ABSTRACT

XUE, SHANG. Genetic Architecture of Domestication and other Complex Traits in Maize. (Under the direction of James B. Holland and Jung-Ying Tzeng).

Research on domestication informs our understanding of the genetic architectures of important traits and can guide efforts to identify causal genetic variants of crop. Numerous morphological traits have changed during domestication from teosinte to maize. However, standing variation still exists for several domestication-related traits in modern maize. The genetic architecture for standing variation remaining in maize for these domestication-related traits is unknown. In Chapter 2, the maize Nested Association Mapping population (NAM) and a diversity panel were used to estimate the proportion of variation due to polygenic, small-effect QTL versus larger effect variants; compare the genomic positions of larger effect variants to the known locations of domestication; and partition the genetic variance using variance components analysis methods. Additive polygenic models explained most of the genotypic variation for domestication-related traits; no large effect loci were detected for any trait. Previously defined improvement sweep regions were associated with more trait variation than expected based on the proportion of the genome they represent. Small effect polygenic variants (enriched in improvement sweep regions of the genome) are responsible for most of the standing variation for domestication-related traits in maize.

Linear mixed models are widely used in humans, animals, and plants to conduct genome-wide association studies (GWAS). Different from human datasets, experimental units for plants are typically multiple-plant plots of families or lines that are replicated across environments. This structure leads to computational challenges to conducting a genome scan

on plot-level data. Two-stage methods have been proposed to reduce the complexity and increase the computational speed for GWAS. However, the appropriate dependent variable to use in the second stage analysis and how to handle unbalanced datasets are not clear. In Chapter 3, I developed a weighted two-stage analysis to reduce bias and improve power of GWAS while maintaining the computational efficiency of two-stage analyses. Power and false discovery rate of one-stage, different two-stage models and new weighted analysis are compared by simulation based on real marker data of a diverse panel of maize inbred lines. Only weighted two-stage GWAS has power and false discovery rate similar to the one-stage analysis for severely unbalanced data in simulation. The weighted two-stage analysis method was implemented in a free open source software TASSEL.

Genetic diversity reduced severely due to selection and bottleneck effect during domestication of maize from teosinte. However, the gene pool of teosinte might harbor agriculturally beneficial alleles. Due to their linkage with a genomic background that is unadapted, measuring the potential benefit of teosinte alleles is challenging. A population called the ZeaSynthetic population was developed as a bridge to investigate teosinte specific alleles in a common maize background. Genotypes of 1846 parents were obtained by genotyping by sequencing and phenotypes of 923 pairs of S1 and S0 full-sib progeny families were measured across six environments. Due to unusual population structure and experimental design, there is no ready to use software available for analysis. In Chapter 4, I used a linear mixed model to estimate additive and dominance effects at each SNP and used simulation to evaluate power and bias of genetic effect estimates from this model. Simulation results showed that the linear mixed model is a reasonable model with low bias estimating genetic effects and high power detecting QTLs. Analysis of the real data showed that loci

with rare alleles and loci in lower recombination regions of the genome tend to have larger additive and dominance effects. These results suggest that recessive deleterious alleles tend to be concentrated in lower recombination regions, and that favorable alleles for agriculture are more likely to be at higher starting frequencies and found at loci in higher recombination regions.

Genetic Architecture of Domestication and other Complex Traits in Maize

by
Shang Xue

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2017

APPROVED BY:

James Holland Committee Co-Chair	Jung-Ying Tzeng Committee Co-Chair
Alison Motsinger-Reif	Ross Whetten

DEDICATION

To my parents, for providing love, care, support and encouragement for over twenty years. Thanks for giving me healthy body and free mind, so that I can explore new knowledge in new environment.

谨以此文献给我的父母，感谢你们给予我二十多年的关心和爱护，支持和鼓励。感谢你们给我健康的身体和自由的思想，让我能在新的环境里探索新知。

BIOGRAPHY

Shang Xue was born on October 7, 1991 at a small town in Linyi, China. She graduated from Shandong University with Bachelor of Science in Biological Science. During the undergraduate study there, she was attracted by amazing genetics and the power of computer to analyze genomic data. Later she came to the U.S. in 2012 as a Ph.D. graduate student in Bioinformatics at North Carolina State University to continue exploring new techniques in the field of genomics. She met great advisors and instructors there, and enjoyed both her study and life. After graduation, she will work as a bioinformatician at a technology company and she is looking forward to use the knowledge she learned to make contributions.

ACKNOWLEDGMENTS

There are so many people that influenced my academic life and personal life. First, I would like to especially thank my Ph.D. advisor, Dr. Jim Holland. Along the way, you gave me tremendous help and endless patience in my research. You are not only providing me academic guidance, but also setting an inspiring example for personal life and future career. Great thanks also go to my co-advisor, Dr. Jung-Ying Tzeng. You gave me solid support on statistical analysis and taught me how to solve problems with rigorous attitude. Your patience, thoughtfulness, solicitude, encouragement brought warmth to my overseas graduate life and your professionalism impacted my academic life profoundly. All those good qualities you gave me will keep motivating me to work harder and explore bigger world as a foreign scholar.

Thanks also go to other advisory committee members, Drs. Alison Motsinger-Reif and Ross Whetten, for providing me comments and suggestions. Many numbers of Holland Lab have contributed their insights into this dissertation including: Drs. Bode Olukolu, Funda Ogut, Charlie Zila, Tiffany Jamann, Zhou Fang, Luis Fernando Samayoa, Jeffrey Dunne and Ryan Andres; technicians, Jason Brewer and Josie Bloom; and graduate students, Thiago Marino, David Horne, Matthew Smith and Anna Rogers. I also want to thank Drs. Qin Yang, Peter Balint-Kurti, Peter Bradbury, Ed Buckler, Brian Reich, Howard Bondell, Sherry A. Flint-Garcia and Ginnie Morrison for suggestions and collaborations.

I want to say Thank You to my mom and dad for training my tenacity, patience, curiosity and positive attitude since I was a child. Those good qualities lead me through the way of conducting my research and pursuing the degree.

Finally, thanks all my friends for sharing joys and tears, your support and encouragement gave me courage to find a better me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Literature Review

Maize domestication

The domestication of all major crop plants occurred in a relatively short period in human history starting about 10,000 years ago (Harlan 1992). Crop domestication occurred when humans began to selectively save and replant seeds of preferred forms from wild species. Plant attributes that humans preferred in crops are referred to as the domestication syndrome (Hammer 1984; Harlan et al. 1973). The attributes include reduced number of harvest units (ears in the case of maize), suppression of shattering (seed dispersal), and larger seed size. Maize (*Zea mays* ssp. mays) was domesticated from its progenitor teosinte (*Z. mays* ssp. *parviglumis*) about 6000 to 10000 years ago in southern Mexico (Galinat 1983; Iltis 1983; Matsuoka *et al.* 2002; Piperno *et al.* 2009; van Heerwaarden *et al.* 2010).

The process of domestication involved artificial selection that resulted in radically different floral, ear, and kernel morphologies between teosinte and maize (Beadle 1939; Doebley *et al.* 1990; Doebley and Stec 1993; Iltis 1983). Teosinte plants have elongated lateral branches at many nodes. In contrast, maize plants typically produce a lateral branch at only two or three of the nodes on their main stems, and these are much shorter than teosinte lateral branches, being reduced to a "shank" that terminates at the base of a female ear (Doebley *et al.* 1997). Furthermore, teosinte "ears" are small, with kernels arranged in a distichous (two-ranked) pattern on the ear axis, compared to large ears of maize that typically have from eight to over twenty rows of kernels in four or more ranks. Several major QTL and in some cases, the specific genes, controlling these differences between maize and teosinte have been identified(Briggs *et al.* 2007; Clark *et al.* 2003; Doebley 2004; Weber *et al.* 2008).

Genetic bottlenecks due to domestication

As a consequence of artificial selection, alleles favorable for growth and development under agricultural conditions or for other traits desired by humans (such as flavor) increased in frequency, often reaching fixation and reducing genetic variation very near causal sequence sites (Wang *et al.* 1999). In addition, domestication was often accompanied by severe genetic bottlenecks from the use of small founder populations. The strong directional selection that occurred during the domestication of maize from teosinte reduced genetic diversity most strongly at key genes controlling domestication-related traits. Similar to other crops, previous studies have demonstrated that lower levels of genetic diversity among inbreds than among landrace and teosinte populations are caused by selection and demography (bottlenecks) (Tenaillon *et al.* 2004; Wright *et al.* 2005). Estimates of the amount of diversity lost are generally agreed to be around 30% in maize (BUCKLER *et al.* 2001; Zhang *et al.* 2002). This is similar to other grass species, such as rice (29% reduced; Oka 1988) and sorghum (33% reduced; Morden *et al.* 1990). Additional studies indicated that approximately 2 to 4% of genes (500 to 1000 genes) were targets of artificial selection during domestication and breeding (Hufford *et al.* 2012; Wright *et al.* 2005).

Domesticated plants can be model systems to study the genetic basis of adaptation, in which human artificial selection was imposed in addition to natural selection (Brown 2010; Ross-Ibarra *et al.* 2007). Also, understanding crop origins and domestication can contribute to the identification of useful genetic resources for breeding (HARRIS 1990; Olsen and Wendel 2013). By influencing the levels of nucleotide diversity and patterns of linkage disequilibrium (LD), domestication shapes the genetic variation available to breeders.

Research on domestication also informs our understanding of the genetic architectures of important traits and can guide efforts to identify causal genetic variants of crop.

S*tanding variation in maize for domestication-related traits*

Despite the severe bottleneck that occurred during domestication and strong selection for the maize plant type, standing variability in cob length, kernel row number, and shank length can still be observed among maize breeding lines. In addition, although most maize plants have purely female flowers on their lateral branch termini, some lines of maize produce a spike of staminate florets at the tips of their ears (Holland and Coles 2011), referred to as "masculinized ear tips", revealing variation for this domestication trait as well.

The genetic architecture for standing variation remaining in maize for these domestication-related traits is unknown. Sequence variation at the same set genes that were involved in the conversion of teosinte into domesticated maize may cause some portion of this variation. Several large-effect mutations that cause maize to exhibit teosinte-like morphological characteristics, such as *tb*1 and *gt*1, were later demonstrated to be allelic to the corresponding domestication loci (Doebley *et al.* 2006; Studer *et al.* 2011; Wills *et al.* 2013). Not all domestication alleles are fixed in domesticated species (Meyer and Purugganan 2013; Studer *et al.* 2011), leaving open the possibility that some domestication loci contribute to standing trait variation in domesticated species. Furthermore, a range of allelic series exists at some domestication loci (Studer and Doebley 2012); smaller-effect alleles may segregate in domesticated species even if larger-effect wild-type alleles were lost from the species. Smaller-effect variants could have originated in the wild ancestor and passed through the

domestication bottleneck because of lower selection intensity or may have arisen by mutation after domestication.

Alternatively, the observed phenotypic variation for domestication traits within domesticated species might be due to large-effect genes that are distinct from the known domestication genes. The variants at these genes may have arisen after domestication, or had effects sufficiently small as to avoid purging during domestication. A third possibility is that the observed phenotypic variation for domestication traits is produced by many small-effect variants distributed throughout the genome, resulting in a polygenic genetic architecture. Even if major-effect alleles were fixed during domestication, smaller-effect variants at other loci could cause phenotypic variation in domestication target traits. Again, these variants could have existed in the wild ancestor and passed through the domestication bottleneck due to small selection coefficients, or they may represent new variation that arose from mutation following domestication.

Loss of useful alleles during domestication?

Because selection and bottleneck effects have greatly reduced the genetic diversity among modern maize varieties, some alleles segregating in teosinte did not pass through the domestication bottleneck and do not exist in maize. It is likely that at least a small proportion of these teosinte-specific alleles could have favorable effects if transferred to maize; they could have been lost from maize populations simply by random genetic drift due to the population bottlenecks in domestication. Teosinte populations may harbor unique alleles for disease resistance other useful agricultural traits that could be used for plant breeding and maize improvement. Identifying agriculturally useful alleles in teosinte is challenging,

however, because of the numerous dramatic morphological between maize and teosinte. Phenotypic comparisons between the subspecies will not reveal useful teosinte alleles because they are completely confounded by the agronomically unacceptable wild plant growth habit, morphology, and seed dispersal phenotypes. Furthermore, teosintes are not adapted to temperate growing environments and will not flower under long daylengths of the USA Corn Belt growing season(Hung *et al.* 2012). Therefore, alleles in teosinte that affect productivity are completely masked by the expression of flowering repressors in teosinte. For these reasons, specialized experimental populations derived from crossing between teosinte and maize parents are required to estimate the effects of teosinte-specific alleles in the context of the maize genome, morphology, and modern agricultural management and production methods.

Maize improvement from breeding

From population improvement to inbred-hybrid breeding

Following the domestication of maize in Southern Mexico, humans spread maize from its center of origin and selected for desired forms adapted to dramatically different ecological conditions and for different uses. Maize was spread from Canada to Chile by humans before the arrival of Europeans in the New World (Weatherwax 1954). The result of this spread and divergent selection process over thousands of years is a dramatic variation in many observable phenotypes in maize. Subsequent classification of the sub-populations of maize based on their geography and morphology recognized up to 350 distinct ‘races’ of maize (Goodman 1968; Goodman and Paterniani 1969; Goodman and Bird 1977; Goodman

1988; Goodman *et al.* 1988; Vigouroux *et al.* 2008). Different maize races are adapted to widely different ecological conditions (Ruiz Corral *et al.* 2008).

Since maize is naturally an open-pollinated crop, the traditional populations grown in farmer's fields represent highly variable segregating populations with dynamic genetic characteristics influenced by natural selection for adaptation, human selection for appearance and various uses, and migration due to intentional seed exchange as well as to pollen spread. Due to various individual preferences, production conditions, and production objectives from one farming household to another, farmers' widely observed practice of selecting and saving seed from the previous harvest and occasionally exchanging seeds provided the basis for the initial pool of diverse and locally-adapted maize varieties (Badstue *et al.* 2007; Pressoir and Berthaud 2004).

The earliest scientific breeding efforts in maize were aimed at improving populations through selection on individual plants or seeds. Starting in the 1920s, efforts began to exploit the strong hybrid vigor, or heterosis, of maize, by breeding inbred lines and crossing unrelated inbred lines to produce hybrid varieties (Duvick *et al.* 2010). This method was ultimately highly successful, resulting in nearly complete adoption of hybrid maize varieties by farmers by 1970, and tremendous gains in average yield per hectare of maize in the USA. The effect of this intensive breeding effort is a second phase of reduction in allelic diversity between modern maize inbreds and older outcrossing "landrace" maize populations on top of the loss of diversity due to domestication. This more recent loss of diversity has been referred to as the "improvement bottleneck" and signals of particularly strong selection above the background level of diversity reduction have been observed at numerous genome regions (van Heerwaarden *et al.* 2012).

Modern maize breeders recognize different sub-groups of maize breeding lines referred to as "heterotic groups". A heterotic group is a set of lines related by pedigree that when crossed together will express only limited heterosis. Crosses between heterotic groups, however, express substantial heterosis. Therefore, maize breeders tend to repeatedly cross lines within heterotic groups to produce new breeding lines, whereas they cross lines between heterotic groups to produce hybrid varieties that farmers grow (Tracy and Chandler 2006). The result of this repeated recycling of related lines within heterotic groups has been a strong divergence of allele frequencies genome-wide between the heterotic groups (van Heerwaarden *et al.* 2012).

Genetic architecture of complex traits in maize

Inbreeding depression and heterosis

Although the inbred-hybrid method is almost universally accepted as the best breeding method for improving yield in maize, the initial phases of the inbred-hybrid breeding method were not entirely promising because of the strong inbreeding depression that occurred during the process of repeated self-fertilization out of the original open-pollinated source populations. Inbreeding depression refers to the reduced survival and fertility of offspring of related individuals, it occurs in wild animal and plant populations as well as in humans, indicating that genetic variation in fitness traits exists in natural populations (Charlesworth and Willis 2009). Heterosis refers to the phenomenon that progeny of diverse varieties of a species or crosses between species exhibit greater biomass, speed of development, and fertility than both parents (Birchler *et al.* 2010). Heterosis from

outcrossing is the converse of inbreeding depression. Because intercrossing inbred strains improves yield (heterosis), which is important in crop breeding, the genetic basis of these effects has been debated since the early twentieth century.

There are two primary hypotheses about the genetic mechanisms responsible for heterosis and the success of maize hybrids. First, overdominance models of heterosis predict that at a single locus two distinct alleles confer heterozygote advantage when combined. Alternatively, the dominance model predicts that heterosis is driven by dominance effects and the complementation of favorable alleles in repulsion-phase linkage, particularly in low recombination regions (which can generate "pseudo-overdominance" in the absence of true overdominance) (Gerke *et al.* 2015; Hill and Robertson 1966; McMullen *et al.* 2009). The majority of genetic evidence indicates that the dominance model can explain most of the genetic control of inbreeding depression and heterosis (Crow 2000; Gardner and Lonnquist 1959; Gerke *et al.* 2015; Mezmouk and Ross-Ibarra 2013; Moll *et al.* 1963), although there is still some debate (Birchler *et al.* 2010). It is important to understand genetics underlying inbreeding depression and heterosis to design optimal breeding strategies for hybrid crops. If loci with heterozygote advantage are common, artificial selection in agricultural species should select for strains that manifest substantial heterosis. Crop plants with a uniform, highly heterozygous genotype with high fitness could be desirable, and could perhaps be achieved by asexual seed production (Grossniklaus *et al.* 2001). However, if heterosis in crops is mainly caused by complementation of dominant favorable alleles that mask deleterious mutations, it might be better to exclude these alleles to produce high-yield mutant-free strains. Also understanding the genetic basis of inbreeding depression can help to

address the question of why genetic variation in fitness-related characteristics exists in so many species, including humans (Lewontin 1974).

Recombination rate and genetic load

Recombination has large impact on evolution in general and plant breeding in particular by promoting the diversity necessary to respond to continually changing environments and preventing the build-up of genetic load by decoupling linked deleterious and beneficial alleles. Recombination varies among genomic region in both animal and plants (Anderson *et al.* 2001; Gaut *et al.* 2007; Jensen-Seaman *et al.* 2004) and variation in recombination rates may be associated with the distribution of variants for genetic load or inbreeding depression. There are two main reasons for varying crossover rate at the molecular level. First, the chromatin structure heavily influences recombination rates in plants. Heterochromatic regions generally reduce crossovers, however, knockout of cytosine-DNA-methyl-transferase (MET1) resulted in genome-wide CpG hypomethylation and increased the proportion of crossovers (Colome-Tatche *et al.* 2012; Mirouze *et al.* 2012; Yelina *et al.* 2012). Fu et al.(Fu and Dooner 2002) demonstrated that, in maize, crossovers were suppressed in regions with repetitive DNA sequences derived from retrotransposons compared to sequences including and nearby functional coding gene sequences. Highly repetitive retrotransposon-derived sequences tend to be heterochromatic in maize(Slotkin and Martienssen 2007). Second, nucleotide content may also be associated with recombination rate possibly due to the effect of GC-biased gene conversion (bGC)(Serres-Giardi *et al.* 2012). Maize experienced a modest bottleneck in genetic diversity (Tenaillon *et al.* 2004), although strongly deleterious variants were likely purged during the inbreeding process

leading to the founder lines, many weakly deleterious alleles can be found segregating at low frequencies among inbreds (Mezmouk and Ross-Ibarra 2013).

McMullen et al. (McMullen *et al.* 2009) observed that excess residual heterozygosity was enriched in pericentromeric regions of maize inbred lines, suggesting that selection against deleterious recessive alleles has been less efficient in these regions because of reduced recombination frequency, as predicted by the Hill-Robertson effect (Hill and Robertson 1966). In addition, other studies observed that deleterious alleles enriched in low-recombination regions, as expected because reduced recombination permits deleterious alleles to hitchhike to high frequency during selective sweeps (Hartfield and Otto 2011; Rodgers-Melnick *et al.* 2015). However, a recent study shows that putatively deleterious nonsynonymous polymorphisms in maize were not significantly enriched in regions of low recombination (Mezmouk and Ross-Ibarra 2013).

Methods and populations used for complex trait genetic analysis in maize

Understanding gene to phenotype relationships and discovering genes affecting agronomic traits has great value for increasing the speed of selective breeding programs in agriculturally important plants and animals and for predicting adaptive evolution. The discovered association can be used for marker assisted selection (Collard and Mackill 2007), genomic prediction (Bian and Holland 2017; Spindel *et al.* 2016) and understanding genetic architecture for agriculturally important traits (Buckler *et al.* 2009; Kump *et al.* 2011; Poland *et al.* 2011). However, most important agronomic traits are complex: they are quantitative traits measured on continuous scale, controlled by the joint effects of many genes,

environmental factors, and interactions between genes and the environment. Pinpointing the causal genes underlying quantitative trait loci (QTL) for complex traits can be challenging, and the difficulties are greater as the proportion of phenotypic variance caused by a gene are smaller.

Linkage analysis and association mapping are two important approaches for QTL mapping. Linkage mapping in biparental crosses was the classical approach for QTL mapping in plants; by relying on recent recombination events, it has the advantage of requiring fewer genetic marker to cover the genome and high statistical power per allele (Mauricio 2001)(Doerge 2002). However, it has limitations, including the substantial work required to build mapping populations, limited sampling of allelic variability, and low mapping resolution (Mauricio 2001). The development of high-throughput, dense genotyping has led to a shift from traditional QTL mapping to association or LD mapping to overcome these limitations (Zhu *et al.* 2008). Rather than focusing on two parental lines that differ strongly in phenotype, LD-mapping approaches assess the correlation between phenotype and genotype by taking advantage of LD in populations of unrelated individuals (Morrell *et al.* 2012; Myles *et al.* 2009). First, by relying historical recombination event in natural population, association mapping surveys genetic variation in much larger genetic pools than a biparental cross (Myles *et al.* 2009). Second, it eliminates the time and effort required to create specialized mapping populations because association mapping can be conducted in general populations (Nordborg and Weigel 2008). Thirdly, it allows for high resolution for QTL mapping, potentially to the level of causal nucleotide variants, depending on the extent of local linkage disequilibrium in the population sample (Mackay *et al.* 2009).

Association analysis has limitations as well, in particular the need to correct for population structure effects that can lead to false positives, and lower power to detect relatively rare variants. Previous results showed that genome wide association studies suffered from low power (Asimit and Zeggini 2010; Gibson 2012; Korte and Farlow 2013) and failing to explain missing heritabilities (Eichler *et al.* 2010; Li *et al.* 2010; Manolio *et al.* 2009). Most GWAS studies started first by identifying SNPs that segregate at intermediate frequency in a small population; then SNPs are genotyped in larger samples and phenotypes are measured. The underlying motivation is the assumption that common phenotypic variation will be caused by common genetic variation. In the context of human genetics, this assumption is known as the common disease-common variant hypothesis (Lander 1996). Although thousands of significant associations have been found, they explained low percent of variation of phenotypes (Eichler *et al.* 2010; Manolio *et al.* 2009). GWAS studies in plants in some cases have been more successful than in humans (Brachi *et al.* 2011), however, the missing-heritability issue still occurs (Li *et al.* 2010).

Variance component partitioning can be used as a complementary approach to GWAS to assess enrichment of significant associations of a particular genomic regions or functional class. Previous association studies showed that there is enrichment for significant trait associations in coding regions and UTRs compared to intergenic SNPs (Schork *et al.* 2013). Although it is possible to document relative enrichment of associations in particular genome regions, this can be confounded by correlations among SNPs in different regions. Furthermore, the contribution of different categories of sequence variants to trait heritability cannot be determined from GWAS. To address these limitations variance component analysis was recently proposed to estimate the total additive variance and heritability explained

collectively by groups of SNPs (Cross-Disorder Group of the Psychiatric,Genomics Consortium 2013; Gusev *et al.* 2013; Lee *et al.* 2012; Lee *et al.* 2011; Speed *et al.* 2012; Yang *et al.* 2013; Yang *et al.* 2010; Yang *et al.* 2011; Zaitlen and Kraft 2012; Zaitlen *et al.* 2013). In contrast to genome wide association study, variance components analysis leverages the entire polygenic architecture of each trait and accounts for pervasive linkage disequilibrium (LD) across functional categories. For a single component of genotyped (or imputed) SNPs, h^2 is defined as r^2 between the true phenotype and the best linear prediction over those SNPs. With multiple components, the goal of the partitioned analysis is to quantify the h^2 directly explained by SNPs in each functional category while excluding tagging of SNPs in other categories. Thus h^2 for each functional category is defined as the r^2 between the true phenotype and the prediction only from SNPs in that functional category when all functional categories are jointly analyzed for a best linear prediction. First, SNPs are grouped into different categories according to functional annotation, then genetic relationship matrices (GRM) for SNPs in each category are calculated. The total variance of the phenotype can be modeled as a summation of variance component multiplied by its corresponding GRM. Then h^2 for each category can be calculated after obtaining variance components estimates. Simulation showed that this approach provides accurate genome-wide estimates of heritability in diverse genetic architectures (Gusev *et al.* 2014; Lee *et al.* 2011).

Numerous association mapping panels are already available in maize (Crouch *et al.* 2011). The largest and most diverse such population that is publicly available is a set of 2,815 inbred lines maintained by the U.S. Department of Agriculture North Central Region Plant Introduction Station (NCRPIS) (Romay *et al.* 2013). This collection contains lines representing nearly a century of maize breeding efforts from programs throughout the world

and has been densely genotyped (Romay *et al.* 2013). An alternative to conducting genome-wide association study in samples of breeding lines, which are characterized by complex population structure, is to use multiple parent populations of known pedigree, with known population structure. One such population is the maize Nested Association Mapping (NAM) population, which consists of 4892 recombinant inbred lines (RILs) derived from 25 bi-parental families (Buckler *et al.* 2009; Hung *et al.* 2012; Tian *et al.* 2011; Yu *et al.* 2008). High resolution GWAS can be conducted in NAM while controlling for the known pedigree structure and for genetic variation at unlinked QTL detected by joint multiple population linkage analysis (Kump *et al.* 2011; Tian *et al.* 2011). For diverse maize germplasm sample, most alleles have low frequency (Romay *et al.* 2013). Association mapping in diversity panel relies on natural historical recombination, which has produced reshuffled genomes with relatively short blocks of LD, providing high power for detection of QTL effects. Whereas diversity panels can be obtained directly from seed banks, the NAM population required substantial resource investment to generated thousands of lines by controlled pollinations over several generations. The NAM provides the advantage that frequencies of alleles captured in the parents are balanced in the progeny, the population structure is known, and LD decay is intermediate between a diversity panel and a single biparental cross, resulting in high of power and moderate resolution for QTL detection.

The relative ease of controlled matings in maize permits the development of specialized experimental populations that can combine the high resolution of diversity panels with low levels of population structure. An example of such a population is the recently developed maize ZeaSyn 6 population. ZeaSyn6 was developed by introgressing teosinte genomic regions into a common maize genetic background. Due to numerous cycles of

random mating in the generating process, this population has reduced LD and essentially no population structure. Because of the population makeup, this population contains a large proportion of moderately rare alleles, and can serve as a useful and unique tool to investigate teosinte-specific rare alleles.

Frontiers of complex trait genetic analysis in maize

Comparing results from different methods and population types.

Little work has been done to apply variance component partitioning approaches to address biological hypotheses in plants. Even in human genetics studies, where such methods were pioneered, the results were presented in isolation of GWAS studies. An important objective for contemporary quantitative genetics analysis is the implementation of variance component partitioning analysis on the same data sets as GWAS methods, and comparisons of the results.

GWAS can identify markers linked to individual or clustered sequence variants that have relatively large effects on phenotypes. However, when effect sizes at individual SNPs are so small that they do not pass genome-wide significance thresholds in GWAS or when causal variants are not in sufficiently LD with available SNPs, GWAS will not detect these SNPs or attribute any proportion of the heritability to them (Lee *et al.* 2011). As a complementary approach, both real data analysis and simulations demonstrated the value of the variance-component approach in recovering true additional heritability beyond that explained by individually or jointly significant markers in various disease traits (Yang *et al.* 2010)(Yang *et al.* 2011)(Gusev *et al.* 2013)(Cross-Disorder Group of the

Psychiatric,Genomics Consortium 2013). Therefore, findings combining GWAS analysis and variance components analysis can provide important ramifications for fine-mapping study design and understanding of complex disease architecture.

In addition, no formal tests of significance have been developed for determining if a particular class of variants explain more variation than expected by chance. Such tests are hindered by the complex dependencies between different features of the genome, such that variation attributed to one class of variants may in fact be due to another class with which it is correlated. To address this problem, in Chapter 2 of this thesis, I developed several strategies to estimate the expected variation for random sets of SNPs using resampling methods and simultaneous fitting of covariance matrices defined for different sets of variants.

It will also be important to compare genetic architecture results across populations with different structure and composition. In Chapter 2 of this thesis, I applied both variance component partitioning and GWAS to both a multiparent linkage mapping population (the maize nested association mapping population) and a very genetically diverse panel of inbred lines representing most of the inbred maize line collection of the USDA seed bank. This permitted comparisons of results of very different methods across two populations with contrasting levels of diversity, allele frequency distributions, population structure, and linkage disequilibrium.

Two-step GWAS algorithms

Using a linear mixed model that accounts for extraneous factors and genetic background effects to estimate a single SNP effect can be computationally challenging for large marker dataset with complex experimental designs. Different strategies have been

proposed in human, animal and plant studies to speed up this process. The stage-wise approach is often used where the raw phenotypes are first adjusted by extraneous effects before analyzing them with a linear mixed model in the second step. In human studies, two-stage regression analysis is often used to test SNPs for associations with quantitative diseases (Laird *et al.* 2000; Naylor *et al.* 2009; Zeegers *et al.* 2004). The limitations are biased estimates of genotypic effects and reduced power (Che *et al.* 2012; Demissie and Cupples 2011). In animal studies, a similar strategy is used. First, account for either pedigree relationships or realized genomic relationships by fitting a mixed model to the data on individuals. Then residuals for each individual are used as dependent variable for marker scan (Amin *et al.* 2007; Aulchenko *et al.* 2007; Lam *et al.* 2007). This approach, called 'GRAMMAR', achieved reduced computing time but sacrificed power in some cases (Zhou and Stephens 2012).

In contrast to human and animal studies, plant data are often generated from experimental designs in which the experimental units are field plots composed of multiple plants from a common family or inbred line, and often the designs are replicated across different environments. In order to accounts for environment, genotype, and genotype-by-environment interactions, a typical linear model should include multiple random terms, each associated with a different variance component. Although a full model incorporating these random effects in addition to the effect of a single marker can be specified and fit using a mixed linear model, this approach is too computationally demanding for practical use in scanning thousands or millions of markers in a GWAS. Software such as EMMA (Kang *et al.* 2008), FaSTLMM (Lippert *et al.* 2011), and GEMMA (Zhou and Stephens 2012) were developed to solve the large computational problem in human datasets. EMMA takes

advantage of the specific nature of the optimization problem in applying mixed models for association mapping by leveraging spectral decomposition of the genomic relationship matrix. By substantially decreasing the computational cost of each iteration, it enables convergence to a global optimum of the likelihood in variance-component estimation with high confidence by combining grid search and the Newton–Raphson algorithm. Since repeatedly estimating variance components for each SNP is computationally expensive, approximate algorithms like 'EMMA expedited' (called EMMAX) and 'population parameters previously determined' (called P3D) provide additional computational savings by assuming that variance parameters for each tested SNP are the same (Kang *et al.* 2010; Zhang *et al.* 2010).

More recently, FaST-LMM and GEMMA algorithms were proposed that can perform rapid GWAS analysis without assuming variance parameters to be the same across SNPs. FaST-LMM uses spectral decomposition of the genetic similarity matrix to transform (rotate) the phenotypes, SNPs and covariates. These transformed data are uncorrelated can be analyzed with a linear regression model. Similarly, GEMMA expedites each iteration by optimizing the efficiency of the computations required to evaluate the model likelihood and the first and second derivatives of the likelihood function.

In general, researchers can use all available single-nucleotide polymorphisms (SNPs) to determine relatedness among individuals (Kang *et al.* 2010). Several variants for calculating relatedness matrix have been proposed. One of these variants is called Fast-LMM-Select (Listgarten *et al.* 2012; Listgarten *et al.* 2013). It was shown theoretically and experimentally that carefully selecting a smaller number of SNPs systematically increases power (that is, it jointly reduces false positives and false negatives), improves calibration

(lessens inflation or deflation of the test statistic) and reduces computational cost. Another idea of calculating relatedness matrix is using all SNPs except SNPs on the chromosome being scanned, and power was increased by excluding markers that are in LD with the marker being tested (Cheng *et al.* 2013).

These algorithms and software provided solutions for linear mixed models that only involve two random components: the polygenic background and error variance components. In many plant studies, two-stage analyses are still necessary even with these improvements in algorithms to conduct linear mixed model GWAS. Two-stage approaches to GWAS for plant studies replicated across environments can take various forms. For example, in the first stage, the genotype effects can be fit as fixed or as random effects with no covariances, leading to the marginal prediction of genotype effects as either best linear unbiased estimation (BLUE) or best linear unbiased prediction (BLUP), respectively. In the second stage, the BLUEs or BLUPs of genotype obtained in the first stage may be fit as the dependent variable in a GWAS, in which the genotypes are treated as random with a variance-covariance matrix proportional to an estimated realized genomic relationship matrix (Aranzana *et al.* 2005; Lipka *et al.* 2013; Pasam *et al.* 2012b; Peiffer *et al.* 2013; Zhang *et al.* 2009).

Another approach to two-step GWAS involves using residuals, similar to the GRAMMAR method, but a complication is that replicated trials result in multiple residual values for each family. The first stage residuals could be averaged for each family and used as inputs to the second stage. Alternatively, a term for independent family effects can be fit in the first stage model in addition to the polygenic family effects with covariance proportional to the relationship matrix (Oakey *et al.* 2007) and the independent line effect could be used as the dependent variable in the second stage. Finally, a three-step analysis procedure could

be used. In the first step, BLUEs are computed for each line from plot level data, second BLUEs are fit as dependent variables in a linear mixed model including the relatedness matrix, and in the third step, residuals from second step are used for GWAS.

In addition to more complex experimental designs, another common feature of plant datasets is their unbalanced nature. Balanced data sets contain an equal number of observations for each combination of model factor levels. In contrast, plant breeding data sets often involve a series of trials over locations and years in which the genetic entries differ across environments. In addition, some data are often missing due to practical problems, and even within environments, experimental designs are often not balanced. The lack of balance impacts two-stage analyses in several ways. First, the BLUPs of lines that are represented by fewer records in the data set are shrunk back to the population mean to a greater extent than lines with more records. Second, the BLUEs or BLUPs obtained from the mixed model analysis of an unbalanced data set have variable standard errors. The variation in precision among the BLUEs or BLUPs is ignored in the second stage analysis, resulting in a loss of information. Simulation studies (Wang *et al.* 2011) indicate that unbalanced data in two-stage GWAS can cause more false positives.

Methods for analyzing a series of unbalanced performance evaluations of crops have been considered in detail in the context of maximizing the precision and accuracy of marginal predictions of the genetic entries (Möhring and Piepho 2009; SMITH *et al.* 2009). In this context, single-step analysis is considered optimal, but may have high computational demand. Two-stage analysis of crop performance trials involves analyzing individual trials separately, then using family BLUEs from each trial as dependent variables in a simplified second stage analysis. Two stage analysis methods that use weighted analysis in the second

step, in which weights are proportional to the precision of the BLUEs from the first step, often provide close approximation to the results of a single stage analysis (Möhring and Piepho 2009; SMITH *et al.* 2009). Additional complexity in the two stage analysis occurs when the residual values within environments are not independent, as occurs when spatial correlations are modeled in the residual variance structure. This results in lack of independence among the BLUEs; however approximate and exact methods have been developed to account for this lack of dependence as well as the variable precision among BLUEs in the second stage of analysis (Möhring and Piepho 2009; Piepho *et al.* 2012).

Inspired by previous work on two stage analysis of crop performance trials, George and Cavanagh (2015) proposed a two-stage GWAS approach that weights the BLUEs for families from the first stage in the linear mixed model GWAS scan. Their results indicate that the weighted two-stage GWAS provided comparable results to the single stage GWAS, and suggest that weighted two-stage analysis appears is a useful approach for conducting GWAS using data from multi-environment plant breeding trials. Several questions about the use of two-stage GWAS remain unanswered, however. First, it is unclear which summary variable is appropriate to use as a dependent variable for second-stage GWAS (Pasam *et al.* 2012a). In particular, the use of BLUP in two-stage analyses in which the hypothesis test is conducted only in the second stage has been criticized (Hadfield *et al.* 2010).

Alternate approach of using residuals from a first stage mixed model accounting for genomic relationships as dependent variables in the second stage may also be considered. Second none of the specialized open-source GWAS software packages have the flexibility to incorporate weights in the residual variance structure. In Chapter 3 of this dissertation, I report the results a simulation study to compare the use of different summary variables in the

first stage of multiple stage analyses. This allowed me to draw conclusions regarding the power and bias of QTL effect estimates for three different simulated genetic architectures are drawn. In addition, I proposed a weighted GWAS method and implemented in a free open source software for easy use.

Unusual population structures – correcting for population structure and for segregation within phenotyped families.

The ZeaSyn6 population is a unique and useful resource for genetic analysis of the effects of diverse maize and teosinte alleles in a common gene pool. The numerous cycles of random mating are expected to eliminate population structure and to reduce linkage disequilibrium, potentially resulting in both high power and high resolution GWAS with minimal complications from population structure. However, this population design introduced some novel features that are not handled by standard GWAS analysis packages. First, the phenotypic measurements are averages over many progenies within each family, but the genotypic data are based on sequencing the parents. The progenies are segregating at many loci, so the probabilities of different genotypic classes within each family are needed, and these can be derived based on the parental genotypes. GWAS can be conducted as multiple regression of the family phenotype means on the expected additive and dominance effect coefficients for each family. Second, the many generations of random mating used to generate the ZeaSyn 6^{th} generation population are expected to reduce linkage disequilibrium and eliminate population structure among the parents. However, since the experiment includes one selfed family and one outcross family derived from each of 923 male parents, there are close pedigree relationships within each pair of families from a common parent.

The appropriate adjustment for this population structure is not obvious, since correcting directly for pedigree relationships would absorb much of polygenic variation. Finally, the phenotyped families differ for level of inbreeding, and adjustment for the inbreeding coefficient is necessary to remove confounding effects of inbreeding depression at unlinked regions of the genome. In Chapter 4 of this dissertation, I discuss several different approaches for correcting for population structure and inbreeding incorporated into an efficient GWAS algorithm that simultaneously tests additive and dominance effects from the mean values of segregating progenies.

Testing biological hypotheses with quantitative genetics parameter estimates

Historically, quantitative genetics has provided 'summary statistics' for the combined effects of many genes throughout the genome. This has provided very useful information on overall heritability and genetic correlations among traits, as well as improved breeding value estimates based on expected genetic relationships using pedigree information. These classical ideas can be combined with modern genomics tools to make more refined estimates of the effects and variation associated with specific variants, genes, genomic regions, or classes of genetic variants. The overall goals of this thesis were to implement statistical genetics techniques to address specific biological questions, and to develop new methods as needed to apply such tests to novel data and population structures. My objective was to go beyond summary descriptions of the overall effects of the many genes involved in controlling complex traits in maize to instead describe the effects of different groups of genetic variants.

In Chapter 2, I performed a variety of tests to determine the relative importance of large effect variants, background polygenic variance, and genomic regions implicated as

targets of selection during domestication and improvement phases of maize population history to the standing variation for domestication traits in maize. In Chapter 3, I developed a method to perform GWAS in plant breeding data sets with highly unbalanced structures. In Chapter 4, I address questions of what kinds of genomic variants contribute most to inbreeding depression and trait variation in a population segregating for both maize and teosinte alleles. I relate the distribution of genetic load variants to recombination rate and allele frequency. Future work on this population may also permit the identification of favorable alleles derived from teosinte that could be useful to maize improvement.

References

Amin, N., C. van Duijn M. and Y. S. Aulchenko, 2007 A genomic background based method for association analysis in related individuals. PLoS ONE **2:** e1274.

Anderson, L. K., K. D. Hooker and S. M. Stack, 2001 The distribution of early recombination nodules on zygotene bivalents from plants. Genetics **159:** 1259-1269.

Aranzana, M. J., S. Kim, K. Zhao, E. Bakker, M. Horton *et al*, 2005 Genome-wide association mapping in *arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet **1:** e60.

Asimit, J., and E. Zeggini, 2010 Rare variant association analysis methods for complex traits. Annu. Rev. Genet. **44:** 293-308.

Aulchenko, Y. S., D. de Koning and C. Haley, 2007 Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics **177:** 577-585.

Badstue, L. B., M. R. Bellon, J. Berthaud, A. Ramírez, D. Flores *et al*, 2007 The dynamics of farmers' maize seed supply practices in the central valleys of oaxaca, mexico. World Dev. **35:** 1579-1593.

Beadle, G. W., 1939 Teosinte and the origin of maize. J. Hered. **30:** 245-247.

Bian, Y., and J. Holland, 2017 Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. Heredity .

Birchler, J. A., H. Yao, S. Chudalayandi, D. Vaiman and R. A. Veitia, 2010 Heterosis. Plant Cell **22:** 2105-2112.

Brachi, B., G. P. Morris and J. O. Borevitz, 2011 Genome-wide association studies in plants: The missing heritability is in the field. Genome Biol. **12:** 232-232.

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

Brown, A. H. D., 2010 Variation under domestication in plants: 1859 and today. Philos. Trans. R. Soc. Lond. , B, Biol. Sci. **365:** 2523.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown *et al*, 2009 The genetic architecture of maize flowering time. Science **325:** .

BUCKLER, E. S., J. M. THORNSBERRY and S. KRESOVICH, 2001 Molecular diversity, structure and domestication of grasses. Genet. Res. **77:** 213-218.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown *et al*, 2009 The genetic architecture of maize flowering time. Science **325:** 714-718.

Charlesworth, D., and J. H. Willis, 2009 The genetics of inbreeding depression. Nature Reviews Genetics **10:** 783-796.

Che, R., A. Motsinger-Reif and C. C. Brown, 2012 Loss of power in two-stage residual-outcome regression analysis in genetic association studies. Genet. Epidemiol. **36:** .

Cheng, R., C. C. Parker, M. Abney and A. A. Palmer, 2013 Practical considerations regarding the use of genotype and pedigree data to model relatedness in the context of genome-wide association studies. G3 (Bethesda) **3:** 1861-1867.

Clark, R. M., E. Linton, J. Messing and J. F. Doebley, 2003 Pattern of diversity in the genomic region near the maize domestication gene tb1. Proc. Natl. Acad. Sci. U. S. A. **101:** 700-707.

Collard, B. C. Y., and D. J. Mackill, 2007 Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences **363:** 557-572.

Colome-Tatche, M., S. Cortijo, R. Wardenaar, L. Morgado, B. Lahouze *et al*, 2012 Features of the arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proc. Natl. Acad. Sci. U. S. A. **109:** 16240-16245.

Cross-Disorder Group of the Psychiatric,Genomics Consortium, 2013 Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat. Genet. **45:** 984-994.

Crouch, J., M. Warburton and Y. JianBing.2011 Association Mapping for Enhancing Maize *(Zea Mays L.)* Genetic Improvement.

Crow, J. F., 2000 The rise and fall of overdominance. Plant Breeding Reviews, Volume 17 225-257.

Demissie, S., and L. A. Cupples, 2011 Bias due to two-stage residual-outcome regression analysis in genetic association studies. Genet. Epidemiol. **35:** .

Doebley, J., and A. Stec, 1993 Inheritance of the morphological differences between maize and teosinte: Comparison of results for two F2 populations. Genetics **134:** 559-570.

Doebley, J., A. Stec, J. Wendel and M. Edwards, 1990 Genetic and morphological analysis of a maize-teosinte F2 population: Implications for the origin of maize. Proc. Natl. Acad. Sci. U. S. A. **87:** 9888-9892.

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

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

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

Doerge, R. W., 2002 Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet. **3:** 43-52.

Duvick, D., J. Smith and M. Cooper, 2010 Long-term selection in a commercial hybrid maize breeding program. Janick.I.Plant Breeding Reviews.Part **2:** 109-152.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal *et al*, 2010 Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews.Genetics **11:** 446-450.

Fu, H., and H. K. Dooner, 2002 Intraspecific violation of genetic colinearity and its implications in maize. Proc. Natl. Acad. Sci. U. S. A. **99:** 9573-9578.

Galinat, W. C., 1983 The origin of maize as shown by key morphological traits of its ancestor, teosinte. Maydica .

Gardner, C., and J. Lonnquist, 1959 Linkage and the degree of dominance of genes controlling quantitative characters in maize. Agron. J. **51:** 524-528.

Gaut, B. S., S. I. Wright, C. Rizzon, J. Dvorak and L. K. Anderson, 2007 Recombination: An underappreciated factor in the evolution of plant genomes. Nature Reviews Genetics **8:** 77-84.

Gerke, J. P., J. W. Edwards, K. E. Guill, J. Ross-Ibarra and M. D. McMullen, 2015 The genomic impacts of drift and selection for hybrid performance in maize. Genetics .

Gibson, G., 2012 Rare and common variants: Twenty arguments. Nature Reviews Genetics **13:** 135-145.

Goodman, M. M., 1988 The history and evolution of maize. Crit. Rev. Plant Sci. **7:** 197-220.

Goodman, M. M., 1968 The races of maize: II. use of multivariate analysis of variance to measure morphological similarity. Crop Sci. **8:** 693-698.

Goodman, M. M., and R. M. Bird, 1977 The races of maize IV: Tentative grouping of 219 latin american races. Econ. Bot. **31:** 204-221.

Goodman, M. M., and E. Paterniani, 1969 The races of maize: III. choices of appropriate characters for racial classification. Econ. Bot. **23:** 265-273.

Goodman, M., W. Brown, G. Sprague and J. Dudley, 1988 Corn and corn improvement. American Society of Agronomy 33-79.

Grossniklaus, U., G. A. Nogler and P. J. van Dijk, 2001 How to avoid sex: The genetic control of gametophytic apomixis. Plant Cell **13:** 1491-1498.

Gusev, A., S. . Lee, G. Trynka, H. Finucane, B. Vilhjálmsson *et al*, 2014 Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. The American Journal of Human Genetics **95:** 535-552.

Gusev, A., G. Bhatia, N. Zaitlen, B. J. Vilhjalmsson, D. Diogo *et al*, 2013 Quantifying missing heritability at known GWAS loci. PLOS Genetics **9:** e1003993.

Hadfield, J. D., A. J. Wilson , D. Garant, B. C. Sheldon and L. E. B. Kruuk, 2010 The misuse of BLUP in ecology and evolution. Am. Nat. **175:** 116-125.

Hammer, K., 1984 Das domestikationssyndrom. Die Kulturpflanze **32:** 11-34.

Harlan, J. R., 1992 Crops & Man. American Society of Agronomy, Madison, WI, USA.

Harlan, J. R., de Wet, J. M. J. and E. G. Price, 1973 Comparative evolution of cereals. Evolution **27:** 311-325.

HARRIS, D. R., 1990 3. vavilov's concept of centres of origin of cultivated plants: Its genesis and its influence on the study of agricultural origins. Biol. J. Linn. Soc. **39:** 7-16.

Hartfield, M., and S. P. Otto, 2011 RECOMBINATION AND HITCHHIKING OF DELETERIOUS ALLELES. Evolution **65:** 2421-2434.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269-294.

Holland, J. B., and N. D. Coles, 2011 QTL controlling masculinization of ear tips in a maize (zea mays L.) intraspecific cross. G3: Genes|Genomes|Genetics **1:** 337-341.

Hufford, M. B., X. Xu, J. van Heerwaarden, T. Pyhajarvi, J. Chia *et al*, 2012 Comparative population genomics of maize domestication and improvement. Nat. Genet. **44:** 808-811.

Hung, H., L. M. Shannon, F. Tian, P. J. Bradbury, C. Chen *et al*, 2012 ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proceedings of the National Academy of Sciences **109:** E1913-E1921.

Iltis, H. H., 1983 From teosinte to maize: The catastrophic sexual transmutation. Science **222:** 886-894.

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin *et al*, 2004 Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. **14:** 528-538.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al*, 2008 Efficient control of population structure in model organism association mapping. Genetics **178:** .

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Kong *et al*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42:** 348-354.

Korte, A., and A. Farlow, 2013 The advantages and limitations of trait analysis with GWAS: A review. Plant methods **9:** 29.

Kump, K. L., P. J. Bradbury, E. S. Buckler, A. R. Belcher, M. Oropeza-Rosas *et al*, 2011 Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat. Genet. **43:** .

Laird, N. M., S. Horvath and X. Xu, 2000 Implementing a unified approach to family-based tests of association. Genet. Epidemiol. **19:** S36-S42.

Lam, A. C., M. Schouten, Y. S. Aulchenko, C. S. Haley and D. de Koning, 2007 Rapid and robust association mapping of expression quantitative trait loci. BMC Proceedings **1:** S144-S144.

Lander, E. S., 1996 The new genomics: Global views of biology. Science **274:** 536.

Lee, S. H., T. R. DeCandia, S. Ripke, J. Yang, The Schizophrenia Psychiatric Genome Wide Association,Study Consortium *et al*, 2012 Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. **44:** 247-250.

Lee, S., N. Wray, M. Goddard and P. Visscher, 2011 Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. **88:** 294-305.

Lewontin, R. C., 1974 The Genetic Basis of Evolutionary Change. Columbia University Press New York.

Li, Y., Y. Huang, J. Bergelson, M. Nordborg and J. O. Borevitz, 2010 Association mapping of local climate-sensitive quantitative trait loci in arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. **107:** 21199-21204.

Lipka, A. E., M. A. Gore, M. Magallanes-Lundback, A. Mesberg, H. Lin et al.2013 Genome-Wide Association Study and Pathway-Level Analysis of Tocochromanol Levels in Maize Grain.

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson *et al*, 2011 FaST linear mixed models for genome-wide association studies. Nat Meth **8:** 833-835.

Listgarten, J., C. Lippert and D. Heckerman, 2013 FaST-LMM-select for addressing confounding from spatial structure and rare variants. Nat. Genet. **45:** 470-471.

Listgarten, J., C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin *et al*, 2012 Improved linear mixed models for genome-wide association studies. Nature methods **9:** 525-526.

Mackay, T. F. C., E. A. Stone and J. F. Ayroles, 2009 The genetics of quantitative traits: Challenges and prospects. Nat. Rev. Genet. **10:** 565-577.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al*, 2009 Finding the missing heritability of complex diseases. Nature **461:** .

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, J. Sanchez G., E. Buckler *et al*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. U. S. A. **99:** 6080-6084.

Mauricio, R., 2001 Mapping quantitative trait loci in plants: Uses and caveats for evolutionary biology. Nature Reviews Genetics **2:** 370-381.

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li *et al*, 2009 Genetic properties of the maize nested association mapping population. Science **325:** .

Meyer, R. S., and M. D. Purugganan, 2013 Evolution of crop species: Genetics of domestication and diversification. Nat. Rev. Genet. **14:** 840-852.

Mezmouk, S., and J. Ross-Ibarra, 2013 The pattern and distribution of deleterious mutations in maize. G3: Genes|Genomes|Genetics **4:** 163-171.

Mirouze, M., M. Lieberman-Lazarovich, R. Aversano, E. Bucher, J. Nicolet *et al*, 2012 Loss of DNA methylation affects the recombination landscape in arabidopsis. Proc. Natl. Acad. Sci. U. S. A. **109:** 5880-5885.

Möhring, and Piepho, 2009 Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci. **49:** .

Moll, R. H., M. F. Lindsey and H. F. Robinson, 1963 Estimates of genetic variances and level of dominance in maize. Genetics **49:** 411-423.

Morden, C. W., J. Doebley and K. F. Schertz, 1990 Allozyme variation among the spontaneous species of sorghum section sorghum (poaceae). Theor. Appl. Genet. **80:** 296-304.

Morrell, P. L., E. S. Buckler and J. Ross-Ibarra, 2012 Crop genomics: Advances and applications. Nat. Rev. Genet. **13:** 85-96.

Myles, S., J. Peiffer, P. J. Brown, E. S. Ersoz, Z. Zhang *et al*, 2009 Association mapping: Critical considerations shift from genotyping to experimental design. Plant Cell **21:** .

Naylor, M. G., S. T. Weiss and C. Lange, 2009 Recommendations for using standardised phenotypes in genetic association studies. Hum Genomics **3:** .

Nordborg, M., and D. Weigel, 2008 Next-generation genetics in plants. Nature **456:** 720-723.

Oakey, H., A. P. Verbyla, B. R. Cullis, X. Wei and W. S. Pitchford, 2007 Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. Theor. Appl. Genet. **114:** 1319-1332.

Oka, H. I. 1988 Origin of Cultivated Rice Electronic Resource]. Japan Scientific Societies Press; Elsevier; Exclusive sales rights for the U.S.A. and Canada, Elsevier Science Pub. Co, Tokyo; Amsterdam Netherlands] ; New York; New York, N.Y.

Olsen, K. M., and J. F. Wendel, 2013 A bountiful harvest: Genomic insights into crop domestication phenotypes. Annual Review of Plant Biology **64:** 47-70.

Pasam, R. K., R. Sharma, M. Malosetti, F. A. van Eeuwijk, G. Haseneyer *et al*, 2012a Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biology **12:** 1-22.

Pasam, R., R. Sharma, M. Malosetti, F. van Eeuwijk, G. Haseneyer *et al*, 2012b Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biology **12:** 1-22.

Peiffer, J. A., S. Flint-Garcia, N. De Leon, M. D. McMullen, S. M. Kaeppler *et al*, 2013 The genetic architecture of maize stalk strength. PLoS ONE **8:** e67066.

Piepho, H., J. Möhring, T. Schulz-Streeck and J. O. Ogutu, 2012 A stage-wise approach for the analysis of multi-environment trials. Biometrical Journal **54:** 844-860.

Piperno, D. R., A. J. Ranere, I. Holst, J. Iriarte and R. Dickau, 2009 Starch grain and phytolith evidence for early ninth millennium B.P. maize from the central balsas river valley, mexico. Proceedings of the National Academy of Sciences **106:** 5019-5024.

Poland, J. A., P. J. Bradbury, E. S. Buckler and R. J. Nelson, 2011 Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proc. Natl. Acad. Sci. U. S. A. **108:** 6893-6898.

Pressoir, G., and J. Berthaud, 2004 Patterns of population structure in maize landraces from the central valleys of oaxaca in mexico. Heredity **92:** 88-94.

Rodgers-Melnick, E., P. J. Bradbury, R. J. Elshire, J. C. Glaubitz, C. B. Acharya *et al*, 2015 Recombination in diverse maize is stable, predictable, and associated with genetic load. Proceedings of the National Academy of Sciences **112:** 3823-3828.

Romay, M. C., M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts *et al*, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. **14:** .

Ross-Ibarra, J., P. L. Morrell and B. S. Gaut, 2007 Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc. Natl. Acad. Sci. U. S. A. **104:** 8641-8648.

Ruiz Corral, J. A., N. Durán Puga, Sánchez González, José de Jesús, J. Ron Parra, D. R. González Eguiarte *et al*, 2008 Climatic adaptation and ecological descriptors of 42 mexican maize races. Crop Sci. **48:** 1502-1512.

Schork, A. J., W. K. Thompson, P. Pham, A. Torkamani, J. C. Roddey *et al*, 2013 All SNPs are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLOS Genetics **9:** e1003449.

Serres-Giardi, L., K. Belkhir, J. David and S. Glemin, 2012 Patterns and evolution of nucleotide landscapes in seed plants. Plant Cell **24:** 1379-1397.

Slotkin, R. K., and R. Martienssen, 2007 Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics **8:** 272-285.

SMITH, A. B., B. R. CULLIS and R. THOMPSON, 2009 The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. The Journal of Agricultural Science **143:** 449-462.

Speed, D., G. Hemani, M. Johnson and D. Balding, 2012 Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. **91:** 1011-1021.

Spindel, J., H. Begum, D. Akdemir, B. Collard, E. Redoña *et al*, 2016 Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity .

Studer, A., Q. Zhao, J. Ross-Ibarra and J. Doebley, 2011 Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. **43:** 1160-1163.

Studer, A. J., and J. F. Doebley, 2012 Evidence for a natural allelic series at the maize domestication locus teosinte branched1. Genetics **191:** 951-U533.

Tenaillon, M. I., J. U'Ren, O. Tenaillon and B. S. Gaut, 2004 Selection versus demography: A multilocus investigation of the domestication process in maize. Mol. Biol. Evol. **21:** 1214-1225.

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun *et al*, 2011 Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. **43:** .

Tracy, W., and M. Chandler.2006 The Historical and Biological Basis of the Concept of Heterotic Patterns in Corn Belt Dent Maize. Wiley Online Library.

van Heerwaarden, J., J. Doebley, W. H. Briggs, J. C. Glaubitz, M. M. Goodman *et al*, 2010 Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. U. S. A. **108:** 1088-1092.

van Heerwaarden, J., M. B. Hufford and J. Ross-Ibarra, 2012 Historical genomics of north american maize. Proc. Natl. Acad. Sci. U. S. A. **109:** 12420-12425.

Vigouroux, Y., J. C. Glaubitz, Y. Matsuoka, M. M. Goodman, G. J. Sanchez *et al*, 2008 Population structure and genetic diversity of new world maize races assessed by DNA microsatellites. Am. J. Bot. **95:** 1240-1253.

Wang, H., K. P. Smith, E. Combs, T. Blake, R. D. Horsley *et al*, 2011 Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theor. Appl. Genet. **124:** 111-124.

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

Weatherwax, P., 1954 Indian corn in old america.

Weber, A. L., W. H. Briggs, J. Rucker, B. M. Baltazar, J. de Jesús Sánchez-Gonzalez *et al*, 2008 The genetic architecture of complex traits in teosinte (zea mays ssp. parviglumis): New evidence from association mapping. Genetics **180:** 1221-1232.

Wills, D. M., C. J. Whipple, S. Takuno, L. E. Kursel, L. M. Shannon *et al*, 2013 From many, one: Genetic control of prolificacy during maize domestication. PLoS Genet **9:** e1003604.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al*, 2005 The effects of artificial selection on the maize genome. Science **308:** 1310-1314.

Yang, J., T. Lee, J. Kim, M. Cho, B. Han *et al*, 2013 Ubiquitous polygenicity of human complex traits: Genome-wide analysis of 49 traits in koreans. PLoS Genet **9:** e1003355.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. **42:** 565-569.

Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso *et al*, 2011 Genome-partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. **43:** 519-525.

Yelina, N. E., K. Choi, L. Chelysheva, M. Macaulay, B. De Snoo *et al*, 2012 Epigenetic remodeling of meiotic crossover frequency in arabidopsis thaliana DNA methyltransferase mutants. PLoS Genet **8:** e1002844.

Yu, J., J. B. Holland, M. D. McMullen and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics **178:** 539-551.

Zaitlen, N., and P. Kraft, 2012 Heritability in the genome-wide association era. Hum. Genet. **131:** 1655-1664.

Zaitlen, N., P. Kraft, N. Patterson, B. Pasaniuc, G. Bhatia *et al*, 2013 Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet **9:** e1003520.

Zeegers, M., F. Rijsdijk and P. Sham, 2004 Adjusting for covariates in variance components QTL linkage analysis. Behav. Genet. **34:** 127-133.

Zhang, L., A. S. Peek, D. Dunams and B. S. Gaut, 2002 Population genetics of duplicated disease-defense genes, hm1 and hm2, in maize (zea mays ssp. mays L.) and its wild ancestor (zea mays ssp. parviglumis). Genetics **162:** 851-860.

Zhang, Z., E. S. Buckler, T. M. Casstevens and P. J. Bradbury, 2009 Software engineering the mixed model for genome-wide association studies on large samples. Briefings in Bioinformatics **10:** 664-675.

Zhang, Z., E. Ersoz, C. Lai, R. J. Todhunter, H. K. Tiwari *et al*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. **42:** 355-360.

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. **44:** 821-824.

Zhu, C., M. Gore, E. S. Buckler and J. Yu, 2008 Status and prospects of association mapping in plants. The Plant Genome Journal **1:** .

CHAPTER 2: Genetic architecture of domestication-related traits in maize

Citation:

Xue, S., P. Bradbury, T. Casstevens and J. B. Holland, 2016 Genetic architecture of domestication-related traits in maize. Genetics 204: 99-113.

Genetic architecture of domestication-related traits in maize

Shang Xue[1], Peter J. Bradbury[2], Terry Casstevens[3], and James B. Holland[4]

[1] Bioinformatics Research Center, North Carolina State University, Raleigh, NC,USA

[2] USDA-ARS Plant, Soil, and Nutrition Research Unit, Ithaca, NY 14853, USA

[3] Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853-2703, USA

[4] USDA-ARS Plant Science Research Unit and Department of Crop Science, North Carolina State University, Raleigh, NC 27695-7620, USA. E-mail: james_holland@ncsu.edu

Abstract

Strong directional selection occurred during the domestication of maize from its wild ancestor teosinte, reducing its genetic diversity, particularly at genes controlling domestication-related traits. Nevertheless, variability for some domestication-related traits is maintained in maize. The genetic basis of this could be sequence variation at the same key genes controlling maize-teosinte differentiation (due to lack of fixation or arising as new mutations after domestication), distinct loci with large effects, or polygenic background variation. Previous studies permit annotation of maize genome regions associated with the major differences between maize and teosinte or that exhibit population genetic signals of selection during either domestication or post-domestication improvement. Genome-wide association studies and genetic variance partitioning analyses were performed in two diverse maize inbred line panels to compare the phenotypic effects and variances of sequence polymorphisms in regions involved in domestication and improvement to the rest of the genome. Additive polygenic models explained most of the genotypic variation for domestication-related traits; no large effect loci were detected for any trait. Most trait variance was associated with background genomic regions lacking previous evidence for involvement in domestication. Improvement sweep regions were associated with more trait variation than expected based on the proportion of the genome they represent. Selection during domestication eliminated large effect genetic variants that would revert maize toward a teosinte type. Small effect polygenic variants (enriched in the improvement sweep regions of the genome) are responsible for most of the standing variation for domestication-related traits in maize.

Keywords: Quantitative trait loci, nested association mapping, genome wide association study, variance components, *Zea mays*

Introduction

The domestication of all major crop plants occurred in a relatively short period in human history starting about 10,000 years ago (Harlan 1992). During the domestication process, seeds of preferred forms were selected and saved to plant subsequent generations. Some alleles favored under domestication may have been neutral or even deleterious for the survival of wild plant species; for example, seed shattering promotes seed dispersal in wild grasses, but alleles for non-disarticulating seed structures were strongly selected for under domestication (Galinat 1983). Consequently, rare alleles favorable for growth and development under agricultural conditions or for traits desired by humans increased in frequency, often reaching fixation and reducing genetic variation very near causal sequence sites (Wang *et al.* 1999). In addition, domestication was often accompanied by severe genetic bottlenecks from the use of small founder populations. The reduction in effective population sizes also resulted in reduced genetic diversity genome-wide. Population genetics methods to model the strength and duration of bottlenecks provide a means to distinguish domestication-associated selection sweeps from reduced diversity due to genetic drift (Meyer and Purugganan 2013; Wright *et al.* 2005).

The details of crop demographic histories are generally unknown and may involve factors that complicate the modelling of genetic bottlenecks and selection sweeps. Such complications may include soft sweeps and incomplete fixation at domestication loci, post-domestication gene flow between crops and their wild ancestors, and the balance between post-domestication directional 'improvement selection' versus genetic diversification from selection for adaptation to new environments and distinct crop uses in different populations

(Darwin 1868; Hufford *et al.* 2012; Meyer and Purugganan 2013; van Heerwaarden *et al.* 2011). Integrating information about the genetic architecture of domestication traits with population genetics data can help refine the understanding of the contribution of sequence variation to domestication and post domestication developmental and morphological changes in crops.

Maize was domesticated about 6000 to 10000 years ago from a wild grass, teosinte, in southwestern Mexico (Galinat 1983; Iltis 1983; Matsuoka *et al.* 2002). Numerous morphological traits have changed in maize compared to its wild ancestor, including the floral morphology (Doebley *et al.* 1990; Iltis 1983). Teosinte plants have elongated lateral branches at many nodes. In contrast, maize plants typically produce a lateral branch at only two or three of the nodes on their main stems, and these are much shorter than teosinte lateral branches, being reduced to a "shank" that terminates at the base of a female ear (Doebley *et al.* 1997). Furthermore, teosinte "ears" are small, with kernels arranged in a distichous (two-ranked) pattern on the ear axis, compared to large ears of maize that typically have from eight to over twenty rows of kernels in four or more ranks. Several major QTL and in some cases, the specific genes, controlling these differences between maize and teosinte have been identified (Briggs *et al.* 2007; Clark *et al.* 2003; Doebley 2004; Weber *et al.* 2008).

The strong directional selection that occurred during the domestication of maize from teosinte reduced genetic diversity most strongly at key genes controlling domestication-related traits. Despite the severe bottleneck that occurred during domestication and strong selection for the maize plant type, standing variability in cob length, kernel row number, and shank length can be observed among maize breeding lines. In addition, although most maize plants have purely female flowers on their lateral branch termini, some lines of maize

produce a spike of staminate florets at the tips of their ears (Holland and Coles 2011), referred to as “masculinized ear tips”, revealing variation for this domestication trait as well.

The genetic architecture for standing variation remaining in maize for these domestication-related traits is unknown. Sequence variation at the same set genes that were involved in the conversion of teosinte into domesticated maize may cause some portion of this variation. Several large-effect mutations that cause maize to exhibit teosinte-like morphological characteristics, such as *tb1* and *gt1*, were later demonstrated to be allelic to the corresponding domestication loci (Doebley *et al.* 2006; Studer *et al.* 2011; Wills *et al.* 2013). Not all domestication alleles are fixed in domesticated species (Meyer and Purugganan 2013; Studer *et al.* 2011), leaving open the possibility that some domestication loci contribute to standing trait variation in domesticated species. Furthermore, a range of allelic series exists at some domestication loci (Studer and Doebley 2012); smaller-effect alleles may segregate in domesticated species even if larger-effect wild-type alleles were lost from the species. Smaller-effect variants could have originated in the wild ancestor and passed through the domestication bottleneck because of lower selection intensity or may have arisen by mutation after domestication.

Alternatively, the observed phenotypic variation for domestication traits within domesticated species might be due to large-effect genes that are distinct from the known domestication genes. The variants at these genes may have arisen after domestication, or had effects sufficiently small as to avoid purging during domestication. For example, the Suppressor of sessile spikelets 1 mutation in maize changes the morphology of maize ears by changing the paired spikelets of maize florets into single spikelets, as found in teosinte ears (Doebley *et al.* 1995), making it a candidate domestication gene. However, genetic analysis

of this mutation demonstrated that it does not complement the teosinte allele that controls single versus paired ear spikelets, so it was not under selection during domestication (Doebley *et al.* 1995).

A third possibility is that the observed phenotypic variation for domestication traits is produced by many small-effect variants distributed throughout the genome, resulting in a polygenic genetic architecture. Even if major-effect alleles were fixed during domestication, smaller-effect variants at other loci could cause phenotypic variation in domestication target traits. Again, these variants could have existed in the wild ancestor and passed through the domestication bottleneck due to small selection coefficients, or they may represent new variation that arose from mutation following domestication. To test these hypotheses, phenotypic evaluations of domestication-related traits and genotypic data of two diverse maize populations were combined in this study to facilitate estimation of the proportion of variation due to polygenic, small-effect QTL versus larger effect variants, and to compare the genomic positions of larger effect variants to the known locations of domestication genes.

Numerous association mapping panels are already available in maize (Crouch *et al.* 2011). The largest and most diverse such population that is publicly available is a set of 2,815 inbred lines maintained by the U.S. Department of Agriculture North Central Region Plant Introduction Station (NCRPIS) (Romay *et al.* 2013). This collection contains lines representing nearly a century of maize breeding efforts from programs throughout the world and has been densely genotyped (Romay *et al.* 2013). An alternative to conducting genome-wide association study in samples of breeding lines, which are characterized by complex population structure, is to use multiple parent populations of known pedigree, with known population structure. One such population is the maize Nested Association Mapping (NAM)

population, which consists of 4892 recombinant inbred lines (RILs) derived from 25 bi-parental families (Buckler *et al.* 2009; Hung *et al.* 2012; Tian *et al.* 2011; Yu *et al.* 2008). High resolution GWAS can be conducted in NAM while controlling for the known pedigree structure and for genetic variation at unlinked QTL detected by joint multiple population linkage analysis (Kump *et al.* 2011; Tian *et al.* 2011).

Using these two diverse maize populations, GWAS and joint linkage QTL mapping were conducted to estimate the relative contributions of polygenic additive background and specific QTLs and SNP variants with larger effects on phenotypic variation for domestication-related traits. QTLs and SNP associations were compared to regions previously identified to contain QTL controlling morphological differences between teosinte and maize ('domestication QTL'; Briggs *et al.* 2007; Doebley 2004) or previously shown to exhibit signals of selection during domestication or post-domestication improvement ('domestication or improvement sweep regions'; Hufford *et al.* 2012). The association of multi-SNP haplotypes at several candidate genes with variation in domestication-related traits was also tested. Finally, we estimated the proportion of trait variation associated with additive polygenic variation in candidate QTL, domestication, and improvement regions using variance partitioning methods.

Material and Methods

NCRPIS inbred panel

The U.S. Department of Agriculture North Central Region Plant Introduction Station (NCRPIS) in Ames, IA maintains a public collection of seeds of 2,572 maize inbred line accessions. This represents most of the publicly available maize inbred lines worldwide (Romay *et al.* 2013). In 2010, almost all inbred lines from the USDA seed bank collection (2572 inbred line entries) were evaluated for domestication-related traits in Clayton, North Carolina. The experimental design was a single-replicate, augmented design. Experimental entries were divided into 9 maturity groups of differing sizes. Each maturity group was randomly divided into two sets and one of each set was planted in each of two field blocks. Each set-block combination was augmented by the addition of one B73 inbred check plot and one of five other check inbreds (IL14H, Ki11, P39, SA24, and Tx303, depending on maturity group). The check plots were assigned to random positions within each set-block combination.

In 2012, a subset of 771 diverse inbreds was evaluated at the same location. Sets were randomized within the field within one year, and each block was augmented by a randomly assigned B73 check plot. Five other checks (GE440, NC358, NK794, PHB47, and Tx303) representing different maturities were included once per set. In 2013, two replicates of a core diversity panel were evaluated at the same location using a randomized complete block design. This panel consists of 279 inbred lines representing a large portion of the available geographic and molecular diversity of publicly available maize inbreds (Flint-Garcia *et al.* 2005). The core diversity panel is a subset of the NCRPIS collection and was included in the

771 lines tested in 2012, so we consider the complete data set consists of 2572 entries, but the design is unbalanced, with most lines evaluated in only one year, some lines evaluated in two years, and the lines from the core diversity set evaluated in three years.

Two plants in each plot were measured for several domestication-related traits. Shank length was measured as the length from the bottom of the ear to the main stalk. Cob length was measured as the length from the top of the ear (not including masculinized ear tips) to the bottom of the ear. Masculinized ear tip length was measured as the length of ear segments bearing anthers. Ear row number was counted on a transverse section of each cob.

Genotypic data of diversity panel

Genotyping by sequencing (GBS), a low-cost, high-throughput sequencing approach (Elshire *et al.* 2011; Glaubitz *et al.* 2014) was used to genotype the complete set of lines (Romay *et al.* 2013; Zila *et al.* 2014). The method produced 681,257 single-nucleotide polymorphism (SNP) markers distributed across the entire genome, with the ability to detect rare alleles at high confidence levels (Romay *et al.* 2013). After the initial imputation described in Romay *et al*. (2013), ~16% of line-marker combinations were still missing. Therefore, an additional imputation was performed using Beagle 4.0 (Browning and Browning 2009). After imputation with Beagle, a set of 405,315 SNPs with estimated imputation accuracy more than 0.96 and with minor alleles observed as homozygous in at least 20 lines was used for further analyses.

A subset of 111,282 SNP markers was used to estimate the realized genomic additive genetic relationship matrix (**G**) among the complete set of 2480 lines. This subset of markers had estimated imputation accuracy of at least 0.995 and was subjected to linkage-

disequilibrium pruning by PLINK (Purcell *et al.* 2007) to result in markers with no pairwise genotypic correlation greater than 0.5. The realized additive relationship matrix (File A.6) was estimated using R software version 3.0.0 (R Core Team 2013) based on observed allele frequencies (VanRaden 2008).

NAM population

The maize NAM population consists of 25 bi-parental families, each of which has B73 as a common parent and one of 25 diverse lines as the second parent. Each family has about 200 RILs, resulting in a total population size of 4892 (Bian *et al.* 2014; McMullen *et al.* 2009). Cob length values for NAM RILs were taken from Hung et al (Hung *et al.* 2011). To measure the other domestication-related phenotypes described previously, the NAM population was grown in Clayton, North Carolina in 2012 using an augmented sets design, wherein each family was a set and lines were randomized to sub-blocks of 22 plots, which each contained one plot of each parental line (Hung *et al.* 2011). We measured the domestication-related phenotypes on all lines of five families (B73×B97, B73×CML52, B73×HP301, B73×Il14H, and B73×M162W) and on 40 random lines of the remaining families. All RILs of eight families (B73×CML103, B73×CML247, B73×CML69, B73×Il14H, B73×KI11, B73×M37W, B73×M18W, B73×P39) were evaluated in 2013 using a similar randomized augmented design. These families were chosen because they had the largest genetic variance for shank length among all NAM families based on an analysis of the 2012 data.

NAM genotype data

A refined linkage map of NAM was recently developed using GBS, which produced a total of 600,000 reliable SNPs, but a large proportion of missing SNP data on each line. An iterative process of imputation and linkage mapping was conducted to produce a final consensus linkage map with complete map scores at 7386 psuedo-markers with a uniform resolution of 0.2 cM per marker (Ogut *et al.* 2015; Swarts *et al.* 2014).

Phenotypic data analysis

Log and square root transformations of shank length were used for the NCRPIS collection and NAM populations, respectively, to minimize the relationship between residual variance and predicted value. Trait data were analyzed using ASReml 3.0 software (Gilmour *et al.* 2009). The mixed model analysis fit line as a fixed effect, and block, year and year×line interactions as random effects. We used heterogeneous error variance structures with unique error variances for each year. The best linear unbiased estimates (BLUEs, sometimes referred to as least squares means) for each line were obtained from this model and treated as the input phenotypic value for further analysis (Files A.7, A.8, A.9, A.10, A.11, and A.12).

For the purpose of partitioning total genotypic variation into additive polygenic and other genotypic variances, a second analysis of the NCRPIS data was conducted using the same model as above, except that line effects are modeled as random effects with a variance-covariance structure for lines proportional to the realized additive genomic relationship matrix (File A.6) and adding an additional term for the identically and independently distributed line effects. The variance component associated with the additive genomic

relationship matrix estimates additive polygenic genetic variance. The variance component associated with the identically and independently distributed line effects captures any other genotypic variance, which could include non-additive variance (although dominance variance should be generally very low among highly homozygous lines) and also non-polygenic variance due to individual genes with large effects (Oakey *et al.* 2006; Oakey *et al.* 2007).

For the purpose of estimating heritability of line mean values, both data sets were also analyzed conducted using the following model:

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{e},$$

where **Y** is the vector of Best linear unbiased estimate (BLUE) values of each phenotype; **u** is a vector or inbred line additive effects, **Z** is a design matrix; and **e** is a vector of random residuals. The variance-covariance matrix of **u** is: Var(**u**) = $\mathbf{G}\sigma_A^2$, where σ_A^2 is the additive genetic variance in the non-inbred reference population and **G** is the realized additive relationship matrix based on all markers. Heritability among the inbred lines was estimated as:

$$\hat{h}^2 = \frac{(1+\bar{F})\sigma_A^2}{(1+\bar{F})\sigma_A^2 + \sigma_{residual}^2},$$

Where $\bar{F}$ is the average inbreeding coefficient of the lines in the population estimated from markers, and $1+\bar{F}$ was estimated as the mean of the diagonal elements of **G**.

Genome wide association study in NCRPIS diversity panel

GWAS was conducted in the NCRPIS diversity panel using TASSEL version 5 (Bradbury *et al.* 2007)) using a mixed linear model to test marker effects:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e},$$

where **Y** is the vector of BLUE values of each line; β is the fixed effect of a single SNP being tested, **u** is a vector of random additive (polygenic background) effects for lines, **X** and **Z** are design matrices, and **e** is a vector of residuals. The variance-covariance matrix of **u** is: Var(**u**) = $\mathbf{G}\sigma_A^2$. We used the optimal compression option in TASSEL. The form of the compressed MLM is the same as equation above, except that individuals in **u** are replaced by their corresponding groups, and realized additive relationship among individuals (**G**) is replaced by the realized additive relationship among groups $\bar{k}$, which is defined as $\bar{k}_{i,j}$:

$$\bar{k}_{i,j} = \text{average}(k_{ht}),\ h \in i\ ,\ t \in j.$$

where *i*,*j* = 1 to *s*, for *s* groups (Zhang *et al.* 2010).

GWAS was repeated on 100 subsamples of the inbred line means, in which each subsample contained a random sample of 80% of the inbred lines with phenotypic data. Resample model inclusion probabilities (RMIP) represent the proportion of data samples in which a particular SNP was declared as significantly associated with the trait at $p < 10^{-7}$. In addition, a single GWAS scan was performed for each trait using the entire data set and false discovery rate (FDR) was estimated for each marker association from this analysis using the qvalue package in R (Bass *et al.* 2015).

NAM joint linkage analysis

Joint linkage analysis of NAM was conducted using Proc GLMSelect in SAS version 9.3 (SAS Institute Inc) to scan the genome at each marker locus. Stepwise selection was used to build the model and *p*-value thresholds for markers to enter and stay in the model were

determined by (Buckler *et al.* 2009; Kump *et al.* 2011). The model contained family main effects and marker effects nested within families:

$$\mathbf{Y} = \mathbf{A\mu} + \sum_{i=1}^{k} \mathbf{X_i \beta_i} + \boldsymbol{\varepsilon},$$

where **Y** is a vector of BLUE values for each inbred line for a given phenotype; **A** is an incidence matrix relating RIL to their corresponding population *p*, **μ** is a vector of population main effects; $\mathbf{X_i}$ is a incidence matrix indicating that RIL's genotype score at locus *i*, the elements of $\mathbf{X_i}$ are estimated dosages of the non-B73 parental allele at SNP *i* (coded as "0" for lines homozygous for the B73 reference allele, and "2" for homozygotes with the alternate parental allele, "1" for heterozygotes, and a non-integer between 0 and 2 for the imputed recombinants as described above); β_i is a vector of the family-specific additive effects associated with locus *i* relative to B73, *k* is the number of significant loci in the final model selected via a stepwise selection and optimization process (Bian *et al.* 2014); and ε is the residual vector.

Segregation for the masculinized ear tip traits was restricted to the B73 × CML69 RIL population. Within this population, the distribution of masculinized ear tip lengths was heavily skewed, with most lines having a value of zero. Therefore, we conducted QTL mapping for masculinized ear tip length within this one population using the 'two-part' model of using R/QTL (Broman *et al.* 2003), similar to the analysis performed for this trait by Holland and Coles (2011).

Genome-wide association

The maize HapMap 2 project provided a total of 28.5 M SNPs (Chia *et al.* 2012). For each chromosome separately, residual values were obtained for each inbred line after fitting

QTL on other chromosomes detected in the joint linkage analysis (Files A.13, A.14, and A.15). These residual values were used as phenotype inputs to GWAS for HapMap SNPs on the test chromosome; these residual values represent the phenotype variation remaining after accounting for unlinked QTL. GWAS was conducted separately for each chromosome by regressing chromosome-specific residuals on each SNP marker using forward selection (Tian *et al.* 2011). NAM association analysis was repeated 300 times across random subsets of 80% of the lines within each family. The *p*-value threshold for declaring a SNP to be significantly associated with the traits was $p = 10^{-7}$.

To calculate the mean r^2 of markers in various testing regions, an additional analysis of the complete data set was conducted in which markers were tested one at a time for association with the appropriate chromosome-specific residuals using a general linear model.

Testing if SNP associations are stronger within hypothesis-defined regions

QTL domestication regions were defined by projecting the end points of QTL support intervals reported by Briggs *et al.* (2007) onto the AGPv2 physical map for the following traits: lateral branch length (BRLG), inflorescence length (INFL), and number of internode columns on primary lateral inflorescence (RANK) from Briggs *et al*. (2007) were used to test hypotheses related to mean r^2 of marker associations with shank length, cob length, and kernel row number, respectively, in our maize evaluations. Domestication and sweep regions were taken from Hufford et al. (2012) and used to compare mean r^2 values for SNPs within and outside of these annotated regions.

For the diversity panel and NAM population, the mean r^2 of all SNPs within or outside of genomic regions annotated as domestication QTL, domestication sweep, or

improvement sweep regions, were estimated using GWAS of the full data set. Differences between r^2 of SNPs inside and outside of hypothesis-defined regions were tested with *t*-tests.

Associations between domestication gene haplotypes and domestication-related traits

We tested for associations between a few well-characterized domestication genes and trait variation using multiple SNP haplotypes in NCRPIS panel. If there were more than 8 SNPs in the candidate gene, then the test region was the gene coding region. Otherwise the test region was extended by 5 kbp on both sides of the coding region to capture sufficient SNP variation to define multiple levels of haplotypes (Table 2.4). Lines with rare haplotypes (less than 5 occurrences in the dataset) were removed from the haplotype association tests. The following model was used to test for haplotype associations:

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{Sh} + \mathbf{e},$$

where **Y** is the vector of BLUE values for each line; **u** is a vector of random additive genetic effects from background markers for lines, **h** is a vector fixed effects of haplotypes at the candidate gene, **Z** and **S** are design matrices, and **e** is a vector of random residuals. The variance-covariance matrix of **u** is: Var(**u**) = $\mathbf{G}\sigma_A^2$. The null hypothesis of no haplotype effects was tested with an F-test for the haplotype factor. For haplotypes with significant effects in the previous model, an additional analysis was conducted using the same model, except that haplotype effect is modeled as random, and the variance component associated with the haplotype is calculated. The proportion of variation associated with haplotype differences was estimated as: $\frac{\sigma_{hap}^2}{\sigma_{hap}^2 + (1+\bar{F})\sigma_A^2 + \sigma_{residual}^2}$.

Partitioning variance associated with different genomic regions

To test if trait variation associated with regions annotated as domestication QTL, domestication sweep, or improvement sweep regions is greater than variation associated with random background polygenic variation, we used a procedure to estimate variance components associated with different genome regions (Gusev *et al.* 2014; Speed *et al.* 2012; Speed and Balding 2014). For each hypothesis and panel of inbred lines, we estimated three additive realized relationship matrices, each based on all SNPs within a hypothesis region (domestication QTL, domestication sweep regions, or improvement sweep regions), and a fourth realized additive relationship matrix using disjoint background markers. A mixed model was fit to estimate simultaneously the variances associated with each relationship matrix:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{Z_Q H_Q} + \boldsymbol{Z_D H_D} + \boldsymbol{Z_I H_I} + \boldsymbol{Z_B B} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{H_Q}$, $\boldsymbol{H_D}$, and $\boldsymbol{H_I}$ are the random effects of genome regions within domestication QTL, domestication sweep regions, and improvement sweep regions, respectively. Each of these hypothesis effect vectors is distributed with a variance-covariance matrix proportional to the realized additive relationship matrix estimated using SNPs within the corresponding genomic region: $\boldsymbol{H}_Q \sim N(0, \boldsymbol{G}_Q \sigma^2_{A(Q)})$, $\boldsymbol{H}_D \sim N(0, \mathbf{G}_D \sigma^2_{A(D)})$, and $\boldsymbol{H}_I \sim N(0, \mathbf{G_I} \sigma^2_{A(I)})$, where $\sigma^2_{A(Q)}$, $\sigma^2_{A(D)}$, and $\sigma^2_{A(I)}$ are the additive genetic variances associated with domestication QTL, domestication sweep regions, and improvement sweep regions, respectively. **B** are the polygenic background effects for each line, $\boldsymbol{B} \sim N(0, \boldsymbol{G}_B \sigma^2_{A(B)})$.

Variability in the scaling of the relationship matrices (which occurs simply due to sampling different markers) affects the magnitude of the associated variance components. The product of the mean diagonal element of the relationship matrix and its associated

variance component is constant, however. Therefore, to make fair comparisons among variance components associated with different relationship matrices, we estimated the additive variance accounted for by a particular hypothesis matrix by multiplying the variance component estimate by the mean of the diagonal elements of the relationship matrix. The total additive variance among inbred lines was estimated as:

$$\sigma^2_{A(T)} = \sum_{i=1}^{h}\left(mean\ of\ diag(G_{Hi})\sigma^2_{A(Hi)}\right) + \left(mean\ of\ diag(G_B)\right)\sigma^2_{A(B)}.$$

The proportion of total additive variance attributable to a particular hypothesis-defined relationship matrix *i* was estimated as:

$$\frac{\left(mean\ of\ diag(G_{Hi})\right)\sigma^2_{A(Hi)}}{\sigma^2_{A(T)}}.$$

We compared the proportion of additive variance for each hypothesis region to the proportion of the genome represented by the markers in the region. The heritability associated with a particular relationship matrix *i* is:

$$\hat{h}^2_i = \frac{\left(mean\ of\ diag(G_{Hi})\right)\sigma^2_{A(Hi)}}{\sigma^2_{A(T)} + \sigma^2_{residual}}.$$

For each hypothesis, we also separately estimated 'matched' background matrices based on a random sample of background markers with same proportion of coding region SNPs and the same total number of markers as the hypothesis-defined realized additive relationship matrix.

We resampled the matching background SNPs and re-estimated the matching background realized additive relationship matrix 20 times for each hypothesis. For each

pairing of a hypothesis realized additive relationship matrix and one of 20 distinct background realized additive relationship matrices, we fit a mixed linear model to estimate the variance components associated with the hypothesis and background matrix:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{Z_H H} + \boldsymbol{Z_B B} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y}$ is the vector of line BLUEs, μ is the intercept vector, $\boldsymbol{H}$ and $\boldsymbol{B}$ are vectors of random effects associated with the hypothesis and background genomic regions for each line, $\boldsymbol{Z_H}$ and $\boldsymbol{Z_B}$ are incidence matrices (in this case, identity matrices of dimension equal to the number of lines), and $\boldsymbol{\varepsilon}$ is the vector of residual effects. Random effects $\boldsymbol{H}$ and $\boldsymbol{B}$ are distributed with variance-covariance matrices proportional to their respective realized additive relationship matrices: $\boldsymbol{H} \sim N(0, \mathbf{G}_H \sigma^2_{A(H)})$ and $\boldsymbol{B} \sim N(0, \boldsymbol{G}_B \sigma^2_{A(B)})$, where $\mathbf{G}_\mathrm{H}$ and $\mathbf{G}_\mathrm{B}$ are the realized additive relationship matrices based on SNPs in the hypothesis regions and in the genomic background, respectively, and $\sigma^2_{A(H)}$and $\sigma^2_{A(B)}$ are additive genetic variance components associated with the hypothesis regions and genomic background, respectively.

All realized relationship matrices were estimated using TASSELversion 5 (Bradbury *et al.* 2007) based on HapMap 3.1 SNP data (Bukowski *et al.* 2015). Variance components were estimated using LDAK version 4.9 (Speed *et al.* 2012; Speed and Balding 2014).

Results

Trait distributions and heritability

Shank length, cob length, and kernel row number were approximately normally distributed within both diversity and NAM inbred line panels (Figure 2.1). All traits had greater variability in the diversity panel than in the NAM panel (Figure 2.1). Heritabilities of line means for shank length, cob length and kernel row number ranged from 0.40 to 0.53 in the diversity panel and from 0.38 to 0.70 in NAM (Table 2.1). Masculinized ear tip length displayed much less variation than the other traits, with most lines exhibiting tip lengths of zero (Figure 2.1). Segregation for masculinized ear tip was limited to a single NAM family, B73 × CML69, and in this family 27 lines among 187 lines had non-zero tip lengths. The diversity panel had a higher proportion of lines with non-zero tip lengths and longer maximum tip length than NAM (Figure 2.1).

QTL and association mapping

In the NAM population, we identified 10 QTL for shank length (each associated with 1.8 to 2.8% of the trait variation), 8 QTL for kernel row number (associated with 1.8 to 7.3% variation), and 20 QTL for cob length (associated with 0.8 to 2.9% variation; File A.1). No QTL were detected for masculinized ear tips; power of QTL detection for this trait was limited because its segregation was restricted to one biparental family. QTL analysis within this one family did not identify any QTL passing a genome-wide permutation-based threshold of $\alpha = 0.05$. Comparisons between the positions of domestication-related trait QTL mapped in NAM and previously identified domestication QTL mapped in crosses between

maize and teosinte revealed little correspondence between the two sets of QTL (Figures A.1, A.2, and A.3). Genome wide association scans conducted in the NCRPIS diversity panel identified 0, 5, and 10 SNPs associated with cob length, shank length and kernel row number, respectively at FDR < 0.05 (File A.2). In general, there was limited overlap between known domestication QTL and SNPs associated with domestication-related traits in either panel (Figures A.1, A.2, and A.3), however a few notable correspondences were observed. A SNP 270,000 bp upstream of *fea2* was strongly associated with kernel row number trait, however the one SNP inside of the *fea2* coding region was not significant. Several associations identified for SL in NAM are in the vicinity of *tb1*, but ~2 million bp downstream of the gene (File A.3). By contrast, the known upstream enhancer of *tb1* is ~ 59-69 kbp from the coding start site (Clark *et al.* 2006)(Studer *et al.* 2011).

Testing concordance between loci associated with domestication traits within maize and loci that distinguish maize from teosinte

For each set of trait QTL and SNP associations, we compared the mean r^2 of associations inside versus outside genomic regions previously identified as related to domestication. Domestication QTLs are mapped in a maize-by-teosinte cross population by Briggs *et al*. 2007 (Table A.1) and domestication selection sweep regions are identified from population genetics analyses (Hufford *et al.* 2012). In addition, we compared mean r^2 of associations for SNPs inside or outside regions defined as post-domestication "improvement" selection sweeps from population genetics analyses (Hufford *et al.* 2012). To remove the potentially confounding effect of variability in gene density among regions tested, we tested

within regions defined using both annotation for involvement in domestication or improvement and annotation for coding or non-coding regions.

For the NCRPIS diversity panel, mean marker r^2 values were around 0.0009, and the largest difference between groups was only 0.000032 (Table 2.2). This maximum difference was observed between coding variants inside and outside of domestication QTL for KRN, and the SNPs outside of the domestication QTL were associated with more variation (Table 2.2). In fact, the only significant differences in mean marker r^2 for SNPs classified according to domestication QTL were observed when SNPs outside of QTL were associated with greater mean r^2 values than SNPs within the QTL (Table 2.2). Further, there was no consistent evidence that SNPs inside domestication or improvement sweeps were associated with more variation than SNPs outside of these regions, although non-coding SNPs within sweep regions had significantly higher mean r^2 values for shank length than non-coding SNPs outside those regions (Table 2.2).

For the NAM population, the mean SNP r^2 values were significantly different for each comparison of hypothesis region and grouping based on coding regions (Table 2.3). The largest differences between categories were observed between SNPs inside and outside of domestication QTL for KRN. Domestication QTL SNPs were associated with more variation only for KRN, whereas domestication QTL SNPs had smaller mean r^2 values for SL and CL (Table 2.3). SNP variances were larger inside than outside of hypothesis regions most consistently for domestication sweep regions, but even within this group, SNPs in non-coding domestication sweep regions had lower mean r^2 values associated with CL than SNPs in non-coding regions outside of domestication sweep regions (Table 2.3).

Association of haplotypes at known domestication genes

A number of domestication QTLs have been resolved to individual genes by a combination of high resolution genetic mapping, mutant analysis, and gene expression studies (Table 2.4). We identified SNPs within and nearby these genes and defined haplotypes at each domestication gene based on multiple SNP genotypes. Haplotype tests in NCRPIS panel shows that haplotypes containing *grassy tillers 1 (gt1)* were significantly associated with shank length (1.6% of variation, $p < 0.05$; Table 2.4). Haplotype additive effects on shank length ranged from +32 mm to -26 mm for *gt1* (File A.4), and the inbred lines with haplotype effects that cause the largest increase in shank length represent a set of tropical and exotic germplasm distinct from the major temperate maize breeding pool (CML254, CML270, CML388, CML389, CML419, GE440, NC264, SC276Q2, SC277, SC76, TZEEI17, TZEEI20, and TZEI5). *Zea apetala homolog1 (zap1)* showed a significant association with cob length (5.9% of variation, $p < 0.01$; Table 2.4 and File A.5). No other candidate gene haplotypes had significant effects on trait variation.

Variance component testing

To estimate the proportion of trait genotypic variance associated with additive polygenic versus other genetic effects (such as large effect loci and non-additive variance) in the NCRPIS panel, we simultaneously modeled genotypic effects with variance-covariance relationships proportional to the realized additive relationship matrix and genotypic effects with no pairwise relationships to capture genetic effects unique to each line. Among traits, 92% - 100% of genotypic variance was accounted for by polygenic additive background effects, with the remainder of variance attributable to a combination of non-additive effects and large

effect loci (Table A.2).

To partition total trait variance into components associated with domestication QTL, domestication sweep regions, improvement sweep regions, and the remainder of the genome, we estimated realized additive relationship matrices using SNPs in each of these regions of the genome and estimated the associated variance components in each panel (Figures 2.2 and 2.3; Tables A.3, A.4, and A.5). When effects associated with all four relationship matrices were fit simultaneously in a common mixed model, the background polygenic variance component accounted for 67 – 80% of the total additive genetic variance in NCRPIS (Figure 2.2A; Tables A.3 and A.5) and 71 – 100% in NAM (Figure 2.3A; Tables A.4 and A.5). The increase in total heritability explained by fitting all four categories together was only zero to 1% compared to simply fitting a single relationship matrix based on all SNPs together across all traits and panels (Figures 2.2A and 2.3A; Tables A.3 and A.4).

The relationship matrices were estimated using widely different numbers of markers, which is expected to affect the proportion of variance associated with each matrix under the null hypothesis of equal contributions to the total genetic variance. Therefore, we compared the proportion of additive variance accounted for by each matrix to the proportion of the genome represented by the hypothesis region. The proportion of additive variance associated with QTL-defined and domestication sweep-related hypothesis matrices was smaller than the proportion of genome represented by the SNPs defining those matrices (except for cob length in the NAM population; Figures 2.2 and 2.3; Table A.5). In contrast, the proportion of total additive variance associated with the improvement sweep-defined relationship matrix was two to five times greater than the proportion of genome represented by the improvement sweeps (except for kernel row number variance, which was completely associated with the

genomic background, Figures 2.2 and 2.3; Table A.5).

An alternative approach to account for differences in the proportion of genome represented in each matrix was to fit each hypothesis-based relationship matrix along with a matched background relationship matrix based on a equally sized sample of background SNPs with the same proportion of coding and non-coding variants to estimate variance components. For each combination of hypothesis region, trait, and inbred line panel, we sampled background SNPs and fit the mixed model 20 times to estimate the variability in variance components estimates across samples. Background polygenic effects were consistently associated with more variance than the domestication QTL, domestication sweep, or improvement sweep regions when fitting relationship matrices with matching numbers and genic composition of SNPs. (Figures A.4 and A.5; Tables A.3 and A.4). Among the hypothesis-defined regions, the improvement sweep regions consistently explained the largest proportion of variation, ranging from 8% to 48% of the total heritable variance when fit with a matched background polygenic effect relationship matrix.

Discussion

Heritability and polygenic variation

Heritabilities of the three traits were relatively low to moderate, in part because the large numbers of lines tested precluded evaluating larger numbers of replicates of the experiment. The polygenic relationship matrix was associated with 40%-53% of total phenotypic variation in the NCRPIS panel (Table 2.1). By comparison, the largest amount of variation associated with an individual SNP was estimated to be about 3% (File A.2) and few SNPs passed stringent thresholds for association tests.

Haplotypes at the candidate gene *zap1* were associated with 6% of cob length variation, suggesting that complex variation in a genomic region occasionally may account more variation than can be associated with a single SNP, but this was the exception to the general trend of no obvious haplotype effects. Variants in *zap1* were associated with ear length in teosinte (Weber *et al.* 2008); our results suggest some functional variation at this locus passed through the domestication bottleneck and remains in maize, or new functional variants have arisen within maize. Haplotypes at candidate gene *gt1* were also associated with a small amount of shank length variation in maize. Although this locus was not detected as affecting lateral branch (shank) length in maize-teosinte crosses (Briggs et al, 2007), Wills et al. (Wills *et al.* 2013) identified *gt1* as conferring the major difference in the number of ears produce by maize and teosinte, and observed that haplotypic variation at this locus suggests only a partial sweep due to selection under domestication. Some of the teosinte-type variation at this locus may even have a favorable effect in maize by increasing the number of ears by a small amount, and it is possible that these same variants have small effects on shank

length.

The *tb1* gene and its linked enhancer played a key role in changing the morphology of maize, including reducing the length of lateral branches, during the domestication process (Studer *et al.* 2011; Tsiantis 2011). Thus, *tb1* is an obvious candidate for explaining the variation among shank (lateral branch) lengths in maize. However, we observed no QTL or SNP association in NAM around *tb1*. We also did not identify an association for shank length near the gene in the diversity panel GWAS, and SNPs inside of *tb1* coding region and its enhancer were not significant. Direct testing of haplotypes defined by SNPs surrounding *tb1* (encompassing a 5268 bp region) and encompassing the *tb1* enhancer region suggested that these haplotypes are not significantly associated with shank length for NCRPIS panel.

Although we identified a few individual SNPs and haplotypes associated with significant amounts of variation for domestication traits in maize, their effects were small. The effects and variances associated with SNPs inside domestication QTL regions were no larger than those of random samples of SNPs throughout the rest of the genome. These results suggest that most of genotypic variation for domestication-related traits in maize is explained by a large number of loci with small effects. Therefore, the genetic architecture of variation for domestication traits within maize appears mostly distinct from the genetic control of differences between maize and teosinte, which are dominated by a relatively few large effect loci. We found no evidence that QTL or SNP associations for these traits were more likely to be near domestication QTL or that markers in domestication QTL explained more trait variation than markers outside of these regions (Figures 2.2 and 2.3, Tables 2.2 and 2.3). No consistent pattern of increased SNP effects was observed for SNPs inside domestication or improvement sweep regions (Tables 2.2 and 2.3). The comparison of

average SNP effects averaged over all SNPs in a group has limitations; many of these effect estimates are expected to be poor, and the mean value estimated is expected to be an upwardly biased estimate of the true mean effect size of individual SNPs. However, by averaging over many thousands of loci within each class, we expect the biases to cancel out when comparing mean effect sizes of different classes.

Partitioning of the genetic variance into components due to specific hypothesis-based regions is likely a more reliable method for comparing the influence of different genomic regions that are highly polygenic. Using this approach, we observed that improvement sweep regions showed a consistently higher proportion of the total heritable variance than other hypothesis-defined regions, and often substantially more than the proportion of genome represented by SNPs defining the improvement sweep relationship matrix (Figure 2.2 and 2.3; Tables A.3, A.4, and A.5). When we fit specific hypothesis-based relationship matrices along with background matrices sampled with matching SNP numbers and proportion of coding SNPs the SNP number and proportion of coding SNP, the variance associated with hypothesis-based relationship matrices was always lower than the matching background (Figures A.4 and A.5). However, although we took care to control for the sample size and gene density of the SNPs used to compute the hypothesis and background relationship matrices, we expect that the markers used for the hypothesis matrix are have higher linkage disequilibrium and relatively less explanatory power than equally sized samples of SNPs from the rest of the genome because they were sampled from restricted genomic blocks. The higher levels of linkage disequilibrium expected among the improvement sweep SNPs would downwardly bias the proportion of total additive variance they can explain relative to an equally sized random sample of SNPs from the rest of the genome. Therefore, these results

are congruent with enrichment of improvement sweep-related regions of the maize genome for functional variants affecting domestication-related traits, although the effects of individual variants appear to be quite small and the precise magnitude of the enrichment remains difficult to assess.

The generally reduced contribution of domestication QTL regions, and to a lesser extent the domestication sweep regions, to domestication-related traits variation in maize is likely a direct result of selection purging variants that favor the teosinte morphology in these regions. The increased contribution of improvement sweep regions to variation in these traits may be due to divergent selection for functional alleles in these regions. Although modern inbreds are significantly differentiated from landraces in these regions, the level of differentiation is lower than the mean differentiation between landraces and teosinte in the domestication sweep regions (Hufford et al., 2013). Thus, more sequence variation exists among inbreds in improvement sweep regions than in domestication sweep regions. However, less variation among inbreds exists in both domestication and improvement sweep regions than in the rest of the genome. This suggests that functional variants for domestication traits in improvement sweep regions may be targets of selection, but divergent selection maintains some variation for such variants. For example, some maize varieties have small kernel row numbers (because this is associated with larger seed size); others with small cob lengths are maintained because they have favored kernel types. Historical selection may have favored more kernel rows and longer cobs in general, but diverse inbred lines sampled from different regions may include contributions from populations selected in the opposite direction, resulting in an overall signal of selection near variants that affect these traits at the same time as these variants contribute disproportionately to the observed trait variation.

Acknowledgments

S.X. was supported by National Institutes of Environmental Health Sciences training grant T32 ES007329 to the NCSU Bioinformatics Research Center and National Science Foundation award IOS-1127076, J.H. was supported by National Science Foundation awards IOS-1127076 and IOS-1238014 and by United States Department of Agriculture, Agricultural Research Service. We thank Jeff Glaubitz for help selecting SNPs from the HapMap 3 database for relationship matrix estimation.

Figures and Tables

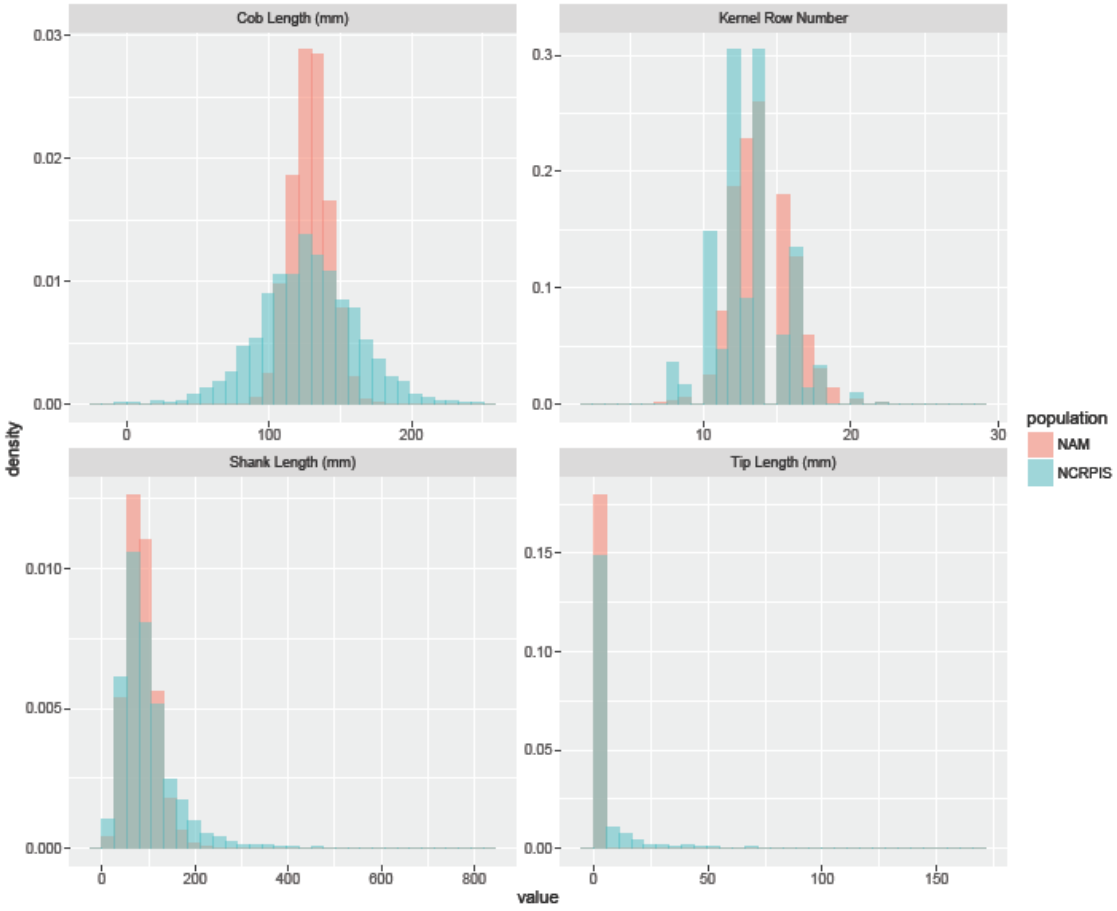

Figure 2.1. Distribution of shank length, cob length, kernel row number and masculinized ear tip length in NCRPIS panel (red) and NAM population (green).

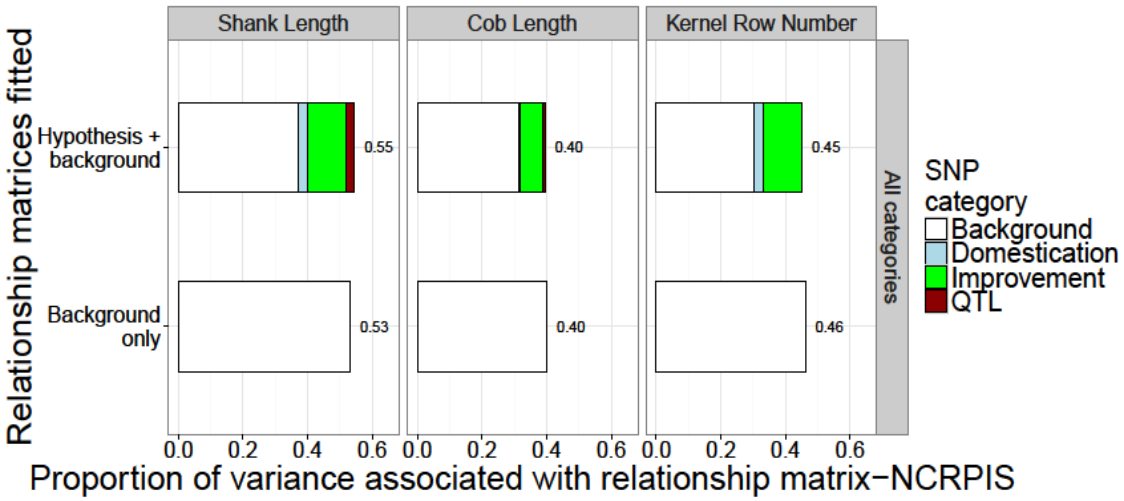

Figure 2.2.A. The proportion of variance for shank length, cob length, and kernel row number among inbred lines of the NCRPIS panel associated with relationship matrices based on all SNPs in hypothesis-defined regions or on background SNPs.

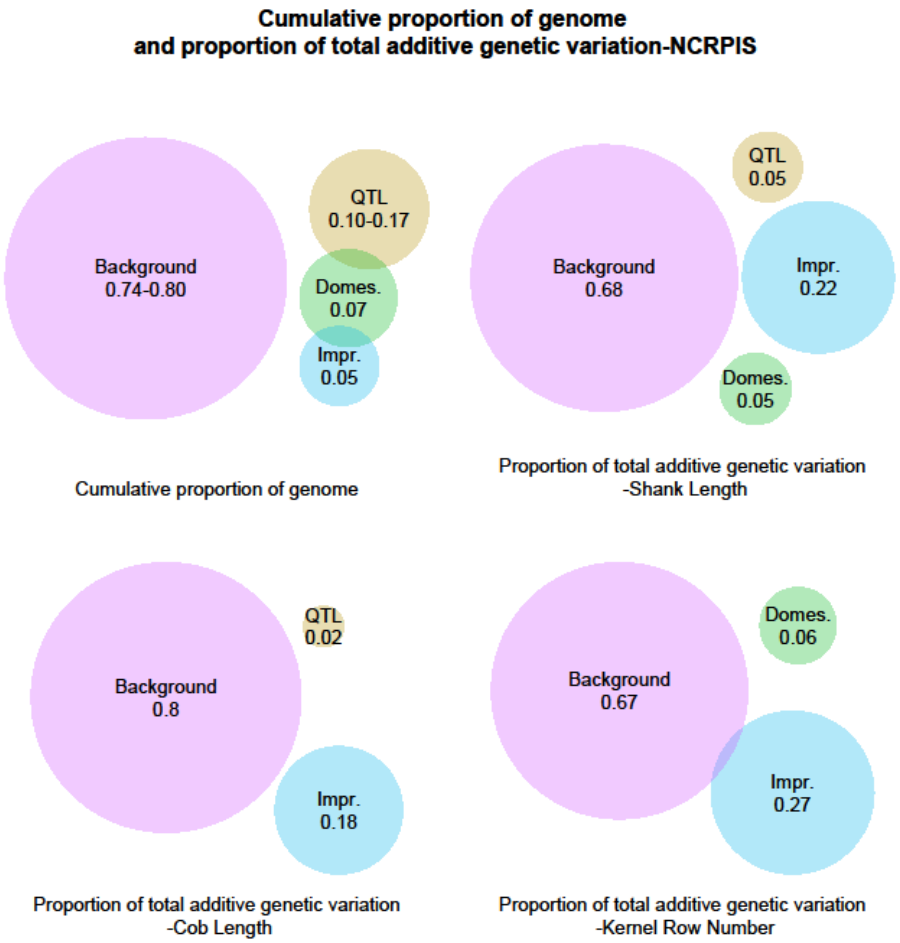

Figure 2.2.B. Cumulative proportion of genome tagged by SNPs defining hypothesis relationship matrices and background matrices, and the proportion of total additive genetic variation associated with each relationship matrix for shank length, cob length, and kernel row number among inbred lines of the NCRPIS panel.

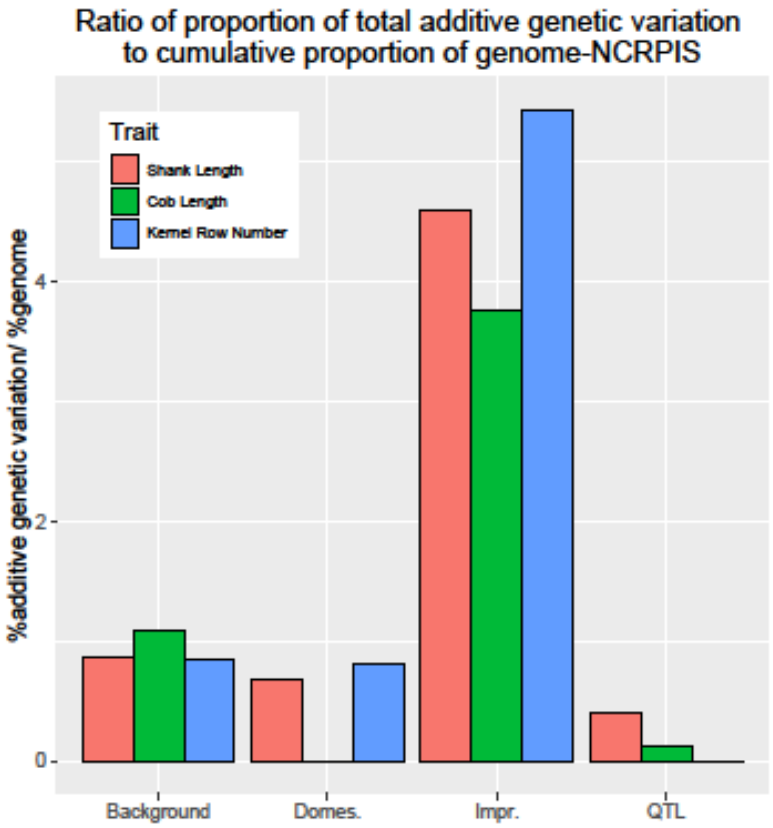

Figure 2.2.C. Ratio of proportion of total additive genetic variation to cumulative proportion of genome tagged by SNPs defining hypothesis and background relationship matrices for shank length, cob length, and kernel row number among inbred lines of the NCRPIS panel.

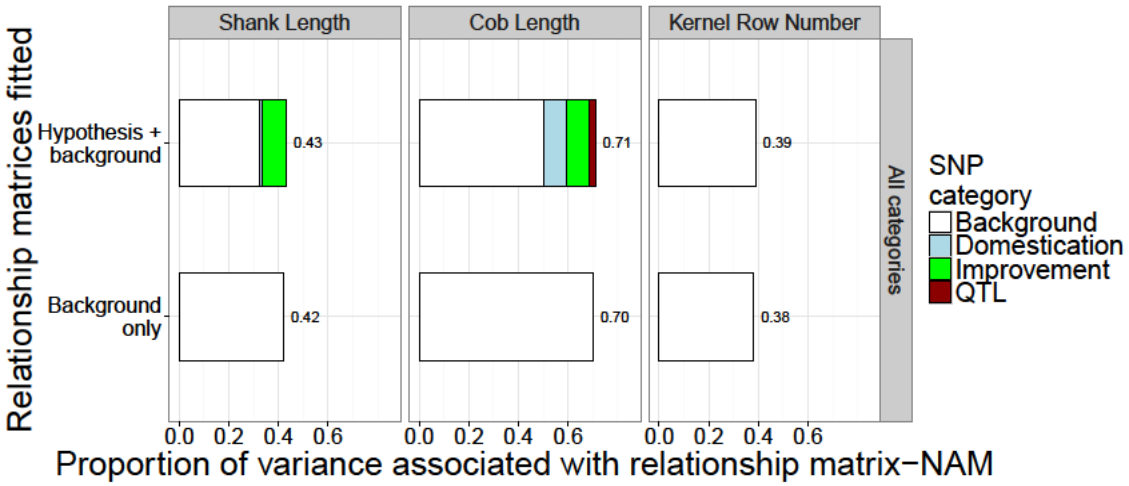

Figure 2.3.A. The proportion of variance for shank length, cob length, and kernel row number among inbred lines of the NAM panel associated with relationship matrices based on all SNPs in hypothesis-defined regions or based on background SNPs.

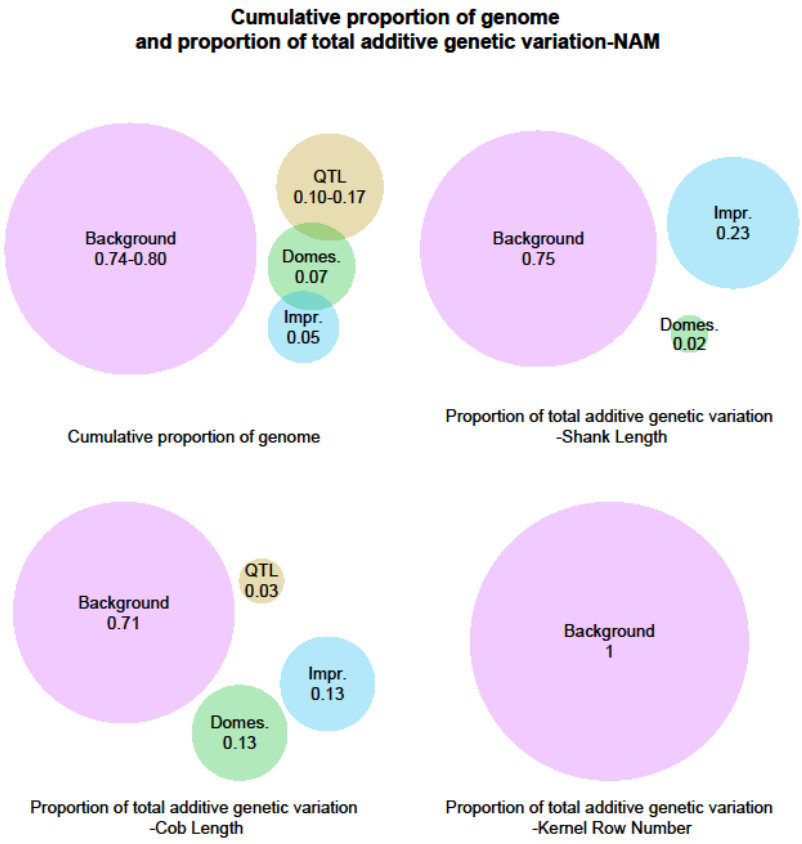

Figure 2.3.B. Cumulative proportion of genome tagged by SNPs defining hypothesis relationship matrices and background matrices, and the proportion of total additive genetic variation associated with each relationship matrix for shank length, cob length, and kernel row number among inbred lines of the NAM panel.

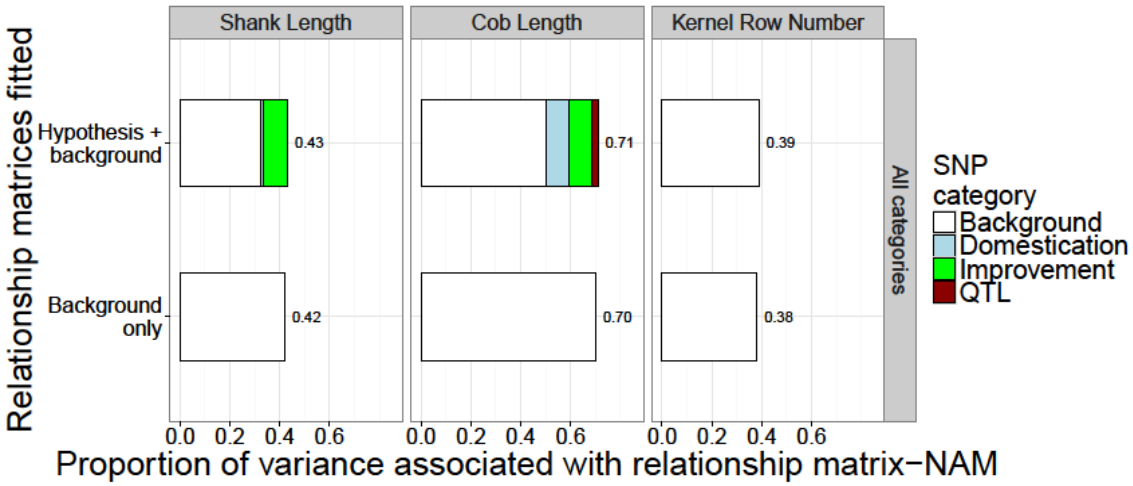

Figure 2.3.C. Ratio of proportion of total additive genetic variation to cumulative proportion of genome tagged by SNPs defining hypothesis and background relationship matrices for shank length, cob length, and kernel row number among inbred lines of the NAM panel.

Table 2.1. Summary statistics and heritability estimates ($\widehat{h^2}$) for three domestication-related traits: shank length (SL), cob length (CL), kernel row number (KRN) in the maize NCRPIS diversity and NAM panels.

	NCRPIS panel			NAM		
	SL[d] (mm)	CL (mm)	KRN	SL[e] (mm)	CL (mm)	KRN
N[a]	5002	3381	4776	6903	32031	6266
Ng[b]	2387	2287	2339	2875	4359	2724
Mean	95	141	14	87	129	14
Min	10	10	6	10	79	4
Max	800	270	28	354	180	24
$\widehat{h^2}$	0.53	0.40	0.46	0.42	0.70	0.38
SE($\widehat{h^2}$)[c]	0.049	0.045	0.049	0.031	0.021	0.031

[a] Total number of plots measured.
[b] Total number of inbred lines measured.
[c] Approximate standard error of heritability estimate.
[d] $\widehat{h^2}$, SE($\widehat{h^2}$) of SL estimated using log-transformed data.
[e] $\widehat{h^2}$, SE($\widehat{h^2}$) of SL estimated using square root-transformed data.

Table 2.2. Mean SNP association r^2 and number of markers (N_m) inside and outside hypothesis-defined testing regions in NCRPIS panel.

Hypothesis or background	Coding or non-coding	Shank length		Cob length		Kernel row number	
		r^2	N_m	r^2	N_m	r^2	N_m
		$\times 10^{-4}$	$N \times 10^3$	$\times 10^{-4}$	$N \times 10^3$	$\times 10^{-4}$	$N \times 10^3$
			Domestication QTL				
Hypothesis	Coding	9.27	21.5	9.18	27.2	9.20	17.8
Background	Coding	9.19	226.4	9.40	220.7	9.52	230.2
Difference	Coding	0.08		-0.22**		-0.32**	
Hypothesis	Non	9.09	15.4	9.16	19.6	9.43	13.6
Background	Non	9.29	141.9	9.38	137.7	9.42	143.7
Difference	Non	-0.20**		-0.22		0.01	
			Domestication sweep				
Hypothesis	Coding	9.30	15.0	9.39	15.0	9.45	15.0
Background	Coding	9.20	232.9	9.37	232.9	9.50	232.9
Difference	Coding	0.10		0.02		-0.05	
Hypothesis	Non	9.52	10.4	9.31	10.4	9.52	10.4
Background	Non	9.25	146.9	9.35	146.9	9.41	146.9
Difference	Non	0.27**		-0.04**		0.11	
			Improvement sweep				
Hypothesis	Coding	9.22	11.9	9.50	11.9	9.60	11.9
Background	Coding	9.20	236.0	9.37	236.0	9.49	236.0
Difference	Coding	0.02		0.13		0.11	
Hypothesis	Non	9.56	9.0	9.53	9.0	9.20	9.0
Background	Non	9.25	148.4	9.34	148.4	9.43	148.4
Difference	Non	0.31**		0.19		-0.23*	

*, ** Significantly different at $p = 0.05$ and $p = 0.01$, respectively.

Table 2.3. Mean SNP association r^2 and number of markers (N_m) inside and outside hypothesis-defined testing regions in NAM panel.

Hypothesis or background	Coding or non-coding	Shank length		Cob length		Kernel row number	
		r^2	N_m	r^2	N_m	r^2	N_m
		$\times 10^{-4}$	$N \times 10^5$	$\times 10^{-4}$	$N \times 10^5$	$\times 10^{-4}$	$N \times 10^5$
				Domestication QTL			
Hypothesis	Coding	9.97	2.1	17.2	2.9	29.7	2.1
Background	Coding	14.3	20.6	20.8	20.3	20.0	20.6
Difference	Coding	-4.3**		-3.6**		9.7**	
Hypothesis	Non	10.1	32.9	18.5	42.5	25.6	28.4
Background	Non	14.9	201.6	25.8	196.9	22.5	206.0
Difference	Non	-4.8**		-7.3*		3.1**	
				Domestication sweep			
Hypothesis	Coding	15.3	1.2	20.7	1.2	29.6	1.2
Background	Coding	13.8	21.5	20.4	21.5	20.0	21.5
Difference	Coding	1.5**		0.3**		9.6**	
Hypothesis	Non	16.1	15.5	21.1	15.5	31.4	15.5
Background	Non	14.2	219.1	25.0	219.1	22.2	219.1
Difference	Non	1.9**		-3.9**		9.2**	
				Improvement sweep			
Hypothesis	Coding	12.8	1.4	25.2	1.4	17.8	1.4
Background	Coding	14.0	21.6	20.2	21.6	20.8	21.6
Difference	Coding	-1.2**		5**		-3**	
Hypothesis	Non	13.9	10.7	31.3	10.7	21.2	10.7
Background	Non	14.3	223.7	24.4	223.7	23.0	223.7
Difference	Non	-0.4**		6.9**		-1.8**	

*, ** Significantly different at $p = 0.05$ and $p = 0.01$, respectively.

Table 2.4. Tests of associations between haplotypes of known domestication genes and domestication-related traits in NCRPIS panel.

Gene name	Chr	Start[a]	End	Extended testing region? [b]	Extended-Start	Extended-End	Num. of SNPs in gene	Num of SNPs in testing region	Num of haplotypes tested for association	Proportion of variance explained (%)		
										SL	CL	KRN
tb1	1	265745979	265747712	Yes	265746572	265751840	5	12	15	NS	NS	-
tb1-enhancer	1	265676479	265687279	No	-	-	9	9	6	NS	NS	-
gt1	1	23241091	23244476	Yes	23236091	23249476	3	13	48	1.6 *[c]	NS	-
zagl1	1	4862047	4877625	Yes	4862244	4862765	5	5	6	-	NS	-
zap1	2	235845160	235853770	No	-	-	21	21	45	-	5.9 **	-
te1	3	165174146	165178071	No	-	-	8	8	17	-	NS	-
fea2	4	133662510	133664998	Yes	133662368	133664252	2	2	6	-	-	NS

[a] Coding sequence start position (AGPv2)

[b] If the region is extended, testing region is 5kbp extended on the left and right side of original position. Sometimes SNPs don't fully spread in whole testing region, so the extended region is actual region for testing.

[c] NS, not significant; *, **, significant at $p < 0.05$ and $p < 0.01$, respectively.

[d] Proportion of variance explained is calculated as $\frac{\sigma^2_{hap}}{\sigma^2_{hap} + (1+F)\sigma^2_A + \sigma^2_{residual}}$

References

Bass, A. J., A. Dabney and D. Robinson.2015 Qvalue: Q-Value Estimation for False Discovery Rate Control. R Package version 2.2.2, Http://github.com/jdstorey/qvalue

Bian, Y., Q. Yang, P. Balint-Kurti, R. J. Wisser and J. B. Holland, 2014 Limits on the reproducibility of marker associations with southern leaf blight resistance in the maize nested association mapping population. BMC Genomics **15:** 1-15.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al*, 2007 TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633-2635.

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

Broman, K. W., H. Wu, S. Sen and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19:** 889-890.

Browning, B. L., and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. **84:** 210-223.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown *et al*, 2009 The genetic architecture of maize flowering time. Science **325:** 714-718.

Bukowski, R., X. Guo, Y. Lu, C. Zou, B. He *et al*, 2015 Construction of the third generation zea mays haplotype map. bioRxiv .

Chia, J., C. Song, P. J. Bradbury, D. Costich, N. de Leon *et al*, 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. **44:** 803-807.

Clark, R. M., T. N. Wagler, P. Quijada and J. Doebley, 2006 A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat. Genet. **38:** 594-597.

Clark, R. M., E. Linton, J. Messing and J. F. Doebley, 2003 Pattern of diversity in the genomic region near the maize domestication gene tb1. Proc. Natl. Acad. Sci. U. S. A. **101:** 700-707.

Crouch, J., M. Warburton and Y. JianBing.2011 Association Mapping for Enhancing Maize (*Zea Mays L.*) Genetic Improvement.

Darwin, C. R., 1868 The variation of animals and plants under domestication. London: John Murray. **2:** .

Doebley, J., A. Stec, J. Wendel and M. Edwards, 1990 Genetic and morphological analysis of a maize-teosinte F2 population: Implications for the origin of maize. Proc. Natl. Acad. Sci. U. S. A. **87:** 9888-9892.

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

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

Doebley, J., A. Stec and B. Kent, 1995 Suppressor of sessile spikelets 1 (Sos1): A dominant mutant affecting inflorescence development in maize. Am. J. Bot. **82:** 571-577.

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

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. Plos One **6:** e19379.

Flint-Garcia, S., A. Thuillet, J. Yu, G. Pressoir, S. Romero *et al*, 2005 Maize association population: A high-resolution platform for quantitative trait locus dissection. Plant Journal **44:** 1054-1064.

Galinat, W. C., 1983 The origin of maize as shown by key morphological traits of its ancestor, teosinte. Maydica .

Gilmour, A. R., B. J. Gogel, B. R. Cullis and R. Thompson, 2009 *ASReml User Guide Release 3.0*. VSN International, Ltd, Hemel Hempstead, UK.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al*, 2014 TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE **9:** 1-11.

Gusev, A., S. . Lee, G. Trynka, H. Finucane, B. Vilhjálmsson *et al*, 2014 Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. The American Journal of Human Genetics **95:** 535-552.

Harlan, J. R., 1992 Crops & Man. American Society of Agronomy, Madison, WI, USA.

Holland, J. B., and N. D. Coles, 2011 QTL controlling masculinization of ear tips in a maize (zea mays L.) intraspecific cross. G3: Genes|Genomes|Genetics **1:** 337-341.

Hufford, M. B., X. Xu, J. van Heerwaarden, T. Pyhajarvi, J. Chia *et al*, 2012 Comparative population genomics of maize domestication and improvement. Nat. Genet. **44:** 808-811.

Hung, H., L. M. Shannon, F. Tian, P. J. Bradbury, C. Chen *et al*, 2012 ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proceedings of the National Academy of Sciences **109:** E1913-E1921.

Hung, H., C. Browne, K. Guill, N. Coles, M. Eller *et al*, 2011 The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. Heredity **108:** 490-499.

Iltis, H. H., 1983 From teosinte to maize: The catastrophic sexual transmutation. Science **222:** 886-894.

Kump, K. L., P. J. Bradbury, E. S. Buckler, A. R. Belcher, M. Oropeza-Rosas *et al*, 2011 Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat. Genet. **43:** .

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, J. Sanchez G., E. Buckler *et al*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. U. S. A. **99:** 6080-6084.

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li *et al*, 2009 Genetic properties of the maize nested association mapping population. Science **325:** 737-740.

Meyer, R. S., and M. D. Purugganan, 2013 Evolution of crop species: Genetics of domestication and diversification. Nat. Rev. Genet. **14:** 840-852.

Oakey, H., A. P. Verbyla, B. R. Cullis, X. Wei and W. S. Pitchford, 2007 Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. Theor. Appl. Genet. **114:** 1319-1332.

Oakey, H., A. Verbyla, W. Pitchford, B. Cullis and H. Kuchel, 2006 Joint modeling of additive and non-additive genetic line effects in single field trials. Theor. Appl. Genet. **113:** 809-819.

Ogut, F., Y. Bian, P. J. Bradbury and J. B. Holland, 2015 Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. Heredity **114:** 552-563.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira *et al*, 2007 PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. **81:** 559-575.

Romay, M. C., M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts *et al*, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. **14:** .

SAS Institute Inc, SAS/STAT® 9.3 User's guide.

Speed, D., and D. J. Balding, 2014 MultiBLUP: Improved SNP-based prediction for complex traits. Genome Res. **24:** 1550-1557.

Speed, D., G. Hemani, M. Johnson and D. Balding, 2012 Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. **91:** 1011-1021.

Studer, A., Q. Zhao, J. Ross-Ibarra and J. Doebley, 2011 Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. **43:** 1160-1163.

Studer, A. J., and J. F. Doebley, 2012 Evidence for a natural allelic series at the maize domestication locus teosinte branched1. Genetics **191:** 951-U533.

Swarts, K., H. Li, J. A. R. Navarro, D. An, M. C. Romay *et al*, 2014 Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome **7:** .

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun *et al*, 2011 Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. **43:** .

Tsiantis, M., 2011 A transposon in tb1 drove maize domestication. Nat. Genet. **43:** 1048-1050.

van Heerwaarden, J., J. Doebley, W. H. Briggs, J. C. Glaubitz, M. M. Goodman *et al*, 2011 Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proceedings of the National Academy of Sciences **108:** 1088-1092.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. **91:** 4414-4423.

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

Weber, A. L., W. H. Briggs, J. Rucker, B. M. Baltazar, J. de Jesús Sánchez-Gonzalez *et al*, 2008 The genetic architecture of complex traits in teosinte (zea mays ssp. parviglumis): New evidence from association mapping. Genetics **180:** 1221-1232.

Wills, D. M., C. J. Whipple, S. Takuno, L. E. Kursel, L. M. Shannon *et al*, 2013 From many, one: Genetic control of prolificacy during maize domestication. PLoS Genet **9:** e1003604.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al*, 2005 The effects of artificial selection on the maize genome. Science **308:** 1310-1314.

Yu, J., J. B. Holland, M. D. McMullen and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics **178:** 539-551.

Zhang, Z. W., E. Ersoz, C. Q. Lai, R. J. Todhunter, H. K. Tiwari *et al*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. **42:** .

Zila, C. T., F. Ogut, M. C. Romay, C. A. Gardner, E. S. Buckler *et al*, 2014 Genome-wide association study of fusarium ear rot disease in the U.S.A. maize inbred line collection. BMC Plant Biology **14:** 1-15.

CHAPTER 3: Comparison of one-stage and two-stage genome-wide association studies

(The manuscript was submitted to G3: Genes | Genomes | Genetics)

Comparison of one-stage and two-stage genome-wide association studies

Shang Xue[1], Funda Ogut[2], Zachary Miller[3], Janu Verma[3], Peter J. Bradbury[4], James B. Holland[5*]

1. Bioinformatics Research Center, North Carolina State University, Raleigh, NC, 27695 USA
2. Faculty of Forestry, Artvin Coruh University, Artvin, Turkey
3. Institute for Genomic Diversity, Cornell University, Ithaca, NY, 14853 USA
4. USDA-ARS Plant Soil and Nutrition Research Unit, Ithaca, NY 14853 USA
5. USDA-ARS and Department of Crop Science, Campus Box 7620, Raleigh, NC 27695-7620.

* corresponding author, e-mail: james_holland@ncsu.edu

Abstract

Linear mixed models are widely used in humans, animals, and plants to conduct genome-wide association studies (GWAS). A characteristic of experimental designs for plants is that experimental units are typically multiple-plant plots of families or lines that are replicated across environments. This structure can present computational challenges to conducting a genome scan on raw (plot-level) data. Two-stage methods have been proposed to reduce the complexity and increase the computational speed of whole-genome scans. The first stage of the analysis fits raw data to a model including environment and line effects, but no individual marker effects. The second stage involves the whole genome scan of marker tests using summary values for each line as the dependent variable. Missing data and unbalanced experimental designs can result in biased estimates of marker association effects from two-stage analyses. In this study, we developed a weighted two-stage analysis to reduce bias and improve power of GWAS while maintaining the computational efficiency of two-stage analyses. Simulation based on real marker data of a diverse panel of maize inbred lines was used to compare power and false discovery rate of the new weighted two-stage method to single-stage and other two-stage analyses and to compare different two-stage models. In the case of severely unbalanced data, only the weighted two-stage GWAS has power and false discovery rate similar to the one-stage analysis. The weighted GWAS method has been implemented in the open-source software TASSEL.

Keywords: GWAS, one-stage analysis, weighted two-stage analysis, unbalanced datasets

Introduction

Genome-wide association studies are widely used to identify genes affecting complex traits in humans, animals, and plants (Huang and Han 2014; Stange *et al.* 2013; Zhang *et al.* 2012). Large sample sizes are required to achieve good statistical power in GWAS (Balding 2006). In addition, the number of markers to be tested is increasing rapidly as next-generation genomics techniques permit the acquisition of dense genome-wide marker data. In addition to the high dimensionality of the data, marker tests need to be adjusted for population structure and genomic relationships (Yu *et al.* 2006a). Combined, these factors result in significant computational burden for GWAS in many instances.

A common problem in GWAS is the need to account for extraneous non-genetic factors that affect the phenotypes. Estimating the effect associated with a single SNP while also including extraneous factors and modeling genetic background effects using a complex variance-covariance structure in linear mixed models can dramatically increase computation time for each marker test. To reduce computational demands for linear mixed model-based GWAS, several different strategies have been proposed in human, animal and plant studies to simplify and speed up computation time for the individual marker tests. In general, these strategies approach the problem using stage-wise procedures that first adjust the phenotypes for extraneous effects and second conduct linear mixed model GWAS on the adjusted phenotypes. For example, in human GWAS studies, two-stage regression analysis is a widely used strategy to test SNPs for associations with quantitative diseases (Laird *et al.* 2000; Naylor *et al.* 2009; Zeegers *et al.* 2004). First, residual effects ('adjusted-outcomes') for each individual are calculated by regressing the raw phenotype on covariates such as

demographic, clinical, and environmental factors. Second, the residual values are used as the dependent variable to test the association with SNP markers using a simple linear regression. Although this approach greatly reduces computational burden, it results in biased estimates of genotypic effects and reduced power (Che *et al.* 2012; Demissie and Cupples 2011).

Similarly, GWAS of domesticated animals often uses stage-wise approaches. In animal studies, the researcher may have available raw phenotypes of individuals and also their estimated breeding values (EBV) from pedigree-based analyses of historical data. Some animals may have genotypes and EBVs based on information from their relatives, but no direct phenotypes. Although EBVs have been used as dependent variables in GWAS (Becker *et al.* 2013; Johnston *et al.* 2011), this approach has a high false positive rate (Ekine *et al.* 2013). Consequences of using EBVs include varying levels of precision and 'shrinkage effect' among the values used to represent phenotypes of different individuals, a reduction in the sample variance of the phenotypes, and 'double-counting' of information from relatives (Garrick *et al.* 2009; Ostersen *et al.* 2011). As an alternative, the EBVs can be 'deregressed' (Garrick *et al.* 2009; Ostersen *et al.* 2011) to standardize the variance and influence of the individuals' EBVs while still accounting for information from relatives. The use of deregressed EBVs as dependent variables can improve the power of GWAS (Sell-Kubiak *et al.* 2015; Sevillano *et al.* 2015). Another alternative is to fit a mixed model to the data on individuals, accounting for either pedigree relationships or realized genomic relationships, then use the residuals for each individual as the dependent variable in a second stage genomic scan (Amin *et al.* 2007; Aulchenko *et al.* 2007; Lam *et al.* 2007). This approach, called 'GRAMMAR', dramatically speeds up computation time for the GWAS because each

marker test is a simple linear regression of residual values on genotype scores at one locus, however, this approach has low power in some cases (Zhou and Stephens 2012).

In contrast to human and animal studies, plant data are often generated from experimental designs in which the experimental units are field plots composed of multiple plants from a common family or inbred line, and often the designs are replicated across different environments. A typical linear model that accounts for environment, genotype, and genotype-by-environment interactions requires multiple random terms, each associated with a different variance component. Although a full model incorporating these random effects in addition to the effect of a single marker can be specified and fit using a mixed linear model, this approach is too computationally demanding for practical use in scanning thousands or millions of markers in a GWAS.

Software such as EMMA (Kang *et al.* 2008a), FaSTLMM (Lippert *et al.* 2011), and GEMMA (Zhou and Stephens 2012) were developed to solve the large computational problem in human datasets. EMMA takes advantage of the specific nature of the optimization problem in applying mixed models for association mapping by leveraging spectral decomposition of the genomic relationship matrix. By substantially decreasing the computational cost of each iteration, it enables convergence to a global optimum of the likelihood in variance-component estimation with high confidence by combining grid search and the Newton–Raphson algorithm. Since repeatedly estimating variance components for each SNP is computationally expensive, approximate algorithms like 'EMMA expedited' (called EMMAX) and 'population parameters previously determined' (called P3D) provide additional computational savings by assuming that variance parameters for each tested SNP are the same (Kang *et al.* 2010; Zhang *et al.* 2010). More recently, FaST-LMM and GEMMA

algorithms were proposed that can perform rapid GWAS analysis without assuming variance parameters to be the same across SNPs. FaST-LMM uses spectral decomposition of the genetic similarity matrix to transform (rotate) the phenotypes, SNPs and covariates. These transformed data are uncorrelated can be analyzed with a linear regression model. Similarly, GEMMA expedites each iteration by optimizing the efficiency of the computations required to evaluate the model likelihood and the first and second derivatives of the likelihood function. However, these software provided solutions for linear mixed models that only involve two random components: the polygenic background and error variance components.

In many plant studies, a full accounting of extraneous variation requires multiple random terms, each with a separate variance component, such that two-stage analyses are still necessary even with these improvements in algorithms to conduct linear mixed model GWAS. Two-stage approaches to GWAS for plant studies replicated across environments can take various forms. For example, in the first stage, the genotype effects can be fit as fixed or as random effects with no covariances, leading to the marginal prediction of genotype effects as either best linear unbiased estimation (BLUE) or best linear unbiased prediction (BLUP), respectively. In the second stage, the BLUEs or BLUPs of genotype obtained in the first stage may be fit as the dependent variable in a GWAS, in which the genotypes are treated as random with a variance-covariance matrix proportional to an estimated realized genomic relationship matrix (Aranzana *et al.* 2005; Lipka *et al.* 2013; Pasam *et al.* 2012b; Peiffer *et al.* 2013; Zhang *et al.* 2009).

Another approach to two-step GWAS involves using residuals, similar to the GRAMMAR method, but a complication is that replicated trials result in multiple residual values for each family. The first stage residuals could be averaged for each family and used

as inputs to the second stage. Alternatively, a term for independent family effects can be fit in the first stage model in addition to the polygenic family effects with covariance proportional to the relationship matrix (Oakey *et al.* 2007) and the independent line effect could be used as the dependent variable in the second stage. Finally, a three-step analysis procedure could be used. In the first step, BLUEs are computed for each line from plot level data, second BLUEs are fit as dependent variables in a linear mixed model including the relatedness matrix, and in the third step, residuals from second step are used for GWAS.

In addition to more complex experimental designs, another common feature of plant datasets is their unbalanced nature. Balanced data sets contain an equal number of observations for each combination of model factor levels. In contrast, plant breeding data sets often involve a series of trials over locations and years in which the genetic entries differ across environments. In addition, some data are often missing due to practical problems, and even within environments, experimental designs are often not balanced. The lack of balance impacts two-stage analyses in several ways. First, the BLUPs of lines that are represented by fewer records in the data set are shrunk back to the population mean to a greater extent than lines with more records. Second, the BLUEs or BLUPs obtained from the mixed model analysis of an unbalanced data set have variable standard errors. The variation in precision among the BLUEs or BLUPs is ignored in the second stage analysis, resulting in a loss of information. Simulation studies (Wang *et al.* 2011) indicate that unbalanced data in two-stage GWAS can cause more false positives.

Methods for analyzing a series of unbalanced performance evaluations of crops have been considered in detail in the context of maximizing the precision and accuracy of marginal predictions of the genetic entries (Möhring and Piepho 2009; SMITH *et al.* 2009). In

this context, single-step analysis is considered optimal, but may have high computational demand. Two-stage analysis of crop performance trials involves analyzing individual trials separately, then using family BLUEs from each trial as dependent variables in a simplified second stage analysis. Two stage analysis methods that use weighted analysis in the second step, in which weights are proportional to the precision of the BLUEs from the first step, often provide close approximation to the results of a single stage analysis (Möhring and Piepho 2009; SMITH *et al.* 2009). Additional complexity in the two stage analysis occurs when the residual values within environments are not independent, as occurs when spatial correlations are modeled in the residual variance structure. This results in lack of independence among the BLUEs; however approximate and exact methods have been developed to account for this lack of dependence as well as the variable precision among BLUEs in the second stage of analysis (Möhring and Piepho 2009; Piepho et al. 2012).

Inspired by previous work on two stage analysis of crop performance trials, George and Cavanagh (2015) proposed a two-stage GWAS approach that weights the BLUEs for families from the first stage in the linear mixed model GWAS scan. Their results indicate that the weighted two-stage GWAS provided comparable results to the single stage GWAS, and suggest that weighted two-stage analysis appears is a useful approach for conducting GWAS using data from multi-environment plant breeding trials. Several questions about the use of two-stage GWAS remain unanswered, however. First, it is unclear which summary variable is appropriate to use as a dependent variable for second-stage GWAS (Pasam et al. 2012a). In particular, the use of BLUP in two-stage analyses in which the hypothesis test is conducted only in the second stage has been criticized (Hadfield et al. 2010).

An alternate approach of using residuals from a first stage mixed model accounting for genomic relationships as dependent variables in the second stage may also be considered. Second, to our knowledge, none of the specialized open-source GWAS software packages have the flexibility to incorporate weights in the residual variance structure.

The objective of this study was to compare one-stage and several different two-stage GWAS methods using simulated real marker data and simulated phenotype data from a large maize diversity panel. Different levels of imbalance were imposed on the data to evaluate the effect of data imbalance. False discovery rate and power of marker-trait association tests and estimates of marker effects were compared for six different methods: one-stage analysis, two-stage unweighted analysis based on BLUPs or BLUEs from the first stage, two-stage weighted analysis based on BLUPs or BLUEs, and analysis of residuals after estimating random family effects with the relationship matrix. A weighted two-stage method that incorporates information on the variance of first-stage marginal predictions was implemented in the publicly available software TASSEL (Bradbury *et al.* 2007).

Material and Methods

Simulation data set

To reflect the real linkage disequilibrium (LD) structure of genome, the genotype used for simulation is from a subset containing 2480 lines representing almost all of the available inbred maize lines from the USDA Plant Introduction collection (Romay *et al.* 2013). After the initial imputation described in Romay et al. (Romay *et al.* 2013), ~16% of line-marker combinations were still missing. An additional imputation was performed using Beagle 4.0 (Browning and Browning 2009). A subset of 111,282 SNP markers was obtained by filtering out markers that have estimated imputation accuracy less than 0.995 and pairwise genotypic correlation greater than 0.5 by linkage-disequilibrium pruning using PLINK (Purcell *et al.* 2007). Data for g = 2480 inbred lines and n_{env} = 10 environments was simulated.

For each simulation data set, q = 10 or q = 50 SNPs were randomly sampled from among markers with minor allele frequency greater than or equal to 0.01, with the restriction that no pairs of markers were within 20 adjacent marker positions to avoid high LD between QTL. Markers selected as causal loci were assigned a constant QTL effect, other markers had zero effect. Genotypic values were created by simulating both QTL and polygenic background effects. The phenotype was simulated as the sum of major gene effects, polygenic genetic background, environmental effect and random error, for i = 1 to g lines, j = 1 to n_{env} environments and k = 1 to q QTL (q = 10 or 50).:

$$y_{ij} = \mu + \sum_{k=1}^{q} X_{ik}\alpha + G_i + \tau_j + \varepsilon_{ij}$$

where polygenic effect $G_i = (G_1, \ldots, G_g) \sim N(0, K\,\sigma_A^2)$, environmental effect $\tau_j = (\tau 1, \ldots, \tau_{nenv}) \sim N(0, \sigma_{env}^2 I)$, and random error $e = (e_1, \ldots, e_{nenv}) \sim N(0, \sigma_\varepsilon^2 I)$, X_{ik} are the coefficients for QTL

effects, reflecting the number of copies (0, 1, or 2) of the minor allele at each QTL in line *i*, and α is the effect of each QTL (which we set constant for all QTL within one replication). By changing σ_A^2, σ_{env}^2, σ_ε^2 and α we were able to simulate a reasonable range of heritabilities (Table 3.1).

We simulated three different genetic architectures varying for the number of QTL and QTL effect sizes: 10 QTL accounting for 69% of total genotypic variation, 10 QTL accounting for 23% of genetic variation, and 50 QTL accounting for 74% of genotypic variation (Table 3.1). The proportion of variance associated with QTL was estimated as the squared correlation between the sum of QTL effects and the phenotypic value of each line. The proportion of variance associated with polygenic background effects was estimated as the squared correlation between the polygenic effects and the phenotypic effects. The QTL and polygenic effects were not independent, so the total heritability was generally less than the sum of the QTL and polygenic variances. Furthermore, we assigned constant effects to all QTL within a genetic architecture setting, but since the QTL were randomly sampled from the true markers, their allele frequencies varied and the heritability due to QTL also varied among datasets. Hereafter, we refer to the average proportion of total heritability explained by QTL across datasets when this proportion is indicated, which means that heritability associate with each QTL is calculated as the average heritability accounted by each QTL.

We simulated three different scenarios for missing data: complete balanced data, randomly missing unbalanced data, and severely unbalanced data (Table 3.1). Balanced datasets had all lines evaluated at all environments with no missing values (24800 records). The two unbalanced datasets were generated from each complete dataset. Randomly unbalanced datasets contained a random subset of 50% of the data of the complete dataset

(12400 records). Severely unbalanced datasets had half of the lines evaluated at only one environment and the other half of lines evaluated at ten environments (13640 records). We generated 50 replicate complete data sets for each genetic architecture and two random subsets of each complete data set (100 replicates total) for each unbalanced data setting.

The realized additive genomic relationship matrix was estimated using R software version 3.0.0 (R Core Team 2013) based on observed allele frequencies (VanRaden 2008; method 1). The dataset for calculating relationship matrix is the whole genotype dataset.

Analysis methods

The simulated datasets were analyzed using each of six methods (Table 3.2).

One-stage model analysis

Suppose that *n* total observations were made on *g* lines so that $\boldsymbol{Y}$ is an $n \times 1$ vector of observed phenotypes. A linear mixed model for single-stage association mapping is expressed as:

$$Y = \mu + \boldsymbol{X_k}\beta_k + E + F + \varepsilon,$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of observed phenotypes, E are random macro-environment main effects, $\boldsymbol{X_k}$ is an $n \times 2$ or $n \times 3$ matrix consisting of a column of ones and one or two columns of dummy variables (depending on the number of genotypes at the marker) indicating the different genotypes at marker *k*. Markers with two genotypic classes require one column, whereas markers with three genotypic classes require two columns of dummy variables in $\boldsymbol{X_k}$. $\beta_k$is a vector of the fixed intercept, μ, and one or two genotypic effect estimates for marker *k*, F are random genetic background effects, and ε are residual effects. The

distributions of random effects are $E \sim N(0, \boldsymbol{M}_{env}\hat{\sigma}^2_{env})$, $F \sim N(0, \boldsymbol{M}_A\hat{\sigma}_A^2)$ and $\varepsilon \sim N(0, \boldsymbol{I}\sigma^2_\varepsilon)$. $\boldsymbol{M}_{env}$ is a block diagonal $n \times n$ matrix, indicating environmental effect correlations of 1 for observations within a common environment and 0 for pairs of observations in different environments. Each block is a $g_i \times g_i$ matrix where every element is 1 and g_i is the number of genotypes evaluated in environment i. $\boldsymbol{M}_{env}$ can be constructed as $\boldsymbol{M}_{env} = \boldsymbol{Z}_{env}\boldsymbol{Z}^T_{env}$, where $\boldsymbol{Z}_{env}$ is the $n \times e$ design matrix for environment effects (e = 10 in all of our data sets). $\mathbf{M}_\mathrm{A}$ is an $n \times n$ matrix, where each element is the realized genomic relationship coefficient for a pair of observations. $\mathbf{M}_\mathrm{A}$ can be constructed as $\boldsymbol{M}_A = \boldsymbol{Z}_g\boldsymbol{K}\boldsymbol{Z}^T_g$, where $\boldsymbol{Z}_g$ is the $n \times g$ design matrix for inbred line effects (g = 2480 in all of our data sets) and $\boldsymbol{K}$ is the $g \times g$ kinship matrix inferred from genotypes based on observed allele frequencies (VanRaden 2008, method1). Since estimating the variance components, particularly σ_A^2, is computationally intensive, we used the parameters previously determined method for the GWAS scan. For each simulation data set, the variance components were estimated once by restricted maximum likelihood from a reduced model with no fixed marker effects using ASReml. The variance components were then fixed at those values while subsequently testing each marker (Zhang *et al.* 2010). After obtaining estimates of σ^2_{env}, σ_A^2, and σ^2_ε using ASReml, the effect of marker k was estimated as:

$$\widehat{\boldsymbol{\beta}}_k = (\boldsymbol{X}_k^T\boldsymbol{V}^{-1}\boldsymbol{X}_\boldsymbol{k})^{-1}\boldsymbol{X}_k^T\boldsymbol{V}^{-1}\boldsymbol{Y}$$

In this formula, $\widehat{\boldsymbol{\beta}}_k$ is a vector of the fixed intercept, μ, and one or two genotypic effect estimates for marker k. Markers with two genotypic classes require only one genotypic effect estimate, whereas markers with three genotypic classes require two effect estimates. $\boldsymbol{X}_\boldsymbol{k}$ is an $n \times 2$ or $n \times 3$ matrix consisting of a column of ones and one or two columns of dummy variables (depending on the number of genotypes at the marker) indicating the different

genotypes at marker *k*. Markers with two genotypic classes require one column, whereas markers with three genotypic classes require two columns of dummy variables in $\boldsymbol{X_k}$. $\boldsymbol{V} = \boldsymbol{M}_{env}\hat{\sigma}^2_{env} + \boldsymbol{M}_A\hat{\sigma}^2_A + \boldsymbol{I}\hat{\sigma}^2_\varepsilon$. Note that this analysis requires inversion of **V** but the inverse can be computed once for a given **Y** vector and used for all marker tests in that data set. F-tests were used to test the null hypotheses of zero effect at each marker separately (Kang *et al.* 2008b; Kennedy *et al.* 1992; Yu *et al.* 2006b). Additive effects of markers were estimated as the appropriate linear combination of homozygous genotype effects.

Unweighted two stage analysis- BLUP

In the first stage, best linear unbiased predictions (BLUPs) for lines are calculated, assuming independence among the lines:

$$Y_{ij} = \mu + E_i + F_j + \varepsilon_{ij}$$

$$BLUP(Y_{.j}) = \mu + \hat{F}_j$$

where $Var(E_i) = \boldsymbol{I}\sigma^2_{env}$, $Var(F_j) = \boldsymbol{I}\sigma^2_f$, and $Var(\varepsilon_{ij}) = \boldsymbol{I}\sigma^2_\varepsilon$.

In the second stage, the $g \times 1$ vector of line BLUPs are treated as the dependent variable in the marker tests:

$$BLUP(Y_{.j}) = \mu + F_j + X_{jk}\beta_k + \varepsilon_{jk}$$

where $F_j \sim N(0, \boldsymbol{K}\sigma^2_A)$, $\boldsymbol{X}$ is a $g \times q$ design matrix for marker effects, β is a $q \times 1$ vector representing coefficients of the fixed marker effects, and $Var(\varepsilon) = I\,\sigma^2_\varepsilon$. The variance components σ^2_A and σ^2_ε are estimated once without fixed marker effects, and the resulting estimates used as fixed values in subsequent tests of each marker.

Unweighted two stage analysis- BLUE

In the first stage, best linear unbiased estimates (BLUEs) of lines are calculated:

$$Y_{ij} = \mu + E_i + F_j + \varepsilon_{ij}$$

$$BLUE(Y_{.j}) = \mu + \hat{F}_j$$

where $Var(E_i) = \boldsymbol{I}\sigma^2_{env}$, $Var(\varepsilon_{ij}) = \boldsymbol{I}\sigma^2_{\varepsilon}$, and F_j is treated as a fixed effect

In the second stage, the $g \times 1$ vector of line BLUEs is treated as the dependent variable:

$$BLUE(Y_{.j}) = \mu + F_j + X_{jk}\beta_k + \varepsilon_{jk}$$

where $F_j \sim N(0, \boldsymbol{K}\sigma^2_A)$, X is an $g \times q$ design matrix for marker effects, β is a $q \times 1$ vector representing coefficients of the fixed marker effects, and $Var(\varepsilon_{ij}) = \boldsymbol{I}\sigma^2_{\varepsilon}$. The variance components σ^2_A and σ^2_{ε} are estimated once without fixed marker effects, and the resulting estimates used as fixed values in subsequent tests of each marker.

Weighted –two stage analysis (BLUE and BLUP)

BLUEs or BLUPs are calculated in the first step, and the variance of each BLUE or BLUP is also recorded for use in the second step. The second step fits the following model:

$$\hat{Y}_{.j} = \mu + F_j + \boldsymbol{X}_{jk}\beta_k + \varepsilon_j$$

where $\hat{Y}_{.j}$ is a $g \times 1$ vector of BLUE or BLUP values for g lines, $F_j \sim N(0, \boldsymbol{K}\sigma^2_A)$, $\mathbf{X}$ is a $g \times q$ design matrix for marker effects, β is a $q \times 1$ vector representing coefficients of the fixed marker effects, and the distribution of residual effects is: $\varepsilon_j \sim N(0, \boldsymbol{I}\boldsymbol{w}\sigma^2_{\varepsilon})$, $\boldsymbol{w}_j = V(\hat{Y}_{.j})$. Thus, the weighted two-stage analyses differ by weighting the diagonal elements of the

residual variance-covariance matrix with the variances of the BLUEs or BLUPs. The variance components σ_A^2 and σ_ε^2 are estimated once without fixed marker effects, and the resulting estimates used as fixed values in subsequent tests of each marker.

Residual three stage

BLUEs are calculated as described above, then the following model is fitted in the second step:

$$BLUE(Y_{.j}) = \mu + F_j + \varepsilon_j$$

where $F_j \sim N(0, \boldsymbol{K}\sigma_A^2)$ and $Var(\varepsilon_{ij}) = \boldsymbol{I}\sigma_\varepsilon^2$. Then the vector of residuals, **ε**, is used as the dependent variable in the last step analysis:

$$resid(Y_{.j}) = \varepsilon_j = \mu + X_{jk}\beta_k + \varepsilon_j^*$$

where $\boldsymbol{X}$ is a $g \times q$ design matrix for marker effects, β is a $q \times 1$ vector representing coefficients of the fixed marker effects, ε_j^* are residuals from the final stage model and are distinguished from the line residuals from the second stage, and $Var(\varepsilon_j^*) = \boldsymbol{I}\sigma_{\varepsilon*}^2$. Each marker was tested separately in the final stage.

In multiple-step models, the initial stage linear mixed models were fitted used ASReml 3.0 software (Gilmour *et al.* 2009). Single-step analysis marker scans were conducted using a custom Python script that uses the variance components for the current data set estimated without fixed marker effects with ASReml, computes $\mathbf{V}^{-1}$, and uses that $\mathbf{V}^{-1}$ for testing each marker. The marker scan steps for other analyses were conducted using TASSEL (Bradbury *et al.* 2007). We implemented the weighted two-stage marker scan as an option in TASSEL.

Power and false discovery rate

Significant association tests were declared based on an empirical false discovery rate estimated for each analysis separately using the "qvalue" package in R (Bass *et al.* 2015). Markers with *q*-values less than or equal to 0.05 were treated as significantly associated with the simulated traits. Power was calculated as the ratio of number of true positive association tests to the total number of true QTLs. We then computed the true false discovery rate for each analysis as the proportion of false positive discoveries among all positive discoveries. False positives were defined as significant markers with small linkage disequilibrium (LD) r^2 values with true QTL. We evaluated two different LD thresholds to declare false positives: $r^2 < 0.1$ and $r^2 < 0.05$. At each of these thresholds, false discovery rate was calculated as the number of false positive association tests divided by the total number of positive (significant) tests.

Bias and mean square error of QTL effect estimation

Bias (b_c) and mean square error (MSE) of effect estimates at causal loci were calculated as follows:

$$b_c = \frac{1}{N_c}\sum_1^{N_c}(\hat{\beta}_c - \beta_c),$$

$$MSE_c = \frac{1}{N_c}\sum_1^{N_c}(\hat{\beta}_c - \beta_c)^2,$$

where N_c is the number of causal loci, $\hat{\beta}_c$ is the estimated additive marker effect for a causal locus and β_c is the true additive marker effect of causal locus.

Results

Power and false discovery rate

We simulated three different genetic architectures based on the observed allele frequencies, population structure, and linkage disequilibrium of a large panel of diverse maize inbred lines (Romay *et al.* 2013). The genetic architectures differed by the number and effect size of QTL (Table 3.1). Within one simulation data set, the QTL effects were constant, but the variation caused by each QTL differed because the allele frequencies differed among the randomly sampled SNPs chosen to represent QTL. The total genetic variation caused by QTL was nearly constant across replicated data sets for a given genetic architecture, however. Therefore, we characterized the genetic architectures by the average genetic variance associated with one QTL, which varied from 1.5% for the situation of 50 QTL to 6.9% for the 10 QTL with large effects (Table 3.1). The average total heritability (due to both QTL and background effects) for all three genetic architectures was consistent, varying only between 72% and 74% (Table 3.1). The sum of heritability due to QTL and heritability due to polygenic effects was larger than total heritability. This occurred because the QTL are not independent of the genetic background effects. As a result, the variance due to line polygenic background effects cannot be entirely separated from the variance due to QTL. The correlation between polygenic effects and total genetic value is influenced by the correlated effects of QTL, and vice-versa. This reflects a realistic and characteristic aspect of genetic architecture in populations with substantial structure and local LD.

For each complete simulated data set, we generated four additional subsets, reflecting two replicated samplings of two different missing data patterns. In addition to the variability

for QTL effects across replicated simulation data sets for a common genetic architecture, the different missing data patterns resulted in variability in the total heritability of line means. On average, however, the total heritabilities were very similar across genetic architectures for a given missing data pattern. On average, 50% randomly missing data resulted in a line mean heritability of about 55%, and the severely unbalanced situation resulted in a line mean heritability of about 35% (Table 3.1).

For balanced datasets, power of association tests was similar for two-stage analyses using BLUEs and BLUPs in the second stage (Figure 3.1). When data are balanced, the variance of BLUEs and BLUPs from the first stage of a two-stage analysis are homogeneous, weighted and unweighted second stage analyses are identical. For randomly unbalanced datasets, weighted and unweighted two-stage analysis using either BLUEs or BLUPs were not identical, but had similar power to detect associations (Figure 3.1). The largest differences among power of different analyses was observed with severely unbalanced datasets, where the weighted BLUE two-stage method had power about equal to the one-stage analysis, weighted and unweighted two-stage analyses using BLUPs were almost as good, but the unweighted two-stage analysis using BLUEs had a notable reduction in power. Analysis using the residuals from a model fitting genetic relationships in a previous step had lowest power among all analyses in all three experimental designs (Figure3.1, File B.1). Power of association tests had strong non-linear relationships with minor allele frequency, QTL effect size, and the proportion of missing data (Figure B.1).

False discovery rate (FDR) was similar for one-stage and all two stage analyses (using BLUEs or BLUPs and weighted or not; Figure 3.2, File B.2). The three-stage analysis based on residuals had very poor power and resulted in very few positive discoveries (File

B.1), so this method is not included in the comparisons of false discovery rate. All methods had an inflated FDR (actual FDR greater than the estimated rate) for the simplest genetic architecture (10 QTL accounting for 69% of true genetic variance; Figure 3.2). The FDR inflation was most severe for balanced data (FDR about 15%), but still exists for the unbalanced cases (Figure 3.2). We excluded markers with LD $r^2 > 0.1$ with causal QTL for the analysis reported in Figure 3.2. When we restricted the computation of FDR to only markers LD $r^2 < 0.05$ with QTL, FDR for balanced data and simplest genetic architecture dropped to about 7% (Figure 3.3). The strong dependency of FDR on LD with QTL in in this case indicates that even low levels of LD with causal markers can inflate FDR when the QTL effects are large.

The one-stage and weighted BLUE two-stage analyses had the best FDR when data were severely unbalanced (Figures 3.2 and 3.3). The unweighted BLUE two-stage analysis had higher FDR than the weighted BLUE two-stage analysis in most unbalanced conditions. Weighted BLUP two-stage analysis had similar FDR to one-stage for balanced and randomly unbalanced data, but had dramatically inflated FDR with severely unbalanced data. The weighted BLUP two-stage method had worse FDR than all other methods, including the unweighted BLUP method, when data were severely unbalanced. This effect remains even when FDR was computed for only markers with LD $r^2 < 0.05$ with QTL (Figure 3.3), and the inflation is strong even in the most polygenic architecture, so it is not simply a function of LD with causal QTL.

Because the power and FDR results suggest that the weighted (but not unweighted) BLUE two-stage analysis has similar properties to the one-stage analysis, we compared the distribution of genome-wide association test *p*-values for the weighted and unweighted

BLUE two-stage methods to the one-stage analysis. The distribution of *p*-values for all SNP tests was nearly identical for the one-stage and weighted BLUE two-stage methods (Figure 3.4). The severely unbalanced data case in particular results in many large deviations of unweighted BLUE two-stage *p*-values compared to the one-stage *p*-values, however (Figure 3.4).

Bias and MSE

Estimated effects of true QTL are biased downward for all methods except the one-stage and the two-stage BLUE methods (Figure 3.5). The residual 3-stage method results in the strongest downward bias under most combinations of genetic architecture and data structure. Methods using BLUPs also have downward bias under all conditions, as a result of the shrinkage of line values that occurs before the final step GWAS scan. For example, for the severely unbalanced data case and large-effect QTL effect, the weighted and unweighted BLUE two-stage analyses estimated the QTL effects with a bias of -0.5 units, or -4% of the true value, whereas the two-stage BLUP methods had downward bias around -6 units, around 50 % of the true value. These trends are also reflected in the mean square error variance for effect estimates (Figure B.2; File B.3).

Computational time

For a single analysis analyzing 2480 lines with ~110,000 markers using a single core (Intel Xeon E5-2680v3), one-step analysis required 146 hours, whereas the GWAS scan (second step) of two-stage methods required 30 hours for unweighted and 32 hours for weighted methods. The two-stage analyses also involved a first step to estimate BLUEs or

BLUPs, which required 0.3 hours. Three-stage analysis using residuals required around 10 hours for the GWAS scan, in addition to 0.4 hours for the first two steps of the analysis.

Discussion

Our simulation used the real marker data on a large and diverse maize inbred line panel with substantial population structure. For each simulation data set, we assigned a very small proportion of markers to have true causal effects, allowing us to test power of association tests directly at causal variants. In real GWAS studies, however, the researcher cannot assume that the causal variants have been genotyped and included in the marker data set. Instead, researchers rely on sufficient marker density and linkage disequilibrium to detect association signals at markers physically linked in close proximity to causal variants, while trying to reduce the influence of longer-range (and unlinked) LD due to population structure on association tests. In general, linkage disequilibrium decays rapidly in diverse maize panels. On average, LD is below $r^2 = 0.2$ for markers separated by more than one kb, but there is a large variance around this average value, such that a small proportion of distantly separated marker pairs may still have high LD (Romay *et al.* 2013).

Therefore, the power of association tests reported in Figure 3.1 and File B.1 represents the optimal but unrealistic situation of having the casual variants in the marker data set. Estimating power at markers linked to causal variants introduces some complication into the concept of power, because different researchers have different criteria for considering an association to be a true positive result or a false positive result, depending on how close the marker is to the true variant in physical or genetic distance. Power of association tests reported here was somewhat lower than power for detecting QTL accounting for similar proportions estimated from a simulation study of the maize nested association mapping (NAM) design (Yu *et al.* 2006a). The lower power of QTL detection in a

diversity panel than in a balanced multiple biparental family design like NAM is expected, as the NAM panel has a simple, known population structure that can be accounted for in the analysis, and more balanced allele frequencies. Power in this study was strongly related to allele frequency, power increased sharply in most cases for minor allele frequencies between 0.05 and 0.20 (Figure B.1). Below 5% minor allele frequency, power was very low (~ 0.12) except for the largest-effect QTL simulated (Figure B.1). Missing data reduced power as expected, and the effect was greater for severely unbalanced data, even though the total proportion of missing data in that case is not highest (Figure B.1). The power of association tests reported here also reflect a rather large sample of inbred lines (2480), which is larger than many association panels currently studied. Thus, in smaller panels, power of detection will be lower than that reported here.

The simulation results clearly demonstrate that the 'GRAMMAR' method for conducting a GWAS scan on residuals from a model including random family effects with covariances proportional to the estimated realized genomic relationship coefficients has worse performance (lower power and higher bias) than other methods evaluated in this study. A similar result was reported by Zhou and Stephens (Zhou and Stephens 2012) based on a comparison between results of GEMMA, EMMAX and GRAMMAR algorithm scans of a human genetics data set. The particularly poor performance of GRAMMAR in our simulation is likely related to numerous close relationships among lines in the collection of maize inbreds studied, coupled with a preponderance of low minor allele frequencies. In this situation, the QTL effects may tend to be restricted to relatively few groups of closely-related lines, and therefore mostly absorbed into the polygenic background effects.

The two-stage methods using line BLUPs from an initial analysis that regards line effects as random but independent had power of association tests about equal to the one-stage analysis (Figure 3.1). Weighting had almost no effect on the power or bias of two-stage BLUP methods (Figures 3.2, 3.3, and 3.4). However, both BLUP methods had considerable bias in estimation of QTL effects, due to the shrinkage of line values that are used as dependent variables in the GWAS scan. In general, the unweighted BLUP two-stage method performed similar to or better than the weighted two-stage BLUP method, because the weighted BLUP two-stage had considerably inflated FDR when data were severely unbalanced (Figures 3.2 and 3.3). This may have occurred because BLUP itself introduces shrinkage toward the mean of the line values, and the shrinkage is greatest for lines with most missing data. Weighting during the GWAS then decreases the relative influence of lines with highest prediction error variance, which are the same lines whose values have shrunk most toward the mean. The double action of shrinking and underweighting the values of lines with least data increases false discoveries. This can happen when, by chance, lines carrying a rare SNP have complete data, whereas much of the rest of the population (half of lines in our simulation) has a large proportion of missing data. Since the SNP alleles are not independent of the background genetic effects, a subgroup carrying the rare allele but having (by chance) little missing data will reflect the average polygenic effect of the subgroup (even though the line relationships were not accounted for in the model). The line BLUPs in such a subgroup are less likely to be shrunk toward the mean, and, in addition, they have higher relative influence on the association test than other lines. The combination of SNP frequencies correlated with polygenic effects and differential shrinkage and weighting between allelic

classes may cause SNP association tests to absorb polygenic effects and produce false positive discoveries.

Users should be cautioned against making inferences about heritability from the relative proportion of genetic and residual variances estimated in the second step of a two-step analysis. The weighting changes the scaling of the residual variance component that is estimated. Therefore, the relative magnitude of the genetic and residual variance components is influenced both by heritability and the scale of the weighting factor.

Our simulation assumed no covariances among the BLUEs computed in the first stage of the analysis. In practice, however, BLUEs may be estimated from complex unbalanced designs such as incomplete block designs, or using models involve spatial correlations among the residuals in first step, leading to correlations among the resulting BLUEs. Various approximate weighting methods have been proposed to handle this situation (Möhring and Piepho 2009; Piepho *et al.* 2012; SMITH *et al.* 2009). Piepho *et al*. (Piepho *et al.* 2012) also developed a method for exact two-stage analyses, and such a method might also be implemented for GWAS studies, as it has for genomic prediction analysis (Schulz-Streeck *et al.* 2013). In many cases, however, a simple weighting method like the one used here is sufficient to recover the properties of a single-stage analysis (Möhring and Piepho 2009).

Our results suggest that the weighted BLUE two-stage analysis can be recommended across a range of genetic architectures and missing data structures. The power, FDR, and bias of weighted BLUE two-stage analysis was very similar to the one-stage analysis, but with substantially reduced computing time. The method has been implemented in version 5 of the publicly available open-source software TASSEL, available from http://www.maizegenetics.net/tassel. Users need to add an additional file containing the

variances of the line BLUEs to the usual two-stage analysis work flow. The file containing the variances of the BLUEs has the same format and header as a TASSEL trait file (so, the same as the file containing the BLUEs themselves). Multiple phenotypes are allowed but they must share the same header and have exactly the same genotype ("taxa") order. In the TASSEL GUI interface, users must first load the data from four files containing BLUEs, variances of BLUEs, genotype scores, and the relationship matrix, respectively. The BLUE and genotype score data need to be joined using "Intersect Join". Then, selecting the joined phenotype and marker score data set, the kinship data set and the BLUE variances data set together, the user can choose "weighted MLM" in analysis tab to perform the analysis. To run the analysis from the command line, the same four files are required. An example of the use of command line execution of a weighted analysis using TASSEL version 5 is provided in File B.4.

Data Availability

All data and software codes used to generate simulation data sets and conduct analyses are available for download at: https://drive.google.com/a/ncsu.edu/file/d/0B7Iwvphs9t5hdkdvWUUyVUY0a2M/view?usp=sharing

If the paper is accepted for publication, we plan to post these files permanently at Dryad or a similar public database.

Acknowledgments

S.X. was supported by National Institutes of Environmental Health Sciences training grant T32 ES007329 to the North Carolina State University Bioinformatics Research Center and National Science Foundation (NSF) award IOS-1127076; J.B.H. was supported by NSF awards IOS-1127076 and IOS-1238014 and by the U.S. Department of Agriculture, Agricultural Research Service. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575, and the North Carolina State University High Performance Computing Center. We thank Jung-Ying Tzeng at Bioinformatics Research Center of North Carolina State University for providing statistical insights on one-stage analysis and evaluating GWAS performance.

Figures and Tables

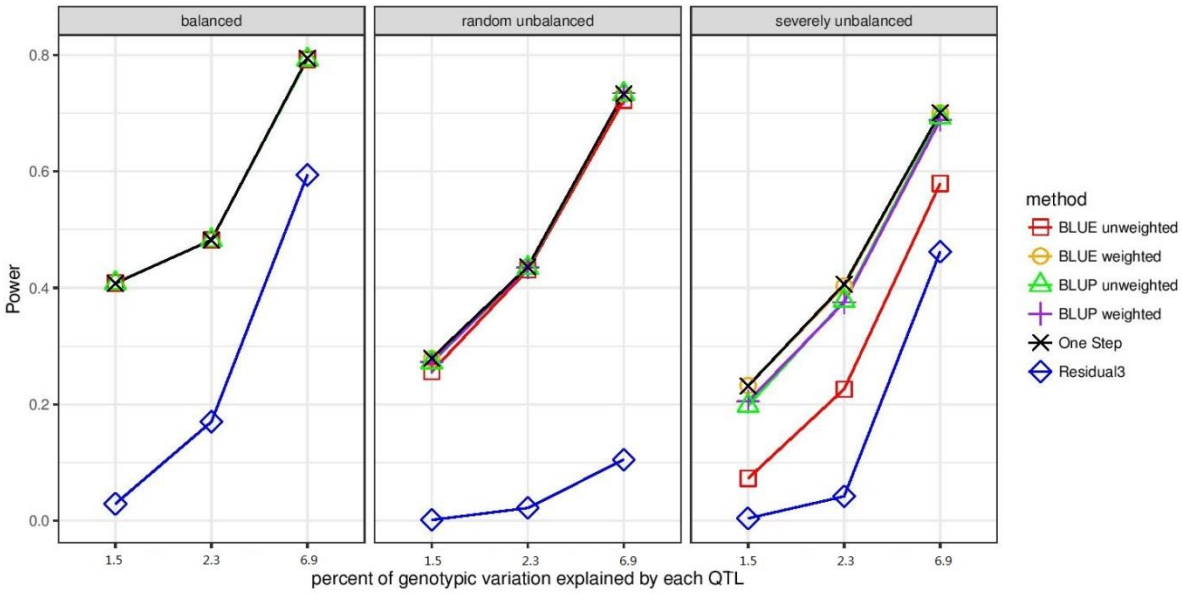

Figure 3.1. Power of six different association testing methods to detect causal variants for three different genetic architectures and three levels of data imbalance. Balanced datasets had all lines evaluated at all environments with no missing values (24800 records). Randomly unbalanced datasets contained a random subset of 50% of the data of the complete dataset (12400 records). Severely unbalanced datasets had half of the lines evaluated at only one environment and the other half of lines evaluated at ten environments (13640 records).

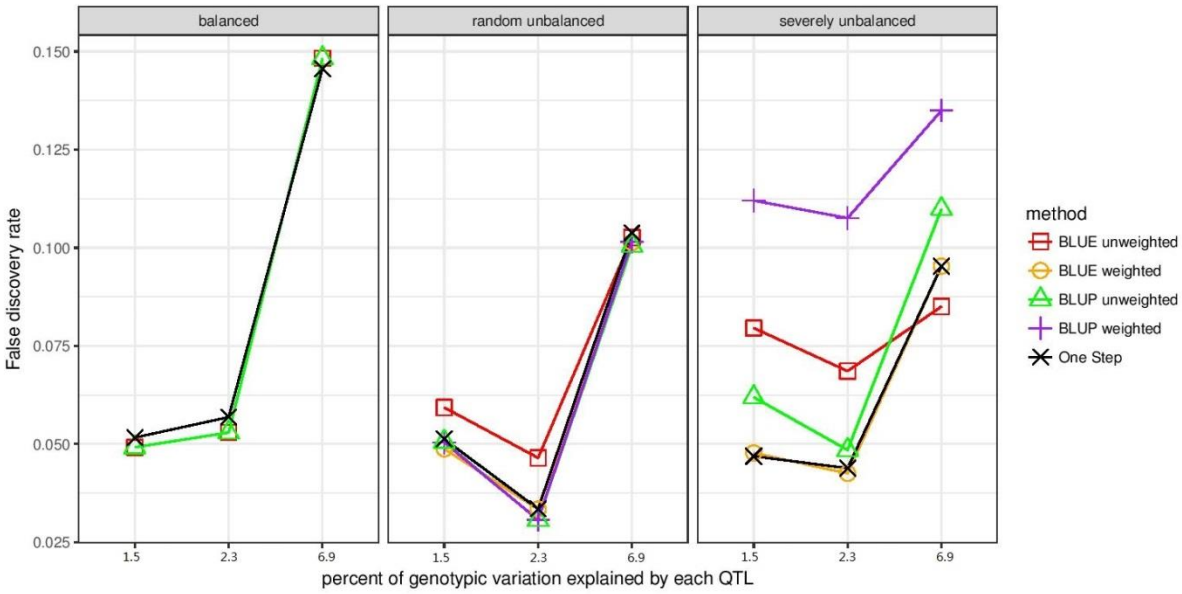

Figure 3.2. False discovery rate when false positives were defined as markers that had LD r2 < 0.1 with true QTL and were declared significant at the empirically estimated q < 0.05. False discovery rate is the proportion of false positives defined this way among all markers declared significant. False discovery rate for residual 3-step method is not shown, since it identified very few significant markers.

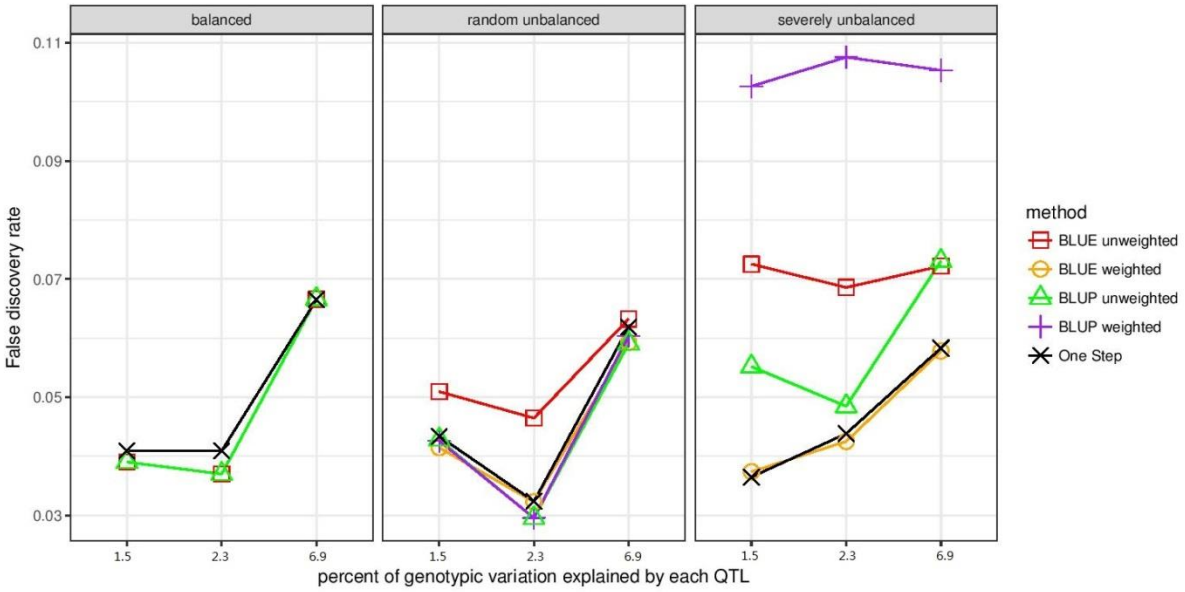

Figure 3.3. False discovery rate when false positives were defined as markers that had LD r2 < 0.05 with true QTL and were declared significant at the empirically estimated q < 0.05. False discovery rate is the proportion of false positives defined this way among all markers declared significant. False discovery rate for residual 3-step method is not shown, since it identified very few significant markers.

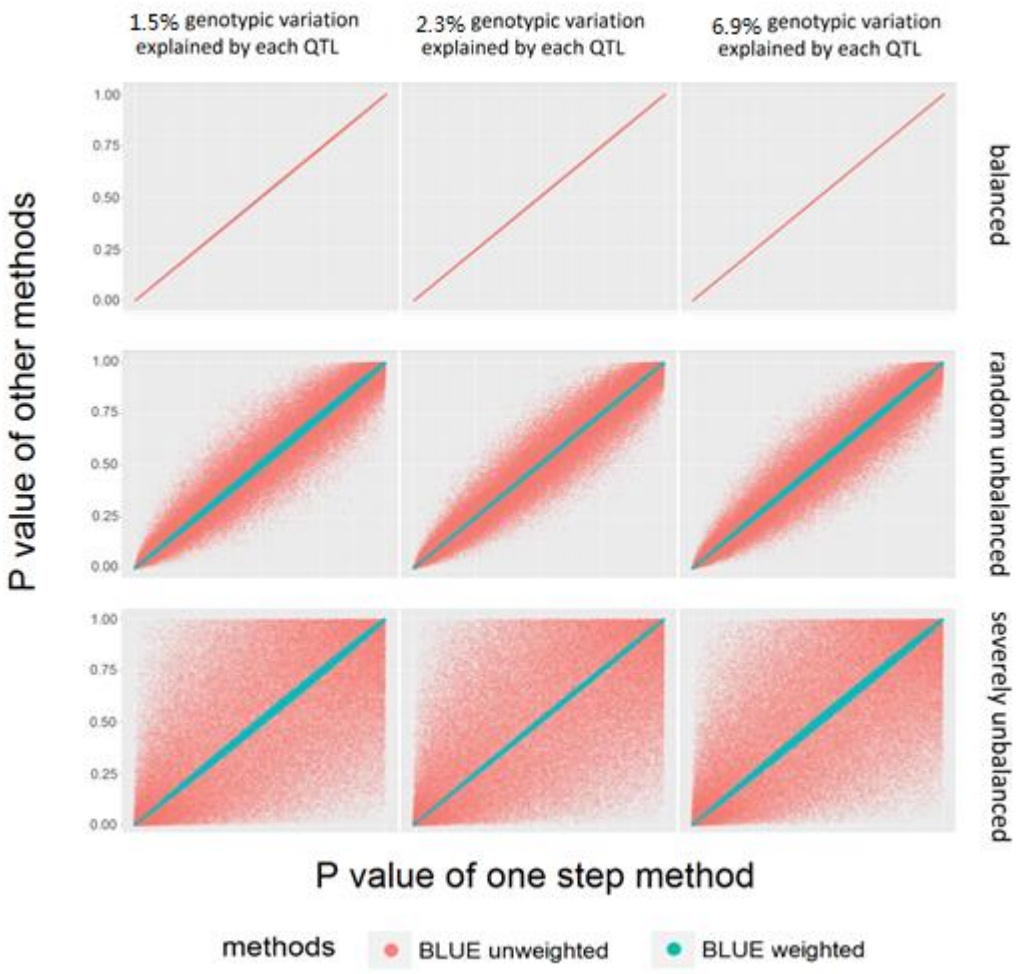

Figure 3.4. Distributions of all genome-wide marker association test p-values using the one-step analysis (Y-axis) or two-step analysis (X-axis) for three different genetic architectures and three levels of data imbalance. *P*-values from weighted BLUE two-stage method are in blue and *p*-values from unweighted BLUE two-stage method are in orange.

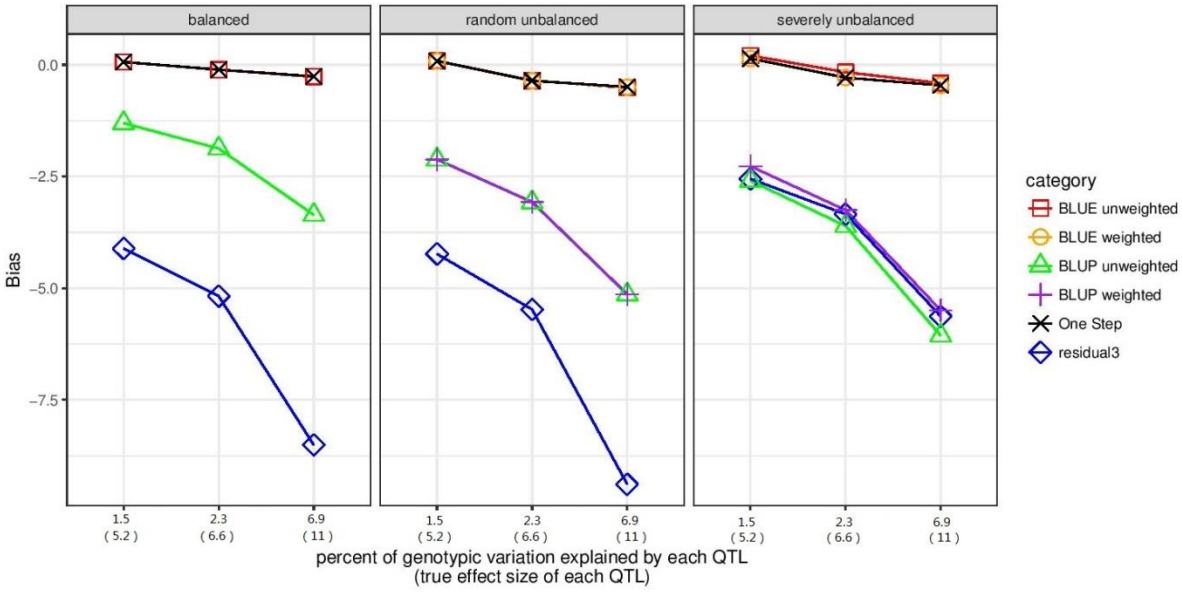

Figure 3.5. Bias of QTL effect estimates from different GWAS methods for three different genetic architectures and three levels of data imbalance.

Table 3.1. Parameter settings for simulation study.

Experimental design	QTL number	QTL effect	σ_A^2	σ_{env}^2	σ_ε^2	Proportion of genotypic variance due to all QTL	Average proportion of genotypic variance due to one QTL (%)	Total heritability	Heritability associated with QTL effects	Heritability associated with polygenic effects	Heritability accounted by each QTL (%)
Balanced	10	11	109	100	2500	0.69	6.9	0.72	0.48	0.22	4.8
Balanced	10	6.6	289	100	2500	0.23	2.3	0.72	0.17	0.55	1.7
Balanced	50	5.2	105	100	2500	0.74	1.5	0.74	0.55	0.21	1.1
Random unbalanced	10	11	109	100	2500	0.69	6.9	0.55	0.37	0.15	3.7
Random unbalanced	10	6.6	289	100	2500	0.23	2.3	0.55	0.13	0.37	1.3
Random unbalanced	50	5.2	105	100	2500	0.74	1.5	0.56	0.42	0.16	0.8
Severely unbalanced	10	11	109	100	2500	0.69	6.9	0.34	0.24	0.09	2.4
Severely unbalanced	10	6.6	289	100	2500	0.23	2.3	0.35	0.09	0.25	0.9
Severely unbalanced	50	5.2	105	100	2500	0.74	1.5	0.36	0.26	0.09	0.5

Table 3.2. GWAS methods used.

Method	First stage Model	First stage distribution of F_j	Summary result of first stage	Second stage model	Second stage distribution of F_j	Summary result of second stage
One-stage	$Y_{ijk} = \mu + E_i + F_j + X_{jk}\beta_k + \varepsilon_{ijk}$	$F_j \sim N(0, K\sigma_A^2)$	Effect of marker $k = \hat{\beta}_k$			
Unweighted BLUE 2-stage	$Y_{ij} = \mu + E_i + F_j + \varepsilon_{ij}$	Fixed effect	$BLUE(Y_{.j}) = \mu + \hat{F}_j$	$BLUE(Y_{.j}) = \mu + F_j + X_{jk}\beta_k + \varepsilon_{jk}$	$F_j \sim N(0, \boldsymbol{K}\sigma_A^2)$	Effect of marker $k = \hat{\beta}_k$
Unweighted BLUP 2-stage	$Y_{ij} = \mu + E_i + F_j + \varepsilon_{ij}$	$F_j \sim N(0, \boldsymbol{I}\sigma_f^2)$	$BLUP(Y_{.j}) = \mu + \hat{F}_j$	$BLUP(Y_{.j}) = \mu + F_j + X_{jk}\beta_k + \varepsilon_{jk}$	$F_j \sim N(0, \boldsymbol{K}\sigma_A^2)$	Effect of marker $k = \hat{\beta}_k$

Table 3.2 continued

Weighted BLUE 2-stage	$Y_{ij} = \mu + E_i + F_j + \varepsilon_{ij}$	Fixed effect	$BLUE(Y_{.j}) = \mu + \hat{F}_j$ $w_j = V(\hat{Y}_{.j})$	$BLUE(Y_{.j}) = \mu + F_j + X_{jk}\beta_k + \varepsilon_{jk}$ $\varepsilon_{jk} \sim N(0, \boldsymbol{I}\boldsymbol{w}\sigma_\varepsilon^2)$	$F_j \sim N(0, \boldsymbol{K}\sigma_A^2)$	Effect of marker $k = \hat{\beta}_k$
Weighted BLUP 2-stage	$Y_{ij} = \mu + E_i + F_j + \varepsilon_{ij}$	$F_j \sim N(0, \boldsymbol{I}\sigma_f^2)$	$BLUP(Y_{.j}) = \mu + \hat{F}_j$ $w_j = V(\hat{Y}_{.j})$	$BLUP(Y_{.j}) = \mu + F_j + X_{jk}\beta_k + \varepsilon_{jk}$ $\varepsilon_{jk} \sim N(0, \boldsymbol{I}\boldsymbol{w}\sigma_\varepsilon^2)$	$F_j \sim N(0, \boldsymbol{K}\sigma_A^2)$	Effect of marker $k = \hat{\beta}_k$
Residual 3-stage	Step 1: $Y_{ij} = \mu + E_i + F_j + \varepsilon_{ij}$ Step 2: $BLUE(Y_{.j}) = \mu + F_j + \varepsilon_j$	Step 1: Fixed effect Step 2: $F_j \sim N(0, \boldsymbol{K}\sigma_A^2)$	Step 1: $BLUE(Y_{.j}) = \mu + \hat{F}_j$ Step 2: $resid(Y_{.j}) = \varepsilon_j$	$resid(Y_{.j}) = \mu + X_{jk}\beta_k + \varepsilon_{jk}$	NA	Effect of marker $k = \hat{\beta}_k$

References

Amin, N., C. van Duijn M. and Y. S. Aulchenko, 2007 A genomic background based method for association analysis in related individuals. PLoS ONE 2: e1274.

Aranzana, M. J., S. Kim, K. Zhao, E. Bakker, M. Horton *et al*, 2005 Genome-wide association mapping in *arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet **1:** e60.

Aulchenko, Y. S., D. de Koning and C. Haley, 2007 Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics **177:** 577-585.

Balding, D. J., 2006 A tutorial on statistical methods for population association studies. Nat. Rev. Genet. **7:** 781-791.

Bass, A. J., A. Dabney and D. Robinson.2015 Qvalue: Q-Value Estimation for False Discovery Rate Control. R Package version 2*.2.2, Http://github.com/jdstorey/qvalue*

Becker, D., K. Wimmers, H. Luther, A. Hofer and T. Leeb, 2013 A genome-wide association study to detect QTL for commercially important traits in swiss large white boars. PLoS ONE **8:** e55951.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al*, 2007 TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics **23:** .

Browning, B. L., and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. **84:** 210-223.

Che, R., A. Motsinger-Reif and C. C. Brown, 2012 Loss of power in two-stage residual-outcome regression analysis in genetic association studies. Genet. Epidemiol. **36:** .

Demissie, S., and L. A. Cupples, 2011 Bias due to two-stage residual-outcome regression analysis in genetic association studies. Genet. Epidemiol. **35:** .

Ekine, C. C., S. J. Rowe, S. C. Bishop and D. de Koning, 2013 Why breeding values estimated using familial data should not be used for genome-wide association studies. G3: Genes|Genomes|Genetics **4:** 341-347.

Garrick, D. J., J. F. Taylor and R. L. Fernando, 2009 Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution **41:** 55.

Gilmour, A. R., B. J. Gogel, B. R. Cullis and R. Thompson, 2009 *ASReml User Guide Release 3.0*. VSN International, Ltd, Hemel Hempstead, UK.

Hadfield, J. D., A. J. Wilson , D. Garant, B. C. Sheldon and L. E. B. Kruuk, 2010 The misuse of BLUP in ecology and evolution

. Am. Nat. **175:** 116-125.

Huang, X., and B. Han, 2014 Natural variations and genome-wide association studies in crop plants. Annu. Rev. Plant Biol. **65:** 531-551.

Johnston, S. E., J. C. Mcewan, N. K. Pickering, J. W. Kijas, D. Beraldi *et al*, 2011 Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. Mol. Ecol. **20:** 2555-2566.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al*, 2008a Efficient control of population structure in model organism association mapping. Genetics **178:** .

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al*, 2008b Efficient control of population structure in model organism association mapping. Genetics **178:** 1709-1723.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Kong *et al*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42:** 348-354.

Kennedy, B. W., M. Quinton and J. A. van Arendonk, 1992 Estimation of effects of single genes on quantitative traits. . **70:** 2000-2012.

Laird, N. M., S. Horvath and X. Xu, 2000 Implementing a unified approach to family-based tests of association. Genet. Epidemiol. **19:** S36-S42.

Lam, A. C., M. Schouten, Y. S. Aulchenko, C. S. Haley and D. de Koning, 2007 Rapid and robust association mapping of expression quantitative trait loci. BMC Proceedings **1:** S144-S144.

Lipka, A. E., M. A. Gore, M. Magallanes-Lundback, A. Mesberg, H. Lin et al.2013 Genome-Wide Association Study and Pathway-Level Analysis of Tocochromanol Levels in Maize Grain.

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson *et al*, 2011 FaST linear mixed models for genome-wide association studies. Nat Meth **8:** 833-835.

Möhring, and Piepho, 2009 Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci. **49:** .

Naylor, M. G., S. T. Weiss and C. Lange, 2009 Recommendations for using standardised phenotypes in genetic association studies. Hum Genomics **3:** .

Oakey, H., A. P. Verbyla, B. R. Cullis, X. Wei and W. S. Pitchford, 2007 Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. Theor. Appl. Genet. **114:** 1319-1332.

Ostersen, T., O. F. Christensen, M. Henryon, B. Nielsen, G. Su *et al*, 2011 Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genetics Selection Evolution **43:** 38.

Pasam, R. K., R. Sharma, M. Malosetti, F. A. van Eeuwijk, G. Haseneyer *et al*, 2012a Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biology **12:** 1-22.

Pasam, R., R. Sharma, M. Malosetti, F. van Eeuwijk, G. Haseneyer *et al*, 2012b Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biology **12:** 1-22.

Peiffer, J. A., S. Flint-Garcia, N. De Leon, M. D. McMullen, S. M. Kaeppler *et al*, 2013 The genetic architecture of maize stalk strength. PLoS ONE **8:** e67066.

Piepho, H., J. Möhring, T. Schulz-Streeck and J. O. Ogutu, 2012 A stage-wise approach for the analysis of multi-environment trials. Biometrical Journal **54:** 844-860.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira *et al*, 2007 PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. **81:** 559-575.

Romay, M. C., M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts *et al*, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. **14:** .

Schulz-Streeck, T., J. O. Ogutu and H. Piepho, 2013 Comparisons of single-stage and two-stage approaches to genomic selection. Theor. Appl. Genet. **126:** 69-82.

Sell-Kubiak, E., N. Duijvesteijn, M. S. Lopes, L. L. G. Janss, E. F. Knol *et al*, 2015 Genome-wide association study reveals novel loci for litter size and its variability in a large white pig population. BMC Genomics **16:** 1049.

Sevillano, C. A., M. S. Lopes, B. Harlizius, E. H. A. T. Hanenberg, E. F. Knol *et al*, 2015 Genome-wide association study using deregressed breeding values for cryptorchidism and scrotal/inguinal hernia in two pig lines. Genetics, Selection, Evolution : GSE **47:** 18.

SMITH, A. B., B. R. CULLIS and R. THOMPSON, 2009 The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. The Journal of Agricultural Science **143:** 449-462.

Stange, M., H. F. Utz, T. A. Schrag, A. E. Melchinger and T. Wuerschum, 2013 High-density genotyping: An overkill for QTL mapping? lessons learned from a case study in maize and simulations. Theor. Appl. Genet. **126:** .

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. **91:** 4414-4423.

Wang, H., K. P. Smith, E. Combs, T. Blake, R. D. Horsley *et al*, 2011 Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theor. Appl. Genet. **124:** 111-124.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki *et al*, 2006a A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38:** .

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki *et al*, 2006b A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38:** 203-208.

Zeegers, M., F. Rijsdijk and P. Sham, 2004 Adjusting for covariates in variance components QTL linkage analysis. Behav. Genet. **34:** 127-133.

Zhang, H., Z. Wang, S. Wang and H. Li, 2012 Progress of genome wide association study in domestic animals. Journal of Animal Science and Biotechnology **3:** 26-26.

Zhang, Z., E. S. Buckler, T. M. Casstevens and P. J. Bradbury, 2009 Software engineering the mixed model for genome-wide association studies on large samples. Briefings in Bioinformatics **10:** 664-675.

Zhang, Z., E. Ersoz, C. Lai, R. J. Todhunter, H. K. Tiwari *et al*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. **42:** 355-360.

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. **44:** 821-824.

CHAPTER 4: Genomic distribution of deleterious recessive variation in the ZeaSynthetic population

(The manuscript is prepared for submission to Genetics)

Genomic distribution of deleterious recessive variation in the ZeaSynthetic population
(The manuscript is prepared for submission to Genetics)

Shang Xue[1], Ginnie Morrison[2], Sherry Flint-Garcia[2,3], James Holland[4*]

1. Bioinformatics Research Center, North Carolina State University, Raleigh, NC, 27695 USA

2. Division of Plant Sciences, University of Missouri, Columbia, MO 65211. USA

3. USDA-ARS Plant Genetics Research Unit, Columbia, MO 65211. USA

4. USDA-ARS Plant Science Research Unit and Department of Crop and Soil Sciences, Campus Box 7620, Raleigh, NC 27695-7620. USA

* corresponding author, e-mail: james_holland@ncsu.edu

Abstract

The gene pool of teosinte, the close wild relative of maize, is a potential resource for discovery of novel and agriculturally beneficial alleles. However, teosinte does not produce maize-like ears or seeds and will not flower under long day lengths of the USA Corn Belt growing season. Therefore, the value of teosinte alleles to agriculture requires testing them in adapted maize genetic backgrounds. As a bridge between teosinte and maize, a ZeaSynthetic population has been developed by introgressing small teosinte genomic regions into a population segregating for diverse maize alleles and random-mating for six generations. A sample of 923 plants from the population was self-fertilized and also crossed as male parents to an equal number of plants used as female parents to produce 923 pairs of S1 and S0 full-sib progeny families. The 1826 parents were genotyped by sequencing to obtain 474690 high quality SNPs on each parent. Progeny families were measured for seven agronomic traits in six environments. Due to several generation of random mating, this population has a unique population structure where only progeny families that share same parents have high relatedness. A linear mixed model was used to estimate additive and dominance effects at each SNP and simulation was conducted to evaluate power and bias of genetic effect estimates from this model. Simulation results shows that the linear mixed model is a reasonable model with low bias estimating genetic effects and high power detecting QTLs. Analysis of the real data showed that loci with rare alleles and loci in lower recombination regions of the genome tend to have larger additive and dominance effects. These results suggest that recessive deleterious alleles (genetic load) tends to be concentrated in lower

recombination regions, and that favorable alleles for agriculture are more likely to be at higher starting frequencies and found at loci in higher recombination regions.

Introduction

Maize was domesticated 6000-10000 years ago from its wild relative *Zea* mays subsp. *parviglumis* in Southern Mexico (Doebley *et al.* 2006; Doebley 2004; Matsuoka *et al.* 2002). During domestication and the subsequent process of strong purging selection during the development of inbred lines, modern maize lines lost genetic diversity. Estimates of the amount of diversity lost are generally agreed to be around 30% in maize (BUCKLER *et al.* 2001; Zhang *et al.* 2002). The teosinte genetic pool might possess agriculturally beneficial alleles that no longer existing in modern maize population. However, discovering beneficial alleles from teosinte is hindered by the inability of teosinte to flower at an appropriate time in temperate growing environments because of its photoperiod sensitivity (Hung *et al.* 2012) and its reproductive structures, which do not resemble ears or seeds of maize (Doebley *et al.* 1990; Doebley and Stec 1993; Doebley *et al.* 1997). The morphology and environmental adaptation of teosinte mask expression of alleles that could be favorable in the context of maize genetic backgrounds and modern agricultural production systems. On the other hand, standard biparental mapping population of maize do not segregate for alleles unique to teosinte, so we cannot infer the potential utility of teosinte alleles from such populations,

To provide a bridge between the gene pools of modern maize and teosinte, Dr. Sherry Flint-Garcia of the USDA-ARS in Missouri created a 'ZeaSynthetic' population by introgressing teosinte genomic fragments into a gene pool segregating for diverse maize

alleles. Therefore, this population segregates for small teosinte genomic segments as well as alleles from diverse maize inbred lines, permitting direct comparisons between the effects of allelelic variants segregating within maize versus variants that are fixed in maize but segregating or fixed for different forms in teosinte. Initially, a "NAM Synthetic" population was created by intermating B73 and the 25 other founder inbred lines of the maize nested association mapping (NAM) population, which were chosen to maximize the representation of genetic diversity (McMullen *et al.* 2009a).Two generations of random mating were conducted within the NAM Synthetic population In parallel, a separate synthetic population was created by intermating 11 BC1 families from backcrosses of 11 teosinte (*Zea mays* subsp. *parviglumis*) donor parents to the recurrent parent B73 (Liu *et al.* 2016). Then the ZeaSynthetic population was initiated by bulk pollen random mating of the teosinte x B73 backcross synthetic population as the pollen donor to the second generation of the NAM synthetic population. After six generation of random mating, 923 randomly-chosen individuals were self-fertilized to create S1 families, and also crossed as male parents to an equal number of randomly-chosen individuals to create 923 full-sib families.

These families sampled from the ZeaSyn6 population can be used to assess the effects of diverse maize and teosinte alleles in a common gene pool, and to identify favorable as well as deleterious teosinte alleles in the context of adapted maize genomic background. However, there are several challenges for the analysis. First, the many generations of random mating in the ZeaSynthetic population is expected to eliminate population structure and reduce linkage disequilibrium. Therefore, we expect few close relationships among families within either the S1 or full-sib families, however, across these two sets, progeny families derived from a common parent will have close relationships. In general, the limited population structure and

lower linkage disequilibrium should improve power and resolution of genome-wide association studies (GWAS), however, the appropriate adjustment for the paired family population structure is not obvious, since correcting directly for pedigree relationships would absorb much of polygenic variation. Another challenge to analysis of this population arises from the fact that the phenotypic measurements are averages over many progenies within each family, but the genotypic data are based on sequencing the parents. The progenies are segregating at many loci, so the probabilities of different genotypic classes within each family are needed, and these can be derived based on the parental genotypes. GWAS can be conducted as multiple regression of the family phenotype means on the expected additive and dominance effect coefficients for each family. Finally, the phenotyped families differ for level of inbreeding, and adjustment for the inbreeding coefficient is necessary to remove confounding effects of inbreeding depression at unlinked regions of the genome.

GWAS of this population will provide genome-wide estimates of additive and dominance effects associated with individual SNPs. These estimates can be used to test what kinds of genomic variants contribute most to inbreeding depression in this population. Inbreeding depression refers to the reduced survival and fertility of offspring of related individuals, it occurs in wild animal and plant populations as well as in humans, indicating that genetic variation in fitness traits exists in natural populations(Charlesworth and Willis 2009). Heterosis refers to the phenomenon that progeny of diverse varieties of a species or crosses between species exhibit greater biomass, speed of development, and fertility than both parents (Birchler *et al.* 2010). Heterosis from outcrossing is the converse of inbreeding depression. Because intercrossing inbred strains improves yield through heterosis, and this is

a major component of yield improvement in maize, the genetic basis of these effects has been debated since the early twentieth century.

Two contrasting models of genetic architecture have been proposed for heterosis and the success of hybrids in maize and other species. The overdominance model of heterosis predicts that distinct alleles at a common locus confer heterozygote advantage when combined, and this effect is aggregated over many loci in the genome (Birchler *et al.* 2006). Alternatively, the dominance model predicts that heterosis is driven by dominance effects and the appearance of overdominance occurs due to repulsion-phase linkage disequilibrium of between deleterious alleles ('pseudo-overdominance';(Gerke *et al.* 2015)). Population genetics models predict that pseudo-overdominance effects will be strongest in regions of low recombination because of the reduced ability to obtain favorable recombinants (Gerke *et al.* 2015; Hill and Robertson 1966; McMullen *et al.* 2009a).The majority of genetic evidence indicates that the dominance model can explain most of the genetic control of inbreeding depression and heterosis (Crow 2000; Gardner and Lonnquist 1959; Gerke *et al.* 2015; Mezmouk and Ross-Ibarra 2013; Moll *et al.* 1963), although there is still some debate(Birchler *et al.* 2010).

It is important to understand the genetic basis underlying inbreeding depression and heterosis because it will influence the choice of optimal breeding strategies. If loci with heterozygote advantage are common, artificial selection in agricultural species should select for strains that manifest substantial heterosis. Crop plants with a uniform, highly heterozygous genotype with high fitness are often desirable, and could perhaps be achieved by asexual seed production (Grossniklaus *et al.* 2001). However, if heterosis in crops is mainly caused by deleterious mutations, it might be better to exclude these alleles to produce high-yield mutant-free strains. Also, understanding the genetic basis of inbreeding depression can

help to address the question of why genetic variation in fitness-related characteristics exists in so many species, including humans (Lewontin 1974).

Recombination has a large impact on evolution in general and plant breeding in particular by promoting the diversity necessary to respond to continually changing environments and preventing the build-up of genetic load by decoupling linked deleterious and beneficial alleles. Recombination varies among genomic regions in both animal and plants (Anderson *et al.* 2001; Gaut *et al.* 2007; Jensen-Seaman *et al.* 2004) and variation in recombination rates may be associated with the distribution of variants for genetic load or inbreeding depression. There are two main reasons for varying crossing-over rates at the molecular level. First, the chromatin structure heavily influences recombination rates in plants. Heterochromatic regions generally reduce crossovers, however, knockout of cytosine-DNA-methyl-transferase (MET1) resulted in genome-wide CpG hypomethylation and increased the proportion of crossovers (Colome-Tatche *et al.* 2012; Mirouze *et al.* 2012; Yelina *et al.* 2012). Fu et al.(Fu and Dooner 2002) demonstrated that, in maize, crossovers were suppressed in regions with repetitive DNA sequences derived from retrotransposons compared to sequences including and nearby functional coding gene sequences. Highly repetitive retrotransposon-derived sequences tend to be heterochromatic in maize (Slotkin and Martienssen 2007). Second, nucleotide content may also be associated with recombination rate possibly due to the effect of GC-biased gene conversion (bGC)(Serres-Giardi *et al.* 2012). Maize experienced a modest bottleneck in genetic diversity (Tenaillon *et al.* 2004), although strongly deleterious variants were likely purged during the inbreeding process leading to the founder lines, many weakly deleterious alleles can be found segregating at low frequencies among inbreds (Mezmouk and Ross-Ibarra 2013).

McMullen et al. (McMullen *et al.* 2009a) observed that excess residual heterozygosity was enriched in pericentromeric regions of maize inbred lines, which suggested that selection against deleterious recessive alleles has been less efficient in these regions because of reduced recombination frequency, as predicted by the Hill-Robertson effect (Hill and Robertson 1966). In addition, other studies observed that deleterious alleles enriched in low-recombination regions, as expected because reduced recombination permits deleterious alleles to hitchhike to high frequency during selective sweeps (Hartfield and Otto 2011; Rodgers-Melnick *et al.* 2015). However, a recent study shows that putatively deleterious nonsynonymous polymorphisms in maize were not significantly enriched in regions of low recombination (Mezmouk and Ross-Ibarra 2013).

The objectives of this study were to obtain reliable estimates of additive and dominance effects throughout the genome estimated from the ZeaSynthetic population, and to use these estimates to explore the genomic distribution of deleterious recessive variation that underlies inbreeding depression. After using simulation studies to verify that a mixed linear model fitting additive and dominance effects of segregating progenies simultaneously had good power and low bias in this population structure, we implemented GWAS utilizing a linear mixed model for seven agriculturally important traits of the ZeaSynthetic population. Given estimates of additive and dominance effects at SNPs distributed throughout the genome, we were able to estimate to relate the distribution of genetic load variants to recombination rate, allele frequency, and to teosinte origins. These comparisons provided insight into the variants underlying genetic load that segregate only in teosinte, an outcrossing species, compared to variants that are shared with maize lines that have been through several generations of inbreeding and purging selection.

Material and Methods

ZeaSynthetic population

In Puerto Rico 2008, ZeaSyn0, the initial generation of the population was created, from a bulk pollen random mating of a teosinte x B73 backcross synthetic population (pollen donor) and the second generation of a nested association mapping panel (NAM) synthetic population (maternal parents). The NAM Synthetic population was created by randomly crossing B73 and the 25 diverse founder inbred lines of the maize Nested Association Mapping population (McMullen *et al.* 2009b) plus Mo17. Before crossing to teosinte backcross synthetic population, the NAM synthetic was intermated at random for two generations after the initial F1 generation. The teosinte backcross synthetic was created by intermating 11 BC1 families from backcrosses of 11 teosinte donor parents (*Zea mays* subsp. *parviglumis* accessions Ames21785, Ames21786, Ames21789, Ames21809, Ames21812, Ames21814, Ames21889, PI384063, PI384065, PI384066, and PI384071 from the USDA Plant Introduction collection) to the recurrent parent B73. Each BC1 individual contained approximately 25% of teosinte alleles in a B73 genetic background (Liu *et al.* 2016). To sample the whole genome of each teosinte donor, a sample of 40 BC1 progenies from each teosinte donor was used in crossing. Individuals in the ZeaSynthetic are expected to contain, on average, 38% B73, 12% teosinte, and 2% of each diverse NAM founder parent.

After the ZeaSynthetic initial population was established, at every generation approximately 250 - 350 random plants were selected and intermated. ZeaSyn6 was formed after the sixth generation of random mating. In the summer of 2013 in Missouri, around 1000 plants were randomly selected from ZeaSyn6 and self-pollinated created selfed (S1) families.

These plants were mated as fathers to an equal number of randomly selected ZeaSyn6 plants, creating full-sibling outcross (S0) families used in this study.

Field evaluations

The 923 pairs of S1 and full-sib families were evaluated in field trials conducted at six environments. To reduce the scale of phenotyping within each environment, the family pairs were randomly assigned to six sets of 153 or 154 family pairs. Five of six sets were evaluated in each environment, and each family was evaluated in five of six environments. The experimental environments were Columbia, MO; Clayton, NC; and Aurora, NY in each of the 2014 and 2105 growing seasons.

A split-plot design was used at each environment. Family pair was the whole-plot factor and inbreeding generation was the sub-plot factor. Each whole plot consisted of two sub-plots of full-sib and S1 families derived from a common parent in side-by-side plots. An alpha-design was also imposed on the whole-plots so that each incomplete block consisted of 20 whole-plots. One whole plot within each incomplete block was assigned a commonly repeated check: the F1 and F2 generations of the hybrid B73 x Mo17. Plots were a minimum of 1.8 m in length, with 0.97 m distance between adjacent rows and a 0.6-m alley at the end of each row. A minimum of 12 seeds were planted in each row, and plots were thinned to approximately 0.3-m distance between plants. Trait measurements included the median days from sowing to anthesis and silk emergence (DTA and DTS), and ear height (distance from ground to the topmost ear-bearing node) and plant height (distance from ground to the tassel-bearing node) measured on three plants per plot. At maturity, all ears were harvested from two randomly selected plants within each plot. The number of ears per plant, the total weight

of seeds from all ears per plot, and the average kernel weight were measured from these sampled plants.

The phenotypic data were analyzed within a mixed model that included random environment macro-effects and within environment spatial variation, plus fixed effects of families. The model used was

$$Y = \mu + E_i + Set_j + E * S_{ij} + Block(Set * Env)_{ijk} + Family_l + Env * Family_l + \varepsilon_{ijkl}$$

Where E_i is the random effect of the *i*th environment, Set_j is the random effect of the *j*th set of entries, $E * S_{ij}$ is the random effect of the interaction between E_i and Set_j, $Block(Set * Env)_{ijk}$ is the effect of the *k*th incomplete block nested within Set_j at environment *i*, $Family_l$ is the fixed effect of the *l*th S0 or S1 generation family, $Env * Family_l$ is the random interaction of environment *i* and family *l*, and ε_{ijkl} is the random residual effect on the *ijkl*th plot. All random effects had independent and identical distributions, except the residual term within each site was modeled with a first order anisotropic autoregressive covariance structure, involving a common error variance component and autoregressive correlations in the row and column directions of the field grids (Gilmour *et al.* 1997). Each environment was allowed to have a unique set of residual variance and spatial correlation coefficients. Best linear unbiased estimators (BLUEs) for the families averaged across all other model effects were estimated and used as input phenotypes to the genome-wide association study.

Genotyping

DNA was isolated from each of the 1846 parental plants. A genotyping-by-sequencing (GBS) protocol based on (Elshire *et al.* 2011; Glaubitz *et al.* 2014) was used to genotype the

parents. Briefly, the DNA from each plant was digested with restriction enzyme ApeK1 and adapters with specific 12-bp DNA adapters ("bar codes") were ligated to digested fragments for each parental plant. Multiplex sequencing libraries were constructed by combining ligated fragments from each of 96 parental plants; each plant within each library had a unique bar code adaptor to identify the origin of the sequenced fragments. DNA was sequenced on an Illumina Hi-Seq 2500 system.

Sequencing reads were deconvoluted to assign reads to plants of origin. Reads were then aligned to the B73 refgen version 2 sequence and SNPs called relative to the B73 sequence. SNPs were then filtered based on minimum depth of three reads per site and the presence of at least one minor allele. This resulted in a set of 474690 high-quality SNPs. Missing data in this set was then imputed using the FILLIN algorithm (Swarts *et al.* 2014) in TASSEL version 5 (Bradbury *et al.* 2007). Reference haplotypes for FILLIN were taken from the filtered samples themselves plus a set of 1941 doubled haploids (DH) derived from the Syn4 stage of the Zea Synthetic population using the FILLINFindHaplotypesPlugin function of TASSEL version 5. Any heterozygote calls in the DH genotypes were changed to missing ('NA'). After imputation, only 56.3% of calls were missing (from 71.1%) on a set of 474690 SNPs. These SNPs were further filtered such that at least 90% of the taxa were genotyped at a site, for a final working set of 30185 SNPs.

Calculating kinship matrix from parental genotypes

TASSEL version 5 was used to calculate an additive kinship matrix for the 1846 parental plants using the method of Endelman and Jannink (Endelman and Jannink 2012). When calculating the kinship matrix, the unimputed, high-quality dataset was further filtered for

sites with 50% or less missing data resulting in a total of 107346 sites used for calculating the kinship matrix. Based on this parental relationship matrix, an expected relationship matrix for progeny families was calculated as the expected value for progenies within that family. Details for the computation of the progeny family relationship matrix are given in File C.1.

Calculating expected additive and dominance coefficient

The expected additive and dominance coefficients at each locus for progeny families were calculated based on the genotypes of the parents according to Table C.1. The additive effect coefficient represents the expected number of minor alleles in each progeny within a family. The dominance effect coefficient represents the expected proportion of heterozygous individuals within a family.

Genome-wide association study

First, five different models differing for population structure and inbreeding effect adjustments (Table C.2) were tested on real data and Q-Q plots inspected for control of false discoveries. The phenotyped families differ for level of inbreeding, and adjustment for the inbreeding coefficient is necessary to remove confounding effects of inbreeding depression at unlinked regions of the genome. A model with no inbreeding effect had the worst deviation from the expected distribution of *p*-values for null hypotheses (Figure C.1).Three fixed effect models using no population structure control or eigenvalues from relationship matrix were also fit to the data (Models 1-3, Table C.2), but all had poor Q-Q plots (Figures C.1, C2 and C3). We tried two different mixed models, one mixed model used a relationship matrix based on all markers; the second used a relationship based on all markers except for those from the

chromosome being scanned. The former mixed model is simpler to implement and provides a consistent background model for all marker tests and it resulted in a better and reasonable Q-Q plot (Figures C.4, C.5 and C.6), so it was chosen for all subsequent GWAS analyses. The final selected model was:

$$Y_i = \mu + F_i + a_{ij} + d_{ij} + g_i + e_i,$$

where Y_i is the BLUE for family *i*, F_i is the expected inbreeding coefficient for family *i*, a_{ij} and d_{ij} are the additive and dominance coefficients for family *i* at SNP *j*, g_i is the polygenic background effect of family *i*, and e_i is the residual effect. All terms except g_i and e_i are fixed effects. The polygenic background effects are distributed as $g_i \sim N(0, \mathbf{K}\sigma_g^2)$, where σ_g^2 is the background polygenic additive variance component and **K** is the realized relationship matrix for the families.

A custom python code (File C.2) was written to implement the EMMAX algorithm for analysis. Variance components estimates for each trait were obtained from a reduced model with no marker effects in ASReml (Gilmour *et al.* 2009):

$$Y_i = \mu + F_i + g_i + e_i,$$

where all terms are equal to the previous model.

The variance component was then considered fixed at the REML estimate from the previous model in subsequent GWAS scan tests. The EMMAX algorithm uses a singular value decomposition of the **K** matrix to eliminate the need for iterations to solve the mixed model equation. A composite F-test for additive and dominance effects together was conducted to determine if one marker have both effects significantly different from zero.

Relationship between genetic effect and allele frequency

To reduce the variation in marker effect estimates due to differences in precision that arise from allele frequency differences, the genetic effects were standardized by dividing the estimates by their standard errors. To quantify the relationship between genetic effect and allele frequency, a simple linear model was fitted to the absolute value of standardized genetic effect and allele frequency. A second-order relationship was also quantified by fitting a simple linear regression of absolute value of standardized genetic effect on squared allele frequency.

Relationship between genetic effect and recombination rate

A refined linkage map of NAM was recently developed using GBS, which produced a total of 3 600,000 reliable SNPs. An iterative process of imputation and linkage mapping was conducted to produce a final consensus linkage map with complete map scores at 7386 psuedo-markers with a uniform resolution of 0.2 cM per marker (Ogut *et al.* 2015; Swarts *et al.* 2014). By using this linkage map, genetic distance per unit physical distance and genetic effect were obtained for SNPs falling in every 0.2 cM interval. Genetic distance per unit physical distance is proportional to recombination rate. As before, genetic effects were standardized by dividing them by their standard errors. To quantify the relationship between genetic effect and recombination rate, a simple linear model was fit to the average absolute value of standardized genetic effect and genetic distance per unit physical distance. A second-order relationship was quantified by fitting a simple linear regression of standardized absolute value of genetic effect on squared genetic distance per unit physical distance.

Simulation to estimate power and bias of association tests

After selecting the mixed model with the common relationship matrix based on the Q-Q plot of real data, we wanted to check power and bias of this method for this data set. The real marker data were used and the real phenotype values for each trait were used as true background polygenic values for each line. In each replication of the simulation, 10 SNPs were chosen randomly from among all SNPs with association test *p*-values > 0.5 for the trait to represent simulated QTL. These SNPs are not likely to be linked to true QTL which could confound the power and bias calculations. The additive effect of each simulated QTL was set as 0.25, 0.5, or 1 standard deviation of the original family BLUEs for each trait. The dominance effect was chosen so that ratio of dominance to additive effects was 0, 0.5, 1, or 2. This resulted in 12 combinations of simulated additive and dominance QTL effect sizes for each trait. Within each setting for QTL effect sizes, all QTL had the same additive and dominance effects. Simulated phenotypes for each family were then calculated by summing the original family BLUE value, each QTL additive effect times the additive effect coefficient for each simulated QTL in the family, and each QTL dominance effect times the dominance effect coefficient for each simulated QTL in the family. For each combination of trait and additive and dominance effect settings, 200 replicated simulation data sets were created and tested.

For each combination of parameter settings, power was calculated as the ratio of number of true positive association tests to the total number of true QTLs. The *p*-value threshold to declare significant association tests was the Bonferroni-corrected alpha value, 0.05 divided by the number of markers (1.6×10^{-6}). Bias was calculated as the mean of

estimated values minus the true simulated effect at simulated QTL. Bias is reported as a proportion of the true causal effect.

Results

Relationships among progeny families

As expected, the random mating conducted for six generations after the establishment of the ZeaSynthetic population almost entirely eliminated population structure among the parental plants. The mean diagonal and off-diagonal values of the parent kinship matrix were 0.9 and -0.0005, respectively.

Using the parental realized relationships, we estimated expected offspring relationships among families. Because there were no strong relationships among any of the parents, the families derived from different parents had near zero realized relationships (Figure C.7). Paired S1 and full-sib families derived from a common parent, however, had strong relationships (Figure C.7).

The estimated inbreeding coefficient for S1 families ranged from 0.11 to 0.80, with a mean near 0.46, close to expected value of F = 0.5. The average estimated inbreeding coefficient for full-sib families ranged from -0.04 to 0.15 with a mean of -0.0003, very close to the expected value of F = 0.

Simulation results for bias

Since the true simulated QTL effect sizes were determined by the standard deviation of each trait, the scale of the bias varied among traits. To normalize results, bias is reported as a proportion of the true effect size (Figure 4.1). In general the bias of estimates of additive effects was acceptable, varying from -10% to 15% of the true value across simulation settings and traits. Similarly, the bias of estimates of dominance effects were acceptable in

general, ranging from -20% to 20%, but usually with an upward bias. Effects on number of ears showed larger bias than other traits, perhaps because the data for this trait are discrete and estimates of genetic effects were not as accurate as for other traits.

Simulation results for power of detecting QTL

As the simulated QTL effects increase, the power of successful QTL detection increases, as expected (Figure 4.3). Power to detect QTL was high (0.74 to 0.99) for QTL with an additive effect of 0.5 or more phenotypic standard deviations. Power was lower for QTL with additive effects of only 0.25 standard deviations, and power was increased for the smaller effect QTL as their dominance effects increased (Figure 4.3). Power was also related to heritability of the trait for QTL with smaller effects, but as the QTL effect became large, heritability had less influence. These results suggest it is possible that we will miss some small effect QTL, and we are more likely to miss QTL with small additive and dominance effects, which could contribute to some bias in the estimation of the importance of dominance when summarizing results from real data.

Surprisingly, minor allele frequency had limited effect on QTL detection power, and the influence of minor allele frequency was greatest for QTL with smallest effects (Figure C.8). The site frequency spectrum of the SNP data reveals that the mode of the distribution is around 0.02 minor allele frequency (Figure C.9). Very rare alleles do not occur in the data because data filtering imposed a lower limit on the minor allele frequency (MAF>0.005). However, a larger proportion of SNPs have minor allele frequencies between 0.1 and 0.5 than is expected for a natural population in drift-mutation equilibrium(Marth *et al.* 2004).

This is due to the relatively balanced contribution of the different parental stocks to the population and the use of large effective population sizes during the intermating generations.

Relationship between additive effect and dominance effect in real data

For loci with less common alleles (MAF < 0.1), additive effects estimates and dominance effects estimates were negatively correlated for all traits (Figure 4.4). For commonly segregating SNPs (MAF > 0.3), however, additive and dominance effects estimates appeared unrelated for all traits except kernel weight (Figure 4.5).

Relationship between MAF and absolute value of genetic effect in real data

For kernel weight, there were no obvious trends between MAF and absolute value of additive effect estimates and dominance effect estimates. However, for all other traits, there was a clear trend that as MAF was smaller, loci tended to have larger effects (Figures 4.6 and 4.7).When minor allele frequencies are lower, their effect estimates are less unreliable, and therefore are expected to have a higher variance. Therefore, genetic effects were standardized by divided them by their standard errors to normalize the effects and reduce the dependence of effect magnitude on minor allele frequency.

Relationship between absolute value of genetic effect and recombination rate

As the local recombination rate around a SNP decreases, allelic effects tended to be larger (Figures 4.8 and 4.9). This was observed for both the actual effect estimates and the scaled effect estimates normalized according to their standard errors.

Discussion

For almost all traits, rare alleles (MAF < 0.1) tended to have larger effects than more common alleles (MAF > 0.3). Furthermore, there was a common pattern of the relationship between additive and dominance effects within the rare allele class. Most rare alleles tended to have negative additive effects and positive dominance effects. The sign of the additive effects is not arbitrary; it indicates whether the minor allele has a positive effect relative to the major allele because the homozygous major allele class was coded as 0 and the homozygous minor allele class was coded as 2. Further, we scaled the additive and dominance effects by their standard errors to remove the expected increase in effect estimate variation for alleles with lower frequency (and thus, lower precision). However, the general trend of the relationship between genetic effect and MAF was the same even without the standardization (Figures C.10-C.13). The rare alleles in this group are recessive and unfavorable because their negative additive effects were paired with larger positive dominance effects. Alleles with more negative additive effects and larger positive dominance effects also tended to occur more frequently than expected by chance in lower recombination regions. These results are congruent with the dominance model of heterosis, as they will tend to generate repulsion phase haplotypes that are not easily disrupted by recombination. The combination of large positive dominance effects and lower recombination "protects" these unfavorable alleles from selection, even under purging selection (since a small proportion of heterozygosity is maintained even under recurrent selfing).

Another group of rare alleles for most traits have positive additive effects and negative dominance effects. This group of rare alleles are more favorable than the common

alleles, and may represent standing variants that are at higher frequency within local populations of maize but did not spread globally. Their correlation with negative dominance effects is surprising, it suggests that these are largely recessive favorable alleles. Their recessivity also limits the effectiveness of selection under outcrossing conditions to increase their allele frequencies.

The genetic architecture of seed weight seems to be distinct from the other traits because, although its heritability was similar to the other traits, additive effects on seed weight were negatively correlated with dominance effects for both common and rare alleles. This result may suggest that alleles affecting seed weight are not under strong selection. Although higher seed weight is expected to confer greater fitness to seedlings, it is generally negatively correlated with numbers of seeds per ear, resulting in a tradeoff between fertility (number of offspring) and fitness of the offspring. This tradeoff may lead to reduced selection pressure or balancing selection for alleles that affect seed weight.

Linear models were fit to the data to quantify the regression of genetic effects on minor allele frequency. The absolute value of genetic (standardized by their SE) effect regressed on MAF (Table C.3). MAF and genetic effects are significantly negatively correlated for most traits. Linear models were also fit to quantify the relationship between recombination rate and genetic effect. Absolute values of average genetic effects of SNPs falling in every 0.2 CM interval were regressed on genetic distance per unit physical distance. The results indicate that recombination rate and genetic effects are significantly negatively correlated for most traits (Table C.4). Since R^2 for these models are very low, a polynomial model of second order relationship and GAM model were also fit to the data, but

the fitted curves were similar to the linear model because most of the values are clustered near the zero on the y-axis.

The recombination rate was quantified by genetic distance per unit physical distance, however, recombination rate in maize is strongly correlated with functional gene density. Much of the maize genome is not coding sequence, and those regions might not have any function. If true, we would expect smaller effect estimates for SNPs in such low gene density regions. Our results contradict this expectation: association effects tended to be greater in low recombination regions! To better understand this relationship, we plan to compute the proportion of bases within functional coding regions for each 0.2 cM interval as an alternative measure that incorporates gene density per cM into the recombination rate estimate. This analysis is ongoing.

We expect alleles derived from teosinte to be minor alleles, but not necessarily rare in this population, as 13% of the parentage of the population traces back to teosinte. Haplotype analysis will be required to identify unique haplotypes in the population that derive from teosinte, we are developing methods to do this. Another possibility to test the effects of teosinte-specific alleles is to identify any individual SNP alleles at loci where the maize parents of the population were fixed for one allele. This would miss some haplotypes that are unique to teosinte but share some marker states with maize, but it may identify a subset whose effect distribution may be representative of the larger set of teosinte alleles.

Acknowledgments

S.X. was supported by National Institutes of Environmental Health Sciences training grant T32 ES007329 to the NCSU Bioinformatics Research Center and National Science Foundation award IOS-1127076, J.H. was supported by National Science Foundation awards IOS-1127076 and IOS-1238014 and by United States Department of Agriculture, Agricultural Research Service.

Figures

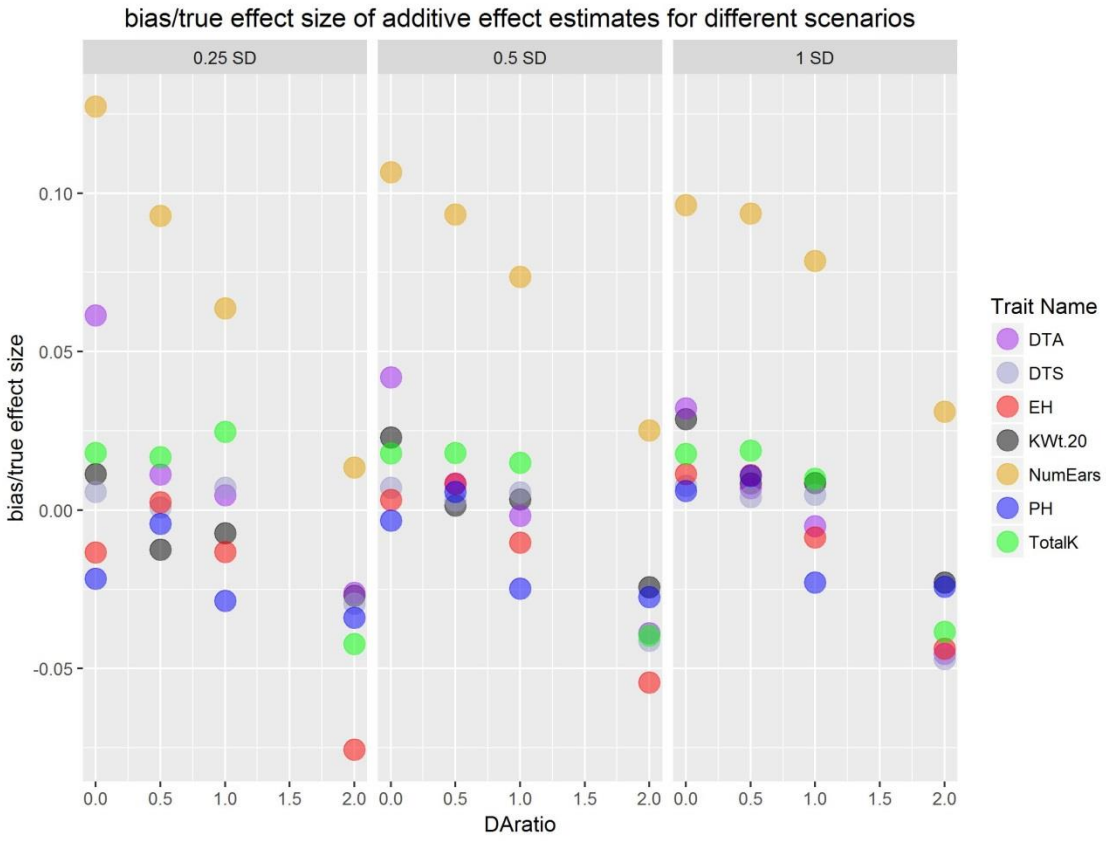

Figure 4.1. Bias of additive effect estimates under 12 different simulation scenarios. Bias was calculated as the mean of estimated values minus the true simulated effect at simulated QTL. Bias is reported as a proportion of the true causal effect. DAratio is ratio of the simulated dominance effect to the additive effect. 0.25 SD, 0.5 SD, 1 SD are simulated additive effect size and SD is the standard deviation of phenotypic value of the corresponding trait.

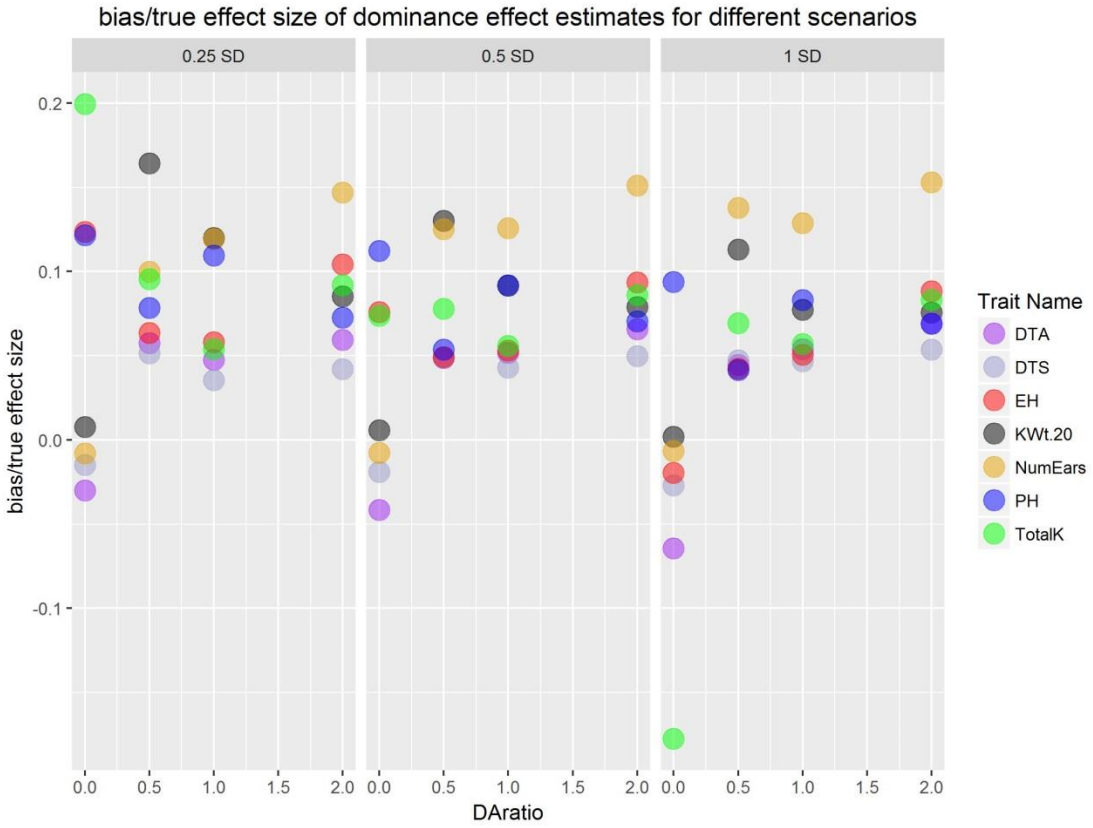

Figure 4.2. Bias of dominance effect estimates under 12 different simulation scenarios. Bias was calculated as the mean of estimated values minus the true simulated effect at simulated QTL. Bias is reported as a proportion of the true causal effect. DAratio is the ratio of the simulated dominance effect to additive effect. 0.25 SD, 0.5 SD, 1 SD are simulated additive effect size and SD is the standard deviation of phenotypic value of the corresponding trait.

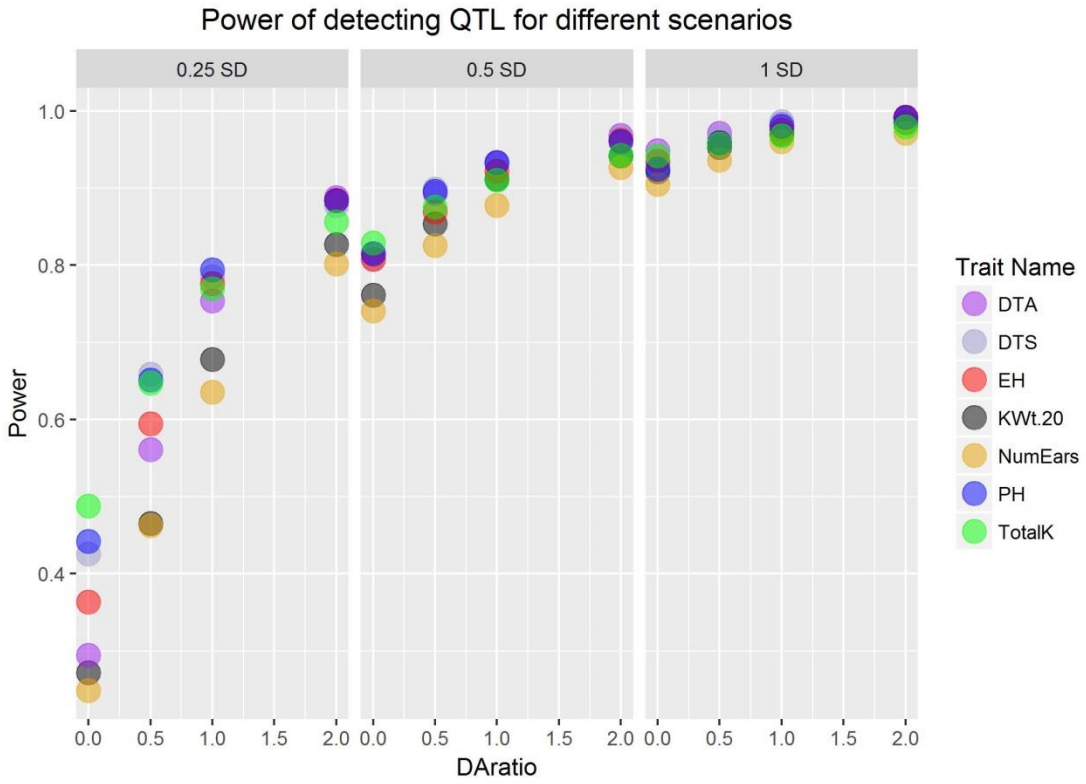

Figure 4.3. Power of detecting QTL under 12 different simulation scenarios. DAratio is the simulated dominance effect to additive effect ratio. 0.25 SD, 0.5 SD, 1 SD are simulated additive effect size and SD is the standard deviation of phenotypic value of the corresponding trait. Significance threshold 1.6×10^{-6} determined by bonferroni correction(0.05 divided by 30185).

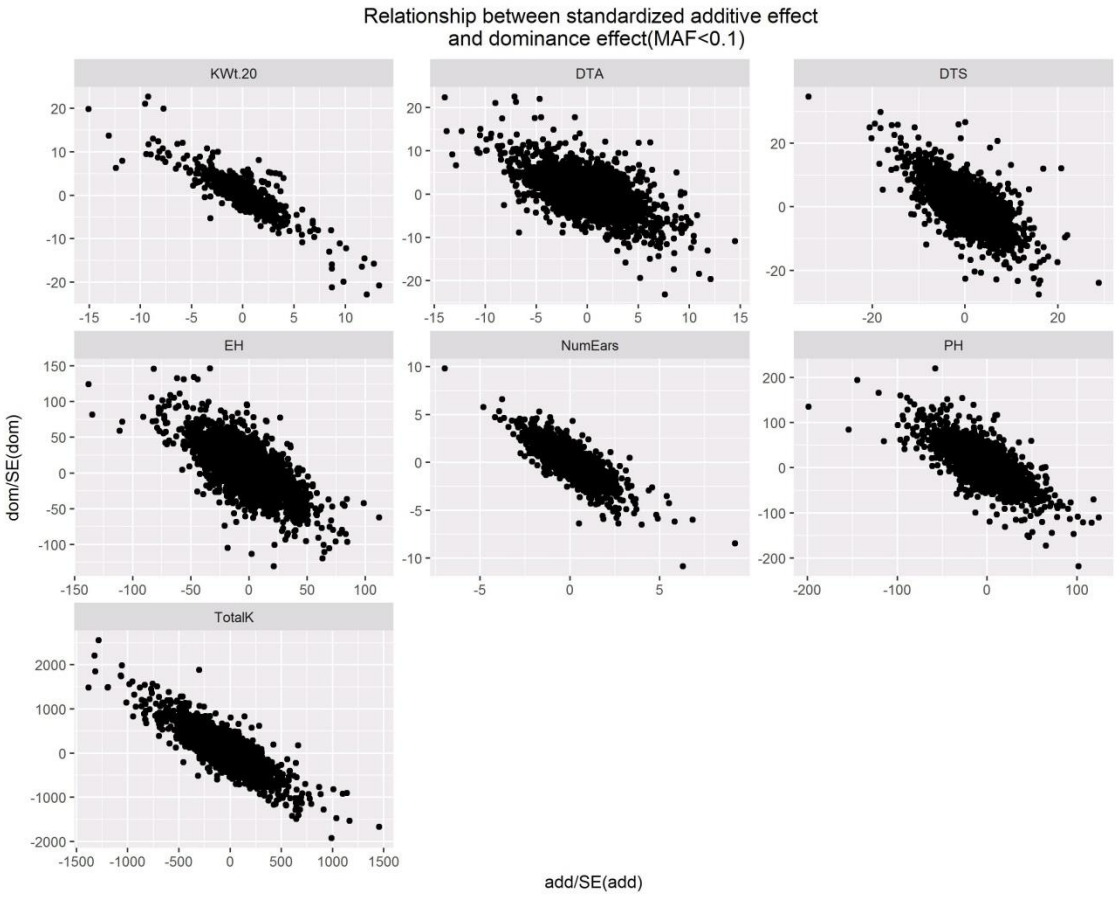

Figure 4.4. Relationship between standardized additive effect (X-axis) and standardized dominance effect (Y-axis) for SNPs with MAF below 0.1. Standardized genetic effect is genetic effect estimate divided by its corresponding standard error.

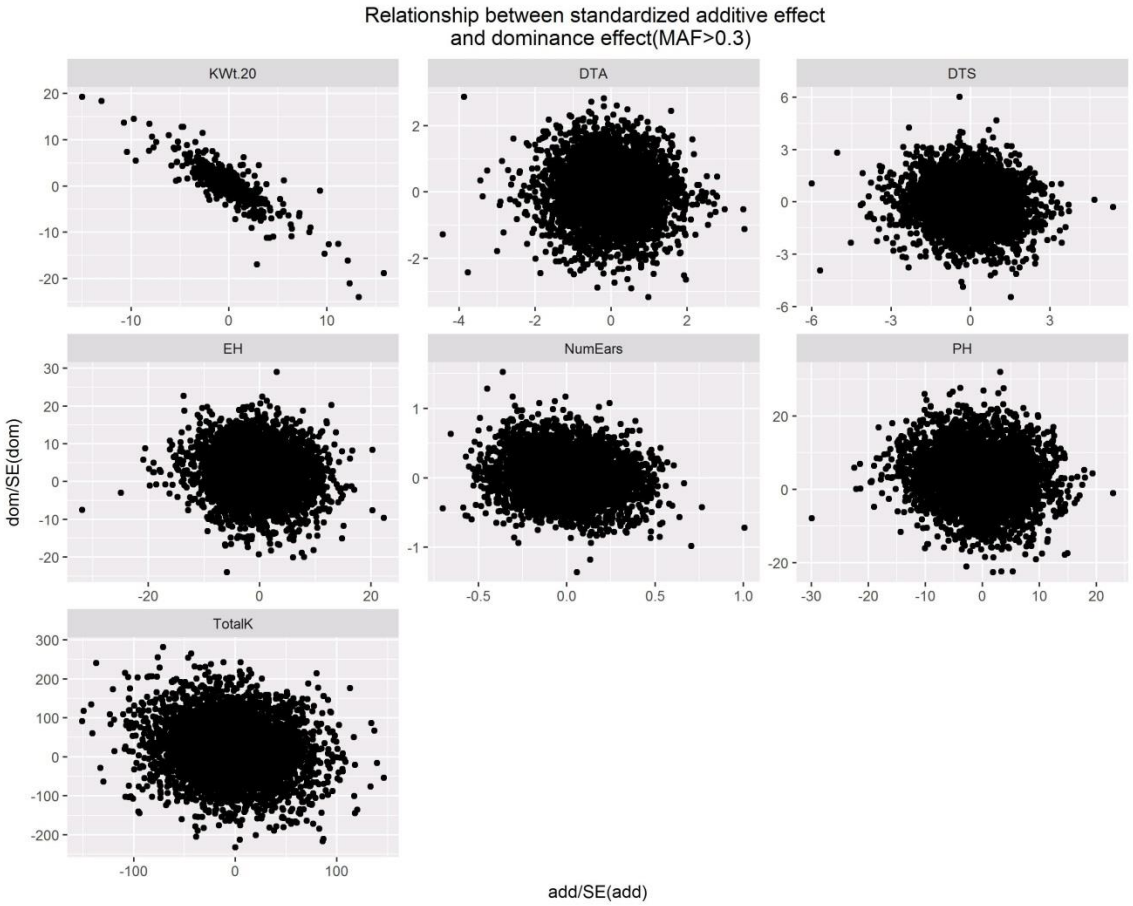

Figure 4.5. Relationship between standardized additive effect (X-axis) and standardized dominance effect (Y-axis) for SNPs with MAF above 0.3. Standardized genetic effect is genetic effect estimate divided by its corresponding standard error.

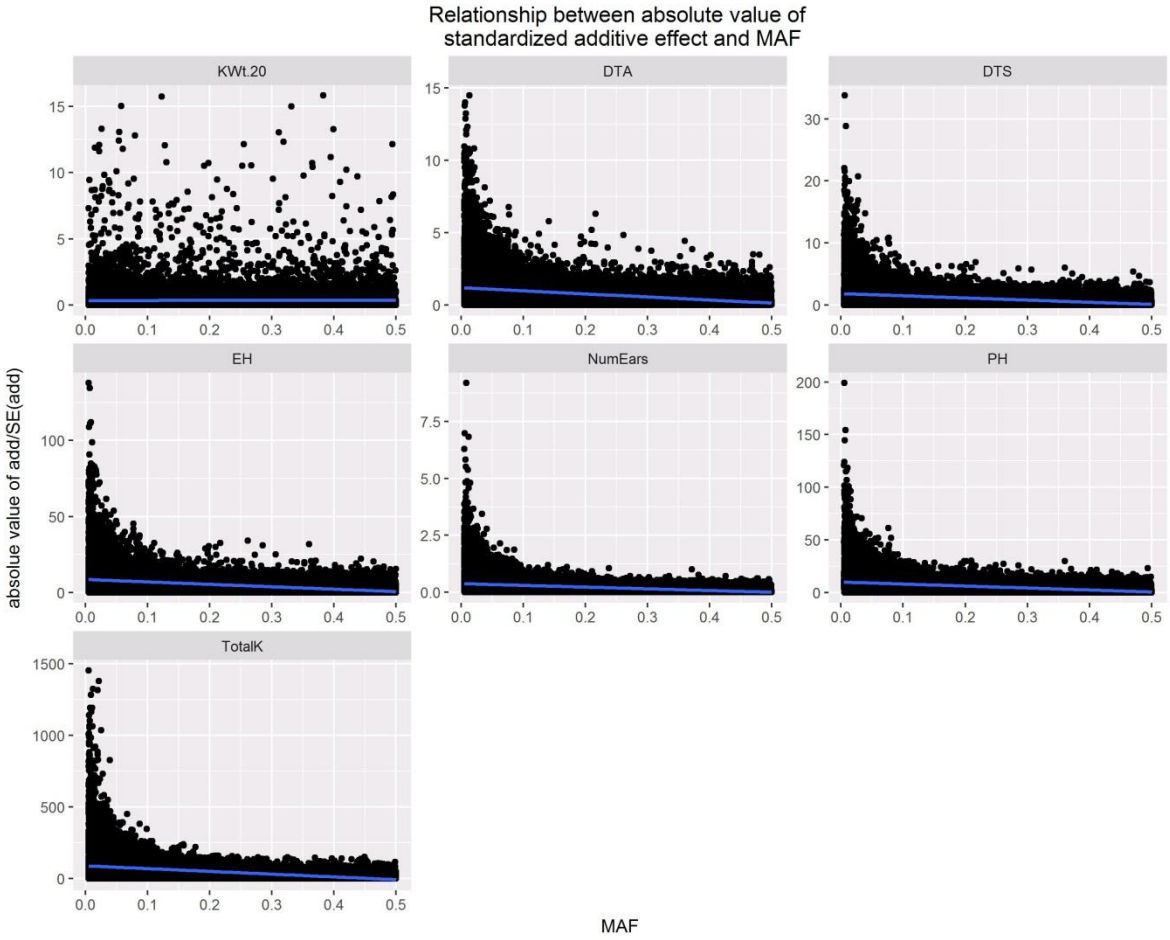

Figure 4.6. Relationship between absolute value of standardized additive effect (Y-axis) and MAF (X-axis)

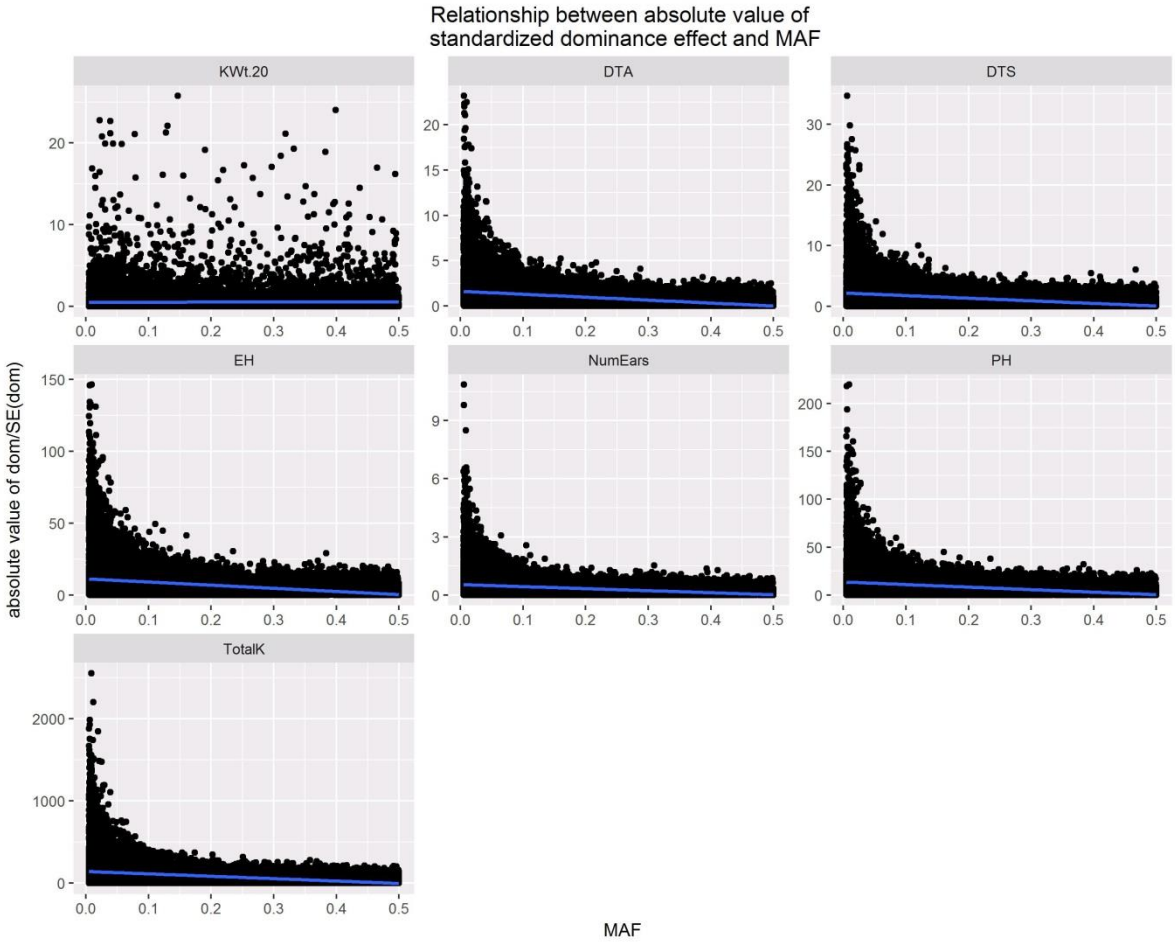

Figure 4.7. Relationship between absolute value of standardized dominance effect (Y-axis) and MAF (X-axis)

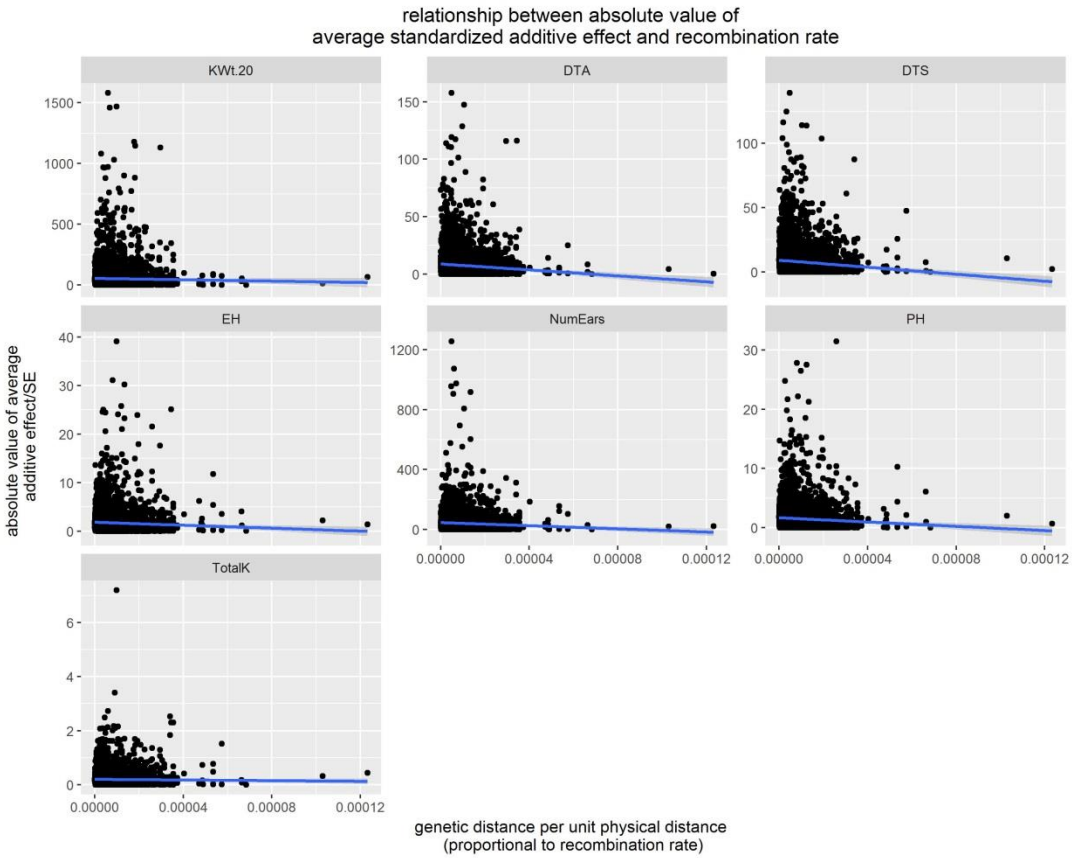

Figure 4.8. Relationship between absolute value of average standardized additive effect (Y-axis) and recombination rate (X-axis)

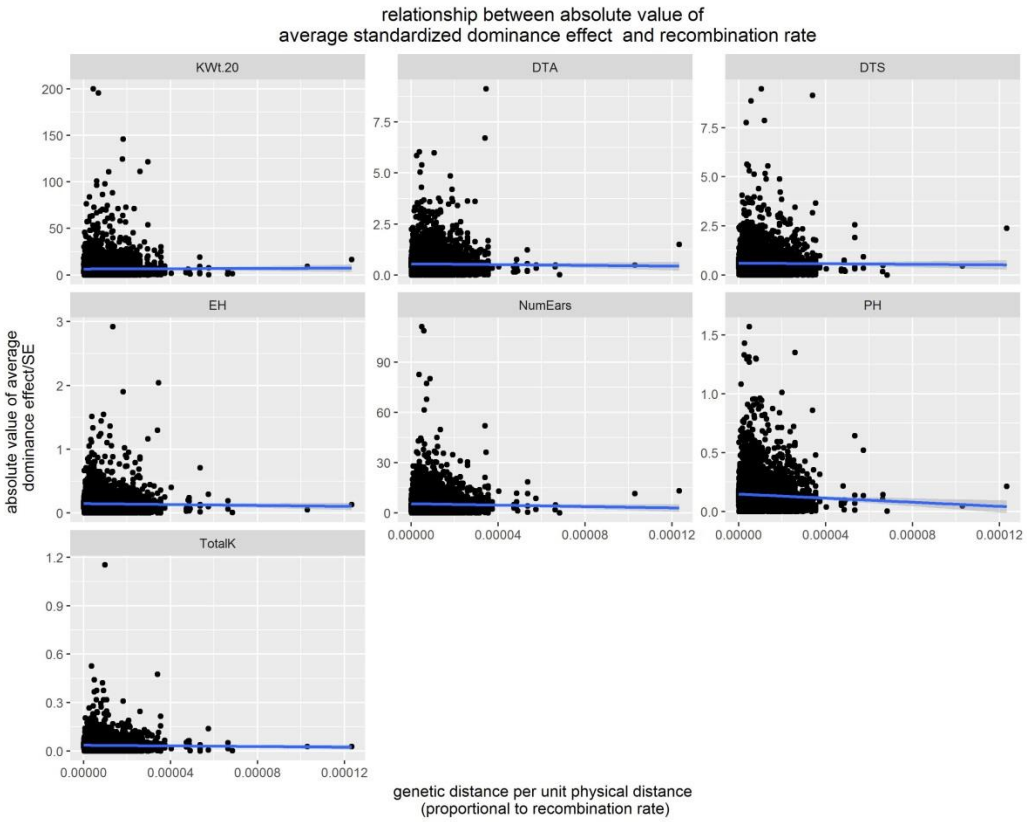

Figure 4.9. Relationship between absolute value of average standardized dominance effect (Y-axis) and recombination rate (X-axis)

References

Anderson, L. K., K. D. Hooker and S. M. Stack, 2001 The distribution of early recombination nodules on zygotene bivalents from plants. Genetics 159: 1259-1269.

Birchler, J. A., H. Yao and S. Chudalayandi, 2006 Unraveling the genetic basis of hybrid vigor. Proc. Natl. Acad. Sci. U. S. A. **103:** 12957-12958.

Birchler, J. A., H. Yao, S. Chudalayandi, D. Vaiman and R. A. Veitia, 2010 Heterosis. Plant Cell **22:** 2105-2112.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al*, 2007 TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics **23:** .

BUCKLER, E. S., J. M. THORNSBERRY and S. KRESOVICH, 2001 Molecular diversity, structure and domestication of grasses. Genet. Res. **77:** 213-218.

Charlesworth, D., and J. H. Willis, 2009 The genetics of inbreeding depression. Nature Reviews Genetics **10:** 783-796.

Colome-Tatche, M., S. Cortijo, R. Wardenaar, L. Morgado, B. Lahouze *et al*, 2012 Features of the arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proc. Natl. Acad. Sci. U. S. A. **109:** 16240-16245.

Crow, J. F., 2000 The rise and fall of overdominance. Plant Breeding Reviews, Volume 17 225-257.

Doebley, J., and A. Stec, 1993 Inheritance of the morphological differences between maize and teosinte: Comparison of results for two F2 populations. Genetics **134:** 559-570.

Doebley, J., A. Stec, J. Wendel and M. Edwards, 1990 Genetic and morphological analysis of a maize-teosinte F2 population: Implications for the origin of maize. Proc. Natl. Acad. Sci. U. S. A. **87:** 9888-9892.

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

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

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

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6:** .

Endelman, J. B., and J. L. Jannink, 2012 Shrinkage estimation of the realized relationship matrix. G3 (Bethesda) **2:** 1405-1413.

Fu, H., and H. K. Dooner, 2002 Intraspecific violation of genetic colinearity and its implications in maize. Proc. Natl. Acad. Sci. U. S. A. **99:** 9573-9578.

Gardner, C., and J. Lonnquist, 1959 Linkage and the degree of dominance of genes controlling quantitative characters in maize. Agron. J. **51:** 524-528.

Gaut, B. S., S. I. Wright, C. Rizzon, J. Dvorak and L. K. Anderson, 2007 Recombination: An underappreciated factor in the evolution of plant genomes. Nature Reviews Genetics **8:** 77-84.

Gerke, J. P., J. W. Edwards, K. E. Guill, J. Ross-Ibarra and M. D. McMullen, 2015 The genomic impacts of drift and selection for hybrid performance in maize. Genetics .

Gilmour, A. R., B. R. Cullis and A. P. Verbyla, 1997 Accounting for natural and extraneous variation in the analysis of field experiments. Journal of Agricultural, Biological, and Environmental Statistics 269-293.

Gilmour, A. R., B. J. Gogel, B. R. Cullis and R. Thompson, 2009 *ASReml User Guide Release 3.0*. VSN International, Ltd, Hemel Hempstead, UK.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al*, 2014 TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE **9:** 1-11.

Grossniklaus, U., G. A. Nogler and P. J. van Dijk, 2001 How to avoid sex: The genetic control of gametophytic apomixis. Plant Cell **13:** 1491-1498.

Hartfield, M., and S. P. Otto, 2011 RECOMBINATION AND HITCHHIKING OF DELETERIOUS ALLELES. Evolution **65:** 2421-2434.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269-294.

Hung, H., L. M. Shannon, F. Tian, P. J. Bradbury, C. Chen *et al*, 2012 ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proceedings of the National Academy of Sciences **109:** E1913-E1921.

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin *et al*, 2004 Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. **14:** 528-538.

Lewontin, R. C., 1974 The Genetic Basis of Evolutionary Change. Columbia University Press New York.

Liu, Z., J. Cook, S. Melia-Hancock, K. Guill, C. Bottoms *et al*, 2016 Expanding maize genetic resources with predomestication alleles: Maize–teosinte introgression populations. The Plant Genome **9:** .

Marth, G. T., E. Czabarka, J. Murvai and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166:** 351-372.

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, J. Sanchez G., E. Buckler *et al*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proceedings of the National Academy of Sciences **99:** 6080-6084.

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li *et al*, 2009a Genetic properties of the maize nested association mapping population. Science **325:** .

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li *et al*, 2009b Genetic properties of the maize nested association mapping population. Science **325:** 737-740.

Mezmouk, S., and J. Ross-Ibarra, 2013 The pattern and distribution of deleterious mutations in maize. G3: Genes|Genomes|Genetics **4:** 163-171.

Mirouze, M., M. Lieberman-Lazarovich, R. Aversano, E. Bucher, J. Nicolet *et al*, 2012 Loss of DNA methylation affects the recombination landscape in arabidopsis. Proc. Natl. Acad. Sci. U. S. A. **109:** 5880-5885.

Moll, R. H., M. F. Lindsey and H. F. Robinson, 1963 Estimates of genetic variances and level of dominance in maize. Genetics **49:** 411-423.

Ogut, F., Y. Bian, P. J. Bradbury and J. B. Holland, 2015 Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. Heredity **114:** 552-563.

Rodgers-Melnick, E., P. J. Bradbury, R. J. Elshire, J. C. Glaubitz, C. B. Acharya *et al*, 2015 Recombination in diverse maize is stable, predictable, and associated with genetic load. Proceedings of the National Academy of Sciences **112:** 3823-3828.

Serres-Giardi, L., K. Belkhir, J. David and S. Glemin, 2012 Patterns and evolution of nucleotide landscapes in seed plants. Plant Cell **24:** 1379-1397.

Slotkin, R. K., and R. Martienssen, 2007 Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics **8:** 272-285.

Swarts, K., H. Li, J. A. R. Navarro, D. An, M. C. Romay *et al*, 2014 Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome **7:** .

Tenaillon, M. I., J. U'Ren, O. Tenaillon and B. S. Gaut, 2004 Selection versus demography: A multilocus investigation of the domestication process in maize. Mol. Biol. Evol. **21:** 1214-1225.

Yelina, N. E., K. Choi, L. Chelysheva, M. Macaulay, B. De Snoo *et al*, 2012 Epigenetic remodeling of meiotic crossover frequency in arabidopsis thaliana DNA methyltransferase mutants. PLoS Genet **8:** e1002844.

Zhang, L., A. S. Peek, D. Dunams and B. S. Gaut, 2002 Population genetics of duplicated disease-defense genes, hm1 and hm2, in maize (zea mays ssp. mays L.) and its wild ancestor (zea mays ssp. parviglumis). Genetics **162:** 851-860.

APPENDICES

APPENDIX A: Supplemental Material for Chapter 2

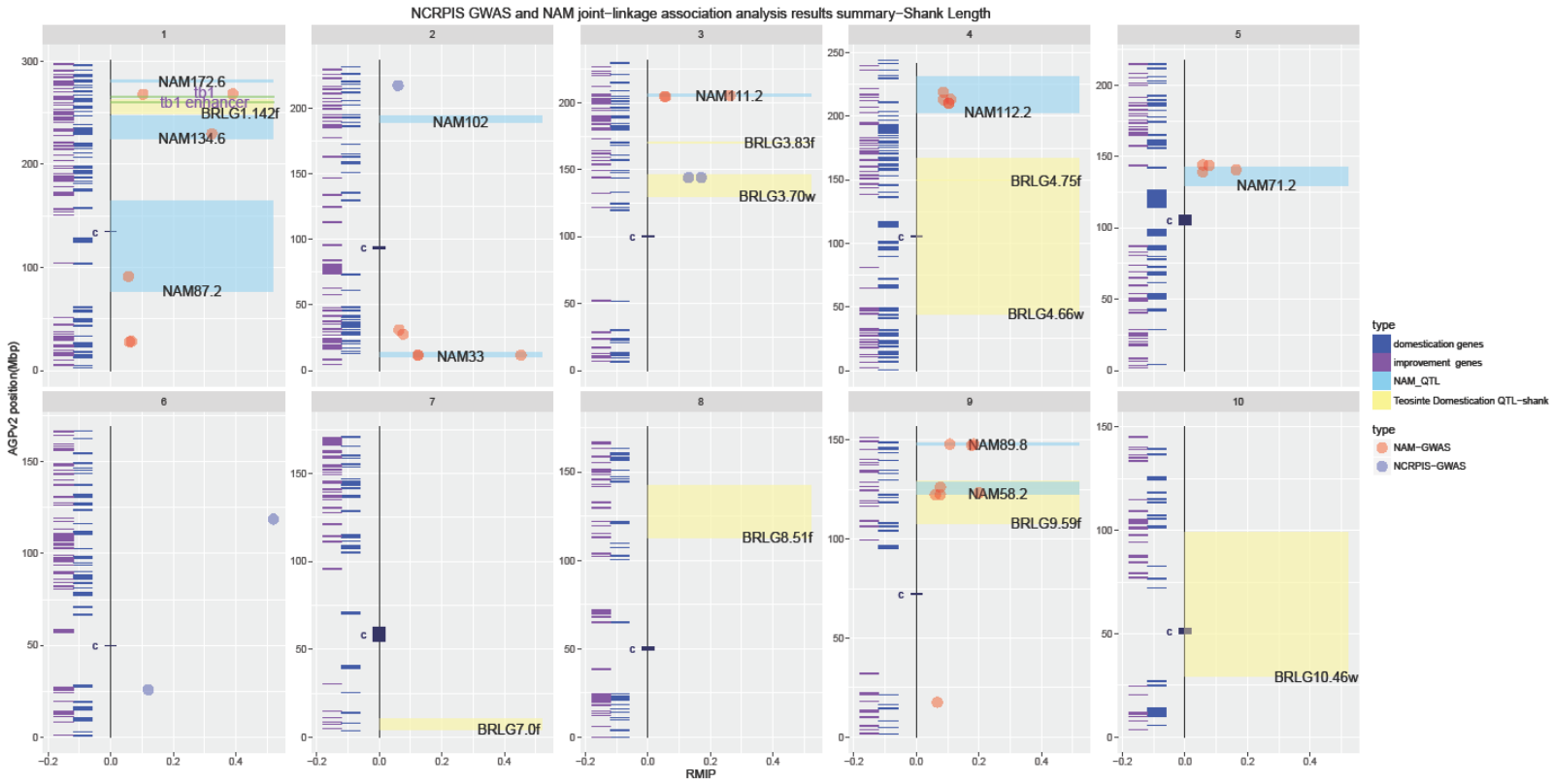
NCRPIS GWAS and NAM joint−linkage association analysis results summary−Shank Length
AGPv2 position(Mbp)
RMIP
NAM172.6
tb1
tb1 enhancer
BRLG1.142f
NAM134.6
NAM87.2
NAM102
NAM33
NAM111.2
BRLG3.83f
BRLG3.70w
NAM112.2
BRLG4.75f
BRLG4.66w
NAM71.2
BRLG7.0f
BRLG8.51f
NAM89.8
NAM58.2
BRLG9.59f
BRLG10.46w
type
domestication genes
improvement genes
NAM_QTL
Teosinte Domestication QTL−shank
type
NAM−GWAS
NCRPIS−GWAS

Figure A.1. Joint-linkage QTL mapping and GWAS results summary for shank length from NCRPIS panel and NAM population. Ten chromosome pairs of maize are displayed, physical position (AGPv2) in Mbp on vertical axis, and the resample model inclusion probability (RMIP) for SNP associations on horizontal axes (red dots are associations in NAM population, blue dots are associations in NCRPIS panel). In both panels, SNP associations with p value $< 10^{-7}$ within a subsample were declared significant. Blue boxes represent NAM QTL intervals, yellow boxes represent QTL mapped for analogous trait in maize-teosinte populations previously. Blue and purple boxes to left of chromosomes represent intervals identified as showing signals of selection during domestication (blue) or post-domestication improvement (purple). Known domestication genes are indicated with green bars. Centromeres are identified by 'C'.

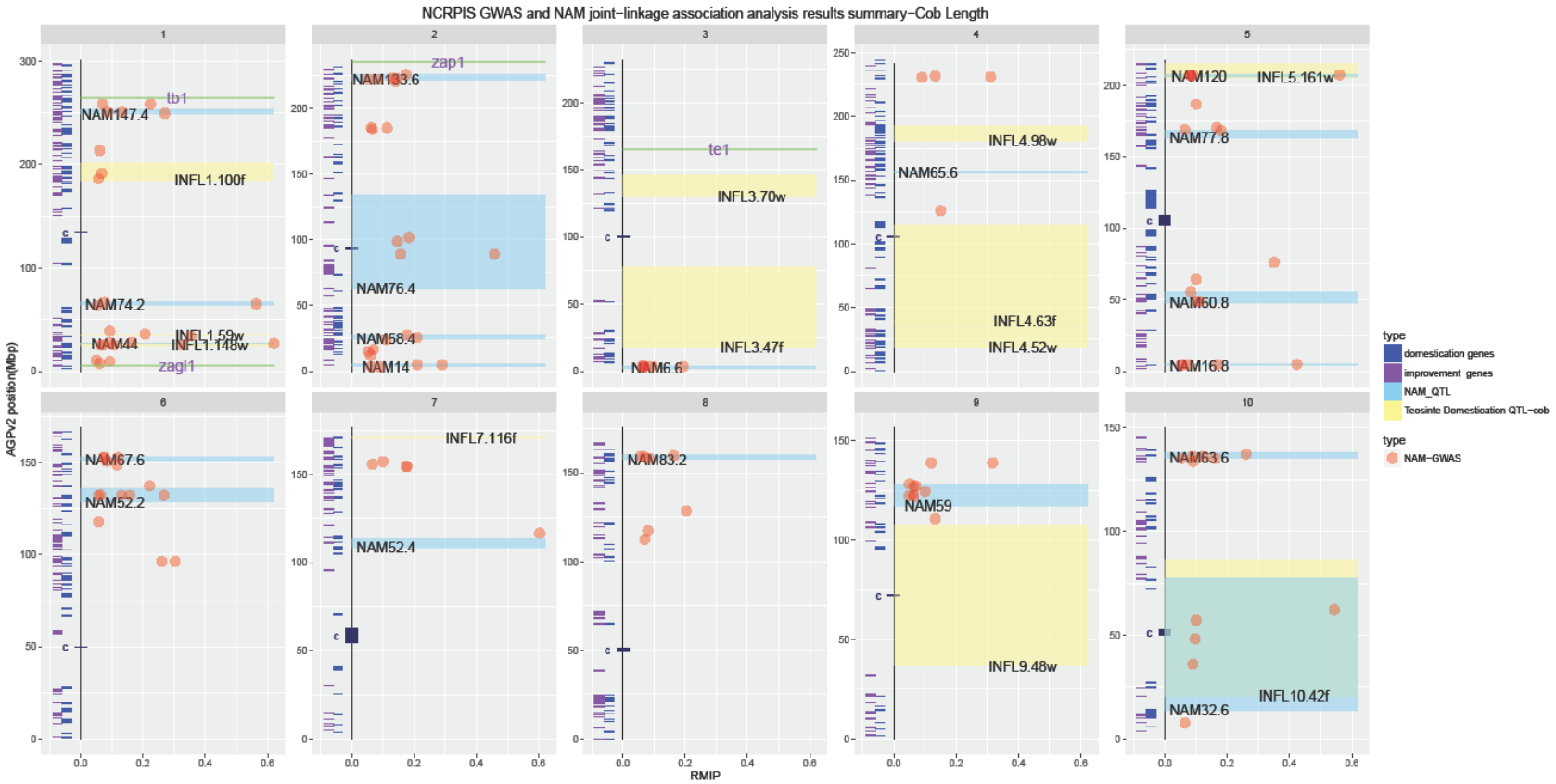
NCRPIS GWAS and NAM joint−linkage association analysis results summary−Cob Length
AGPv2 position(Mbp)
RMIP
tb1
NAM147.4
INFL1.100f
NAM74.2
NAM44
INFL1.59w
INFL1.148w
zagl1
zap1
NAM133.6
NAM76.4
NAM58.4
NAM14
te1
INFL3.70w
INFL3.47f
NAM6.6
INFL4.98w
NAM65.6
INFL4.63f
INFL4.52w
NAM120
INFL5.161w
NAM77.8
NAM60.8
NAM16.8
NAM67.6
NAM52.2
INFL7.116f
NAM52.4
NAM83.2
NAM59
INFL9.48w
NAM63.6
INFL10.42f
NAM32.6
type
domestication genes
improvement genes
NAM_QTL
Teosinte Domestication QTL−cob
type
NAM−GWAS

Figure A.2. Joint-linkage QTL mapping and GWAS results summary for cob length from NCRPIS panel and NAM population. Ten chromosome pairs of maize are displayed, physical position (AGPv2) in Mbp on vertical axis, and the resample model inclusion probability (RMIP) for SNP associations on horizontal axes (red dots are associations in NAM population, blue dots are associations in NCRPIS panel). In both panels, SNP associations with p value $< 10^{-7}$ within a subsample were declared significant. Blue boxes represent NAM QTL intervals, yellow boxes represent QTL mapped for analogous trait in maize-teosinte populations previously. Blue and purple boxes to left of chromosomes represent intervals identified as showing signals of selection during domestication (blue) or post-domestication improvement (purple). Known domestication genes are indicated with green bars. Centromeres are identified by 'C'

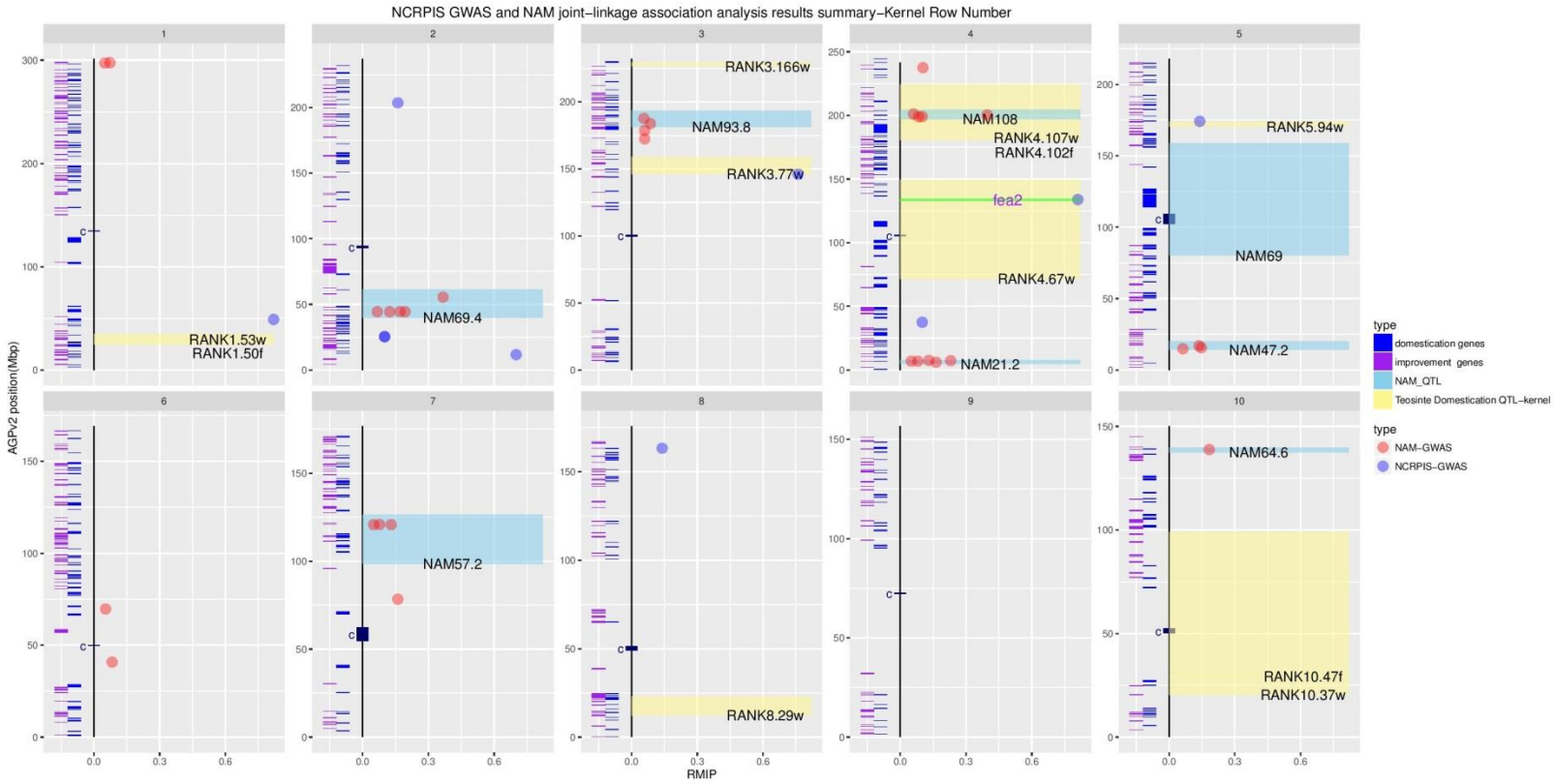
NCRPIS GWAS and NAM joint−linkage association analysis results summary−Kernel Row Number
AGPv2 position(Mbp)
RMIP
RANK1.53w
RANK1.50f
NAM69.4
RANK3.166w
NAM93.8
RANK3.77w
NAM108
RANK4.107w
RANK4.102f
fea2
RANK4.67w
NAM21.2
RANK5.94w
NAM69
NAM47.2
NAM57.2
RANK8.29w
NAM64.6
RANK10.47f
RANK10.37w
type
domestication genes
improvement genes
NAM_QTL
Teosinte Domestication QTL−kernel
type
NAM−GWAS
NCRPIS−GWAS

Figure A.3. Joint-linkage QTL mapping and GWAS results summary for kernel row number from NCRPIS panel and NAM population. Ten chromosome pairs of maize are displayed, physical position (AGPv2) in Mbp on vertical axis, and the resample model inclusion probability (RMIP) for SNP associations on horizontal axes (red dots are associations in NAM population, blue dots are associations in NCRPIS panel). In both panels, SNP associations with p value $< 10^{-7}$ within a subsample were declared significant. Blue boxes represent NAM QTL intervals, yellow boxes represent QTL mapped for analogous trait in maize-teosinte populations previously. Blue and purple boxes to left of chromosomes represent intervals identified as showing signals of selection during domestication (blue) or post-domestication improvement (purple). Known domestication genes are indicated with green bars. Centromeres are identified by 'C'.

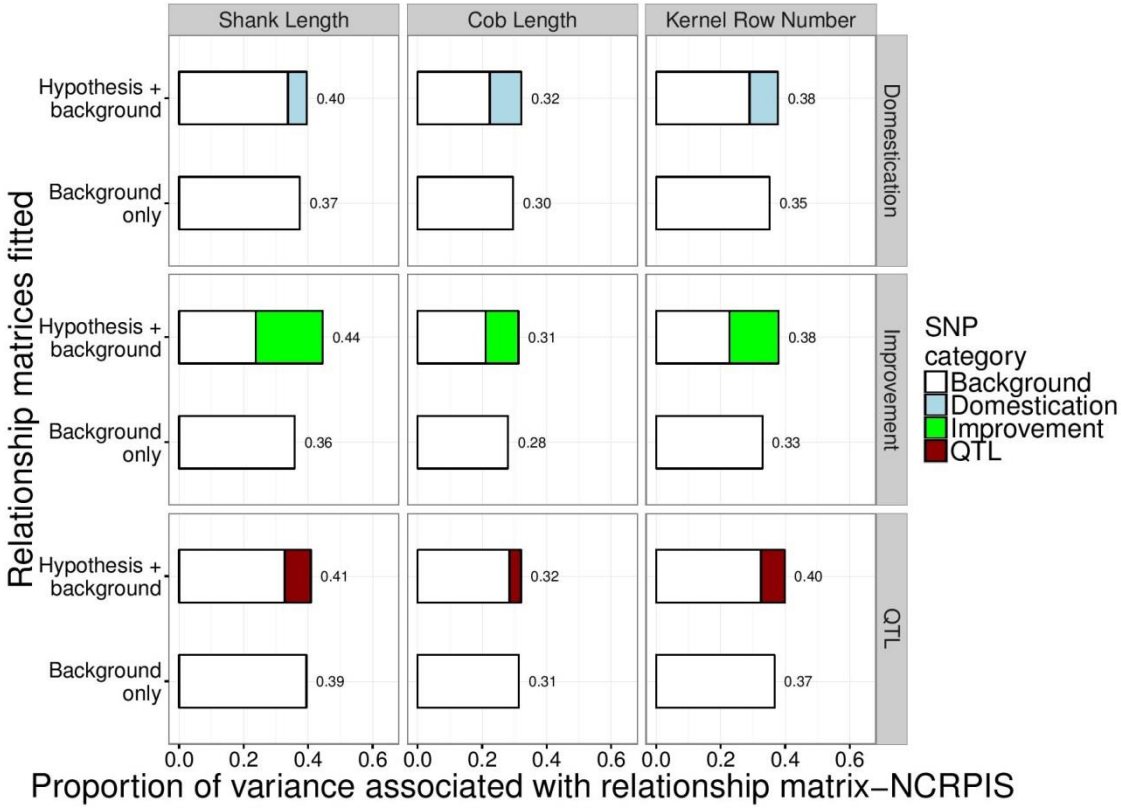

Figure A.4. The proportion of variance among inbred lines of the NCRPIS panel for shank length, cob length, and kernel row number associated with relationship matrices based on SNPs in hypothesis-defined regions or based on a sample of background SNPs with matching number of SNPs and proportion of coding SNPs.

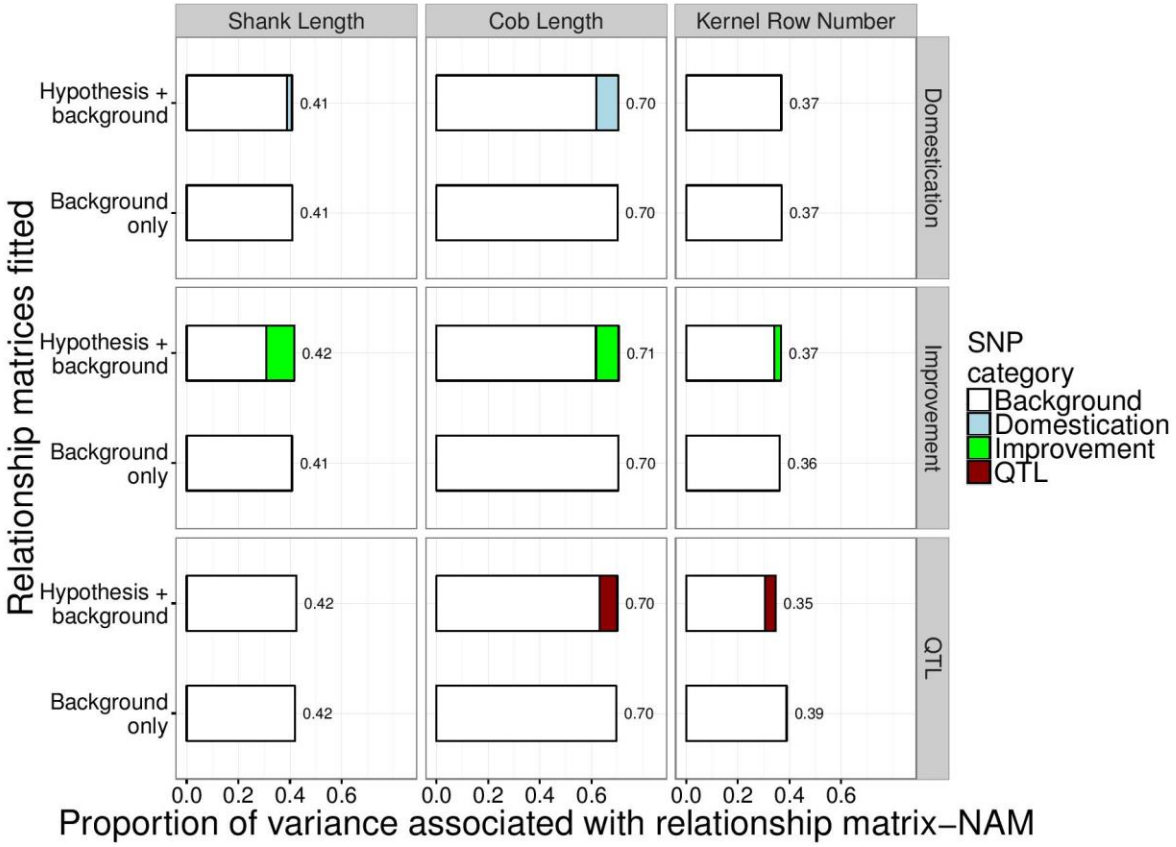

Figure A.5 The proportion of variance among inbred lines of the NAM panel for shank length, cob length, and kernel row number associated with relationship matrices based on SNPs in hypothesis-defined regions or based on a sample of background SNPs with matching number of SNPs and proportion of coding SNPs.

Table A.1 AGPv2 positions of intervals containing teosinte domestication QTL mapped in a maize × teosinte population by Briggs et al (2007).

Trait		Candidate QTL	Chr	Start	End
Shank Length	Length of the primary lateral branch	BRLG1.142f	1	248,738,069	262,720,351
		BRLG3.70w	3	129,976,755	146,252,963
		BRLG3.83f	3	170,110,525	170,365,685
		BRLG4.66w	4	44,508,376	149,276,909
		BRLG4.75f	4	149,276,909	167,190,374
		BRLG7.0f	7	4,238,741	10,620,195
		BRLG8.51f	8	112,677,295	142323391
		BRLG9.59f	9	129,409,633	107,625,029
		BRLG10.46w	10	29,222,679	98,999,036
Cob Length	Length of the primary lateral inflorescence	INFL1.100f	1	185016727	201,486,994
		INFL1.148w	1	24,941,000	26,671,972
		INFL1.59w	1	34,673,066	35,579,277
		INFL3.47f	3	17,430,424	77,678,189
		INFL3.70w	3	129,976,755	146,252,963
		INFL4.52w	4	18,682,264	39,092,497
		INFL4.63f	4	39,092,497	114,693,724
		INFL4.98w	4	180,785,427	192,205,587
		INFL5.161w	5	205440973	215,340,163
		INFL7.116f	7	170,656,427	171,232,753
		INFL9.48w	9	36,546,991	107,625,029
		INFL10.42f	10	20,409,576	86,424,870
Kernel Row Number	Number of internode columns (ranks) on the primary lateral inflorescence	RANK1.50f	1	24,941,000	26,671,972
		RANK1.53w	1	26,671,972	34,673,066
		RANK3.166w	3	226,481,779	229,251,792
		RANK3.77w	3	146,252,963	158,982,309
		RANK4.102f	4	180,785,427	192,205,587
		RANK4.107w	4	192,205,587	224245251
		RANK4.67w	4	71,640,363	149,276,909
		RANK5.94w	5	170,416,093	173,805,422
		RANK8.29w	8	12,285,916	22,978,660
		RANK10.37w	10	20,409,576	29,222,679
		RANK10.47f	10	29,222,679	98,999,036

Table A.2. Variance components associated with additive and non-additive polygenic effects in NCRPIS panel, the proportion of genetic variance explained by additive effects, and heritability due to additive and total genetic effects.

Trait name	$(\hat{\sigma}^2_{non\text{-}add})$[a]	$(\hat{\sigma}^2_{add})$[b]	Additive proportion of genotypic variance[c]	Additive heritability[d]	Total heritability[e]
SL	1.27E-03	2.40E-02	95%	57%	60%
CL	2.99E-05	3.76E+02	100%	40%	40%
KRN	0.193251	2.22E+00	92%	55%	60%

a. Variance component of genetic effect that is not explained by realized additive relationship matrix.

b. Variance component of genetic effect that is explained by realized additive relationship matrix.

c. Proportion is calculated as $(\hat{\sigma}^2_{add})/((\hat{\sigma}^2_{add}) + (\hat{\sigma}^2_{non\text{-}add}))$

d. Additive heritability is calculated as $(\hat{\sigma}^2_{add})/((\hat{\sigma}^2_{add}) + (\hat{\sigma}^2_{non\text{-}add})+ (\hat{\sigma}^2_{error}))$

e. Total heritability is calculated as $((\hat{\sigma}^2_{add})+ (\hat{\sigma}^2_{non\text{-}add}))/((\hat{\sigma}^2_{add}) + (\hat{\sigma}^2_{non\text{-}add})+ (\hat{\sigma}^2_{error}))$

Table A.3. Proportion of phenotypic variance (heritability, $\hat{h}^2_{Ai}$), proportion of total heritability, and proportion of genome associated with hypothesis-defined and background genetic relationship matrices in NCRPIS population. Relationship matrices were defined either using all SNPs and fitting all hypothesis matrices simultaneously, or by fitting one hypothesis-defined matrix at a time with a background relationship matrix defined using an equal number of SNPs.

	Shank Length						Cob Length						Kernel Row Number					
Category	All SNPs[a]			Matched backgrounds[b]			All SNPs			Matched backgrounds			All SNPs			Matched backgrounds		
	$\hat{h}^2_{Ai}$	Prop. of $\hat{\sigma}^2_{A(T)}$ [c]	Prop. of genome [d]	QTL	Domes.	Impr.	$\hat{h}^2_{Ai}$	Prop. of $\hat{\sigma}^2_{A(T)}$ [c]	Prop. of genome	QTL	Domes.	Impr.	$\hat{h}^2_{Ai}$	Prop. of $\hat{\sigma}^2_{A(T)}$ [c]	Prop. of genome	QTL	Domes.	Impr.
QTL	0.03	0.05	0.12	0.08			0.01	0.02	0.17	0.04			0.00	0.00	0.10	0.07		
Domestication	0.03	0.05	0.07		0.06		0.00	0.00	0.07		0.10		0.03	0.06	0.07		0.09	
Improve	0.12	0.22	0.05			0.21	0.07	0.18	0.05			0.10	0.12	0.27	0.05			0.15
Background	0.37	0.68	0.78	0.33	0.34	0.24	0.31	0.80	0.74	0.28	0.22	0.21	0.30	0.67	0.80	0.33	0.29	0.23
Total	0.55			0.41	0.40	0.44	0.40			0.32	0.32	0.31	0.45			0.40	0.38	0.38
Background Only	0.53			0.39	0.37	0.36	0.40			0.31	0.30	0.28	0.46			0.37	0.35	0.33

[a] "All SNPs": all markers were partitioned into three hypothesis-defined groups and remaining background markers; variance components were estimated from fitting four relationship matrices simultaneously, with no subsampling.

[b] "Matched backgrounds": for each hypothesis testing region, we calculated one additive realized relationship matrix using all SNPs within regions identified by the hypothesis (e.g., domestication QTL, domestication sweep regions, or improvement sweep regions) and a second realized additive relationship matrix using disjoint background markers. The background relationship matrix was estimated from a subset background markers with the same proportion of coding region SNPs and the same total number of markers as the hypothesis-defined realized additive relationship matrix. The background marker set was resampled twenty times and the mean variance components estimates from fitting the hypothesis defined relationship matrix along with one of the resampled background matrices are reported.

[c] The proportion of total additive variance attributable to a particular hypothesis-defined relationship matrix is:

$$\frac{(mean\ of\ diag(G_{Hi}))\sigma^2_{A(Hi)}}{\sigma^2_{A(T)}}, \quad \sigma^2_{A(T)} = \sum_{i=1}^{h}\left(mean\ of\ diag(G_{Hi})\sigma^2_{A(Hi)}\right) + \left(mean\ of\ diag(G_B)\right)\sigma^2_{A(B)}.$$

[d] The proportion of genome physical sequence accounted for by intervals defining the hypothesis and background relationship matrix.

Table A.4. Proportion of phenotypic variance (heritability, $\hat{h}^2_{Ai}$), proportion of total heritability, and proportion of genome associated with hypothesis-defined and background genetic relationship matrices in NAM population. Relationship matrices were defined either using all SNPs and fitting all hypothesis matrices simultaneously, or by fitting one hypothesis-defined matrix at a time with a background relationship matrix defined using an equal number of SNPs.

	Shank Length						Cob Length						Kernel Row Number					
Category	All SNPs[a]			Matched backgrounds[b]			All SNPs			Matched backgrounds			All SNPs			Matched backgrounds		
	$\hat{h}^2_{Ai}$	Prop. of $\hat{\sigma}^2_{A(T)}$ [c]	Prop. of genome[d]	QTL	Domes.	Impr.	$\hat{h}^2_{Ai}$	Prop. of $\hat{\sigma}^2_{A(T)}$ [c]	Prop. of genome	QTL	Domes.	Impr.	$\hat{h}^2_{Ai}$	Prop. of $\hat{\sigma}^2_{A(T)}$ [c]	Prop. of genome	QTL	Domes.	Impr.
QTL	0.00	0.00	0.12	0.00			0.02	0.03	0.17	0.07			0.00	0.00	0.10	0.04		
Domestication	0.01	0.02	0.07		0.02		0.09	0.13	0.07		0.09		0.00	0.00	0.07		0.00	
Improve	0.10	0.23	0.05			0.11	0.10	0.13	0.05			0.09	0.00	0.00	0.05			0.03
Background	0.32	0.75	0.78	0.42	0.39	0.31	0.50	0.71	0.74	0.63	0.62	0.62	0.39	1.00	0.80	0.30	0.37	0.34
Total	0.43			0.42	0.41	0.42	0.71			0.70	0.70	0.71	0.39			0.35	0.37	0.37
Background Only	0.42			0.42	0.41	0.41	0.70			0.70	0.70	0.70	0.38			0.39	0.37	0.36

[a] "All SNPs": all markers were partitioned into three hypothesis-defined groups and remaining background markers; variance components were estimated from fitting four relationship matrices simultaneously, with no subsampling.

[b] "Matched backgrounds": for each hypothesis testing region, we calculated one additive realized relationship matrix using all SNPs within regions identified by the hypothesis (e.g., domestication QTL, domestication sweep regions, or improvement sweep regions) and a second realized additive relationship matrix using disjoint background markers. The background relationship matrix was estimated from a subset of background markers with the same proportion of coding region SNPs and the same total number of markers as the hypothesis-defined realized additive relationship matrix. The background marker set was resampled twenty times and the mean variance components estimates from fitting the hypothesis defined relationship matrix along with one of the resampled background matrices are reported.

[c] The proportion of total additive variance attributable to a particular hypothesis-defined relationship matrix is:

$$\frac{(mean\ of\ diag(G_{Hi}))\sigma^2_{A(Hi)}}{\sigma^2_{A(T)}}, \quad \sigma^2_{A(T)} = \sum_{i=1}^{h}\left(mean\ of\ diag(G_{Hi})\sigma^2_{A(Hi)}\right) + \left(mean\ of\ diag(G_B)\right)\sigma^2_{A(B)}.$$

[d] The proportion of genome physical sequence accounted for by intervals defining the hypothesis and background relationship matrices.

Table A.5. Cumulative proportion of genome tagged by SNPs defining hypothesis relationship matrices and background matrices, and the proportion of total additive genetic variation associated with each relationship matrix.

		Total size of genomic regions (Mbp)	Proportion of genome	Proportion of additive variance in NCRPIS panel			Proportion of additive variance in NAM population		
				SL	CL	KRN	SL	CL	KRN
QTL	SL	251.4	0.12	0.05	-	-	0.00	-	-
	CL	349.8	0.17	-	0.02	-	-	0.03	-
	KRN	207.1	0.10	-	-	0.00	-	-	0.00
Domestication sweep		151.6	0.07	0.05	0.00	0.06	0.02	0.13	0.00
Improvement sweep		98.7	0.05	0.22	0.18	0.26	0.23	0.13	0.00
Background		1518.6 - 1650.8	0.78 - 0.80	0.68	0.80	0.67	0.75	0.71	1.00

Files A.1-A.15

Supporting data

Available for download at:

http://www.genetics.org/content/early/2016/07/11/genetics.116.191106.supplemental

File A.1. QTL mapped in NAM population. The affected trait, chromosome, start and end positions (AGPv2 bp) and length (bp) of support intervals, and proportion of variation among line means associated with the QTL are presented.

File A.2. Variants significantly ($p < 10^{-7}$) associated with domestication traits in at least 5% of data subsamples (RMIP > 0.05) of the NAM population. The affected trait, chromosome, AGPv2 position, and segregating alleles are presented for each SNP. '+/-'segregating alleles refer to single base indels and 'CNV-' refers to copy number variants inferred from read depth variation. The effect estimatse and *p*-values are averaged over data subsamples in which the SNP was included in the model. For intragenic variants, the gene model and annotation information for the gene containing the variant is presented. If the SNP is not in a coding region ('intergenic'), annotation information for the nearest gene is presented. Segregation ratio is the proportion of NAM families in which the variant segregates.

File A.3. SNPs significantly ($p < 10^{-7}$) associated with domestication traits in at least 5% of data subsamples (RMIP > 0.05) of the NCRPIS panel. The affected trait, chromosome, AGPv2 position, and segregating alleles are presented for each SNP. '+/-' segregating alleles refer to single base indels and 'CNV-' refers to copy number variants inferred from read depth variation. The r^2 and *p*-value for each marker is based on their association in the full data set. For intragenic variants, the gene model and annotation information for the gene containing the variant is presented. If the SNP is not in a coding region ('intergenic'), annotation information for the nearest gene is presented.

File A.4. Haplotypes in *gt1* region represented by at least 5 lines in NCRPIS population and their estimated effects on shank length.

File A.5. Haplotypes in *zap1* gene regions represented by at least 5 lines in NCRPIS population and their estimated effects on cob length.

File A.6 Realized additive relationship matrix for 2480 inbred lines NCRPIS panel based on a subset of 111,282 SNPs selected from Romay et al (2013) data. (.zip, 51.11 MB)
File A.7 Best linear unbiased estimates (BLUEs) for shank length of NCRPIS lines. (.zip, 13 KB)
File A.8 Best linear unbiased estimates (BLUEs) for cob length of NCRPIS lines. (.zip, 13 KB)

File A.9 Best linear unbiased estimates (BLUEs) for kernel row number of NCRPIS lines. (.zip, 12 KB)

File A.10 Best linear unbiased estimates (BLUEs) for shank length of NAM lines. (.csv, 70 KB)

File A.11 Best linear unbiased estimates (BLUEs) for cob length of NAM lines.(.csv, 88 KB)

File A.12 Best linear unbiased estimates (BLUEs) for kernel row number of NAM lines. (.csv, 66 KB)

File A.13 Chromosome-specific residuals for shank length of NAM lines, adjusted for the effects of QTL off the target chromosome, used for genome-wide association study. (.csv, 399 KB)

File A.14 Chromosome-specific residuals for cob length of NAM lines, adjusted for the effects of QTL off the target chromosome, used for genome-wide association study. (.csv, 673 KB)

File A.15 Chromosome-specific residuals for kernel row number of NAM lines, adjusted for the effects of QTL off the target chromosome, used for genome-wide association study.

APPENDIX B: Supplemental Material for Chapter 3

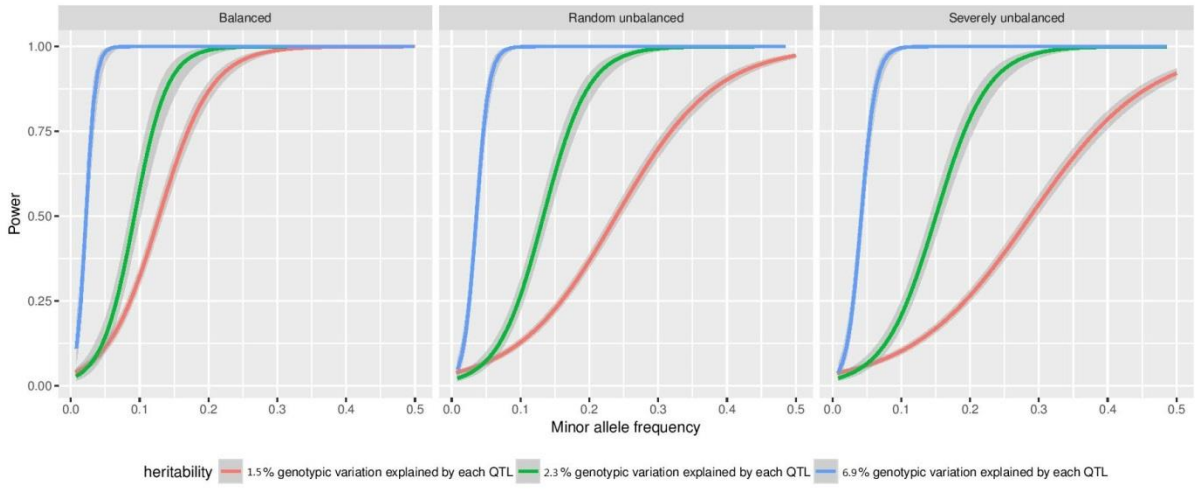

Figure B.1. Power (Y-axis) and minor allele frequency (X-axis) of single-stage association tests for three different genetic architectures and three levels of data imbalance. Curves represent loess estimates of the mean and 95% confidence interval of power at each minor allele frequency.

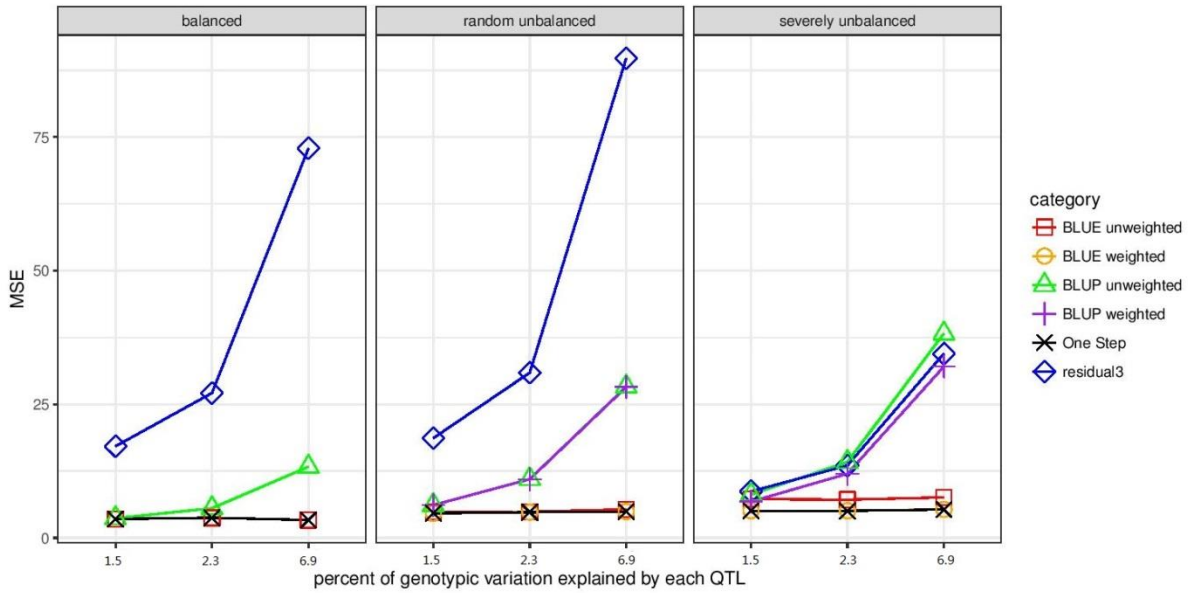

Figure B.2. Mean squared error of QTL effect estimates for different GWAS methods under three different genetic architectures and three levels of data imbalance.

Files B.1-B.3

Supporting data

Available for download at:

https://drive.google.com/drive/folders/0By4mQNWXnpaacmwyNzZyclF1TFE

File B.1. Power of six different GWAS methods to detect causal variants under three different genetic architectures and three levels of data imbalance.

File B.2. False discovery rate of six different GWAS methods under three different genetic architectures and three levels of data imbalance. Two different thresholds for declaring false positives are reported.

File B.3. Bias and mean squared errors of causal loci effect estimates for six different GWAS methods under three different genetic architectures and three levels of data imbalance.

File B.4. Example of command line execution

APPENDIX C: Supplemental Material for Chapter 4

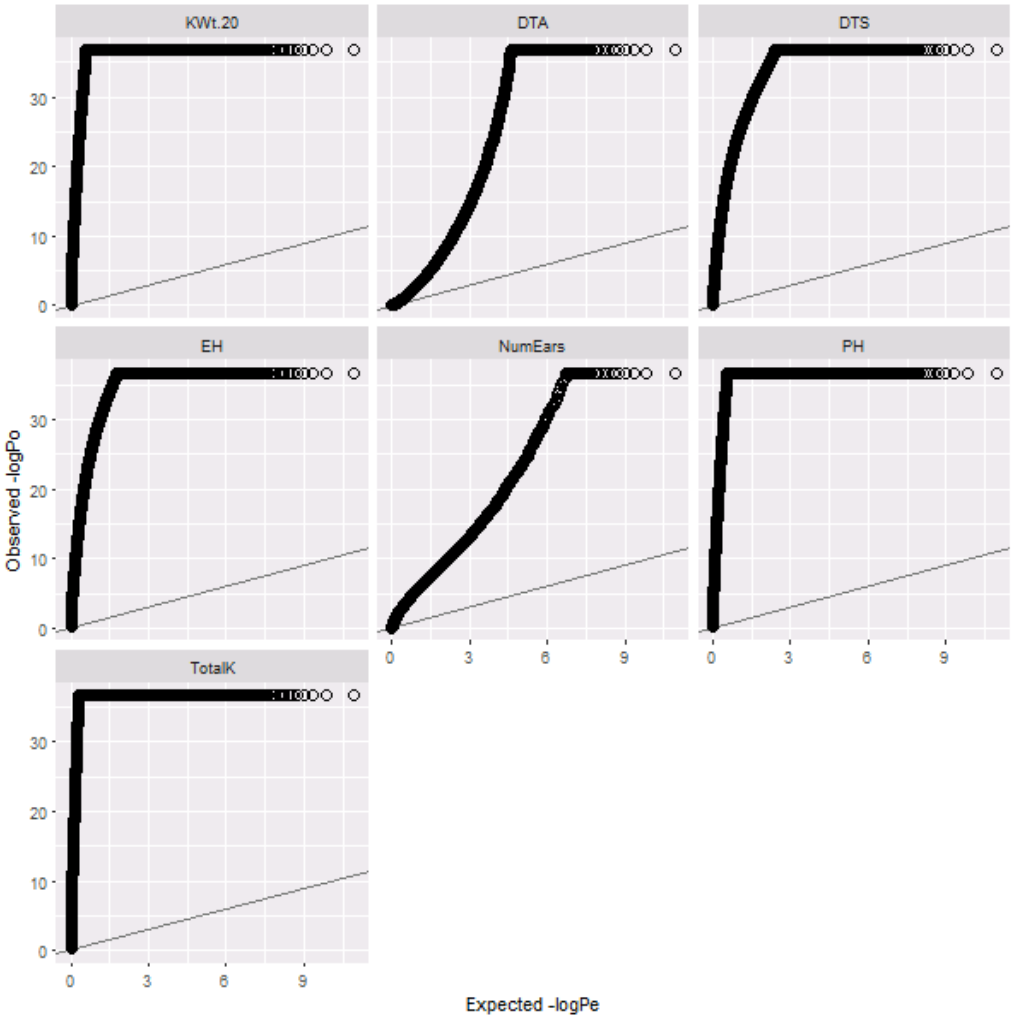

Figure C.1. Q-Q plot of P values for model 1 for all traits. Y-axis indicates observed –logP values. X-axis indicates expected –logP values from uniform distribution.

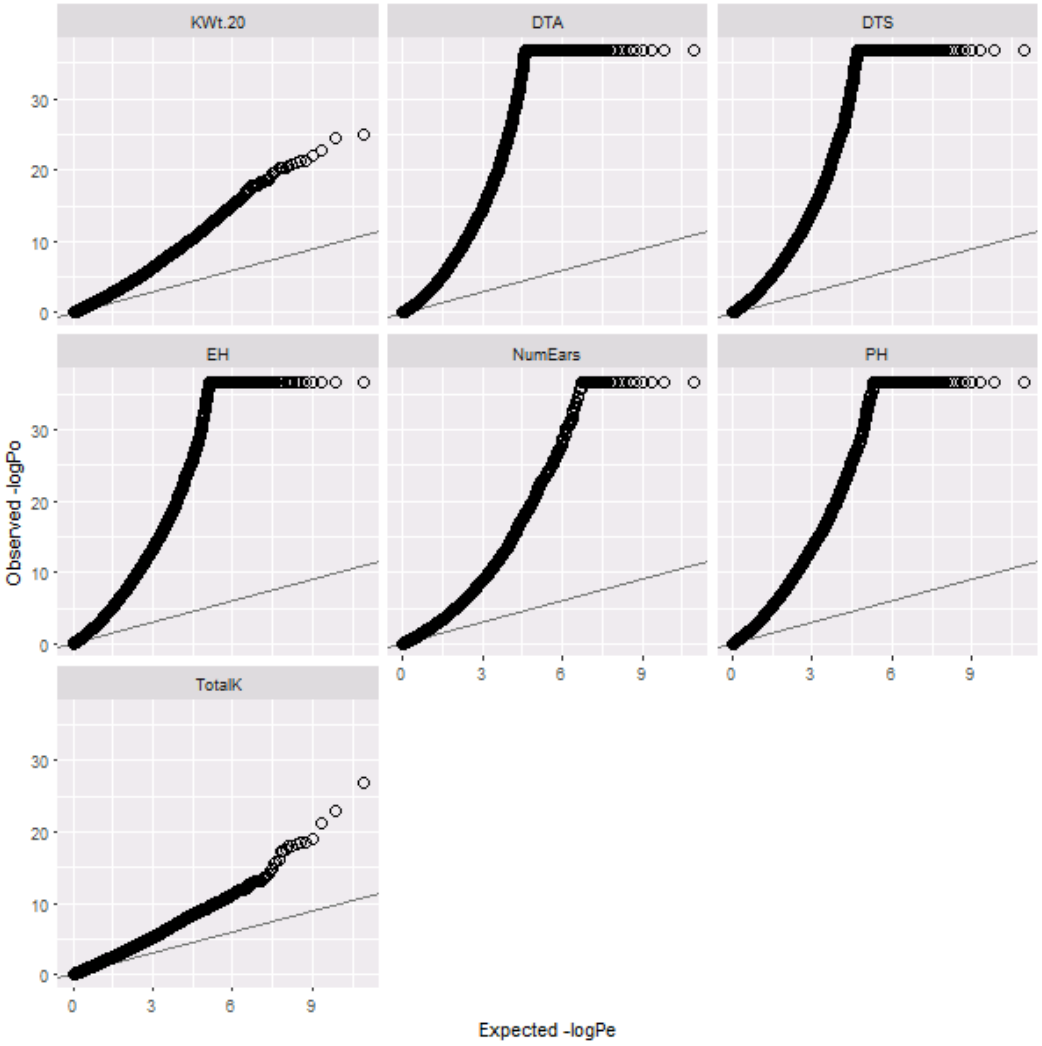

Figure C.2. Q-Q plot of P values for model 2 for all traits. Y-axis indicates observed –logP values. X-axis indicates expected –logP values from uniform distribution.

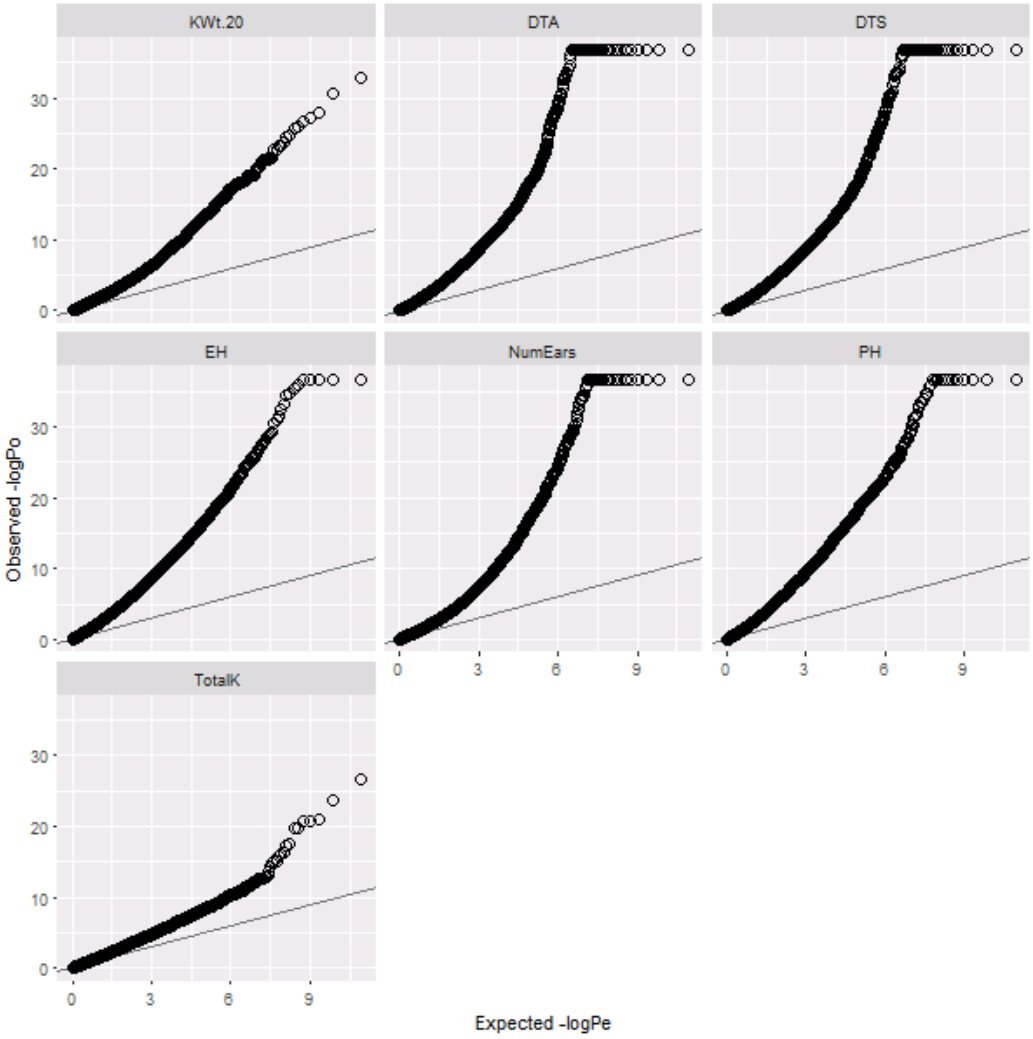

Figure C.3. Q-Q plot of P values for model 3 for all traits. Y-axis indicates observed –logP values. X-axis indicates expected –logP values from uniform distribution.

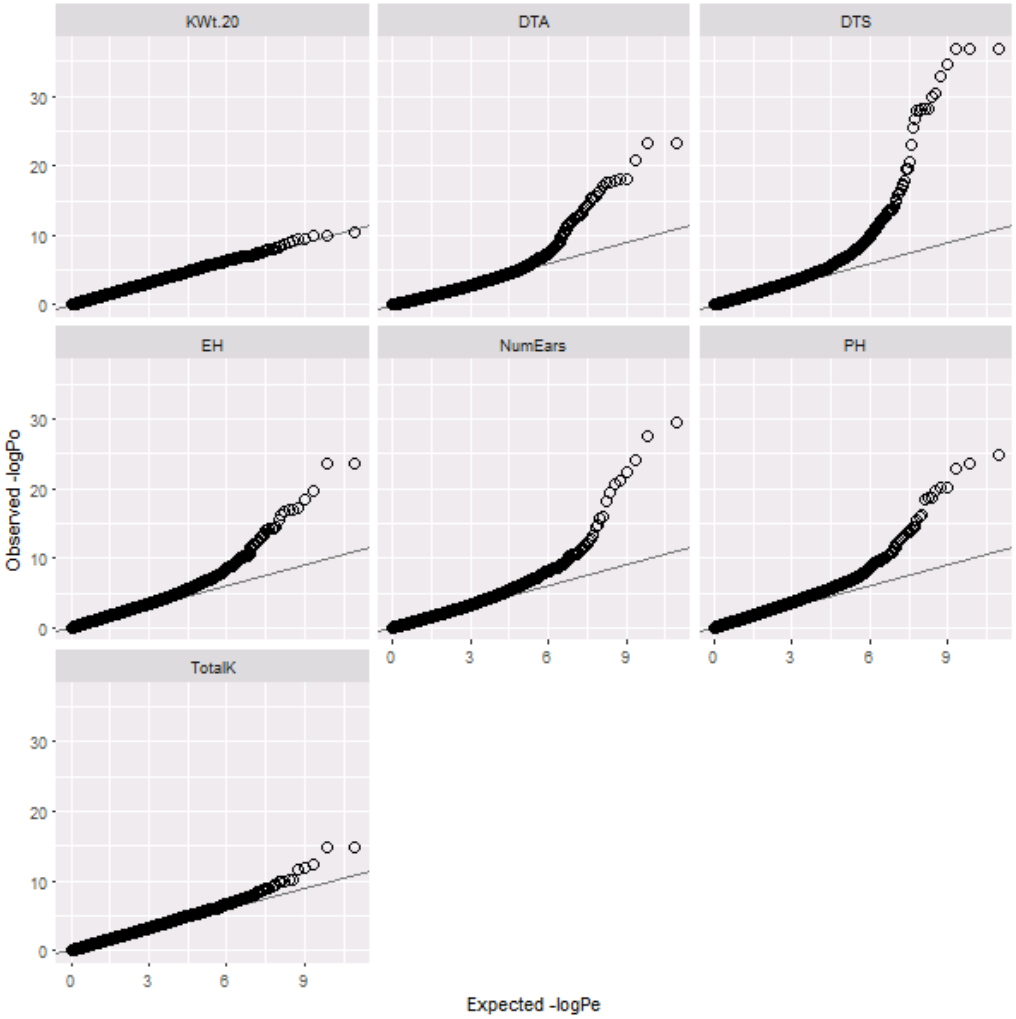

Figure C.4. Q-Q plot of P values for model 4 for all traits. Y-axis indicates observed –logP values. X-axis indicates expected –logP values from uniform distribution.

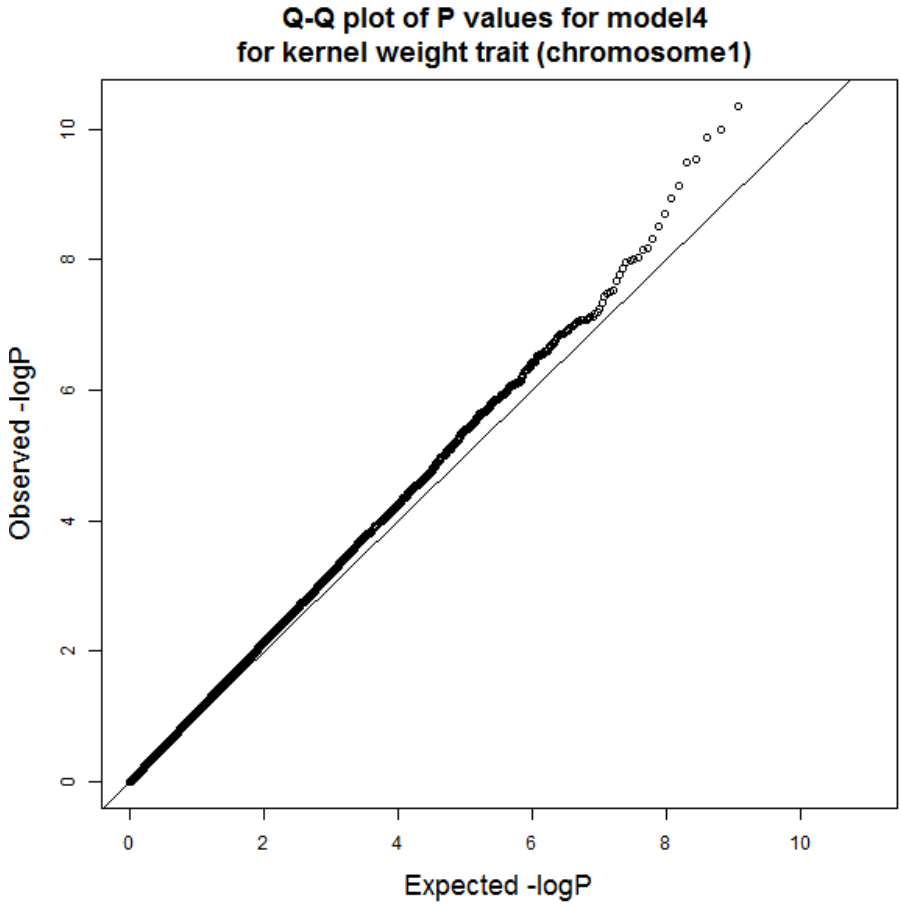

Figure C.5. Q-Q plot of P values for model 4 for kernel weight trait at chromosome 1. Y-axis indicates observed –logP values. X-axis indicates expected –logP values from uniform distribution.

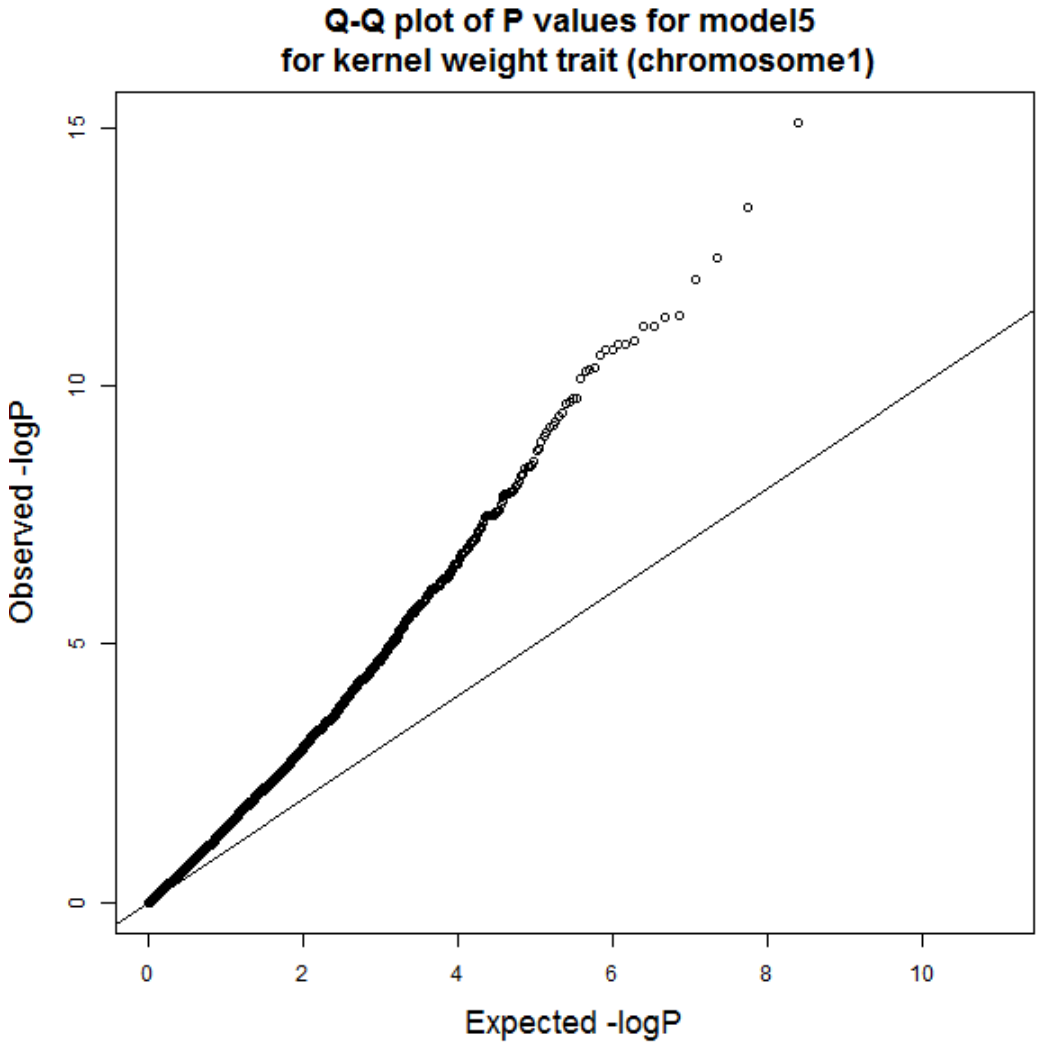

Figure C.6. Q-Q plot of P values for model 5 for kernel weight trait at chromosome 1. Y-axis indicates observed –logP values. X-axis indicates expected –logP values from uniform distribution.

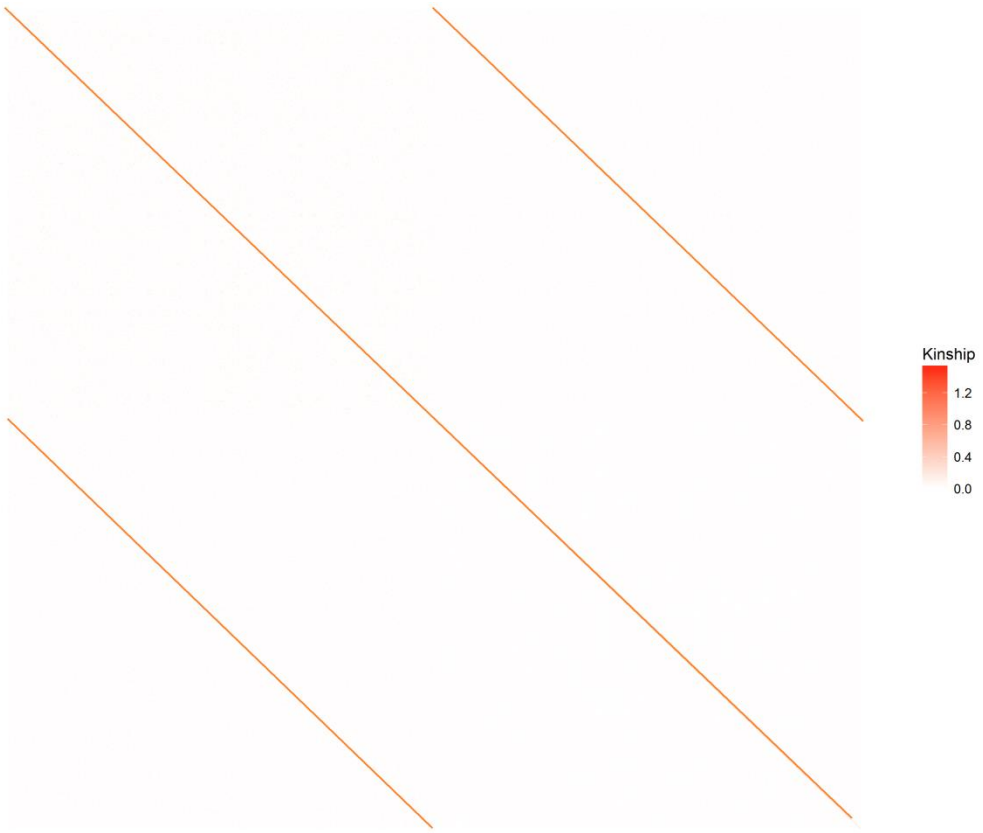

Figure C.7. Heatmap of kinship of progeny families

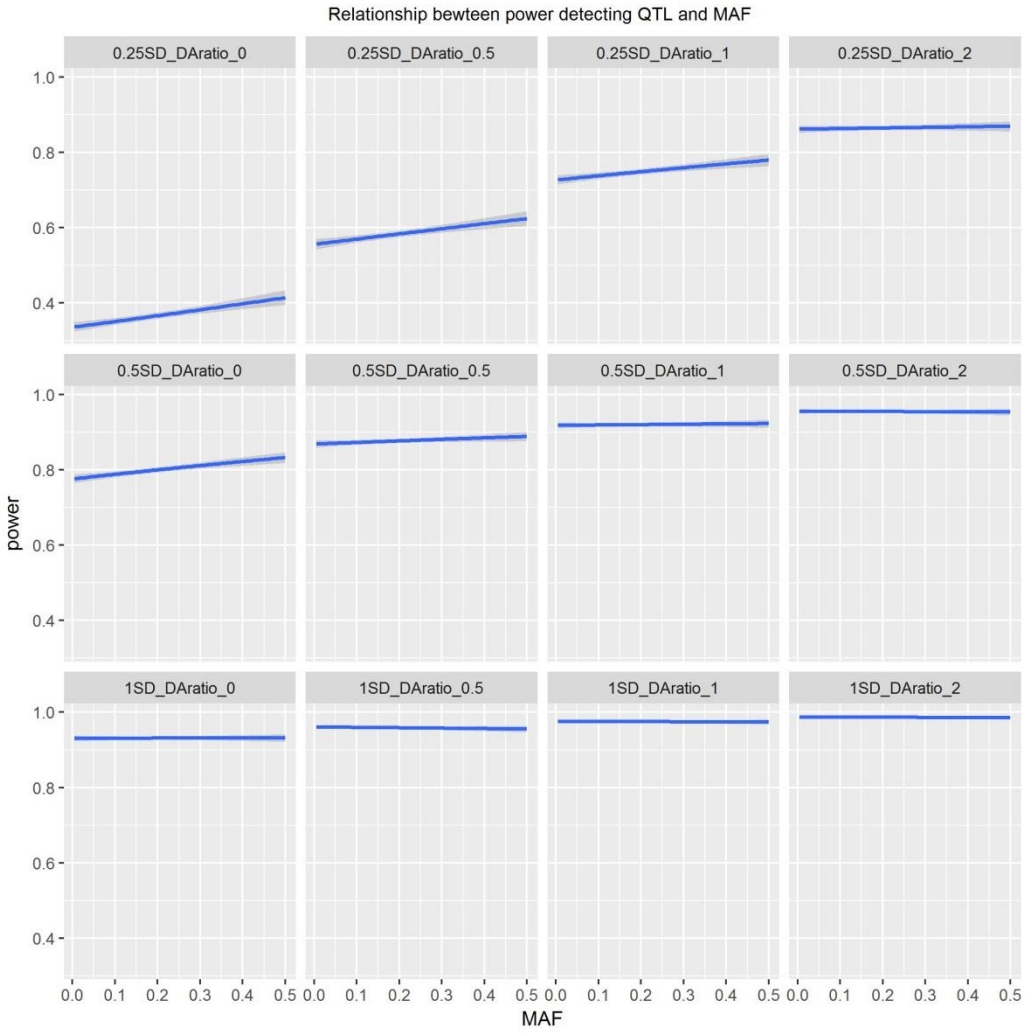

Figure C.8. Relationship between power detecting QTL and MAF

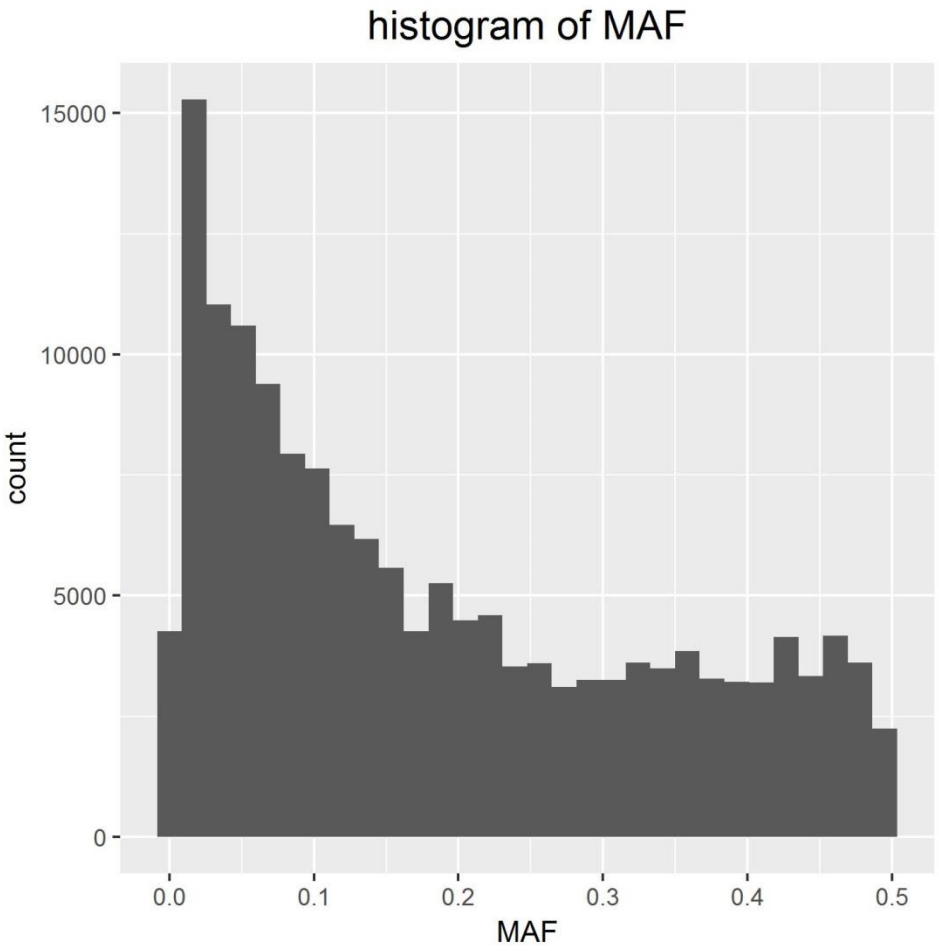

Figure C.9. Histogram of MAF for the simulated QTLs

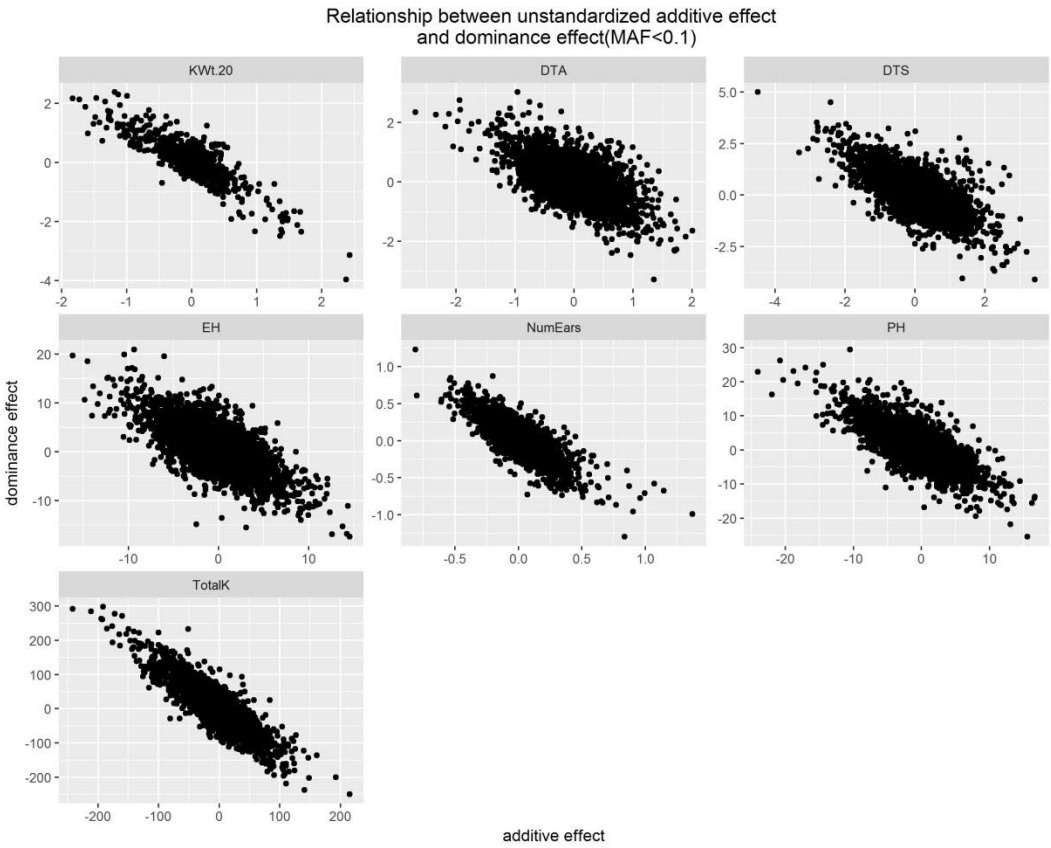

Figure C.10. Relationship between unstandardized additive effect and unstandardized dominance effect for markers with MAF below 0.1

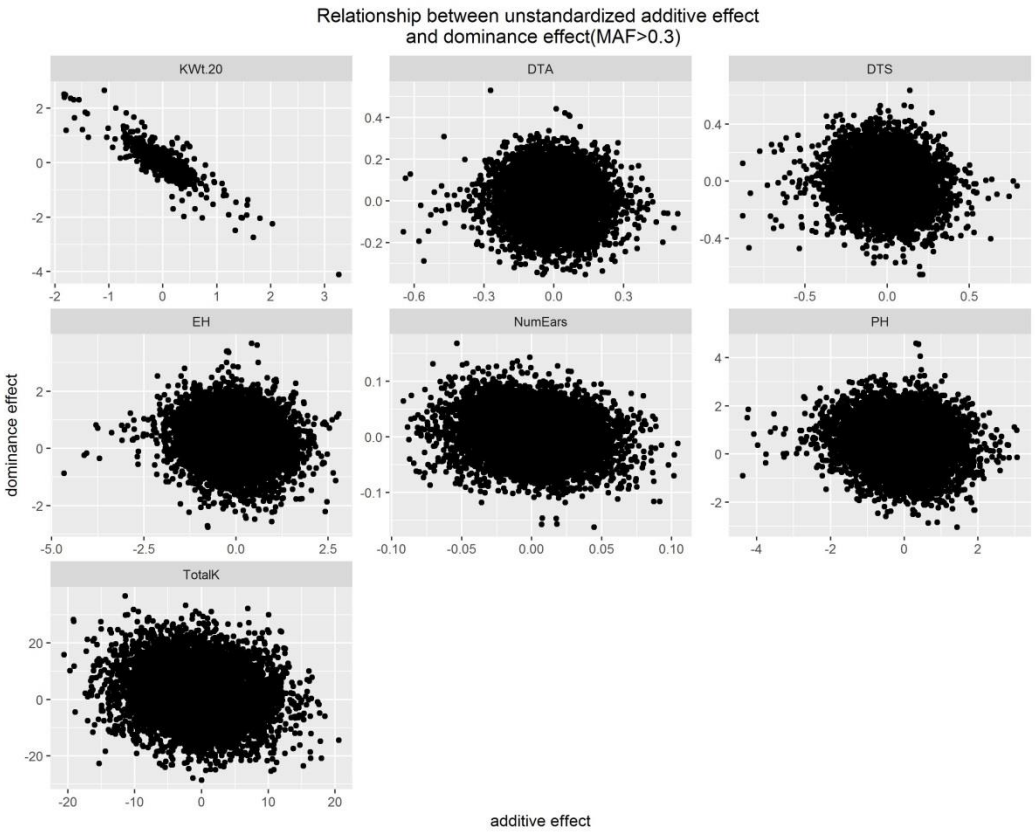

Figure C.11. Relationship between unstandardized additive effect and unstandardized dominance effect for markers with MAF above 0.3

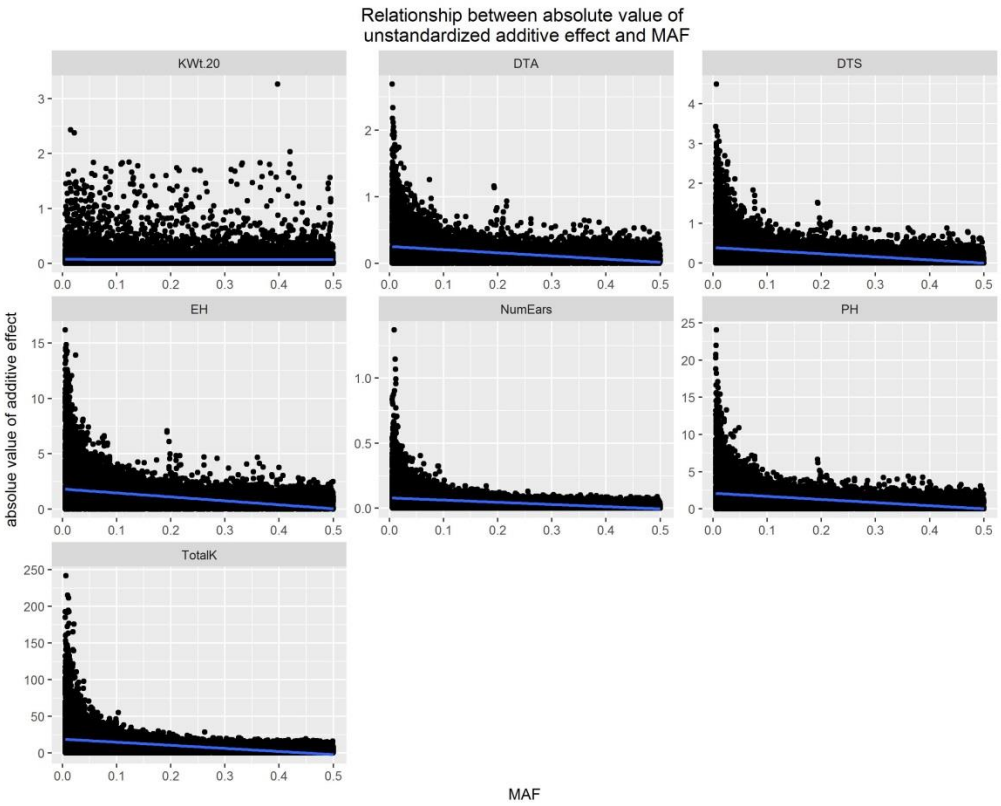

Figure C.12. Relationship between absolute value of unstandardized additive effect and MAF

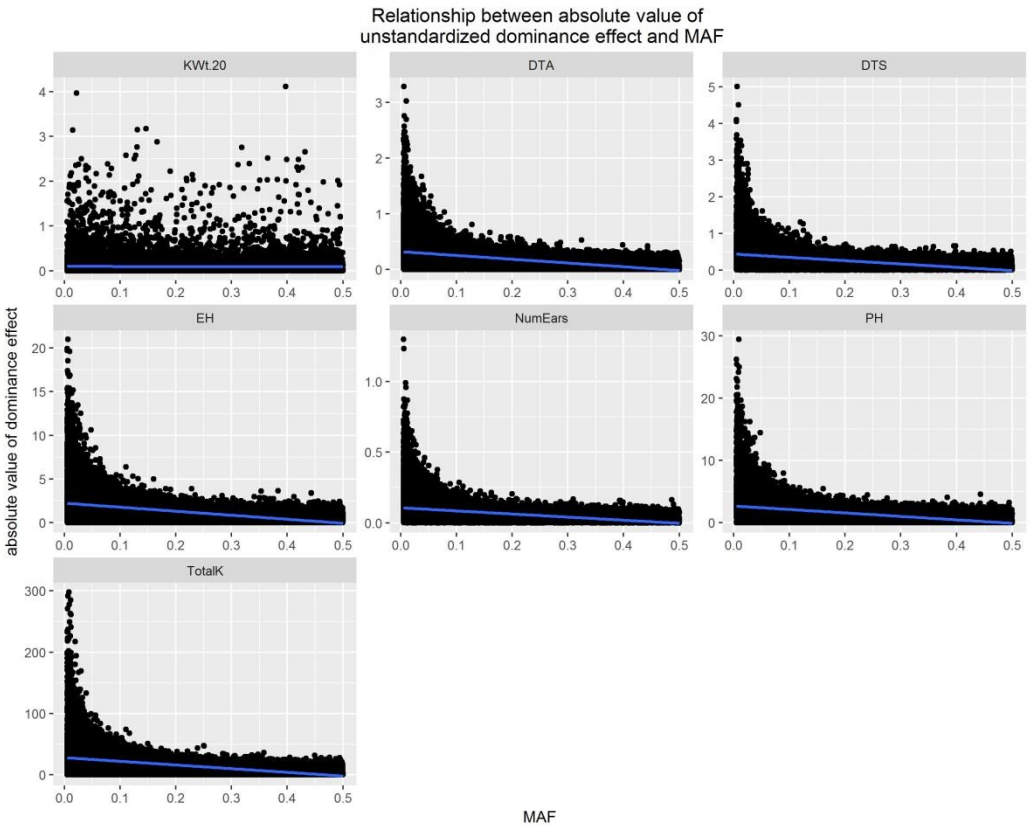

Figure C.13. Relationship between absolute value of unstandardized dominance effect and MAF

Table C.1. Calculating expected additive and dominance coefficient

Parent1	Parent2	Progeny expected additive coefficient	Progeny expected dominance coefficient
0	0	0	0
0	1	0.5	0.5
0	2	1	1
1	1	1	0.5
1	2	1.5	0.5
2	2	2	0

Table C.2. Five model for analysis

Model	
1	$Y = \mu + X_{add} * \boldsymbol{a} + X_{dom} * \boldsymbol{d} + \varepsilon$
2	$Y = \mu + F_{exp} * f + X_{add} * \boldsymbol{a} + X_{dom} * \boldsymbol{d} + \varepsilon$
3	$Y = \mu + F_{exp} * f + X_{add} * \boldsymbol{a} + X_{dom} * \boldsymbol{d} + G_{eigen} * E + \varepsilon$
4	$Y = \mu + F_{exp} * f + X_{add} * \boldsymbol{a} + X_{dom} * \boldsymbol{d} + G_{minus} + \varepsilon$
5	$Y = \mu + F_{exp} * f + X_{add} * \boldsymbol{a} + X_{dom} * \boldsymbol{d} + G + \varepsilon$

Y represent mean of phenotypic value for each family, F_{exp} is expected inbreeding coefficient (0 or 0.5), *f* is the estimated inbreeding effect, X_{add} expected additive effect coefficient, *a* is the estimated additive effect, X_{dom} expected dominance effect coefficient, *d* is the estimated dominance effect, G family polygenic effects used for correcting for population structure, cov(G) = $\boldsymbol{K_A}*\sigma_g^2$, G_{minus} family polygenic effects used for correcting for population structure, cov(G_{minus}) = $\boldsymbol{K_{minus}}*\sigma_g^2$, $\boldsymbol{K_{minus}}$ is calculated from all SNPs from other chromosomes except the chromosome that is being tested, E is the estimated fixed effects for controlling population structure when eigenvectors of $\boldsymbol{K_A}$ are included, ε is random error, cov(ε) = $\boldsymbol{I}*\sigma_e^2$.

Table C.3. Linear regression coefficient for regressing absolute value of genetic effects on MAF

Trait Name	Beta of add/SE(add)	R square (additive)	Beta of dom/SE(dom)	R square (dominance)
KWt.20	0.86***	0.001	0.22***	0.0004
DTA	-1.87***	0.035	-0.89***	0.10
DTS	-2.65***	0.043	-1.05***	0.09
EH	-2.60***	0.043	-1.15***	0.09
NumEars	-3.24***	0.057	-1.56***	0.09
PH	-2.80***	0.045	-1.26***	0.09
TotalK	-3.88***	0.072	-2.16***	0.10

Table C.4. Linear regression coefficient for regressing absolute value of genetic effects on genetic distance per unit physical distance (proportional to recombination rate)

Trait Name	Beta of add/SE(add)	R square (additive)	Beta of dom/SE(dom)	R square (dominance)
KWt.20	-279823	0.0006	7449.7	3.7E-05
DTA	-129251***	0.0089	-921.9	2.0E-04
DTS	-136189***	0.0103	-706.5	9.0E-05
EH	-15252.7***	0.0027	-350.5	3.9E-04
NumEars	-18000.8***	0.0047	-20789.1 *	8.9E-04
PH	-525627***	0.0047	-868.5 ***	2.8E-03
TotalK	-604.178	0.0004	-96.5	4.3E-04

Files C.1-C.2
Supporting data
Available for download at:
https://drive.google.com/drive/folders/0By4mQNWXnpaab192Ny16Zk9CbmM

File C.1. Calculating progeny kinship from parental genotype.docx

File C.2. Python codes for GWAS