Y, Goodman MM, Sanchez G J, Doebley J. 2008. Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. *American Journal of Botany* **95**(10): 1240-1253.

Wenger JW, Schwartz K, Sherlock G. 2010. Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from Saccharomyces cerevisiae. *PLoS Genet* **6**(5): e1000942.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**(7): 873-881.

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J. 2009. Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PloS one* **4**(12): e8451.

Yi G, Lauter AM, Scott MP, Becraft PW. 2011. The thick aleurone1 mutant defines a negative regulation of maize aleurone cell fate that functions downstream of defective kernel1. *Plant Physiology* **156**(4): 1826-1836.

## Figure Legends

**Figure 1. KRN phenotype of diverse germplasm accessions.** Histogram distribution and density plot of germplasm accessions (N = 6,952) in GRIN database and density plots of three selected KRN phenotypic pools of low (blue), high (red) and random (blue).

**Figure 2. Histograms of depth of sequencing and coverage for filtered gene set (FGSv2) of three phenotypic pools.** In the left panels (A, B and C), vertical dashed lines indicate 50X depth sequencing; in the right panels (D, E and F), vertical dashed lines indicate 90% of gene coverage.

**Figure 3. Manhattan plot of XP-GWAS results.** Horizontal red dash line indicates the 5% FDR threshold.

**Figure 4. XP-GWAS and independent pairwise Chi-square test results in the region of Chr4:175-176Mb**. Panel (A) shows the XP-GWAS results; panel (B), (C) and (D) are the independent pairwise Chi-square tests of high KRN pool vs. low KRN pool, high KRN pool vs. random KRN pool and low KRN pool vs. random KRN pool, respectively. Red dashed lines indicate the threshold of FDR < 0.05. Panel (E) shows the read depth of three KRN pools using a bin size = 1000, where blue lines are for high KRN pool, red lines are for low KRN pool and yellow lines are for random KRN pool.

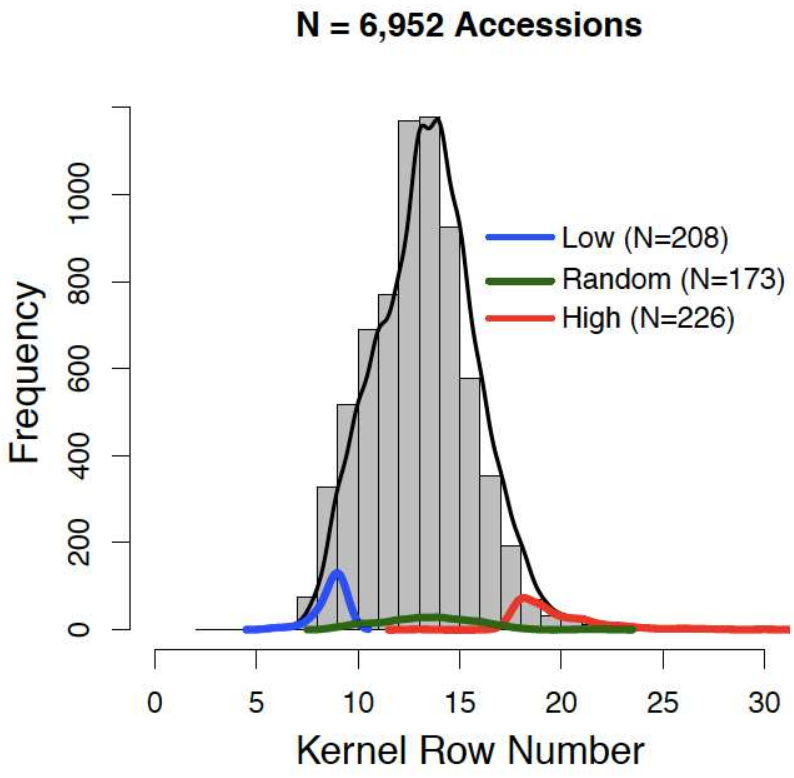
N = 6,952 Accessions
Frequency
0
200
400
600
800
1000
Low (N=208)
Random (N=173)
High (N=226)
0
5
10
15
20
25
30
Kernel Row Number

**Figure 1**

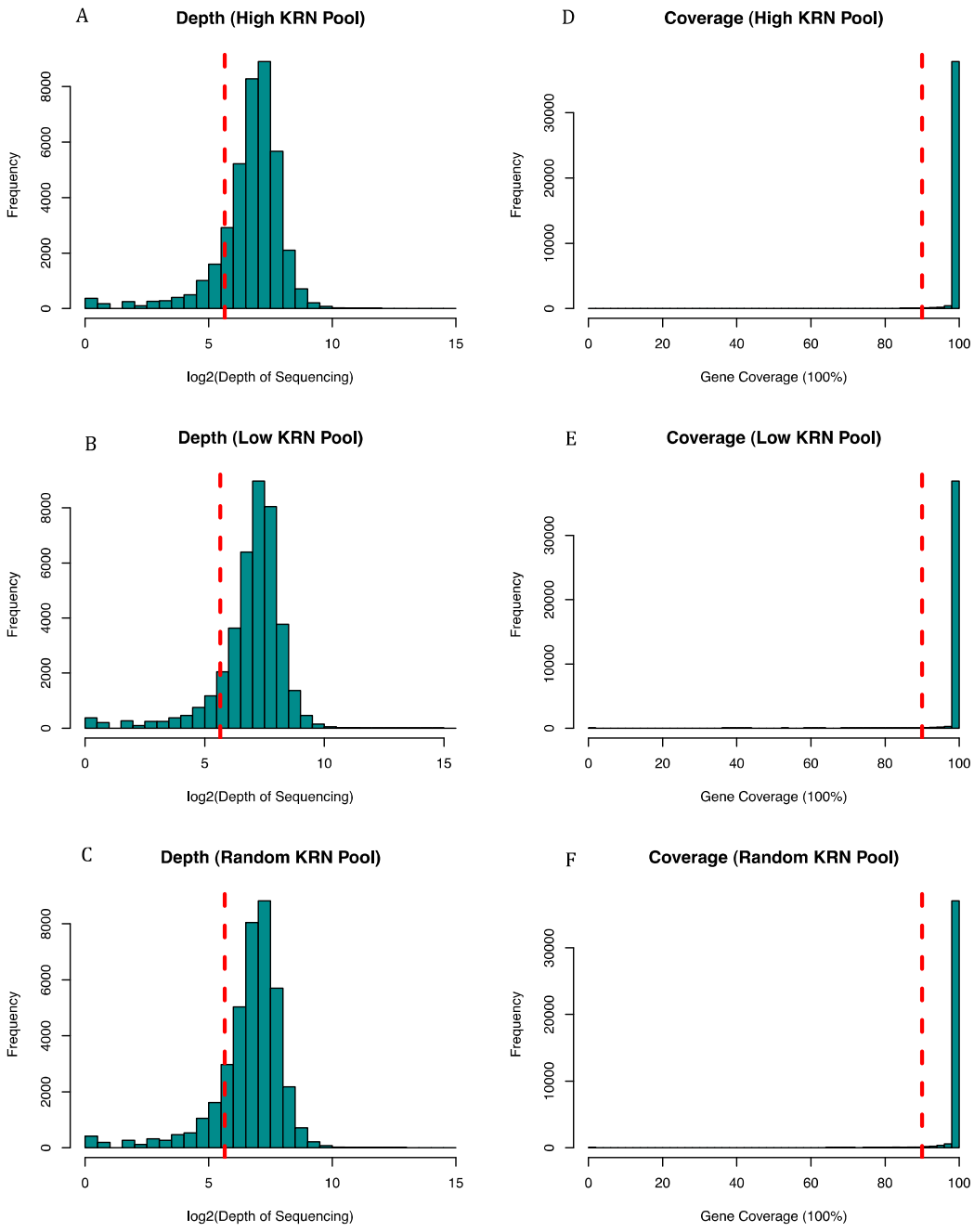
A
Depth (High KRN Pool)
Frequency
log2(Depth of Sequencing)
D
Coverage (High KRN Pool)
Frequency
Gene Coverage (100%)
B
Depth (Low KRN Pool)
Frequency
log2(Depth of Sequencing)
E
Coverage (Low KRN Pool)
Frequency
Gene Coverage (100%)
C
Depth (Random KRN Pool)
Frequency
log2(Depth of Sequencing)
F
Coverage (Random KRN Pool)
Frequency
Gene Coverage (100%)

**Figure 2**

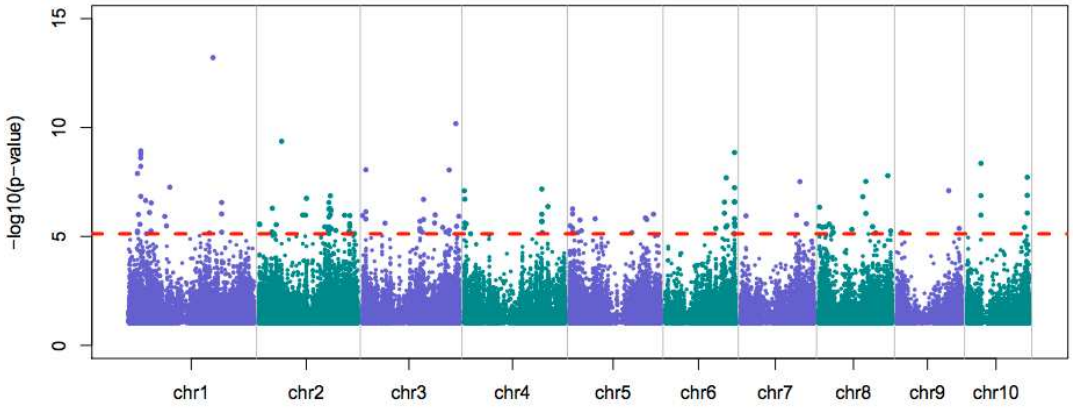
−log10(p−value)
0
5
10
15
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10

**Figure 3**

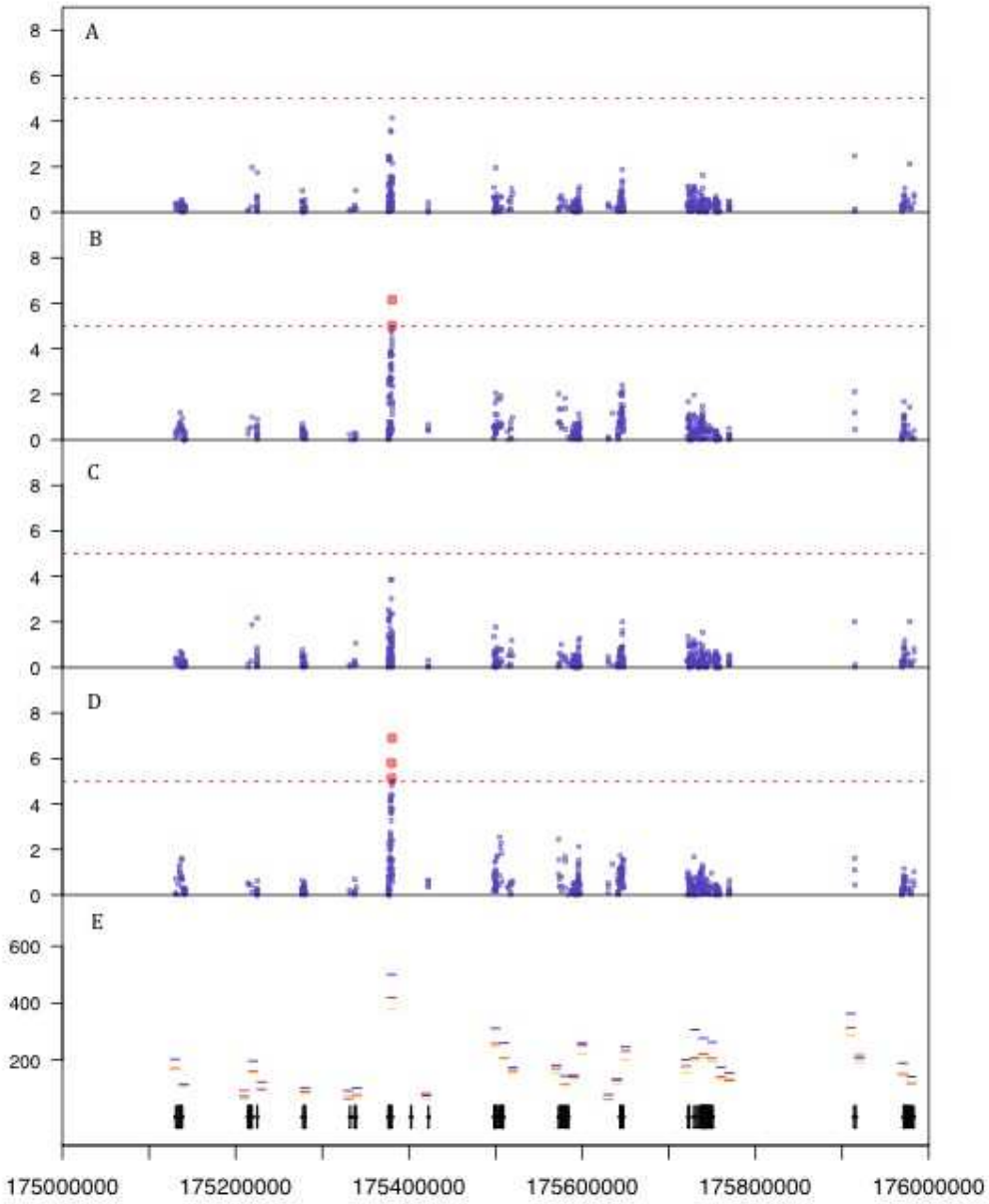
A
B
C
D
E
8
6
4
2
0
600
400
200
175000000
175200000
175400000
175600000
175800000
176000000

**Figure 4**

# CHAPTER 5. GENERAL CONCLUSIONS

Identification of trait-associated variants (TAVs) is the first step in investigating genetic determinants of quantitative traits and biotechnological interventions. Over the last few decades, several approaches were developed to detect TAVs, such as QTL linkage mapping (Morton 1955), bulked segregant analysis (BSA) (Michelmore et al. 1991), and recently developed Genome-wide Association Studies (GWAS) (Klein et al. 2005). The pros and cons of the above methods were discussed in Chapter 4. The reducing cost of genotyping enabled by NGS technologies greatly facilitates the applications of GWAS in identifying genetic controls of agronomical important traits. In Chapter 2, we collected kernel row number (KRN) trait data from a set of 6,230 maize lines. Using these data, three distinct approaches for conducting GWAS were compared: 1) single-variant, 2) stepwise regression, and 3) Bayesian-based multi-variant approaches. By analyzing subset of the identified TAVs using three unrelated populations that were not included in the GWAS; approximately 50% of the successfully genotyped TAVs were cross-validated in at least one unrelated population. Importantly, ~60% of the cross-validated TAVs were identified by only one of the three statistical approaches, demonstrating that the three approaches were complementary. To the best of our knowledge, this is the first time in plant studies to provide a lower bound for true positive GWAS results. This is particularly important given that many plant studies using GWAS have been reported in the literature.

In addition, the cross-validated TAVs identified in Chapter 2 provided useful information to enhance our understanding of the maize domestication and the developmental steps involved in ear development. For example, we found that allele frequencies of KRN associated variants were higher in maize compared with that of the teosinte, the wild ancestor of maize, and this is consistent with changes in morphological development that occurred during domestication. Meanwhile, two

biological pathways were enriched by surveying genes near these TAVs, which also provide insights into the understanding of biological processes during ear development.

In Chapter 3, six other yield-related traits in addition to KRN were analyzed via GWAS and their genetic architectures were dissected. Because these seven yield-related traits exhibit varied levels of heterosis, the GWAS results provided us the opportunity to compare different patterns among their genetic components. Importantly, by studying the traits using two multi-variant approaches, we found that the phenotypic variance explained by genome-wide markers were negatively correlated with the levels of heterosis. In contrast, the level of heterosis for a trait was positively correlated with the number and magnitudes of TAVs that exhibiting dominance gene action. These findings provide strong support for the view that heterosis is at least a partial consequence of the positive dominant gene action. Furthermore, adding dominance when conducting GWAS was able to recover up to 45% of the missing heritability, and the cumulative effects of both additive and dominant gene action could account for 86%-96% of the heritability for the seven traits.

Genomic selection has been recently introduced in plant breeding. As compared to conventional marker-assisted selection (MAS) using only the significant markers, genomic selection is able to estimate all the markers across the genome regardless of their magnitudes of effects (Jannink et al. 2010). Theoretically, it has more power than conventional method, although empirical large-scale genomic selection experiments have not been reported in plants yet. In Chapter 3, simulation studies for genomic selection were conducted with the implementation of the dominant effects using 'BayesC'. The results showed that the prediction accuracies ranged from 0.8 to 0.9 using only the additive model, while it increased to 0.85 to 0.95 after adding the dominance model. A livestock study with 3,500 Holstein bulls achieved

accuracies of 0.44 to 0.79 for traits with different heritability (VanRaden et al. 2009). The accuracies achieved in our simulations were relative higher compared to this study, probably due to the high-density markers we have employed (~13M variants). However, it is worth to note that the phenotypic values in our study were observed in replicated trials and adjusted after applying the mixed linear model.

Although QTL mapping and genome-wide association study (GWAS) were well developed, such as the QTL and the GWAS approaches used in Chapter 2 and Chapter3, it remains challenging to detect trait-marker association expeditiously and cost-efficiently, especially for species without well-established mapping population. In Chapter 4, an XP-GWAS method was developed as an attempt to meet these needs. XP-GWAS interrogates allele frequencies in different extreme phenotypic pools and tests their associations with the trait. This method was applied to the KRN trait in maize. After comparisons among pools, 145 TAVs were identified as being associated with the KRN phenotype. These trait-associated variants were significantly enriched in regions identified by a conventional GWAS. We also demonstrated the high resolution of XP-GWAS by resolving linked QTLs and detecting TAVs within a single gene under a QTL peak.

To date, at least 55 plant genomes have been sequenced, which represent ~50 different species (Michael and Jackson 2013). To meet the increasing demands of food, clothing and energy in the future, breeding plants better adapted to human needs would be necessary. Establishment of trait-marker association is the first step towards this goal. In this study, we explored QTL linkage mapping (Chapter 2 and Chapter 3), QTL cloning (Chapter 4), GWAS (Chapter 2 and Chapter 3) and genomic selection (Chapter 4) in the model species maize. And we also developed a high-resolution mapping method (termed as XP-GWAS) for a rapid trait-marker association (Chapter 4). The knowledge learned here will shine a light on other plant species. For example, the XP-GWAS will be of particular valuable for detecting

genes or alleles responsible for quantitative variation in minor crops, such as cucumber or papaya, which do not have extensive genotyping resources. In addition, the comparisons of the genetic architecture controlling seven yield-related traits provide insights into heterosis and missing heritability. This knowledge will be of great interests across species.

Even with the big progress been made during the past few decades, there is still a long way to go before fully understanding of the functional features of the plant genomes. As we can learn from the human ENCODE projects (Consortium 2012), the NGS enabled "omics" approaches will add increasing number of layers to the genome. Combining all layers of genomic information with advanced phenotyping technologies, we hope in the near future, GWAS or XP-GWAS will help to better annotate the functional aspects of the plant genomes, and genomic selection will assist breeders to delivery valuable varieties to the world.

## References

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Jannink JL, Lorenz AJ, Iwata H. 2010. Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* **9**(2): 166-177.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**(5720): 385-389.

Michael TP, Jackson S. 2013. The First 50 Plant Genomes. *Plant Genome-Us* **6**(2).

Michelmore RW, Paran I, Kesseli RV. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc Natl Acad Sci U S A* **88**(21): 9828-9832.

Morton NE. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet* **7**(3): 277-318.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci* **92**(1): 16-24.

## APPENDIX A. SUPPLEMENTAL INFORMATION FOR CHAPTER 2

**Figure S1** High parent heterosis (HPH)(A) and mid-parent heterosis (MPH)(B) of the KRN trait in three hybrid populations. Vertical dashed lines indicate no heterosis.

**Figure S2** Phenotypic distribution of the B73, Mo17 and their reciprocal $F_1$ hybrids. Replicated observations of B73 x Mo17 (KNR = 15.5 ± 1.4) and Mo17 x B73 (KRN = 15.6 ± 1.4) exhibited similar mean and variance of phenotypic distributions, where the mean values of KRN fall in between of the mean KRN values from B73 (KRN = 17.1) and Mo17 (KRN = 10.8). Density curves were smoothed with default bandwidth parameters.

**Figure S3** Venn diagram comparing KAV bins identified by the three different approaches. After applying a variant thinning procedure, 192 100-kb bins (257 variants), 296 bins (300 variants) and 343 bins (442 variants) were identified using arbitrary thresholds for the single-variant, stepwise regression and Bayesian-based multi-variant approaches, respectively. Note in the parentheses listed the unique KAV numbers in the bins.

**Figure S4** Single variant effect and heritability of the 231 KAVs. In panel (A) and (B), histogram distributions of single variant effect and heritability were computed for the 231 KAVs individually. Note that because of linkage disequilibrium (LD), effect and heritability might be repeatedly calculated.

**Figure S5** Characteristics of the ~13M variants and 231 KAVs. In upper panels (A, B and C), pie charts show the physical distance to the nearest gene, origin and type of variation of the ~13M variants used for the GWAS, respectively; in lower panels (D, E and F), show the same characteristics of 231 KAVs. The star indicates significant difference after Chi-square tests.

**Table S1** KRN trait data of the 6,230 lines.

**Table S2** Separate QTL results of IBM and NAM RILs.

**Table S3** Four QTL regions exhibit opposite effects in different NAM RIL subpopulations.

**Table S4** Joint QTL results of 25 NAM RIL subpopulations.

**Table S5** KAVs identified by three GWAS approaches. Column "value" are significant measurements of different approaches, where $-\log_{10}(P)$ values for single-variant approach, *F*-test statistic values for stepwise regression approach, and posterior model frequencies (MF) for Bayesian-based multi-variant approach. Binsize = 100-kb.

**Table S6** Effects of KAVs in the four QTL regions. In QTL regions of order 1 and 2, both positive and negative effects were observed for identified KAVs.

**Table S7** The 231 KAVs selected for cross-validation. Note some of the KAVs were identified by multiple approaches.

**Table S8** Favorable allele frequencies and statistical test results of historical lines.

**Table S9** Cross-validation samples.

**Table S10** Genotypic data for cross-validation population of elite inbred lines. Genotypes were coded using 3 for reference-like, 2 for heterozygotes, 1 for non-reference-like variant types and 0 for missing.

**Table S11** Genotypic data for cross-validation population of USDA PI accessions. Genotypes were coded the same as Table S10.

**Table S12** Genotypic data for cross-validation population of BSLE**.** Genotypes were coded the same as Table S10.

**Table S13** Summary of cross-validation results. Columns of "Elite1_Qvalue" and "Elite2_Qvalue" are cross-validation results using elite inbred population with and without population structure control. Columns of "USDA_Qvalue" and "BSLE_Qvalue" are cross-validation results using USDA germplasm accessions and BSLE population. Columns "Elite1_DOE", "Elite2_DOE", "USDA_DOE" and "BSLE_DOE" are products of direction of effects (DOE) of KAVs in the GWAS population and the

cross-validation populations, where positive numbers indicated consistency and negative number indicated inconsistency of DOE. “NA” in the table indicated the data are not available. Significant KAVs are indicated by stars (*) in the column of “SNPID”.

**Table S14** List of candidate genes within KAV-associated chromosomal bins that aligned to evidence supported genes.

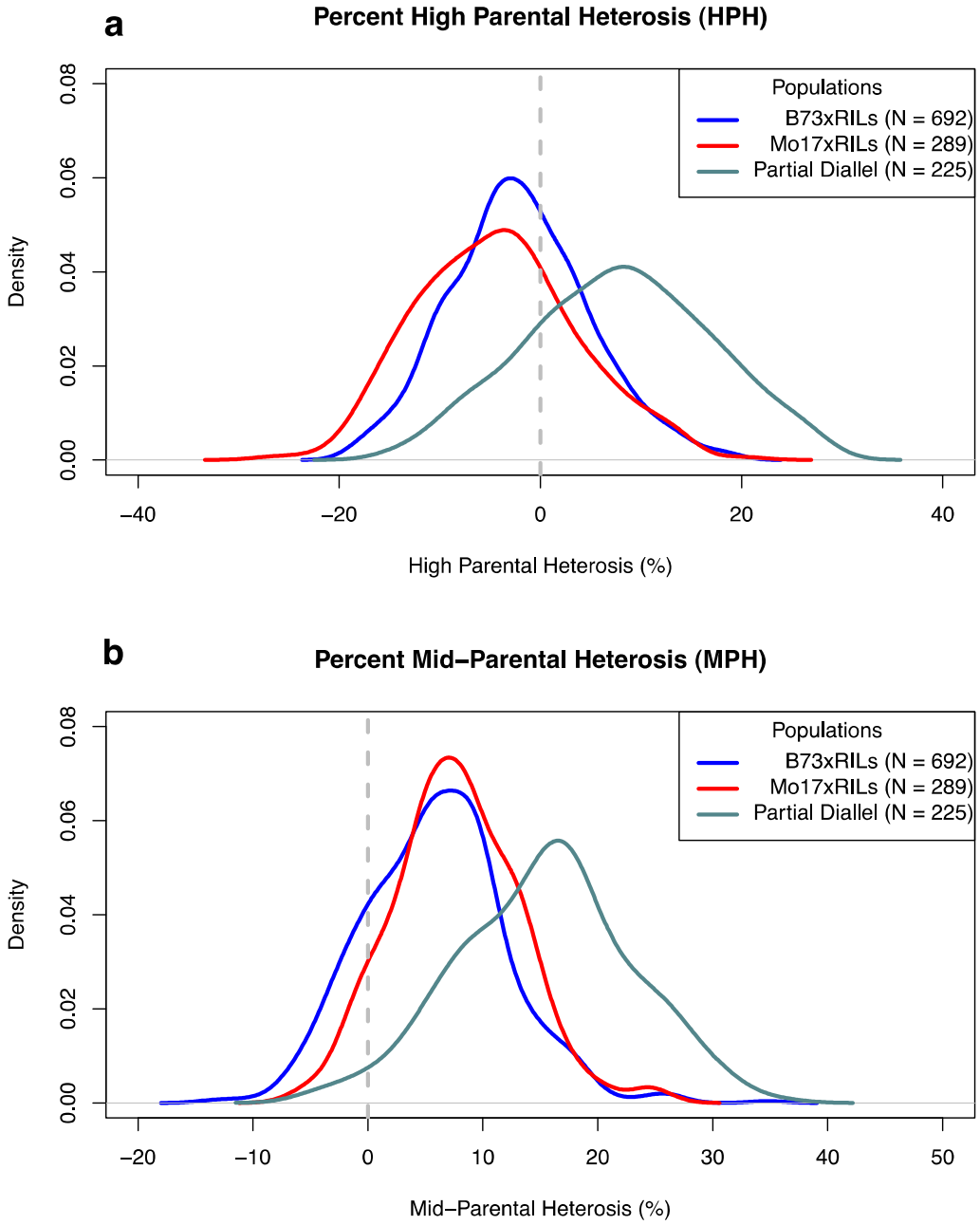
a
Percent High Parental Heterosis (HPH)
Populations
B73xRILs (N = 692)
Mo17xRILs (N = 289)
Partial Diallel (N = 225)
Density
0.00
0.02
0.04
0.06
0.08
−40
−20
0
20
40
High Parental Heterosis (%)
b
Percent Mid−Parental Heterosis (MPH)
Populations
B73xRILs (N = 692)
Mo17xRILs (N = 289)
Partial Diallel (N = 225)
Density
0.00
0.02
0.04
0.06
0.08
−20
−10
0
10
20
30
40
50
Mid−Parental Heterosis (%)

**Figure S1**

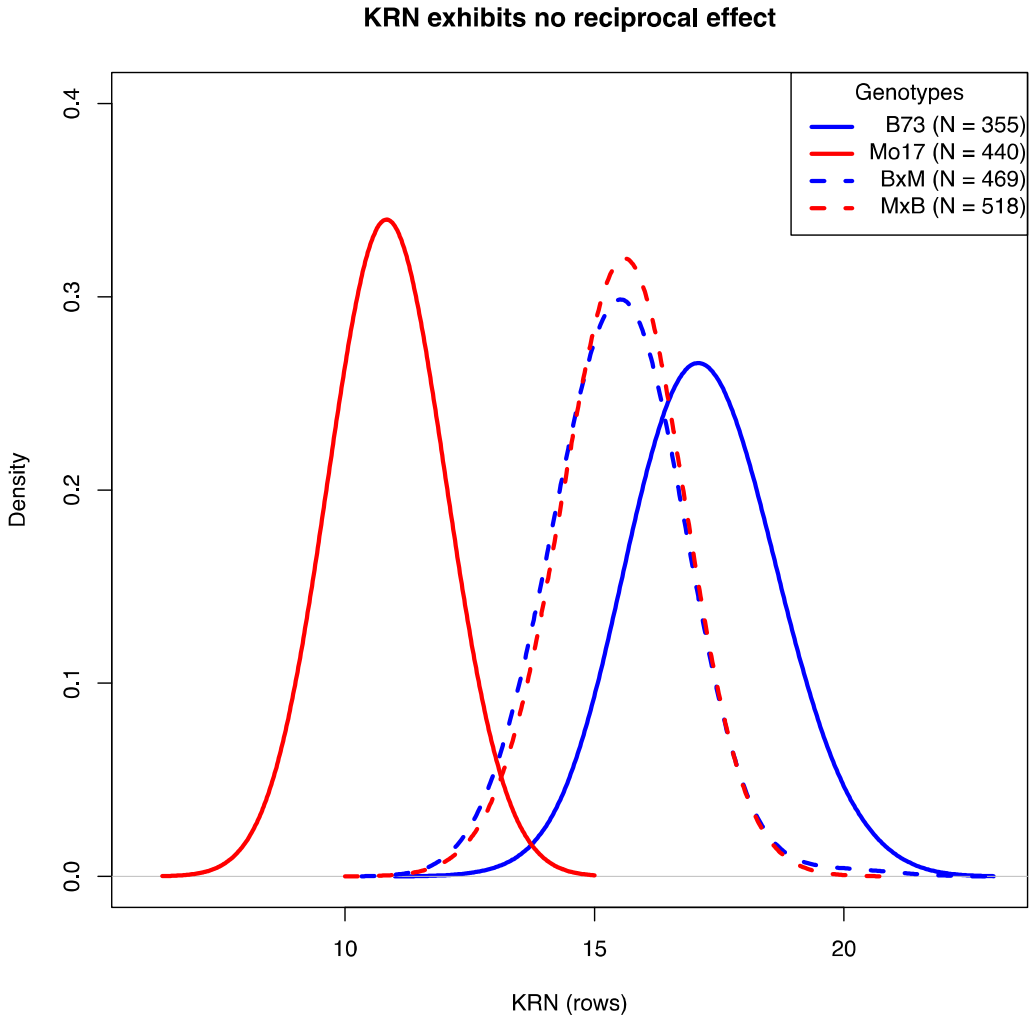
KRN exhibits no reciprocal effect
Genotypes
B73 (N = 355)
Mo17 (N = 440)
BxM (N = 469)
MxB (N = 518)
Density
0.0
0.1
0.2
0.3
0.4
10
15
20
KRN (rows)

**Figure S2**

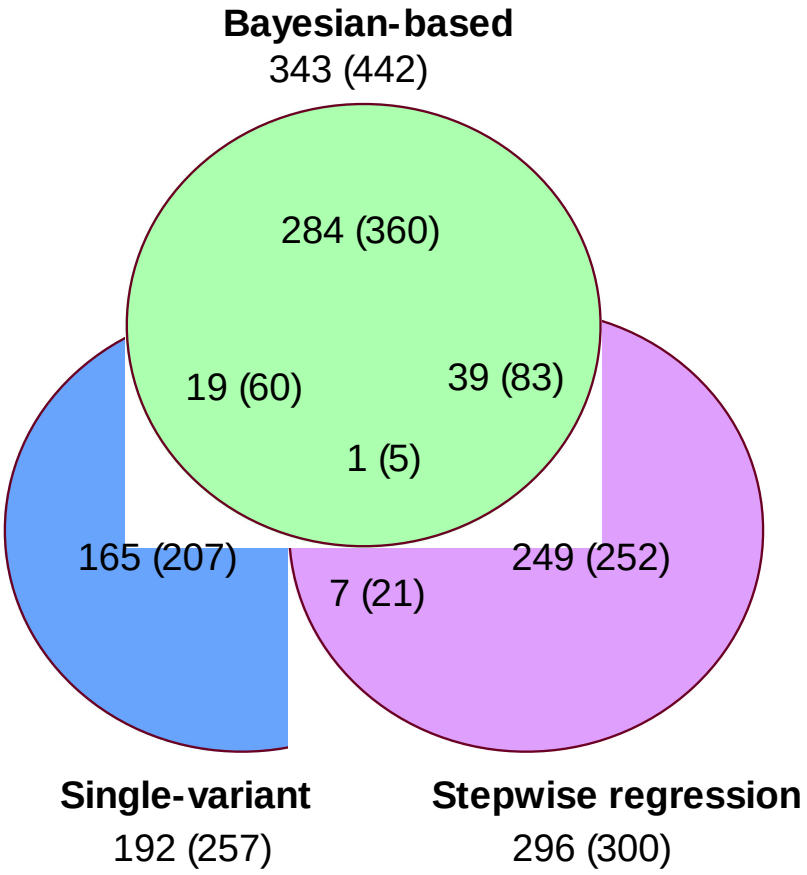
Bayesian-based
343 (442)
284 (360)
19 (60)
39 (83)
1 (5)
165 (207)
249 (252)
7 (21)
Single-variant
192 (257)
Stepwise regression
296 (300)

**Figure S3**

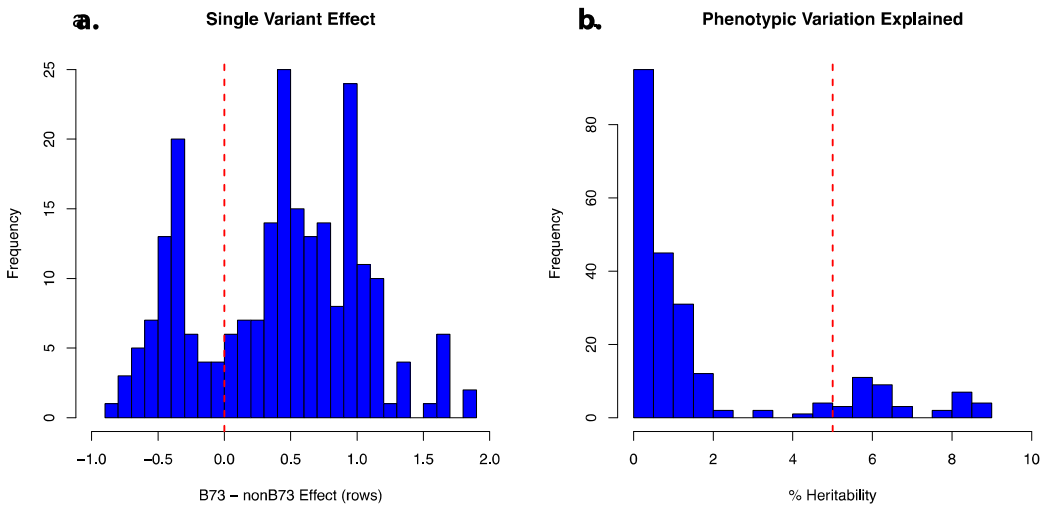
a.
Single Variant Effect
Frequency
0
5
10
15
20
25
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
B73 − nonB73 Effect (rows)
b.
Phenotypic Variation Explained
Frequency
0
20
40
60
80
0
2
4
6
8
10
% Heritability

**Figure S4**

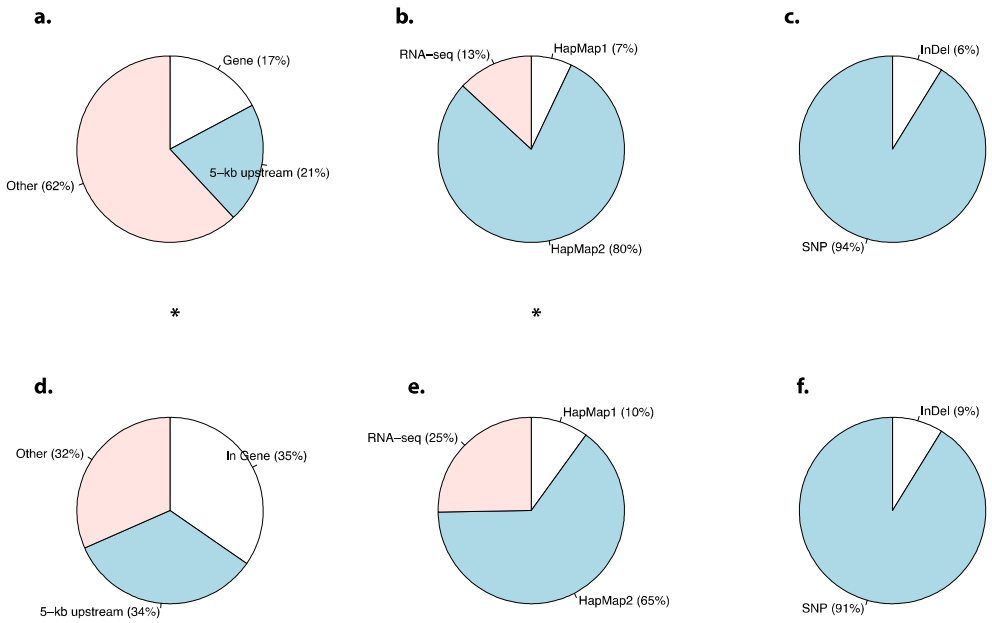
a.
Gene (17%)
5–kb upstream (21%)
Other (62%)
b.
RNA–seq (13%)
HapMap1 (7%)
HapMap2 (80%)
c.
InDel (6%)
SNP (94%)
*
*
d.
In Gene (35%)
Other (32%)
5–kb upstream (34%)
e.
HapMap1 (10%)
RNA–seq (25%)
HapMap2 (65%)
f.
InDel (9%)
SNP (91%)

**Figure S5**

**Table S1** KRN trait data of the 6,230 lines.

*See separate file.*

**Table S2** Separate QTL results of IBM and NAM RILs.

*See separate file.*

**Table S3** Four QTL regions exhibit opposite effects in different NAM RIL subpopulations.

| Order | Subpopulation | Marker | Chr | Genetic [1] (cM) | Physical [2] (Mb) | LOD | Confidence Interval (Mb) | Effect [3] |
|---|---|---|---|---|---|---|---|---|
| 1 | IL14H | PZA03551.1 | 1 | 24.4 | 12.2 | 3.6 | 6.2 - 77.2 | 1.17 |
| | MS71 | PZA02094.9 | 1 | 31.7 | 15.8 | 4.9 | 11 - 26.2 | -0.69 |
| | KI11 | PZA01030.1 | 1 | 36.3 | 17.7 | 5 | 15 - 28.4 | -1.08 |
| | CML103 | PZB01957.1 | 1 | 47.2 | 26.2 | 6.4 | 22.6 - 34.7 | -0.86 |
| | B97 | PZA02490.1 | 1 | 52.7 | 32.6 | 3.3 | 26.2 - 45.5 | -0.68 |
| | KI3 | PZA02490.1 | 1 | 53.7 | 32.6 | 3.6 | 22.9 - 50.1 | 0.9 |
| 2 | CML322 | PZA02014.3 | 1 | 124.2 | 213.2 | 5.7 | 204.3 - 223 | -0.82 |
| | OH43 | PZA01039.1 | 1 | 125.6 | 213.7 | 4.3 | 204.8 - 217.6 | 0.83 |
| | M162W | PZA00658.21 | 1 | 132.9 | 225.1 | 4.6 | 204.8 - 231.7 | -0.58 |
| 3 | CML69 | PZA01530.1 | 5 | 58.3 | 38.5 | 4 | 11.5 - 58.6 | 0.43 |
| | P39 | PHM12992.5 | 5 | 58.4 | 39.2 | 5.2 | 15.1 - 58.2 | 1.05 |
| | CML322 | PZA02207.1 | 5 | 60.6 | 49.9 | 5 | 38.5 - 60.8 | -0.5 |
| | CML322 | PZA01303.1 | 5 | 66 | 73.1 | 5.2 | 60.8 - 84.3 | -0.82 |
| | KY21 | PZA02862.3 | 5 | 66.8 | 75.9 | 4.2 | 49.9 - 84.3 | 0.79 |
| 4 | IL14H | PHM4341.42 | 10 | 46.7 | 117.6 | 4.1 | 98.7 - 130.1 | -1.02 |
| | CML322 | PZA01141.1 | 10 | 47.6 | 120.8 | 3.3 | 72.8 - 130.1 | 1 |
| | CML247 | PZA01005.1 | 10 | 49.2 | 124.9 | 4.6 | 111.8 - 132.6 | 0.62 |
| | MO18W | PZA00647.9 | 10 | 52.2 | 130.1 | 3.7 | 124.5 - 140 | 0.78 |
| | CML333 | PZA02320.1 | 10 | 56.1 | 132.6 | 14.1 | 124.5 - 136.3 | 1.37 |
| | M162W | PZB01111.8 | 10 | 61.4 | 134.4 | 3.5 | 124.9 - 142.2 | 0.67 |

[1] Genetic positions according to IBM[31] and NAM[32] genetic map.
[2] Physical positions according to B73 RefGen_v2.
[3] QTL effects were calculated by using B73 subtracting non-B73 alleles.

**Table S4** Joint QTL results of 25 NAM RIL subpopulations.

*See separate file.*

**Table S5** KAVs identified by three GWAS approaches. Column "value" are significant measurements of different approaches, where $-\log_{10}(P)$ values for single-variant approach, *F*-test statistic values for stepwise regression approach, and posterior model frequencies (MF) for Bayesian-based multi-variant approach. Binsize = 100-kb.

*See separate file.*

**Table S6** Effects of KAVs in the four QTL regions. In QTL regions of order 1 and 2, both positive and negative effects were observed for identified KAVs.

*See separate file.*

**Table S7** The 231 KAVs selected for cross-validation. Note some of the KAVs were identified by multiple approaches.

*See separate file.*

**Table S8** Favorable allele frequencies and statistical test results of historical lines.

*See separate file.*

**Table S9** Cross-validation samples.

*See separate file.*

**Table S10** Genotypic data for cross-validation population of elite inbred lines. Genotypes were coded using 3 for reference-like, 2 for heterozygotes, 1 for non-reference-like variant types and 0 for missing.

*See separate file.*

**Table S11** Genotypic data for cross-validation population of USDA PI accessions. Genotypes were coded the same as **Table S10**.

*See separate file.*

**Table S12** Genotypic data for cross-validation population of BSLE**.** Genotypes were coded the same as **Table S10**.

*See separate file.*

**Table S13** Summary of cross-validation results. Columns of "Elite1_Qvalue" and "Elite2_Qvalue" are cross-validation results using elite inbred population with and without population structure control. Columns of "USDA_Qvalue" and "BSLE_Qvalue" are cross-validation results using USDA germplasm accessions and BSLE population. Columns "Elite1_DOE", "Elite2_DOE", "USDA_DOE" and "BSLE_DOE" are products of direction of effects (DOE) of KAVs in the GWAS population and the cross-validation populations, where positive numbers indicated consistency and negative number indicated inconsistency of DOE. "NA" in the table indicated the data are not available. Significant KAVs were indicated by stars (*) in the column of "SNPID".

*See separate file.*

**Table S14** List of candidate genes within KAV-associated chromosomal bins that aligned to evidence supported genes.

| Gene ID[1] | Chr[2] | Start[2] | End[2] | Query[3] | Gene Set[3] |
|---|---|---|---|---|---|
| GRMZM2G035243 | 1 | 15168789 | 15171560 | SAUR23 | auxin |
| GRMZM2G303463 | 2 | 227264500 | 227268560 | ARF12 | auxin |
| GRMZM2G412085 | 3 | 2811806 | 2813735 | SAUR23 | auxin |
| GRMZM2G168704 | 3 | 5823074 | 5823678 | crr1 | cytokinin |
| GRMZM2G177220 | 3 | 165897260 | 165901223 | ARR11 | cytokinin |
| GRMZM2G033359 | 3 | 194424615 | 194428239 | GH3.8 | auxin |
| GRMZM2G176841 | 3 | 201182487 | 201185704 | AUX1 | auxin |
| GRMZM2G471304 | 4 | 236378470 | 236379262 | SAUR14 | auxin |
| GRMZM2G113135 | 5 | 22998056 | 22998869 | SAUR56 | auxin |
| GRMZM2G053338 | 7 | 162669667 | 162672151 | GH3.8 | auxin |
| GRMZM2G143187 | 9 | 154096524 | 154099391 | SAUR23 | auxin |
| GRMZM2G456644 | 10 | 141552329 | 141553228 | SAUR11 | auxin |

[1] Candidate genes located within KAV-associated chromosomal bins and hit by evidence-supported genes after BLAST.

[2] Physical positions according to B73 RefGen_v2.

[3] Query genes and gene sets they belong to.

## APPENDIX B. SUPPLEMENTAL INFORMATION FOR CHAPTER 3

**Figure S1** Phenotypic distributions of seven yield-related traits in four GWAS populations. Panels (A), (B), (C), (D), (E), (F) and (G) are KRN, CD, AKW, CL, CW, KC and TKW, respectively.

**Figure S2** Pairwise correlations of the seven yield-related traits. The upper right panels show the scatter plots of the pairwise correlations, the red lines are the fitted smooth curves; the lower left panels show correlation coefficients (*r*), asterisks (*) indicate that the correlation coefficients are statistically significant based on Pearson correlation tests ($P < 0.05$).

**Figure S3** Synergistic effect plot for a pair of negatively correlated traits of KRN vs. AKW.

**Figure S4** Synergistic effect plot for a pair of negatively correlated traits of KRN vs. CL traits.

**Figure S5** Synergistic effect plot for a pair of negatively correlated traits of CD and CL.

**Table S1** Phenotypic values for seven yield-related traits of 6,230 lines.
*See separate file.*

**Table S2** Separate linkage analyses identified 425 QTLs for the six yield-related traits. The 126 QTLs controlling for KRN trait were not included.

*See separate file.*

**Table S3** Synergistic effect QTLs for the three pairs of negatively correlated traits.

*See separate file.*

**Table S4** Table of pleiotropic QTLs.

*See separate file.*

**Table S5** Joint linkage analyses identified 128 QTLs for the six yield-related traits. The 28 joint QTLs controlling for KRN trait were not included.

*See separate file.*

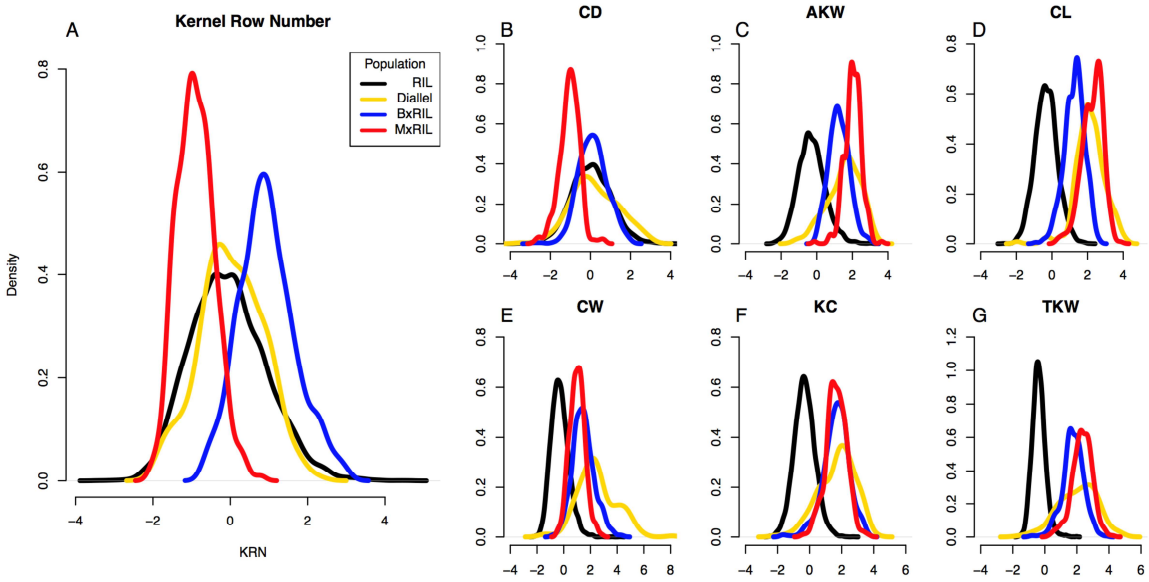
A
Kernel Row Number
Population
RIL
Diallel
BxRIL
MxRIL
Density
KRN
B
CD
C
AKW
D
CL
E
CW
F
KC
G
TKW

**Figure S1**

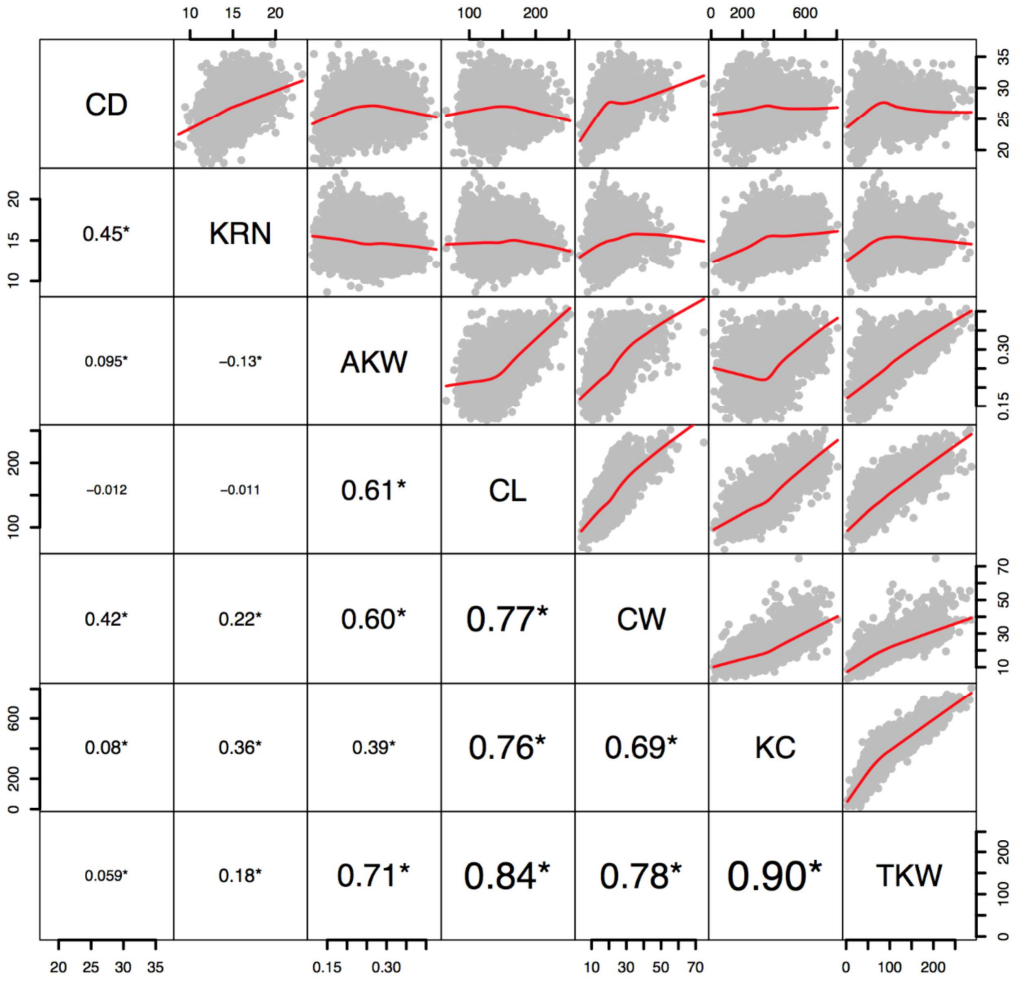
CD
KRN
AKW
CL
CW
KC
TKW
0.45*
0.095*
−0.13*
−0.012
−0.011
0.61*
0.42*
0.22*
0.60*
0.77*
0.08*
0.36*
0.39*
0.76*
0.69*
0.059*
0.18*
0.71*
0.84*
0.78*
0.90*
10 15 20
100 200
0 200 600
20 25 30 35
0.15 0.30
10 30 50 70
0 100 200

**Figure S2**

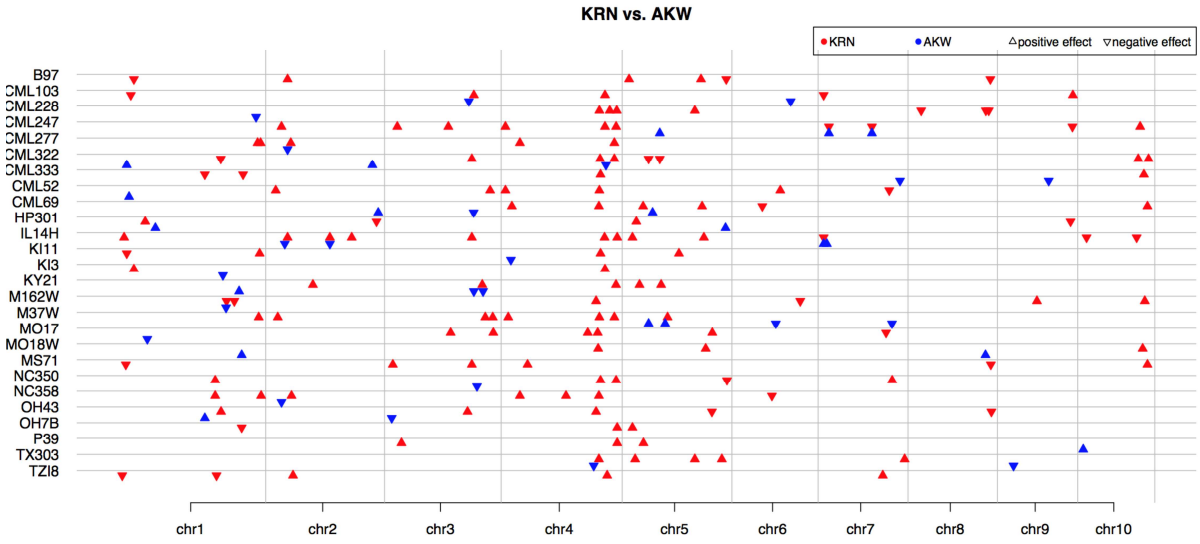
KRN vs. AKW
KRN
AKW
positive effect
negative effect
B97
CML103
CML228
CML247
CML277
CML322
CML333
CML52
CML69
HP301
IL14H
KI11
KI3
KY21
M162W
M37W
MO17
MO18W
MS71
NC350
NC358
OH43
OH7B
P39
TX303
TZI8
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10

**Figure S3**

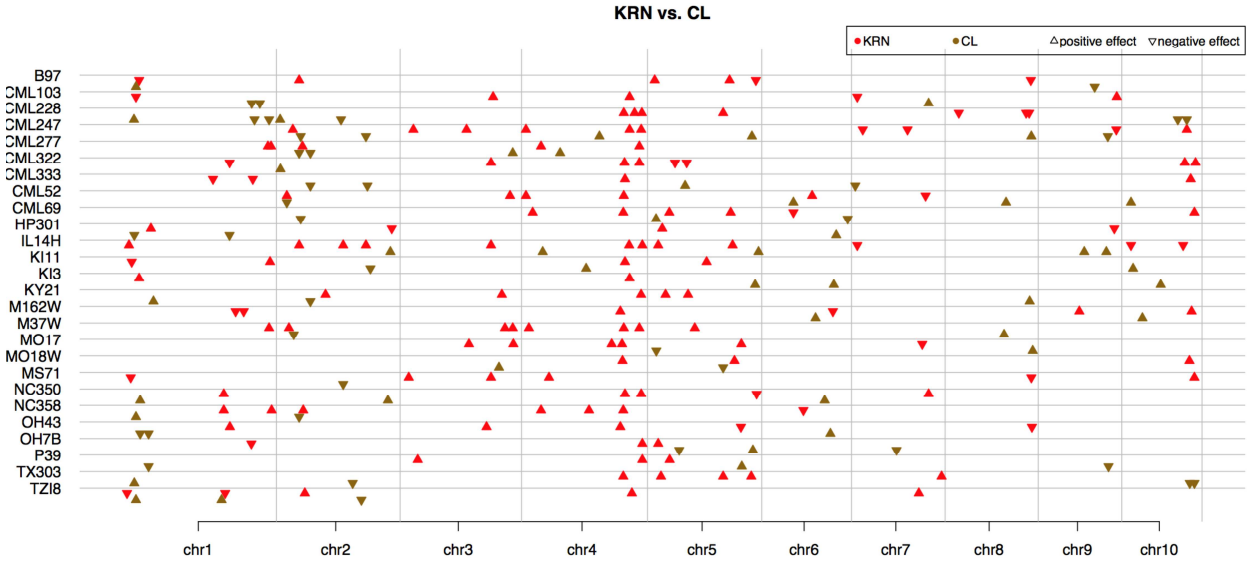
KRN vs. CL
KRN
CL
positive effect
negative effect
B97
CML103
CML228
CML247
CML277
CML322
CML333
CML52
CML69
HP301
IL14H
KI11
KI3
KY21
M162W
M37W
MO17
MO18W
MS71
NC350
NC358
OH43
OH7B
P39
TX303
TZI8
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10

**Figure S4**

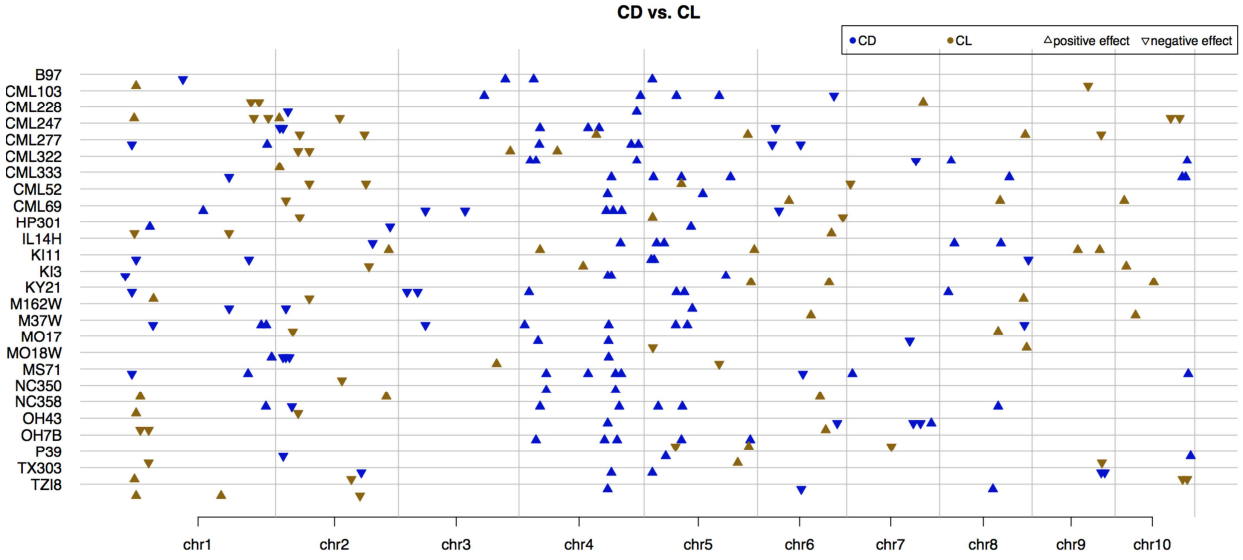
CD vs. CL
CD
CL
positive effect
negative effect
B97
CML103
CML228
CML247
CML277
CML322
CML333
CML52
CML69
HP301
IL14H
KI11
KI3
KY21
M162W
M37W
MO17
MO18W
MS71
NC350
NC358
OH43
OH7B
P39
TX303
TZI8
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10

**Figure S5**

## APPENDIX C. SUPPLEMENTAL INFORMATION FOR CHAPTER 4

**Figure S1** Plots of number of variants and the required depth of support for variants calling. The inside plot zooms in the depth of coverage > 50.

**Figure S2** Simulation of random sampling error with different depth of coverage. With different minor allele frequencies (MAFs) using different colors, sampling errors decrease as the depth of coverage increase.

**Figure S3** Distribution of off-target variants identified through Exome-sequencing. On the left side of the vertical dashed line shows the regions upstream of gene models; and on the right side of the vertical dashed line shows the regions downstream of gene models. Each dot represents number of variants detected in a bin region (bin size = 100-bp).

**Figure S4** Quantile-quantile plots of the $\chi^2$ distribution of the XP-GWAS. Panel (A) shows the result before the genomic control and panel (B) shows the result after the genomic control. Black lines are the diagonal lines.

**Figure S5** Reference allele frequencies of the identified TAVs. Panel (A) shows the pattern of high > random > low (N = 118 TAVs); panel (B) shows the pattern of high < random > low (N = 23 TAVs); panel (C) shows the pattern of high < random < low (N = 2 TAVs); and panel (D) shows the pattern of high < random > low (N = 1 TAVs).

**Figure S6** QTL fine mapping results. In the upper panel, interval mapping results using 26 heterozygous recombinant families (black solid line) and 220 homozygous recombinant families (red solid line). In the lower panel shows the identified recombinants and their genotyping results (blue denotes B73-like genotype and red denotes Mo17-like genotype). Horizontal dashed lines indicate 10% cut-off after 1000 permutation tests and vertical dashed lines indicate SNP markers of SNP173687206 and SNP176134968.

**Figure S7** XP-GWAS and independent pairwise Chi-square test results in the region of Chr4:185-186Mb. Panel (A) shows the XP-GWAS results; panel (B), (C) and (D) are the independent pairwise Chi-square tests of high KRN pool vs. low KRN pool,

high KRN pool vs. random KRN pool and low KRN pool vs. random KRN pool, respectively. Red dashed lines indicate the threshold of FDR < 0.05. Panel (E) shows the read depth of three KRN pools using a bin size = 1000, where blue lines are for high KRN pool, red lines are for low KRN pool and yellow lines are for random KRN pool.

**Figure S8** XP-GWAS and independent pairwise Chi-square test results in the region of Chr4:186-187Mb. Panel (A) shows the XP-GWAS results; panel (B), (C) and (D) are the independent pairwise Chi-square tests of high KRN pool vs. low KRN pool, high KRN pool vs. random KRN pool and low KRN pool vs. random KRN pool, respectively. Red dashed lines indicate the threshold of FDR < 0.05. Panel (E) shows the read depth of three KRN pools using a bin size = 1000, where blue lines are for high KRN pool, red lines are for low KRN pool and yellow lines are for random KRN pool.

**Figure S9** XP-GWAS and independent pairwise Chi-square test results in the region of Chr4:199.5-200.5Mb. Panel (A) shows the XP-GWAS results; panel (B), (C) and (D) are the independent pairwise Chi-square tests of high KRN pool vs. low KRN pool, high KRN pool vs. random KRN pool and low KRN pool vs. random KRN pool, respectively. Red dashed lines indicate the threshold of FDR < 0.05. Panel (E) shows the read depth of three KRN pools using a bin size = 1000, where blue lines are for high KRN pool, red lines are for low KRN pool and yellow lines are for random KRN pool.

**Figure S10** Validation of KRN phenotypic values in the database using our replicated field trial observations.

**Table S1** Accessions ID and KRN values of the selected high KRN, low KRN and random KRN lines.

**Table S2** Summary of exome-sequencing results of the three KRN pools. For each phenotypic pool, four technical replications of sequencing were conducted.

**Table S3** KRN phenotypic values of the identified homozygous recombinant families.

**Table S4** Genotypic data of the identified recombinants for Chr4:169-180Mb QTL fine mapping using 26 SNPs.

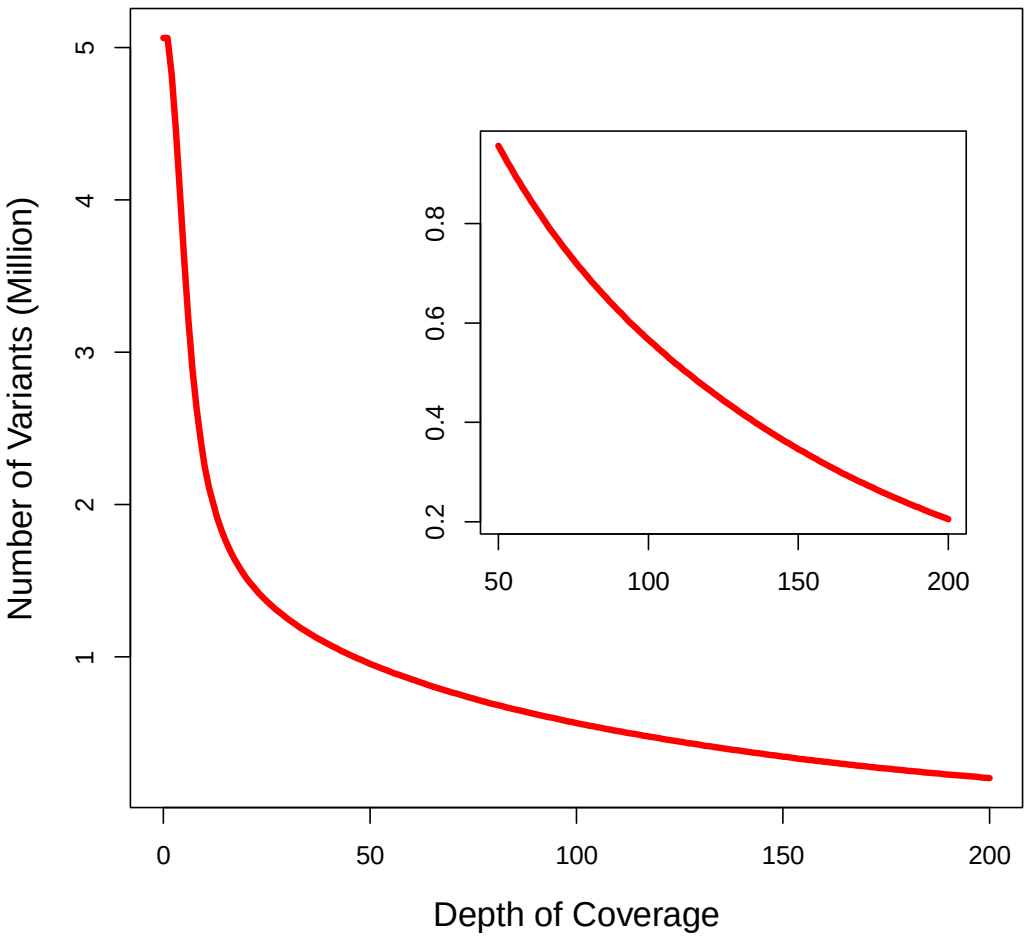
Number of Variants (Million)
1
2
3
4
5
0
50
100
150
200
Depth of Coverage
0.2
0.4
0.6
0.8
50
100
150
200

**Figure S1**

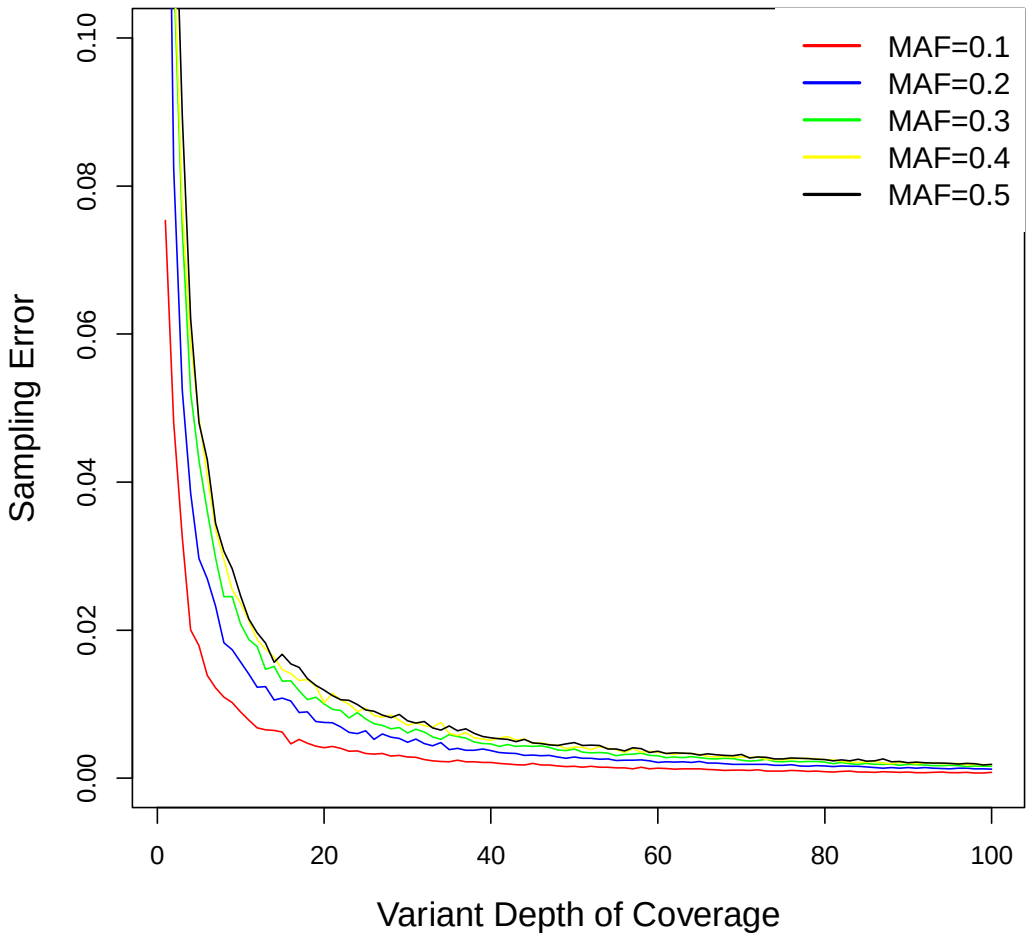
MAF=0.1
MAF=0.2
MAF=0.3
MAF=0.4
MAF=0.5
Sampling Error
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
80
100
Variant Depth of Coverage

**Figure S2**

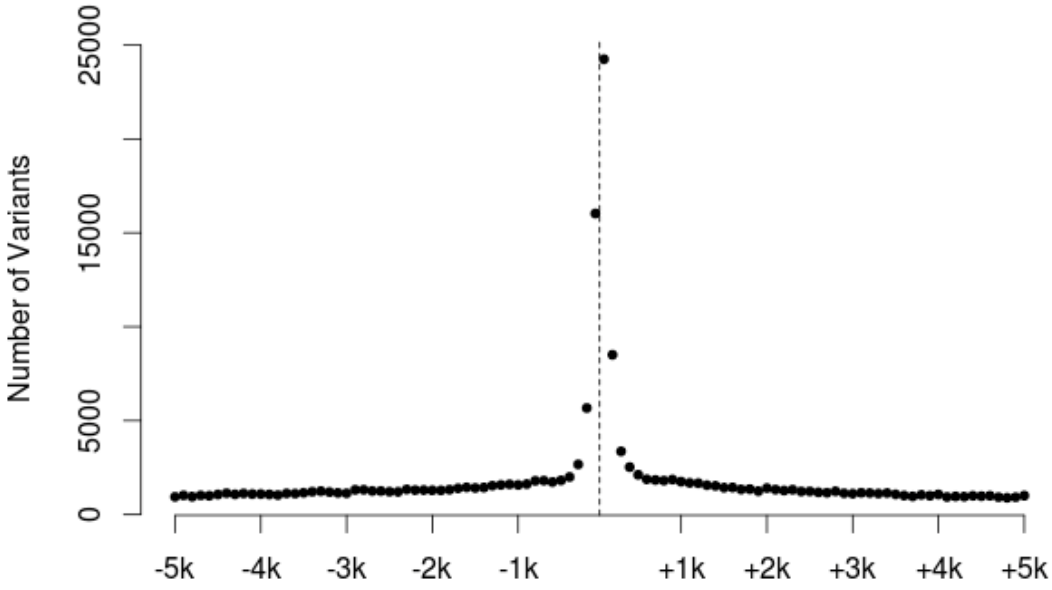
Number of Variants
0
5000
15000
25000
-5k
-4k
-3k
-2k
-1k
+1k
+2k
+3k
+4k
+5k

**Figure S3**

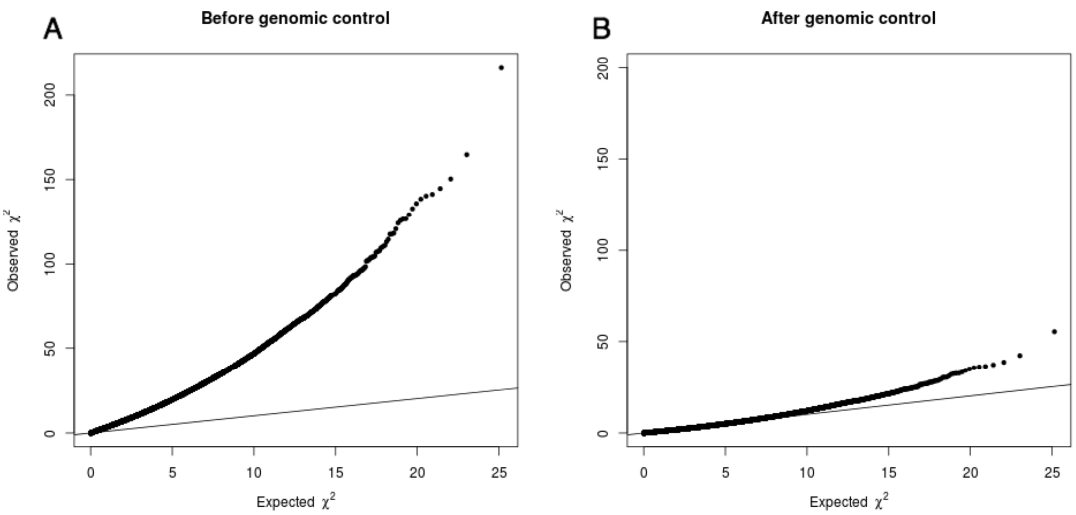
A
Before genomic control
Observed $\chi^2$
Expected $\chi^2$
0
50
100
150
200
0
5
10
15
20
25
B
After genomic control
Observed $\chi^2$
Expected $\chi^2$
0
50
100
150
200
0
5
10
15
20
25

**Figure S4**

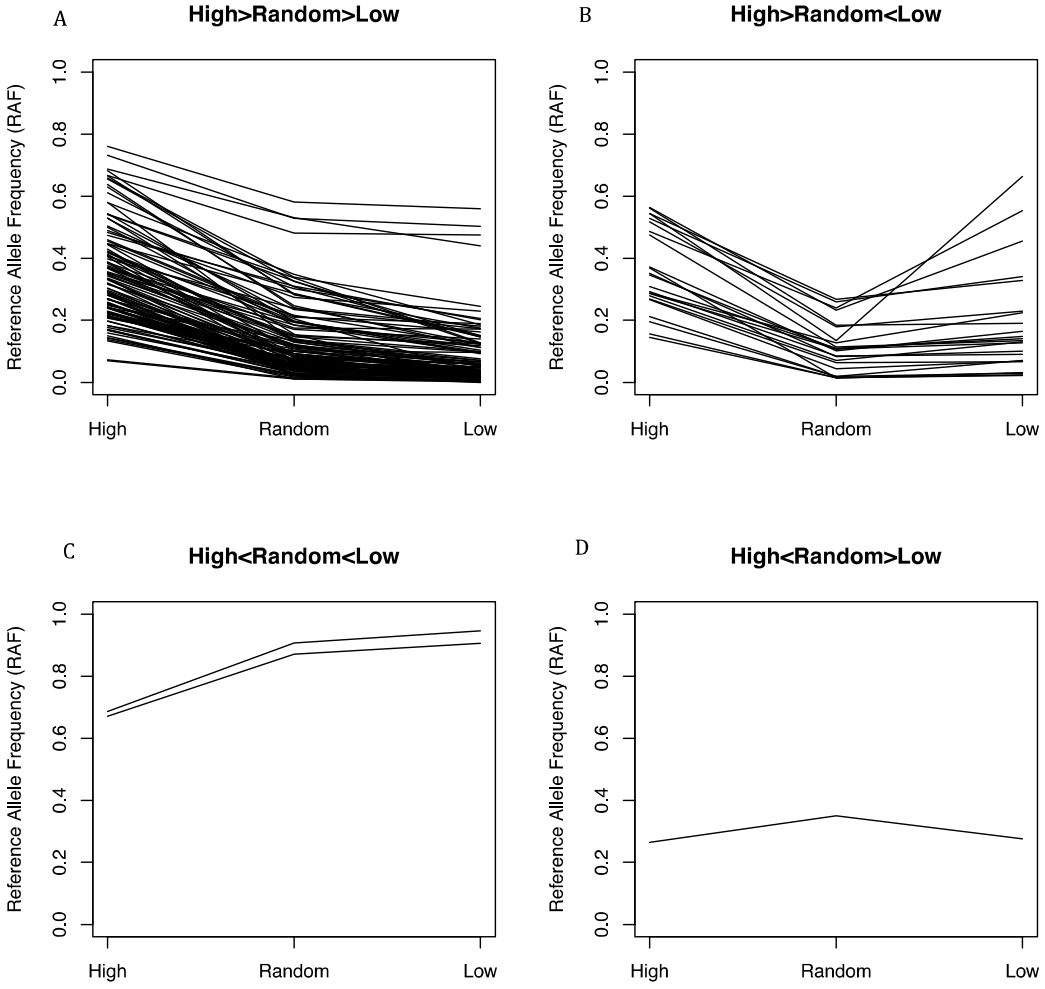
A
High>Random>Low
B
High>Random<Low
C
High<Random<Low
D
High<Random>Low
Reference Allele Frequency (RAF)
0.0
0.2
0.4
0.6
0.8
1.0
High
Random
Low

**Figure S5**

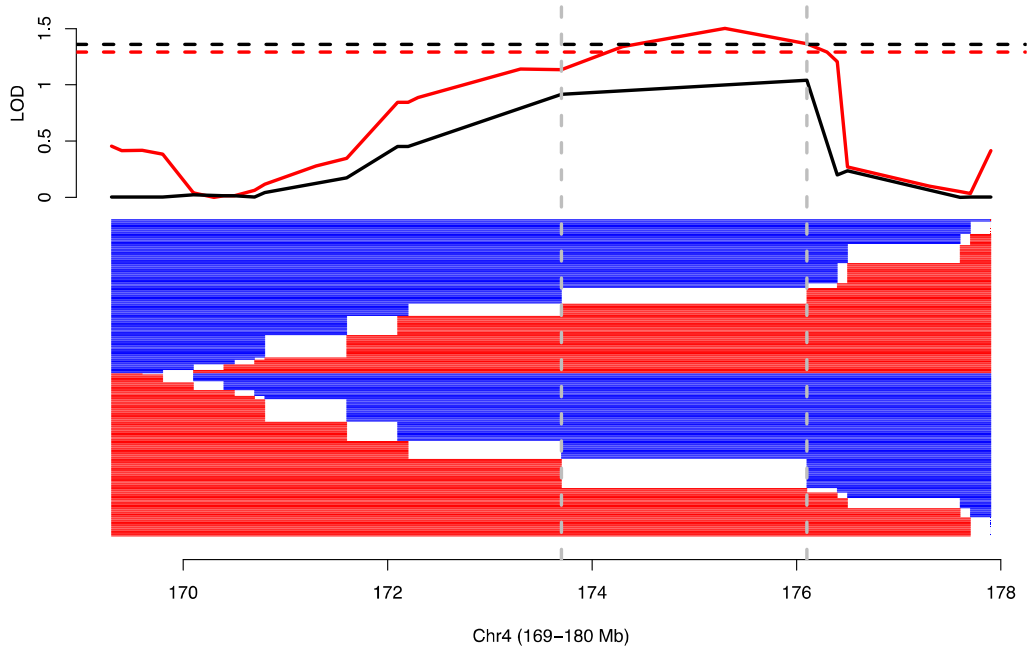


**Figure S6**

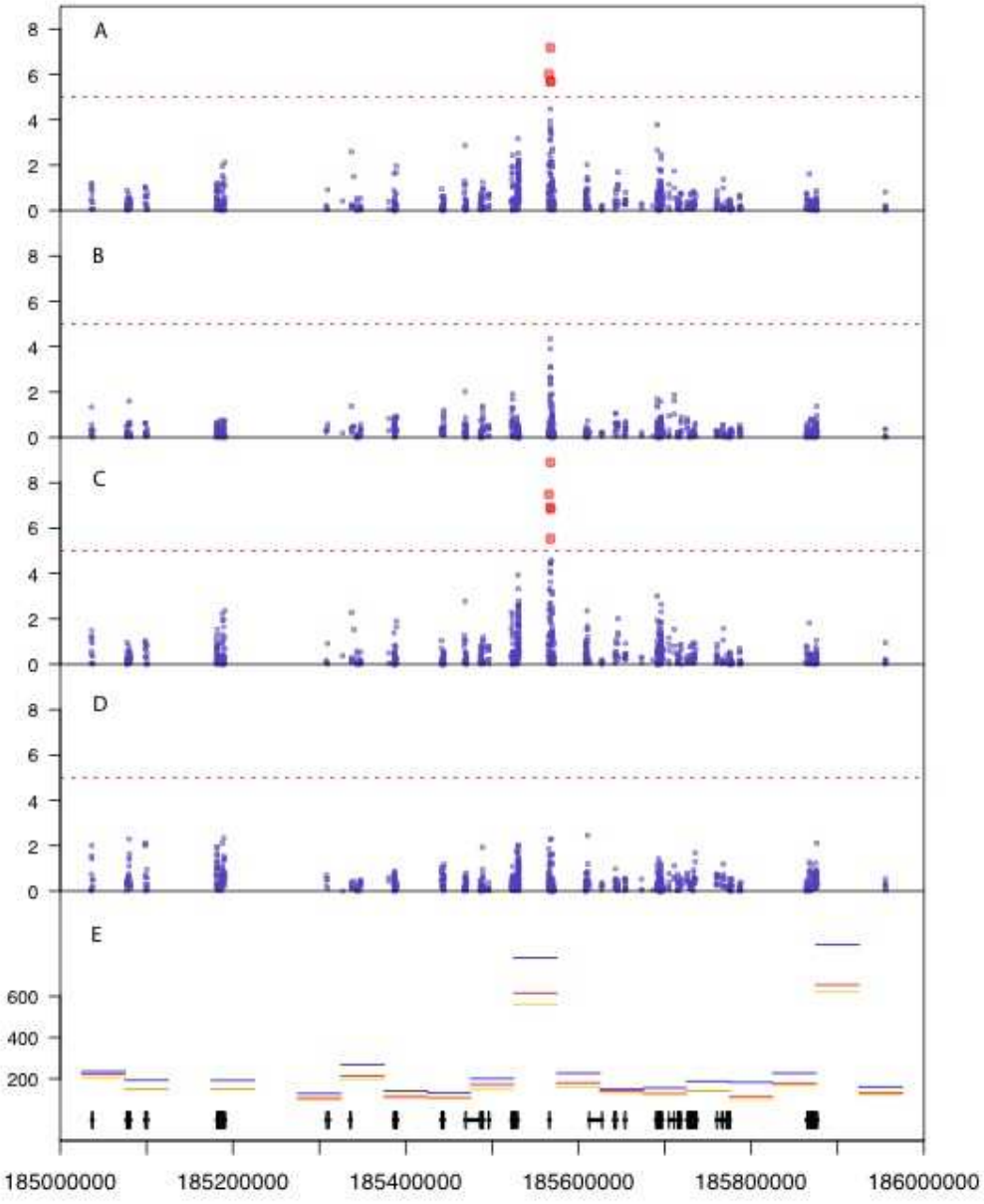
A
B
C
D
E
8
6
4
2
0
600
400
200
185000000
185200000
185400000
185600000
185800000
186000000

**Figure S7**

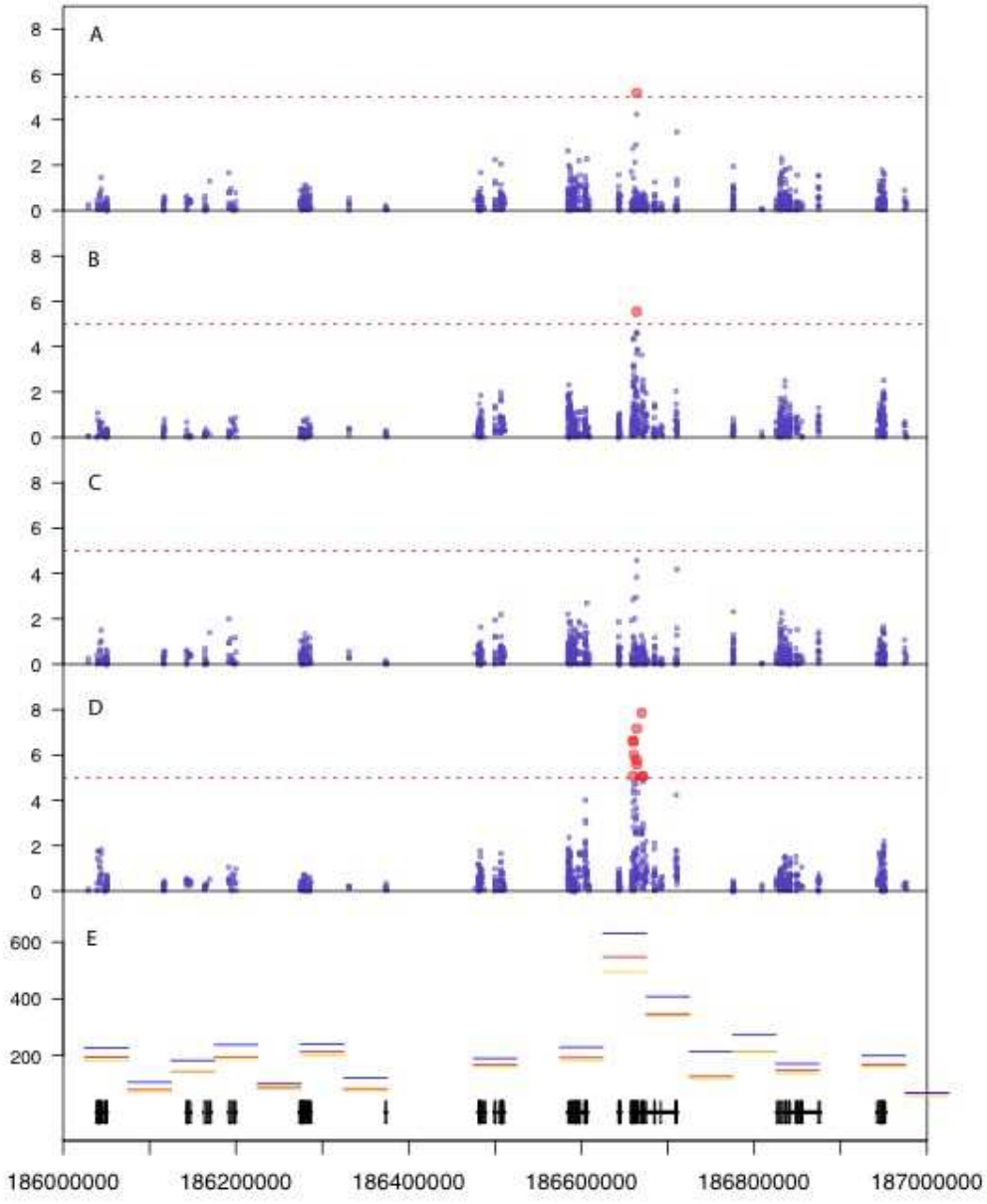
A
B
C
D
E
8
6
4
2
0
600
400
200
186000000
186200000
186400000
186600000
186800000
187000000

**Figure S8**

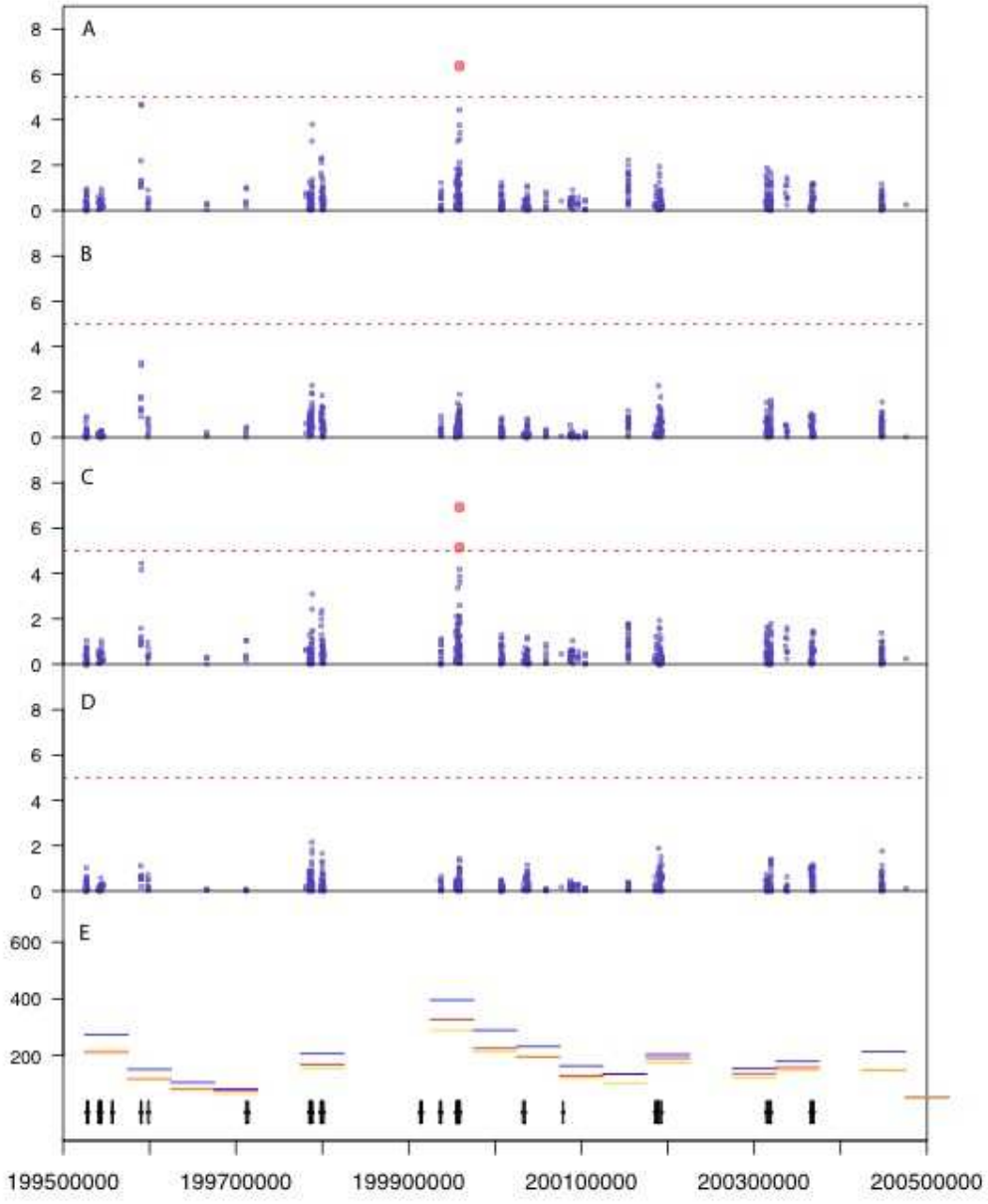
A
B
C
D
E
8
6
4
2
0
600
400
200
199500000
199700000
199900000
200100000
200300000
200500000

**Figure S9**

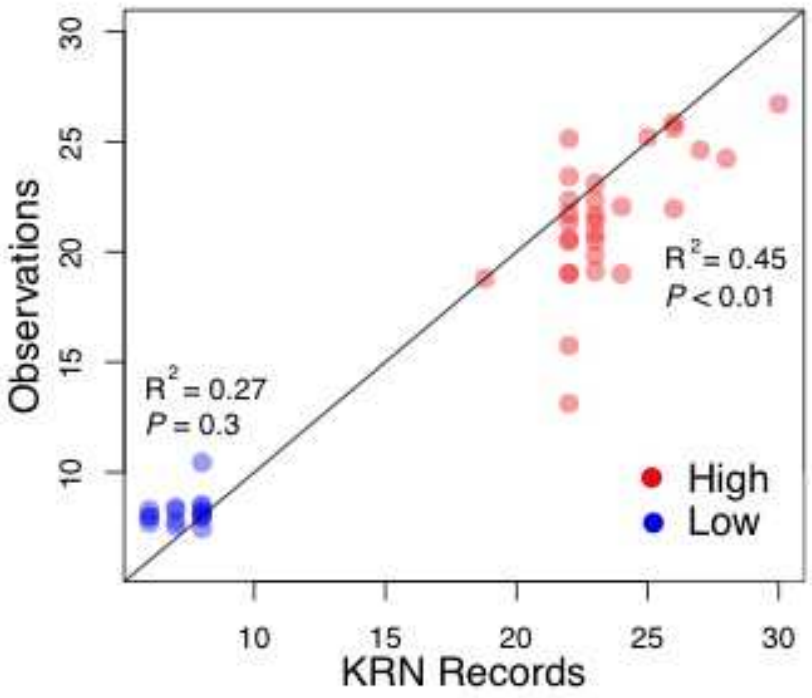
Observations
KRN Records
10
15
20
25
30
R² = 0.27
P = 0.3
R² = 0.45
P < 0.01
High
Low

**Figure S10**

**Table S1** Accessions ID and KRN values of the selected high KRN, low KRN and random KRN lines.

*See separate file.*

**Table S2** Summary of exome-sequencing results of the three KRN pools. For each phenotypic pool, four technical replications of sequencing were conducted.

*See separate file.*

**Table S3** KRN phenotypic values of the identified homozygous recombinant families.

*See separate file.*

**Table S4** Genotypic data of the identified recombinants for Chr4:169-180Mb QTL fine mapping using 26 SNPs.

*See separate file.*

## ACKNOWLEDGEMENTS

First of all, I would like to appreciate Dr. Patrick Schnable for sharing his enthusiasm for science and for providing guidance in both my professional and personal development. I also thank Dr. Dan Nettleton for the invaluable time he has dedicated towards this project. And I also thank other members of my POS committee: Drs. Erik Vollbrecht, Nick Lauter, and Jack Dekkers for their guidance and helpful suggestions. I appreciate Drs. Sanzhen Liu, Wei Wu, Ruth Swanson Wagner, Yan Fu, Kai Ying, Yi Jia and Heng-Cheng "Alvis" Hu for their discussions and helpful suggestions. I also appreciate Eddy Yeh, Dr. An-Ping Hsia, Alina Ott, Xiao Li, Lisa Coffey, Mitzi Wilkening, Dr. Haiying Jing, Dr. Li Li and all other lab members for their technical supports. Finally, I would like to give my special thanks to my family who have been supportive throughout my graduate studies: my parents and my parents-in-law for their support and encouragement; my wife, Jingjie Hao, for her patience, love, encouragement and constant support; and my daughter Olivia Yang, for her loveliness that makes my hard-working meaningful.