View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Illinois Digital Environment for Access to Learning and Scholarship Repository

GENOME-WIDE ASSOCIATION STUDY FOR NON-NORMALLY DISTRIBUTED TRAITS: A CASE STUDY FOR STALK LODGING IN MAIZE

BY

ESPERANZA M. SHENSTONE

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Crop Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Adviser:

Assistant Professor Alexander E. Lipka

ABSTRACT

The abundance of new genomic information available has increased the ability of computational tools to study the genetic basis of agricultural traits, notably with the application of the Genome-Wide Association Study (GWAS). A limitation of GWAS is that the assumptions underlying the linear model typically used to conduct the analysis are often violated in nature, and in such cases, the linear model is inappropriate to use. Alternatively, the mixed logistic regression model is well-suited for a genome-wide association study of binomially distributed agronomic traits because it can include fixed and random effects that account for spurious associations. However, the computational burden associated with fitting this model renders it inefficient to use at every genetic marker that are analyzed in the genome-wide association study. Therefore, the purpose of this work was to assess the ability of simpler statistical models to identify promising subsets of genome-wide markers to apply to the mixed logistic regression model. We tested this approach on stalk lodging, a binomially distributed trait measured on a maize (*Zea mays* L.) diversity panel. This analysis culminated in the mixed logistic regression model identifying genomic regions coinciding with signals associated with closely related quantitative traits. Using genomic data from the same panel, we conducted a simulation study to determine which parameters of the binomial distribution most likely contribute to the detection of quantitative trait nucleotides. The results suggest that the discovery of such signals is maximized when the probability of a successful Bernoulli trial is 0.5. Based on our findings, we present an analytical framework that involves phenotyping binomially distributed traits so that the possibility of identifying associated markers is maximized and then prioritizes subsets of genome-wide markers for fitting the mixed logistic regression model; such prioritization should make it practical to use the mixed logistic regression model to test for marker-trait associations on an average computer.

ACKNOWLEDGEMENTS

I would first like to thank my advisor Dr. Alexander E. Lipka whose guidance over the last few years was crucial to my success as a graduate student, and will have a lasting impact on my future endeavors. I would also like to thank my thesis committee members: Dr. Tiffany Jamann, Dr. Martin Bohn, and Dr. Patrick Brown. Their insights and feedback were invaluable throughout the course of this project. Additionally, I would like to extend thanks to Julian Cooper and the Jamann Lab, for generously allowing me to use their field trials to collect data. Furthermore, I would like to thank the members of the Lipka Lab: Angela Chen, Brian Rice, and Amanda Owings. Their feedback and support has been much appreciated.

To my parents, Mark and Amanda Shenstone, for always supporting me

TABLE OF CONTENTS

CHAPTER 1: LITERATURE REVIEW

Introduction

The United States grows and exports more maize than any other country in the world (U.S. Grains Council, 2015). Maize (*Zea mays* L.) is the most widely grown feed grain in the United States, with an estimated 90 million acres of land planted every year (USDA Economic Research Service 2015). One major impediment towards maize production is stalk lodging, which is defined as the collapse of a cereal stem when it is no longer able to support its own weight. It is estimated that between 5-25% of maize yield will be lost to lodging on a yearly basis (Flint-Garcia *et al.* 2003). In an industry valued at $65 billion (Barton and Clark 2014) this is a significant economic loss, prompting the need for further investigation of stalk lodging.

Maize Agronomic Practices for Lodging Prevention and Management

Currently, many growers employ agronomic management practices that may help reduce the presence of the factors most responsible for stalk lodging. The Extension of Purdue (Nielsen and Colville 1988) lists plant stress, plant sugars, and stalk rot as three interrelated factors causing stalk lodging. The onset of environmental stressors can affect sugar mobilization within the plant, which ultimately results in the incidence of stalk rot. From a management standpoint, each of these factors can be taken into consideration during crop production, as growing practices can be instrumental in avoiding undue environmental stress. For example, planting date has the potential to affect the success of a crop; maize planted earlier in the growing season may more efficiently use solar radiation and soil nutrients. This improves the overall health of the plant, causing increased standability at the end of the growing season and thus reducing the chance of stalk lodging. Another factor known to affect lodging is planting density; higher densities may lead to increased competition for light and soil, causing nutrient deficiencies. This was observed in a recent study where the effects of planting density and nitrogen rate on lodging were analyzed. Overall, at increased planting density, and decreased nitrogen rate, more lodging was observed among the tested hybrids (Shi *et al.* 2016). Furthermore, the quality of the soil can affect the susceptibility of maize plants to lodging. For example, depleted nutrient levels in the soil can cause a reduction in stalk strength, leaving plants more susceptible to lodging. Poorly drained soils also affect root growth, which in turn affects the plants' susceptibility to lodging. Moreover, the lack of moisture in soil induces drought stress in the plants, causing sugar build-ups within the xylem. Consequently, the plant becomes prone to stalk rot, thus increasing the chances for stalk lodging (Nielsen and Colville 1988).

Present strategies for predicting the occurrence of stalk lodging are conducted at the phenotypic level. For example, a simple method used to predict stalk lodging is known as the squeeze test, where the stalk is squeezed at two nodes (Thomison and Paul 2012). If squeezed easily, there is likely to be stalk rot within the stem and thus greater susceptibility to lodging. Another method for quantifying stock lodging is the push test, which involves pushing the stalks at ear level 6-8 inches from the upright position (Thomison and Paul 2012). Breakage in the stalk between the ear and the lowest node is an indicator of stalk rot. Such methods are used to identify the adverse effects of stalk lodging and then to harvest fields that show susceptibility as soon as possible, which would reduce the possibility of grain loss due to lodging (Thomison and Paul 2012). Current management techniques are time sensitive; that is, the risk of lodging not being detected in time or being checked for too soon in the growing season is high.

Theoretically, the utilization of genomic sources underlying the incidence of stalk lodging could substantially aid in predicting when lodging could occur during a field season. However, predicting stalk lodging from genomic information will likely be difficult, as many factors can contribute to stalk lodging, including overall stalk strength, stalk composition, rot issues, and pests. In addition, adverse weather events such as strong winds and rain can exacerbate the susceptibility of maize to lodging (Thomison and Paul 2012). These various contributing factors suggest that stalk lodging might be a complex trait. By dissecting the genetic architecture of this trait, it could be possible to determine the number of loci that contribute to the genetic sources of its phenotypic variability, as well as their effect sizes and types (i.e., additive, dominance, or epistatic effects). Because of this putatively complex genetic architecture, it is possible that there are many small effect loci contributing to the genetic variation underlying stalk lodging, making it difficult to identify quantitative trait loci (QTL).

Genome-wide Association Study

The genome-wide association study (GWAS) has the potential to facilitate the identification of genomic loci associated with stalk lodging. Used as a QTL discovery tool, the GWAS utilizes genome-wide marker sets to search the genome for polymorphisms that are associated with a phenotype of interest (Lipka *et al.* 2015; Ogura and Busch 2015). A factor underlying the ability of a GWAS to successfully identify marker-trait associations is linkage disequilibrium (LD), defined as the non-random association of alleles at different loci (Chakravarti 2014). In GWAS, genetic markers, such as single nucleotide polymorphisms (SNPs), spanning the entire genome of a species are genotyped in every individual considered for GWAS. Then, a statistical model is used to search for indirect associations between SNPs and the trait of interest, relying on LD to infer the location of the causal variant. The most commonly used statistical approach for a plant GWAS is to fit a model at each marker, where the trait of interest is the response variable and the additive effects of the tested marker is an explanatory variable. (Lipka *et al.* 2015).

Association mapping does not require population development; rather it makes use of a diversity panel, a previously existing set of individuals meant to capture the allelic diversity of a species' genome. Two widely studied diversity panels in maize include the Goodman-Buckler diversity panel (Flint-Garcia *et al.* 2005) and the Ames diversity panel (Romay *et al.* 2013). The individuals in a diversity panel are assumed to capture the majority of historical recombination events that occurred since a theoretical ancestor population, thus allowing higher genomic resolution to identify causal mutations relative to individuals derived from biparental crosses (Lipka *et al.* 2015). To adequately cover the LD structure in these maize diversity panels, at least hundreds of thousands of genomic markers are needed (Lipka *et al.* 2013, 2015). The resolution offered by such markers translates to a state-of-the-art GWAS being able to identify genomic regions with moderate to strong associations with a trait (Spain and Barrett 2015).

Fine-mapping techniques are often implemented to elucidate the causal mutations underlying genomic regions identified in a GWAS (Spain & Barrett, 2015). Once genetic associations are identified in GWAS, fine mapping is used to further discern the casual variant associated with the trait of interest, and identify the target gene. This process involves the development of a new population that segregates for your genomic locus of interest. These techniques require accurate genotyping, high-quality data, and large sample sizes (Spain and Barrett 2015). Once the fine-mapping process is completed, the results must be confirmed. Briefly, the genetic region of interest is transformed into a near isogenic line (NIL), which is the grown out, increased for seed, and then phenotyped for the trait of interest. The combination of GWAS and fine-mapping creates an efficient gene discovery technique that has contributed to the further identification of novel genes in maize.

Linear GWAS Models

The most widely used statistical model in plant GWAS is the unified mixed linear model (MLM; Yu *et al.* 2006), which uses fixed and random effect covariates to control for population structure and familial relatedness. Specifically, population structure is controlled for through the incorporation of fixed effect covariates (e.g., principal components from a principal component analysis, as described in Lipka et al., 2015). To account for relatedness, the individuals are included as a random effect in the GWAS model, and then an additive genetic relatedness matrix (i.e., a kinship matrix) is used to estimate the variance-covariance between the individuals. When conducting a GWAS, it is imperative that the sample size of the diversity panel is sufficiently large to ensure adequate statistical power to detect associated loci. Nevertheless, it is common for many traits to be regulated by small effect loci, many of which are undetectable because of the inherent conservativeness of the GWAS (Ogura and Busch 2015).

Traditional iterative algorithms used to fit the unified MLM to the data are computationally intensive. To ease this computational burden, several approaches including the compressed mixed linear model (CMLM; Zhang *et al.* 2010), efficient mixed model association expedited (EMMAX; Kang *et al.* 2010) population parameters previously determined (P3D; Zhang *et al.* 2010), factored spectrally transformed linear mixed models(FaST-LMM; Lippert *et al. 2011),* Enriched CMLM (Li, Liu, *et al.* 2014) and genome-wide efficient mixed model association (GEMMA; Zhou and Stephens 2012) have been implemented into software specifically designed to conduct a GWAS. Collectively these approaches reduce computational time by either reducing the dimensionality of the variance-covariance matrix between the individuals or utilize mathematical algorithms that approximate or provide the most statistically appropriate parameter estimates (Lipka *et al.* 2012).

Logistic GWAS Models

One straightforward manner for quantifying stalk lodging in a statistical framework of a maize plant is as a Bernoulli trial, where a success is if the plant lodges and a failure is that it does not lodge. An important ramification of this quantification is that from a statistical perspective it is inappropriate to use the unified MLM, which assumes that the error terms are normally distributed. An alternative to the unified MLM to test for genotype-phenotype associations is the mixed logistic regression model (Agresti and Kateri 2011). Similar to the unified MLM, the mixed logistic regression model uses fixed and random effect covariates to control for population structure and familial relatedness. However, this model is used to test for associations between SNPs and either a Bernoulli- or binomial (i.e., number of successes in a series of independent, identical Bernoulli trials) distributed trait. Because the expected value and variance depends on the probability of a successful Bernoulli trial π_i, a change in the value of π_i will also change the variance. This makes the analysis of binary data using the standard unified mixed linear model inappropriate, as the assumption of constant variance is not met.

By nature, fitting a mixed logistic regression model bears a higher computational load compared to a mixed linear model. Unlike linear models (which use least squares to estimate β), logistic regression uses an iterative algorithm to obtain maximum likelihood estimates for logistic regression parameters (e.g., marker effect estimates and the effects of principal components accounting for population structure). The computational load associated with maximum likelihood functions is compounded by the introduction of a random effect accounting for the individuals. This computational burden limits its use in GWAS as the corresponding logistic regression model is fit at potentially hundreds of thousands of genetic markers, and commercially available computers are unable to complete the analysis in a reasonable timeframe. An R package, titled Generalized Linear Mixed Model Association Tests (GMMAT; Chen *et al.*

2016), serves to reduce this computational burden via the implementation of score tests. Because the score test statistic is calculated by looking at the first derivative of the likelihood function of the data under H_0: $no\ association\ between\ marker\ i\ and\ the\ trait$, there is no need to refit a separate mixed logistic regression model at each marker. Thus, the computationally intensive model fitting procedure for a mixed logistic regression model only needs to be done once for a null logistic regression model with principal components of the markers included as fixed effects to account for population structure, and the individuals as random effects. The R package GMMAT then conducts a score test at each genetic marker to test for statistical association between the marker and binary trait. Currently, the GMMAT R package only runs on the UNIX operating system, limiting its widespread implementation in the scientific community.

The logistic regression model has been successfully implemented for GWASs of human disease, where the case-control nature of the data requires the use of non-linear models (Li *et al.* 2016). For example, logistic regression GWAS was used to identify alleles associated with sporadic post-menopausal breast cancer (Hunter *et al.* 2007). It is important to note that these studies did not include a random effect and that the incorporation of a random effect into our GWAS models will pose an additional challenge. However, based on the success of these studies, it is plausible that agronomic binary traits such as stalk lodging can also be analyzed using these methodologies.

Traits Related to Lodging

Currently, few published studies have investigated stalk lodging by directly phenotyping for lodging. Instead, previous studies have investigated the genetic architecture of stalk lodging in maize indirectly by assessing closely related quantitative traits. For example, Peiffer et al. (2013) assert that stalk strength is directly correlated with lodging. Thus, they explore the genetic architecture of stalk strength. There was extensive data collection from the 4,692 RILs that comprise the US nested association mapping (NAM panel) (Yu *et al.* 2006; McMullen *et al.* 2009) an intermated B73×Mo17 population (IBM) of 196 RILs (Lee *et al.* 2002) as well as the Ames panel (Romay *et al.* 2013). Rind penetrometer resistance was used to quantify stalk strength of the maize; however, two different tools were used for phenotyping. Because of this, a potential source of additional variability is added to the reported stalk strength phenotypes. However, Peiffer et al. (2013) were able to account for environmental variability in stalk strength by replicating the study in three different environments. Broad sense heritabilities were calculated and were found to vary by population; 0.20 for the NAM, 0.34 for the IBM, and 0.54 for the NCPRIS. It was found that 37% of stalk strength variation in the NAM and IBM populations could be explained by variation in environment, 15% was due to genetic factors, 11% due to the genotype by environment (GxE) interaction, and the remaining 37% was attributed to unknown sources and error (Peiffer *et al.* 2013).

A similar study by Li et al. (2014) also investigates the genetic architecture of rind penetrometer resistance in maize. Two RIL populations with parental lines of variable stalk strength were developed for this study. Linkage mapping was used to identify potential QTL associated with rind penetrometer resistance (RPR); with phenotypic variance percentages ranging from 4.4% to 18.9%. The largest QTL identified in this study were also identified in previous studies, Flint-Garcia et al. (2003), and Hu et al.(2012). Despite this commonality, of the 33 populations of maize studied for RPR (that were found in primary literature as of 2014), 69 QTL have been identified, and only 10 were found to occur in at least two populations (Li *et al. 2014)*. More recently, a stalk strength study focused on the morphological characteristics of the maize stem and how they relate to stalk lodging (Robertson *et al.* 2017). In this study, the elliptical section modulus (a morphological trait) was found to explain 80% of the variation in stalk strength, whereas in the same study RPR (a material trait) was only found to explain 18% of the variation in stalk strength. Multiple morphological and material traits were analyzed, and it was found that overall morphological traits explained more variation in stalk strength than their material counterparts did. In light of these new findings, the heritabilities calculated from RPR may not accurately represent the heritability of stalk strength. Thus, future studies that examine the relationship of stalk strength and stalk lodging, may be more informative when morphological proxy traits rather than material traits are used.

Another class of traits associated with stalk lodging in maize includes biotic stresses such as rot and pests. One such pest is the European corn borer (ECB), which causes damage to the corn plant by laying eggs and feeding within the leaves and stalks. The burrowing by larvae from ECB increases susceptibility to infection by creating entry points for fungus to infect the plant (Munkvold and Hellmich 1999). For example, the ECB can serve as a vector for *Fusarium*, a fungus that causes stalk rot of maize. As the larvae burrow into the plant, they can carry *Fusarium* spores from the exterior with them. Infestation by ECB is estimated to cause about one billion dollars of loss every year (Ostlie *et al.* 1997). Janvis et al. (1984) studied the relationship between ECBs and stalk rot. The main findings of this study were that the presence of stalk rot did not affect the incidence of ECBs; however, the presence of ECBs led to an increase in stalk rot susceptibility (Janvis *et al.* 1984).

Interestingly, it is also possible that diseases not associated with stalk rot may indirectly increase the incidence stalk rot as result of the underlying symptoms of the disease. One such disease of interest is Goss's Wilt. Goss's Wilt is a bacterial blight, *Calvibacter michiganenis*, that affects maize crops in the United States. The disease originated in Nebraska in 1969, and spread throughout the Midwest over the next two decades. Reported symptoms of the disease include leaf blight, necrosis, tissue death, and vascular wilt. Wilt disrupts the vascular system resulting, resulting in stalk rot and death of plants (Harveson 2011). This disease has been previously linked to stalk lodging, however the lack of published research on this topic warrants further investigation. Other maize diseases, such as anthracnose, are also believe play a role in lodging susceptibility(Jirak-Peterson and Esker 2011), however more research is needed to better quantify the relationship between this diseases and stalk lodging.

A major concern in using breeding to improve stalk lodging resistance in maize is the potential for loss in quality of the grain (Nielsen and Colville 1988). Albrecht et al. (1985) conducted experiments to determine if recurrent selection for stalk strength and stalk rot resistance would alter the composition of the stalk in a way that would negatively affect forage quality. It was revealed that three cycles of recurrent selection for stalk rot resistance and stalk strength resulted in an overall increase in vitro digestible dry matter (Albrecht *et al.* 1986). This result implies that the forage quality of maize will not be degraded as a result of stalk lodging breeding efforts, suggesting that breeding efforts towards stalk lodging will not be compromised in instances where forage quality is desired.

Dissecting the genetic architecture of stalk lodging in maize has potentially major implications for the entire corn industry and through extension crop breeding efforts as a whole. Many components can contribute to lodging. By exploring these components individually and analyzing where they intersect, there will be great potential to obtain a greater understanding of stalk lodging.

Conclusion

Few studies have analyzed traits related to stalk lodging, but few to none have studied the genetic basis of stalk lodging in maize. This can possibly be attributed to the complex nature of this trait as well as the statistical ramifications of analyzing a non-normally distributed trait. The purpose of this study is evaluate a new approach to mixed logistic regression GWAS and further explore the usefulness of GWAS for binary traits. We hypothesize that by using our methodology, GWAS results will be able to be obtained in a reduced time-frame.

We also hypothesize that inherent properties of the phenotypic data may affect our ability to detect genomic signals. By investigating these applications of mixed logistic regression GWAS we can use the information obtained from this study to improve upon future genomic studies of binary traits.

CHAPTER 2: MATERIALS AND METHODS

Germplasm

To investigate the genetic variability of stalk lodging in maize, the Goodman-Buckler diversity panel (Flint-Garcia *et al.* 2005) was used in these research trials. This diversity panel contains 281 unique maize lines, capturing 75% of all allelic diversity in maize (Romay *et al.* 2013). The panel consists of stiff stalk, non-stiff stalk, tropical, and popcorn lines. Seed was obtained from GRIN in 2016 and increased in 2016 via sib mating for use in 2017.

Experimental Design

Field trials were conducted in 2016 and 2017 at the Crop Sciences Research and Education Center in Urbana, IL. In both years, the population was planted in single row plots, 3.2m long, with a spacing of 0.76m between rows. Each row was planted using a vacuum planter at a density of 20 kernels/ row. The 2016 trial consisted of two inoculated replications of the Goodman-Buckler diversity panel (Flint-Garcia *et al.* 2005) in an incomplete block design. In 2017, two experiments of two replications each were conducted in an incomplete block design. One experiment was inoculated with Goss's Wilt, while the other served as a control and was not inoculated. The control experiment was added in 2017 to further examine the influence of Goss's wilt on stalk lodging in maize. In all experiments, check lines (FR4326, CQ184A, CQ183) were included in each incomplete block. The incomplete block design was created using the *agricolae* package (Mendiburu 2017) by Cooper et al. in R (R Core Team 2017).

Phenotypic Data

Phenotypic data were collected on the Goodman-Buckler diversity panel (Flint-Garcia *et al.* 2005) in Urbana, Il during the summers of 2016 and 2017. Stand count (the number of plants standing per plot) was recorded for each plot at 42 days after planting (DAP) in 2016 and 41

DAP in 2017. Stalk lodging measurements were taken 115 DAP in 2016 and 114 DAP in 2017. Stalk lodging was recorded by counting the number of plants stalk lodged within each plot. A plant was considered stalk lodged when the stalk is broken below the ear node. Root lodging was not examined in this study.

Goss's Wilt Inoculation

Goss's wilt is a foliar disease caused by the bacterium *Clavibacter michiganenis* subsp. *nebraskensis* that affects maize crops in the United States. The disease is known to cause leaf blight, necrosis, tissue death, and vascular wilt. Inoculant was prepared using a protocol described in Cooper at al. (in minor revision). Briefly, a *Clavibacter michiganensis* subsp. *nebraskensis* strain *16Cmn001* was used to prepare a solution for inoculation. Two rounds of inoculations were administered via pinprick one week apart. Disease ratings were then conducted by Cooper et al. to determine the area of leaf infected by Goss's Wilt, and then converted to an area under disease progress curve (AUDPC) (Wilcoxson *et al.* 1975).

Genome-Wide Association Study

The phenotypic data were formatted to include stand count, number of plants lodged, number of plants not lodged, and percent plants lodged per plot. As a result, we were able to set up a binomial experiment, where within each plot each plant is a Bernoulli trial that has two outcomes (lodged/not lodged). Furthermore, we assumed that the probability of a plant lodging is the same within a plot, and one plant lodging will not change the likelihood of another plant lodging. Thus, the trait considered for GWAS, the number of plants that lodge in a plot, follows a binomial distribution (Ott and Longnecke 2008). Consequently, it was determined that a mixed logistic regression model that includes fixed and random effect covariates to account for population structure and familial relatedness, would be the most appropriate model to fit for a GWAS for stalk lodging (Chen *et al.* 2016).

Genotypic information for the diversity panel had previously been obtained using the Illumina MaizeSNP50 BeadChip (referred to as 55K SNPs)(Cook *et al.* 2012) as well other genotyping technologies (referred to as 4K SNPs) (McMullen *et al.* 2009)(Yu *et al.* 2006). Principal components and kinship matrix were previously calculated in (Lipka *et al.* 2013) using the non-industry subset of 34,368 SNPs, and incorporated as fixed effects and random effects respectively.

Multi-Model Analysis

A consequence of analyzing binary data is the computational burden associated with using logistic regression to analyze random and fixed effects (Kiernan *et al.* 2012). Accordingly, a three-pronged modeling approach was developed to reduce computational time. The intention behind this multi-model approach was to identify a subset of SNPs most likely to be associated with lodging. The computationally intensive mixed logistic regression model could then be run on only this subset. Consequently, the time to complete such an analysis would be reasonable on a computer with average memory (8GB RAM) and processing capabilities (Intel Core Duo Processor). The following describes the three models used in this approach.

Model 1

Model 1 was fit using the 2016 field data in R Version 3.31 using a logistic regression model that accounts for population structure by incorporating the first three principal components of the 34,368 non-industry SNPs from the Illumina MaizeSNP50 BeadChip as covariates. Consider the i^{th} plot (consisting of a set of genotypically unique individuals) in the

j^{th} incomplete block, which consists of n_{ij} plants observed during the stand count. Then Model 1 can be written as follows

Model 1:

Y_i are independent binomial random variables with expected values

$$E\{Y_i\} = n_{ij} * \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged)$$

and variance of

$$Var(Y_i) = n_{ij} * \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged) * (1 - \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged))$$

and,

$$\log\left(\frac{\pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged)}{\pi(1 - \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged)}\right) = \mu + \sum_{k=1}^{3} \beta_k PC_{ik} + \alpha x_i + Block_j$$

Where:
$\pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged) = probability\ that\ a\ plant\ with\ i^{th} genotype\ in\ the\ j^{th}\ block\ lodges$
$\mu = the\ grand\ mean$

$\beta_k = fixed\ effect\ of\ the\ k^{th}\ principal\ component\ (PC)$

$PC_{ik} = value\ of\ the\ k^{th} PC\ for\ plant\ with\ i^{th}\ genotype$

$\alpha = fixed\ additive\ effect\ of\ the\ tested\ marker$

$x_i = observed\ genotype\ of\ tested\ marker\ for\ plant\ with\ i^{th}\ genotype$

$$= \begin{cases} 0, if\ aa \\ 1, if\ Aa\ or\ aA \\ 2, if\ AA \end{cases}$$

$Block_j = fixed\ effect\ of\ the\ j^{th} block$

Model 2

Model 2 fits the unified mixed linear model (Yu *et al.* 2006) using the R package GAPIT (Lipka *et al.* 2012). The use of this R package allowed us to implement the ‘population parameters previously determined' (P3D) function, meaning variance components were only estimated once. Best linear unbiased predictors (BLUPs) were calculated from a mixed model that incorporated random block effects that were then used as the phenotypic data in GAPIT. Additionally, principal components were incorporated as fixed effects to account for population structure and were the same as described in Model 1. Additionally this model incorporated an additive genetic relatedness matrix (kinship matrix)(Loiselle at al. 1995) to account for familial relatedness, that was calculated with the same subset of non-industry SNPs used to calculate the principal components.

Model 2:

$$Y_i = \mu + \sum_{k=1}^{3} \beta_k PC_{ik} + \alpha x_i + Line_i + \varepsilon_i$$

Where:

Y_i = *The phenotype of the ith individual*

μ = *The grand mean*

β_k= *the fixed effect of the kth PC*

PC_{ik} = *value of the kth PC at the ith genotype*

$\alpha = fixed\ additive\ effect\ of\ the\ tested\ marker$

$x_i = observed\ genotype\ of\ tested\ marker\ for\ plant\ with\ i^{th}\ genotype$

$$= \begin{cases} 0, if\ aa \\ 1, if\ Aa\ or\ aA \\ 2, if\ AA \end{cases}$$

ε_i= *Random error term associated with the ith individual*

and,

$Line_i = The\ random\ effect\ of\ the\ i^{th} genotype\ \ where$

(Line$_1$,…, Line$_n$) ~ MVN(0, 2Kσ_G^2)

K = kinship matrix

ε_i *~ i.i.d. N(0, σ_E^2)*

Model 3

Model 3 fits a mixed logistic regression model that controls for both population structure and relatedness(Chen *et al.* 2016) using the PCs and kinship matrix described previously. Due to the computational load associated with logistic regression and random effects, only a subset of markers exhibiting peak associations with lodging when fitted to Models 1 and/or 2 were used. This model was fit in SAS using PROC GLMMIX due to the option for a user-inputted kinship matrix. Consider the k^{th} line, in the j^{th} plot , in the i^{th} incomplete block consisting of n_{ij} plants observed during the stand count. Then Model 3 can be written as follows:

Model 3:

Y_i are independent binomial random variables with expected values

$$E\{Y_i\} = n * \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged)$$

and variance of

$$Var(Y_i) = \text{n*}\ \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged)\ (1 - \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged))$$

and,

$$\log\left(\frac{\pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged)}{\pi(1 - \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged)}\right) = \mu + \sum_{k=1}^{3} \beta_k PC_{ik} + \alpha x_i + Line_i + Block_j$$

Where:
$\pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged) =$
$probability\ that\ a\ plant\ with\ i^{th} genotype\ in\ the\ j^{th}\ block\ lodges$
$\mu = the\ grand\ mean$

$\beta_k =\ fixed\ effect\ of\ the\ k^{th}\ principal\ component\ (PC)$

$PC_{ik} = value\ of\ the\ k^{th} PC\ for\ plant\ with\ i^{th}\ genotype$

$\alpha = fixed\ additive\ effect\ of\ the\ tested\ marker$

$x_i = observed\ genotype\ of\ tested\ marker\ for\ plant\ with\ i^{th}\ genotype$

$$= \begin{cases} 0, if\ aa \\ 1, if\ Aa\ or\ aA \\ 2, if\ AA \end{cases}$$

$Block_j = fixed\ effect\ of\ the\ j^{th} block$

$Line_i = Random\ effect\ of\ the\ i^{th} genotype\ where$

$(Line_1, \dots, Line_n) \sim$ MVN(0, 2Kσ_G^2)
K = kinship matrix

Simulation Study

In the natural world, various evolutionary processes contribute to the genetic diversity we see today, however many of these processes have unforeseeable outcomes that limit our understanding of underlying genetic functions. Simulation studies circumvent the unpredictability of processes that occur in nature (Lipka *et al.* 2012) and are therefore useful in evaluating methodologies that can be used to analyze real data (Hoban *et al.* 2012). To further explore the efficacy of mixed logistic regression GWAS for binomial data we conducted a simulation study using R Version 3.31 (R Core Team 2017). The population used in this simulation study was the previously described Goodman-Buckler Diversity Panel(Flint-Garcia *et al.* 2005) that was genotyped using the 4k SNPs.

The factors that we hypothesized to be crucial for identifying genomic signals associated with binomially distributed traits were the baseline probability of a successful Bernoulli trail (e.g., the baseline probability of an individual plant lodging within the context of stalk lodging) and the number of Bernoulli trials considered (e.g., the number of plants in a plot recorded during the stand count prior to measuring stalk lodging). Thus in the simulation settings, we simulated binomially distributed phenotypes with and a grand mean (i.e., intercept of a logistic regression model; this controls the baseline probability of a successful Bernoulli trial) of zero, one, three or five and a total of 10, 15, 20, and 25 independent Bernoulli trials at each plot.

Because the purpose of this simulation study was to assess how these factors contribute to the success of identifying genomic loci associated with a binomially distributed trait, one SNP from the 4k marker set was randomly selected to be the quantitative trait nucleotide (QTNs) with a large additive effect of 0.9. With the exception of the settings where no QTNs were no simulated (which was ran to study the false positive detection rate), the same SNP with the same effect size was considered across all simulation settings. At each of these settings, a total of 100 traits were simulated.

Each simulated population was then fit to Model 1, Model 2, and Model 3 following the protocol previously described. To assess the utility of the three-model approach, we examined the proportion of times QTN were successfully detected by our models. A QTN was considered successfully detected if a marker was identified as significant at 5% FDR within 250 kb from the QTN. Due to the high volume of markers, the top 100 SNPs with a FDR of 5% were recorded for each setting.

Setting	Intercept β_o	Stand Count	Additive effect size
1	0	10	0.9
2	1	10	0.9
3	3	10	0.9
4	5	10	0.9
5	0	15	0.9
6	0	20	0.9
7	0	25	0.9

Table 2.1: Simulation settings used in simulation are listed in this table.

CHAPTER 3: RESULTS

Variability of Stalk Lodging in the Field

During the 2016 field season two replications of 299 inbred maize lines (as well as 86 plots of check lines) were phenotyped for stalk lodging. When examining variation between all plots, 217 of the 684 plots that were phenotyped experienced at least one lodged plant, with 51 of those plots having a proportion-lodged greater than 0.5. When examining variability on a within-taxa basis there appeared to be low repeatability across replications (Pearson correlation coefficient = 0.249). Overall, the distribution of the proportion of plants lodged per plot was highly skewed to the right, with the majority of the plots experienced no lodging (Figure 3.1), therefore reducing the variability of lodging in the field.

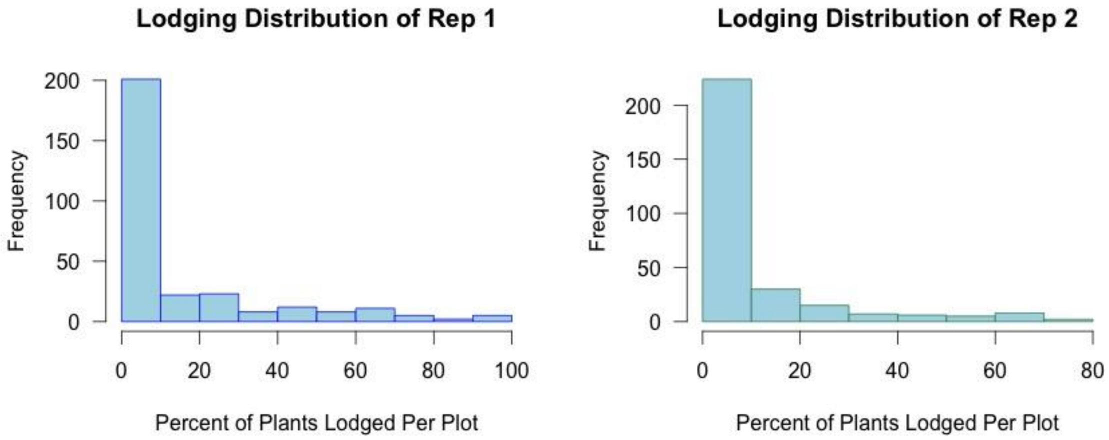


Figure 3.1: Histograms representing the distribution of the percent of plants lodged per pot on a replicate basis. The number of times a percentage was observed is reported on the y-axis, and the percentage is reported on the x-axis for each graph.

Multi-model Mixed Logistic Regression Identifies Peak SNPs Associated with Stalk Lodging in Maize

To examine the genomic underpinnings of stalk lodging in maize, we conducted a GWAS on n = 262 of the inbred maize lines from the Goodman-Buckler diversity panel that had previously been genotyped. After removing SNPs with MAF < 5%, a total of 49,332 SNPS were considered for the GWAS. Due to the binomial distribution of the stalk lodging data a novel multi-model approach was used to identify peak SNPs.

The first model (Model 1) was a binomial logistic regression model that controls for population structure, to perform GWAS on the maize stalk lodging data. In total, 24,211 SNPs were declared significant at an FDR of 5%, with the most significant SNP occurring on chromosome 1 (ss196523926, 286,987,962 bp, P-Value 3.3E-41). A Manhattan plot consisting of the results of Model 1 is in Figure 3.2. The top 2,796 SNPs (the number of SNPs that could be fit within a 24 hour time period) were subset from the output of Model 1 and used as the genotypic input for Model 3. In the second part of the multi-model approach, Model 2, a unified-mixed linear model was fit using GAPIT. Given that lodging is assumed to follow a binomial distribution, the model assumptions of the MLM an BLUP calculations were violated. The results of this model yielded no significant SNPs (Figure 3.3), and thus none were subset for use in Model 3.

Model 3 fit a mixed logistic regression model that accounted for population structure and relatedness using the 2,796 SNPs identified in the Model 1 as the genotype input file. At and FDR of 5%, 1,906 SNPS were declared to significant, with
the most significant SNP located on chromosome 5 (ss196463892, 83,398,133 bp, P-value 7.02E-09) (Figure 3.4). As previous genomic studies on lodging are not publicly available, the results of Model 3 were compared to those from association studies conducted on traits related to lodging.

Two traits that are frequently mentioned in relation to maize stalk lodging are stalk strength and rind penetrometer resistance (RPR). From four different previously published studies on stalk strength and RPR (Flint-Garcia *et al.* 2003; Hu *et al.* 2012; Peiffer *et al.* 2013; Li, Yan, *et al.* 2014) we identified regions of the genome that may also be associated with stalk lodging. Peak associations for these traits were previously found on chromosome 7 in the region 159.4Mb (Peiffer *et al.* 2013), chromosome 2 in the region of 236.4-237.0 Mb, and chromosome 3 in the region of 181.1-184.7 Mb (Flint-Garcia *et al.* 2003; Hu *et al.* 2012; Li, Yan, *et al.* 2014). Using the results in Model 3 (Figure 3.4), we were able to identify six significant SNPs that fell near these regions. Specifically, the three most significant SNPs on chromosome 7 were found at 161.9 Mb, 155.8 Mb, 164.9 Mb, the 14th most significant SNP on Chr 2 was found at 236.8 Mb, and the 92nd and 98th most significant SNPs on chromosome 3 were found at 181.7 Mb and 182.0 Mb respectively (Table 3.1).

Based on these findings we chose to look at the rate of LD decay in the region surrounding the most significant SNP on chromosome 7 (ss196481136, 164,952,176 bp, P-value 6.46E-08). The rate of LD decay in this region is presented in Figure 3.5, where LD between our SNP of interest and SNPs found in the region of 160.0 Mb and 168.0 Mb is plotted against physical location within the genome. The results of this plot indicate that there is not a high level of LD within this specified region.

Model 1 Results

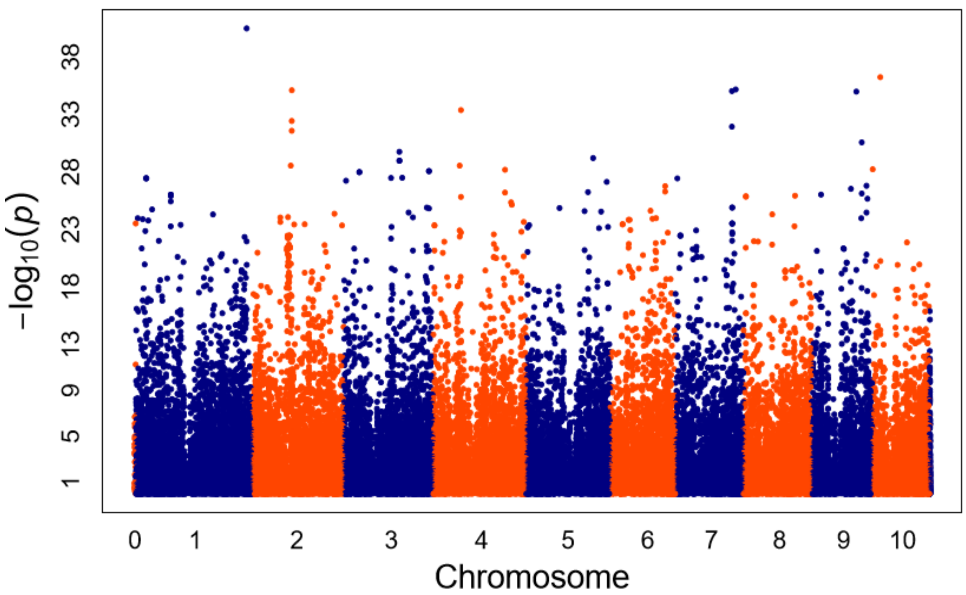


Figure 3.2: A genome-wide association study (GWAS) for stalk lodging in maize. A Manhattan plot of association results from binomial logistic regression model that included principal components representative of population structure as covariates. The –log$_{10}$ P Values are plotted on the y axis, and the physical location in the genome is plotted on the x-axis. Orange and blue dots represent the 55K SNPs used in this model.

Model 2 Results

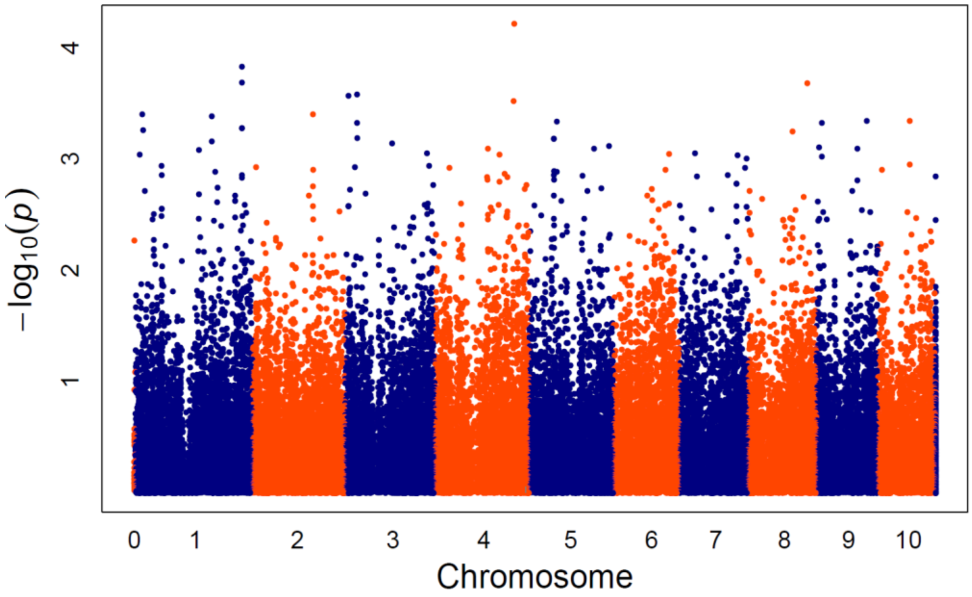


Figure 3.3: a genome-wide association study (GWAS) of best linear unbiased predictors (BLUPs) for stalk lodging in maize. A Manhattan plot of association results from unified mixed linear model. The –log$_{10}$ P Values are plotted on the y axis, and the physical location in the genome is plotted on the x-axis. Orange and blue dots represent the 55K SNPs used in this model.

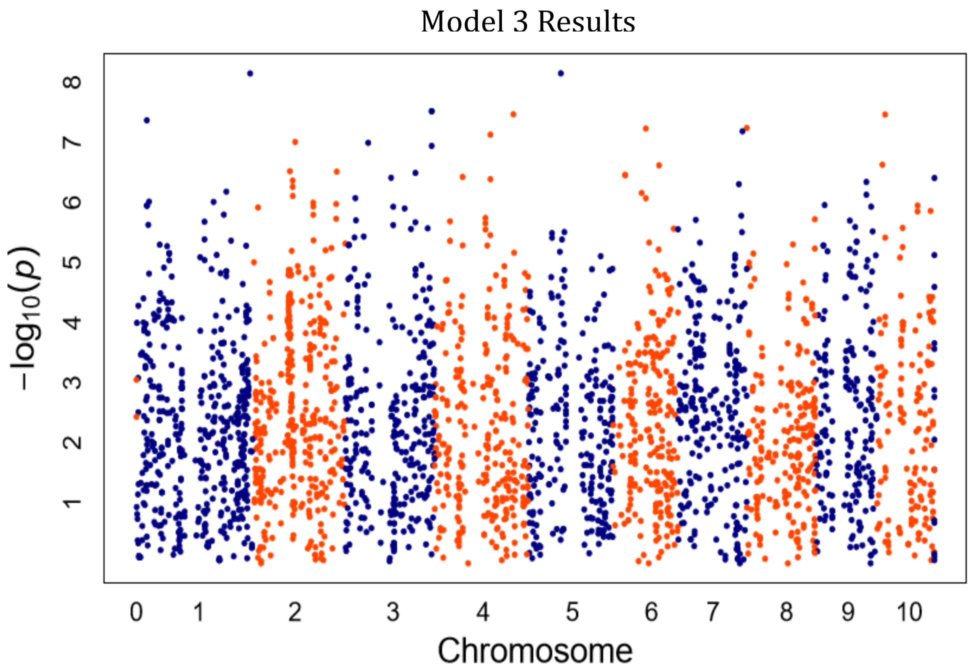


Figure 3.4: A genome-wide association study for stalk lodging in maize. A manhattan plot of association results from a binomial mixed logistic regression model. The –log10 P Values are plotted on the y axis, and the physical location in the genome is plotted on the x-axis. Orange and blue dots represent the top 2,794 significant SNPs at 5% FDR from Model 1 (No SNPs from Model 2 were included as there were no significant SNPs at 5% FDR).

Peak SNPs of Interest				
Type of Region identified	Chr	Location in Literature	Location in Model 3	Notes
Marker	7	159.4 Mb	161.9 Mb, 155.8, Mb 164.9 Mb	Three most significant SNPs on Chr 7
qRPR2 QTL	2	236.4-237.0 Mb	236.8 Mb	14th most significant SNP on Chr 2
qRPR3-1 QTL	3	81.1 Mb-184.7	181.7 Mb, 182.0 Mb	92nd and 98th most significant SNP On Chr 3
Marker	1	NA	290.85 Mb	Most significant SNP on chromosome 1
Marker	5	NA	83.39 Mb	Most significant SNP on chromsome 5

Table 3.1: Table Representing SNPs identified in Model 3, and SNPs that have been previously identified to be associated with stalk strength and rind penetrometer resistance. Peak SNPs on Chromosome 7 were in the same location as the most robust marker association with RPR (Pieffer et al., 2013). Additionally, Model 3 was able to identify two significant SNPs in the BP region of Maize Stalk Strength QTL identified in Li et al., 2014, Flint-Garcia et al., 2003, and Hu et al., 2012. The top two significant SNPs from this model were also identified. Locations were determined using the B73 RefGen_v2 coordinates.

LD Decay Plot of Region Surrounding Peak SNP on Chromosome 7

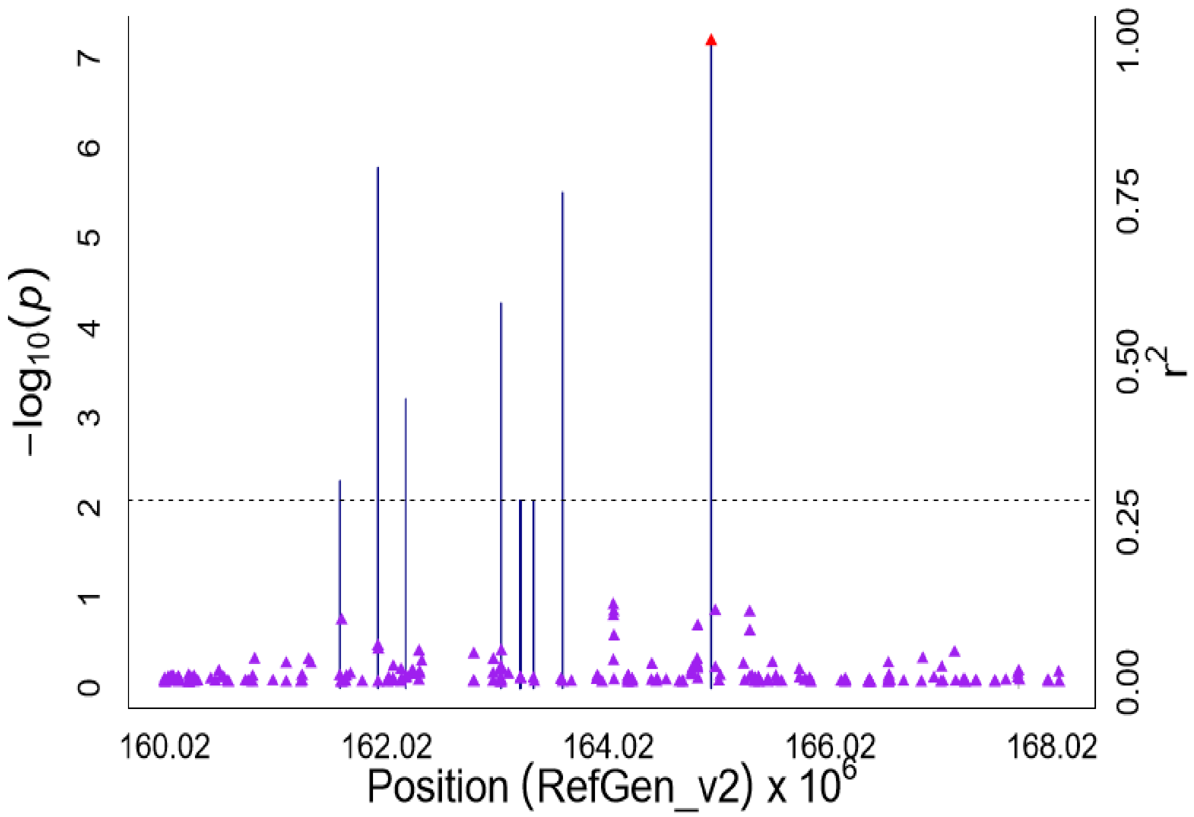


Figure 3.5: Scatterplot representing results of binomial mixed logistic regression model in genomic region surrounding the peak SNP on chromosome seven. The left-sided y-axis represents the –log10 P Values and the right-sided y-axis represents the r2 value (measure of LD). These values are plotted against the physical location in the genome on the x-axis. The blue vertical lines represent SNPs that were significant at 5% FDR in the 8MB region surrounding the peak SNP on chromosome seven and their values correspond the left-sided y-axis. The purple triangles are the r2 values of each SNP and the SNP of interest located at 164.9 Mb (denoted by red triangle), and their values correspond the right-sided y-axis. The dashed line represents the 5% FDR cutoff of the –log10 P-values.

Simulation Studies Evaluate the Efficacy of Multi-model Approach to GWAS on Binomial Trait

We conducted a simulation study to further evaluate how well this multi-model GWAS approach performs. Additionally, the use of simulated data allowed us to create different settings in which we could experiment with changing variables, and evaluate the effect of these variables on model performance. As we conducted Models 1, 2, and 3 we also evaluated how these variable changes affected our results. Variables that we evaluated include intercept (which translates to a baseline probability of stalk lodging) and stand count.

Settings were simulated using the Goodman-Buckler diversity panel (Flint-Garcia *et al.* 2005). To investigate the effect of intercept on model performance, four settings were created, each with a stand count of 10, with the different intercepts of 0,1,3, and 5. The intercept is representative of the baseline trait value (i.e. a higher intercept will increase probability that a plant will be simulated as lodged). Three additional settings were simulated, this time varying stand count between 15, 20, and 25, with a constant intercept of 0. The SNP chosen to be the additive QTN remained constant across simulation settings 1-7. These settings are presented in tabular form in Table 3.2.

At each of these settings, 100 traits were simulated and Model 1 was fitted at each of the 55K SNPs that had MAF < 0.05. The results of all 100 traits of for each setting were compiled, and the top 100 SNPS from each trait were extracted (maximum 10,000 SNPs per setting). For each setting, the proportion of times each of these top SNPs were identified as significant was plotted against their physical location. Figure 3.6 compares the results of Model 1 for settings where the intercept varies (Setting 1-4). From these graphics, it appears that as the intercept increases, SNPs nearby the QTN are identified a lower proportion of times. Interestingly, it also appears that with an intercept of $\beta_o = 3$, SNPs near the QTN are still identified a relatively high proportion of times (albeit a lower proportion than the intercept values closer to zero), whereas SNPs not nearby are being found significant a very low proportion of times (lower than the lesser intercepts). This may indicate a trade-off between the intercept value and the number of false positives, as well as a trade-off between the intercept value and the proportion of times a QTN is correctly identified. Figure 3.7 compares the results of Model 1 when the stand count varies (Settings 1, 5-7). From these graphs there does not appear to be a relationship between stand count and the ability to detect the QTN. Additionally, we did not observe a difference between the results of Model 1 and the results of Model 2 (Figure 3.8). Model 3 was unable to be fit with these data as the model failed to converge in SAS as a result of the relatedness matrix not being positive definite.

Setting	Intercept β_o	Baseline Probability	Stand Count	QTN	Minor Allele Frequency	Chr.	Additive Effect Size
1	0	0.50	10	PHM4757.14	0.30	8	0.9
2	1	0.73	10	PHM4757.14	0.30	8	0.9
3	3	0.95	10	PHM4757.14	0.30	8	0.9
4	5	0.99	10	PHM4757.14	0.30	8	0.9
5	0	0.50	15	PHM4757.14	0.30	8	0.9
6	0	0.50	20	PHM4757.14	0.30	8	0.9
7	0	0.50	25	PHM4757.14	0.30	8	0.9

Table 3.2: Table describing the components of each simulation setting that fit to Models 1,2,and 3.

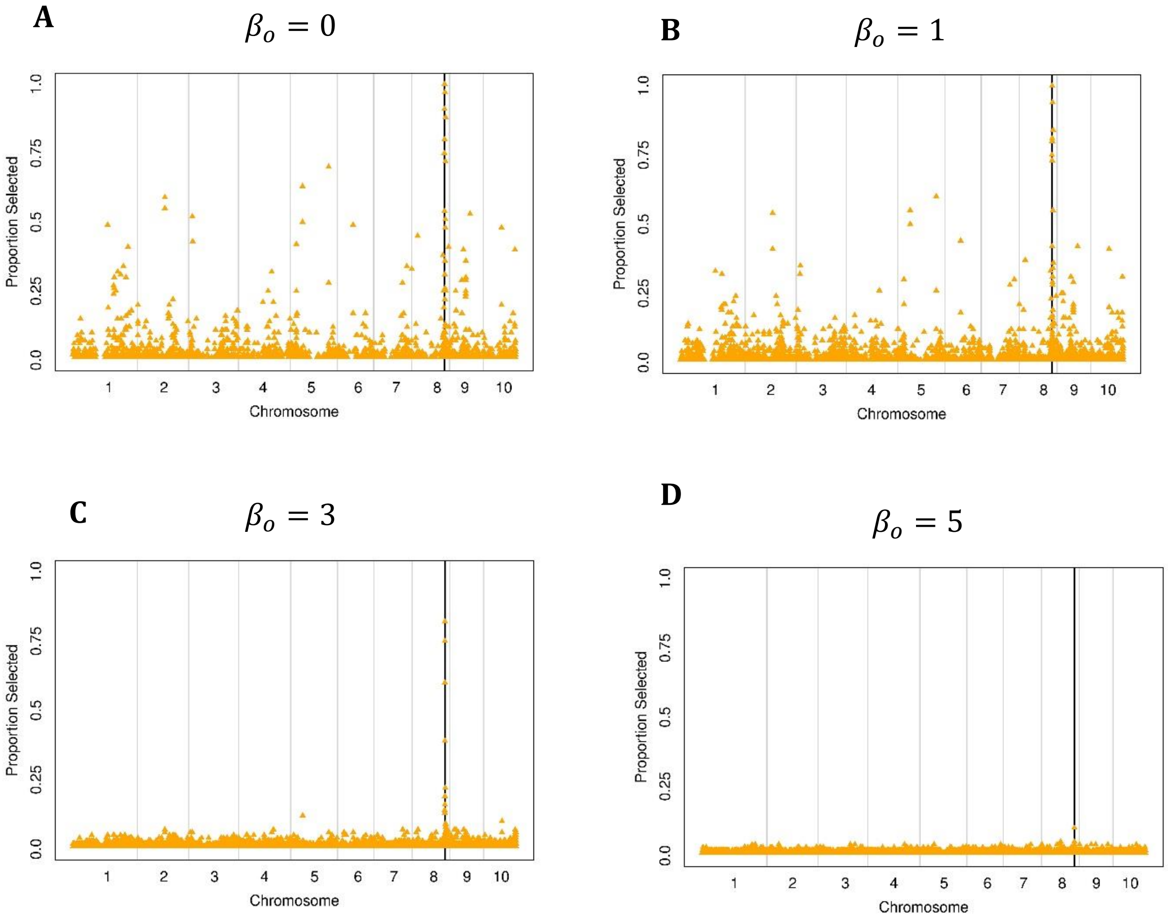


Figure 3.6: Model 1 results of simulation study of one large-effect (0.9) additive QTN PHM4757.14 with a stand count of 10 plants per plot. (A) Scatterplot of association results from top 100 SNPs of each of the 100 simulated traits obtained using Model 1 with an intercept β_o= 0. Proportion of times the QTN was located is on the y-axis, while the physical location of the SNP is on the x-axis. The black vertical line indicates the actual location of the simulated QTN (152.75 Mb, Chr 8). The triangles are representative of each of the SNPs used in this figure. (B) Scatterplot of association results from top 100 SNPs of each of the 100 simulated traits obtained using Model 1 with an intercept β_o= 1, depicted as in described in (A). (C)) Scatterplot of association results from top 100 SNPs of each of the 100 simulated traits obtained using Model 1 with an intercept β_o= 3, depicted as in described in (A).(D)) Scatterplot of association results from top 100 SNPs of each of the 100 simulated traits obtained using Model 1 with an intercept β_o= 5, depicted as in described in (A).

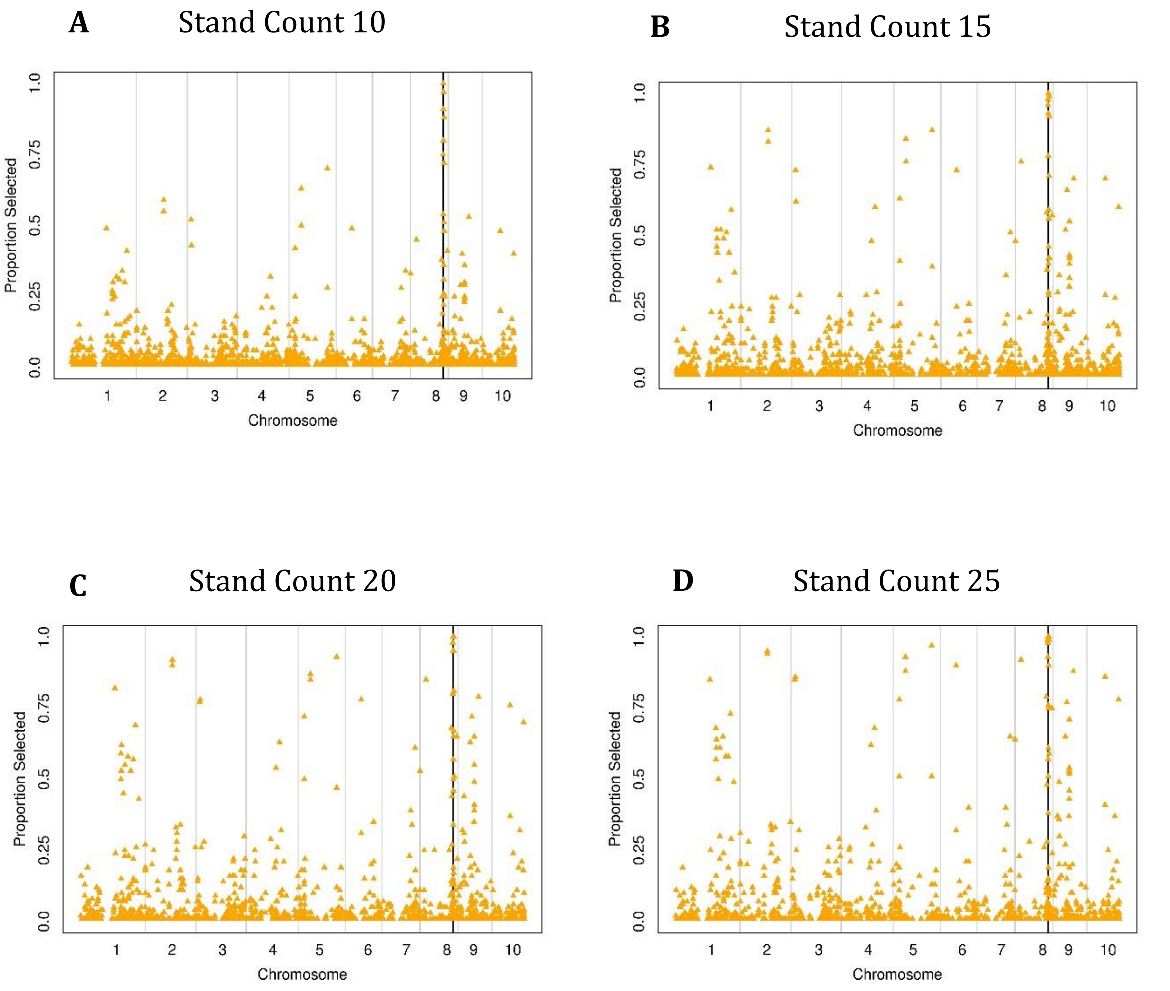


Figure 3.7: Model 1 results of simulation study of one large-effect (0.9) additive QTN PHM4757.14 with an intercept of $\beta_o = 0$. (A) Scatterplot of association results from top 100 SNPs of each of the 100 simulated traits obtained using Model 1 with a stand count of 10 plants per plot. Proportion of times the QTN was located is on the y-axis, while the physical location of the SNP is on the x-axis. The black vertical line indicates the actual location of the simulated QTN (152.75 Mb, Chr 8). The triangles are representative of each of the SNPs used in this figure. (B) Scatterplot of association results from top 100 SNPs of each of the 100 simulated traits obtained using Model 1 with a stand count of 15 plants per plot, depicted as in described in (A). (C)) Scatterplot of association results from top 100 SNPs of each of the 100 simulated traits obtained using Model 1 with a stand count of 20 plants per plot, depicted as in described in (A).(D)) Scatterplot of association results from top 100 SNPs of each of the 100 simulated traits obtained using Model 1 with a stand count of 25 plants per plot, depicted as in described in (A).

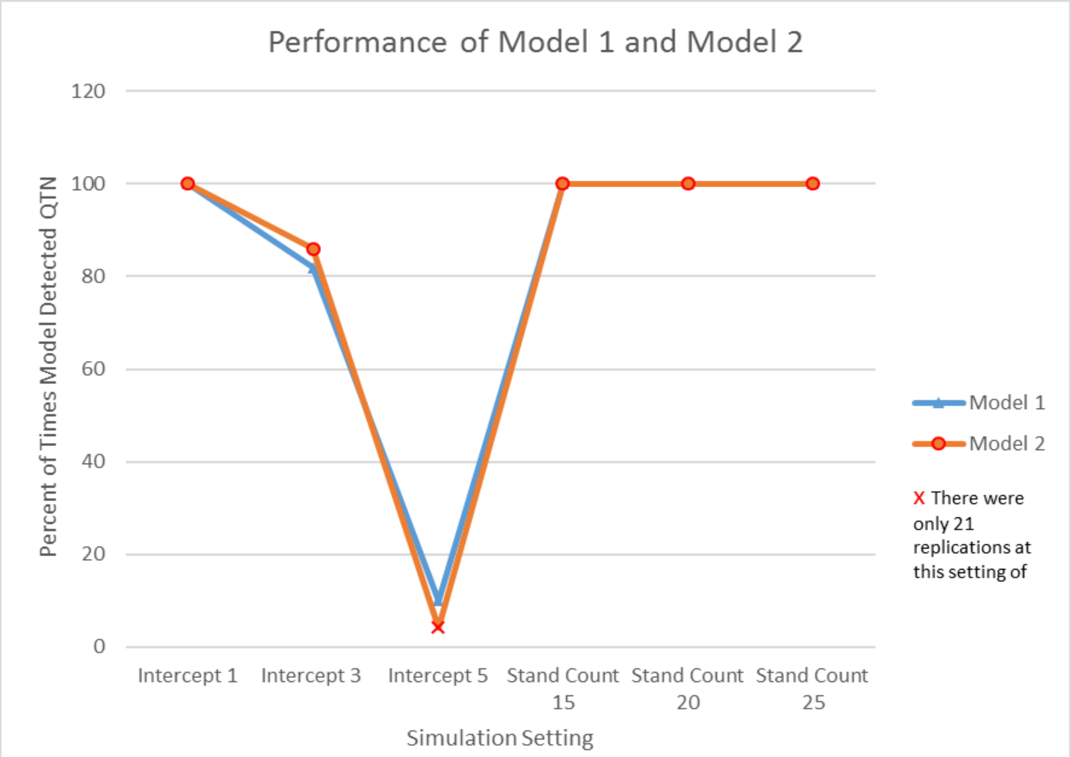


Figure 3.8: Line Graph comparing the performance of Model 1 and Model 2. The percentage of time the QTN was successfully detected (y-axis) was plotted against the setting that was ran (x-axis). The blue line indicates the results of Model 1 and the orange line indicates the results of Model 2. The red X indicates a setting that only had 21 replicates, whereas the rest of the settings had 100 replicates.

Goss's Wilt is Associated with an Increase in the Prevalence of Stalk Lodging in Maize

In 2017 two lodging trials were planted in the same location to evaluate the effect of Goss's Wilt on stalk lodging. One trial was inoculated with Cmn which led to Goss's wilt, while the other served as a control and received no inoculum. To evaluate the prevalence of lodging we looked at the proportion of plots lodged in both fields, which is presented in Figure 3.9. To see if these proportions of lodging were significantly different we conducted a two-sample test for the equality of proportions , which tests the null hypothesis of H_0: $p_{inoculated}$=$p_{non\text{-}inoculated}$ (Naranjo 2003) (Figure 3.10). Based on the P-value of 1.384e-08 obtained for this hypothesis test we reject the null hypothesis at $\alpha = 0.05$ and conclude that there is sufficient evidence to conclude that the proportion of lodging in the inoculated field differs from that of the control field. Based on these data we conclude that it is likely that the presence of Goss's wilt is associated with the likelihood of observing stalk lodging.

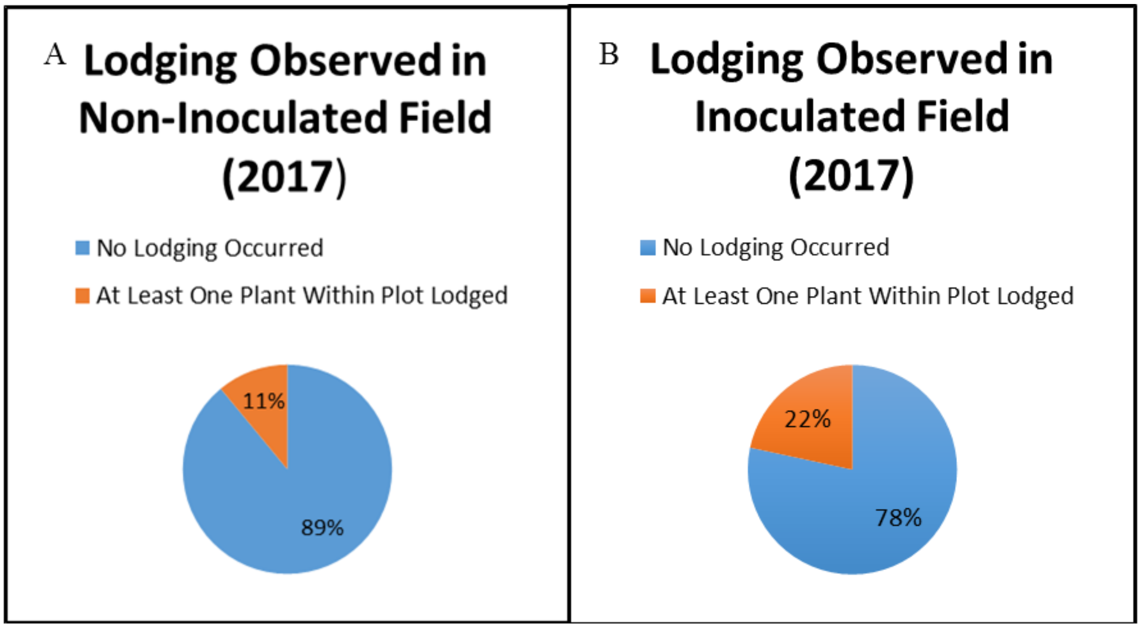
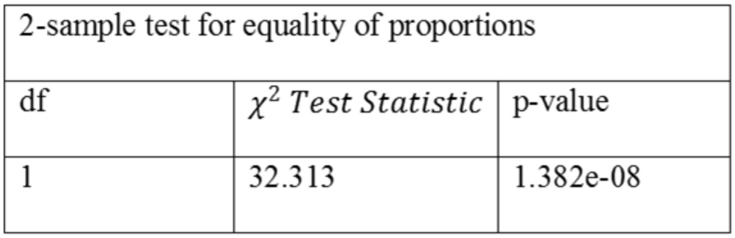


Figure 3.9: (A) Pie chart representing the percentage of plots that experienced lodging in the non-inoculated field (2017 field season). (B) Pie chart representing the percentage of plots that experienced lodging in the non-inoculated field (2017 field season)



2-sample test for equality of proportions		
df	χ^2 *Test Statistic*	p-value
1	32.313	1.382e-08

Figure 3.10: Results of 2-sample test for equality of proportions conducted in R using the prop.test function.

CHAPTER 4: DISCUSSION

Although GWAS has been readily used for over a decade, few GWAS models have been developed specifically for the analysis of binomially distributed agronomic traits such as stalk lodging in maize. With the advent of large-scale phenotyping projects, such as the Genomes to Fields Initiative where binary traits such as lodging are directly quantified, the need for such models will become increasingly necessary. This study rigorously evaluated the use of a logistic regression model to detect genomic signals associated with a binomial trait using real and simulated data. From these analyses we were able to demonstrate that it is possible to use logistic regression-based GWAS to identify genomic signals underlying binomial traits, however the properties of the tested phenotypic data influence the ability to correctly detect a QTL.

A challenge with analyzing stalk lodging is that the lodging must be induced by an external factor. Although most reported incidents of widespread stalk lodging are due to weather related factors, such as high winds (Nielsen and Colville 1988), we did not have the capabilities to simulate such conditions in the field. Therefore, for the purposes of this study we chose to phenotype lodging in a field trial that had been inoculated with Goss's wilt. This bacterial blight is known to affect the vascular system, resulting in symptoms such as stalk rot, potentially making plants more susceptible to stalk lodging (Nielsen and Colville 1988; Harveson 2011). Using the 2017 field data, we were able to compare the amount of lodging that occurred in an inoculated field and a control (non-inoculated) field. We found that a greater proportion of plots were lodged in the inoculated field and that there was a significant difference between the proportions of lodging in each field.

Due to missing observations, the association analyses (conducted on 2016 data only) were conducted with 86% of the phenotyped taxa which reduced our power to detect marker-trait associations (Long and Langley 1999). Another factor influencing the power of the association test is heritability. The narrow-sense heritability of lodging was calculated using GAPIT (Lipka *et al.* 2012) to be 0.09. It is important to note however that this result was calculated using a model where assumptions are violated, meaning the value reported may be biased. Accordingly, across the two replications of each taxa we observed low repeatability, meaning that across replicates of taxa we did not see the same prevalence of lodging (correlation=0.249). This suggests that a large proportion of the phenotypic variability can be attributed to environmental factors, rather than the corresponding genetic variation. Due to these limitations, we focused our efforts on evaluating the analytical pipelines developed in this project, rather than dissecting the genetic architecture of stalk lodging.

The primary objective of this experiment was to develop a model that could be used to perform GWAS on binomial traits using the computational bandwidth that is available on a typical laptop or desktop. Accordingly, we proposed a three-model approach for analyzing the stalk lodging data from the 2016 field trial. Briefly, Model 1 is a logistic regression model with principal components as covariates, Model 2 is a unified mixed linear model that accounts for population structure and relatedness, and Model 3 is a mixed logistic regression model that accounts for population structure and covariates. In Model 1, 24,210 SNPs were found to be significant at 5% FDR. We hypothesized that this large amount of significant results could be attributed to various factors, including spurious associations due to relatedness and multiple genes underlying the trait. In Model 2, no SNPs were found to be significantly associated with stalk lodging at an FDR of 5%. It is possible that due to the underlying assumptions of the unified mixed linear model (normality, equal variances, independence), this non-significance was the result of the violation of the model assumptions. In Model 3, 1,905 significant SNPs were identified to be significant at 5% FDR. One scenario that explains these results is that there might be many small effect loci associated with the stalk lodging trait. Although this is a possible explanation, the lower than expected amount of significant marker-trait associations from Model 3 may also be explained by a low power to detect associations as a result of the inherent properties of this data set.

Based on the results of Model 3, we decided to investigate the linkage disequilibrium (LD) surrounding our strongest signals. The most peak signals were not found in any area known to be associated with lodging. Additionally, there were no documented candidate genes in these regions. Therefore, we chose to focus on the region of chromosome 7 where we had peak associations with stalk lodging in locations similar to those identified in previous studies on RPR (Peiffer *et al.* 2013). We plotted the LD in terms of r^2 between the most significant SNP on chromosome 7 (ss196481136, 164,952,176 bp, p-value 6.46E-08) and the SNPs in the region of 160.0 Mb- 168.0 Mb of chromosome seven. Overall, there appears to be a high rate of LD decay in this genomic region, with LD measuring 0.11 within 35 kb of our SNP of interest. Of the genotyped SNPs in this region, the highest observed r^2 value was 0.12 at 164.06 Mb, which suggests that all genotyped SNPs in this region are in low LD with our SNP of interest. This result was unusual, as one might expect significant SNPs within the same region to be in LD with each other. A possible explanation for this result is that not all polymorphisms in the genome have been genotyped and that several ungenotyped markers in the surrounding genomic region may, in fact, be in LD with the markers that we analyzed. Regardless, this low level of LD could limit the ability of a genotyped marker to be in LD with the true casual mutation, bringing into question whether the significant associations on chromosome 7 is indicative of the genetic basis of stalk lodging.

Overall, the three-model approach demonstrated the ability to successfully fit a mixed logistic regression model with finite computing resources. The signals identified in regions associated with RPR and stalk strength reflect favorably on the ability of this model to accurately detect QTL associated with stalk lodging. However, without the availability of previously published literature, the GWAS results of Model 3 would be relatively uninformative of the genetic basis of stalk lodging, as highly significant results spanned the entire genome, making it difficult to identify a specific region of interest. It is possible that the power to successfully detect QTL was diminished by the protocols used to quantify stalk lodging, resulting in a large amount of significant results. As research going into accurately quantifying phenotypes that approximate stalk lodging continues to be developed and refined (e.g., the morphological study by Robertson et al. (2017)), we expect that we would have a greater ability to compare and contrast the genomic signals identified from a GWAS of stalk lodging directly (i.e., directly phenotyping lodging as done in this study) to those from studies that approximate lodging via a quantitative trait. The ability to detect genomic signals as a result of phenotyping lodging directly was further explored throughout the simulation study.

Real data have an extra element of uncertainty in that not all sources of genomic variation underlying a studied trait are known. This presents a challenge when evaluating the ability of our GWAS approach to identify QTN. To further evaluate the three-model approach we took advantage of the certainty of simulated data, simulated stalk lodging data from the Goodman-Buckler diversity panel (Flint-Garcia *et al.* 2005). Within the context of this simulation we were able to control the number of replications, the stand count of each plot, the intercept of the model (which translates to a baseline probability of a plant lodging), the SNP(s) assigned to be QTN, and the additive effect size of the QTN(s). The power of an association test is affected by many factors, one of them being the allele frequency of the QTL. Within the context of the simulation study, we found that QTN with higher minor allele frequency increased the resolution of our results. Therefore, when comparing different variables (such as stand count), we kept the SNP that was chosen to be the QTN (MAF of 0.30) constant over compared settings.

Overall there was little variability in the ability of Model 1 and Model 2 to detect the simulated QTN. We were unable to fit these data to Model 3 as the model failed to converge as a result of the relatedness matrix not being positive definite. It is possible that this error occurred due to not enough variation in the response variable. One possible explanation for similarities in performance between Model 1 and Model 2 is that Model 2 may have had enough power to successfully detect the QTN despite model assumption violations that may lead to empirical type I error rates that differ substantially from α. Another possible explanation for this result is a previous study (Pirinen *et al.* 2013) showed that linear models can be approximated by logistic regression models when the effect size of the genetic variant is small, and population structure has been removed (Chen *et al.* 2016). In the case of these models the population structure has been removed via the incorporation of PCs. However, the additive effect size was quite large in almost all cases, which may affect the legitimacy of this argument. Regardless of these similar outcomes, we were able to prove that a logistic regression model could accurately detect QTN in certain conditions.

When we simulated these binomially distributed traits, we were able to specify the intercept that was used in simulating the lodging phenotypes as well as the stand count for each plot. Changing the intercept among settings directly corresponds to the baseline probability of observing stalk lodging in a maize plant. For instance, a population simulated with an intercept of 0 had a baseline probability of 0.5 of lodging, while a population simulated with an intercept of 5 had a baseline probability of 0.99 of lodging. Essentially, a more extreme value of the intercept would translate to either a very high rate of lodging, or in the cases of an extremely negative intercept, a very low rate of lodging. Moreover, since the variance of a binomial random variable ($n\pi(1-\pi)$) is maximized when the probability of a success is π = 0.50, we both theoretically expected and empirically observed lesser variability in the simulated phenotypes at larger intercept parameter values. In examining the simulated phenotypes it is apparent that as the absolute value of β_o increases the simulated phenotypes became less variable, and consequently the rate at which a QTN is successfully detected decreases. In changing the stand count between settings we found no notable difference in the proportion of times a QTN was successfully detected. Considering these findings in the context of the 2016 field data, it is possible that our model's ability to accurately detect QTL was compromised, as in the 2016 field season we observed an overall low rate of lodging. Consequently, it is possible that the baseline probability of lodging in the field was a value considerably different from 0.5. If this was, in fact, the case, then the inability of our model to detect QTL may have been exacerbated by an intercept value that is far removed 0. To investigate this hypothesis we calculated the intercept of the 2016 field data to be -2.3. This value could provide a possible explanation for the results obtained when the 2016 lodging data was fit to Model 1 and Model 3.

Many plots (0.60 of all plots (NAs removed)) from the 2016 field season experienced no lodging. This low rate of phenotypic variability could explain the nature of the results of Models 1,2,3 for the 2016 field data. Remedying this issue of low variability in the future will be difficult; we are limited in our ability to induce lodging at the rate needed to achieve the ideal phenotypic variability. In general, this is an issue that will need to be taken into consideration for future analysis of binomial traits with a logistic regression model.

This research used stalk lodging as case study to further explore the application of logistic regression models for GWAS. To address the computational burden associated with logistic regression models and random effects we developed a three-model approach to conduct mixed logistic regression GWAS. This pipeline was developed with the intention of reducing the number of mixed logistic regression models that must be fit by identifying SNPs most likely to be associated with lodging in the first two models. From this approach, we demonstrated that logistic regression GWAS could successfully detect QTL under certain specified conditions. Most notably, the baseline trait value of the data set appeared to greatly affect the accuracy of the GWAS results. This point of discussion should be taken into consideration when future experiments on binomial traits are designed, as extreme baseline trait values will negatively affect one’s ability to detect regions of the genome associated with a trait of interest. In conclusion, mixed binomial logistic regression is a viable option for QTL discovery, however computational limitations and baseline trait values need to be taken into consideration when using this methodology.

REFERENCES

Agresti, A., and M. Kateri, 2011 Categorical Data Analysis, pp. 206–208 in *International Encyclopedia of Statistical Science*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Albrecht, K. A., M. J. Martin, W. A. Russel, W. F. Wedin, and D. R. Buxton, 1986 Chemical and in Vitro Digestible Dry Matter Composition of Maize Stalks after Selection for Stalk Strength and Stalk-Rot Resistance1. Crop Sci. 26: 1051.

Annual Report 2015 | U.S. Grains Council.

Barton, B., and S. Clark, 2014 Water &amp; Climate Risks Facing U.S. Corn Production.:

Chakravarti, A., 2014 Linkage Disequilibrium, in *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, Chichester, UK.

Chen, H., C. Wang, M. P. Conomos, A. M. Stilp, Z. Li *et al.*, 2016 Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models.

Cook, J. P., M. D. McMullen, J. B. Holland, F. Tian, P. Bradbury *et al.*, 2012 Genetic Architecture of Maize Kernel Composition in the Nested Association Mapping and Inbred Association Panels. Plant Physiol. 158:.

Cooper, J., B. Rice, E. Shenstone, A. E. Lipka, and T. M. Jamann, 2017 Genome-wide analysis and genomic selection for Goss's wilt in maize. G3 Genes, Genomes, Genet.

Flint-Garcia, S. A., M. D. McMullen, and L. L. Darrah, 2003 Genetic Relationship of Stalk Strength and Ear Height in Maize. Crop Sci. 43: 23–31.

Flint-Garcia, S. A., A.-C. Thuillet, J. Yu, G. Pressoir, S. M. Romero *et al.*, 2005 Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. 44: 1054–64.

Harveson, R., 2011 A Historical Perspective of Goss' Wilt in Nebraska | Plant Pathology Department | University of Nebraska–Lincoln.

Hoban, S., G. Bertorelle, and O. E. Gaggiotti, 2012 Computer simulations: tools for population and evolutionary genetics. Nat. Rev. Genet. 13: 110.

Hu, H., Y. Meng, H. Wang, H. Liu, and S. Chen, 2012 Identifying quantitative trait loci and determining closely related stalk traits for rind penetrometer resistance in a high-oil maize population. Theor. Appl. Genet. 124: 1439–1447.

Hunter, D. J., P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager *et al.*, 2007 A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat. Genet. 39: 870–4.

Janvis, S., W. Guthrie, E. Berry, and R. Clark, 1984 The relationship between second-generation European corn borers and stalk rot fungi in maize hybrids. Maydica.

Jirak-Peterson, J. C., and P. D. Esker, 2011 Tillage, Crop Rotation, and Hybrid Effects on Residue and Corn Anthracnose Occurrence in Wisconsin. Plant Dis. 95: 601–610.

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.

Kiernan, K., J. Tao, and P. Gibbs, 2012 Tips and Strategies for Mixed Modeling with SAS/STAT ® Procedures.

Lee, M., N. Sharopova, W. D. Beavis, D. Grant, M. Katt *et al.*, 2002 Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. Plant Mol. Biol. 48: 453–461.

Li, M., X. Liu, P. Bradbury, J. Yu, Y.-M. Zhang *et al.*, 2014 Enrichment of statistical power for genome-wide association studies. BMC Biol. 12: 73.

Li, Y., L. Si, Y. Zhai, Y. Hu, Z. Hu *et al.*, 2016 Genome-wide association study identifies 8p21.3 associated with persistent hepatitis B virus infection among Chinese. Nat. Commun. 7: 11664.

Li, K., J. Yan, J. Li, and X. Yang, 2014 Genetic architecture of rind penetrometer resistance in two maize recombinant inbred line populations. BMC Plant Biol. 14: 152.

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. G3 Genes, Genomes, Genet. 3:.

Lipka, A. E., C. B. Kandianis, M. E. Hudson, J. Yu, J. Drnevich *et al.*, 2015 From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr. Opin. Plant Biol. 24: 110–118.

Lipka, A. E., F. Tian, Q. Wang, J. Peiffer, M. Li *et al.*, 2012 GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397–2399.

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. Methods 8: 833–835.

Loiselle, B. A., V. L. Sork, J. Nason, and C. Graham, 1995 Spatial Genetic Structure of a Tropical Understory Shrub, Psychotria officinalis (Rubiaceae). Am. J. Bot. 82: 1420.

Long, A. D., and C. H. Langley, 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 9: 720–731.

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: 737–40.

Mendiburu, F. de, 2017 Statistical Procedures for Agricultural Research.

Munkvold, G. P., and R. L. Hellmich, 1999 Genetically modified insect resistant corn:

Implications for disease management. APSnet Featur. Artic.

Naranjo, J., 2003 Testing for Equality of Proportions Between 2 Samples. Dep. Stat. West. Michigan Univ.

Nielsen, B., and D. Colville, 1988 Stalk Lodging in Corn: Guidelines for Preventive Management. Agron. Guid.

Ogura, T., and W. Busch, 2015 From phenotypes to causal sequences: using genome wide association studies to dissect the sequence basis for variation of plant development. Curr. Opin. Plant Biol. 23: 98–108.

Ostlie, K. R., W. D. Hutchison, and R. L. Hellmich, 1997 Bt corn and European corn borer : Pest management : Corn Production : University of Minnesota Extension. Univ. Minnesota, Ext.

Ott, R. L., and M. T. Longnecke, 2008 *An Introduction to Statistical Methods and Data Analysis*.

Peiffer, J. A., S. A. Flint-Garcia, N. De Leon, M. D. McMullen, S. M. Kaeppler *et al.*, 2013 The Genetic Architecture of Maize Stalk Strength (I. De Smet, Ed.). PLoS One 8: e67066.

Pirinen, M., P. Donnelly, and C. C. A. Spencer, 2013 Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Ann. Appl. Stat. 7: 369–90.

R Core Team, 2017 R Core Team (2017). R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria. URL http//www.R-project.org/. R Foundation for Statistical Computing.

Robertson, D. J., M. Julias, S. Y. Lee, and D. D. Cook, 2017 Maize Stalk Lodging: Morphological Determinants of Stalk Strength. Crop Sci. 57: 926.

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:

R55.

Shi, D. Y., L. YH, Z. JW, P. Liu, B. Zhao *et al.*, 2016 Effects of plant density and nitrogen rate on lodging-related stalk traits of summer maize. Plant, Soil Environ. 62: 299–306.

Spain, S. L., and J. C. Barrett, 2015 Strategies for fine-mapping complex traits. Hum. Mol. Genet. 24: R111-9.

Thomison, P., and P. Paul, 2012 C.O.R.N. Newsletter 2012-30 | Agronomic Crops Network.

USDA Economic Research Service - Background.

Wilcoxson, R. D., B. Skovmand, and A. H. Atif, 1975 Evaluation of wheat cultivars for ability to retard development of stem rust. Ann. Appl. Biol. 80: 275–281.

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

Zhang, Z., 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: 355–360.

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